



# 自然语言处理导论

张奇 桂韬 黄萱菁

2023 年 2 月 19 日



# 前言

时光荏苒，自 2003 年我师从吴立德教授，开启自然语言处理学习与研究之路，转眼已近二十载春秋。回想当年第一次听到自然语言处理的目标——“让机器理解人类语言”时的兴奋，第一次看到《大规模中文文本处理》教材时的茫然，仿佛黄萱菁教授对我研究生入学的电话面试就在昨天，每周与吴老师固定交流前的紧张感依然清晰。从求学到任教，深刻感受到自然语言处理的快速发展，从基于特征的统计机器学习方法到深度神经网络模型，再到大规模预训练方法，自然语言处理研究范式的更新迭代速度也在不断加快。在本科生和研究生的自然语言处理课程教学过程中，虽然通过不断补充国际国内的近期研究进展，将最新的理论和方法通过课件和面授的形式介绍给同学们，但是系统全面的书籍仍然是不可或缺的重要资料。于是，自 2020 年起与黄萱菁教授和桂韬研究员一起开始着手本书的准备，在经过几十次的讨论和大纲和结构反复修改后，自 2021 年暑假起开始了本书的写作。2022 年本书入选复旦大学七大系列百本精品教材项目和复旦大学研究生规划系列教材项目，进一步督促我们加快进度。从规划到完成，历时近三年之久，这本拙作终于完成。

自然语言处理研究融合了语言学、计算机科学、机器学习等多学科内容。自然语言处理的研究内容从语言单位上划分涵盖字、词、短语、句子、段落到篇章等不同粒度，从类型上划分包含处理、理解到生成等不同种类。研究内容涉及的知识点多且复杂。自然语言研究大体经历了 20 世纪 50 年代末到 80 年代基于规则的研究范式，20 世纪 90 年代到 2010 年前基于特征的统计机器学习研究范式，2010 年到 2018 年基于深度神经网络研究范式，以及 2018 年至今基于大规模和超大规模预训练模型的研究范式等几个阶段。每个阶段的研究范式都有非常鲜明的特点，但也与机器学习研究有着十分紧密的联系。自然语言处理研究内容繁杂以及与机器学习方法交织导致本书的写作难度远超最初的预想。由于很多自然语言处理任务都转换为了机器学习问题，因此很多机器学习算法可以应用于多个自然语言处理任务。比如，条件随机场模型可用于中文分词，也可以用于词性标注，还可以用于命名实体识别。在这些任务中，条件随机场模型也都取得了不错的效果。我们花费大量的时间讨论如何设计本书的结构，在避免重复的同时能够使得读者更好的了解更多的自然语言处理研究内容和算法。

本书的目标是介绍自然语言处理的基本任务和主要处理算法。为了能够让读者更好的了解任务的特性和算法设计的主要目标，在介绍每个自然语言处理任务时，除了介绍任务的目标之外，还会介绍该任务所涉及的主要语言学理论知识以及任务的主要难点。针对自然语言处理历史发展过

程中的不同研究范式，选择不同类型的算法进行介绍。因此，大多数情况下每个章节分为如下几个部分：任务概述、相关语言学知识、基于规则的方法、基于特征的机器学习方法、基于深度神经网络的算法，任务评测指标和常见数据集合。针对同一机器学习算法可以应用于不同任务的问题，为了避免重复，我们在不同任务的介绍中选择同一类别的不同机器学习算法进行介绍，并说明该算法还可以应用于哪些任务，以及该类型的任务应该采用哪种类别的机器学习算法。尽量使得读者能够建立起自然语言处理任务和机器学习算法之间的关系，即如何将自然语言处理任务转化为机器学习问题，如何选择合适的机器学习算法，如何根据任务特性设计机器学习算法。希望读者通过本书的阅读能够了解不同任务的难点和算法设计的要点，明确自然语言处理方法和机器学习算法之间的关系。虽然我们在这个问题上花费了大量的时间对本书的结构进行了设计，但是对于初学者来说这仍然是需要相当多的实践才能更深入领悟的部分。

本书主要面向高年级本科生和研究生作为自然语言处理相关课程教材使用，也可以作为对自然语言处理感兴趣的读者入门之用。在撰写过程中，尽量平衡学生的知识储备水平与内容完备性之间的关系。在内容选择上，主要针对计算机和人工智能领域学生的基础知识特点，因此语言学理论介绍略显单薄，语言学理论内容的选择上也偏重经典，对于不同语言学理论之间的关系以及最新的语言学前沿研究介绍较为缺乏。对于有志于从事自然语言处理研究的读者，可以进一步的拓展语言学相关领域的阅读。由于很多自然处理任务都转化为了机器学习问题，采用各类型的统计机器学习算法进行解决，因此本书的介绍必然涉及到机器学习中的模型选择、学习准则设定以及优化算法使用等问题。本书在相关算法介绍时，以如何将特定自然语言处理任务转化为机器学习问题为重点，对于优化算法选择等基础问题需要读者参考机器学习和深度学习书籍。也建议读者在阅读本书前，系统地学习机器学习和深度学习的相关课程。

在内容组织方面，本书主要包含基础技术、核心技术以及模型分析三个部分。基础技术部分主要介绍自然语言处理的基础任务和底层技术，主要包含词汇处理、句法分析、语义分析、篇章分析和语言模型。核心技术部分主要介绍自然语言处理应用任务和相关技术，主要包括信息抽取、机器翻译、情感分析、文本摘要、知识图谱。模型分析部分主要介绍基于机器学习的自然语言处理模型的鲁棒性和可解释性问题。教学课时安排上，可以满足 32 学时到 56 学时的教学安排。模型稳健性和模型可解释性是近年来人工智能领域的研究热点，但是也涉及到各类自然语言处理任务和模型，需要读者花费更多时间在相关任务实践中学习。

本书的写作过程得到了众多专家和同学的大力支持和帮助。张翀博士、马若恬博士、周鑫博士、赵君博士、周杰博士、费子楚博士、邹易澄博士、王枭博士、郑锐博士分别为本书的第 4-13 章提供了部分基础素材。张奇撰写了第 1、2、3、4、5、6、10、11、12、13 章，桂韬副研究员撰写了第 7、8、9 章，吴苑斌副教授撰写了第 14 章，黄萱菁教授审阅了全书。尽管从本书的提纲结构讨论开始，我们就保持着最严肃认真的态度对待这项工作，但是越是临近本书付梓之际，越是惶恐不安。自然语言处理涉及文理工多学科交叉，研究内容又极其繁杂，受限于我们的认知水平和所从事的研究工作的局限，对其中一些任务和工作的细节理解可能存在不少错误，也恳请专家、

读者批评指正，您的意见对我们非常重要。

最后，衷心地感谢我的导师吴立德教授，他不仅带领我走进了自然语言处理之门，更重要的是他严谨求真的治学态度和高瞻远瞩的研究视野使我受益终身。感激我的家人给予的支持，为了能够提供我专心的写作环境，他们承担了几乎全部孩子教育、家务等繁琐而辛苦的事务，才使我能够完成本书的写作。他们默默地牺牲了自己的休息时间甚至是事业，才让我可以任性地追求自己的梦想。欲报之德，昊天罔极。

张奇

2023年1月于复旦曦园

# 数学符号

## 数与数组

$\alpha$	标量
$\alpha$	向量
$A$	矩阵
$\mathbf{A}$	张量
$I_n$	$n$ 行 $n$ 列单位矩阵
$v_w$	单词 $w$ 的分布式向量表示
$e_w$	单词 $w$ 的独热向量表示: $[0,0,\dots,1,0,\dots,0]$ , $w$ 下标处元素为 1

## 索引

$\alpha_i$	向量 $\alpha$ 中索引 $i$ 处的元素
$\alpha_{-i}$	向量 $\alpha$ 中除索引 $i$ 之外的元素
$w_{i:j}$	序列 $w$ 中从第 $i$ 个元素到第 $j$ 个元素组成的片段或子序列
$A_{ij}$	矩阵 $A$ 中第 $i$ 行、第 $j$ 列处的元素
$A_{i:}$	矩阵 $A$ 中第 $i$ 行
$A_{:j}$	矩阵 $A$ 中第 $j$ 列
$A_{ijk}$	三维张量 $\mathbf{A}$ 中索引为 $(i, j, k)$ 处元素
$\mathbf{A}_{::i}$	三维张量 $\mathbf{A}$ 中的一个二维切片

## 集合

$\mathbb{A}$	集合
$\mathbb{R}$	实数集
$\mathbb{C}$	复数集
$\{0, 1, \dots, n\}$	含 0 和 $n$ 的正整数的集合
$[a, b]$	$a$ 到 $b$ 的实数闭区间
$(a, b]$	$a$ 到 $b$ 的实数左开右闭区间

## 线性代数

$\mathbf{A}^\top$	矩阵 $\mathbf{A}$ 的转置
$\mathbf{A} \odot \mathbf{B}$	矩阵 $\mathbf{A}$ 与矩阵 $\mathbf{B}$ 的 Hadamard 乘积
$\det \mathbf{A}^\top$	矩阵 $\mathbf{A}$ 的行列式
$[x; y]$	向量 $x$ 与 $y$ 的拼接
$[\mathbf{U}; \mathbf{V}]$	矩阵 $\mathbf{A}$ 与 $\mathbf{V}$ 沿行向量拼接
$x \cdot y$ 或 $x^\top y$	向量 $x$ 与 $y$ 的点积

## 微积分

$\frac{dy}{dx}$	$y$ 对 $x$ 的导数
$\frac{\partial y}{\partial x}$	$y$ 对 $x$ 的偏导数
$\nabla_{\mathbf{x}} y$	$y$ 对向量 $\mathbf{x}$ 的梯度
$\nabla_{\mathbf{X}} y$	$y$ 对矩阵 $\mathbf{X}$ 的梯度
$\nabla_{\mathbf{x}} y$	$y$ 对张量 $\mathbf{X}$ 的梯度

## 概率与信息论

$a \perp b$	随机变量 $a$ 与 $b$ 独立
$a \perp b \mid c$	随机变量 $a$ 与 $b$ 关于 $c$ 条件独立
$P(a)$	离散变量概率分布
$p(a)$	连续变量概率分布
$a \sim P$	随机变量 $a$ 服从分布 $P$
$\mathbb{E}_{x \sim P}(f(x))$ 或 $\mathbb{E}(f(x))$	$f(x)$ 在分布 $P(x)$ 下的期望
$\text{Var}(f(x))$	$f(x)$ 在分布 $P(x)$ 下的方差
$\text{Cov}(f(x), g(x))$	$f(x)$ 与 $g(x)$ 在分布 $P(x)$ 下的协方差
$H(f(x))$	随机变量 $x$ 的信息熵
$D_{KL}(P \parallel Q)$	概率分布 $P$ 与 $Q$ 的 KL 散度
$\mathcal{N}(\mu, \Sigma)$	均值为 $\mu$ 、协方差为 $\Sigma$ 的高斯分布

## 数据与概率分布

$\mathbb{X}$ 或 $\mathbb{D}$	数据集
$x^{(i)}$	数据集中第 $i$ 个样本（输入）
$y^{(i)}$ 或 $y^{(i)}$	第 $i$ 个样本 $x^{(i)}$ 的标签（输出）

## 函数

$f : \mathcal{A} \longrightarrow \mathcal{B}$	由定义域 $\mathcal{A}$ 到值域 $\mathcal{B}$ 的函数（映射） $f$
$f \circ g$	$f$ 与 $g$ 的复合函数
$f(x; \theta)$	由参数 $\theta$ 定义的关于 $x$ 的函数（也可以直接写作 $f(x)$ , 省略 $\theta$ ）
$\log x$	$x$ 的自然对数函数
$\sigma(x)$	Sigmoid 函数 $\frac{1}{1 + \exp(-x)}$
$\ x\ _p$	$x$ 的 $L^p$ 范数
$\ x\ $	$x$ 的 $L^2$ 范数
$1^{\text{condition}}$	条件指示函数：如果 condition 为真，则值为 1；否则值为 0

## 本书中常用写法

- 给定词表  $\mathbb{V}$ , 其大小为  $|\mathbb{V}|$
- 序列  $x = x_1, x_2, \dots, x_n$  中第  $i$  个单词  $x_i$  的词向量  $v_{x_i}$
- 损失函数  $\mathcal{L}$  为负对数似然函数： $\mathcal{L}(\theta) = -\sum_{(x,y)} \log P(y|x_1 \dots x_n)$
- 算法的空间复杂度为  $\mathcal{O}(mn)$

# 目 录

<b>1 終論 .....</b>	<b>1</b>
<b>1.1 自然語言處理基本概念 .....</b>	<b>1</b>
1.1.1 自然語言處理簡史 .....	2
1.1.2 自然語言處理的主要研究內容 .....	4
1.1.3 自然語言處理的主要難點 .....	6
<b>1.2 自然語言處理的基本范式 .....</b>	<b>10</b>
1.2.1 基于規則的方法 .....	10
1.2.2 基于機器學習的方法 .....	11
1.2.3 基于深度學習的方法 .....	13
1.2.4 基于大模型的方法 .....	14
<b>1.3 本書的內容安排 .....</b>	<b>15</b>
<b>2 詞彙分析 .....</b>	<b>17</b>
<b>2.1 语言中的词汇 .....</b>	<b>17</b>
2.1.1 词的形态学 .....	17
2.1.2 词的词性 .....	18
<b>2.2 词语规范化 .....</b>	<b>22</b>
2.2.1 词语切分 .....	22
2.2.2 词形还原 .....	23
2.2.3 词干提取 .....	23
<b>2.3 中文分词 .....</b>	<b>24</b>
2.3.1 中文分词概述 .....	24
2.3.2 基于最大匹配的中文分词 .....	27
2.3.3 基于线性链条件随机场的中文分词 .....	28
2.3.4 基于感知器的中文分词 .....	30
2.3.5 基于双向长短句记忆网络的中文分词 .....	33

2.3.6 中文分词评价方法 .....	35
2.3.7 中文分词语料库 .....	36
<b>2.4 词性标注 .....</b>	<b>37</b>
2.4.1 基于规则的词性标注 .....	37
2.4.2 基于隐马尔可夫模型的词性标注 .....	39
2.4.3 基于卷积神经网络的词性标注 .....	41
2.4.4 词性标注评价方法 .....	44
2.4.5 词性标注语料库 .....	44
<b>2.5 延伸阅读 .....</b>	<b>46</b>
<b>2.6 习题 .....</b>	<b>47</b>
<b>3 句法分析 .....</b>	<b>48</b>
<b>3.1 句法概述 .....</b>	<b>48</b>
3.1.1 成分语法理论概述 .....	49
3.1.2 依存语法理论概述 .....	51
<b>3.2 成分句法分析 .....</b>	<b>53</b>
3.2.1 基于上下文无关文法的成分句法分析 .....	54
3.2.2 基于概率上下文无关文法的成分句法分析 .....	61
3.2.3 成分句法分析评价方法 .....	69
<b>3.3 依存句法分析 .....</b>	<b>70</b>
3.3.1 基于图的依存句法分析 .....	72
3.3.2 基于神经网络的图依存句法分析 .....	77
3.3.3 基于转移的依存句法分析 .....	82
3.3.4 基于神经网络的转移依存句法分析 .....	84
3.3.5 依存句法分析评价方法 .....	87
<b>3.4 句法分析语料库 .....</b>	<b>88</b>
<b>3.5 延伸阅读 .....</b>	<b>91</b>
<b>3.6 习题 .....</b>	<b>92</b>
<b>4 语义分析 .....</b>	<b>94</b>
<b>4.1 语义学概述 .....</b>	<b>94</b>
4.1.1 词汇语义学 .....	95
4.1.2 句子语义学 .....	99

<b>4.2 语义表示</b>	101
4.2.1 谓词逻辑表示法	102
4.2.2 框架表示法	103
4.2.3 语义网表示法	105
<b>4.3 分布式表示</b>	107
4.3.1 单词分布式表示	107
4.3.2 句子分布式表示	118
4.3.3 篇章分布式表示	120
<b>4.4 词义消歧</b>	123
4.4.1 基于目标词上下文的词义消歧方法	123
4.4.2 基于词义释义匹配的词义消歧方法	126
4.4.3 基于词义知识增强预训练的消歧方法	130
4.4.4 词义消歧评价方法	132
4.4.5 词义消歧语料库	132
<b>4.5 语义角色标注</b>	136
4.5.1 基于句法树的语义角色标注方法	136
4.5.2 基于深度神经网络的语义角色标注	139
4.5.3 语义角色标注评价方法	144
4.5.4 语义角色标注语料库	144
<b>4.6 延伸阅读</b>	147
<b>4.7 习题</b>	148
<b>5 篇章分析</b>	149
<b>5.1 篇章理论概述</b>	149
5.1.1 篇章的衔接	150
5.1.2 篇章的连贯	152
5.1.3 篇章的结构	153
<b>5.2 话语分割</b>	157
5.2.1 基于词汇句法树的统计话语分割	158
5.2.2 基于循环神经网络的话语分割	159
<b>5.3 篇章结构分析</b>	161
5.3.1 修辞结构篇章分析	161
5.3.2 浅层篇章分析	165

5.4 指代消解 .....	171
5.4.1 基于表述对的指代消解 .....	172
5.4.2 基于表述排序的指代消解 .....	175
5.4.3 基于实体的指代消解 .....	179
5.5 延伸阅读 .....	183
5.6 习题 .....	184
<b>6 语言模型 .....</b>	<b>185</b>
6.1 语言模型概述 .....	185
6.2 n 元语言模型 .....	187
6.2.1 加法平滑 .....	188
6.2.2 古德-图灵估计法 .....	188
6.2.3 Katz 平滑 .....	189
6.2.4 平滑方法总结 .....	191
6.3 神经网络语言模型 .....	192
6.3.1 前馈神经网络语言模型 .....	192
6.3.2 循环神经网络语言模型 .....	193
6.4 预训练语言模型 .....	195
6.4.1 动态词向量算法 ELMo .....	195
6.4.2 生成式预训练语言模型 GPT .....	198
6.4.3 掩码预训练语言模型 BERT .....	200
6.4.4 序列到序列预训练语言模型 BART .....	203
6.4.5 预训练语言模型的应用 .....	205
6.5 大规模语言模型 .....	207
6.5.1 基础大模型训练 .....	209
6.5.2 指令微调 .....	211
6.5.3 人类反馈 .....	212
6.6 语言模型评价方法 .....	214
6.7 延伸阅读 .....	214
6.8 习题 .....	215
<b>7 信息抽取 .....</b>	<b>216</b>
7.1 信息抽取概述 .....	216
7.2 命名实体识别 .....	218
7.2.1 非嵌套命名实体识别 .....	219

7.2.2 嵌套命名实体识别 .....	228
7.2.3 多规范命名实体识别 .....	233
7.2.4 命名实体识别评价方法 .....	236
7.2.5 命名实体识别语料库 .....	237
<b>7.3 关系抽取 .....</b>	<b>238</b>
7.3.1 有监督关系抽取 .....	239
7.3.2 远程监督关系抽取 .....	244
7.3.3 开放关系抽取 .....	248
7.3.4 关系抽取评价方法 .....	252
7.3.5 关系抽取语料库 .....	253
<b>7.4 事件抽取 .....</b>	<b>254</b>
7.4.1 限定域事件抽取 .....	254
7.4.2 开放域事件抽取 .....	258
7.4.3 事件抽取评价方法 .....	262
7.4.4 事件抽取语料库 .....	264
<b>7.5 延伸阅读 .....</b>	<b>265</b>
<b>7.6 习题 .....</b>	<b>265</b>
<b>8 机器翻译 .....</b>	<b>267</b>
<b>8.1 机器翻译概述 .....</b>	<b>267</b>
8.1.1 机器翻译发展历程 .....	268
8.1.2 机器翻译现状与挑战 .....	269
<b>8.2 基于统计的机器翻译方法 .....</b>	<b>270</b>
8.2.1 任务定义与基本问题 .....	270
8.2.2 IBM 模型 I .....	274
8.2.3 IBM 模型 II .....	278
8.2.4 IBM 模型 III .....	279
8.2.5 IBM 模型 IV .....	280
8.2.6 IBM 模型 V .....	281
<b>8.3 基于神经网络的机器翻译方法 .....</b>	<b>282</b>
8.3.1 循环神经网络翻译模型 .....	283
8.3.2 卷积神经网络翻译模型 .....	286
8.3.3 自注意力神经网络翻译模型 .....	289

8.4 机器翻译语料库 .....	293
8.5 延伸阅读 .....	295
8.6 习题 .....	296
<b>9 情感分析 .....</b>	<b>297</b>
9.1 情感分析概述 .....	297
9.1.1 情感模型 .....	298
9.1.2 情感分析主要任务 .....	301
9.2 篇章级情感分析 .....	305
9.2.1 基于支持向量机的篇章级情感分析 .....	305
9.2.2 基于层次结构的篇章级情感分析 .....	308
9.2.3 篇章级情感分析语料库 .....	310
9.3 句子级情感分析 .....	312
9.3.1 基于词典的句子级情感分析 .....	312
9.3.2 基于递归神经张量网络的句子级情感分析 .....	313
9.3.3 基于情感知识增强预训练的句子级情感分析 .....	315
9.3.4 句子级情感分析语料库 .....	316
9.4 属性级情感分析 .....	317
9.4.1 情感信息抽取 .....	318
9.4.2 属性级情感分类 .....	323
9.4.3 属性级情感分析语料库 .....	333
9.5 延伸阅读 .....	335
9.6 习题 .....	336
<b>10 智能问答 .....</b>	<b>337</b>
10.1 智能问答概述 .....	337
10.1.1 智能问答发展历程 .....	338
10.1.2 智能问答主要类型 .....	339
10.2 阅读理解 .....	341
10.2.1 基于特征的阅读理解算法 .....	342
10.2.2 基于深度神经网络的阅读理解算法 .....	344
10.2.3 阅读理解语料库 .....	351
10.3 表格问答 .....	352
10.3.1 基于特征的表格问答方法 .....	353
10.3.2 基于深度学习的表格问答模型 .....	354

10.3.3 表格问答语料库 .....	355
<b>10.4 社区问答 .....</b>	<b>356</b>
10.4.1 基于特征的语义匹配 .....	357
10.4.2 基于深度神经网络的问题匹配 .....	358
10.4.3 社区问答数据集 .....	361
<b>10.5 开放领域问答 .....</b>	<b>362</b>
10.5.1 检索-阅读理解架构的开放问答模型 .....	363
10.5.2 端到端架构的开放问答模型 .....	365
10.5.3 开放领域问答语料库 .....	367
<b>10.6 延伸阅读 .....</b>	<b>368</b>
<b>10.7 习题 .....</b>	<b>369</b>
<b>11 文本摘要 .....</b>	<b>370</b>
<b>11.1 文本摘要概述 .....</b>	<b>370</b>
11.1.1 文本摘要发展历程 .....	371
11.1.2 文本摘要主要任务 .....	372
<b>11.2 抽取式文本摘要 .....</b>	<b>373</b>
11.2.1 基于排序的方法 .....	373
11.2.2 基于序列标注的方法 .....	378
<b>11.3 生成式文本摘要 .....</b>	<b>382</b>
11.3.1 序列到序列生成文本摘要 .....	383
11.3.2 抽取与生成结合式文本摘要 .....	390
<b>11.4 文本摘要的评测 .....</b>	<b>392</b>
11.4.1 人工评测 .....	394
11.4.2 自动评测 .....	395
<b>11.5 文本摘要语料库 .....</b>	<b>398</b>
11.5.1 单文档摘要语料库 .....	398
11.5.2 多文档摘要语料库 .....	398
11.5.3 对话摘要语料库 .....	399
11.5.4 多模态文本摘要语料库 .....	399
11.5.5 跨语言文本摘要语料库 .....	399
<b>11.6 延伸阅读 .....</b>	<b>400</b>
<b>11.7 习题 .....</b>	<b>400</b>

<b>12 知识图谱 .....</b>	<b>402</b>
<b>12.1 知识图谱概述 .....</b>	<b>402</b>
12.1.1 知识图谱发展历程 .....	404
12.1.2 知识图谱研究内容 .....	405
<b>12.2 知识图谱表示与存储 .....</b>	<b>407</b>
12.2.1 知识图谱的符号表示 .....	407
12.2.2 知识图谱的向量表示 .....	410
12.2.3 基于表的知识谱图谱存储 .....	413
12.2.4 基于图的知识谱图谱存储 .....	416
<b>12.3 知识图谱获取与构建 .....</b>	<b>419</b>
12.3.1 属性补全 .....	420
12.3.2 实体链接 .....	423
12.3.3 实体对齐 .....	426
<b>12.4 知识图谱推理 .....</b>	<b>432</b>
12.4.1 基于符号逻辑的知识图谱推理 .....	433
12.4.2 基于表示学习的知识图谱推理 .....	436
<b>12.5 知识图谱问答 .....</b>	<b>440</b>
12.5.1 基于语义解析的知识图谱问答 .....	441
12.5.2 基于信息检索的知识图谱问答 .....	443
12.5.3 基于深度神经网络的知识图谱问答 .....	447
12.5.4 知识图谱问答语料库 .....	451
<b>12.6 延伸阅读 .....</b>	<b>452</b>
<b>12.7 习题 .....</b>	<b>453</b>
<b>13 模型稳健性 .....</b>	<b>454</b>
<b>13.1 稳健性概述 .....</b>	<b>454</b>
13.1.1 稳健性基本概念 .....	455
13.1.2 稳健性主要研究内容 .....	456
<b>13.2 数据偏差消除 .....</b>	<b>457</b>
<b>13.3 文本对抗攻击方法 .....</b>	<b>459</b>
13.3.1 字符级别攻击方法 .....	460
13.3.2 词级别攻击方法 .....	461
13.3.3 句子级别攻击方法 .....	463

13.3.4 后门攻击 .....	464
<b>13.4 文本对抗防御方法 .....</b>	<b>468</b>
13.4.1 基于对抗训练的文本防御方法 .....	468
13.4.2 基于表示压缩文本防御方法 .....	470
13.4.3 基于数据增强的文本防御方法 .....	471
13.4.4 对抗样本检测 .....	473
<b>13.5 模型稳健性评价基准 .....</b>	<b>474</b>
13.5.1 特定任务稳健性评价基准 .....	475
13.5.2 模型稳健性通用评价基准 .....	477
<b>13.6 延伸阅读 .....</b>	<b>482</b>
<b>13.7 习题 .....</b>	<b>483</b>
<b>14 模型可解释性 .....</b>	<b>484</b>
<b>14.1 可解释性概述 .....</b>	<b>484</b>
14.1.1 可解释的分类 .....	485
14.1.2 解释的评价 .....	486
<b>14.2 解释性分析方法 .....</b>	<b>488</b>
14.2.1 局部分析方法 .....	488
14.2.2 全局分析方法 .....	494
<b>14.3 自然语言处理算法解释分析方法 .....</b>	<b>497</b>
14.3.1 模型解释性分析算法 .....	497
14.3.2 数据解释分析方法 .....	501
14.3.3 可解释评估 .....	503
<b>14.4 延伸阅读 .....</b>	<b>505</b>
<b>14.5 习题 .....</b>	<b>505</b>



# 1. 绪论

---

自然语言处理（Natural Language Processing, NLP）是计算机科学领域和人工智能领域的重要研究方向之一，旨在探索实现人与计算机之间用自然语言进行有效交流的理论与方法<sup>[1]</sup>。它融合了语言学、计算机科学、机器学习、数学、认知心理学等多学科内容，涉及从字、词、短语到句子、段落、篇章的多种语言单位，以及处理、理解、生成等不同层面的知识点，研究内容涉及的知识点多且复杂。自 20 世纪 90 年代以来，自然语言处理发展迅猛，各类任务和算法和研究范式层出不穷，在搜索引擎、医疗、金融、教育、司法等众多领域展示出重要作用。

本章主要介绍自然语言处理的基本概念和研究内容，并对自然语言处理范式进行总结和介绍。

## 1.1 自然语言处理基本概念

语言是人类与其他动物最重要的区别，而人类的多种智能也与此密切相关。逻辑思维以语言的形式表达，大量的知识也以文字的形式记录和传播。如今，互联网上已经拥有数万亿以上的网页资源，其中大部分信息都是以自然语言描述的。因此，如果人工智能想要获取知识，就必须懂得如何理解人类使用的不太精确、可能有歧义、混乱的语言。

自然语言处理目标就是实现人机之间的有效通信，意味着要使计算机能够理解自然语言的意义，也能以自然语言文本来表达给定的意图、思想等<sup>[1]</sup>。前者称为自然语言理解（Natural Language Understanding, NLU），后者称为自然语言生成（Natural Language Generation, NLG）。需要说明的是，自然语言处理、自然语言理解以及计算语言学这些概念并没有严格统一的定义。本书采用吴立德教授在 1997 年所著的《大规模中文文本处理》中所给出的定义。无论是自然语言理解还是自然语言生成，目前都是开放性问题（Open Problem），通用的高精度高鲁棒自然语言处理系统还没有解决方案，仍然需要长期研究。但是针对特定领域的应用，很多具有自然语言处理能力的系统已经有产业化应用，例如：智能客服系统、机器翻译系统、语音助手、电子邮件筛选、新闻写作、智慧教育、司法辅助等。

### 1.1.1 自然语言处理简史

自然语言处理的研究历史可以追溯到 1947 年，当时第一台通用计算机 ENIAC 也才刚刚问世一年，Warren Weaver 就提出了利用计算机翻译人类语言的可能，并于 1949 年发布了著名的《Translation》(翻译)备忘录。1950 年，Alan Turing 发表了著名的具有划时代意义的论文《Computing Machinery and Intelligence》(计算机器与智能) [2]，提出了使用图灵测试 (Turing Test) 对机器是否具备智能进行评测，即如果一台机器能够与人类展开对话而不能被辨别出其机器身份，那么这台机器具有智能。1951 年语言学家 Yehoshua Bar-Hillel 在麻省理工学院开始了机器翻译研究。1954 年乔治城大学 (Georgetown University) 与 IBM 合作的机器翻译演示系统将 60 多个俄语句子翻译成了英文。研究者们当时期望通过三到五年的时间完全解决机器翻译问题。20 世纪 50 年代初是自然语言处理的萌芽期。自然语言处理简史的时间线如图 1.1 所示。大体来看自然语言处理经历了 20 世纪 50 年代末到 60 年代的初创期、20 世纪 70 年代到 80 年代的理性主义时代、20 世纪 90 年代到 21 世纪初的经验主义时代以及 2006 年至今的深度学习时代。

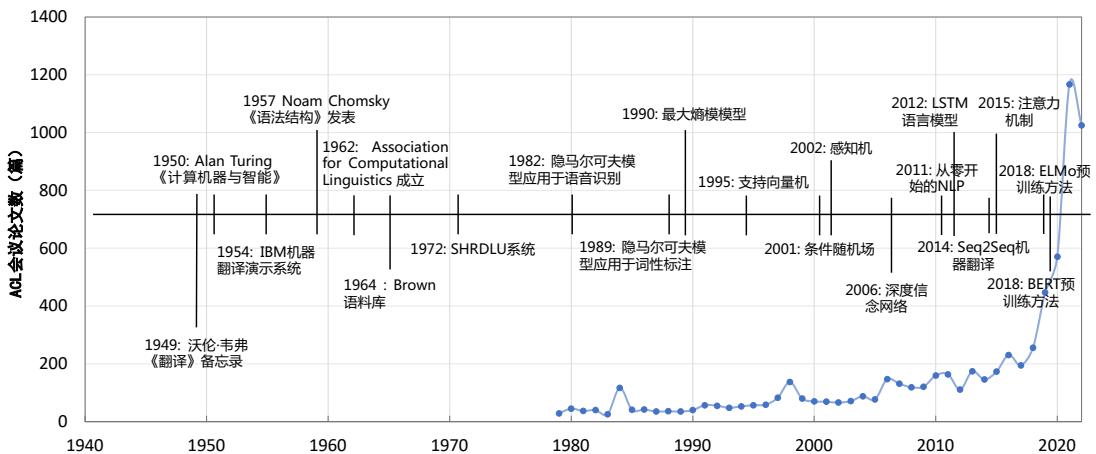


图 1.1 自然语言处理简史时间线

20 世纪 50 年代末到 60 年代，大量的研究不断涌现，并且形成了两大流派：符号学派 (Symbolic) 和随机学派 (Stochastic)。以美国语言学家 Noam Chomsky 为代表的符号学派提出了形式语言理论，基于 1957 年发表的《Syntactic Structures》(句法结构) 介绍了生成语法的概念，并提出了一种特定的生成语法称为转换语法。开启了使用数学方法研究语言的先河。随机学派则是以 1959 年 Bledsoe 和 Browning 将贝叶斯方法 (Bayesian method) 应用于字符识别问题为代表。试图通过贝叶斯方法来解决自然语言处理中的问题。这期间计算语言学 (Computational Linguistics) 概念也被正式提出。1962 年美国成立了机器翻译和计算语言学学会 (Association for Machine Translation

and Computational Linguistics)。1965 年国际计算语言学委员会 (The International Committee on Computational Linguistics, ICCL) 成立，并于当年召开了第一届国际计算语言学大会 (The International Conference on Computational Linguistics, COLING)。20 世纪 60 年代还出现了第一个大规模语料库，布朗美国英语语料库 (Brown Corpus)，包含来自不同文体的 500 多篇书面文本，超过一百万单词，涉及新闻、小说、科技文化等。自此，自然语言处理研究全面开启。

20 世纪 70 年代到 80 年代，更多的工作从不同角度开展了系统的研究，也产生了一系列的研究范式，至今仍对自然语言处理研究起着重要作用。这些范式主要包括：基于逻辑的范式 (Logic-based Paradigm)、基于规则的范式 (Rule-based Paradigm) 和随机范式 (Stochastic Paradigm)。1970 年 Colmerauer 等人使用逻辑方法研制 Q 系统 (Q-system) 和变形语法 (Metamorphosis Grammar) 并在机器翻译中得到应用。1980 年 Pereira 和 Warren 提出的定子句语法 (Definite Clause Grammar) 是逻辑范式成功应用的范例之一。基于规则的范式是这个时代最典型的模式之一，1972 年研制的 SHRDLU 系统是其中一个代表性工作。该系统模拟了一个玩具积木世界，能够接受自然语言的书面指令（例如：Pick up a big red block.），指挥机器人移动玩具积木块。1970 年，William A. Woods 提出了扩充转移网络 (Augmented Transition Network) 用来描述自然语言输入，并用于自然语言处理若干任务中。受到 20 世纪 80 年代初隐马尔可夫模型 (Hidden Markov Model) 和噪声信道与解码模型 (Noisy Channel Model and Decoding Model) 在语音识别中的成功应用，随机范式也逐渐在自然语言处理任务中崭露头角，包括词性标注<sup>[3]</sup>、姓名检索<sup>[4]</sup>等。

从 20 世纪 90 年代开始，自然语言处理开启了繁荣发展的时代。自 1989 年机器翻译任务中引入语料库方法之后，这种建立在大规模真实语料上的研究方法将自然语言处理研究推向了新的高度。从 90 年代后期开始，基于机器学习和数据驱动的方法取代了早期基于规则和基于逻辑的方法，成为自然语言处理的标准模式。自然语言处理的各类任务，包括词法分析、词性标注、句法分析、文本分类、机器翻译等都开始引入机器学习算法。这期间朴素贝叶斯 (Naive Bayes)<sup>[5]</sup>、K 近邻 (K-nearest neighbor)<sup>[6]</sup>、支撑向量机 (Support Vector Machine, SVM)<sup>[7]</sup>、最大熵模型 (Maximum Entropy, ME)<sup>[8]</sup>、神经网络 (Neural Network)<sup>[9]</sup>、条件随机场 (Conditional Random Fields)<sup>[10]</sup>、感知机 (Perceptron)<sup>[11]</sup> 等方法也都在自然语言处理不同任务上进行了尝试并取得了一定的成功。这种以大规模数据为基础进行分析的方法称为经验主义 (Empiricism)。随着数据驱动方法的发展，大部分关于自然语言处理的早期理论都大打折扣，特别是数据量的不断增加以及计算能力的不断提高，经验主义方法直到现在也还在主导着自然语言处理领域。从当前自然语言处理领域重要会议 EMNLP (Empirical Methods in Natural Language Processing) 的名称和发展也可以看到经验主义的发展过程。

2006 年加拿大多伦多大学教授 Geoffrey Hinton 和他的学生 Ruslan Salakhutdinov 在《科学》杂志上发表了基于深度信念网络 (Deep Belief Networks, DBN) 以及无监督预训练结合有监督训练微调的方法解决深层神经网络训练中梯度消失问题<sup>[12]</sup>，将神经网络重新拉回到机器学习研究者的视野中。2012 年基于卷积神经网络 (Convolutional Neural Network, CNN) 网络的 AlexNet 在图像识

别领域 ImageNet 竞赛中取得惊人的效果，开启了深度学习在学术界和工业界的浪潮<sup>[13]</sup>。2011 年论文《Natural language processing (almost) from scratch》(从零开始的 NLP) 引起了极大的关注，深度神经网络可以在不使用人工特征的情况下，用一个统一的网络架构在词性标注、组块分析、命名实体识别、语义角色标注等任务中都取得了很好的效果<sup>[14]</sup>。2014 年 Seq2Seq (序列到序列) 的模型<sup>[15]</sup>在机器翻译任务上取得了非常好的效果，并且完全不依赖任何人工特征，推动了神经机器翻译的广泛落地。这种端到端的方式进行编码和解码的方式不仅有效推动了包括生成式摘要<sup>[16]</sup>、对话系统<sup>[17, 18]</sup>等在内的自然语言生成问题上取得了突破，还应用于自然语言处理中的很多任务，包括句法分析<sup>[19]</sup>、问题回答<sup>[20]</sup>、中文分词<sup>[21]</sup>等。此外，循环神经网络 (Recurrent neural network, RNN)<sup>[22]</sup>、长短时记忆网络 (Long Short Term Memory Network, LSTM)<sup>[23]</sup>、递归神经网络 (Recursive Neural Network)<sup>[24]</sup>、卷积神经网络 (Convolutional Neural Network, CNN)<sup>[25]</sup>、图神经网络 (Graph Neural Networks, GNN)<sup>[26, 27]</sup> 等神经网络模型也都成功应用于自然语言处理各个任务中。

2018 年美国艾伦人工智能研究所 (Allen Institute for AI) 和华盛顿大学 (Washington University) 联合发表的论文中提出了名为 ELMo 的上下文相关的文本表示方法，首先利用语言模型或其他自监督任务进行预训练，此后在处理下游任务时，从预训练的网络中提取对应单词的网络各层的单词嵌入作为新特征补充到下游任务中，在多个自然语言处理任务上表现非常突出<sup>[28]</sup>。此后，深度学习开启了预训练模型 (Pre-trained Models, PTM) 结合任务微调的新范式。谷歌、OpenAI、微软、清华大学、百度、智源研究院等先后提出了 BERT<sup>[29]</sup>，GPT<sup>[30]</sup>，XLNet<sup>[31]</sup>、ERNIE(THU)<sup>[32]</sup>、ERNIE(Baidu)<sup>[33]</sup>、悟道等大规模预训练模型，在几乎所有自然语言处理任务中都取得了非常好的效果，甚至在很多任务的标准评测集合上取得了超越人类准确率的水平。尤其是在类似阅读理解、常识推理等任务上有惊人的效果提升。与此同时，预训练模型的规模也越来越大，2018 年谷歌开发的 BERT-Base 模型有 1.1 亿参数，BERT-Large 模型有 3.4 亿参数，到了 2019 年 OpenAI 开发的 GPT-2 模型就达到了 15 亿参数量。2021 年 GPT-3 模型参数量更是达到了 1750 亿，而同年谷歌开发的 Switch Transformer 模型参数量首次超过万亿，达到了 1.6 万亿。在此之后不久，北京智源研究院所发布的“悟道 2.0”模型就刷新了上述记录，模型参数量达到了 1.75 万亿。虽然预训练大模型取得了巨大的成功，但是仍然面临模型鲁棒性亟待提升、超大规模模型如何高效适配下游任务、大模型的理论解释等诸多问题。

### 1.1.2 自然语言处理的主要研究内容

自然语言处理的研究内容十分庞杂，整体上可以分为基础算法研究和应用技术研究。基础算法研究又可以细分为自然语言理解和自然语言生成。从语言单位角度看，涵盖了字、词、短语、句子、段落以及篇章等不同粒度。从语言学研究角度看则涉及形态学、语法学、语义学、语用学等不同层面。此外，由于目前绝大多数自然语言处理算法采用基于机器学习的方法，针对特定的自然语言处理任务，以有监督、无监督、半监督、强化学习等不同的机器学习算法为基础进行构建。因此，自然语言处理研究又与机器学习和语言学研究交织在一起，使得自然语言处理的研究内容

涉及范围广，学科交叉度大。

自然语言处理研究与语言学密切相关，语言学研究可以划分为形态、语法、语义、语用等几个层面。形态学（Morphology）主要研究单词的内部结构和构成方式。语法学（Syntax）主要研究句子、短语以及词等语法单位的语言结构与语法意义的规律。语义学（Semantics）主要研究语言的意义，目标是发现和阐述关于意义的知识。语用学（Pragmatics）是从使用者的角度来研究语言，研究在一定的上下文环境下的语言如何理解和使用。在实际的任务中，上述几个层面的问题往往相互关联，并不能完全独立。语法结构的分析需要词汇形态学的支撑，语法结构也影响着词汇的形态，语法结构和语义也是相互交织，而下上文环境又对语义有重要的影响，因此很多自然语言处理任务并不是完全独立的。但是为了简化任务处理难度，通常处理不同层面的任务时仍然独立考虑。从自然语言处理研究内容的难度来看，从形态、语法、语义到语用是逐层递增的。目前基于机器学习的自然语言算法处理涉及语义的相关任务都较为困难，因此语用层面的自然语言处理算法研究相对较少，大多数的研究集中于形态、语法和语义三个层面。从语言单元粒度和语言学研究层次两个维度进行归类，自然语言处理主要研究内容如图1.2所示。

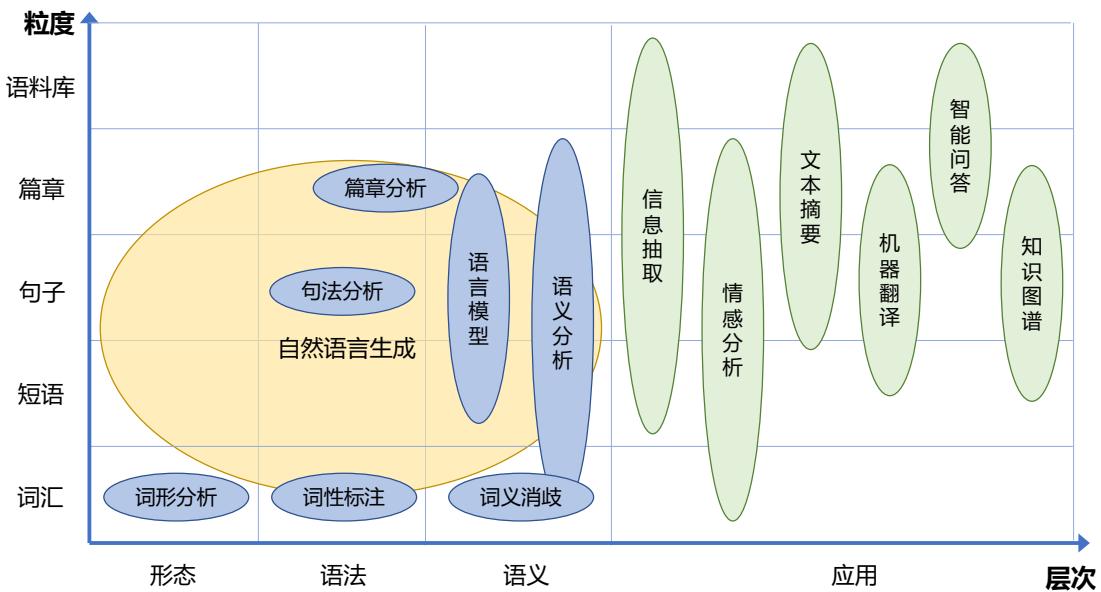


图 1.2 自然语言处理主要研究内容

自然语言处理主要研究内容在词汇粒度的研究内容主要包括：词形分析、词性标注、词义消歧，分别针对单词的词性、语法、语义开展研究。句法分析则是主要针对句子根据语法进行结构分析。篇章分析核心是对篇章的连贯性和衔接性进行分析，涉及到篇章级别语法结构，同时也包

含部分语义的内容。而语义分析研究则涉及到从词汇、短语、句子到篇章等各个粒度。语言模型主要聚焦于句子粒度，但是也包含部分短语和篇章级别的研究。以上研究内容主要围绕自然语言理解的基础问题开展。自然语言生成则主要研究利用常识、逻辑和语法等知识自动生成文本，涉及形态、语法和语义层面，同时也涵盖从短语到篇章多个粒度。在自然语言处理基础研究内容之上，信息抽取、情感分析、文本摘要、机器翻译、智能问答、对话系统等任务则围绕自然语言处理的应用开展，所处理的语言单元也根据任务特性而不尽相同。

整体上来看，自然语言处理的主要研究内容围绕语言学基础理论，在形态、语法以及语义等层面开展自然语言理解基础算法和自然语言生成基础算法研究。在此基础上围绕自然语言处理的重要应用场景开展一系列的应用技术研究。这些研究内容也已经深度应用于信息检索、虚拟助理、推荐系统、量化交易、智能问诊、精准医疗等众多系统中。

### 1.1.3 自然语言处理的主要难点

自然语言理解和自然语言生成都是十分困难的任务，这种困难的根本原因是自然语言在各个层面都广泛存在的各种各样的歧义性或多义性 (Ambiguity)。自然语言文本从形式上是由字符（包括中文汉字、英文字母、符号）组成的字符串。由字母或者汉字可以组成词，由词可以组成词组，由词组可以组成句子，进而组成段落、篇章。无论哪种粒度的语言单元，还是从一个层级向上一个层级转变中都存在歧义和多义现象。形式上一样的字符串，可以理解为不同的词串、词组串，并有不同的意义<sup>[1]</sup>。Joseph F. Kess 和 Ronald A. Hoppe 甚至还提出了“语言无处不歧义”的理论<sup>[34]</sup>。在某种程度上，我们也可以说明自然语言处理基础任务的核心就在于解决歧义问题。

#### 1. 语音歧义

语音歧义 (Phonetic Ambiguity) 主要体现在口语中，是由于语言中同音异义词 (Homophone)、爆破音不完全、重音位置不明确等原因造成的。汉字的同音异义现象则更加严重，在汉语中只有 413 个不同的音（节），如果结合声调的变化组合，也仅有 1277 个音（节），而汉字则多达数万个，因此同音字非常多。英语中虽然同音异义的词语相对汉语要少得多，但是由于连读、爆破音、重音位置等造成的语音异义也非常常见。

例如：请问您贵姓？

免贵姓 zhang。

这组对话中“zhang”既可以是“张”，也可以是“章”。汉语中同音异义词也有非常多，例如：“chéng shì：城市、程式、成事、城事”、“jìn shì：近视、进士、尽是”、“shǒu shì：首饰、手势”等。

在英语中语音歧义的现象虽然没有汉语中这么严重，但是也是普遍存在的现象。

例如：Please hand me the flower. 请把花递给我。

Please hand me the flour. 请把面粉递给我。

这两句话中“flower”和“flour”的发音相同，由同音异义词造成了歧义。类似的情况还包括“see (看见) 与 sea (大海)”、“son (太阳) 与 sun (儿子)”等。

## 2. 词语切分歧义

词语切分歧义 (Word Segmentation Ambiguity) 是由字符组成词语时的歧义现象。对于英语等印欧语系的语言来说，绝大部分单词之间都由空格或标点分割。但是对于汉语、日语等语言来说，单词之间通常没有分隔符。对于这些语言来说，这些连续的字符切分为单词时就会产生歧义。

例如：语言学是一门基础学科。

这门语言学起来很困难。

该例句中“语言学”、“语言”都是词语，在同一个句子中就会出现多种切分方法。这种切分歧义在汉语中普遍存在。我们将在第 2 章详细讨论词语切分歧义的问题以及词语切分的方法。

## 3. 词义歧义

词义歧义 (Word Sense Ambiguity) 是指词语具有相同形式但是不同意义。这种歧义在各种语言中都广泛存在，通常越是常见的词语其词义数量就越多。例如“打”字在《现代汉语词典（第七版）》中，有两个读音“dá”和“dǎ”，分别作为量词、动词和介词，在作为动词时“打”字有 24 个意项<sup>[35]</sup>。

例如：打 dǎ 动词：

- (1) 用手或器具撞击物体：～门 |～鼓
  - (2) 器皿、蛋类等因撞击而破碎：碗～了 | 鸡飞蛋～
  - (3) 殴打；攻打：～架 |～援
  - (4) 发生与人交涉的行为：～官司 |～交道
  - (5) 汲取；盛取：～米 |～酱油
- ...

英语中存在大量类似的情况，例如根据 WordNet 中给出的定义，单词“bank”具有名词和动词两种词性，作为名词时具有 10 种词义<sup>[36]</sup>。

例如：Bank 名词：

- (1) sloping land (especially the slope beside a body of water  
“they pulled the canoe up on the bank”)
  - (2) a financial institution that accepts deposits and channels the money into lending activities  
“he cashed a check at the bank”
  - (3) a long ridge or pile  
“a huge bank of earth”
- ...
- (10) a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)  
“the plane went into a steep bank”

我们将在第 4 章中详细讨论词汇的语义歧义问题以及消除词汇语义歧义的方法。

## 4. 结构歧义

结构歧义 (Structural Ambiguity) 是由词组成词组或者句子时, 由于其组成的词或词组间可能存在不同的语法或语义关系而出现的 (潜在) 歧义现象。结构歧义有时也称为语法歧义 (Grammatical Ambiguity)。冯志伟教授在文献 [37] 中对结构歧义进行了系统的描述, 其中一些典型的结构歧义如下:

- “VP+ 的 + 是 +NP” 型歧义结构:

例如: 反对 | 的 | 是 | 少数人

该类型歧义中, VP 是一个双向动词, “VP+ 的”是主语, “是 +NP”是谓语, 整个句式是个一个主谓结构。由于主语部分的“VP+ 的”既可以是施事, 也可以是受事, 因而会产生歧义。这个例子中既可以理解为“提反对意见的是少数人”, 也可以理解为“所反对的是少数人”。

- “VP+N1+ 的 +N2” 型歧义结构:

例如: 咬死了 | 猎人 | 的 | 狗

该类型歧义中, N1 作为 VP 的宾语, 述宾结构“VP+N1”加上“的”之后, 作为名词 N2 的定语, 整个结构是一个定中结构。但是 N1 又可以与“的”结合在一起作为 N2 的定语, 构成“N1+ 的 +N2”, 这个名词词组作为 VP 的宾语, 整个结构构成一个述宾结构。这个例子中既可以理解为“咬死了一只猎人的狗”, 也可以理解为“一只把猎人咬死的狗”。

- “N1+ 和 +N2+ 的 +N3” 型歧义结构:

例如: 桌子 | 和 | 椅子 | 的 | 腿

该类型歧义是由于连词“和”的管辖范围的不同造成的潜在歧义。这个例子中既可以理解为“桌子和 (椅子的腿)”, 也可以理解为“ (桌子和椅子) 的腿”。

类似的结构歧义类型有很多, 例如: “ADJ+N1+N2”、“VP+ADJ+ 的 +N”等。这些歧义的不同理解会造成不同的句法结构以及语义上的不同。句法分析的主要难度就是解决结构歧义问题。我们将在第 3 章对结构歧义以及如何进行句法分析进行详细介绍。

### 5. 指代和省略歧义

在由多个句子组成的段落或篇章中, 各种歧义依然存在, 例如指代歧义和省略歧义。指代歧义 (Demonstrative Ambiguity) 是指代词 (如我, 你, 他等) 和代词词组 (如“那件事”, “这一点”等) 所指的事件可能存在歧义。

例如: 猴子吃了香蕉, 因为它 饿了。

猴子吃了香蕉, 因为它 熟透了。

上述两个句子的前半句完全相同, “它”可以指代“猴子”和“香蕉”, 需要具体是后半句的谓词决定指代关系。

省略歧义 (Ellipsis Ambiguity) 是指自然语言中由于省略所产生的歧义。省略是自然语言中的一种重要的语言现象, 尤其在汉语中省略现象非常常见。省略掉一些成分, 在绝大部分情况下不会影响句子的表达, 但是还是存在一些由于省略造成歧义的问题。

例如: 县政府同意乡政府报告。

这个例子中省略了助词“的”，因此使得该句具有两种解释，一个是县政府同意乡政府的那份报告，另外一个是县政府同意乡政府作出报告。

## 6. 语用歧义

语用歧义（Pragmatic Ambiguity）是指由于上下文、说话人属性、场景等语用方面的原因造成的歧义。一句话在不同的场合、由不同的人说、不同的语境，都可能产生不同的理解。

例如：下例由于场景的不同，同样的句子可以有不同的意义。

句子：你知道南京路怎么走吗？

(1) 如果说话人是游客，说话的对象是警察，那么这句话的含义就是问路。

(2) 如果说话人同样是游客，但是说话的对象换成出租车司机，那么这句话的含义就是询问出租车司机是否可以送他到南京路。

再比如，由于上下文的不同，同样的句子也可以有不同的意义。

句子：女子致电男友：地铁站见。如果你到了我还没到，你就等着吧。如果我到了你还没到，你就等着吧！！

这个例子中，同样的句子“你就等着吧”，前一个的含义是请耐性等待，后一个的含义是你要有麻烦了。

从上述介绍中，可以看到自然语言中存在大量的歧义现象。对人类而言，这些歧义在绝大多数的情况下都可以根据上下文以及相应的语境和场景得到解决。这也就是为什么我们平时使用自然语言交流并没有感知到语言的歧义。但是，为了消解这些歧义，需要使用大量的知识进行推理才能完成。而如何表示知识和使用知识、如何完整收集和整理知识以及常识都是极其困难的问题。莫拉维克悖论（Moravec's paradox）对自然语言处理也依旧适用。也正是由于这些问题，才使得消解歧义是自然语言处理中最大的难点之一。

此外，自然语言并不是一成不变的，而是在动态发展中，存在大量未知语言现象。新词汇、新含义、新用法、新句型等层出不穷<sup>[38]</sup>。

例如：新词汇：双碳、双减、绝绝子、社恐、元宇宙

新含义：躺平、打工人、凡尔赛、青蛙、潜水、盖楼

新用法：走召弓虽、YYDS、回忆杀、求扩列、orz

新句型：纠结的说、看书ing、一整个无语住

这些层见迭出的语言现象对于自然语言处理系统来说也是巨大的挑战。无论是自然语言处理基础任务还是应用系统，如何应对这些未知的情况都是巨大的挑战。

总而言之，自然语言处理的困难来源于非常多的方面，即面临来自于语言本身所不可避免的根本性问题，也缺乏通用的语义表示以及语言意义的理论支撑。同时，现阶段自然语言处理算法所依赖的机器学习方法，还存在需要大规模标注数据、跨领域效果差、泛化能力和鲁棒性弱、模型不可解释等诸多问题。也正因此，自然语言处理研究极具挑战，能够称得上“人工智能皇冠上的明珠”。

## 1.2 自然语言处理的基本范式

自然语言处理的发展经历了从理性主义到经验主义，再到深度学习三个大的历史阶段。在发展过程中也逐渐形成了一定的范式，主要包括：基于规则的方法、基于机器学习的方法以及基于深度学习的方法。这三种范式也基本对应了自然语言处理的不同发展阶段的重点。需要特别说明的是，虽然以上三种范式来源于自然语言处理的不同发展阶段，有明显的发展先后顺序，并且在大部分自然语言处理任务的标准评测集合中基于深度学习的方法都好于基于机器学习的方法，更优于基于规则的方法很多。但是，这三种范式各有利弊，在实际应用中需要根据任务的特点、计算量、可控制性以及可解释性等具体情况进行选择。

上述三种范式虽然有很大的不同，但是都有一个相同点就是需要针对特定任务进行构建。面向不同的任务，按照不同的范式构建数据、模型等不同方面，所得到的算法或者系统仅能够处理特定的任务。在机器学习和深度学习范式下，甚至对模型预测目标进行微小修正，通常都需要对模型进行重新训练。对于未知任务的零样本学习（Zero-shot Learning）能力很少在上述范式中进行讨论和研究。基于机器学习和深度学习范式也很难实现模型对未知任务的泛化。2022年11月随着ChatGPT的发布，大模型所展现出来的文本生成能力以及对未知任务的泛化能力使得未来的自然语言处理的研究范式很可能会发生非常大的变化。因此，本节中也将简要介绍大模型研究范式的雏形。

### 1.2.1 基于规则的方法

基于规则的自然语言处理方法的主要思想是通过词汇、形式文法等制定的规则引入语言学知识，从而完成相应的自然语言处理任务。这类方法在自然语言处理早期受到了很大的关注，包括机器翻译在内的很多自然语言处理任务都采用此类方法。甚至目前仍有很多系统还在使用基于规则的方法。基于规则的方法基本流程如图1.3所示，主要包含：数据构建、规则构建、规则应用和效果评价等四个部分。

基于规则的方法核心是规则形式定义，其目标是使得语言学家可以在不了解计算机程序设计的情况下，能够容易地将知识转换为规则。这就要求规则描述要具有足够的灵活性并易于使用和理解。规则引擎的目标是高效地解析这些人工定义的大量规则，针对输入数据根据规则库进行解释执行，从而完成特定任务。这种方式可以使得语言学家不需要编写代码就可以完成规则库构建。

常见的规则包括产生式、框架、自动机、谓词逻辑、语义网等形式。例如，产生式规则是以“IF-THEN”形式构造，表示如果满足条件，则执行相应的语义动作。举例来说，对于机器翻译任务可以构造如下规则库：

IF 源语言主语 = 我 THEN 英语译文主语 =I

IF 英语译文主语 =I THEN 英语译文 be 动词为 am/was

IF 源语言 = 苹果 AND 没有修饰量词 THEN 英语译文 =apples

条件判断中也可以结合正则表达式，增强规则的泛化能力。再比如，可以根据英语的词典，构造

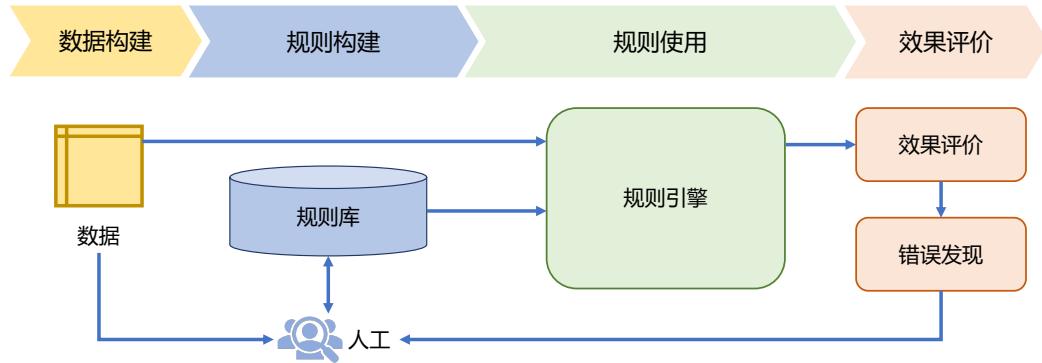


图 1.3 基于规则的自然语言处理算法基本流程

有限状态自动机 (Finite State Automaton, FSA) 进行英语单词的拼写检查。除此之外，非确定有限状态自动机 (Nondeterministic Finite Automaton, NFA)、有限状态转录机 (Finite State Transducers, FST) 还广泛应用于词法分析、词性标注、句法分析、机器翻译等众多方面。

基于规则的方法从某种程度上可以说是在试图模拟人类完成某个任务时的思维过程。这类方法主要优点是直观、可解释、不依赖大规模数据。利用规则所表达出来的语言知识具有一定的可读性，不同的人之间可以相互理解。规则分析引擎通过规则库所得到的分析结果，也具有很好的解释性。所使用的规则就可以作为系统作出判断的依据。规则库的构造也能够完全不依赖于大规模的有标注数据，可以仅根据人类背景知识进行构建。但是，基于规则的方法也有明显的缺点，主要包括覆盖率差、大规模规则构建代价大、难度高等。人工构建规则可以较为容易处理常见现象，但是对于复杂的语言现象难以描述。由于语言现象的复杂性，使得基于规则的方法整体覆盖率很难提升到非常高的程度。并且，规则库达到一定数量之后维护困难，新增加的规则与已有规则也容易发生冲突。不同人对于同一问题的解决思路的不同，也造成了大规模规则库中规则的不一致性，从而使得维护难度进一步提高。

## 1.2.2 基于机器学习的方法

基于机器学习的自然语言处理算法绝大部分采用有监督分类算法，将自然语言处理任务转化为某种分类任务，在此基础上根据任务特性构建特征表示，并构建大规模的有标注语料，完成模型训练。其基本流程如图1.4所示，通常分为四个步骤：数据构建、数据预处理、特征构建以及模型学习。

(1) 数据构建阶段主要工作是针对任务的要求构建训练语料，也称为语料库 (Corpus)。随着自然语言处理研究的不断发展，很多任务都有公开的基准测试集合 (Benchmark)，可以方便地用来进行模型训练以及模型之间的横向对比。针对没有公开数据的任务，也可以采用人工标注的方法构建训练语料。

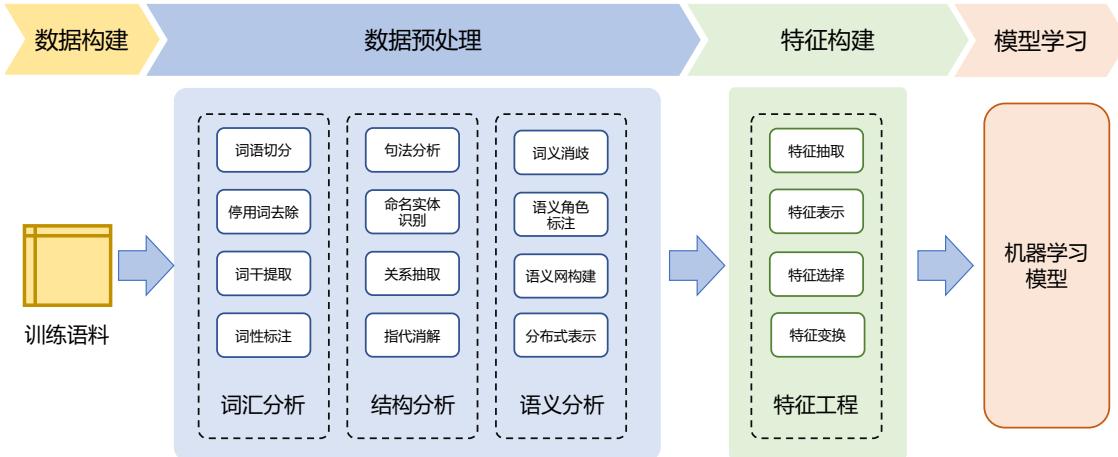


图 1.4 基于机器学习的自然语言处理算法基本流程

(2) 数据预处理阶段主要工作是利用自然语言处理基础算法对原始输入，从词汇、句法、结构、语义等层面进行处理，为特征构建提供基础。根据所处理语言和针对任务的不同，采用不同的模块和流程。对于汉语通常需要进行分词，对于英语通常需要进行词干提取和单词的规范化。在此之后，根据特征构建的需求，还可能需要进行词性标注、句法分析、语义角色标注等。

(3) 特征构建阶段主要工作是针对不同任务从原始输入、词性标注、句法分析、语义分析等结果和数据中提取对于机器学习模型有用的特征。例如，针对属性级情感倾向分析任务，需要根据目标属性，从句法分析结果提取该属性在对应句子中的评价词等信息。特征定义一般都是由人工完成，根据经验选取适合的特征，这项工作又被称为特征工程（Feature Engineering）。由于针对自然语言任务构建的特征通常维数非常高又非常稀疏，因此还会利用特征选择算法降低特征维度。也可以通过特征变换，根据人工设计的准则进行有效特征提取，例如：主成分分析、线性判别分析、独立成分分析等。

(4) 模型学习阶段主要工作是根据任务，选择合适的机器学习模型，确定学习准则，采用相应的优化算法，利用语料库训练模型参数。机器学习模型有很多类型，从不同的维度可以分为：分类模型、回归模型、排序模型、生成式模型、判别式模型、有监督模型、无监督模型、半监督模型、弱监督模型等等类别，需要根据任务的目标以及特性选择适合的模型。学习准则是机器学习模型中重要的因素，包括 0-1 损失函数（0-1 Loss Function）、平方损失（Quadratic Loss Function）、交叉熵损失函数（Cross-Entropy Loss Function）、Hinge 损失函数（Hinge Loss Function）等。针对所选择的模型和学习准则需要选择相应的优化算法，包括梯度下降（Gradient Descent Method）、牛顿法（Newton Method）、拟牛顿法（Quasi Newton Method）、随机梯度下降（Stochastic Gradient Descent, SGD）等。机器学习三要素：模型、学习准则、优化算法的选择都会对算法的效果产生影响。此外，

模型中通常包含一些可以调整的超参数（Hyper-parameters），也需要通过实验和经验进行选择。

通过整体流程可以看到，基于机器学习方法的自然语言处理算法需要针对任务构建大规模训练语料，以人工特征构建为核心，针对所需的信息利用自然语言处理基础算法对原始数据进行预处理，并需要选择合适的机器学习模型，确定学习准则，以及采用相应的优化算法。整个流程中需要人工参与和选择的环节非常多，从特征设计到模型，再到优化方法以及超参数，并且这些选择非常依赖经验，缺乏有效的理论支持。也使得基于机器学习的方法需要花费大量的时间和工作在特征工程上。开发一个自然语言处理算法的主要时间消耗在数据预处理、特征构建以及模型选择和实验上。此外，对于复杂的自然语言处理任务需要在数据预处理阶段引入很多不同的模块，这些模块之间需要单独优化，其目标并不一定与任务总体目标一致。其次多模块的级联会造成错误传播，前一步错误会影响后续的模型，这些问题都提高了基于机器学习的方法实际应用的难度。

### 1.2.3 基于深度学习的方法

深度学习（Deep Learning）方法通过构建有一定“深度”的模型，将特征学习和预测模型融合，通过优化算法使得模型自动地学习出好的特征表示，并基于此进行结果预测。基于深度学习方法的自然语言处理算法基本流程框架如图1.5所示。与传统机器学习算法的流程相比，基于深度学习方法的流程简化很多，通常仅包含数据构建、数据预处理和模型学习三个部分。同时，在数据预处理方面也大幅度简化，仅包含非常少量的模块。甚至目前很多基于深度学习的自然语言处理算法可以完全省略数据预处理阶段，对于汉语直接使用汉字作为输入，不提前进行分词，对于英语也可以省略单词的规范化步骤。

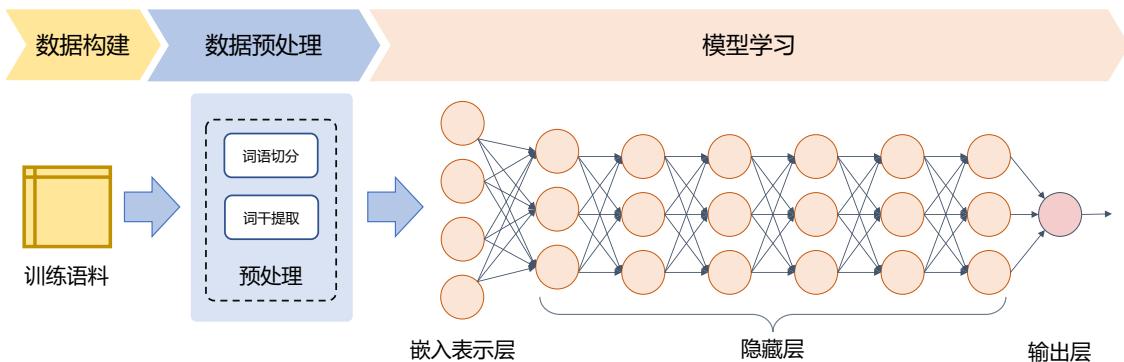


图 1.5 基于深度学习的自然语言处理算法基本流程

深度学习是机器学习的一个子集，通过多层的特征转换，将原始数据转换为更抽象的表示。这些学习到的表示可以在一定程度上完全代替人工设计的特征，这个过程也叫做表示学习（Representation Learning）。与基于特征工程的方法所通常采用的离散稀疏表示不同，深度学习算法通常使

用分布式表示（Distributed Representation），特征表示为低维稠密向量。分布式表示通常需要从底层特征开始，经过多次非线性变换得到。由于深层结构可以增加特征的重用性，从而使得表示能力指数级增加。因此，表示学习的关键是构建具有一定深度的多层次特征表示<sup>[39]</sup>。随着深度学习研究的不断深入和计算能力的快速发展，模型深度也从早期的 5 到 10 层增加到现在的数百层。随着模型深度的不断增加，其特征表示能力也不断增强，从而也使得深度学习模型中的预测部分更加简单，预测也更加容易。

自 2018 年 ELMo 模型<sup>[28]</sup> 提出之后，基于深度学习的自然语言处理范式又进一步演进为预训练微调范式。首先利用自监督任务对模型进行预训练，通过海量的语料学习到更为通用的语言表示，然后根据下游任务对预训练网络进行调整。这种预训练范式在几乎所有自然语言处理任务上都表现非常出色。预训练模型在模型网络结构上可以采用 LSTM、Transformer 等具有较好序列建模能力的模型，预训练任务可以采用语言模型、掩码语言模型（Masked Language Model）、机器翻译等自监督或有监督方式，还可以引入知识图谱、多语言、多模态等扩展任务。自 2018 年以来有非常多的相关研究，取得了非常好的效果，但仍然面临模型稳健性提升、模型可解释性等诸多问题亟待解决。第 6.4 节将对预训练模型进行详细介绍。

### 1.2.4 基于大模型的方法

大模型是大规模语言模型（Large Language Model）的简称。2018 年开始以 BERT<sup>[29]</sup>、GPT<sup>[30]</sup>为代表预训练语言模型相继推出，在各种自然语言处理任务上都得到了非常好的效果。此后，语言模型的规模不断扩大，2020 年 Open AI 发布的 GPT-3 模型<sup>[40]</sup>的规模达到了 1750 亿，Google 发布的 PaLM 模型<sup>[41]</sup>的参数量达到了 5400 亿。这种参数量级的语言模型很难再延续此前针对不同的任务而使用的预训练微调范式。因此，研究人员们开始探索使用采用提示词（Prompt）模式完成各类型自然语言处理任务。此后又提出了指令微调（Instruction Finetuning）<sup>[42]</sup> 方案，将大量各类型任务，统一为生成式自然语言理解框架，并构造训练语料进行微调。2022 年 ChatGPT 所展现出来的通用任务理解和未知任务泛化能力，使得未来自然语言处理的研究范式可能进一步发生变化。如图 1.6 所示，基于大模型的自然语言处理的流程转换为：大规模语言模型构建、通用任务能力训练以及特定任务使用三个主要步骤。

在大规模语言模型构建阶段，通过大量的文本内容，训练模型长文本的建模能力，使得模型具有语言生成能力，并使得模型获得隐式的世界知识。由于模型参数量和训练数据量都十分庞大，普通的服务器单机无法完成训练过程，因此需要解决大模型的稳定分布式架构和训练问题。在通用能力注入阶段，利用包括阅读理解、情感分析、信息抽取等现有任务的标注数据，结合人工设计的指令词对模型进行多任务训练，从而使得模型具有很好的任务泛化能力，能够通过指令完成未知任务。特定任务使用阶段则变得非常简单，由于模型具备了通用任务能力，只需要根据任务需求设计任务指令，将任务中所需处理的文本内容与指令结合，然后就可以利用大模型得到所需结果。

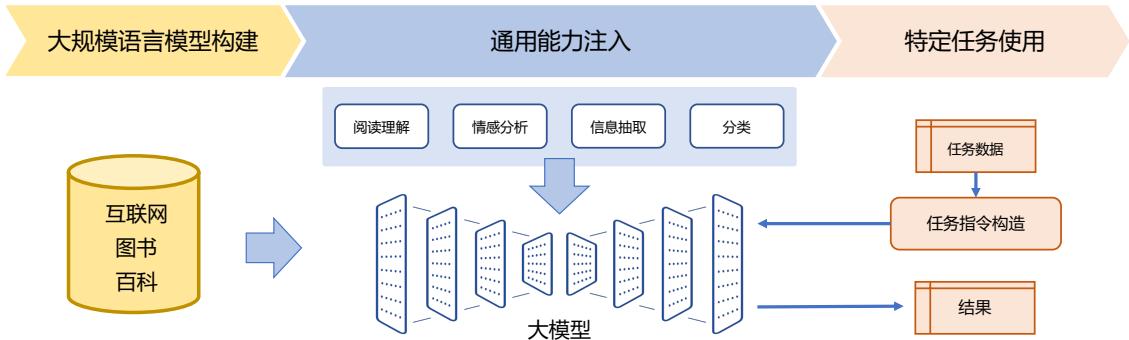


图 1.6 基于大模型的自然语言处理算法基本流程

如果该范式在非常多任务上都达到了目前基于预训练微调范式的结果，那么该范式会使得自然语言处理产生质的飞跃。突破了传统自然语言处理需要针对不同任务进行设计和训练的瓶颈，任务可以不需要预先给定，仅依赖很少的任务特定标注数据，或者完全不依赖任何任务的有监督数据就可以得到相应结果。当然，这种方法也仅仅是刚刚展露出一定的希望，当前使用该范式的大模型在绝大部分任务上所得到的效果仍然具有预训练微调范式有很大差距，模型参数量太大导致训练和使用成本过高，等等这些问题都亟待研究。

### 1.3 本书的内容安排

本书共分为 14 章，主要包含三个部分：第一部分主要介绍自然语言处理的基础技术，包括词汇处理、句法分析、语义分析、篇章分析和语言模型；第二部分主要介绍自然语言处理的一系列核心技术，包括信息抽取、机器翻译、情感分析、文本摘要、知识图谱；第三部分主要介绍基于机器学习的自然语言处理模型的稳健性和可解释性问题。本书章节安排如图 1.7 所示。

第 2 章到第 6 章从词汇、句法到篇章三个不同粒度的语言单位，从形态、结构到语义三个不同语言层面，对自然语言处理的基础技术进行介绍。第 2 章主要介绍语言学中词汇相关的基本概念，以及词语规范化、中文分词、词性分析等词汇分析主要任务和相关算法。第 3 章主要介绍语言学中句法基本概念，以及成分句法分析算法、依存句法分析算法。第 4 章主要介绍语义学和语义表示的基本概念和语义知识的表示方法，以及词义消歧、语义角色标注等语义分析主要任务和相关算法。第 5 章主要介绍篇章结构基础理论和基本概念，以及话语分割、话语分析和指代消解等篇章分析的主要任务和相关算法。第 6 章主要介绍语言模型基本概念，以及  $n$  元语言模型、神经语言模型以及预训练语言模型的常见算法。

第 7 章到第 12 章主要介绍自然语言处理支撑各种应用的核心技术。第 7 章主要介绍信息抽取的基本任务和相关算法，包括命名实体识别、关系抽取和事件抽取。第 8 章主要介绍机器翻译的



图 1.7 本书章节安排

基本概念和常见方法，包括基于统计和基于神经网络的机器翻译方法。第 9 章主要介绍情感倾向分析的基本概念和主要任务，包括文档、句子、属性三个不同粒度的分析算法。第 10 章主要介绍智能问答基本任务和分析算法，包括阅读理解、表格问答、社区问题、开发问答等。第 11 章主要介绍文本摘要的相关任务和基本算法，包括生成式文本摘要、抽取式文本摘要等。第 12 章主要介绍知识图谱相关概念和基本任务，包括知识表示学习、知识图谱构建和知识图谱应用。

第 13 章和第 14 章针对基于机器学习模型的自然语言处理算法所面临的模型稳健性问题和可解释性问题进行讨论。第 13 章主要介绍自然语言处理模型稳健性的基本概念，以及数据偏差消除、文本攻击方法、文本防御方法以及模型稳健性评价基准。第 14 章主要介绍自然语言处理模型的可解释性问题，主要包括解释性分析工具和可解释自然语言处理。

此外，还需要说明特别强调的是自然语言处理中很多任务都转换为了机器学习问题，因此很多机器学习算法可以应用于多个自然语言处理任务。比如，条件随机场模型（Conditional Random Fields, CRF）可以用于中文分词，也可以用于词性标注，还可以用于命名实体识别。为了避免重复，我们仅在第2.3.3节详细介绍了如何使用线性链条件随机场模型进行中文分词，在词性标注、命名实体识别等章节选择了不同的算法进行介绍。需要读者朋友能够融会贯通，在本书学习结束时对特定机器学习模型可以适用于哪些自然语言处理任务有清晰的了解。

## 2. 词汇分析

---

词汇是语言知识中的重要环节，在语言学中，词（Word）是形式和意义相结合的单位<sup>[43]</sup>，也是语言中能够独立运用的最小单位。掌握一个词汇意味着知道其读音和语义。在书面语中，正字法（Orthography）也是词形式的一种表达。例如：英文单词“cat”具有的语义是“猫”，读音为“/kæt/”。由于词是语言运用的基本单位，自然语言处理算法中词通常也是基本单元。因此，词的处理也是自然语言处理中重要的底层任务，是句法分析、文本分类、语言模型等任务的基础。

本章首先介绍语言学中词相关的基本概念，在此基础上再介绍词语规范化相关算法，中文分词算法，以及词性分析算法。

### 2.1 语言中的词汇

词通常是由语素（Morpheme）构成。语素又称词素，是语言中意义的最小单元。语素与词不同，语素不能够独立运用而词可以。只包含一个语素的词语称为简单词（Simple word），而包含多个语素的词称为复杂词（Complex word）。例如：“电灯”，包含“电”和“灯”两个语素此外，根据词在语言中的用途的不同，词还可以被划分为实义词（Content words）和功能词（Function words）。实义词包含事物、行为、属性和观念等概念。功能词则是指没有清楚词汇意义或与之有关的明显概念的词。本节将分别针对语素如何构成词以及如何对词进行分类进行介绍。

#### 2.1.1 词的形态学

虽然单词的形式和意义之间的关系本质上是任意的，但是由于社会的约定俗成，词的形式具有服从于某种规则的内在结构。在语言学中，研究单词的内部结构和其构成方式的学科称为形态学（Morphology），又称构词学。词是由一个或多个语素构成，语素主要分成两类：词根（Lemma）和词缀（Affix）。词根也称为原形或字典形，是指能在字典中查到的语素，通常是一个词最主要的语素。词缀是其他附着在原形上的语素，帮助在原形基础上衍生出新词，包含前缀、中缀、后缀等。

例如：英语单词 unhappy 中， happy 为原形， -un 为前缀

邦托克语单词 fumikas (是强壮的) 中， fikas (强壮) 为原形， -um- 为中缀

俄语单词 barabanshchik (鼓手) 中, baraban (鼓) 为原形, -shchik 为后缀 Morphology 本身就是由两个语素构成: morph+ology, 后缀-ology 表示“关于... 的科学”。一个词也可以包含多个词缀, 例如: unhappiness 包含前缀 “un-” 和后缀 “-ness”。同样, 一个词也可以包含多个词根, 例如: homework 包含词根 “home” 和 “work”。

有些语言的单词通常只包含一个或者两个语素, 但是有一些语言的单词则包含多达十个以上的语素。汉语中每个单词的语素都很少, 也不会根据性、数、格、人称等发生形态变化。但是对于英语单词 dog, 在末尾添加 s 可以将它从单数名词变成复数名词 dogs。对于德语单词 bäcker, 在末尾添加 in 可以将它从阳性词 (男面包师) 变为阴性词 bäckerin (女面包师)。不同语言的词形变化差别非常大, 以英语为例, 很多英语词都包含两个或两个以上的语素, 其词形变化如表2.1所示。

表 2.1 英语中常见词形变换

词形变化	说明	举例
屈折 Inflection	通过“词根 + 词缀”的方式构成和原形“同一类型”的词	名词后加 -s 后缀 复数名词 (cat+s) 动词后加 -ed 后缀 动词的过去式 (walk+ed)
派生 Derivation	通过“词根 + 词缀”的方式构成和原形“不同类型”的词	employ 添加后缀 -ee 变为 employee meaning 添加后缀 -less 变为 meaningless
复合 Compounding	通过组合多个词根构成一个新词, 也称组合词	home + work → homework water + proof → waterproof
附着 Cliticization	通过“词根 + 附着语”的方式“附着”在词根上	I'm 中的'm 代表 am 附着在 I 上 We're 中're 代表 are
截搭 Blending	两个词语各自的一部分拼接起来构成新词	smoke (烟) + fog (雾) → smog (烟雾) spoon (勺子) + fork (叉子) → spork (叉勺)
缩略 Acronym	短语中多个单词首字母组合在一起组合成词	NLP 代表 Natural Language Processing IT 代表 Information Technology
截短 Clipping	将长的单词截为较短的单词	demonstration 简化为 demo refrigerator 简化为 fridge

通过语素组成词汇也可以反映语言的一个重要特性: 创造性。我们可以理解从未见过的词, 也可以通过新颖的方法将语素结合起来创造新词。如果能够自动将词汇分解为语素, 可以更好地对词汇进行进一步的分析。

### 2.1.2 词的词性

词性 (Part of Speech, POS) 也称词类, 是根据词在句子中扮演的语法角色以及与周围词的关系对词的分类。例如: 表示事物的名字 (“钢琴”), 地点 (“上海”) 通常被归为名词, 而表示动作 (“踢”), 状态 (“存在”) 的词被归为动词。对词性进行划分时通常要综合考虑词的语法特性的各个方面, 以某一个标准为主, 同时参照其他标准进行。通过词性可以大致圈定一个词在上下文环境

中有可能搭配的范围，例如：介词“in”后面通常跟名词短语。通过词性可以为语法分析、语义理解提供帮助。由此，词性也被称为带有“分布式语法”信息 (Syntactic distributional properties)。

现在语言学中一个重要的词的分类是区分实义词(Content Words)和功能词(Function Words)。实义词表达具体的意义。由于实义词可以不断地增加，因此这类词又被称作开类词(Open class words)。实义词主要包含名词、动词、形容词等。功能词则主要是为了满足语法功能需求。由于功能词相对比较稳定，一个语言中通常很少增加新的功能词，因此功能词又被称作闭类词(Close Class Words)。功能词主要包含代词、冠词、指示词等。

以英语为例，词性主要包含以下几种：

名词(Noun)是表示人、物、地点以及抽象概念的一类词。名词按其意义又可以分为专有名词(Proper Noun)和普通名词(Common Noun)。普通名词还可以再细分为类名词(Class Noun)、集体名词(Collective Noun)、物质名词(Material Noun)和抽象名词(Abstract Noun)。名词还可以按照其可数性分为可数名词(Countable Noun)和不可数名词(Uncountable Noun)。

- 例如：
- 1) 专有名词: Shanghai (上海) New York (纽约)
  - 2) 类名词: city (城市) bird (鸟)
  - 3) 集体名词: family (家庭) army (军队)
  - 4) 物质名词: water (水) light (光)
  - 5) 抽象名词: music (音乐) honesty (诚实)

动词(Verb)是表示动作或状态的一类词，是英语中最复杂的一类词。动词除了具有人称和数的变化之外，还具备一些语法特征，包括：时态(tense)、语态(voice)、语气(mood)、体(aspect)等。动词可以进一步细分为及物动词(Transitive verb)、不及物动词(Intransitive verb)、连系动词(Linking verb)、助动词(Auxiliary verb)、限定动词(Finite verb)、不限定动词(Non-finite verbs)、短语动词(Phrasal verb)等。例如：

- 例如：
- 1) Boys **fly** kites. (男孩们放风筝)
  - 2) 不及物动词: Birds **fly**. (鸟会飞)
  - 3) 连系动词: The rose **smells** sweet. (玫瑰花香)
  - 4) 助动词: I **may** have meet him before. (我以前应该见过他)
  - 5) 限定动词: John **reads** papers every day. (约翰每天都读论文)
  - 6) 不限定动词: I hope **to see** you this morning. (我希望早上见到你)
  - 7) 短语动词: Tom **called up** George. (汤姆给乔治打了电话)

形容词(Adjective)是用来描写或修饰名词的一类词。按照构成，形容词可以分为简单形容词和复合形容词。按照与其所修饰的名词的关系，形容词还可以分为限制性形容词(Restrictive adjective)和描述性形容词(Descriptive adjective)。例如：

- 例如：
- 1) 简形容词：
    - a) 由一个单词构成 good (好的) long (长的)

- b) 由现在分词构成 interesting (令人感兴趣的)
- c) 由过去分词构成 learned (博学的)
- 2) 复形容词: duty-free (免税的) hand-made (手工制作的)
- 3) 限制性形容词: an **Italian** dish (一道意大利菜)
- 4) 描述性形容词: a **delicious** Italian dish (一道美味的意大利菜)

副词 (Adverb) 是用来修饰动词、形容词、其他副词以及全句的词。按照形式，副词可以分为简单副词、复合副词和派生副词。按照意义，副词可以被细分为方式副词、方向副词、时间副词、强调副词等。按照句法作用，可以分为句子副词、连接副词、关系副词等。例如：

- 例如: 1) 简单副词: just (刚刚) only (仅仅)
- 2) 复合副词: somehow (不知怎地) somewhere (在某处)
- 3) 派生副词: interesting → interestingly (有趣地)
- 4) 方式副词: quickly (迅速) awkwardly (笨拙地)
- 5) 方向副词: outside (外面) inside (里面)
- 6) 时间副词: recently (最近) always (总是)
- 7) 强调副词: very (很) fairly (相当)

数词 (Numeral) 是表示数目多少或者先后顺序的一类词。表示数目多少的叫做基数词 (Cardinal numeral)。表示顺序先后的叫做序数词 (Ordinal numeral)。

- 例如: 1) 基数词: one (1) nineteen (19)
- 2) 序数词: first (第一) fiftieth (第五十)

代词 (Pronoun) 是代替名词以及起名词作用的短语、子句和句子的一类词。代词的词义信息较弱，必须通过上下文来确定。代词主要可以分为人称代词 (Personal Pronoun)、物主代词 (Possessive Pronoun)、自身代词 (Self Pronoun)、相互代词 (Reciprocal Pronoun)、指示代词 (Demonstrative Pronoun)、疑问代词 (Interrogative Pronoun)、关系代词 (Relative Pronoun) 和不定代词 (Indefinite Pronoun)。

- 例如: 1) 人称代词:
  - a) 主格: I, you, he, she, it, we, they
  - b) 宾格: me, you, him, her, it, us, them
- 2) 物主代词:
  - a) 形容词性物主代词: my, your, his, her, its, our, their
  - b) 名词性物主代词: mine, yours, his, hers, its, ours, theirs
- 3) 自身代词: myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself
- 4) 相互代词: each other, one another
- 5) 指示代词: this, that, these, those

- 6) 疑问代词: who, whom, whose, which, what
- 7) 关系代词: who, whom, whose, which, that, as
- 8) 不定代词: some, something, somebody, someone, any, anything, anybody, anyone, no, nothing, nobody, no one

冠词 (Article) 是置于名词之前, 说明名词所指的人或事物的一种功能词。冠词不能够离开名词而独立存在。英语中冠词有三种冠词: 定冠词 (Definite article) “the”、不定冠词 (Indefinite article) “a/an”和零冠词 (Zero article)。

介词 (Preposition) 又称前置词, 是用于表示名词或相当于名词的词语与句中其它词语的关系的一类词。介词在句子中不单独作为任何句子成分。介词后面的名词或者相当于名词的词语叫做介词宾语, 与介词共同组合成介词短语。从介词的构成来看, 其主要包含简单介词 (Simple Preposition)、复合介词 (Compound Preposition)、二重介词 (Double Preposition)、短语介词 (Phrasal Preposition)、分词介词 (Participle Preposition)。

- 例如:
- 1) 简单介词: at, in, of, since
  - 2) 复合介词: as for, as to, out of
  - 3) 二重介词: from under, from behind
  - 4) 短语介词: according to, because of
  - 5) 分词介词: including, regarding

连词 (Conjunction) 是连接单词、短语、从句或句子的一类词。在句子中也不单独作为句子成分。按照其构成可以细分为简单连词 (Simple Conjunction)、关联连词 (Correlative Conjunction)、分词连词 (Participial Conjunction)、短语连词 (Phrasal Conjunction)。连词按照其性质可以分为并列连词 (Coordinative Conjunction)、从属连词 (Subordinative Conjunction)。

- 例如:
- 1) 简单连词: and, or, but, if
  - 2) 关联连词: both ... and, not only ... but also
  - 3) 分词连词: supposing, considering
  - 4) 短语连词: as if, as long as, in order that
  - 5) 并列连词: and, or, but, for
  - 6) 从属连词: that, whether, when, because

感叹词 (Interjection) 是用来表示喜怒哀乐等情绪或情感的一类词。感叹词也没有实际意义, 也不能在句子中构成任何句子成分, 但是与全句有关联。

- 例如: ‘**Oh**‘, it’s you. 啊, 是你  
 ‘**Ah**‘, how pitiful! 呀, 多可惜!

在语言学研究中, 对于词性划分的标准、依据甚至目的等都存在大量分歧。到目前为止, 还没有一个被广泛认可的统一划分标准。在不同的语料集中所采用的划分粒度和标记符号也都不尽相同。英语宾州树库 (Penn TreeBank) 使用了 48 种不同的词性, 汉语宾州树库 (Chinese Penn Treebank)

中汉语词性被划分为 33 类，而布朗语料库（Brown Corpus）<sup>[44]</sup> 中则使用了具有 87 个词性。虽然在语言学中词性还有很多需要研究的内容，但是由于词性可以提供关于单词和其周边邻近成分的大量有用信息，词性分析也是自然语言处理中重要的基础任务之一。

## 2.2 词语规范化

在对自然语言文本进行分析前，通常需要对文本进行规范化的处理。文本的规范化处理主要包含句子切分、词语切分、词语规范化等步骤。由于绝大部分语言的句子结束符数量有限，符号歧义性相对容易处理，因此句子切分通过词典结合模板或者有监督分类方法都可以达到较高的准确率。词语规范化（Word Normalization）任务是将单词或词形转化为标准形式，针对有多种形式的单词使用一种单一的形式进行表示。本章中主要讨论词语的规范化问题，包括词语切分、词形分析和词干提取。

### 2.2.1 词语切分

对于绝大部分的印欧语系语言来说，词语之间通常由分隔符区分开来。英语是印欧语系（Indo-European languages）的典型代表，英语句子中绝大部分单词之间都由空格或标点分割。但是以汉语为代表的汉藏语系（Sino-Tibetan languages）的语言中，单词之间通常没有分隔符。因此在对文本进行分析前，通常需要将句子切分为单词序列，称之为词语切分（Word Tokenization）。

词语切分任务可以定义为：给定一个符号串  $x = c_1c_2 \dots c_n$ ，（其中  $c_i$  对于英文来说是字母、数字、标点符号等，对于中文来说是汉字、数字、标点符号等），输出一个词形（Token）序列  $y = t_1t_2 \dots t_m$ ，可能会省略或删除其中的部分标点符号。例如：

输入：Let's first understand what's NLP.

输出：Let<sub>↓</sub>'s<sub>↓</sub>first<sub>↓</sub>understand<sub>↓</sub>what<sub>↓</sub>'s<sub>↓</sub>NLP<sub>↓</sub>.

通过上面的例子可以看到虽然英语句子中绝大部分的单词可以通过空格和标点符号为分隔符进行识别，但是还是存在一些例外情况，例如：缩写（Prof.），日期（02/18/2022），数字（562,000），连字符（upper-case）等。此外，还可以看到“Let's”被划分为了“Let”和“'s”。正是因此，在词语切分的定义中使用了词形。词形（Token）指的是在一个特定文档中的某个能够表达语义含义的字符序列。虽然在大部分情况下词形和单词没有区别，但对于某些场景和算法有必要对单词和词形进行区分。

在英语中，一些特殊的符号和数字也需要完整的保留到一起。比如数字（“67.20”）、时间（22:37）、微博话题标签（# 北京 2022 年冬奥会 #）、Email 地址（cs\_nlp@fudan.edu.cn）等。在特定的应用中有时也会将“Hong Kong”，“Head, Shoulders, Knees and Toes”划分为一个词形。这也使得在某些应用中词语切分与命名实体识别任务（将在第 7 章信息抽取中进行详细介绍）紧密相关。

通常情况下针对英语等印欧语系语言的词语切分任务可以采用基于有限状态自动机（Finite State Automata）融合正则表达式的方法完成。但是针对汉语、日语、阿拉伯语等词语中间没有分

隔符的语言，词语切分问题更加复杂，在后续章节中我们将以中文分词为例进行详细介绍。

## 2.2.2 词形还原

词形还原 (Lemmatization) 是将词的各种变化形式还原其词根的过程。通过词形还原可以实现词语的规范化，单词的不同变化形式统一为词根。

例如：原始输入句：They are working on interesting tasks

词形还原后：they be work on interesting task

词形还原可以通过词形分析 (Morphological Parsing) 完成。词形分析是将一个词分解成为语素的过程。最简单的方法是词典查表法，将每一个词的所有词形变换都存储下来，使用时直接匹配查找。对于英语来说，构造包含所有绝大多数词形的词典能够有效地支撑许多应用场景。由于用词方式的变化和新词的不断出现，需要对该字典进行及时维护。但是，对于某些语言（特别是土耳其语、阿拉伯语等黏着语系的语言）枚举所有词的词形变换则是不可能的。

例如：土耳其语词汇 *uygarlaştıramadıklarımızdanmışsınızcasına* 是由以下 10 项变换组合而成<sup>[45]</sup>：

uygar	+1a	+tr	+ama	+dk	+lar	+mz	+dan	+m	+snz	+casna
civilized	+BEC	+CAUS	+NABL	+PART	+PL	+P1PL	+ABL	+PAST	+2PL	+AsIf

其中除了词根 *uygar* 以外，其他语素的含义如下：

+BEC	“变成”(become)
+CAUS	标识使役动词
+NABL	“不能”(not able)
+PART	过去分词
+PL	名词复数
+P1PL	第一人称复数所有格
+ABL	表来源的离格 (ablative (from/among) case maker)
+PAST	带过去时的间接引语 (indirect/inferential past)
+AsIf	从限定动词 (finite verb) 派生出的副词

可以看到，在一些语言中由于词形变换的复杂性，一个词的原形可能衍生出很多不同的词。采用词典匹配的方法很难达到较好的分析效果。因此，需要更有效率的词形分析算法。典型的词形分析算法包括基于有限状态转换机 (Finite State Transducer) 方法，融合词典和有限状态转换机的方法以及统计机器学习方法等。

## 2.2.3 词干提取

词干提取 (Stemming) 是词形分析的简化版本，其目标是将具有词形变化（通常是屈折或派生）的词语还原为其词干 (Word Stem)。与词形分析不同，词干提取并不要求还原的词干一定与其语言学词根完全一致，只需要将相关的单词映射为统一的词干。甚至词干本身可能并不是一个单词。例如：词干提取算法 Porter Stemmer<sup>[46]</sup> 将 *argue, argued, argues, arguing, argus* 都转换为 *argu.*

最简单的词干提取算法可以通过查询词表的方法获得，这种方法依赖词典所能覆盖的单词数量，并且需要及时更新以应对不断出现的新词。另外一种常见的算法是后缀剥离 (Suffix-stripping)，通过定义一组规则，将特定的后缀从词形中删除。

例如：如果单词以'ed'结尾，则删除'ed'

如果单词以'ing'结尾，则删除'ing'

如果单词以'ly'结尾，则删除'ly'

后缀剥离方法虽然可以很好地处理词语的规则变形，但是无法处理特殊变形（如：ran, took等）。后缀替代 (Suffix Substitution) 算法可以在一定程度上解决上述问题。与后缀剥离不同，后缀替代是定义规则将单词后缀替换为另外一个后缀。

例如：如果单词以'ational'结尾，则替换为'ate' (relational → relate)

如果单词以'ing'结尾，则替换为'ε' (working → work)

如果单词以'zzes'结尾，则替换为'Z' (quizzes → quiz)

Porter Stemmer 就采用了这种后缀替换的方法进行词干提取。

## 2.3 中文分词

以英语为代表的印欧语系中词之间通常有分隔符（空格等）来区分，词可以比较容易地从句子中分割得到。但是以汉语为代表的汉藏语系，以及以阿拉伯语为代表的闪-含语系 (Semitic-Hamitic languages) 中却不包含明显的词之间的分隔符，而是由一串连续的字符构成。因此，针对汉语等语言的处理算法通常首先需要进行词语切分。

本节将以汉语为例介绍词语切分的基本概念以及所面临的主要问题，然后介绍基于词典、基于字统计、基于词统计以及基于神经网络的分词算法，最后介绍常见的中文分词数据集合。

### 2.3.1 中文分词概述

中文分词 (Chinese Word Segmentation, CWS) 是指将连续字序列转换为对应的词序列的过程，也可以看做在输入的序列中添加空格或其他边界标记的过程。中文分词任务可以形式化表示为：给定中文句子  $c_1, c_2, \dots, c_n$ ，其中  $c_i$  为单个字符，输出词序列  $w_1, w_2, \dots, w_m$ ，其中  $w_j$  是中文单词。

例如：复旦大学是中国人自主创办的第一所高等院校。

分词结果：复 | 島 | 大 | 学 | 是 | 中 | 国 | 人 | 自 | 主 | 创 | 办 | 的 | 第 | 一 | 所 | 高 | 等 | 院 | 校 |。

由于汉语中语素绝大部分是单个汉字，很多情况下单独使用时是词，不单独使用时又是构词成分，这使得汉语构词具有很大的灵活性和很强的组词能力。对于新概念的表示不需要创造新的汉字，仅需使用现有汉字就可以灵活地创造新词。但是，正是因为汉语的这些特点，中文分词任务面临了巨大的挑战，主要困难来自以下三个方面：分词规范、歧义切分和未登录词识别。

## 1. 分词规范

汉语中对词的具体界定目前还没有定论。1992 年国家标准局颁布的《信息处理用现代汉语分词规范》中大部分规定都是通过举例和定性描述来体现。例如：“二字或三字词，以及结合紧密、使用稳定的二字或三字词组，一律为分词单位。”然而在实际应用中对“紧密”与“稳定”都很难界定，不可直接用于计算。

北京大学计算语言学研究所为了构造包含 2600 多万字《人民日报》基本标注语料库，制订了词语切分和词性标注规范《北京大学语料库加工规范》<sup>[47]</sup>。针对国家标准分词规范，对分词单位进行了定义和解释。针对人名、地名、机构名、其他专有名词、数词、数量词组、时间词、区别词、述补结构、成语、习用语、非汉字的字符串等情况分别进行了详细的说明。部分标注规范如下所示：

- (1) 人名 (nr)：汉族方式的“姓”和“名”单独切分，“姓”标注为 nrf，“名”标注为 nrg。例如：李/nrf 明/nrg，欧阳/nrf 洪涛/nrg；
- (2) 地名 (ns)：国名不论长短，作为一个切分单位，地名后有“省”、“市”等单字的现代行政区划名称时，不切分开，如果地名后的行政区划有两个以上的汉字，则将地名同行政区划名称切开。例如：中华人民共和国/ns，上海市/ns，[深圳/ns 特区/n]ns；
- (3) 机构名 (nt)：一般是短语型的，较长，且含有地名或人名等专名，按照参考文献 [47] 给出的规范需要先切分，再组合，加方括号标注为 nt。例如：[中国/ns 中文/n 信息/n 学会/n]nt，[复旦/ns 大学/n]nt；
- (4) 数词与数量词组：基数、序数、小数、分数、百分数一律不予切分，约数，前加副词、形容词或后加“来、多、左右”等助数词的应予切分。例如：一百二十三/m，约/d 一百/m 多/m 万/m；
- (5) 时间词：年月日时分秒，按年、月、日、时、分、秒切分，“牛年、虎年”等一律不予切分，标注为 t。例如：2021 年/t 9 月/t 16 日/t，牛年/t；
- (6) 成语习语：四个字的成语或习用语为一个切分单位，除标注其词类标记 i 或 l 外，还要求根据其在句子中的功能进一步标注子类，超过四个字的成语或习用语，一般不予切分，不分子类。例如：胸有成竹/iv，近水楼台先得月/i；

需要注意的是，不同的分词规范之间也存在一定的不同，微软亚洲研究院<sup>[48]</sup>所给出的分词标注规范与《北京大学语料库加工规范》存在很多不同。例如，微软亚洲研究院给出的规范中姓名需要整体标出，含有外文和数字的命名实体应整体一起标注<sup>[48]</sup>。但是《北京大学语料库加工规范》中姓名要分为姓和名两个词。此外，虽然标注规范中尽可能给出了详尽的细节，但是其中还存在一些弹性，由于中文词汇本身具有开放性和动态性，不同人之间也存在认同差异，通用分词标准也是中文分词的难题。

## 2. 切分歧义

由于汉语构词方式的灵活性，使得同一个汉语句子很可能产生多个不同的分词结果，这些不同的分词结果也被称为切分歧义。

例如：南京市长江大桥

切分方式 1：南京市 | 长江大桥

切分方式 2：南京 | 市长 | 江大桥

该例句中“南京”、“南京市”、“市长”、“长江”都是词语，因此同样一个句子可以出现多种切分方式。这种切分歧义在汉语中普遍存在。通常汉语中常见的切分歧义可以归纳为三类：交集型切分歧义、组合型切分歧义和真歧义。

交集型切分歧义是指汉字串 AJB 中，AJ、JB 都可以分别组成词汇，则汉字串 AJB 被称为交集型切分歧义，此时汉字串 J 称作交集串。交集型切分歧义也被称为偶发歧义，当两个有交集的词“偶然”的相邻出现时这样的歧义才会发生。

例如：乒乓球拍卖完了。

切分方式 1：乒乓 | 球 | 拍卖 | 完 | 了 |。

切分方式 2：乒乓 | 球拍 | 卖 | 完 | 了 |。

该例句中存在交集型切分歧义，A、J、B 分别代表“球”、“拍”和“卖”。“球拍”和“拍卖”同时都为合法词汇，它们之间存在有一个交集串。类似的例子还包括：“今天下雨”，“很多云彩”，“北京城市规划”，“中国产品质量”等。

组合型切分歧义是指如果汉字串 AB 满足 A, B, AB 同时为词，则汉字串 AB 被称为组合型切分歧义。组合性切分歧义也称为固有歧义，指的是词固有的属性，不依赖于“偶然”发生的上下文。

例如：他马上过来。

切分方式 1：他 | 马上 | 过来 |。

切分方式 2：他 | 马 | 上 | 过来 |。

该例句中“马上”为组合型切分歧义。A, B, AB 分别代表“马”，“上”和“马上”。类似的情况还包括：“才能”，“应对”，“学会”等。

真歧义是指如果汉字串 ABC 满足多种切分方式下语法和语义均没有问题，只有通过上下文环境才能给出正确的切分结果，则汉字串 ABC 被称为真歧义。

例如：白天鹅在水里游泳。

切分方式 1：白天 | 鹅 | 在 | 水 | 里 | 游泳 |。

切分方式 2：白天鹅 | 在 | 水 | 里 | 游泳 |。

对于这个句子来说，以上两种切分方式在语法和语义上都是正确的，需要考虑句子上下文环境，甚至是篇章内容才能进行正确判断。

上述歧义切分的定义都是从机器识别的角度出发的。而事实上，许多歧义切分在真实的上下文环境中通常不能成立。例如，“平淡”根据定义属于组合型切分歧义，但实际上“平 | 淡”这样的切

分方式能够符合上下文语境的情况非常罕见。根据刘开瑛教授在《中文文本自动分词和标注》中给出的统计，汉语新闻文本中每 1000 个词约出现 16 次交集型切分歧义<sup>[49]</sup>。

### 3. 未登录词识别

未登录词（Out Of Vocabulary, OOV）又称生词（Unknown Words），是指在训练语料中没有出现或者词典当中没有，但是在测试数据中出现的词。根据分词算法所采用的技术不同，未登录词所代表的含义也稍有区别。基于词典的分词方法所指的未登录词是指所依赖的词典中没有的单词。对于完全基于统计方法不依赖词典特征的方法，未登录词则是指训练语料中没有出现的单词。而对于融合词典特征的统计方法，未登录词则是指训练语料和词典中均未出现的词。

汉语具有很强的灵活性，未登录词的类型也十分复杂，可以粗略地将汉语文本中常见的未登录词分为以下类型：

- 新出现的普通词汇：语言的使用会随着时代的变化而演化出新的词，这个过程在互联网环境中显得更为快速。例如：下载，给力，点赞，人艰不拆等。
- 命名实体 (Named Entity)：
  - ①人名（如：杰辛达，周杰伦）；
  - ②地名（例如：新江湾，张江）；
  - ③组织机构名（例如：中国中文信息学会，复旦大学）；
  - ④时间和数字（例如：2021-09-16，正月初四，110 亿人民币）；
- 专业名词：出现在专业领域的词语（例如：图灵机，偶氮二甲酸二乙酯，胞质溶胶）；
- 其他专有名词：新出现的产品名、电影名、书籍名等。

针对中文分词中歧义切分和未登录词造成的损失情况，黄昌宁教授和赵海教授在 Bakeoff-2003 的四个中文分词语料库，对当年最好的多种中文分词算法进行了统计，结果均标明未登录词造成的分词精度损失比歧义切分造成的精度损失至少大 10 倍左右<sup>[50]</sup>。宗成庆教授在新闻领域的语料也进行了类似的实验，结果发现未登录词造成的分词错误超过 98%，其中由命名实体引起的分词错误占到了 55% 左右<sup>[38]</sup>。由此可见，未登录词是中文分词的一个重要难题。

#### 2.3.2 基于最大匹配的中文分词

最大匹配（Maximum Matching）分词算法主要包含前向最大匹配，后向最大匹配以及双向最大匹配等三类。这些算法试图根据给定的词典，利用贪心搜索策略找到分词方案。

前向最大匹配算法的基本思想是，从左向右扫描句子，选择当前位置与词典中最长的词进行匹配，对于句子中的一个位置  $i$ ，依次考虑子串  $c[i : i + L - 1], c[i : i + L - 2], \dots, c[i : i]$ ，其中  $c[i : j] \triangleq c_i c_{i+1} \dots c_j$  表示从第  $i$  个字到第  $j$  个字构成的字串（每一个这样的字串对应于一个候选词）， $L$  表示词典中词的最大长度。当某一个  $c[i : j]$  能够对应字典中的一个词时，输出这个词并从  $j + 1$  开始重复以上的过程直至整个句子被遍历完成。

例如：针对句子“他是研究生物化学的一位科学家”，前向最大分词的过程如表2.2所示。为简

单起见，词典中词语最大长度假设为 4，词表为 {"他", "是", "研究", "生物化学", "的", "一", "位", "科学家"}。

表 2.2 前向最大匹配分词过程示例

时间步	开始位置	候选匹配	输出
1	1	他是研究, 他是研, 他是, 他	他
2	2	是研究生, 是研究, 是研, 是	是
3	3	研究生物, 研究生, 研究, 研	研究
4	5	生物化学, 生物化, 生物, 生	生物化学
5	9	的一位科, 的一位, 的一, 的	的
6	10	一位科学, 一位科, 一位, 一	一
7	11	位科学家, 位科学, 位科, 位	位
8	12	科学家, 科学, 科	科学家

后向最大匹配和正向最大匹配思想相同，区别在于对句子从右向左扫描。双向最大匹配则是同时进行前向最大匹配和反向最大匹配，当两者的分词结果不同时，可以使用启发式的规则决定选取哪一个结果作为最终的输出（例如选择平均词长较大的一个）。

可以看到，基于词典的分词方法具有简单、快速、可控、仅依赖词表等优点。但对于没有在词典中出现的词没有很好的处理方案，同时对于分词歧义的处理能力也不足。

### 2.3.3 基于线性链条件随机场的中文分词

根据中文分词任务定义，我们可以将分词过程看做是对于字的分类。具体来说，对于输入句子中的每一个字  $c_i$ ，根据它在分词结果中的位置赋予不同的标签。可以假设一个字在词语中有四个位置：开始 (B)、中间 (I)、结尾 (E) 以及单独成词 (S)。

例如：输入句子：他是研究生物化学的一位科学家。

分词结果：他 | 是 | 研究 | 生物化学 | 的 | 一 | 位 | 科学家 |。

对应标记：他/S 是/S 研/B 究/E 生/B 物/I 化/I 学/E 的/S 一/B 位/E 科/B 学/I 家/E。/S  
这里的“字”不仅包含汉字，还包含英文字母、数字、标点符号等所有可能出现在汉语文本中的符号。

通过 BIES 标签可以将分词问题转换为字的分类问题。此外，一个字  $c_i$  的分类结果与其周边的字的分类结果有关联。比如  $c_i$  被分类为 B 标签表示一个单词的开头时， $c_{i+1}$  标签就不应该分类为 S 标签表示一个成词的字。因此，中文分词任务也是典型的序列标注问题。可以采用条件随机场等结构化机器学习方法进行解决。

条件随机场 (Conditional Random Field, CRF) 试图对多个变量在给定观测值后的条件概率进行建模。 $x = \{x_1, x_2, \dots, x_n\}$  为观测序列， $y = \{y_1, y_2, \dots, y_n\}$  为对应的标记序列，条件随机场的目标是构建条件概率模型  $P(y|x)$ 。在中文分词任务中，观察序列  $x$  对应输入的字序列  $\{c_1, c_2, \dots, c_n\}$ ，

标记序列为每个字对应的BIES标签。在实际应用中，对序列任务进行建模时，通常使用如图2.1所示的链式结构，即线性链条件随机场（Linear-chain CRF）。

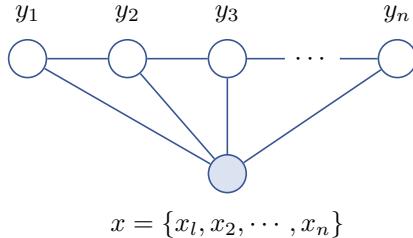


图 2.1 线性链条件随机场结构图

条件随机场使用势函数和图结构上的团来定义条件概率  $P(y|x)$ 。给定观测序列  $x$ ，线性链式条件随机场主要包含两种关于标记变量的团：单个标记变量  $y_i$  和相邻的标记变量  $y_{i-1}, y_i$ 。选用指数势函数并引入特征函数（Feature Function），条件概率则定义为：

$$P(y|x) = \frac{1}{Z(x)} \exp \left( \sum_j \sum_{i=2}^n \lambda_j t_j(x, y_i, y_{i-1}, i) + \sum_k \sum_{i=1}^n \mu_k s_k(x, y_i, i) \right) \quad (2.1)$$

$$Z(x) = \sum_y \exp \left( \sum_j \sum_{i=2}^n \lambda_j t_j(x, y_i, y_{i-1}, i) + \sum_k \sum_{i=1}^n \mu_k s_k(x, y_i, i) \right) \quad (2.2)$$

其中  $t_j(x, y_i, y_{i-1}, i)$  是转移特征函数（Transition feature function），用于刻画相邻标记之间的相关关系以及观测序列对它们的影响； $s(x, y_i, i)$  是状态特征函数（Status feature function），用于刻画观测序列对标记变量的影响； $\lambda_j$  和  $\mu_k$  为参数； $Z(x)$  为规范化因子，在所有可能的输出序列上进行求和，用于确保公式2.1是正确定义的概率。通常转移特征函数  $t_j$  和状态特征函数  $s_k$  的取值为 0 或 1，当满足特征条件时取值为 1，否则为 0。线性链式条件随机场完全由特征函数  $t_j$ 、 $s_k$  以及其对应的参数  $\lambda_j$  和  $\mu_k$  决定。

针对中文分词任务，典型的转移特征如下：

$$t_j(x, y_i, y_{i-1}, i) = \begin{cases} 1 & \text{if } x_i = \text{“复”} \text{ 并且 } y_i = \text{“B”} \text{ 并且 } y_{i-1} = \text{“E”} \\ 0 & \text{otherwise} \end{cases}$$

表示第  $i$  个观测值为“复”时，相应的标记  $y_i$  和  $y_{i-1}$  很可能分别为 B 和 E。典型的状态特征如下：

$$s_j(x, y_i, i) = \begin{cases} 1 & \text{if } x_i = \text{"上"} \text{ 并且 } y_i = \text{"B"} \\ 0 & \text{otherwise} \end{cases}$$

表示第  $i$  个观测值为“上”时，相应的标记  $y_i$  很可能为 B。

如何设计有效的特征函数对于序列标注任务是至关重要的。针对中文分词问题，可以使用模板的方式从当前字的上下文中构建。表2.3列出了中文分词任务常用的模板。其中  $T(c)$  表示字符  $c$  的类型，包括阿拉伯数字、中文数字、标点符号、英文字母等。基于特征模板和训练语料，可以自动生成转移特征以及状态特征。

表 2.3 基于线性链条件随机场的中文分词常见模板

模板名	描述
$c_k(k = -2, -1, 0, 1, 2)$	$c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}$ 位置字符
$c_k c_{k+1}(k = -2, -1, 0, 1)$	$c_{i-2}c_{i-1}, c_{i-1}c_i, c_ic_{i+1}, c_{i+1}c_{i+2}$ 双字组
$c_{-1}c_1$	$c_{i-1}c_{i+1}$ 双字组
$T(c_k)(k = -2, -1, 0, 1, 2)$	$T(c_{i-2}), T(c_{i-1}), T(c_i), T(c_{i+1}), T(c_{i+2})$ 位置字符类型

基于线性链条件随机场中文分词方法可以有效地平衡训练语料中出现的词语和未登录词，并且可以使用模板特征引入词典信息。相较于基于词典的方法，基于线性链条件随机场中文分词方法通常可以省略未登录词的识别模块。关于线性链条件随机场的模型训练和预测算法可以参阅李航博士《统计学习方法（第二版）》第 11 章中的相关内容<sup>[51]</sup>。

### 2.3.4 基于感知器的中文分词

中文分词可以定义为将连续字序列转换为对应的词序列的过程。使用  $x = \{c_1, c_2, \dots, c_n\}$  表示输入字序列， $y = \{w_1, w_2, \dots, w_m\}$  表示输出词序列， $F(x)$  表示最优分词结果。根据上述定义中文分词可以形式化地表达为：

$$F(x) = \arg \max_{y \in \text{GEN}(x)} \text{SCORE}(y) \quad (2.3)$$

其中  $\text{GEN}(x)$  代表对于每一个输入句子  $x$  可能的所有候选输出， $\text{SCORE}(y)$  为针对分词结果  $y$  的评分函数。

基于感知器中文分词方法，将每一个分词后的单词序列  $y$  定义为一个特征向量  $\Phi(x, y) \in \mathbb{R}^d$ ，其中  $d$  代表模型中的特征数量，评分函数  $\text{SCORE}(y)$  定义为由向量  $\Phi(x, y)$  与参数  $\alpha \in \mathbb{R}^d$  间的点积：

$$\text{SCORE}(y) = \Phi(x, y) \cdot \alpha \quad (2.4)$$

将中文分词任务转化为上述问题后，需要解决如下三个问题：

**问题 1：词序列预测** 在给定模型参数  $\alpha$  和输入字序列  $x = \{c_1, c_2, \dots, c_n\}$  的情况下，如何得到最优的词序列  $y = \{w_1, w_2, \dots, w_m\}$ ，即模型解码算法。

给定模型参数情况下，对输入句子的词序列预测问题，根据公式2.3的定义需要计算所有候选分词结果得分。但是，每一个句子都有数量十分庞大的候选分词结果，如果将所有可能的结果都枚举一遍的话，搜索空间将变得非常巨大，使得我们无法有效地进行训练与推断。针对这一问题，常见的解决方式是使用集束搜索（Beam Search）算法进行解码。集束搜索是一种常用的限制搜索空间的启发式算法，在每一步解码过程中，从上一步解码的所有候选集中选取前  $K$  个得分最高的结果继续解码，而舍弃得分排在第  $K$  名之后的所有候选结果。集束搜索可以理解为一种“松弛”过的贪心算法，它并不能保证一定会得到得分最高的候选序列。算法2.1给出了基于感知器中文分词集束搜索算法详细流程。

---

#### 代码 2.1: 基于感知器中文分词解码算法

---

```

输入: 待分词汉字序列  $x = \{c_1, c_2, \dots, c_n\}$ 
输出: 分词结果  $y = \{w_1, w_2, \dots, w_m\}$ 
src = [[]], tgt=[]; // 初始化;
for i = 1 to n do
    foreach item ∈ src do
        item1 =  $c_i$ ;                      // 当前字作为新词的开始;
        item2 = item[item.length]+ $c_i$ ;      // 当前字附加到 item 最后一个候选词上;
        tgt 中添加 item1 和 item2 ;
    end
    使用评分函数 SCORE 对 tgt 中所有分词结果进行打分;
    对 tgt 中的评分结果进行排序，保留前 K 个;
    src = tgt ;
    tgt =[];
end
return src[1]; // 返回 src 中最好结果

```

---

基本思路是：针对输入的句子  $x$ ，解码器每次读入一个字  $c_i$ ，根据候选词队列，采用两种方法扩充候选结果：(1) 作为下一个词的开始；(2) 添加到上一个候选词的末尾。对现有的候选分词结果进行评分，保留得分最高的前  $K$  个候选分词结果。重复上述过程，直到句子结束，输出得分最高的分词结果。

**问题 2：模型参数学习** 在给定训练语料  $\{x_i, y_i\}$  的情况下，如何调整模型参数  $\alpha$ ，使其能针对训练语料得到最好的分词结果，即模型参数学习算法。

对于模型参数  $\alpha$  的学习问题，可以使用感知器算法进行训练。对训练语料中每一个句子，根据现

有模型参数进行解码得到分词结果，与正确答案进行比对，如果结果错误则更新参数  $\alpha$ 。算法对整个训练语料迭代  $T$  轮。算法2.2给出了训练算法的详细流程。也可以采用平均感知器（Average perceptron）算法避免训练过程中的过拟合问题。

---

**代码 2.2: 基于感知器算法的评分函数训练算法**


---

```

输入: 训练数据  $D = (x_i, y_i)$ 
输出: 模型参数  $\alpha$ 
for  $i = 1$  to  $T$  do //  $T$  轮迭代;
  foreach  $(x, y) \in D$  do
     $z = \arg \max_{y \in \text{GEN}(x)} \text{SCORE}(y);$            // 使用算法 2.1 给出的解码算法 ;
    if  $z \neq y$  then
       $\alpha = \alpha + \Phi(x, y) - \Phi(x, z);$ 
    end
  end
end
return  $\alpha$ 

```

---

**问题 3: 特征定义** 给定输入字序列  $x$  和分词后的词序列  $y$ ，如何定义特征对词序列进行描述，并能够区分分词序列的优劣，即构建特征向量  $\Phi(x, y) \in \mathbb{R}^d$ 。

针对特征向量  $\Phi(x, y)$  的定义问题，感知器算法所需的特征由一系列人工选取的特征值组成，包含字、词以及长度信息。在训练和解码时会使用特征模板将解码得到的序列映射到特征向量。Zhang 和 Clark 在其论文中所使用的具体特征模板<sup>[52]</sup> 如表2.4所示。

表 2.4 基于感知器的中文分词算法所使用特征模板样例<sup>[52]</sup>

编号	模板	编号	模板
1	单词 $w$	8	所有单词的第一个与最后一个字符 $c_1$ 和 $c_2$
2	二元单词 $w_1 w_2$	9	字符 $c$ 的前一个词 $w$
3	单字符单词 $w$	10	单词 $w$ 之后的第一个字 $c$
4	初始字符 $c$ 以及长度 $l$	11	两个连续单词的第一个字符 $c_1$ 和 $c_2$
5	终止字符 $c$ 以及长度 $l$	12	两个连续单词的最后一个字符 $c_1$ 和 $c_2$
6	由空格隔开的字符 $c_1$ 和 $c_2$	13	单词长度 $l$ 以及之前的词 $w$
7	二元字符 $c_1 c_2$	14	单词的长度 $l$ 以及之后的单词 $w$

比较基于线性链条件随机场的中文分词的方法所使用的特征模板（如表2.3所示），以及本节所介绍的基于感知器的中文分词算法所使用的特征模板（如表2.4所示），可以看到基于感知器的

方法可以使用词作为特征，而基于线性链条件随机场的方法只能使用字作为特征。因此在 2.3.3 节所介绍的以字为单位作为分类目标的方法也称为基于字的中文分词算法，本节所介绍的以词为基础的方法也称为基于词的中文分词算法。

### 2.3.5 基于双向长短期记忆网络的中文分词

随着深度学习技术的发展，很多中文分词算法也采用了基于神经网络模型。循环神经网络（Recurrent Neural Network, RNN）相较于前馈神经网络等要求固定输入长度的神经网络结构而言，更适用于处理长度不固定的序列数据。特别符合文本、语音等在内的数据特性，广泛应用于自然语言处理任务的很多任务中。长短期记忆网络（Long Short-Term Memory, LSTM）<sup>[53, 54]</sup> 是循环神经网络的一个变体，可以在一定程度上缓解简单循环神经网络的梯度消失和梯度爆炸问题。

LSTM 网络循环单元结构如图 2.2 所示，网络引入了新的内部状态（Internal State） $c_t \in \mathbb{R}^D$ ，专门用来进行信息传递。此外，LSTM 网络还引入了门控机制（Gating Mechanism）来控制信息传递路径。通过遗忘门  $f_t$  控制上一个时刻的内部状态  $c_{t-1}$  需要遗忘多少信息。输入门  $i_t$  用来控制当前时刻的候选状态  $\tilde{c}_t$  有多少信息需要保存。输出门  $o_t$  控制当前时刻内部状态  $c_t$  有多少信息需要输出给外部状态  $h_t$ 。

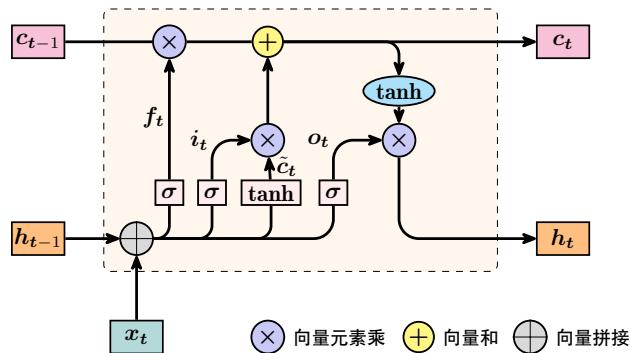


图 2.2 LSTM 网络循环单元结构

输入门  $i_t$ 、输出门  $o_t$  和遗忘门  $f_t$  的计算方式为：

$$i_t = \sigma(\mathbf{W}_i x_t + \mathbf{U}_i h_{t-1} + \mathbf{b}_i) \quad (2.5)$$

$$f_t = \sigma(\mathbf{W}_f x_t + \mathbf{U}_f h_{t-1} + \mathbf{b}_f) \quad (2.6)$$

$$o_t = \sigma(\mathbf{W}_o x_t + \mathbf{U}_o h_{t-1} + \mathbf{b}_o) \quad (2.7)$$

其中  $\sigma(\cdot)$  为 Logistic 函数。候选状态  $\tilde{c}_t$ 、内部状态  $c_t$  以及隐藏输出  $h_t$  通过如下公式计算：

$$\tilde{c}_t = \tanh(\mathbf{W}_c x_t + \mathbf{U}_c h_{t-1} + \mathbf{b}_c) \quad (2.8)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (2.9)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (2.10)$$

更为详细的介绍请参阅邱锡鹏教授《神经网络与深度学习》的第六章<sup>[55]</sup>。

在自然语言处理的很多任务中，一个时刻的输出不但与过去某个时刻的信息相关，也与后续时刻的信息相关。双向长短记忆网络（Bidirectional LSTM, BiLSTM）是用来建模上述问题的一种方法。BiLSTM 是由两层长短记忆网络组成，它们结构相同但是信息传递的方向不同。双向长短记忆网络还可以结合条件随机场，更有效地利用结构化学习和神经网络的特点，在很多自然语言处理任务上都取得了很好的效果。图2.3给出了一个使用 BiLSTM 网络结合条件随机场（BiLSTM+CRF）进行分词的模型框架。

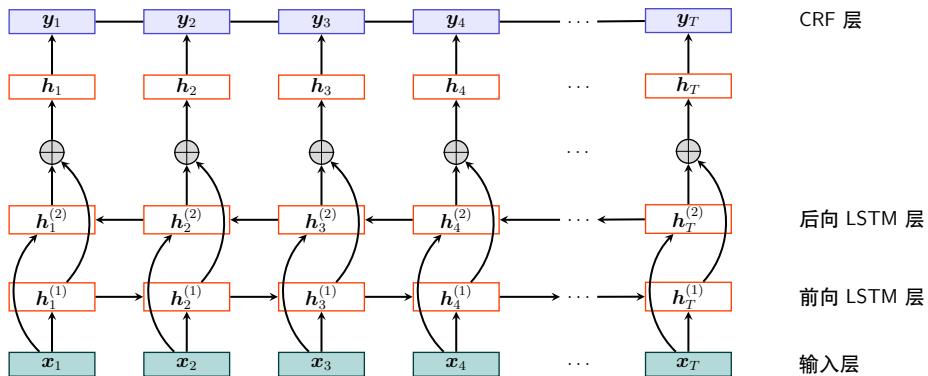


图 2.3 基于 BiLSTM+CRF 的神经网络分词模型框架

在基于神经网络的分词算法中，通常采用与基于字的统计方法类似的问题建模方法，将分词任务转换为字的序列标注任务，对于给定一个中文句子  $x = \{c_1, c_2, \dots, c_T\}$ ，根据它在分词结果中的位置以及所采用的标签系统（例如：“BIES”等），输出标签序列  $y = \{y_1, y_2, \dots, y_T\}$ 。具体模型如图2.3所示，BiLSTM-CRF 主要包含三层：输入层、双向长短记忆网络层和 CRF 层。在输入层，需要将每个字转换为低维稠密的字向量（Character Embedding） $x_i$ 。

BiLSTM 层采用双向 LSTM，其主要作用是提取句子特征。将句子中的每个字向量序列  $x_1, x_2, \dots, x_T$  输入到双向 LSTM 各个时间步，再将正向 LSTM 输出的隐状态序列  $h_1^{(1)}, h_2^{(1)}, \dots, h_T^{(1)}$  与反向 LSTM 隐状态序列的  $h_1^{(2)}, h_2^{(2)}, \dots, h_T^{(2)}$  按位置进行拼接  $h_i = h_i^{(1)} \oplus h_i^{(2)}$ ，从而得到完整的隐

状态序列。

对于给定的长度为  $T$  的输入  $[\mathbf{x}]_1^T$ , 定义网络的输出矩阵为  $f_{\theta}([\mathbf{x}]_1^T)$  (简写为  $f_{\theta}$ ) , 其中  $[f_{\theta}]_{i,t}$  表示参数为  $\theta$  的网络对于句子  $[\mathbf{x}]_1^T$  的第  $t$  个字的第  $i$  标签的打分。同时定义转移值矩阵  $\mathbf{A}$ , 其中  $[\mathbf{A}]_{i,j}$  为相邻的两个字的标签从  $i$  标签到第  $j$  标签的值,  $[\mathbf{A}]_{i,0}$  为开始标签为第  $i$  标签的值。由于转移值矩阵也是模型参数的一部分, 因此整个模型的参数  $\tilde{\theta} = \theta \cup \{[\mathbf{A}]_{i,j}, \forall i, j\}$ 。对于输入句  $[\mathbf{x}]_1^T$  的某个特定标签序列  $[i]_1^T$  定义得分为转移值和网络值的和, 具体公式如下:

$$s([\mathbf{x}]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^T ([\mathbf{A}]_{[i]_{t-1}, [i]_t} + [f_{\theta}]_{[i]_t, t}) \quad (2.11)$$

通过 Softmax 函数可以将某个标签序列的得分根据所有可能标签序列  $[j]_1^T$  的得分进行归一化, 得到标签序列的条件概率:

$$P([i]_1^T | [\mathbf{x}]_1^T, \tilde{\theta}) = \frac{e^{s([\mathbf{x}]_1^T, [i]_1^T, \tilde{\theta})}}{\sum_{\forall [j]_1^T} e^{s([\mathbf{x}]_1^T, [j]_1^T, \tilde{\theta})}} \quad (2.12)$$

由此可以进一步得到对于输入  $[\mathbf{x}]_1^T$  的正确标签序列  $[y]_1^T$  的条件概率的对数似然 (log-likelihood):

$$\log P([y]_1^T | [\mathbf{x}]_1^T, \tilde{\theta}) = s([\mathbf{x}]_1^T, [y]_1^T, \tilde{\theta}) - \log \left( \sum_{\forall [j]_1^T} e^{s([\mathbf{x}]_1^T, [j]_1^T, \tilde{\theta})} \right) \quad (2.13)$$

基于最大化对数似然目标, 以及公式2.13的线性计算方法<sup>[14, 56]</sup>, 可以根据标注语料训练得到模型参数  $\tilde{\theta}$ 。根据模型参数, 使用维特比 (Viterbi) 算法可以对任意句子预测每个字的标签序列, 从而得到分词结果。

### 2.3.6 中文分词评价方法

中文分词算法效果评测通常也采用统计机器学习算法评测中常用的指标进行对比, 包括: 精确率 (Precision, P)、召回率 (Recall, R)、F 值 (F-Measure)。各指标在中文分词任务中的具体计算方法如下:

$$\text{精确率 (P)} = \frac{\text{算法输出的正确分词结果个数}}{\text{算法输出的全部分词结果个数}} \times 100\% \quad (2.14)$$

$$\text{召回率 (R)} = \frac{\text{算法输出的正确分词结果个数}}{\text{测试集合中全部答案个数}} \times 100\% \quad (2.15)$$

$$F_{\beta} = \frac{(\beta^2 + 1) \times P \times R}{P + R} \times 100\% \quad (2.16)$$

通常 F 值计算时设置  $\beta$  为 1, 因此 F 值又称为 F1 值。

在本章第 2.3.1 节中文分词概述中提到, 未登录词是中文分词任务的难点之一, 也是影响中文

分词算法效果的重要因素。因此，在中文分词评测中通常还会对召回率进一步细分为未登录词召回率 ( $R_{OOV}$ ) 和词典词召回率 ( $R_{IV}$ )。

$$\text{未登录词召回率 } (R_{OOV}) = \frac{\text{算法输出的未登录词正确结果个数}}{\text{测试集合中未登录词个数}} \times 100\% \quad (2.17)$$

未登录词召回率也是评价一个中文分词算法的主要指标。

此外，由于有些中文分词算法会利用词典、无标注数据等除训练数据外的资源，为了能够更好地对模型本身的效果进行评价，评测有时还会区分封闭测试（Closed Test）和开放测试（Open Test）。封闭测试仅允许使用给定的训练语料，而开放测试可以使用任意资源。

### 2.3.7 中文分词语料库

中文分词算法的训练通常依赖大规模标注语料。大规模中文分词语料集的建设也是推动中文分词算法快速发展的一个不可或缺的因素。SIGHAN 2005 和 SIGHAN 2008 是两组最常用的中文分词语料集合。SIGHAN 是国际计算语言学协会中文处理特别兴趣组，组织了多次包含多家机构的数据的中文处理相关评测（International Chinese Language Processing Bakeoff）。常见中文分词语料库如表2.5所示。本节将介绍目前较为广泛使用的部分中文分词语料集合。

表 2.5 常用中文分词语料库汇总

语料库名称	单词数量	简/繁体
北京大学分词语料库 (PKU)	110 万	简体
香港城市大学分词语料库 (CITYU)	145 万	繁体
微软研究院分词语料库 (MSR)	237 万	简体
Academia Sinica (AS)	545 万	繁体
中文宾州树库 6.0 (CTB 6.0)	78 万	简体
中文宾州树库 7.0 (CTB 7.0)	120 万	简体
中文宾州树库 8.0 (CTB 8.0)	162 万	简体
中文宾州树库 9.0 (CTB 9.0)	208 万	简体
微博分词语料库 (WordSeg-Weibo)	46 万	简体

#### 1. 北京大学分词语料库 (PKU)

北京大学分词语料库（也称为人民日报语料库）是由北京大学计算语言学研究所与富士通公司（Fujitsu）合作发布的包含 110 万字的分词语料集合。数据来源为《人民日报》，字符总数约为 182 万。同时还制定了《现代汉语语料库加工规范》，规定了分词要与词性标注进行结合的原则。例如，“复合”方式可将两个构词成分结合成一个新词。规范中规定了许多新词的构词方式，也规定了一般性名词和专有名词切分的规范。

## 2. 香港城市大学分词语料库 (CITYU)

香港城市大学分词语料库是香港城市大学语言资讯科学研究中心制作的繁体中文分词数据集，对包含 145 万字的原始数据进行了切分。他们制定了相关的切词规则，在名词、数词、时间词、略语、二字结构、三字复合词、四字词、短语、叠词、非汉字部分等十个方面进行了详细的规范。另外还对其他方面进行了补充，古语方言和熟语等不进行切分，例如，“踏破铁鞋无觅处”这句话不进行分词。

## 3. 微软研究院分词语料库 (MSR)

微软研究院分词语料库是由微软亚洲研究院 (MSRA) 整理，在 230 万字的简体中文原始语料上进行划分，采用 CP936 的编码方式。该语料库将词汇分为三大类，词汇词（如：教授，高兴，吃饭），命名实体（如：蒙特利尔，中央民族乐团）和陈述词。其中陈述词类别较多，包含日期、时间、持续时间、量词、电话号码等。

# 2.4 词性标注

词性是词语的基本属性，根据其在句子中所扮演的语法角色以及与周围词的关系进行分类。词性标注 (Part-of-speech Tagging, POS Tagging) 是指在给定的语境中确定句子中各词的词性<sup>[1]</sup>。词性标注是句法分析的基础，也是自然语言处理中一项重要的基础任务。

词性标注的主要难点在于歧义性，即一个词可能在不同的上下文中具有不同的词性。这些具有多个词性的词语称为兼类词。例如：“book”可以表示名词“书”，也可以表示动词“预定”，“good”可以表示形容词“好”，也可以表示名词“货物”，“China”可以表示专有名词“中国”，也可以表示普通名词“瓷器”等等。因此需要结合上下文来确定词在句子中所对应的词性。另一方面，兼类词多为常用词，而且越是常用词，其用法就越多。英语“like”具有动词、名词、介词等多种词性。针对北京大学计算语言学研究院 200 万字语料库统计，发现兼类词所占比例仅有 11%，但是出现的次数却达到了 47%<sup>[57]</sup>。对 Brown 语料库的统计也发现超过 80% 的词通常只有一个词性。

此外，由于在语言学研究中，还没有一个被广泛认可的统一词性划分标准，在不同的语料集中所采用的划分粒度和标记符号也都不尽相同，这也一定程度上对词性标注问题的研究造成了困难。表2.6列出了在宾州树库 (Penn Treebank, PTB) 中所使用的词性。宾州大学汉语树库 (Chinese Penn TreeBank, CTB) 中汉语词性被划分为 33 类，北京大学计算语言学研究所给出的语料库加工规范中包含 26 个基础词性，74 个扩展词性。由于词性表以及词性定义有许多不同的变种，词性标注的结果与这些标注密切相关。本节中将主要以 PTB 标准为例。

### 2.4.1 基于规则的词性标注

基于规则的词性标注算法是最早应用于词性标注任务的一类方法，其核心思想是利用词典和搭配规则针对词语和上下文进行分析，从而得到句子中每个词语词性的方法。早期通常采用人工

表 2.6 宾州树库中的词性标签

标签	描述	标签	描述
CC	并列连词	CD	数字
DT	限定词	EX	<u>there</u>
FW	外来词	IN	介词或从属连词
JJ	形容词	JJR	形容词比较级
JJS	形容词最高级	LS	列表项标记
MD	情态助动词	NN	名词单数
NNS	名词复数	NNP	专有名词单数
NNPS	专有名词复数	PDT	前限定词
POS	所有格结束词	PRP	人称代名词
PRP\$	物主代词	RB	副词
RBR	副词比较级	RBS	副词最高级
RP	小品词	SYM	符号
TO	to	UH	叹词
VB	动词	VBD	动词过去式
VBG	动词现在进行式	VBN	动词过去分词
VBP	动词一般现在式 非第三人称单数	VBZ	动词一般现在式 第三人称单数
WDT	Wh-限定词	WP	Wh-代词
WP\$	所有格 Wh-代词	WRB	Wh-副词

的方法来构建规则，随着机器学习算法的不断发展以及资源的不断完善，也出现了一些基于机器学习方法的规则自动学习算法。在本节中我们将重点介绍基于转换的 Brill Tagger 方法<sup>[58]</sup>。

Brill Tagger 是一种利用错误驱动方法学习转换规则的词性标注算法。在 Brown 语料库上仅使用 71 个规则就得到接近 95% 的分析准确率。其分析算法的主要过程如下：

- (1) **初始化：**对于词典中包含的词语，根据词语最常使用的词性设置初始值；对于词典中没有的单词根据词性分析结果设置初始值（例如：以大写字母开头的设置为专有名词）。
- (2) **规则转换：**根据补丁规则对初始标注进行转换，补丁规则包含以下三类：
  - (a) 如果某单词词性为 a，并且其所在上下文为 C，那么将其词性转换为 b；
  - (b) 如果某单词词性为 a，并且其具有词汇属性 P，那么将其词性转换为 b；
  - (c) 如果某单词词性为 a，并且其周边范围 R 内有一个词汇具有属性 P，那么将其词性转换为 b；

例如：补丁规则“NN VB PREV-TAG TO”表示，如果一个单词被标注为了 NN（名词），并且它前面的单词标注为了 TO（不定式“to”），那么将这个单词的词性转换为 VB（动词）。可以用用于解决类似“to book a hotel”中对于单词 book 的词性默认标注错误的问题。

Brill Tagger 中对于补丁规则的学习方法采用了基于错误驱动的有监督模板学习方法。首先根据现有的初始词典和补丁模板对训练语料进行分析，将错误的分析结果汇总为三元组  $\langle tag_a, tag_b, n \rangle$  形式，表示一个单词的词性应该为  $tag_b$ ，但是在评测语料中有  $n$  次都被标注为了词性  $tag_a$ 。根据所得到的三元组，利用以下模板生成补丁规则：

- 前一个（或者后一个）单词被标注为了 z
- 前面第二个（或者后面第二个）单词被标注为了 z
- 前面两个（或者后面两个）单词某一个被标注为了 z
- 前面三个（或者后面三个）单词某一个被标注为了 z
- 前一个单词被标注为了 z，并且后一个单词被标注为了 w
- 前一个单词被标注为了 z，并且前面第二个（或者后面第二个）单词被标注为了 w
- 当前单词是（不是）首字母大写
- 前一个单词是（不是）首字母大写

根据每个  $\langle tag_a, tag_b, n \rangle$  三元组，以及利用上述模板得到的补丁规则，可以计算利用该规则可以修复的错误标记数，以及利用该规则所引入的新的错误数。根据上述数值，选择改进最大的补丁规则加入规则列表中，并进行新一轮的分析和规则生成。

基于错误驱动的规则方法可以在一定程度上缓解人工规则抽取的时间成本和人力成本。在词性标注问题中取得了不错的效果。但是其效果严重依赖于训练语料的规模和质量，同时也较难处理未登录词。此外，受到规则模板复杂度的限制，其效果通常也低于基于统计机器学习的方法。

#### 2.4.2 基于隐马尔可夫模型的词性标注

隐马尔可夫模型（Hidden Markov Model, HMM）又称隐马尔科夫模型，是马尔可夫过程扩充而来的一种随机过程，其基本理论是由数学家 Baum 及其同事构建并逐步完善。随着隐马尔可夫模型在语音识别领域取得巨大成功<sup>[59]</sup>，其在自然语言处理众多序列标注任务中也得到了广泛应用并取得了非常好的效果。一个隐马尔可夫模型可用如下 5 个参数定义：

- $N$ : 状态数。所有的状态记为  $S = \{s_1, s_2, \dots, s_N\}$ 。系统在  $t$  时刻的状态记为  $q_t$ 。 $Q = \{q_1, q_2, \dots, q_T\}$ ，为长度为  $T$  的状态序列。
- $M$ : 观察值数。所有的可能观察值记为  $V = \{v_1, v_2, \dots, v_M\}$ 。系统在  $t$  时刻的观测值记为  $o_t$ 。 $O = \{o_1, o_2, \dots, o_T\}$ ，为长度为  $T$  的观测序列。
- $\pi$ : 初始状态概率。 $\pi = [\pi_i]_{1 \times N}, \pi_i = P(q_1 = s_i), 1 \leq i \leq N$ ，表示初始时刻  $t = 1$  时处于某个状态  $s_i$  的概率。
- $A$ : 状态转移概率矩阵。 $A = [a_{ij}]_{N \times N}, a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq N$ ，表示在时刻  $t$  处于状态  $s_i$  的条件下，下一时刻  $t + 1$  转移到状态  $s_j$  的概率。
- $B$ : 观测概率矩阵。 $B = [b_j(k)]_{N \times M}, b_j(k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq N, 1 \leq k \leq M$ ，表示在时刻  $t$  处于状态  $s_j$  的条件下，观测到  $v_k$  的概率。

为了简化起见, 隐马尔可夫模型可以表示成  $\lambda = (A, B, \pi)$ 。 $M, N$  也隐含地已经包含在  $A, B, \pi$  中。隐马尔可夫模型的三个主要问题是:

**问题 1: 观测概率计算** 在给定模型  $\lambda = (A, B, \pi)$  的情况下, 如何根据观测序列  $O = \{o_1, o_2, \dots, o_T\}$  计算  $P(O|\lambda)$ , 即在给定模型情况下, 如何观测序列的概率。

**问题 2: 状态序列预测** 在给定模型  $\lambda = (A, B, \pi)$  和观测序列  $O = \{o_1, o_2, \dots, o_T\}$  的情况下, 如何得到与该观测序列最匹配的状态序列  $Q = \{q_1, q_2, \dots, q_T\}$ , 即如何根据观测序列推断出隐藏的状态序列。

**问题 3: 模型参数学习** 在给定观测序列  $O = \{o_1, o_2, \dots, o_T\}$  情况下, 如何调整模型参数  $\lambda = (A, B, \pi)$  使得该序列的  $P(O|\lambda)$  最大, 即如何训练模型使其能最好地建模观测序列。

关于问题 1, 问题 2 以及问题 3 的求解方法可以参阅李航博士《统计学习方法 (第二版)》第 10 章中的相关内容<sup>[51]</sup>。

针对词性标注任务, 使用隐马尔可夫模型可以按照如下方式构建和学习模型。 $N$  为词性数,  $S = \{s_1, s_2, \dots, s_N\}$  为词性表, 包含所使用到的所有词性信息。 $M$  为单词数,  $V = \{v_1, v_2, \dots, v_M\}$  为单词词表, 包含所有单词。给定一个由  $T$  个单词组成的句子  $W = w_1, w_2, \dots, w_T$ , 即相当于观测序列  $O = \{o_1, o_2, \dots, o_T\}$ ,  $o_i$  为句子中第  $i$  个单词  $w_i$ 。状态序列  $Q = \{q_1, q_2, \dots, q_T\}$  则表示输入句子中单词对应的词性。根据训练语料, 可以使用最大似然估计的 Baum-Welch 方法高效地得到模型参数。在此基础上, 针对输入的句子可以利用维特比 (Viterbi) 算法应用动态规划求解状态路径, 从而得到对应的词性。图 2.4 给出了基于词性标注的隐马尔可夫模型概率图模型样例。

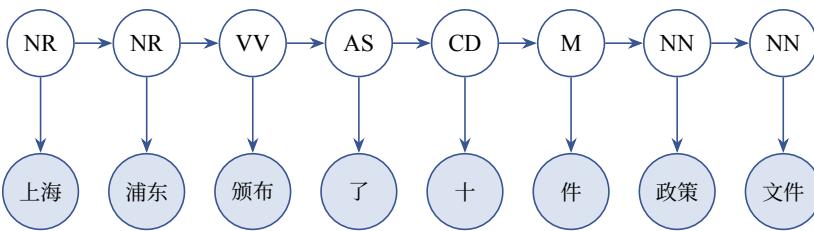


图 2.4 词性标注隐马尔可夫模型概率图模型样例

在实际应用过程中, 使用隐马尔可夫模型进行词性标注, 通常还需要解决两个问题: 长句子和未登录词。在《人民日报》语料库中, 有些句子非常长, 甚至会超过 120 个字。Gabriel García Márquez 所著的魔幻现实主义小说《族长的秋天》中, 很多句子“一逗到底”, 长度甚至超过 1000 个词。虽然这种长句子在真实环境中很少出现, 但是对于模型的设计和实现都带来了一定的挑战。因此, 通常会限定一个句子中单词的最大数量。如果一个句子超过了所设定的最大长度, 则寻找距离最大长度最近的标点, 并在标点处将句子截断。对于词典当中没有出现的未登录词, 由于观测概率矩阵  $B$  中不存在, 也需要进行特殊处理。第一种做法是在单词表中增加一个“未登录词”项,

同时在观测概率矩阵中设置该词以同样的概率观察到所有标记类别。这种做法较为粗糙，在本章中我们介绍过词的一个重要分类角度是开类词和闭类词。未登录词通常属于名词、动词、形容词等开放类词语。其中人名、地名、机构名等名词又占据了很大的比例。因此第二种做法是引入词法规则，对人名、地名、数词、副词等进行判断。此外，还可以根据更大规模的未登录词的词性统计，从而设定更合理的观测概率。

### 2.4.3 基于卷积神经网络的词性标注

在深度神经网络应用于自然语言处理任务之前，绝大多数自然语言处理算法依赖于特征工程。Collobert 等人<sup>[14]</sup>在 2011 年所提出的“从零开始的 NLP”框架利用统一的具有多个隐藏层的神经网络解决了多个自然语言处理任务，省去了特征工程的步骤，推动了深度学习在自然语言处理任务中的快速发展。在本节我们以词性标注任务为例介绍该方法。基于卷积神经网络的词性标注算法神经网络结构如图2.5所示。

首先通过表查询（Lookup Table） $\text{LT}_{\mathbf{W}}(\cdot)$ 将单词通过表查询转换为词性向量表示，词向量的维度是 $d_{wrd}$ 。

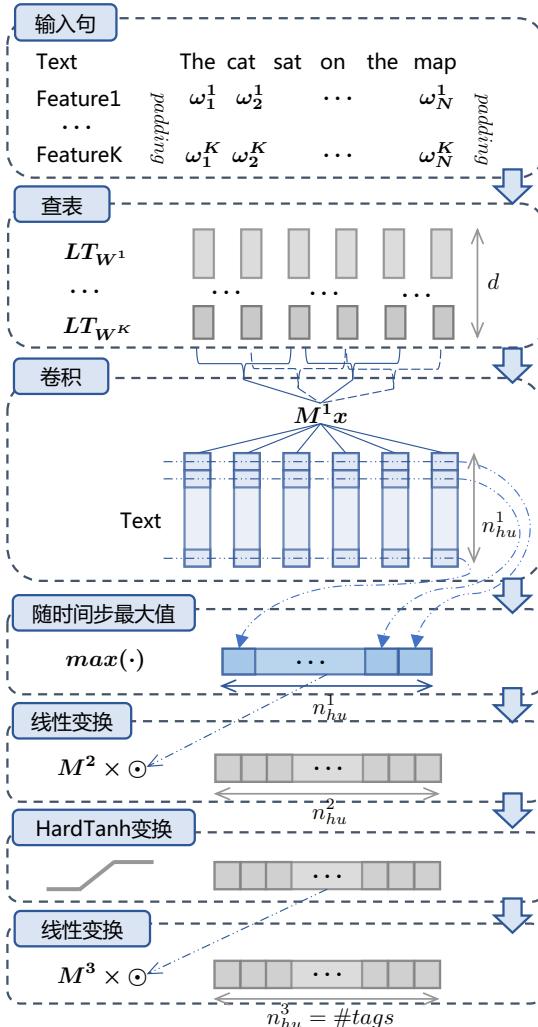
$$\text{LT}_{\mathbf{W}}(w) = \langle \mathbf{W} \rangle_w^1 \quad (2.18)$$

其中 $\mathbf{W} \in \mathbb{R}^{d_{wrd} \times |\mathbb{D}|}$ ， $\mathbb{D}$ 是包含有限个单词的字典， $\langle \mathbf{W} \rangle_w^1 \in \mathbb{R}^{d_{wrd}}$ 表示 $\mathbf{W}$ 矩阵的第 w 列。 $\mathbf{W}$ 矩阵也是要进行学习的参数。对于给定的任意一句包含 $T$ 个单词的句子 $[w]_1^t$ ，通过表查询层对序列中的每个单词进行转换，得到如下表查询层输出矩阵：

$$\text{LT}_{\mathbf{W}}([w]_1^T) = \left( \langle \mathbf{W} \rangle_{[w]_1}^1, \langle \mathbf{W} \rangle_{[w]_2}^1 \dots \langle \mathbf{W} \rangle_{[w]_T}^1 \right) \quad (2.19)$$

除了单词本身之外，还可以提供一些其他特征，例如该单词在词典中最常见的词性等信息。因此，可以将单词更一般地表示为 $K$ 个离散特征 $w = \mathcal{D}^1 \times \mathcal{D}^2 \times \dots \mathcal{D}^K$ ， $\mathcal{D}^k$ 是第 k 维特征的字典。 $\text{LT}_{W^k}(\cdot)$ 是每维特征的查询表， $\mathbf{W}^k \in \mathbb{R}^{d_{wrd}^k \times |\mathcal{D}^k|}$ 是第 k 维特征的嵌入向量矩阵， $d_{wrd}^k \in \mathcal{N}$ 是用户给定的向量维度。对于一个单词 $w$ ，其特征向量的维度 $d_{wrd} = \sum_k d_{wrd}^k$ ，通过表查询得到连接后的向量：

$$\text{LT}_{\mathbf{W}^1, \dots, \mathbf{W}^K}(w) = \begin{pmatrix} \text{LT}_{\mathbf{W}^1}(w_1) \\ \vdots \\ \text{LT}_{\mathbf{W}^K}(w_K) \end{pmatrix} = \begin{pmatrix} \langle \mathbf{W}^1 \rangle_{w_1}^1 \\ \vdots \\ \langle \mathbf{W}^K \rangle_{w_K}^1 \end{pmatrix} \quad (2.20)$$

图 2.5 基于卷积神经网的词性标注模型结构<sup>[14]</sup>

由此，可以得到如下表查询层输出矩阵：

$$LT_{W^1, \dots, W^K}([w]_1^T) = \begin{pmatrix} \langle W^1 \rangle_{[w_1]_1}^1 & \dots & \langle W^1 \rangle_{[w_1]_T}^1 \\ \vdots & & \vdots \\ \langle W^K \rangle_{[w_K]_1}^1 & \dots & \langle W^K \rangle_{[w_K]_T}^1 \end{pmatrix} \quad (2.21)$$

在表查询层后连接的是卷积层 (Convolutional Layer)，根据所设置的窗口大小  $d_{win}$ ，将每个单词周边的单词拼接起来构成具有  $d_{wrd}d_{win}$  维度的向量：

$$f_\theta^1 = \langle \text{LT}_W([w]_1^T) \rangle_t^{d_{win}} = \begin{pmatrix} \langle \mathbf{W} \rangle_{[w]_{t-d_{win}/2}}^1 \\ \vdots \\ \langle \mathbf{W} \rangle_{[w]_t}^1 \\ \vdots \\ \langle \mathbf{W} \rangle_{[w]_{t+d_{win}/2}}^1 \end{pmatrix} \quad (2.22)$$

$f_\theta^1$  会被送入单层或者多层的卷积层，第  $l$  层的第  $t$  列向量可以根据如下公式计算得到：

$$\langle f_\theta^l \rangle_t^1 = \mathbf{W}^l \langle f_\theta^{l-1} \rangle_t^{d_{win}} + b^l \quad \forall t \quad (2.23)$$

在同一层中  $\mathbf{W}^l$  为相同参数。对于  $f_\theta^l$  中每一维在公式2.23计算完成后，都要进行非线性变化，可以采用如下方式：

$$[f_\theta^l]_i = \text{HardTanh}([f_\theta^l]_i), \quad (2.24)$$

$$\text{HardTanh}(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases} \quad (2.25)$$

通过公式2.23得到的特征向量更多地反映了局部特征，并且数量与句子长度相关。为了得到全局特征并且维度固定的特征向量，需要引入池化层 (Pooling Layer)，在这里使用的是随时间推移最大化 (Max Over Time) 方法。给定通过卷积层计算得到的矩阵  $f_\theta^{l-1}$ ，池化层输出的向量  $f_\theta^l$  计算如下：

$$[f_\theta^l]_i = \max_x [f_\theta^{l-1}]_{i,t} \quad 1 \leq i \leq n_{hu}^{l-1} \quad (2.26)$$

针对通过池化层计算得到的向量  $f_\theta^l$  需要继续进行线性变换，再利用公式2.24进行非线性变换，之后再叠加新的线性层后完成特征提取工作。线性变换层对于输入  $f_\theta^{l-1}$  利用如下公式计算得到其输出  $f_\theta^l$ ：

$$f_\theta^l = \mathbf{W}^l f_\theta^{l-1} + b^l \quad (2.27)$$

在分类阶段，采用句子级别对数似然方法 (Sentence-Level Log-Likelihood)。相关算法以及公式在第2.3.5节基于 BiLSTM-CRF 方法进行分词部分进行了详细介绍，也可通过参考文献 [14] 查看详细算法和公式推导。除了网络输出矩阵  $f_\theta ([x]_1^T)$  (简写为  $f_\theta$ ) 之外，引入转移值矩阵  $A$ ，对

于输入句  $[x]_1^T$  的某个特定标签序列  $[i]_1^T$  定义为转移值和网络值的和。基于最大化对数似然目标，可以根据标注语料训练得到模型参数  $\tilde{\theta}$ 。根据模型参数，使用维特比（Viterbi）算法可以获得任意句子中每个词的词性。

#### 2.4.4 词性标注评价方法

词性标注问题通常转换为多分类问题，并且每个单词仅会输出一个词性结果。因此，词性标注算法的评测通常采用准确率（Accuracy）和宏平均 F1（Macro-F1）两种评测指标。

准确率具体计算方法如下：

$$\text{准确率 (Accuracy)} = \frac{\text{算法输出的正确结果个数}}{\text{算法输出的全部结果总数}} \times 100\% \quad (2.28)$$

宏平均 F1（Macro-F1）需要首先计算每个词性标签的 F1 值，再计算所有词性标签 F1 值的平均值。以名词词性标签的精确率和召回率为例，具体计算方法如下：

$$P_{\text{名词}} = \frac{\text{算法输出的正确的名词结果个数}}{\text{算法输出的名词结果总数}} \times 100\% \quad (2.29)$$

$$R_{\text{名词}} = \frac{\text{算法输出的正确的名词结果个数}}{\text{测试集合中正确的名词结果总数}} \times 100\% \quad (2.30)$$

$$F1_{\text{名词}} = \frac{2 \times P_{\text{名词}} \times R_{\text{名词}}}{P_{\text{名词}} + R_{\text{名词}}} \quad (2.31)$$

宏平均 F1（Macro-F1）的具体计算方法如下：

$$\text{Macro-}F1 = \frac{1}{n} \sum_{i \in POS} F1_i \quad (2.32)$$

其中  $n$  为词性标签的数量，POS 为词性标签集合。

从上述计算公式中可以看到，宏平均 F1 对含有较少单词的标签类别更敏感，每个类别标签的 F1 值同等重要。而准确率对于仅含有少量单词的标签类别的效果不敏感。从衡量算法的能力角度，宏平均 F1 相对可以更好地反映类别不均衡情况下的算法性能。

#### 2.4.5 词性标注语料库

通过本章的介绍可以知道词性标注算法的训练过程都依赖于标注语料集合。对不同算法的效果进行对比也依赖于标准测试集合。常见词性标注语料库如表2.7所示。本节将介绍几种常见的包含词性标签的语料库。

表 2.7 常见词性标注语料库汇总

语料库名称	单词数量	语言
英语宾州树库 (PTB)	117 万	英文
通用依存树库 (UD V2.0 CoNLL 2017)	281 万	多语言
RIT-Twitter	1 万	英文
ARK-Twitter	3 万	英文
中文宾州树库 6.0 (CTB 6.0)	78 万	中文
中文宾州树库 7.0 (CTB 7.0)	120 万	中文
中文宾州树库 8.0 (CTB 8.0)	162 万	中文
中文宾州树库 9.0 (CTB 9.0)	208 万	中文

### 1. 英语宾州树库

英语宾州树库 (English Penn Treebank, PTB) 是最知名和最常用的短语结构句法树库之一。在对句子的语法树标注的同时，也标注了句子中单词的词性信息。因此，英语宾州树库也是最常使用的词性语料库之一。该语料库中包含多个部分，其中 WSJ-PTB 是最常用于词性标注评测的集合，其原始数据来自于 1989 年的华尔街日报文章，按照 PTB(V2) 的标注策略进行标注，包含 49208 个句子，1173766 个单词，48 种不同的词性标签。

### 2. 中文宾州树库

中文宾州树库 (Chinese Penn Treebank, CTB) 是目前最常用的大规模中文短语结构句法标注语料库之一。1998 年开始构建，2016 年发布的最新的 Chinese Treebank 9.0 版本，包含中文新闻网站、政府文书、杂志文章、新闻群组、广播对话节目、博客等各类不同来源的 3726 篇文章，共计 132076 个句子，2084387 个单词，3247331 个中文和外文字符。在 CTB 中，汉语词性被划分为 33 类，包括 4 类动词和谓语形容词，3 类名词，1 类处所词，1 类代词，3 类限定词和数词，1 类量词，1 类副词，1 类介词，8 类语气词和 8 类其他词。

### 3. 通用依存树库

通用依存树库 (Universal Dependencies, UD) 是一个为多种语言开发的跨语言一致的依存句法树库项目。其词性标注采用了 Google 通用词性标签<sup>[60]</sup>，由十二个通用词类构成的标记集，包括 NOUN (名词), VERB (动词), ADJ (形容词), ADV (副词), PRON (专有名词), DET (限定词和冠词), ADP (介词和后置词), NUM (数字), CONJ (连接词), PRT (小品词), ‘.’ (名词所有格) 和 X (其他)。除了标记集之外，还为来自 22 个语言的 25 个不同的树库开发了一个从细粒度词性标记到这个通用标记集的映射。

## 2.5 延伸阅读

本章中介绍了中文分词和词性标注任务，这两个任务都是典型的序列标注任务，除了基于词典的中文分词算法之外，本章中介绍的其他算法都采用有监督分类算法。因此，这些方法通常都面临跨领域处理效果差、依赖大规模训练语料等挑战。此外，还存在基于特征表示的方法依赖人工设计特征函数，而经典深度学习模型无法有效利用知识等问题。针对上述问题，近年来有大量工作从不同方面开展研究。

针对中文分词任务，为了解决有监督分类算法依赖大规模训练数据的问题，大量的研究工作针对不同分类算法，开展了融合有标注数据和无标注数据的半监督算法研究，包括基于部分标签学习的条件随机场算法（Partial-label Learning with CRF）<sup>[61]</sup>、非参数贝叶斯模型（Nonparametric Bayesian）<sup>[62]</sup>、基于图的标签传播算法（Graph-based label propagation）<sup>[63]</sup>、协同训练方法（Co-training）<sup>[64]</sup>等。为了解决深度学习方法不能有效利用已有知识的问题，近年来一些研究试图将词典信息融合到深度学习模型，包括将字在词典中的特征信息编码<sup>[65]</sup>、通过损失函数编码词典特征<sup>[66]</sup>、词典增强的自适应注意力机制<sup>[67]</sup>等方法。针对中文分词问题中标注规范不统一的问题，多标准中文分词研究近年来也受到了越来越多的重视，研究者提出了对抗多标准学习（Adversarial Multi-Criteria Learning）<sup>[68]</sup>、转换长短时记忆网络（Switch-LSTMs）<sup>[69]</sup>等方法试图同时学习多种分词标准。为了使得深度学习模型可以更好地融合词语级别特征，研究者提出了基于转移中文分词模型<sup>[70]</sup>、栅格化循环神经网络<sup>[71]</sup>等方法。针对如何更好地评价中文分词算法在不同情况下的性能以及不同场景下的鲁棒性问题，一些工作从细粒度的中文分词算法评价<sup>[72]</sup>以及中文分词算法鲁棒性研究<sup>[73]</sup>等方面开展研究。

针对词性标注任务，为了能够利用大规模的无标注数据，先后提出了期望正规化（Expectation Regularization）<sup>[74]</sup>、基于平均感知器半监督算法<sup>[75]</sup>、半监督密集近邻（Semi-supervised Condensed Nearest Neighbor）<sup>[76]</sup>、基于图信号的半监督主动学习（Active Semi-supervised Learning）<sup>[77]</sup>等方法。针对算法训练数据和应用领域不同时，性能大幅度降低的问题，跨领域和领域自适应方法在词性标注任务中也有大量的研究工作，包括基于鲁棒表示的方法<sup>[78]</sup>、基于熵的数据选择方法<sup>[79]</sup>、分层贝叶斯（Hierarchical Bayesian）<sup>[80]</sup>、基于对抗神经网络算法<sup>[81]</sup>、基于强化学习的训练数据选择方法<sup>[82]</sup>等。此外，针对汉语的词性标注问题，通常采用流水线方式，首先对汉语句子进行分词，之后再对分词结果进行词性标注。由于流水线模式会造成错误的传递问题，因此也有一些工作将中文分词和词性标注问题联合建模，包括分类目标合并<sup>[83]</sup>、级联线性模型（Cascaded Linear Model）<sup>[84]</sup>、词栅格重排序（Word Lattice Reranking）算法<sup>[85]</sup>、堆叠子词模型<sup>[86]</sup>、双向注意力机制<sup>[87]</sup>等方法。

中文分词和词性标注是自然处理的底层任务，其分析效果对后续任务有很大影响，长期以来都是自然语言处理研究的重点。此外，中文分词和词性标注也是典型的序列标注任务，结构化机器学习领域很多工作也将上述任务作为算法效果验证目标。

## 2.6 习题

- (1) 语言学中词和语素的定义分别是什么？其主要的不同是什么？
- (2) 英语中句子切分主要解决什么歧义问题？如何使用有监督分类算法进行句子切分？
- (3) 中文分词中歧义切分包含几种主要的类别？针对每种歧义类别试举几例，并说明具有歧义的切分方式。
- (4) 如何在基于线性链条件随机场的中文分词算法中引入词典特征？
- (5) 如何处理词性标注算法中的未登录词？
- (6) 如何同时 BiLSTM-CRF 方法进行分词和词性标注联合建模？

## 3. 句法分析

任何人类语言都具备构造无限数量句子的能力<sup>[43]</sup>，可以通过增加形容词、副词、关系小句、介词短语等方法把任意的句子进一步地创造。因此，无法将一种语言按照词典的方式把所有句子存储起来。但是，通过语言学研究发现，句子并非词语的随意组合，而是按照一定规则结合起来的离散单位组成<sup>[88]</sup>。句法（Syntax）就是研究这些自然语言中不同成分组成句子的方式以及支配句子结构并决定句子是否成立的规则。句法反映了句子中词、词序以及层级结构之间的关系。通过句法规则，可以将词语组合成短语，将短语组合成句子，或者确定某个句子是否符合某一语言正确的词序。句法规则还阐释了词的分组与其意义的相关性，并在一定程度上解释了语言的无限性。句法分析具有非常重要的作用，是自然语言处理中的经典问题。

本章首先介绍语言学中句法基本概念，在此基础上介绍成分句法分析算法、依存句法分析算法以及常用的句法分析语料库。

### 3.1 句法概述

句法是现代语言学研究中的重要课题，有大量的句法理论（Syntactic Theory）相关研究。句法理论在很多语法理论研究中也都占据了主要部分。每种自然语言都可以看作由正确句子构成的集合。通常一种自然语言中包含的词和语素可以达到几万甚至几十万，而句子的长度可以超过 100 个单词，甚至可以认为是无限长。因此，不可能采用穷举的方式将一个语言中所有正确的句子保存起来。语法（Grammar）就是指自然语言中句子、短语以及词等语法单位的语法结构与语法意义的规律<sup>[89]</sup>。根据语法就可以判断不同成分组成句子的方式以及决定句子是否成立。语法学的外延相关广泛，甚至有些语言学家认为音系学和语义学都包含在语法学中。在本书中我们采用比较狭义的语法学定义，即认为语言学不包含音系学和语义学，因此狭义的语法学的研究基本等同于句法学。语言学家自 19 世纪 50 年代以来，构建了大量表达明确并且形式化的语法理论，对自然语言句法分析提供了理论支撑。

语法理论在构建时，一个重要的问题是该理论是基于成分关系还是基于依存关系。如果一种语法理论基于成分关系，那么该理论就属于成分语法（Constituent Grammar），也称短语结构语法（Phrase Structure Grammar）。成分语法主要包含：范畴语法（Categorical Grammar）、词汇功能语

法 (Lexical Functional Grammar)、最简方案 (the Minimalist Program) 等。如果一个语法理论基于依存关系，那么该理论就属于依存语法 (Dependency Grammar)。依存语法主要包含：文本-意义理论 (Meaning Text Theory)、词格理论 (Lexicase)、功能生成描述理论 (Functional Generative Description) 等。

本节将分别针对成分语法和依存语法理论进行简单介绍。

### 3.1.1 成分语法理论概述

Noam Chomsky 于 1957 年发表的《Syntactic Structures》奠定了成分语法的基础<sup>[90]</sup>。成分 (Constituent) 又称短语结构，是指一个句子内部的结构成分。成分可以独立存在，或者可以用代词替代，又或者可以在句子中的不同位置移动。

例如：他正在写一本小说。

“一本小说”是一个成分

在此基础上，根据不同成分之间是否可以进行相互替代而不会影响句子语法正确性，可以进一步的将成分进行分类，某一类短语就属于一个句法范畴 (Syntactic Category)。比如“一本小说”、“一所大学”等都属于一个句法范畴：名词短语 (None Phrase, NP)。句法范畴不仅仅包含名词短语 (NP)、动词短语 (VP)、介词短语 (PP) 等短语范畴，也包含名词 (N)、动词 (V)、形容词 (Adj) 等词汇范畴。除此之外还包含功能范畴 (包括冠词、助动词等)。

句法范畴之间不是完全对等的，而是具有层级关系。例如：一个句子可以由一个名词短语和一个动词短语组成，一个名词短语可以由一个限定词和一个名词组成，一个动词短语又可以由一个动词和一个名词短语组成。我们可以定义短语结构规则 (Phrase Structure Rules) 又称改写规则或重写规则，是对句法范畴间的关系进行形式化描述。通常可以用  $X \rightarrow Y Z W \dots$  表示，其中  $X$  表示短语名称，“ $\rightarrow$ ”表示“改写为”，“ $Y Z W \dots$ ”定义了短语  $X$  的结构，如果  $Y Z W$  是短语，则需要构造出它们的规则。

例如：

- (1)  $S \rightarrow NP VP$
- (2)  $NP \rightarrow Det N$
- (3)  $VP \rightarrow V NP$

成分语法就是由句法范畴以及短语结构规则定义的语法。由于短语结构规则具有递归性，可以使短语和句子无限循环组合。这也说明了语言的创造性和无限性。如图3.1和图3.2所示，一个句子根据成分语法分析得到的层级结构，可以使用成分语法树进行表示。

由于成分语法局限于表层结构分析，不能彻底解决句法和语义问题，因此存在非连续成分、结构歧义等问题。如图3.3和图3.4所示，对于句子“The boy saw the man with telescope”有两种可能的合法的句法结构树，不同的树结构对应不同的语义。第一种句法树表示“男孩使用望远镜看到了这个男人”，第二种句法树表示“男孩看到了一个拿着望远镜的男人”。

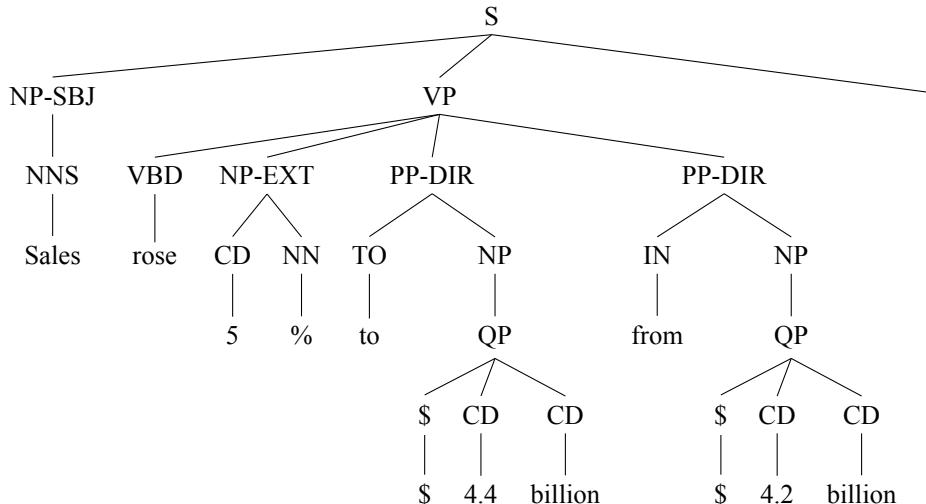


图 3.1 Penn Treebank 3.0 成分句法树样例

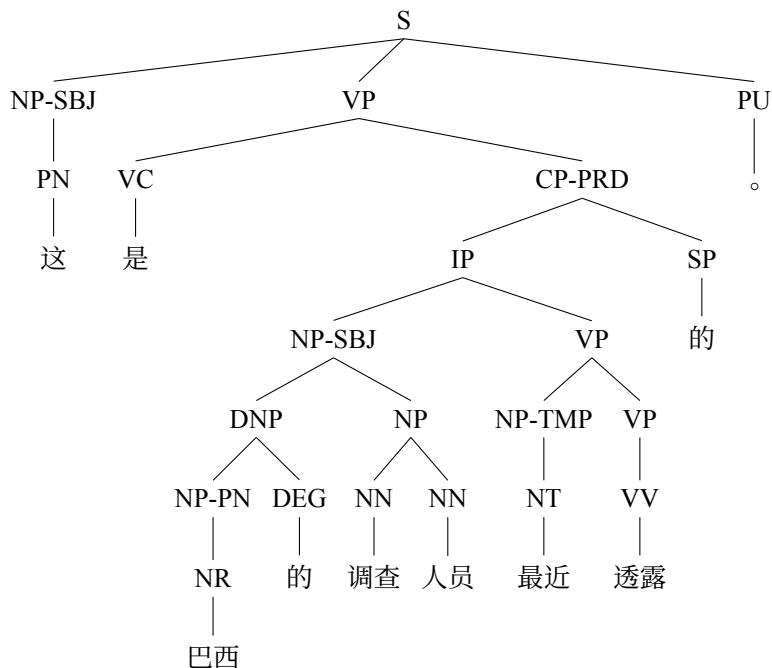


图 3.2 Chinese Treebank 7.0 成分句法树样例

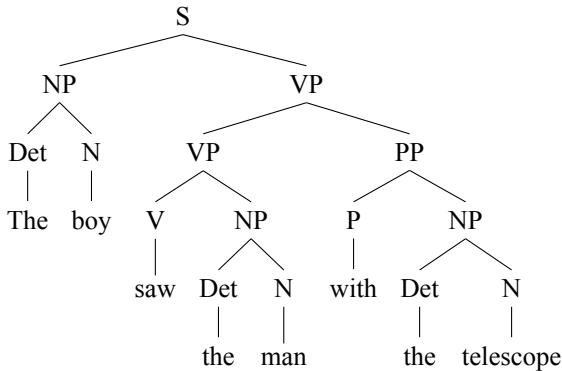


图 3.3 句子 The boy saw the man with telescope 的第一种成分句法树

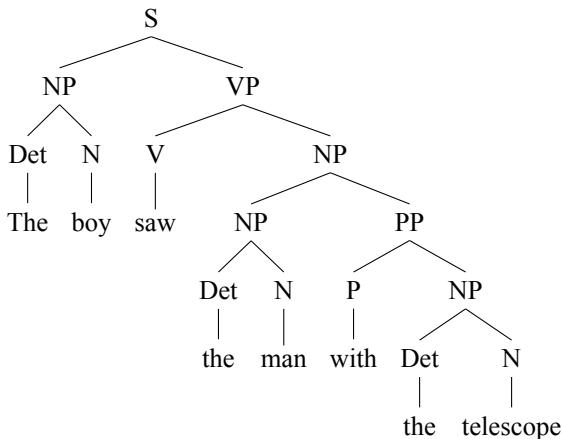


图 3.4 句子 The boy saw the man with telescope 的第二种成分句法树

### 3.1.2 依存语法理论概述

Lucien Tesnière 于 1959 年发表的《*Eléments de syntaxe structurale*》奠定了句法依存关系研究的基础<sup>[91]</sup>。在基于依存关系的语法中，句子中的每个成分对应句法结构中的唯一一个节点。两个成分之间的依存关系是二元的非对称关系，具有方向性，一个成分是中心语，另一个成分依附于中心语存在，关系从中心语成分指向依存成分。中心成分称为中心词或支配者（Governor, Regent, Head），依存成分也称为修饰词或从属者（Modifier, Subordinate, Dependency）。例如：“读书”中“读”是中心语，“书”依存于“读”，有多种方式可以用于表示依存关系，如图3.5所示。

两个单词之间是否存在依存关系？单词之间谁处于支配地位？谁处于从属地位？建立这些词与词之间关系的依据是什么？很多依存句法理论从不同方面对上述问题进行了回答。配价（Valency）

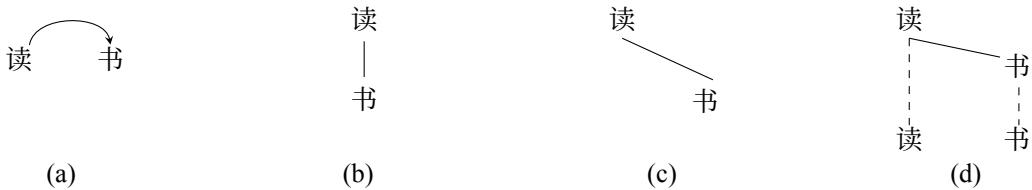


图 3.5 “读书”的依存关系表示实例

理论是其中最为经典的论著之一。这里“价”的概念是由 Lucien Tesnière 从化学中的“化合价”的概念引入语言学研究。价是词语的一个属性，表示某个词语与其他词语结合的能力。配价模式 (Valency pattern) 则是描述了某一个具有特定意义的词的出现语境，以及当一个词出现在一个特定的模式下时，还有哪些词语会出现在这个模式下及其语义角色。通过对不同词类的支配和被支配能力（配价）以及词类间依存关系类型的定义，就可以得出某种具体的依存句法。

利用配价理论，可以将汉语里的动词 V 根据其价数，即属于几价动词，分为以下四类：

- (1) 零价动词不强制与某个行动元关联的动词，例如：地震、刮风、下雨、下雪……
- (2) 一价动词强制与一个行动元关联的动词，例如：病、醉、休息、咳嗽、游泳
- (3) 二价动词强制与两个行动元关联的动词，例如：爱、采、参观、讨论、改良……
- (4) 三价动词强制与三个行动元关联的动词，例如：给、送、告诉、退还、赔偿……

有一种特殊的关系是并列关系，构成这种关系的元素之间不存在支配与被支配关系。但是为了能够使得与其他的关系统一，也通过并列关系标记将其纳入句法树中，通常将第一个词作为中心语，其余的词作为该词的从属成分。这种情况下，在树结构上表现出来的层级关系是一种伪层级。

词语间的依存关系还可以根据语法关系定义为不同的类型，Carroll 等人<sup>[92]</sup> 将依存关系细分为 20 种，并给出了关系之间的层级结构。Marneffe 等人<sup>[93]</sup> 在上述工作的基础上对依存关系进行了进一步的细化，定义了 48 种依存关系，主要分为论元依存关系和修饰语依存关系两大类。图 3.6 给出了一个包含依存关系类型的句法树的样例。

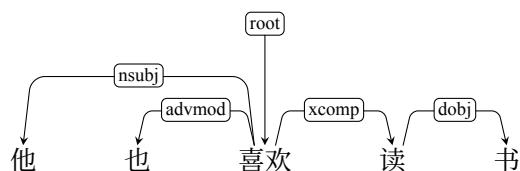


图 3.6 句子“他也喜欢读书”的依存句法树

通常依存句法树引入 *root* 做为句子或者单棵树的主要支配者，是树的根节点。一般情况下动词是 *root* 节点的直接从属成分，其余节点应该直接或者间接依存于动词节点。没有动词的句子除外，但是应该存在一个句子成分做为 *root* 的从属成分。依存句法树中的每个节点只能有一个支配者，但是可以有多个从属者。一个符合语法的句子中，所有节点应该是相连的，不允许存在游离

在外的节点。

依存语法中根据依存成分与中心语或姐妹成分在语序上的关系，可以分为符合投射性原则和违反投射性原则两类。在依存层级中如果每个依存成分与其中心或姐妹成分相邻，那么该依存层级就是符合投射性原则（Projectivity），如图3.7所示，依存树中没有任何两线交叉，因此是投射性的。如果依存成分的相邻成分使其与中心语分类，那么就是违反投射性原则的。如图3.8所示，成分“about this book”与其中心语不相邻，因此存在交叉线，违反了投射原则。这种现象在依存句法中通常称为远距离依赖（Long-distance Dependencies）。

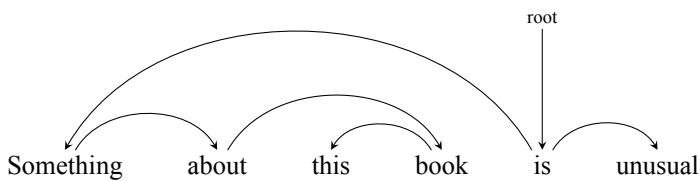


图 3.7 符合投射原则依存句法树样例

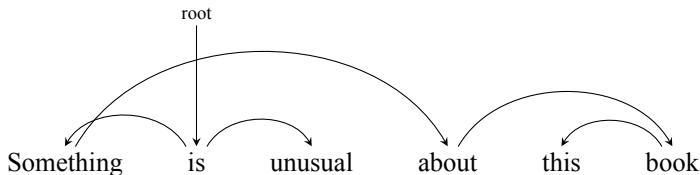


图 3.8 违反投射原则依存句法树样例

## 3.2 成分句法分析

成分句法分析（Constituency Parsing）是对给定句子根据成分语法中制定的规则构建其所对应的结构树的过程。在本章第3.1节中介绍了关于成分语法的基本概念和组成部分，成分语法主要包含句法范畴和短语结构规则两个部分。句法范畴又包含短语范畴、词汇范畴和功能范畴。短语结构规则通常可以由  $X \rightarrow Y Z W \dots$  形式进行表示。成分语法与数学系统中的上下文无关文法（Context-Free Grammar, CFG）组成非常类似，因此上下文无关文法是目前最常用的模拟英语以及其他语言的成分语法的数学系统。上下文无关文法  $G$  包含以下四个部分：

终结符  $\Sigma$  (Terminal Symbols)：与语言中词汇对应的符号，用  $u, v, w$  等小写罗马字母表示；

例如：上海， walk

非终结符  $N$  (Non-terminals)：对应词组等聚集类或概括性符号，用  $A, B, C$  等大写字母表示；

例如：NP (名词短语)， VP (动词短语)， N (名词)， 介词 (P)

初始符  $S$  (Start symbol): 语法中指定的初始符，通常用  $S$  表示；

规则  $R$  (rules): 对应语法中的短语结构规则，从初始符号开始可以造出合法句子的规则集合，在上下文无关语法中，一个规则的左边是一个单独的非终结符，右边是一个或者多个终结符或非终结符组成的有序序列  $(\Sigma \cup N)^*$ 。用  $\alpha, \beta, \gamma$  等小写希腊字母表示  $(\Sigma \cup N)^*$ 。

例如:  $S \rightarrow NP VP$ ,  $NP \rightarrow Det Adj N$ ,  $Det \rightarrow the$

对一个句子进行句法分析的过程可以看做对一个句子搜索所有可能的路径空间，从中发现正确的句法树的过程。搜索过程受到两个约束限制，一个是句子本身，另外一个是语法。根据成分句法的定义，叶子节点一定是句子中的单词，中间节点与其子节点需要符合语法定义。这两种约束也产生了大多数分析算法所采用的搜索策略：自底向上 (Bottom-up) 和自顶向下 (Top-down)。

由于句法结构具有歧义，因此句法分析中最重要的工作之一也是如何消除歧义。成分语法中的结构歧义主要有两种：附着歧义 (Attachment ambiguity) 以及并列连接歧义 (Coordination ambiguity)。图3.3和图3.3所给出的关于句子“The boy saw the man with telescope”的两种分析结果就是附着歧义的一个例子，其核心就是“with telescope”附着于动词 saw 还是名词短语 the man。并列连接歧义也是常见的结构歧义之一，例如，短语“重要政策和措施”中“重要”可以修饰“政策和措施”整体，也可以修饰“政策”，就可以表示为两种层级结构，“[重要 [政策和措施]]”和“[重要政策] 和 [措施]”。但是由于结构歧义通常伴随语义上的变化，因此句法结构歧义的消除通常需要依赖统计知识、语义知识甚至是语用知识才能更好的完成。

### 3.2.1 基于上下文无关文法的成分句法分析

在给定上下文无关文法  $G$  的情况下，对于给定的句子  $W = \{w_1, w_2, \dots, w_n\}$ ，输出其对应的句法结构，通常有两大类搜索方法：自顶向下和自底向上。自顶向下搜索试图从根节点  $S$  出发，搜索语法中的所有规则直到叶子节点，并行构造所有的可能的树。自底向上的方法是从输入的单词开始，每次都是用语法规则，直到成功构造了以初始符  $S$  为根的树。针对这两大类算法，本节中将分别介绍 CYK 分析算法和移进-规约分析算法。

#### 1. CYK 成分句法分析算法

CYK 算法 (CYK 是 Cocke–Younger–Kasami 的缩写，有时也称为 CKY) 是由 John Cocke、Daniel Younger 以及 Tadao Kasami 分别独立提出的基于动态规划思想的自底向上语法分析算法<sup>[94–96]</sup>。CYK 算法要求所使用的语法必须符合乔姆斯基范式 (Chomsky Normal Form, CNF)，其语法规则被限制为只具有  $A \rightarrow BC$  或  $A \rightarrow w$  这种形式。由于任何上下文无关语法都可以转换为相应的 CNF 语法，因此这种限制并不会对表达能力造成损失。但是这种限制使得基于表的分析算法变得简单和非常易于实现。

根据 CNF 语法形式，句法树的叶子节点为单词，单词的父节点为词性符号，在词性符号层之上每一个非终结符都有两个儿子节点。因此 CYK 算法采用了二维矩阵对整个树结构进行编码。对于一个长度为  $n$  的句子，构造一个  $(n + 1) \times (n + 1)$  的二维矩阵  $T$ ，矩阵主对角线以下全部为 0，

主对角线上的元素由输入句子的终结符号(单词)构成, 主对角线以上的元素  $T_{ij}$  包含由文法 G 的非终结符构成的集合, 这个集合表示输入句子中横跨在位置 i 到 j 之间的单词的组成成分。输入句子中索引从 0 开始, 索引位于输入句子的单词之间, 也可以看成单词之间的间隔指针(例如: <sub>0</sub> 她 <sub>1</sub> 喜欢 <sub>2</sub> 跳 <sub>3</sub> 芭蕾 <sub>4</sub>)。

CYK 算法按照平行于主对角线的方向, 逐层向上填写矩阵中的各个元素  $T_{ij}$ 。如果存在一个正整数 k ( $i + 1 \leq k \leq j - 1$ ), 在文法规则集中具有产生式  $A \rightarrow BC$  并且  $B \in T_{ik}, C \in T_{kj}$ , 那么将 A 合并到矩阵表的  $T_{ij}$  中。为了方便根据矩阵 T 输出句法分析结果, 在矩阵 T 每个单元  $T_{ij}$  中不仅记录非终结符的集合, 还要记录推导路径。逐层计算完成后, 判断一个句子由文法 G 所产生的充要条件是:  $T_{0n} = S$ 。具体过程如算法3.1所示。

---

### 代码 3.1: CYK 句法分析算法

---

```

输入: 语法信息 G 和句子  $w_1 w_2 \dots w_n$ 
输出: 句法树矩阵 T
// 初始化
for i = 1 to n do
     $T_{ii} = w_i;$  // 主对角线上依次放入单词  $w_i$ ;
    foreach  $A | A \rightarrow w_i$  do
         $| T_{(i-1)i} = T_{(i-1)i} \cup A, (A \rightarrow w_i \in G);$  // 依次放入单词  $w_i$  的所有词性;
    end
end
// 平行于主对角线逐层计算
for d = 2 to n do
    for i = 0 to n - d do
        j = i+d;
        for k = i + 1 to j - 1 do
            if  $A \rightarrow BC \in G$  and  $B \in T_{ik}$  and  $C \in T_{kj}$  then
                 $| T_{ij} = T_{ij} \cup \{A, (k, B, C)\};$  // (k, B, C) 用于保存推导路径;
            end
        end
    end
end
return T

```

---

以下通过一个实例来说明 CYK 算法的具体流程。给定如下符合乔姆斯基范式的文法 G 如下:

非终结符号集合:  $N = \{S, N, P, V, VP\}$

终结符号集合:  $\Sigma = \{她, 喜欢, 跳, 芭蕾\}$

规则集合:  $R = \{ (1) S \rightarrow P VP; (2) VP \rightarrow V V; (3) VP \rightarrow VP N;$   
 $(4) P \rightarrow \text{她}; (5) V \rightarrow \text{喜欢}; (6) V \rightarrow - > \text{读}; (7) N \rightarrow \text{芭蕾} \}$

图3.9给出了CYK算法分析句子“她喜欢跳芭蕾”的过程。首先初始化矩阵主对角线上的单词以及单词所对应的词性。在此基础上进行第二层  $T_{02}, T_{13}, T_{24}$  元素的计算, 通过规则集合中  $VP \rightarrow VV$  推导规则, 可以得到在  $T_{13}$  添加  $VP$  以及相应的路径信息。针对第三层  $T_{03}, T_{14}$  元素, 虽然针对  $T_{03}$  元素, 可以根据推导规则  $S \rightarrow P VP$  添加非终结符  $S$ , 但是由于非终结符  $S$  是只能出现在  $T_{0n}$  中, 因此  $T_{03}$  不添加  $S$ 。 $T_{14}$  元素中根据通过规则  $VP \rightarrow VP N$  添加  $VP$  信息。最后一层  $T_{04}$  根据  $S \rightarrow P VP$  规则, 添加  $S$  及其路径信息并完成分析。

## 2. 移进-规约成分句法分析算法

移进-规约成分句法分析算法的基本思想是从左到右扫描输入的包含单词词性对的句子, 使用堆栈和一系列的移进(Shift)和规约(Reduce)操作序列构建句法树<sup>[97]</sup>。

算法初始时堆栈  $S$  为空, 队列  $Q$  中包含整个句子所有单词。最终通过一系列的操作, 在算法结束时堆栈  $S$  中包含一个完整的句法树, 队列  $Q$  为空。所采用的操作包含以下四个:

**移进(Shift):** 将非空队列  $Q$  最左端的单词移入堆栈  $S$  中。

**规约(Reduce):** 根据推导规则, 根据推导规则右侧所包含非终结符数量, 将堆栈  $S$  中的最顶端相应数量元素移出, 然后将利用推导规则产生的新结构压入堆栈中。

**接受(Accept):** 队列中所有单词都已被移到堆栈中, 并且堆栈中只剩下一个由非终结符  $S$  为根的树, 表示分析成功。

**拒绝(Reject):** 队列中所有单词都已被移到堆栈中, 但是堆栈中并非只有一个以非终结符  $S$  为根的树, 并且无法继续规约, 表示分析失败。

如何根据当前堆栈  $S$  和队列  $Q$  中的状态, 选择下一步的操作是移进-规约分析算法中最重要的部分。由于移进和规约操作并不是完全互斥的, 在很多状态下两种操作都可以选择, 这就造成了移进规约冲突(Shift Reduce Conflict)。在自然语言这种非确定性语法下, 这种冲突不可避免。在基于规则和策略的移进-规约分析方法中, 当有多种规约可能时分析算法需要选择其中某个, 当移进和规约都可以时也需要选择执行哪个动作。分析算法可以通过设置执行策略来解决这些冲突, 例如采用深度优先搜索策略, 只有在不能规约时才移进, 有多种规约时选择最长的文法规则等策略。由于移进和规约操作不可撤销, 因此即便输入的句子是符合语法的情况下, 由于策略选择的问题, 也可能分析失败, 堆栈中不能规约到只有  $S$  为根的树的情况。为了解决这个问题也可以采用回溯策略, 当分析失败时回溯顺次尝试不同选择。

为了方便理解移进-规约分析算法, 以下使用在与介绍CYK算法时相同的文法  $G$ , 也以“她喜欢跳芭蕾”句子为例, 介绍移进-规约算法的分析具体过程, 如图3.10和3.11所示。在初始化时, 队列  $Q$  中保存句子中的所有单词, 堆栈  $S$  为空。第1步进行移进操作, 将句子的第一个单词“她”压入堆栈中。第2步利用文法  $G$  中的“ $P \rightarrow \text{她}$ ”规则生成子树, 并将以  $P$  为根节点的子树压入堆栈中。第3-6步依次移进“喜欢”和“跳”并利用文法规则生成子树。接下来可以进行移进操作将句子中“芭

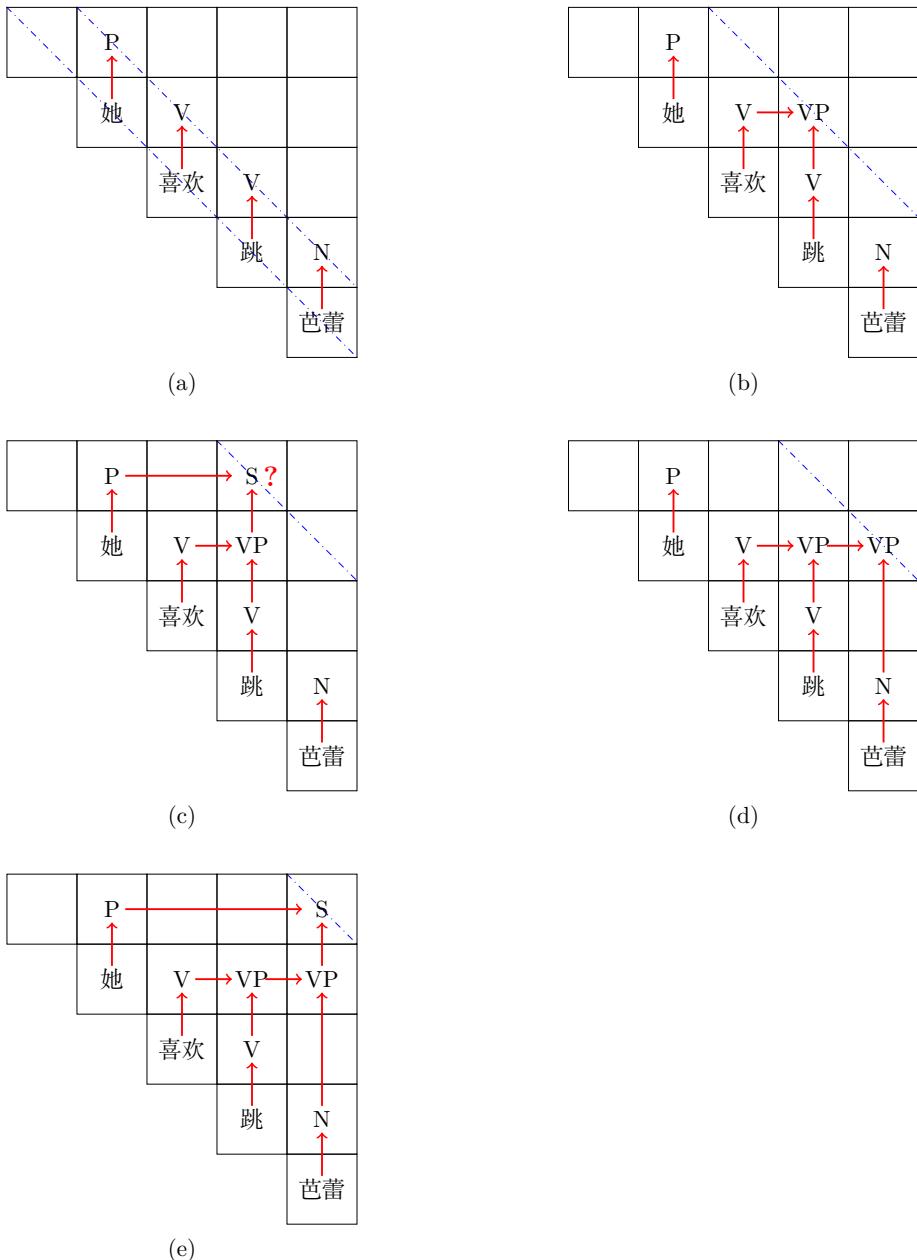


图 3.9 CYK 算法分析实例

蕾”压入栈顶，也可以利用文法“ $VP \rightarrow V V$ ”进行规约操作，面临了移进规约冲突问题。在这里我们采用深度有限搜索策略，选择规约操作，并将生成的以  $VP$  为顶点的子树压入栈顶。第 8 步采用移进操作，将最后一个单词压入栈顶。此后通过第 9-11 步一系列的规约操作，最终生成的以  $S$  为根节点的句法树，并保存在栈中。此时队列为空，最后执行接受操作，分析成功。

堆栈		动作	队列
0.	Ø	Ø	她 喜欢 跳 芭蕾
1.	她	shift	喜欢 跳 芭蕾
2.	P   她	reduce	喜欢 跳 芭蕾
3.	P   她 喜欢	shift	跳 芭蕾
4.	P   她 喜欢 V	reduce	跳 芭蕾
5.	P   她 喜欢 V 跳	shift	芭蕾
6.	P   她 喜欢 V V 跳	reduce	芭蕾
7.	P   她 VP   V V   喜欢 跳	reduce	芭蕾
8.	P   她 VP   V V   喜欢 跳 芭蕾	shift	Ø

图 3.10 移进规约成分句法分析算法分析过程实例

	堆栈	动作	队列
9.	<pre> graph TD     P1[P] --- VP1[VP]     VP1 --- V1[V]     VP1 --- V2[V]     V1 --- J1[喜欢]     V2 --- J2[跳]     N1[N] --- B1[芭蕾]   </pre>	reduce	$\emptyset$
10.	<pre> graph TD     P2[P] --- VP2[VP]     VP2 --- V3[V]     VP2 --- V4[V]     V3 --- J3[喜欢]     V4 --- J4[跳]     N2[N] --- B2[芭蕾]   </pre>	reduce	$\emptyset$
11.	<pre> graph TD     S1[S] --- P1[P]     P1 --- VP1[VP]     VP1 --- V1[V]     VP1 --- V2[V]     V1 --- J1[喜欢]     V2 --- J2[跳]     N1[N] --- B1[芭蕾]   </pre>	reduce	$\emptyset$
12.	<pre> graph TD     S2[S] --- P2[P]     P2 --- VP2[VP]     VP2 --- V3[V]     VP2 --- V4[V]     V3 --- J3[喜欢]     V4 --- J4[跳]     N2[N] --- B2[芭蕾]   </pre>	accept	$\emptyset$

图 3.11 移进规约成分句法分析算法分析过程实例 (续)

### 3.2.2 基于概率上下文无关文法的成分句法分析

上下文无关文法虽然比较简单且容易理解，但是对于句子结构歧义无法很好的处理。比如句子“*He eat soup with spoon*”具有两种可能的句法结构树，并且两种树结构都符合语法，如图3.12所示。使用上下文无关文法很难从这两种句法树中进行选择。基于概率上下文无关文法（Probabilistic Context-Free Grammar, PCFG）的句法分析则可以结合规则方法和统计方法，在一定程度上缓解了上述歧义问题。

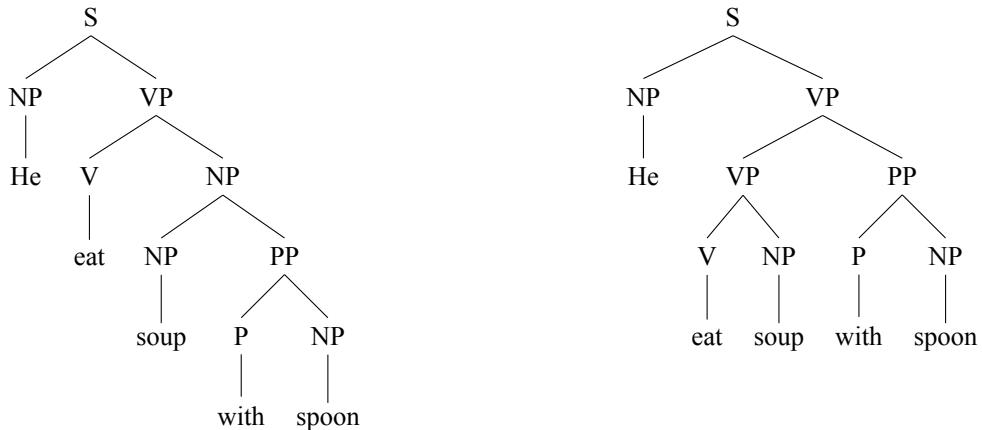


图 3.12 句子 “He eat soup with spoon” 的成分语法树

PCFG 是 CFG 的扩展，因此 PCFG 的文法也是由终结符集合  $\Sigma$ 、非终结符集合  $N$ 、初始符  $S$  以及规则集合  $R$  组成。只是在 CFG 的基础上对每条规则增加了概率，其规则用如下形式表示：

$$A \rightarrow \alpha, p$$

其中  $A$  为非终结符， $\alpha \in (\Sigma \cup N)^*$  为终结符和非终结符组成的有序序列集合， $p$  为  $A$  推导出  $\alpha$  的概率，即  $p = P(A \rightarrow \alpha)$ ，该概率分布要满足如下条件：

$$\sum_{\alpha} P(A \rightarrow \alpha) = 1 \quad (3.1)$$

也就是说，每一个非终结符展开得到的概率之和为 1。

由于 PCFG 中每个规则中包含了概率信息  $P(A \rightarrow \alpha)$ ，因此可以根据一个句子及其句法分析树计算特定句法分析树的概率、句子的概率以及句子片段的概率。利用特定分析树的概率可以用于消除分析树的歧义。接下来面临的问题就是如何利用 PCFG 构建句子的最佳树结构以及如何得到规则的概率参数。本节将针对上述问题进行介绍。

### 1. PCFG 句法分析树概率计算

句法分析树概率计算是指在给定 PCFG 文法  $G$ 、句子  $W = w_1 w_2 \dots w_n$  以及句法树  $T$  的情况下，计算句法分析树概率  $P(T)$ 。为了简化句法树概率计算，句法树概率计算还要应用以下三个独立假设：

- (1) 位置不变性 (Place in-variance)：子树的概率不依赖于该子树所在的位置；
- (2) 上下文无关性 (Context-free)：子树的概率不依赖于子树以外的单词；
- (3) 祖先无关性 (Ancestor-free)：子树的概率不依赖于子树的祖先节点。

基于上述独立性假设，一个特定句法树  $T$  的概率定义为该句法树  $T$  中用来得到句子  $W$  所使用的  $m$  个规则的概率乘积：

$$P(T) = \prod_{i=1}^m P(A_i \rightarrow \alpha) \quad (3.2)$$

假设给定如下 PCFG 文法：

$G(S):$	$S \rightarrow NP\ VP$	1.0	$NP \rightarrow He$	0.3
	$VP \rightarrow VP\ PP$	0.8	$NP \rightarrow soup$	0.3
	$VP \rightarrow V\ NP$	0.2	$NP \rightarrow spoon$	0.2
	$NP \rightarrow NP\ PP$	0.2	$V \rightarrow eat$	1.0
	$PP \rightarrow P\ NP$	1.0	$P \rightarrow with$	1.0

针对对于句子“ $He\ eat\ soup\ with\ spoon$ ”的两种可能的句法结构包含概率的形式如图3.13所示。

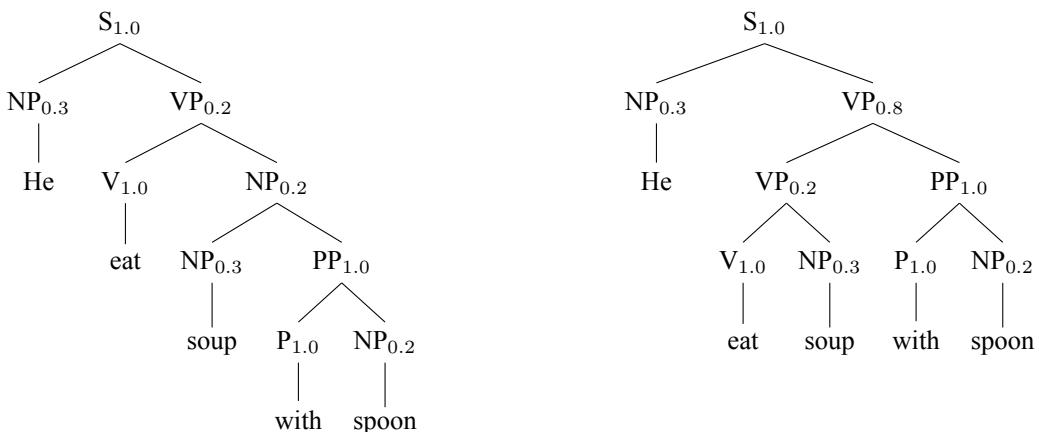


图 3.13 句子 “ $He\ eat\ soup\ with\ spoon$ ” 标注推导规则概率的成分语法树

根据句法树概率计算公式3.2, 图3.13所示的两个句法树的概率分别为:

$$\begin{aligned} P(T_{Left}) &= P(S \rightarrow NP VP) \times P(NP \rightarrow He) \times P(VP \rightarrow V NP) \times \\ &\quad P(V \rightarrow eat) \times P(NP \rightarrow NP PP) \times P(NP \rightarrow soup) \times \\ &\quad P(PP \rightarrow P NP) \times P(P \rightarrow with) \times P(NP \rightarrow spoon) \\ &= 1.0 \times 0.3 \times 0.2 \times 1.0 \times 0.2 \times 0.3 \times 1.0 \times 1.0 \times 0.2 = 0.00072 \end{aligned}$$

$$\begin{aligned} P(T_{Right}) &= P(S \rightarrow NP VP) \times P(NP \rightarrow He) \times P(VP \rightarrow VP PP) \times \\ &\quad P(VP \rightarrow V NP) \times P(V \rightarrow eat) \times P(NP \rightarrow soup) \times \\ &\quad P(PP \rightarrow P NP) \times P(P \rightarrow with) \times P(NP \rightarrow spoon) \\ &= 1.0 \times 0.3 \times 0.8 \times 0.2 \times 1.0 \times 1.0 \times 0.3 \times 1.0 \times 0.2 = 0.00288 \end{aligned}$$

由此可以选择图3.13右侧树结构, 从而解决针对句子“He eat soup with spoon”的句法树歧义问题。

## 2. PCFG 的句子概率计算

句子概率计算是指在给定 PCFG 文法  $G$  的情况下, 计算给定句子  $W$  的概率  $P(W|G)$ 。在第二章中我们介绍了 N 元语言模型, 由于 N 元语言模型仅考虑上下文中相邻的单词, 缺乏对远距离单词以及句法信息的考虑。基于 PCFG 文法的句子概率计算则可以将这些信息引入概率计算。 $P(W|G)$  表示句子  $W$  所有的句法分析树的概率之和。可以采用内向算法 (Inside Algorithm) 或外向算法 (Outside Algorithm) 通过动态规划算法计算得到句子的概率。

采用内向算法, 首先定义内变量  $\alpha_{ij}(A)$  为非终结符  $A$  推导出  $W$  中子串  $w_i w_{i+1} \cdots w_j$  的概率, 即:

$$\alpha_{ij}(A) = P(A \rightarrow w_i w_{i+1} \cdots w_j) \quad (3.3)$$

句子  $W$  的概率则相应的记为  $\alpha_{1n}(S)$ , 即:

$$P(W|G) = P(S \rightarrow W|G) = \alpha_{1n}(S) \quad (3.4)$$

$\alpha_{ij}(A)$  可通过如下递归公式计算得到:

$$\alpha_{ii}(A) = P(A \rightarrow w_i) \quad (3.5)$$

$$\alpha_{ij}(A) = \sum_{B,C} \sum_{i \leq k \leq j} P(A \rightarrow BC) \alpha_{ik}(B) \alpha_{(k+1)j}(C) \quad (3.6)$$

上述公式表示当  $i \neq j$  时, 子串  $w_i w_{i+1} \cdots w_j$  可以被切分成两端:  $w_i \cdots w_k$  和  $w_{k+1} \cdots w_j$ , 前半部分  $w_i \cdots w_k$  是由非终结符  $B$  推导出, 后半部分  $w_{k+1} \cdots w_j$  是由非终结符  $C$  推导出, 即  $A \rightarrow BC \rightarrow w_i \cdots w_k w_{k+1} \cdots w_j$ , 这一推导过程的概率为  $P(A \rightarrow BC) \alpha_{ik}(B) \alpha_{(k+1)j}(C)$ 。具体过程如

算法3.2所示。

---

**代码 3.2:** 面向 PCFG 句子概率求解的内向算法

---

```

输入: PCFG G(S) 和句子  $w_1 w_2 \cdots w_n$ 
输出:  $\alpha_{ij}(A), i \leq i \leq j \leq n$ 
for  $i = 1$  to  $n$  do
     $\alpha_{ii}(A) = P(A \rightarrow w_i), 1 \leq i \leq n;$ 
end
for  $j = 1$  to  $n$  do
    for  $i = 1$  to  $n - j$  do
         $\alpha_{i(i+j)}(A) = \sum_{B,C} \sum_{i \leq k \leq i+j-1} P(A \rightarrow BC) \alpha_{ik}(B) \alpha_{(k+1)(i+j)}(C);$ 
    end
end
return  $\alpha$ 

```

---

计算给定句子  $W$  的概率  $P(W|G)$  也可以采用外向算法利用动态规划方法得到。外向变量  $\beta_{ij}(A)$  定义为初始非终结符  $S$  在推导出句子  $W$  的过程中产生符号串  $w_1 w_2 \cdots w_{i-1} A w_{j+1} \cdots w_n$  的概率 ( $A \rightarrow w_i \cdots w_j$ )：

$$\beta_{ij}(A) = P(S \rightarrow w_1 w_2 \cdots w_{i-1} A w_{j+1} \cdots w_n) \quad (3.7)$$

$\beta_{ij}(A)$  可通过如下递归公式计算得到：

$$\beta_{1n}(A) = \begin{cases} 1, & A = S \\ 0, & A \neq S \end{cases} \quad (3.8)$$

$$\begin{aligned} \beta_{ij}(A) &= \sum_{B,C} \sum_{k \geq j} P(B \rightarrow AC) \alpha_{(j+1)k}(C) \beta_{ik}(B) \\ &\quad + \sum_{B,D} \sum_{k \leq i} P(B \rightarrow DA) \alpha_{k(i-1)}(D) \beta_{kj}(B) \end{aligned} \quad (3.9)$$

上述公式表示当  $i = 1, j = n$  时, 如果  $A = S$ , 那么  $\beta_{1n}(A) = P(S \rightarrow W)$  按照 PCFG 的定义  $P(S \rightarrow W) = 1$ , 即  $\beta_{1n}(A) = 1$ 。如果  $A \neq S$  同样按照 PCFG 的定义以及规范语法中不存在  $S \rightarrow A$  的推导规则, 则  $\beta_{1n}(A) = 0$ 。当  $i \neq 1$  或  $j \neq n$  时, 如果在  $S$  推导出  $W$  的过程中出现了符号串  $w_1 w_2 \cdots w_{i-1} A w_{j+1} \cdots w_n$ , 那么根据上下文无关语法的性质一定存在  $B \rightarrow AC$  或者  $B \rightarrow DA$  的规则。因此构造上述递推公式, 具体过程如算法3.3所示。

---

**代码 3.3: 面向 PCFG 句子概率求解的外向算法**


---

输入: PCFG G(S) 和句子  $w_1 w_2 \dots w_n$   
 输出:  $\beta_{ij}(A), i \leq i \leq j \leq n$

```

// 初始化
 $\beta_{1n}(A) = \begin{cases} 1, & A = S \\ 0, & A \neq S \end{cases};$ 
for  $j = n - 1$  to 0 do
  for  $i = 1$  to  $n - j$  do
     $\beta_{ij}(A) = \sum_{B,C} \sum_{k \geq j} P(B \rightarrow AC) \alpha_{(j+1)k}(C) \beta_{ik}(B)$ 
    +  $\sum_{B,D} \sum_{k \leq i} P(B \rightarrow DA) \alpha_{k(i-1)}(D) \beta_{kj}(B);$ 
  end
end
return  $\beta$ 

```

---

### 3. PCFG 的最佳树结构求解

最佳树结构求解是指对于给定句子  $W = \{w_1, w_2, \dots, w_n\}$  和 PCFG 文法  $G$ , 求解该句子的最佳树结构, 即如何选择句法结构树使得其概率最大:

$$\hat{t} = \arg \max_{t \in \mathcal{T}_G(W)} P(t|W, G) \quad (3.10)$$

$\mathcal{T}_G(W)$  表示句子  $W$  所有符合文法  $G$  的句法树。该问题可以通过利用基于概率的 CYK 算法进行。与 CYK 算法一样, 概率 CYK 算法也要求所使用的语法必须符合乔姆斯基范式, 其规则被限制为  $A \rightarrow BC$  或  $A \rightarrow w$  两种形式。与 CYK 算法一样, 概率 CYK 算法也将输入的句子表示为单词之间带有索引号的句子形式。例如:  $_0 \text{He}_1 \text{eat}_2 \text{soup}_3 \text{with}_4 \text{spoon}_5$ 。

针对句子长度为  $n$  的句子, 概率 CYK 算法同样也使用  $(n+1) \times (n+1)$  的矩阵  $T$  的上三角形部分进行存储。 $T_{ij}$  表示输入句子中横跨在位置  $i$  到  $j$  之间的单词的组成成分, 包含由文法  $G$  的非终结符构成的集合以及以该非终结符为根的句法树的概率。为了方便期间这里我们使用  $T_{ij,A}$  表示矩阵  $T_{ij}$  保存的非终结符集合中  $A$  的概率。算法3.4给出了概率 CYK 算法的基本流程。

以下通过一个实例来说明使用 CYK 算法求解 PCFG 的最佳树结构具体流程。给定如下符合乔姆斯基范式的文法  $G$  如下:

非终结符号集合:  $N = \{S, P, V, NP, PP, VP\}$

终结符号集合:  $\Sigma = \{ \text{eat}, \text{he}, \text{soup}, \text{spoon}, \text{with} \}$

规则集合:  $R = \{ (1) S \rightarrow NP VP \ 1.0; \ (2) VP \rightarrow VP PP \ 0.8; \ (3) VP \rightarrow V NP \ 0.2; \ (4) NP \rightarrow NP PP \ 0.2; \ (5) PP \rightarrow P NP \ 1.0 \ (6) NP \rightarrow He \ 0.3; \}$

**代码 3.4: 概率 CYK 句法分析算法**


---

**输入:** 语法信息  $G$  和句子  $w_1 w_2 \cdots w_n$   
**输出:** 句法树矩阵  $T$

```

// 初始化
for i = 1 to n do
     $T_{ii} = w_i;$  // 主对角线上依次放入单词  $w_i$ ;
    foreach  $A | A \rightarrow w_i$  do
         $T_{(i-1)i} = T_{(i-1)i} \cup (A, p), (A \rightarrow w_i, p \in G);$  // 依次放入单词  $w_i$  的所有词性及其
        // 概率;
    end
end

// 平行于主对角线逐层计算
for  $d = 2 to n$  do
    for  $i = 0 to n - d$  do
        j = i+d;
        for  $k = i + 1 to j - 1$  do
            if  $(A \rightarrow BC, p) \in G$  and  $T_{ik,B} > 0$  and  $T_{kj,C} > 0$  then
                if  $T_{ij,A} < p \times T_{ik,B} \times T_{kj,C}$  then
                     $T_{ij} = T_{ij} \cup \{A, p \times T_{ik,B} \times T_{kj,C}, (k, B, C)\};$  // (k, B, C) 用于保存推
                    // 导路径;
                end
            end
        end
    end
end

return  $T$ 

```

---

- (7)  $NP \rightarrow soup \ 0.3$ ; (8)  $NP \rightarrow spoon \ 0.2$ ; (9)  $V \rightarrow eat \ 1.0$ ;  
 (10)  $P \rightarrow with \ 1.0 \}$

图3.14和图3.15给出了概率 CYK 算法分析句子“*He eat soup with spoon*”的过程。首先初始化矩阵主对角线上的单词以及单词所对应的词性和概率。之后进行  $T_{02}, T_{13}, T_{24}, T_{35}$  元素的计算，通过规则集合中  $VP \rightarrow V NP \ 0.2$  的推导规则，可以得到在  $T_{13}$  添加  $VP$  以及相应的路径信息，其概率为  $0.2 \times 1.0 \times 0.3 = 0.06$ 。根据  $PP \rightarrow P NP \ 1.0$  的推导规则，在  $T_{35}$  添加  $PP$  以及相应的路径信息，其概率为  $1.0 \times 1.0 \times 0.2 = 0.2$ 。依次逐层进行分析，在对  $T_{1,6}$  进行分析时，分别根据  $VP \rightarrow VP PP \ 0.8$  和  $VP \rightarrow V NP \ 0.2$  两个不同的推导规则得到  $VP_1$ ，其概率为  $0.8 \times 0.06 \times 0.2 = 0.0096$ ， $VP_2$  的概率为  $0.2 \times 1.0 \times 0.012 = 0.0024$ 。据此在  $T_{05}$  节点可以根据  $S \rightarrow NP VP \ 1.0$ ，得到两个不同分析结

果:  $S_1$  的概率为  $1.0 \times 0.3 \times 0.0096 = 0.00288$  和  $S_2$  的概率为  $1.0 \times 0.3 \times 0.0024 = 0.00072$ 。根据不同所得到的不同分析树的概率大小, 选择最终结果。

通过上述例子我们可以看到, 根据构建好的 PCFG 文法, 利用概率 CYK 分析等最佳树结构求解算法, 可以在一定程度上对句法分析中的歧义进行很好的处理, 根据概率选择可能性较大的树结构。

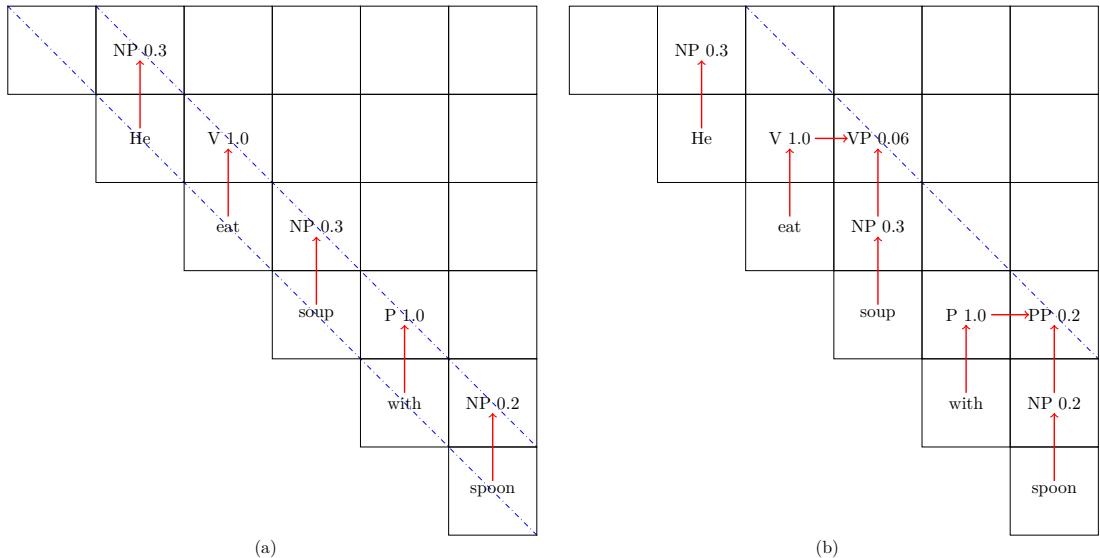


图 3.14 使用 CYK 算法求解 PCFG 的最佳树结构分析实例

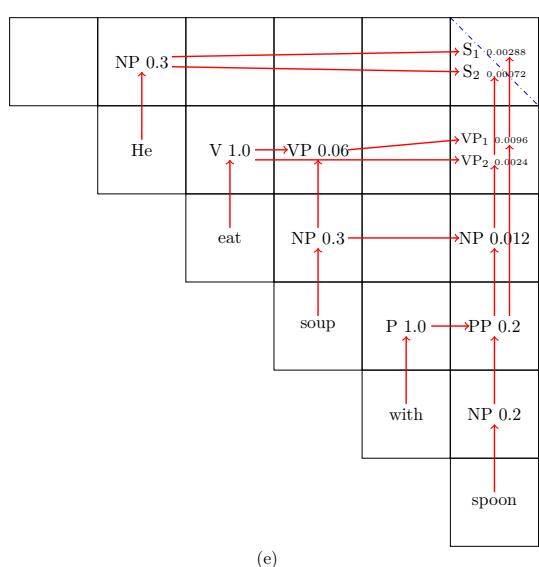
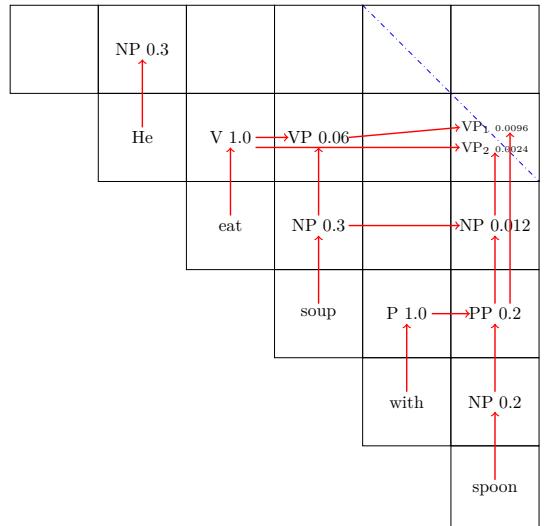
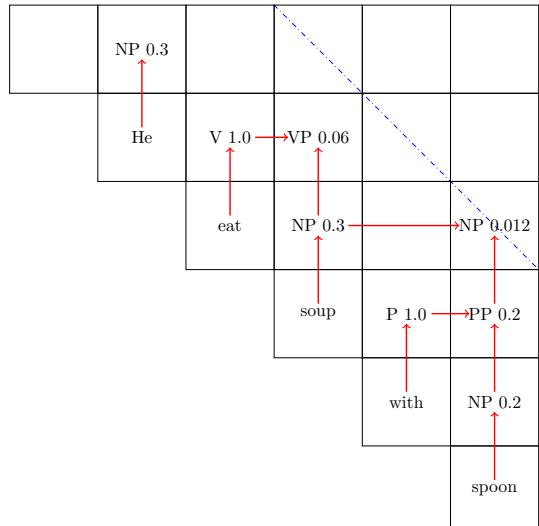


图 3.15 使用 CYK 算法求解 PCFG 的最佳树结构分析实例 (续)

#### 4. PCFG 的模型参数学习

模型参数学习是指给定 CFG 文法  $G$  下, 如何学习  $G$  中规则的概率参数构建 PCFG 文法。有两种常见的方法用于学习语法规则概率: (1) 有大规模句法树标注数据 (树库); (2) 有大规模的没有标注句法树的句子。

在有大规模标注句法树标注的情况下, 可以基于最大似然估计, 统计非终结符的出现次数进行概率参数估计:

$$P(A \rightarrow \alpha) = \frac{Count(A \rightarrow \alpha)}{\sum_{\lambda} Count(A \rightarrow \lambda)} = \frac{Count(A \rightarrow \alpha)}{Count(A)} \quad (3.11)$$

$Count(A \rightarrow \alpha)$  是指规则  $A \rightarrow \alpha$  在整个树库中出现的次数,  $Count(A)$  是指在树库中非终结符  $A$  出现的次数。

在仅有大规模无标记句子的情况下, 也可以通过期望最大化算法 (Expectation Maximization, EM) 估计规则的概率参数。基本思想是首先对给定的 CFG 文法  $G$  中的每个规则, 在满足归一化条件的情况下随机赋值概率值, 构造文法  $G_0$ 。在此基础上利用  $G_0$  对所有句子进行分析, 计算出每条规则使用次数的期望值。之后利用期望次数根据最大似然估计原理, 得到文法  $G$  的概率参数更新, 记为  $G_1$ 。循环执行该过程直到  $G$  的概率参数收敛于最大似然估计值。利用当前的文法  $G_i$  估算每条规则出现的期望值是期望步 (Expectation Step, E-步骤), 重新估算概率得到  $G_{i+1}$  的步骤是最大化步 (Maximization Step, M-步骤)。具体的计算公式可以参考文献 [98, 99]。

#### 3.2.3 成分句法分析评价方法

成分句法分析算法的性能评价, 目前通常采用的是方法是 PARSEVAL 方法<sup>[100]</sup>, 主要包含以下三个指标:

- (1) 标记精确率 (Labeled Precision, LP): 句法器分析结果中正确的短语结构所占比例, 即分析结果与标准句法树中短语匹配的个数占分析器所有输出结果中短语个数的比例, 具体计算公式如下:

$$LP = \frac{\text{分析结果中正确的短语个数}}{\text{分析结果中短语总数}} \times 100\% \quad (3.12)$$

- (2) 标记召回率 (Labeled Recall, LR): 分析结果中正确的短语结构占标准句法树中短语总数的比例, 具体计算公式:

$$LR = \frac{\text{分析结果中正确的短语个数}}{\text{标准结果中短语总数}} \times 100\% \quad (3.13)$$

- (3) 交叉括号数 (Crossing brackets, CBs): 一颗成分句法树中所包含的与标准句法树中边界相交叉的短语个数。

此外, 通常还会根据标记精确率 (LP) 和标记召回率 (LR), 利用与标准 F1 值计算公式一致的方法得到标记 F1 值, 综合 LP 和 LR 得分。

以句子“`He eat soup with spoon`”的正确答案和某成分句法分析器结果为例，如图3.16所示，对上述评测指标进行解释。

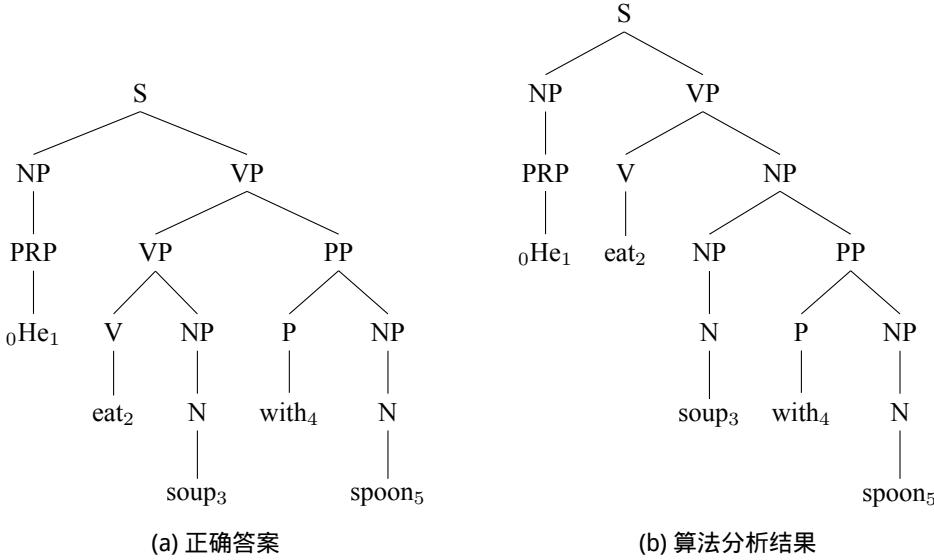


图 3.16 成分句法分析结果评测样例

为了方便评测，分析树中除词性之外的非终结符节点通常会增加起始位置和结束位置的方式进行表示： $XP$ (起始位置，结束位置)。其中  $XP$  表示短语名称，例如：动词短语  $VP$ ，名词短语  $NP$  等。根据这种表示形式，图3.16所示的两棵树的结果表示为：

- (a) 正确结果： $S(0,5)$ ,  $NP(0,1)$ ,  $VP(1,5)$ ,  $VP(1,3)$ ,  $NP(2,3)$ ,  $PP(3,5)$ ,  $NP(4,5)$
- (b) 算法分析结果： $S(0,5)$ ,  $NP(0,1)$ ,  $VP(1,5)$ ,  $NP(2,5)$ ,  $NP(2,3)$ ,  $PP(3,5)$ ,  $NP(4,5)$

根据公式3.12、公式和公式可以计算得到该句法分析算法的如下评估指标：

$$LP = \frac{6}{7} \times 100\% = 85.7\%$$

$$LR = \frac{6}{7} \times 100\% = 85.7\%$$

算法分析结果中的  $NP(2,5)$  与正确结果的  $PP(3,5)$  和  $VP(1,3)$  都发生了边界交叉，因此交叉括号  $CBs$  值为 2。

### 3.3 依存句法分析

依存句法分析（Dependency Parsing）任务目标是依据依存语法理论分析输入句子得到其依存句

法结构树。依存句法理论的基本假设是句法结构由单词和单词之间的依存关系组成。依存关系具有方向性, 从中心语成分指向依存成分。依存关系根据中心成分和依存成分之间的关系又可以被定义为不同的依存关系类型。依存句法结构使用依存图 (Dependency Graph) 进行表示。如图3.17所示。

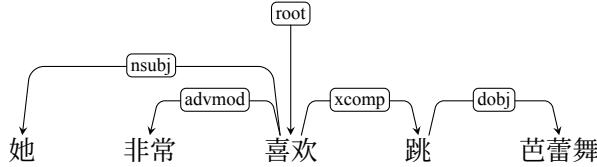


图 3.17 句子“她非常喜欢跳芭蕾舞”的依存句法树

在本节中, 使用  $S = w_0 w_1 \dots w_n$  表示输入的句子, 其中  $w_0 = \text{root}$  表示虚拟根节点,  $w_1 \dots w_n$  为输入句子中的  $n$  个单词。 $R = \{r_0, r_1, \dots, r_m\}$  表示依存关系类型集合,  $r \in R$  表示在句子中单词之间的依存关系, 也叫做边标签 (Arc Label)。例如在图3.17中, “喜欢”和“她”之间的依存关系类型  $r = \text{nsubj}$ 。依存图  $G = (V, A)$  是一个有标记的有向图, 顶点  $V$  和边  $A$  针对句子  $S$  和依存关系集合符合以下要求:

- (1)  $V \subseteq \{w_0, w_1, \dots, w_n\}$
- (2)  $R \subseteq V \times R \times V$
- (3) 如果  $(w_i, r, w_j) \in A$ , 那么  $(w_i, r', w_j) \notin A, \forall r' \neq r$

边  $(w_i, r, w_j)$  表示以  $w_i$  为头  $w_j$  为尾边标记为  $r$  的边, 其表示以  $w_i$  为中心词,  $w_j$  为修饰词, 类型为  $r$  的依存关系。依存图  $G$  表示了一组句子  $W$  中的单词之间的有标记的依存关系。依存图中的节点集合  $V$  通常包含句子中的所有单词, 针对句子  $S$  通常用  $V_S = \{w_0, w_1, \dots, w_n\}$  表示。

根据上述定义, 图3.17可以表示为:

- (1)  $G = (V, A)$
- (2)  $V = V_W = \{\text{ROOT}, \text{她}, \text{非常}, \text{喜欢}, \text{跳}, \text{芭蕾舞}\}$
- (3)  $A = \{(\text{ROOT}, \text{root}, \text{喜欢}), (\text{喜欢}, \text{nsubj}, \text{她}), (\text{喜欢}, \text{advmod}, \text{非常}), (\text{喜欢}, \text{xcomp}, \text{跳}), (\text{跳}, \text{dobj}, \text{芭蕾舞})\}$

如果依存图  $G = (V, A)$  对于输入句子  $S$  和关系集合  $R$ , 是一个从  $w_0$  出发的有向树, 并且包含句子中的所有单词, 那么这个依存图  $G$  就成为形式良好的依存图 (Well-formed Dependency Graph), 也称为依存树 (Dependency Tree)。对于输入句子  $W$  和关系集合  $R$ , 所有形式良好的依存图集合记为  $\mathcal{G}_S$ 。依存句法分析就是对输入句子  $W$  根据一定的原则从  $\mathcal{G}_S$  中选择得分最高的依存句法树。

### 3.3.1 基于图的依存句法分析

基于图的依存句法分析核心是构造评分函数，对句子  $S$  所有依存句法树  $G = (V, A) \in \mathcal{G}_S$  进行评分。这个评分代表了一个句法树作为句子分析正确结果的可能性。不同的基于图的分析方法采用不同的假设来计算得分。基于图的依存句法分析算法通常将对依存句法树的  $G = (V, A)$  的评分转化为对其树上的边的评分：

$$\text{score}(G_S) = \sum_{(w_i, r, w_j) \in A} \lambda_{(w_i, r, w_j)} \quad (3.14)$$

$\lambda$  表示所有边的评分， $\lambda_{(w_i, r, w_j)}$  表示边  $(w_i, r, w_j)$  的得分，其具体计算方法将在本节后续部分进行详细介绍。在图得分计算基础上，可以将基于图的依存句法分析形式化表示为：

$$\hat{G} = \arg \max_{G=(V,A) \in \mathcal{G}_S} \sum_{(w_i, r, w_j) \in A} \lambda_{(w_i, r, w_j)} \quad (3.15)$$

可以证明在依存句法树不考虑投射性（Projectivity）的情况下，对于输入句子  $S$  的依存句法分析问题等价于基于边评分  $\lambda_{(w_i, r, w_j)}$  的图  $G_S$  的最大生成树（Maximum Spanning Tree）寻找问题<sup>[101]</sup>。此外，可以定义  $\lambda_{(w_i, w_j)} = \max_r \lambda_{(w_i, r, w_j)}$ ，并利用  $\lambda_{(w_i, w_j)}$  作为边的权重构建图  $G'$ 。通过图  $G'$  得到的最大生成树  $\hat{G}'$ ，可以证明与通过  $\lambda_{(w_i, r, w_j)}$  构建得到的最大生成树  $\hat{G}$  相同<sup>[101]</sup>。由此，可以将包含依存关系类别的分析转换为无依存关系类别的分析，进一步降低分析算法的复杂性。

需要注意的是，利用最大生成树算法得到的依存句法树不具备投射性。针对具有投射性要求的依存句法树，可以利用其与上下文无关语法之间的强相关性，利用基于 CYK 算法等上下文无关语法分析算法进行依存句法树分析。本节将针对上述两种情况分别进行介绍。

#### 1. 非投射性依存句法分析方法

由上节的介绍我们可以知道，非投射性的依存句法分析等价于图的最大生成树寻找问题。朱-刘/埃德蒙兹算法（Chu-Liu/Edmonds）方法<sup>[102, 103]</sup> 是一种常见的带权有向图最小/大生成树寻找算法，因此也常被应用于非投射性依存句法分析。该算法的核心是采用贪心的思想对边进行选择，利用递归方法来实现整个过程。

该算法用于非投射性依存句法分析的输入是待分析句子  $S = w_0 w_1 \cdots w_n$  和边之间的权重  $\lambda_{(w_i, w_j)} \in \lambda$ 。根据定义， $w_0$  是句子的虚拟根节点，依存句法树中不存在指向  $w_0$  的边，因此设置所有  $\lambda_{(w_i, w_0)} = -\infty$ 。首先根据句子中的单词和权重组成有向图  $G_S = (V_S, A_S)$ ， $V_S = \{w_0, w_1, \dots, w_n\}$ ， $A_S = \{(w_i, w_j) | \forall w_i, w_j \in V_S\}$ 。接下来针对图  $G$  中每个顶点选择入边权重最大的边构建子图  $G' = (V_S, A')$ 。如果该子图中没有环，那么该子图就是图  $G$  的最大生成树。否则，说明图  $G'$  中至少包含一个环。那么选择其中任意一个环  $C$ ，其边集合为  $A_C$ ，将环  $C$  用一个节点  $w_C$  来代表，图  $G_C$  中包含所有在图  $G$  中但是不在环  $C$  中的节点，以及  $w_C$ ， $G_C$  的边通过以下规则构建：

- (1) 对于所有不在环  $C$  的顶点  $w_j$ , 如果存在一个边  $(w_i, w_j)$ ,  $w_i$  在环  $C$  中, 那么就添加边  $(w_c, w_j)$  到  $G_C$  中, 定义  $ep(w_c, w_j) = \arg \max_{w_i \in C} \lambda_{(w_i, w_j)}$ , 用于记录环  $C$  中相对应的顶点, 相应的修改  $(w_c, w_j)$  的边权重  $\lambda_{(w_c, w_j)} = \lambda_{(ep(w_c, w_j), w_j)}$ ;
- (2) 对于所有不在环  $C$  的顶点  $w_i$ , 如果存在一个边  $(w_i, w_j)$ ,  $w_j$  在环  $C$  中, 那么添加边  $(w_i, w_c)$  到  $G_C$  中, 定义  $ep(w_i, w_c) = \arg \max_{w_j \in C} [\lambda_{(w_i, w_j)} - \lambda_{(a(w_j), w_j)}]$ , 相应的修改  $(w_i, w_c)$  的边权重  $\lambda_{(w_i, w_c)} = \lambda_{(w_i, ep(w_i, w_c))} - \lambda_{(a(ep(w_i, w_c)), (ep(w_i, w_c)))} + \sum_{w \in C} \lambda_{(a(w), w)}$ ;
- (3) 对于所有边  $(w_i, w_j)$ , 其顶点  $w_i$  和  $w_j$  都不在环  $C$  中, 直接添加该边到图  $G'$  中, 保持原有权重不变。

将图  $G_C$  作为输入递归调用上述算法得到其最大生成树  $G = (V, A)$ 。之后根据所返回的最大生成树信息对原始图信息进行修正, 移除环  $C$ , 并且将  $w_c$  所指向的节点的实际边还原, 返回输入图所对应的最大生成树。朱-刘/埃德蒙兹算法的伪代码如算法3.5所示。

## 2. 投射性依存句法分析方法

设置虚拟根节点在句子首位的情况下, 投射性依存句法分析树等价于嵌套依存句法分析树 (Nested Dependency Trees)。因此投射性依存句法分析与上下文无关语法具有非常强的关系, 很多用于上下文无关语法分析的算法也可以应用于投射性依存句法分析。在 4.2 节中介绍了 CYK 算法用于 CFG 和 PCFG 的句法分析, 本节将介绍基于 CYK 算法改进的依存句法分析算法。

首先定义动态规划表  $C[s][t][i]$  ( $s \leq i \leq t$ ), 表示投射性句法树以单词  $w_i$  为根节点覆盖从单词  $w_s$  到单词  $w_t$  的句子片段的最高得分。由此可以得到  $C[0][n][0]$  表示输入句子  $S = w_0, w_1, \dots, w_n$  的依存句法树的最高得分。任何以  $w_i$  为根节点覆盖  $w_s$  到  $w_t$  的投射性句法树都是由更小的子树构成。同时更大的子树是通过相邻的子树由内向外添加依存关系得到。因此对于  $C[s][t][i]$  可以构造如下递推公式:

$$C[s][t][i] = \max_{s \leq k \leq t, s \leq j \leq t} \begin{cases} C[s][k][i] + C[k+1][t][j] + \lambda(w_i, w_j), & \text{如果 } j > i \\ C[s][k][j] + C[k+1][t][i] + \lambda(w_i, w_j), & \text{如果 } j < i \end{cases} \quad (3.16)$$

图3.18给出了合并过程示意图。

由于  $C[s][t][i]$  中仅保存了子树的最高得分, 但是没有保存子树的结构。因此在实际构建过程中, 还需要增加树结构矩阵  $A[s][t][i]$ , 用于保存依存句法树结构。通过公式3.16计算得到最优的  $k$  和  $j$  之后,  $A[s][t][i]$  按照如下公式记录树结构:

$$A[s][t][i] = \begin{cases} A[s][k][i] \cup A[k+1][t][j] \cup (w_i, w_j), & \text{如果 } j > i \\ A[s][k][j] \cup A[k+1][t][i], & \text{如果 } j < i \end{cases} \quad (3.17)$$

针对句子  $S$  得到的依存句法树  $G = (V, A[0][n][0])$  组成。

**代码 3.5: 朱-刘/埃德蒙兹算法求解依存句法分析树**

**输入:** 图  $G = (V, A)$ ,  $\lambda_{(w_i, w_j)} \in \lambda$

**输出:** 最大生成树图  $G$

**Function Chu-Liu-Edmonds( $G, \lambda$ ):**

$$A' = \{(w_i, w_j) | w_j \in V, w_i = \arg \max_{w_i} \lambda_{(w_i, w_j)}\};$$

$$G' = (V, A');$$

**if**  $G'$  中没有环 **then**

| **return**  $G'$ ;

**end**

$A'_C$ =图  $G'$  中任意一个环  $C$  的边集合;

$V'_C$ =环  $C$  顶点集合;

$$V_C = V \cup \{w_c\} - V'_C;$$

**foreach**  $w_j \in V - V'_C : \exists_{w_i \in V'_C} (w_i, w_j) \in A$  **do**

$$| A_C = A_C \cup \{(w_c, w_j)\};$$

$$| ep(w_c, w_j) = \arg \max_{w_i \in C} \lambda_{(w_i, w_j)};$$

$$| \lambda_{(w_c, w_j)} = \lambda_{(ep(w_c, w_j), w_j)};$$

**end**

**foreach**  $w_i \in V - V'_C : \exists_{w_j \in V'_C} (w_i, w_j) \in A$  **do**

$$| A_C = A_C \cup \{(w_i, w_c)\};$$

$$| ep(w_i, w_c) = \arg \max_{w_j \in C} [\lambda_{(w_i, w_j)} - \lambda_{(a(w_j), w_j)}];$$

$$| \lambda_{(w_i, w_c)} = \lambda_{(w_i, ep(w_i, w_c))} - \lambda_{a(ep(w_i, w_c)), (ep(w_i, w_c))} + \sum_{w \in C} \lambda_{(a(w), w)};$$

**end**

**foreach**  $w_i \in V - V'_C : \exists_{w_j \in V - V'_C} (w_i, w_j) \in A$  **do**

$$| A_C = A_C \cup \{(w_c, w_j)\};$$

**end**

$G_C = (V_C, A_C);$

$G = (A, V) = \text{Chu-Liu-Edmonds}(G_C, \lambda);$

寻找  $A$  中指向  $w_c$  的边  $(w_i, w_c)$ ,  $w_j = ep(w_i, w_c)$ ;

寻找环  $C$  中指向  $w_j$  的边  $(w_k, w_j)$ ;

寻找  $A$  中所有以  $w_c$  为头节点的边  $(w_c, w_l)$ ;

$$A = A \cup (ep(w_c, w_l), wl)_{forall (w_c, w_l) \in A} \cup A_c \cup (w_i, w_j) - (w_k, w_j);$$

**return**  $G$ ;

**return**

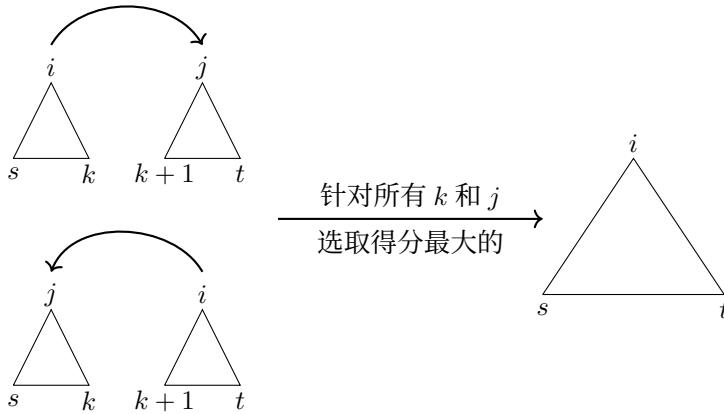


图 3.18 面向依存句法分析的 CYK 算法递推样例

### 3. 边评分模型学习方法

边评分模型可以使用基于高维特征向量的线性函数进行建模。使用  $f(w_i, r, w_j) \in \mathbb{R}^m$  表示特征函数，其输出是高维特征向量， $w$  是  $f(\cdot)$  的对应的权重向量。边评分参数  $\lambda_{(w_i, r, w_j)}$  可以表示为：

$$\lambda_{(w_i, r, w_j)} = w \cdot f(w_i, r, w_j) \quad (3.18)$$

$f(\cdot)$  包含了边和输入句子  $S$  中各类型相关特征，例如：

- $w_i$ = 喜欢
- $w_j$ = 跳
- $w_i$  的词性 =V
- $w_j$  的词性 =V
- $r$  的依存关系类型 =xcomp
- $w_{i-1}$  的词性 =ADV
- $w_{j+1}$  的词性 =N
- $w_i$  和  $w_j$  之间距离 =1

这些特征还可以组合为更复杂的类型例如：

- $w_i$  的词性 =V &  $w_j$  的词性 =V &  $w_i$ = 喜欢
- $w_i$ = 喜欢 &  $w_j$ = 跳 &  $w_i$  和  $w_j$  之间距离 =1

类别特征通常也会通过二值化操作转换为 0/1 特征，具有  $m$  个类别的特征通常会转化为  $m$  个 0/1 特征，表示某个类别出现或不出现。从上述特征的构成我们可以看到，特征向量维度通常会非常高，但是对于一个边来说特征向量非常稀疏，仅有非常小部分的特征不是 0。因此可以利用稀疏性进行表示和计算。

基于特征向量的线性函数建模的假设下，对于输入句子的依存句法分析转换为了如下问题：

$$\hat{G} = \arg \max_{G=(V,A) \in \mathcal{G}_S} \sum_{(w_i, r, w_j) \in A} \lambda_{(w_i, r, w_j)} = \arg \max_{G=(V,A) \in \mathcal{G}_S} \sum_{(w_i, r, w_j) \in A} \mathbf{w} \cdot f(w_i, r, w_j) \quad (3.19)$$

$\mathcal{D} = \{(S_d, G_d)\}_{d=1}^{|\mathcal{D}|}$  表示训练语料集合，其中  $S_d$  表示输入句子，所对应的正确的依存句法树用  $G_d$  表示。针对输入句子可以利用特征函数  $f(\cdot)$  构造句子中所有单词对和根据不同依存关系类型输出特征向量。边评分模型的学习就转换为了权重向量  $\mathbf{w}$  的学习问题。该问题可以采用感知器算法 (Perceptron Algorithm) 完成。感知器算法是一种典型的基于推理 (Inference-based Learning) 的在线学习方法，也称为错误驱动 (Error-driven) 的学习算法，伪代码如算法3.6所示。

---

#### 代码 3.6: 感知器算法学习参数向量 $\mathbf{w}$

---

输入: 训练语料  $\mathcal{D} = \{(S_d, G_d)\}_{d=1}^{|\mathcal{D}|}$

输出: 权重向量  $\mathbf{w}$

Function Perceptron( $\mathcal{D}$ ) :

```

 $w=0;$ 
for  $n : 1 \dots N$  do
    for  $d : 1 \dots |\mathcal{D}|$  do
         $G' = \arg \max_{G=(V,A) \in \mathcal{G}_{S_d}} \sum_{(w_i, r, w_j) \in A} \mathbf{w} \cdot f(w_i, r, w_j);$ 
        if  $G \neq G'$  then
             $| \quad \mathbf{w} = \mathbf{w} + \sum_{(w_i, r, w_j) \in A_d} f(w_i, r, w_j) - \sum_{(w_i, r, w_j) \in A'} f(w_i, r, w_j);$ 
            end
        end
    end
return  $w$ 
return

```

---

感知器算法每次仅考虑一个训练样本，利用当前的权重向量  $\mathbf{w}$  和依存树构建算法针对当前样本  $S_d$  构建依存句法树  $G'$ 。如果  $G'$  与标准标注  $G_d$  不同，则通过增加正确依存树所对应的边的特征向量并减去不正确的句法树分析结果的边，从而更新当前的权重向量  $\mathbf{w}$ 。最大生成树依存句法分析器 (Maximum Spanning Tree Dependency Parser, MSTParser)<sup>[104]</sup> 采用了上述方法，并在上述基础上增加了最大间隔 (Max Margin) 目标函数，并引入了边缘注入松弛算法 (Margin-Infused Relaxed Algorithm, MIRA)，进一步提升了所学习得到的权重向量的泛化能力。

### 3.3.2 基于神经网络的图依存句法分析

基于图的依存句法分析主要包含边评分模型和句法树生成算法两个部分组成。其中边评分模型对于分析效果具有决定性的影响。传统的分析方法依赖人工设计的数百万甚至数千万特征，模型的泛化能力不高，很容易出现过拟合问题。此外，由于特征函数中使用大量的单词关联信息，也很容易使得特征空间巨大。神经网络方法可以在一定程度上缓解上述问题，同时也不需要人工设计特征函数。在本节中，将介绍三种基于神经网络图依存句法分析算法。

#### 1. 基于前馈神经网络的方法

文献 [105] 提出的依存句法分析器采用最大生成树方法构造依存句法树。但是在边评分模型方面，他们提出了仅利用最基本信息作为输入的神经网络方法。公式3.14给出了基于图的依存句法分析的树得分计算方法，将树的得分转换为了树上边的得分。传统方法基于公式3.18将边得分转换为特征设计和对应权重学习问题。该算法的输入仅为包括单个单词、单个单词词性、单词之间距离等在内的基本信息，来计算边  $(h, m)$  的得分  $\lambda_{(h,m)}$ ， $h$  表示依存关系的中心词（head）， $m$  表示该依存关系中的修饰词（modifier）。所采用的神经网络结构如图3.19所示。

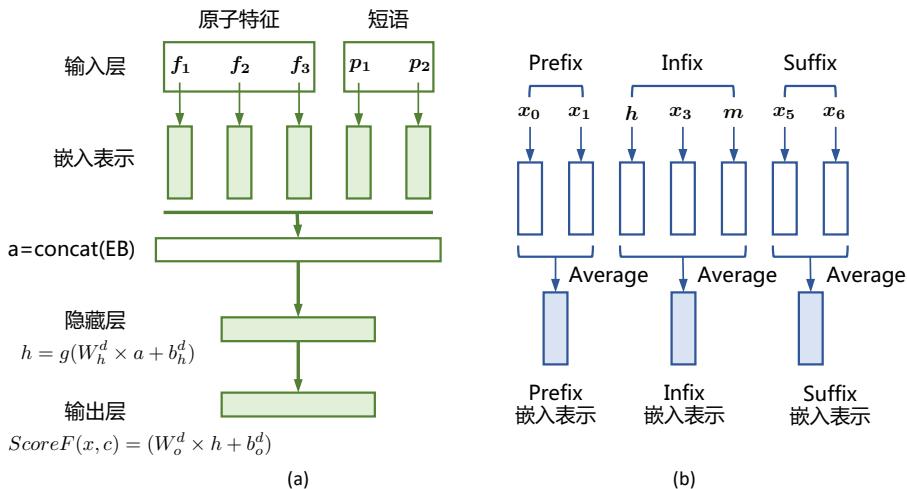


图 3.19 基于神经网络的边评分模型网络结构图<sup>[105]</sup>

模型输入包含两大部分：原子特征（Atomic Features）和短语结构（Phrases）。原子特征部分使用单个词和单个词性。通过查找嵌入矩阵  $M = \mathbb{R}^{d \times |\mathcal{D}|}$  转换为相应的嵌入（Embedding）表示。 $|\mathcal{D}|$  是特征字典大小， $d$  是特征嵌入表示的维度。为了更好的建模依存关系的上下文信息，根据所要预测中心词和修饰词在句子中的位置，将句子切分为前缀（Prefix）、中缀（Infix）和后缀（Suffix）。如图3.19的右边部分所示。采用平均池化（Average Pooling）的方法构建前缀嵌入表示

(Prefix Embedding)、中缀嵌入表示 (Infix Embedding) 和后缀嵌入表示 (Suffix Embedding) 等成短语嵌入 (Phrase Embedding)，并与原子特征一起作为模型输入。

模型第二层将原子特征的嵌入表示和短语嵌入合并到单个向量  $\mathbf{a}$ 。第三层隐藏层利用激活函数 tanh-cube 学习线性变换后的向量  $\mathbf{a}$  多个特征的综合。

$$\mathbf{h} = g(\mathbf{W}_h^d \mathbf{a} + \mathbf{b}_h^d) \quad (3.20)$$

$$g(l) = \tanh(l^3 + l) \quad (3.21)$$

模型最后一层输出层利用线性变换得到边评分函数：

$$\text{Score}_F(x, c) = \mathbf{W}_o^d \mathbf{h} + \mathbf{b}_o^d \quad (3.22)$$

通过最大间隔标准进行模型参数  $\{\mathbf{W}_h^d, \mathbf{b}_h^d, \mathbf{W}_o^d, \mathbf{b}_o^d, \mathbf{M}\}$  进行训练。

## 2. 基于双仿射变换的方法

Deep Biaffine Parser<sup>[106]</sup> 也是采用最大生成树方法构造依存句法树，在文献 [105] 工作的基础上引入了双向长短时记忆网络（BiLSTM），更好的建模句子上下文信息。同时也考虑到直接使用 BiLSTM 的每个时刻隐藏状态序列作为单词表示可能带来的信息过多，容易造成过拟合并需要更多数量的训练语料的问题，提出了双仿射变换的方法。Deep Biaffine Parser 的边评分模型神经网络结构如图 3.20 所示。

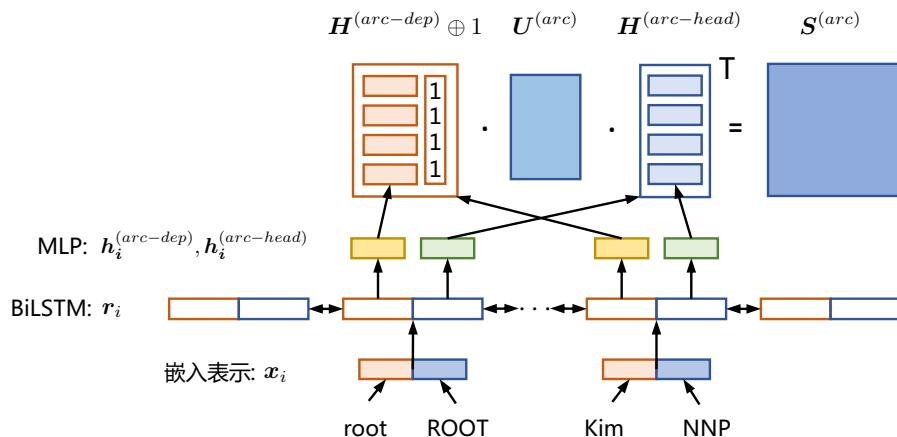


图 3.20 Deep Biaffine Parser 边评分模型神经网络结构图<sup>[106]</sup>

句子中所有单词的词嵌入和对应的词性嵌入合并作为双向长短时记忆网络的输入，记为  $x_i =$

$\mathbf{v}_i^{word} \oplus \mathbf{v}_i^{POS}$ 。所有单词经过 BiLSTM 层计算后得到每个时刻的输出记为  $\mathbf{r}_i$ 。接下来，使用多层感知器 (MLP) 对  $\mathbf{r}_i$  用于中心词和修饰词的不同，分别进行两种不同的降维  $\mathbf{h}_i^{(arc-head)}$  和  $\mathbf{h}_i^{(arc-dep)}$ ：

$$\mathbf{h}_i^{(arc-head)} = \text{MAP}^{(arc-head)}(\mathbf{r}_i) \quad (3.23)$$

$$\mathbf{h}_i^{(arc-dep)} = \text{MAP}^{(arc-dep)}(\mathbf{r}_i) \quad (3.24)$$

所有时刻的  $\mathbf{h}_i^{(arc-head)}$  和  $\mathbf{h}_i^{(arc-dep)}$  分别合并组成矩阵  $\mathbf{H}^{(arc-head)}$  和  $\mathbf{H}^{(arc-dep)}$ 。之后利用不定类别双仿射分类器 (Variable-class Biaffine Classifier) 对  $\mathbf{H}^{(arc-dep)}$  额外拼接了一个单位向量，利用矩阵  $\mathbf{U}^{(arc)}$  进行仿射变换，得到边评分矩阵：

$$\mathbf{S}^{(arc)} = \mathbf{H}^{(arc-dep)} \oplus \mathbf{1} \cdot \mathbf{U}^{(arc)} \cdot \mathbf{H}^{(arc-head)}^\top \quad (3.25)$$

由于依存关系类别数目是确定的，针对类别的分类问题，可以采用下述公式计算单词  $w_i$  作为修饰词  $w_{y_i}$  做为中心词的依存关系类别得分：

$$\mathbf{s}_i^{(label)} = \mathbf{r}_{y_i}^\top \mathbf{U}^{(1)} \mathbf{r}_i + (\mathbf{r}_{y_i} \oplus \mathbf{r}_i)^\top \mathbf{U}^{(2)} + \mathbf{b} \quad (3.26)$$

### 3. 基于图神经网络的方法

文献 [107] 同样采用了在对边评分的基础上，利用贪心或者最大生成树等算法构建依存句法生成树的框架，构造了基于图神经网络的依存句法分析算法 (Graph-based Dependency Parsing with Graph Neural Networks, GNNDP)，试图将更多的结构化信息引入到结点表示，算法的网络架构如图3.21所示。GNNDP 算法从表层的单词序列和深层的树结构两个角度进行节点编码。

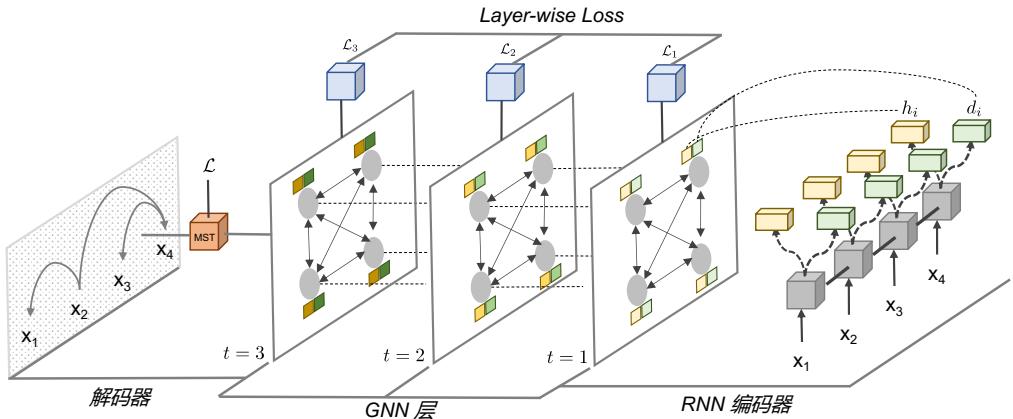


图 3.21 基于图神经网络的依存句法分析网络结构图<sup>[107]</sup>

单词序列表示使用双向长短时记忆网络（BiLSTM）编码单词序列。在每个句子位置  $i$ , 一个前向 LSTM 链（参数为  $\theta^f$ ）通过收集从句子开头到当前位置  $i$  的信息计算出隐向量  $c_i^f$ ; 同样，一个后向 LSTM 链（ $\theta^b$ ）收集从句子结尾到位置  $i$  的信息  $c_i^b$ :

$$c_i^f = \text{LSTM}(\mathbf{x}_i, c_{i-1}^f; \theta^f) \quad (3.27)$$

$$c_i^b = \text{LSTM}(\mathbf{x}_i, c_{i+1}^b; \theta^b) \quad (3.28)$$

$$\mathbf{c}_i = \mathbf{c}_i^f \oplus \mathbf{c}_i^b \quad (3.29)$$

其中  $\mathbf{x}_i$  是 LSTM 单元的输入，由三部分拼接而成：随机初始化的单词嵌入  $e_{w_i}$ ，使用 Glove 的预训练单词嵌入  $e'_{w_i}$  以及随机初始化的词性标签嵌入  $e_{pos_i}$ 。归一化双线性函数基于结点表征来定义得分函数  $\sigma$ 。由于依存边有方向性，因此采用两个多层感知器来生成不同的向量从而区分这两种角色<sup>[106]</sup>:

$$\mathbf{h}_i = \text{MLP}_h(\mathbf{c}_i) \quad (3.30)$$

$$\mathbf{d}_i = \text{MLP}_d(\mathbf{c}_i) \quad (3.31)$$

$$\begin{aligned} \sigma(i, j) &= \text{Softmax}_i(\mathbf{h}_i^\top \mathbf{A} \mathbf{d}_j + \mathbf{b}_1^\top \mathbf{h}_i + \mathbf{b}_2^\top \mathbf{d}_j) \\ &\triangleq P(i | j) \end{aligned} \quad (3.32)$$

其中  $\mathbf{A}, \mathbf{b}_1, \mathbf{b}_2$  是可训练模型参数，输出的得分实际上也是依存边 ( $w_i$  支配  $w_j$ ) 的概率。

树结构表示采用基于图神经网络（Graph Neural Networks, GNNs）进行建模。首先，我们以图注意力网络（Graph Attention Networks, GATs）为例图神经网络介绍的一般框架<sup>[26]</sup>。给定有向图  $G$ ，图神经网络是一个多层网络。它在每一层通过聚合其邻居的信息来维护一组结点的表示。对于结点  $i$ ，用  $\mathcal{N}(i)$  表示  $G$  中结点  $i$  的邻居，用  $\mathbf{v}_i^t$  表示结点  $i$  在图神经网络第  $t$  层的隐向量。 $\mathbf{v}_i^t$  的计算如下：

$$\mathbf{v}_i^t = g\left(W \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^t \mathbf{v}_j^{t-1} + B \mathbf{v}_i^{t-1}\right). \quad (3.33)$$

其中  $g$  是一个非线性激活函数 LeakyReLU， $\mathbf{W}$  和  $\mathbf{B}$  是参数矩阵。边权  $\alpha_{ij}^t$  表示邻居结点  $j$  在构建  $\mathbf{v}_i^t$  时的贡献值。可以看到，图神经网络自然地捕捉到了多跳（即高阶）关系。以前两层为例，对于第二层的每个结点  $i$ ， $\mathbf{v}_i^2$  包含其 1 跳邻居  $\mathbf{v}_j^1$  的信息。由于  $\mathbf{v}_j^1$  已经在第一层编码了它自己的 1 跳邻居， $\mathbf{v}_i^2$  实际上编码了它的 2 跳邻居的信息。因此，图神经网络可以有助于编码高阶结构特征。

GNNDP 将图神经网络应用到依存句法分析任务上。如图 3.21 所示，解析任务需要在图  $G$  上同时处理支配词表征  $\mathbf{h}_i$  和从属词表征  $\mathbf{d}_i$ ，而不是为每个结点编码一个向量。此外，为了近似精确的高阶解析，解析器需要每个 GNNs 网络层有关于句子的具体含义。因此，GNNDP 采用完全图

(即所有结点都是连接的), 并用依存边条件概率 (公式 3.32) 设置边权。

$$\alpha_{ij}^t = \sigma^t(i, j) = P^t(i | j) \quad (3.34)$$

GNNDP 关注三类高阶信息, 即祖父母、孙子和兄弟关系, 如图 3.22 所示, 其中灰色的阴影表示哪些结点表征已经存在于一阶特征中。橙色阴影表示在高阶特征中应该包括哪些结点表征。需要调整一般的 GNNs 更新公式, 将它们适当地编码到节点表示中。首先, 为了纳入祖父母的信息 (图 3.22(a)), 使得  $\sigma^t(j, i)$  不仅考虑一跳父子关系 ( $j, i$ ), 还考虑两跳  $i$  的祖先结点 (用  $k$  表示)。同样, 为了在  $\sigma^t(j, i)$  中对两跳的  $j$  的孙子结点进行编码 (也用  $k$  表示), 需要聚合邻居结点的从属词表示。

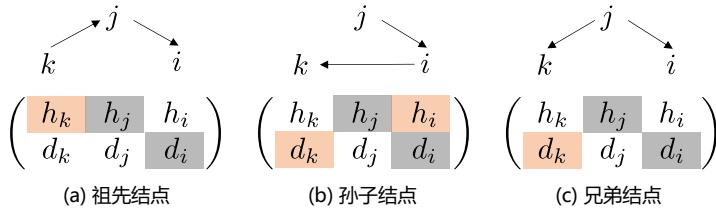


图 3.22 三种类型的高阶信息集成在依存边  $(j, i)$  中

GNNDP 采用以下协议:

$$\begin{cases} \mathbf{h}_i^t = g\left(W_1 \sum_{j \in \mathcal{N}(i)} \alpha_{ji}^t \mathbf{h}_j^{t-1} + B_1 \mathbf{h}_i^{t-1}\right) \\ \mathbf{d}_i^t = g\left(W_2 \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^t \mathbf{d}_j^{t-1} + B_2 \mathbf{d}_i^{t-1}\right) \end{cases} \quad (3.35)$$

为了纳入  $i$  的兄弟结点 (同样用  $k$  表示) 的信息 (图 3.22(c)),  $\mathbf{h}_j^t$  的更新涉及到  $\mathbf{d}_k^{t-1}$ , 这是第二个更新协议:

$$\begin{cases} \mathbf{h}_i^t = g\left(W_1 \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^t \mathbf{d}_j^{t-1} + B_1 \mathbf{h}_i^{t-1}\right) \\ \mathbf{d}_i^t = g\left(W_2 \sum_{j \in \mathcal{N}(i)} \alpha_{ji}^t \mathbf{h}_j^{t-1} + B_2 \mathbf{d}_i^{t-1}\right) \end{cases} \quad (3.36)$$

将公式 3.35 和 3.36 整合到一个更新中, 以统一的方式处理三种高阶信息。与一般的 GNNs 相比, 上述的更新方式是为解析任务定制的。最后, 除了默认的同步设置, GNNDP 还研究了异步更新版本, 最终的更新协议如下:

$$\begin{cases} \mathbf{h}_i^{t-\frac{1}{2}} = g\left(W_1 \sum_{j \in \mathcal{N}(i)} (\alpha_{ji}^t \mathbf{h}_j^{t-1} + \alpha_{ij}^t \mathbf{d}_j^{t-1}) + B_1 \mathbf{h}_i^{t-1}\right) \\ \mathbf{d}_i^t = g\left(W_2 \sum_{j \in \mathcal{N}(i)} (\alpha_{ji}^t \mathbf{h}_j^{t-\frac{1}{2}} + \alpha_{ij}^t \mathbf{d}_j^{t-1}) + B_2 \mathbf{d}_i^{t-1}\right) \end{cases} \quad (3.37)$$

训练目标由两部分组成。第一部分是 GNNs 层后面有一个解码器 (用  $\tau$  表示), 它将对树结构

(使用  $P^\tau(i|j)$ ) 和边标签 (使用  $P(r|i,j)$ ) 进行解码。为了预测边标签, 使用用另一个 MLP 来衡量依存边  $(i,j)$  具有关系  $r$  的可能性。其交叉熵损失函数定义如下:

$$\mathcal{L}_0 = -\frac{1}{n} \sum_{(i,j,r) \in T} (\log P^\tau(i | j) + \log P(r | i, j)) \quad (3.38)$$

第二个部分是每个 GNNs 层的  $P^t(i|j)$  的准确率进行监督 (只对树结构进行监督, 忽略边的标签)。每层交叉熵损失函数定义如下:

$$\mathcal{L}' = \sum_{t=1}^{\tau} \mathcal{L}_t = \sum_{t=1}^{\tau} -\frac{1}{n} \sum_{(i,j,r) \in T} \log P^t(i | j) \quad (3.39)$$

最终目标是最小化两者的加权组合  $L = \lambda_1 \mathcal{L}_0 + \lambda_2 \mathcal{L}'$ 。训练完成后, 解析器在所有的依存边得分上使用最大生成树算法形成一棵有效的依存树。

### 3.3.3 基于转移的依存句法分析

转移系统 (Transition System) 包含状态集合 (State 或 Configuration) 以及状态之间的转移动作集合 (Transition)。有限状态自动机就是属于转移系统的最简单的一种, 但是应用于依存句法分析的转移系统更为复杂。标准弧 (Arc-Standard) 转移系统<sup>[108]</sup> 是其中最常用的投射性依存句法分析转移系统之一。按照标准弧转移系统定义, 状态  $c = (\sigma, \beta, A)$  由三个部分组织,  $\sigma$  表示已处理的单词的堆栈,  $\beta$  表示尚未处理的单词的缓冲器,  $A$  表示已经构建的依存关系边集合。对于给定的句子  $S = w_0, w_1, \dots, w_n$ , 其初始状态  $c_0(S) = ([w_0]_\sigma, [w_1, \dots, w_n]_\beta, [\emptyset]_A)$ ,  $(\sigma, [], \beta, A)$  则对应终止状态的形式。标准弧转移系统中转移动作包含以下三类:

- (1) 左弧 (LA<sub>r</sub>):  $([\sigma|w_i, w_j|\beta, A) \Rightarrow (\sigma, w_j|\beta, A \cup \{(w_j, r, w_i)\})$ ,  $w_i$  是堆栈  $\sigma$  最上面的单词,  $w_j$  是缓冲器  $\beta$  中最前面的单词, 将  $w_i$  从堆栈中弹出并增加从  $w_j$  到  $w_i$  关系类型为  $r$  的依存关系。前置条件是  $i \neq 0$  并且  $\neg \exists w_k \exists r' [(w_k, r', w_i) \in A]$ 。
- (2) 右弧 (RA<sub>r</sub>):  $(\sigma|w_i, w_j|\beta, A) \Rightarrow (\sigma, w_i|\beta, A \cup \{(w_i, r, w_j)\})$ ,  $w_i$  是堆栈  $\sigma$  最上面的单词,  $w_j$  是缓冲器  $\beta$  中最前面的单词, 将  $w_i$  从堆栈中弹出, 将缓冲器中最前面的单词  $w_j$  替换为  $w_i$ , 增加从  $w_i$  到  $w_j$  关系类型为  $r$  的依存关系。前置条件是  $\neg \exists w_k \exists r' [(w_k, r', w_j) \in A]$ 。
- (3) 移进 (SH):  $(\sigma, w_i|\beta, A) \Rightarrow (\sigma|w_i, \beta, A)$ , 从缓冲器中移除最前面的单词  $w_i$ , 并压入堆栈中。

针对一个句子的转移动作序列(Transition Sequence)可以对应状态序列  $C_{0,m} = (c_0, c_1, \dots, c_m)$ 。 $c_0$  表示起始状态,  $c_0(S)$  表示针对句子  $S$  的起始状态,  $c_m$  是终止状态。转换动作为  $t \in \mathcal{T}$  使得状态  $c_{i-1}$  转换到状态  $c_i$  的, 使用  $c_i = t(c_{i-1})$  表示。针对句子“她非常喜欢跳芭蕾”的依存句法分析转移动作序列如表3.1所示。

假设存在函数  $o$ , 可以根据当前的状态  $c$  正确的确定下一步的转移动作  $t$ , 即  $o(c) = t$ 。那么整个句法分析的过程就可以使用非常简单的贪心算法完成。针对输入句子  $S$ , 首先, 构造初始状

表 3.1 依存句法分析转移动作序列样例

	$\sigma$	$\beta$	$A$
	([ROOT],	[她, ...],	$\emptyset$ )
SH $\Rightarrow$	([ROOT, 她],	[非常, ...],	$\emptyset$ )
SH $\Rightarrow$	([ROOT, 她, 非常],	[喜欢, ...],	$\emptyset$ )
LA <sub>advmod</sub> $\Rightarrow$	([ROOT, 她],	[喜欢, ...],	$A_1 = \text{喜欢, advmod, 非常}$ )
LA <sub>nsubj</sub> $\Rightarrow$	([ROOT],	[喜欢, ...],	$A_2 = A_1 \cup (\text{喜欢, nsubj, 她})$
SH $\Rightarrow$	([ROOT, 喜欢],	[跳, ...],	$A_2$ )
SH $\Rightarrow$	([ROOT, 喜欢, 跳],	[芭蕾舞],	$A_2$ )
RA <sub>dobj</sub> $\Rightarrow$	([ROOT, 喜欢],	[跳],	$A_3 = A_2 \cup (\text{跳, dobj, 芭蕾舞})$
RA <sub>xcomp</sub> $\Rightarrow$	([ROOT],	[喜欢],	$A_4 = A_3 \cup (\text{喜欢, xcomp, 跳})$
RA <sub>root</sub> $\Rightarrow$	([],	[ROOT],	$A_5 = A_4 \cup (\text{ROOT, root, 喜欢})$
SH $\Rightarrow$	([ROOT],	[]	$A_5$ )

态  $c_0(S)$ , 调用函数  $o$  得到下一步转移动作  $t = o(c_0)$ 。之后, 利用根据转移动作得到下一个状态  $c = t(c_0)$ 。如此循环直达到到终止状态形式  $(\sigma, []_\beta, A)$ 。

转移动作预测函数可以使用分类器进行建模。构造一个分类器  $f(c)$ , 输入为状态  $c$ , 输出为其所对应的标准转移动作  $o(c)$ 。由此, 基于转移的依存句法分析问题转化为了典型的机器学习问题。如果采用有监督机器学习算法, 那么需要解决如下三个基本问题: (1) 如何表示状态  $c$ ; (2) 如何构造训练语料; (3) 如何选择和训练分类器。

针对状态  $c$  的表示问题, 可以采用从堆栈  $\sigma$ 、缓存器  $\beta$  以及边集合  $A$  中对特定位置的单词、单词词性、树结构等抽取信息构造特征向量的方法。Maltparser<sup>[108]</sup> 针对标准弧转移系统利用如表3.2所示的信息构造针对状态  $c$  的特征向量。 $\beta[0]$  表示缓存器中最前面的单词,  $\sigma[0]$  表示堆栈最顶端的单词,  $ld(x)$  表示  $x$  最左面的修饰词,  $rd(x)$  表示  $x$  最右面的修饰词。利用这些特征模板, 可以构造状态的高维向量表示。使用  $f(c)$  表示特征函数, 其输出是状态  $c$  的特征向量。

依存句法树库  $\mathbb{D} = \{(S_d, G_d)\}_{d=0}^{|\mathbb{D}|}$  是由大规模的句子  $S_d$  以及所对应的正确的依存句法树  $G_d$  组成。但是用于转移动作预测的分类器所需的训练数据集  $\mathbb{D}'$  需要由状态的特征表示  $f(c)$  和对应的正确的转移动作  $t$  组成,  $\mathbb{D}' = \{(f(c_d), t_d)\}_{d=0}^{|\mathbb{D}'|}$ 。因此, 需要将原始依存句法树库  $\mathbb{D}$  中每个句子和对应依存句法树  $(S_d, G_d)$  转换为转移序列  $C_{0,m}^d = (c_0^d, c_1^d, \dots, c_m^d)$ 。其中每个非终结状态  $c_i^d \in C_{0,m}^d$

表 3.2 Maltparser<sup>[108]</sup> 针对标准弧转移系统采用的特征向量

	单词	词元	粗粒度词性	细粒度词性	词形特征	依存关系类型
$\beta[0]$	×	×	×	×	×	
$\beta[1]$	×			×		
$\beta[2]$				×		
$\beta[3]$				×		
$ld(\beta[0])$						×
$rd(\beta[0])$						×
$\sigma[0]$	×	×	×	×	×	
$\sigma[1]$				×		
$ld(\sigma[0])$						×
$rd(\sigma[0])$						×

都可以根据如下方式得到其对应的转移动作  $t_i^d = o(c_i^d)$ :

$$o(c = (\sigma, \beta, A)) = \begin{cases} LA_r & \text{如果 } (\beta[0], r, \sigma[0]) \in A_d \\ RA_r & \text{如果 } (\sigma[0], r, \beta[0]) \in A_d \text{ 并且} \\ & \text{如果 } (\beta[0], r', w) \in A_d \text{ 那么 } (\beta[0], r', w) \in A \\ SH & \text{其他情况} \end{cases} \quad (3.40)$$

基于上述公式，可以根据非终结状态  $c_i^d$  和对应的转移动作  $t_i^d$ ，添加实例  $(\mathbf{f}(c_i^d), t_i^d)$  到训练数据集合  $\mathbb{D}'$  中。

利用训练数据集合  $\mathbb{D}' = \{(\mathbf{f}(c_d), t_d)\}_{d=0}^{|\mathbb{D}'|}$  可以选用各种类型的分类器，已经成功应用于该任务的分类器包括：SVM<sup>[109]</sup>、基于记忆的学习<sup>[110]</sup>、最大熵<sup>[111]</sup>等。相关的有监督分类算法在机器学习相关书籍和本书其他章节中也多有介绍，这里就不再赘述。

### 3.3.4 基于神经网络的转移依存句法分析

基于转移的依存句法分析通过定义转移系统，将依存句法分析转换为了根据当前状态  $c = (\sigma, \beta, A)$  预测转移动作  $t$  的分类问题。上节介绍了依赖特征工程的分类方法，本节将介绍两种基于神经网络转移依存句法分析算法。

#### 1. 基于前馈神经网络的方法

文献 [112] 基于 MaltPraser 的基本思想，提出了将当前状态  $c$  中的堆栈  $\sigma$  和缓冲器  $\beta$  利用神经网络提取特征，并预测下一步的转移动作的方法，其神经网络结构如图3.23所示。神经网络结构由三层组成：输入层、隐藏层和输出层。

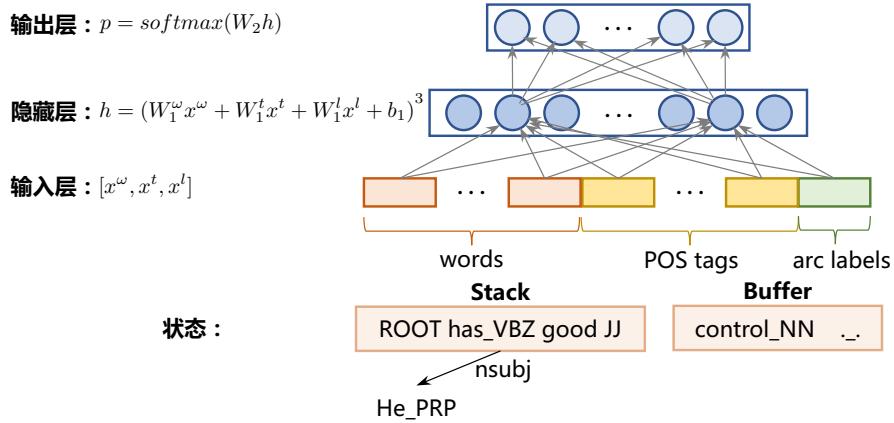


图 3.23 基于神经网络的依存句法分析算法神经网络结构图<sup>[112]</sup>

输入层是由单词嵌入、词性嵌入以及边类别嵌入。单词嵌入表示 (word embedding) 由  $d$  维向量  $e_i^w \in \mathbb{R}^d$  表示, 整个单词嵌入矩阵使用  $\mathbf{E}^w \in \mathbb{R}^{d \times N_w}$  表示, 其中  $N_w$  表示词典中单词数量。同样的使用  $e_i^t$  和  $e_j^l$  表示第  $i$  个词性标签的嵌入和第  $j$  个边类别的嵌入表示。对应的词性标签嵌入矩阵使用  $\mathbf{E}^t \in \mathbb{R}^{d \times N_t}$  表示, 其中  $N_t$  表示词性数量, 边类别标签嵌入矩阵使用  $\mathbf{E}^l \in \mathbb{R}^{d \times N_l}$  表示, 其中  $N_l$  表示边类型数量。

单词嵌入、词性嵌入以及边类别嵌入根据设定的上下文分别连接起来构成  $\mathbf{x}^w, \mathbf{x}^t, \mathbf{x}^l$ .  $\mathbf{x}^w$  是由堆栈最顶端的 3 个单词和缓冲器中最前面的 3 个单词:  $\sigma[0], \sigma[1], \sigma[2], \beta[0], \beta[1], \beta[2]$ , 以及堆栈中最顶端的 2 个单词的最左、次左、最右和次右的儿子节点:  $lc_1(\sigma[i]), lc_2(\sigma[i]), rc_1(\sigma[i]), rc_2(\sigma[i]), i = 1, 2$ , 再加上堆栈中最顶端的 2 个单词的最左的孙子节点和最右的孙子节点:  $lc_1(lc_1(\sigma[i])), rc_1(rc_1(\sigma[i])), i = 1, 2$ , 共 18 个单词嵌入连接组成。 $\mathbf{x}^t$  是由组成  $\mathbf{x}^w$  的 18 个单词的词性嵌入组成。 $\mathbf{x}^l$  是  $\mathbf{x}^w$  中除了堆栈和缓冲器的 6 个单词之外的其他 12 个单词对应的边类别嵌入连接组成。

隐藏层为了有效的融合单词、词性以及边类别信息, 采用立方激活函数结合线性变换:

$$\mathbf{h} = (\mathbf{W}_1^w \mathbf{x}^w + \mathbf{W}_1^t \mathbf{x}^t + \mathbf{W}_1^l \mathbf{x}^l + \mathbf{b}_1)^3 \quad (3.41)$$

其中,  $\mathbf{W}_1^w \in \mathbb{R}^{d_h \times (d \cdot n_w)}$ ,  $\mathbf{W}_1^t \in \mathbb{R}^{d_h \times (d \cdot n_t)}$ ,  $\mathbf{W}_1^l \in \mathbb{R}^{d_h \times (d \cdot n_l)}$ ,  $\mathbf{b}_1 \in \mathbb{R}^{d_h}$ ,  $n_w$  是  $\mathbf{x}^w$  中单词的数量,  $n_t$  是  $\mathbf{x}^t$  中词性的数量,  $n_l$  是  $\mathbf{x}^l$  中边标签的数量,  $d_h$  是隐藏单元  $\mathbf{h}$  的维度。

输出层则是采用标准的 softmax 函数决定转移动作  $P = \text{softmax}(\mathbf{W}_2 \mathbf{h})$ , 其中  $\mathbf{W}_2 \in \mathbb{R}^{|\mathcal{T}| \times d_h}$ ,  $\mathcal{T}$  是转移动作集合。

最后, 可以根据上节所介绍的通过依存句法树库构造训练语料  $\{(c_i, t_i)\}_{i=1}^m$ , 其中  $c_i$  是状态,

$t_i \in \mathcal{T}$  是对应的转移动作。利用交叉熵损失做为训练目标：

$$\mathcal{L}(\theta) = -\sum_i \log P(t_i|c_i) + \frac{\lambda}{2} \|\theta\|^2 \quad (3.42)$$

利用优化算法学习参数集合  $\theta = \{\mathbf{W}_1^w, \mathbf{W}_1^t, \mathbf{W}_1^l, \mathbf{b}_1, \mathbf{W}_2, \mathbf{E}^w, \mathbf{E}^t, \mathbf{E}^l\}$ 。

## 2. 基于堆栈长短时记忆网络的方法

文献 [113] 针对基于转移的依存句法分析中如何对堆栈表示的问题，提出了一种堆栈长短时记忆网络（Stack-LSTM）方法。不同于标准长短时记忆网络的序列是从左到右的顺序，基于转移的依存句法分析中需要使用堆栈完成句法分析，因此堆栈长短时记忆网络针对堆栈的压栈（push）和出栈（pop）操作，增加了“栈指针”（stack pointer）来扩充 LSTM。针对出栈操作，只是将堆栈指针移动到栈顶的前一个元素。针对压栈操作，需要添加对应的输入  $x_t$ ，长短时记忆网络的内部状态计算  $c_t$  所依赖的前一个状态的  $c_{t-1}$  和  $h_{t-1}$  都指向栈指针所指向的位置。堆栈长短时记忆网络压栈和出栈操作样例如图3.24所示。

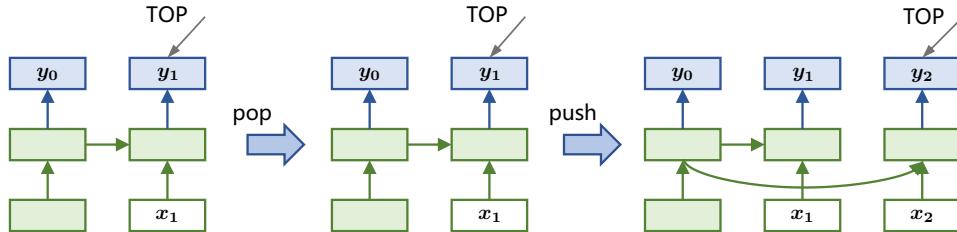


图 3.24 堆栈长短时记忆网络压栈和出栈操作样例<sup>[113]</sup>

Stack-LSTM 在原始从左至右 LSTM 的基础上，增加了堆栈指针（图中标记为 TOP）。图3.24给出了三种状态：带有单个元素的堆栈（左侧）、堆栈执行出栈操作的结果（中部），以及此后堆栈使用压栈操作的结果（右侧）。最底层代表堆栈的内容，是 Stack-LSTM 的输入，中间层是 LSTM 单元，上层是输出层。使用堆栈指针指向的位置作为输出表示。通过上述操作我们可以看到，在执行出栈操作时，Stack-LSTM 中间层 LSTM 单元并不发生改变，仅是修改堆栈指针 TOP 的位置。在执行压栈操作时，新增加的元素增加在最右端，并与前一个 TOP 指针指向的 LSTM 单元连接。

分析算法采用标准弧转移系统，状态  $c = (\sigma, \beta, A)$  包含已处理的单词的堆栈（ $S$ ），尚未处理的单词的缓冲器（ $B$ ），以及分析算法所执行的转移动作历史（ $A$ ）。利用三个堆栈长短时记忆网络分别对状态中堆栈、缓存器以及转移动作历史进行表示，其神经网络架构如图3.25所示。

在  $t$  时刻，句法分析器状态的表示  $p_t$  根据三个堆栈长短时记忆网络的输出经过线性变换和非

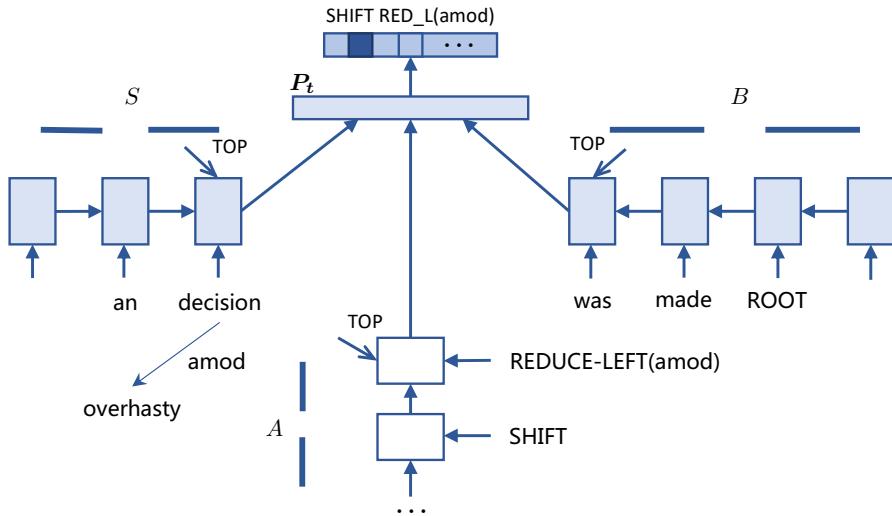


图 3.25 基于堆栈长短时记忆网络的状态表示神经网络结构图<sup>[113]</sup>

线性 ReLU 函数转换后得到。 $\mathbf{p}_t$  具体计算公式如下：

$$\mathbf{p}_t = \max\{\mathbf{0}, \mathbf{W}[\mathbf{s}_t, \mathbf{b}_t, \mathbf{a}_t] + \mathbf{d}\} \quad (3.43)$$

由此可以得到在  $t$  时刻在每个转移动作的概率：

$$P(z_t | \mathbf{p}_t) = \frac{\exp \mathbf{g}_{z_t}^\top \mathbf{p}_t + q_{z_t}}{\sum_{z' \in \mathcal{A}(S, B)} \exp \mathbf{g}_{z'}^\top \mathbf{p}_t + q_{z'}} \quad (3.44)$$

其中  $\mathbf{g}_z$  表示转移动作  $z$  的向量表示， $q_z$  是转移动作  $z$  对应的偏置项， $\mathcal{A}(S, B)$  是给定当前堆栈  $S$  和缓冲器  $B$  状态时所有可能的转移动作集合。根据每个转移动作的概率可以选择下一步的转移动作，从而完成依存句法分析。

### 3.3.5 依存句法分析评价方法

依存句法分析算法的性能评价，目前较为常用的指标是由 Yamada 和 Matsumoto 在文献 [114] 采用的三种指标：

- (1) 依存准确率 (Dependency Accuracy, 简称 DA)：正确分析得到其中心词的非根节点词语个数占总非根节点词数的百分比；
- (2) 根准确率 (Root Accuracy, 简称 RA)：正确根节点的个数与句子个数的百分比；

(3) 完全匹配率 (Complete Match, 简称 CM): 无标记依存结构完全正确的句子占句子总数的百分比;

此外, 如果考虑所有词的依存准确率, 根据是否考虑依存标记, 还可以使用无标记依存准确率 (Unlabeled attachment score, 简称 UAS) 和有标记依存准确率 (Labeled attachment score, 简称 LAS) 两个指标。UAS 表示正确分析得到其修饰词的词语个数占总词数的百分比。LAS 表示正确分析得到其修饰词以及依存关系的词语个数占总词数的百分比。

图3.26给出了依存句法分析样例, 图3.26(a)表示句子“他非常喜欢跳芭蕾舞”和正确分析结果, 图3.26(b)表示算法分析结果。

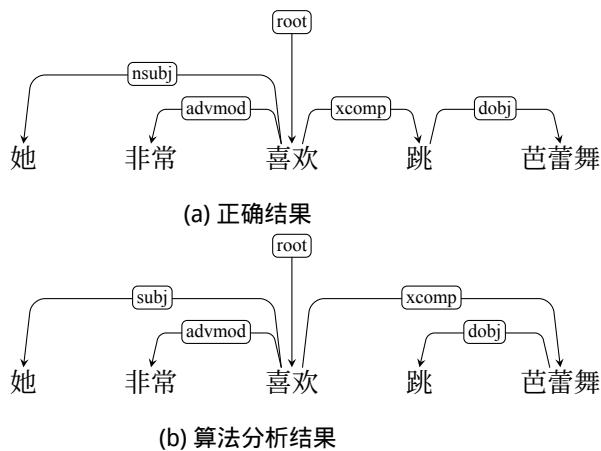


图 3.26 依存句法分析结果评测样例

根据上述评价指标定义, 可以得到如下评价结果:

$$\text{依存准确率 (DA)} = \frac{2}{4} \times 100\% = 50\%$$

$$\text{无标记依存准确率 (UAS)} = \frac{3}{5} \times 100\% = 60\%$$

$$\text{有标记依存准确率 (LAS)} = \frac{2}{5} \times 100\% = 40\%$$

## 3.4 句法分析语料库

句法分析语料库也称为句法树库, 包含大规模句子以及其对应句法树的集合。通过本章的介绍, 可以知道很多句法分析方法都转换为了有监督机器学习算法, 因此句法分析算法的训练和评测都依赖语料库的建设。常见句法分析语料库如表3.3所示。本节将介绍几种常见的句法分析语料库。

表 3.3 常见句法分析语料库汇总

语料库名称	单词数量	语法类型	语言
英语宾州树库 (PTB)	117 万	成分语法	英文
通用依存树库 (UD V2.0 CoNLL 2017)	281 万	依存语法	多语言
通用依存树库 (UD V2.2 CoNLL 2018)	1714 万	依存语法	多语言
组合范畴语法树库 (CCGBank)	116 万	组合范畴语法	英文
中文宾州树库 6.0 (CTB 6.0)	78 万	成分语法	中文
中文宾州树库 7.0 (CTB 7.0)	120 万	成分语法	中文
中文宾州树库 8.0 (CTB 8.0)	162 万	成分语法	中文
中文宾州树库 9.0 (CTB 9.0)	208 万	成分语法	中文
中文语义依存树库 (SDP)	52 万	语义依存	中文

### 1. 英语宾州树库

英语宾州树库 (English Penn Treebank, PTB) 是最知名和最常用的短语结构句法树库之一。语言资源联盟 (Linguistic Data Consortium, LDC) 1999 年的发行版 Treebank-3 是 PTB 最常用的发行版本。Treebank-3 中包含四个部分: ATIS (Air Travel Information System 标注样例)、BROWN (布朗语料库标注)、SWBD (Switchboard 口语数据标注)、WSJ (华尔街日报标注)。WSJ 是英语宾州树库中最常用的部分, 其原始数据来自于 1989 年华尔街日报文章, 按照 PTB(V2) 的标注策略进行标注, 使用括号表示树结构。WSJ 部分总计包含 49208 个句子, 1173766 个单词, 分成 25 个章 (section), 通常使用 0 到 18 章作为训练集合, 19 到 21 章作为验证集合, 22 到 24 章作为测试集合。

示例:

```
( (S
  (NP-SBJ (DT Both) (NNS distributions) )
  (VP (VBP are)
    (ADJP-PRD (JJ payable)
      (NP-TMP (NNP Dec.) (CD 4) )
      (PP (TO to)
        (NP
          (NP (JJ limited) (NNS partners) )
          (PP (IN of)
            (NP (NN record) ))
          (NP-TMP (NNP Nov.) (CD 3) )))))
      (. .) )))
```

## 2. 中文宾州树库

中文宾州树库 (Chinese Penn Treebank, CTB) 是目前最常用的大规模中文短语结构句法标注语料库之一。1998 年开始构建, Chinese Treebank 1.0 规模达到 10 万词规模, 其原始数据来源于新华社新闻报道文章。2016 年发布了最新的 Chinese Treebank 9.0 版本, 包含中文新闻网站、政府文书、杂志文章、新闻群组、广播对话节目、博客等各类不同来源的 3726 篇文章, 共计 132076 个句子, 2084387 个单词, 3247331 个中文和外文字符。对这些文章中的句子进行了分词、词性标注和成分句法树标注。采用了英语宾州树库的括号结构对树结构进行表示。

示例:

```
( ( IP (NP-SBJ (DNP (NP-PN (NR 北海市))
          (DEG 的)))
        (NP (NN 崛起 ))))

(PU , )
(VP (VC 是)
  (NP-PRD (CP-APP (IP (IP-SBJ (LCP-TMP (NP (NT 近年 ))
          (LC 来)))
        (NP-PN-SBJ (NR 广西)
          (NN 壮族)
          (NN 自治区)))
        (VP (PP-DIR (P 对)
          (NP (NN 外))))
          (VP (VV 开放 )))))

( VP (VV 取得)
  (NP-OBJ (ADJP (JJ 卓著))
    (NP (NN 成就 )))))

(DEC 的)
(ADJP (JJ 重要))
(NP (NN 标志)
  (NN 之一 ))))

(PU 。)) )
```

## 3. 通用依存树库

通用依存树库 (Universal Dependencies, UD) 是一个为多种语言开发的跨语言一致的依存句法树库项目。标注方案采用斯坦福通用依存标签<sup>[93, 115, 116]</sup>、Google 通用词性标签<sup>[60]</sup>以及形态句法标签集<sup>[117]</sup>。目前共有超过 300 个贡献者提供了 100 多种语言的 200 多个依存句法树库。树库中包含了原始句子、词语、词性、依存关系以及依存关系类型等信息。

示例:

1	同样	同样	ADJ	JJ	_	3	amod	_	SpaceAfter=No
2	的	的	PART	DEC	Case=Gen	1	case	_	SpaceAfter=No
3	手法	手法	NOUN	NN	_	5	nsubj:pass	_	SpaceAfter=No
4	被	被	VERB	BB	Voice=Pass	5	aux:pass	_	SpaceAfter=No
5	用	用	VERB	VV	_	0	root	_	SpaceAfter=No
6	于	于	VERB	VV	_	5	mark	_	SpaceAfter=No
7	现代	现代	NOUN	NN	_	10	nmod	_	SpaceAfter=No
8	的	的	PART	DEC	Case=Gen	7	case	_	SpaceAfter=No
9	摄影	摄影	NOUN	NN	_	10	nmod	_	SpaceAfter=No
10	技术	技术	NOUN	NN	_	5	obj	_	SpaceAfter=No
11	中	中	ADP	IN	_	10	acl	_	SpaceAfter=No
12	。	。	PUNCT	.	_	5	punct	_	SpaceAfter=No

#### 4. 中文语义依存树库

中文语义依存树库（Chinese Semantic Dependency Parsing, SDP）是由哈尔滨工业大学车万翔教授在 SemEval-2016 发布。语料库含包含从新闻中选取的 10068 个句子和从小学课文中选取的 14793 个句子。新闻句子平均长度是 31 个词，课文句子平均长度是 14 个词。与传统的依存句法树库不同，中文语义依存树库从语义角度构建依存关系，并依存结构扩展到符合汉语特点的有向无环图-汉语语义依存图结构<sup>[118]</sup>。定义了 45 个标签用来描述论元 (Argument) 之间的语义关系，19 个标签用来描述谓词 (Predicate) 之间的关系，以及 17 个标签用来提供更丰富的谓词描述。

示例：

1	他	他	PN	PN	_	2	Exp	_	_
2	是	是	VC	VC	_	0	Root	_	_
3	研究	研究	NN	NN	_	6	rAgt	_	_
4	机器人	机器人	NN	NN	_	3	Datv	_	_
5	的	的	DEG	DEG	_	3	mAux	_	_
6	专家	专家	NN	NN	_	2	Clas	_	_

## 3.5 延伸阅读

句法分析句法分析是句子结构和语义之间的桥梁，具有非常重要的作用，很多自然语言处理算法需要依赖句法分析结果，因此句法分析效果也直接影响到很多自然语言处理应用。句法分析是自然语言处理中长期关注的核心问题之一。

在本章中，句法分析任务限定在得到完整的句法分析树，重点介绍了基于有监督机器学习算法的句法分析方法。在实际应用中，有很多任务并不需要使用完整的句法分析树，仅依赖部分或者粗粒度的分析结果，这种句法分析称之为部分句法分析 (Partial Parsing) 又称浅层句法分析 (Shallow

Parsing)。组块分析 (Chunking) 是最常用的浅层句法分析任务，目标是将句子分解为不重叠的片段，这些片段是由主要内容词的非递归短语组成，包括：名词短语、动词短语、形容词短语、介词短语等。例如：句子“*He eat soup with spoon.*”的组块分析结果为 [*NP He*] [*VP eat soup*] [*PP with spoon.*] 组块分析通常也转换为分类问题，包括规则<sup>[119]</sup>、最大熵<sup>[120]</sup>、支撑向量机<sup>[121]</sup>、条件随机场<sup>[122]</sup>、卷积神经网<sup>[14]</sup> 等方法都针对组块分析任务开展了研究。

在成分句法分析方面，除了本章中所介绍的算法之外，还有一些工作从如何对句法分析历史进行表示对子树评分<sup>[123]</sup>、基于神经网络的方法对子树和标签评分<sup>[124, 125]</sup>、将句法分析任务转换为机器翻译任务<sup>[126, 127]</sup>、利用特定的注意力机制<sup>[128, 129]</sup>、多语言预训练<sup>[130]</sup> 等方面开展了深入研究。此外针对大规模句法树标注困难的问题，一些研究工作引入了无标注数据，设计了多种半监督算法以降低对标注数据的需求，包括基于期望最大化算法 (Expectation Maximization)<sup>[131, 132]</sup>、自监督方法<sup>[133]</sup>、半监督重排序<sup>[134]</sup> 等。还有一些研究工作试图通过跨语言迁移方法，将标注资源丰富的汉语、英语语料库或者模型迁移到资源缺乏的语言中，采用了包括标注迁移<sup>[135]</sup>、结合机器翻译算法<sup>[136]</sup>、模型迁移等方法<sup>[137]</sup>。另外一些方法利用无标记数据，包括基于语言模型预训练<sup>[138]</sup>、递归自编码器 (Recursive Auto-Encoders)<sup>[139]</sup>、标准化流 (Normalizing Flow)<sup>[140]</sup> 等方法。

在依存句法分析方面，针对如何减少大规模训练语料问题，很多工作从无监督方法、半监督方法以及迁移方法三个方面开展研究。在无监督方法方面，研究人员们提出了基于有价的依存关系理论<sup>[141–143]</sup>，维特比期望最大化 (Viterbi Expectation Maximization)<sup>[144]</sup>、声学线索<sup>[145]</sup>、条件随机场自编码器<sup>[146]</sup> 等方法。在半监督方法方面，研究人员们针对无监督数据特征提取<sup>[147]</sup>、协同训练<sup>[148]</sup>、特征转换<sup>[149]</sup>、歧义感知的集成学习<sup>[150]</sup>、自监督训练<sup>[151]</sup>、弧因子变分自编码器 (Arc-factored variational autoencoding)<sup>[152]</sup>、跨语言训练<sup>[153–155]</sup> 等方法开展了一系列的工作。在迁移方法方面，提出了跨语言上下文表示对齐<sup>[156]</sup>、分布迁移<sup>[157]</sup>、跨语言 BERT 模型迁移<sup>[158]</sup> 等方法。此外，还有一些工作面向社会媒体非规范语言的依存句法分析<sup>[159]</sup>、针对混合语言句子的依存句法分析<sup>[160]</sup>、基于序列生成方法的依存句法分析<sup>[161]</sup>、基于二阶树条件随机场的依存句法分析<sup>[162]</sup> 等围绕依存句法分析的各个方面开展研究。

## 3.6 习题

- (1) 如何判断一个语法理论属于成分语法还是依存语法？
- (2) 试举例说明什么是句法范畴以及句法范畴之间的层级关系。
- (3) 除了本章中介绍的标准弧转移系统，还有什么转移系统可以应用于依存句法分析？试说明该种转移系统的优缺点。
- (4) 基于转移的依存句法分析相较于基于图的依存句法分析有什么优缺点？
- (5) 通常情况下对中文句子进行句法分析时，需要首先进行分词，并在此基础上对词语进行词性标注，然后进行句法分析。这种流水线方法有什么优点和缺点？如何设计一种方法可以同时进行中文分词、词性标注和句法分析？

(6) 试比较几种开源句法分析器在常见数据集合上的性能。

## 4. 语义分析

掌握一种语言意味着懂得如何产生并理解数量无限的该种语言句子的意义。研究语言意义的科学被称为语义学（Semantics）。语义问题也被大多数语言学家认为是语言的核心问题，同时也受到了包括哲学、逻辑学、心理学以及计算机等众多学科的广泛关注。自然语言处理目标就是要使计算机具有理解和运用自然语言的能力。因此，语义也是自然语言处理的关键问题和难点问题。从计算角度，语义研究需要以语义的形式化结构表示为基础。这种形式化结构表示称之为语义表示（Semantic Representation）。自然语言处理中语义分析（Semantic Analysis）则是指解释各粒度的语言单位，并将其转换为对应语义表示。

本章首先介绍语义学和语义表示的基本概念和主要研究内容，在此基础上语义和知识的表示方法，词义消歧算法以及语义角色标注算法。

### 4.1 语义学概述

什么是“意义”是一个困扰了哲学家和语言学家数千年的问题。我们可以非常容易地理解中文，并且用汉字组成对其他人来说也是有意义的句子。我们也可以知道某个词语、句子是否有意义，还可以通过一个句子衍推出另外一个句子。意义从何而来？语言的意义的本质又是什么？学术界对这些问题众说纷纭没有定论。中国古代以“字”为核心的训诂语义研究达到了很高的水准，公元前2世纪就有了专门解释词义的专著《尔雅》。先秦时期，荀子和墨子就开始对“名”与“实”的关系进行讨论。古希腊哲学家苏格拉底、亚里士多德等也都在其哲学著作中探讨过语言的意义。

语义学的研究目标就是发现和阐述关于意义的知识。1883年由法国语言学家 Michel Bréal 发表的论文中首次提出了语义学的概念，1897年出版了《语义学探索》对语义学的研究对象和可能采取的研究方法进行了系统地阐述，从此语义学逐渐成为语言学中一门独立的学科。语义学的研究已经成为语言学、逻辑学、哲学、心理学、认知科学、人工智能等多门学科的研究热点和难点。也因此语义学研究十分庞杂，理论和流派层出不穷，不同学科关于语义学的研究范畴、关注角度、重点问题都有很大差异。本章侧重从语言学和自然语言处理角度，对语义问题的基本概念和任务进行介绍。

从语言表达层面划分，语义学的研究大致可以分为三个层面：(1) 词汇语义学 (Lexical Semantics)

主要包括词义问题、词汇间关系、词汇场、成语的语义等；(2) 句子语义学 (Sentential Semantics) 主要以真值条件语义理论、配价理论、生成理论等为基础研究句义关系以及语序等问题；(3) 话语语义学 (Discourse Semantics) 主要研究句子以上层次结构的意义，包括话语衔接、话语连贯、语用过程解释等。本节中针对自然语言处理领域关注较多的词汇语义学和句子语义学基本语言学理论进行介绍。

### 4.1.1 词汇语义学

词是语言中能够独立运用的最小的单位，也是音、形、义的结合体。词语通过搭配组合，可以构建出短语、句子、篇章等复杂的语言结构。语义学自创建之初，就将词汇语义作为重要的研究目标。词汇语义学主要研究单个词语的意义以及词汇之间的相互关系。

#### 1. 词汇语义理论

词义 (Word Meaning) 有很多的方面，可以从不同的角度分析和定义，因而出现了包括语义场理论、语义成分分析、并置理论、框架语义理论等众多词汇语义理论。

语义场理论 (Semantic Field) 也称作词义场理论 (Lexical Field) 认为语言中词汇的意义是相互联系的，构成一个完整的系统和网络，具有某些相同语义特征的一组词聚而成场。如表示苹果、香蕉、橘子等都有一个共同的义素 [+ 水果]，组成了语义场中的“水果场” (Fruit Field)。水果、肉、蔬菜、谷物等又可以构成食物语义场，如表4.1所示。根据语义场理论，不能够孤立的研究一个词的词义，只有通过分析比较词与词之间的关系，才能确定一个词的真正意义。除了词汇的上下位关系外，同义关系、反义关系等都构成语义场。因此，语义场也可以认为是研究词与词之间的聚合关系 (Paradigmatic Relation)。

表 4.1 食物语义场示例

食物	水果	苹果
		香蕉
		...
	肉	牛肉
		羊肉
		...
	蔬菜	白菜
		菠菜
		...
	谷物	大米
		小麦
		...

语义成分分析 (Componential Analysis) 理论认为词义可以由最小的语义成分组合而成。这种最小的语义成分又被成为语义特征。

例如：可以定义 ADULT、YOUNG、MALE、FEMALE 为语义特征，根据这些特征可以表达词汇的意义：

man: ADULT + MALE

woman: ADULT + FEMALE

boy: YOUNG + MALE

girl: YOUNG + FEMALE

词的语义特征可以从语法-语义特征、内在语义特征以及感受性语义特征等三个部分进行考察。语法-语义特征主要指明语法标注，如人称、性、数、语态等；内在语义特征指直接反映的客观事物本质的语义特征；感受性语义特征指带有主观色彩和表示内涵的语义特征<sup>[163]</sup>。

义元理论 (Theory of Lexcial Primitives) 的核心思想是自然语言中包含非常少部分的词语，这些词语可以用于解释绝大部分词汇的意义。这些语义上不能分解的最小的意义组成单元称为义元 (Lexcial Primitives)。例如：man 和 fish 是义元，而 fishy 和 manliness 则是衍生词。可以使用义元对其他词语进行解释<sup>[164]</sup>。

例如：根据文献 [164] 中的定义，boy、girl、woman、man 使用义元解释如下：

boy: young human being that one thinks of as becoming a man.

girl: young human being that one thinks of as becoming a woman.

woman: human being that could be someone's mother.

man: human being that could cause a woman to be someone's mother.

董振东教授所创建的知网 (HowNet)<sup>[165]</sup> 也结合了义元理论，构建了包含 2540 多个义元的精细的语义描述体系，并为 237974 个汉语和英语词所代表的概念进行了标注。

例如：Hownet 中美味、难题的定义如下所示：

美味: edible| 食物:modifier=GoodTaste| 好吃

难题: problem| 问题:modifier=difficult| 难

HowNet 中义元采用中英双语的形式进行描述。上例子中“edible| 食物”、“GoodTaste| 好吃”是义元。“难题”是由核心义元“problem| 问题”以及对核心义元的附加描述义元“difficult| 难”组成。

框架语义学 (Frame Semantics) 则认为词义只能在相应的知识框架背景中才能得到理解。在意义的理解过程中，概念并不是杂乱无章的，很多概念往往具有一种同现的趋势，例如：顾客、服务员、吃饭、账单等概念都与饭店相关，是理解“饭店”的框架。此外，在意义的理解过程中，并不一定需要激活一个语义框架的全部成分，往往只需要激活部分框架。

例如：文献 [166] 中定义的“RISK”的框架是由如下成分组成：

RISK frame:

Chance (uncertainty about the future)

Harm  
 Victim (of the harm)  
 Valued Object (potentially endangered by the risk)  
 Situation (which gives rise to the risk)  
 Deed (that brings about the Situation)  
 Actor (of the Deed)  
 Gain (by the Actor in taking the risk)  
 Purpose (of the Actor in the Deed)  
 Beneficiary and motivation (for the Actor)

对于包含动词 risk 的句子，可以根据 RISK 框架对其进行理解。例如：

- (1) You've (Actor/Victim) risked your health (Valued Object) for a few cheap thrills (Gain).
- (2) She (Actor/Victim) has risked so much (Valued Object) for the sake of vanity (Motivation).

框架网络 (FrameNet)<sup>[167]</sup> 就是根据框架语义学理论，依靠语料库的支持构建的词汇语义知识库。截止 2022 年 3 月，FrameNet 针对 13685 个词元 (lexical unit) 构建了 1224 语义框架。

## 2. 词汇间的关系

词汇之间的关系 (Lexcial Relations) 是词汇语义学研究的另一个重点问题。关系类型可以分为三大类：形体关系、意义关系和实体关系<sup>[168]</sup>。形体关系 (Form Relations) 主要研究词汇的声音形体和拼写之间的关系。意义关系 (Sense Relations) 主要关注词汇意义之间的关联性、相似性、对立性等关系。实体关系 (Object Relations) 则主要研究词汇之间的客观关系。表4.2给出了词汇之间关系的主要类型和例子。

表 4.2 词汇之间的主要关系

词汇之间的关系类型		示例	
形体关系	Homonymy	同音异义关系	ring 与 wring 发音相同
		同形异义关系	bank (银行) 与 bank (河岸) 词形相同
意义关系	Synonymy	同义关系	“电脑”与“计算机”是同义词
	Antonymy	反义关系	“大”与“小”之间是反义关系
	Hyponymy	下位关系	“燕子”是“鸟”的下位词
	Hypernymy	上位关系	“动物”是“老虎”的上位词
实体关系	Meronymy	部分整体关系	“发动机”与“汽车”

同音异义关系和同形异义关系不涉及词汇的意义，表示具有相同发音或者相同形式但是意义不同的词汇。同义关系 (Synonymy) 表示两个词汇含有相同或相近的意义 (即义项)。通常认为两个词只有一个义项相同，就可以被认为是同义词。反义关系 (Antonymy) 表示两个词之间意义相对立或相反。上位关系 (Hypernymy) 和下位关系 (Hyponymy) 表示词的意义包含另外一个词，或

者词的意义包含在另外一个词中。许多下位词可以属于同一个上位词。部分整体关系（Meronymy）表示客观实体之间的组成部分和整体之间的关系。同形异义关系和同义关系也反映了词汇的形式（Word Form）和义项（Sense）之间的分离。

根据词汇间关系的研究，美国普林斯顿大学 George A. Miller 教授领导构建了 WordNet<sup>[36]</sup>，是目前最常用的英语词汇知识资源库。在其中词汇按照义项组合成同义集（Synset），每个义项表达不同的概念。名词、动词、形容词和副词各自独立的组合成网络。WordNet 的语义关系不是在单词之间建立的，而是在义项之间建立的。义项之间的关系包括：同义关系、反义关系（Inconsistency）、上下位关系、部分与整体关系、对等关系（Coordinate terms）、继承关系（Entailment）等类型。WordNet 3.1 版本包含 155327 个单词，175979 个同义集，共组成了 207016 对单词和义项对。

“bank”做为名词和动词在 WordNet 中的部分词条如下：

### Noun

1. bank (sloping land (especially the slope beside a body of water)) “they pulled the canoe up on the bank”; “he sat on the bank of the river and watched the currents”
2. depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) “he cashed a check at the bank”; “that bank holds the mortgage on my home”
3. bank (a long ridge or pile) “a huge bank of earth”
4. bank, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)

### Verb

1. bank (tip laterally) ”the pilot had to bank the aircraft”
2. bank (enclose with a bank) ”bank roads”
3. bank (do business with a bank or keep an account at a bank) ”Where do you bank in this town?”
4. deposit, bank (put into a bank account) ”She deposits her paycheck every month”

在 WordNet 中单词的每个义项都包含采用字典风格的注解和该义项的同义集，部分义项还包含用例。例如：bank<sup>4</sup> (bank 的第 4 个义项) 与 cant<sup>4</sup> 和 camber<sup>2</sup> 组成了一个同义集，对这个同义集的描述是 a slope in the turn of a road or track。

此外，在 WordNet 中名词和动词可以根据上下位关系或者部分整体关系构成层级结构。例如，bank<sup>4</sup> 的上下位关系链：

```

bank, cant, camber (a slope in the turn of a road or track)
=> slope, incline, side
=> geological formation, formation
=> object, physical object

```

=> physical entity

=> entity

在后续的发展中，WordNet 逐渐发展为支持 200 多种语言的语言知识库。

### 4.1.2 句子语义学

句子语义学主要是在句子层面对意义的研究。人们通常通过句子来表达完整语义，相较于词汇句子也复杂得多，因此非常多的工作都是围绕句子语义学从各个角度开展，包括语音、语法、逻辑、认知、心理学等等。本节中，从语言学角度对句子语义学的主要理论进行简要介绍。

#### 1. 句子语义理论

语言是对外部世界的编码，句子就是人们对客观世界的概念表征，人们对句子意义的认知始于真假判断。真值条件语义学（Truth-conditional Semantics）核心就是将意义定义为一个句子或句子所表达的命题为真时所必须满足的一系列条件<sup>[169]</sup>。该理论试图通过解释句子何时为真来定义给定句子或命题的意义。提出了一个检验句子真值的通用公式—T 公式：S is true iff P，S 代表某个句子。P 代表句子的真值条件，iff 表示“if and only if”。例如：他是学生，S 表示这个句子，P 表示“他”所代表的人并且真的是学生的列表。真知条件语义学开创了用数理逻辑方法解释自然语言的语义，用严格数学方法研究自然语言语义的方向。但自然语言中很多句子并不能判断真假，如疑问句和所使句等，这也在一定程度上限制了真知条件语义学的应用。

在词汇语义理论中语义成分分析理论认为词义可以由最小的语义成分组合而成，在句子层面同样也存在语义成分，这种语义成分通常称作语义格（Semantic Case）。格语法（Case Grammar）以及从格语法发展而来的框架语义学（Frame Semantics）都是以语义格为基础。语义格也称语义角色（Semantic Roles），又称语义关系、主题关系（Thematic Relations）。美国语言学家 Charles J. Fillmore 对 Noam Chomsky 的转换语法进行了延伸，提出了格语法，认为句子中名词短语总是与动词相关，并且以唯一可以识别的方式表示了名词短语的语义格。他指出“主语”、“宾语”等语法关系实际上都是表层结构上的概念，语言的底层是用“施事”、“受事”、“工具”等概念所表示的句法语义关系。

例如：The key opened the door.

The boy opened the door with a key.

上述例子中的“key”在深层句法语义上始终是“工具”，但是它可以是主语，也可以是介词 with 的宾语。常见的语义格如表4.3所示。语义格的数量以及定义并没有定论，甚至在菲尔墨的不同文章中也有不同。

在格语法中对于词库中词汇的每个词条需要标明其格特征，对于名词标明其可以作为的语义格（例如：“街道”需要标明 [+LOCATION]），对于动词需要标明其对应的格框架。格框架由主要概念和辅助概念组成，主要概念通常为动词，辅助概念为施事格、受事格、方位格、工具格等语义深层格。

表 4.3 常见语义格定义

名称	定义	示例
施事格 AGENT	动作的发起者	<b>He</b> wrote the book.
受事格 PATIENT	受到动作或状态影响的实体	He wrote <b>the book</b> .
工具格 INSTRUMENT	对动作或状态而言作为某种因素而牵涉到的无生命的客体	He cleaned the table with <b>an antiseptic wipe</b> .
方位格 LOCATION	动作或状态的处所或空间位置	They are in the <b>building</b>
来源格 SOURCE	动作所作用到的事物的来源或发生位置变化过程中的起始位置	He bought a book from <b>Jeremy</b> .
目标格 GOAL	动作所作用到的事物的终点或发生位置变化过程中的终端位置	He sold a book to <b>Jeremy</b> .
时间格 TIME	动作或状态的时间	<b>In the winter</b> , he saw snow.
伴随格 COMITATIVE	与施事共同完成动作的伴随者	He sang a song with <b>Jeremy</b> .
受益者格 BENEFICIARY	因动作的实行而受益的实体	He sang a song for <b>Jeremy</b> .
感受格 EXPERIENCER	知道谓语所描述的动作或状态，但不受该动作或状态控制	<b>He</b> saw snow.

例如：BREAK 可以放入如下格框架：

[(施事格)(受事格)(工具格)(方位格)]

格框架可以帮助解决之前语法中隐藏的某些歧义，可以提取意义相同但是结构不同的句子。

例如：他在房间里用锤子打破了玻璃杯。

根据 BREAK 框架得到：

[BREAK [ Case-frame:

  [AGENT: 他

  [PATIENT: 玻璃杯

  [INSTRUMENT: 锤子

  [LOCATION: 房间]]]

## 2. 句义关系

句子之间也存在各种语义关系，把句子当做一个整体，句子和句子之间的语义关系可以包含同义、反义、蕴含等。

同义关系（Synonym）表示两个不同的句子表达相同的意义。

- 例如：a. 他打碎了玻璃杯。  
b. 玻璃杯被他打碎了。

上述两个句子表达了相同的含义，具有相同的真值。

反义关系（Inconsistency）表示两个句子的意义只能有一个与客观事实相符。

- 例如：a. 他打碎了玻璃杯。  
b. 玻璃杯完好的放在橱窗里。

上述两个句子表达的含义，一句为真时，另外一句一定为假。

蕴含关系（Entailment）表示两个句子的意义，前者为真时后者必然为真，前者为假时后者可能为真也可能为假。

- 例如：a. 他拿着一本书去了校门口。  
b. 书在他手里。

上述例子中，句子 a 为真，那么句子 b 也必然为真，我们可以说 a 蕴含 b。

预设关系（Presupposition）表示一个句子的意义是另外一个句子的前提。

- 例如：a. 复旦大学江湾校区管委会举办了迎新活动。  
b. 复旦大学有多个校区。

上述例子中，句子 a 为真，那么句子 b 也必然为真，如果 a 句为假，b 句仍然为真。预设关系通常认为是语用关系，两者有前后的时间关系，属于历时关系。

## 4.2 语义表示

语义表示（Semantic Representation）是语义的符号化和形式化的过程，主要研究语义表示的通用原则和方法。为了使得计算机能够处理自然语言的语义，就需要用恰当的模式对语义进行表示，因此语义表示方法也是自然语言理解的基础。从语义学基础理论介绍，可以看到目前关于意义的定义和本质还没有定论，大量的语义学理论从不同的角度开展了一系列的讨论。已有的语义表示方法大多都是根据不同的语义学理论针对某项具体研究所提出的，有一定的针对性和局限性，适用于词汇、句子、篇章等各个层面各种应用的通用语义表示方法还是一个亟待解决的问题。本节中介绍常见的一阶谓词逻辑、框架、语义网等语义表示方法，分布式表示方法在下节中单独介绍。

### 4.2.1 谓词逻辑表示法

数理逻辑（Mathematical Logic）在知识的形式化表示和机器的自动定理证明方面都有广泛的应用和很好的表现，真值条件语言学中也是使用数理逻辑来研究自然语言的语义。自然语言的语义表示中也经常采用数理逻辑的方法。其中常用的是谓词逻辑（Predicate Logic）和命题逻辑（Propositional Logic）。谓词逻辑可以更细致的刻画语义，可以表示事物的状态、属性、概念等事物性语义，也可以表示因果关系等规则性语义，同时命题逻辑也可以认为是谓词逻辑的一种特例，因此本节中重点介绍谓词逻辑。

在谓词逻辑中，研究对象全体所构成的非空集合称为论域（个体域）。论域中的元素称为个体或个体词。论域中包含的个体数量可以是无限的也可以是有限的。个体可以是常量、变量或函数。个体常量表示具体的或特定的个体，个体变量表示抽象的或泛指的个体，个体函数表示一个个体到另一个个体的映射。用于刻画个体的性质、状态或个体之间关系的词项则称为谓词。这些常量、变量、函数和谓词也都需要有明确的语义解释。

谓词一般用  $P(x_1, x_2, \dots, x_n)$  表示， $P$  是谓词名， $x_1, x_2, \dots, x_n$  表示某个独立存在的事物或某个抽象的概念。如果谓词  $P$  中的所有个体都是常量、变量或函数，则称该谓词为一阶谓词（First Order Predicate Logic）。如果某个个体本身又是一个一阶谓词，则称  $P$  为二阶谓词。

例如：

谓词：Teacher( $x$ ) 表示  $x$  是教师，是一阶谓词。

句子：“老张是一名老师”可以表示为 Teacher(老张)

除了直接使用单个谓词和指代对象的常量、变量或者函数组成原子公式之外，还可以使用 5 种逻辑连接词和量词构造复杂的表示，就是谓词逻辑中的公式。原子公式是谓词演算的基本组块，运用连接词可以组合多个原子公式，以构成更加复杂的公式。具体连接词和量词定义如下：

#### (1) 连接词

- $\neg$ : “否定”（Negation）或“非”
- $\vee$ : “析取”（Disjunction）或“或”
- $\wedge$ : “合取”（Conjunction）或“与”
- $\rightarrow$ : “蕴含”（Implication）或“条件”
- $\leftrightarrow$ : “等价”（Equivalence）或“双向蕴含”

连接词的真值表如表4.4所示。连接词的优先级从高到底排列为： $\neg$ 、 $\wedge$ 、 $\vee$ 、 $\rightarrow$ 、 $\leftrightarrow$ 。

#### (2) 量词

$\forall$ : 全称量词（Universal Quantifier），表示对个体域中的所有（或任意一个）个体  $x$

$\exists$ : 存在量词（Existential Quantifier），表示在个体域中存在个体  $x$

可以利用上述谓词和逻辑公式的定义对如下句子的语义进行符号化表示：

a. “有机器人都是红色的”

谓词定义：ROBOT( $X$ ) 表示  $X$  是机器人；COLOR( $X, Y$ ) 表示  $X$  的颜色为  $Y$

表 4.4 连接词真值表

P	Q	$\neg P$	$P \vee Q$	$P \wedge Q$	$P \rightarrow Q$	$P \leftrightarrow Q$
T	T	F	T	T	T	T
T	F	F	T	F	F	F
F	T	T	T	F	T	F
F	F	T	F	F	T	T

谓词公式:  $(\exists X)[\text{ROBOT}(X) \wedge \text{COLOR}(X, \text{RED})]$

b. “人人都爱护环境”

谓词定义:  $\text{MAN}(X)$  表示 X 人;  $\text{PROTECT}(X, Y)$  表示 X 保护 Y

谓词公式:  $(\forall X)[\text{MAN}(X) \rightarrow \text{PROTECT}(X, \text{ENVIRONMENT})]$

c. “小明不在 3 号房间”

谓词定义:  $\text{INROOM}(X, Y)$  表示 X 在 Y 中

谓词公式:  $\neg \text{INROOM}(\text{XIAOMING}, \text{ROOM3})$

由于谓词逻辑具有扎实的数学基础, 一阶谓词逻辑具有充分的表达能力和完备的逻辑推理算法, 其推理过程和结果的准确性可以得到有效保证, 因此可以精密地表达语义, 有很广泛的应用领域。但是使用一阶谓词逻辑表示语义并不简单, 通常需要如下步骤:

- (1) 定义谓词及个体: 确定每个谓词及个体的确切含义。
- (2) 变量赋值: 根据所要表达的事物或概念, 为每个谓词中的变量赋予特定的值。
- (3) 谓词公式构造: 根据所表达的语义, 用适当的连接符号和量词将各谓词连接起来。

可以看到如果使用一阶谓词逻辑清晰表达语义, 需要对大量谓词以及个体, 不仅涉及到领域内特定知识的定义, 还涉及到领域的通用知识甚至是世界知识的定义, 过程十分庞杂并且需要领域专家的协助。此外, 一阶谓词逻辑不能很好的表达非精确的语义, 以及在推理过程中很可能产生的组合保证问题也都限制了其应用范围。

### 4.2.2 框架表示法

框架 (Frame) 表示法是以框架语义理论为基础发展起来的一种语义表示方法。框架用来表示所讨论对象 (一个事物、概念或者事件) 的语义。每个框架由若干槽 (Slot) 组成, 描述框架所讨论对象的某一方面的属性。每个槽根据实际情况可以赋值一定类型的实例或若干数据, 称为槽值。每个槽还可以划分为若干侧面 (Facet), 描述相应属性的一个方面。每个侧面也可以赋值一定类型的实例或若干数据, 称为侧面值。一个框架通常包含多个不同的槽和侧面, 分别用框架名、槽名和侧面名表示。典型的框架结构如下所示:

<框架名>

<槽名 1>

<侧面名 1-1><值 1-1> ...

<侧面名 1-2> <值 1-2> ...  
...  
<槽名 2>  
<侧面名 2-1> <值 2-1> ...  
<侧面名 2-2> <值 2-2> ...  
...  
...  
<槽名 n>  
<侧面名 n-1> <值 n-1> ...  
<侧面名 n-2> <值 n-2> ...  
...

MUC-3 事件语义理解的评测集合上所定义的“恐怖袭击事件”框架<sup>[170]</sup> 如下所示：

框架名：恐怖袭击事件  
槽 1：事件发生时间  
槽 2：事件类型  
槽 3：事件种类  
槽 4：犯罪者  
    侧面 4-1：个人  
    侧面 4-2：组织  
    侧面 4-3：置信度  
槽 5：实物目标  
    侧面 5-1：名称列表  
    侧面 5-2：数量  
    侧面 5-3：类型  
槽 6：人目标  
    侧面 5-1：姓名列表  
    侧面 5-2：数量  
    侧面 5-3：类型  
槽 7：事件发生地点  
槽 8：对实物目标的影响  
槽 9：对人目标的影响

利用“恐怖袭击事件”框架，句子“在位于巴黎 11 区的巴塔克兰剧院，多名武装分子在巴黎当地时间 13 日晚劫持了正在剧院观看演出的大约 1500 名观众并与警方展开对峙。”的语义可以表示为：

事件发生时间：巴黎当地时间 13 日晚

事件类型：劫持

事件种类：恐怖袭击

人目标：

类型：观众

数量：约 1500

事件发生地点：巴黎 11 区的巴塔克兰剧院

框架表示方法可以有效的表达结构性语义，并能够将语义的内部结构关系及语义间的联系表示出来。此外，框架表示法可以将槽位设置为另一个框架，从而实现框架间的联系，构建更加复杂的框架网络，还可以实现框架之间的继承关系。但是，精确表达语义需要非常复杂详细的框架以及很多嵌套层级的框架结构。

### 4.2.3 语义网表示法

语义网络（Semantic Network）是一种用实体及其语义关系来表达知识和语义的网络图。语义网络由节点和弧组成：节点表示各种事件、事物、概念、属性、动作等，也可以是一个语义子网络；弧表示节点之间的语义关系，并且是有方向和标注的，方向表示节点间的主次关系且方向不能随意调换。图4.1给出了“大学”的语义网表示样例。

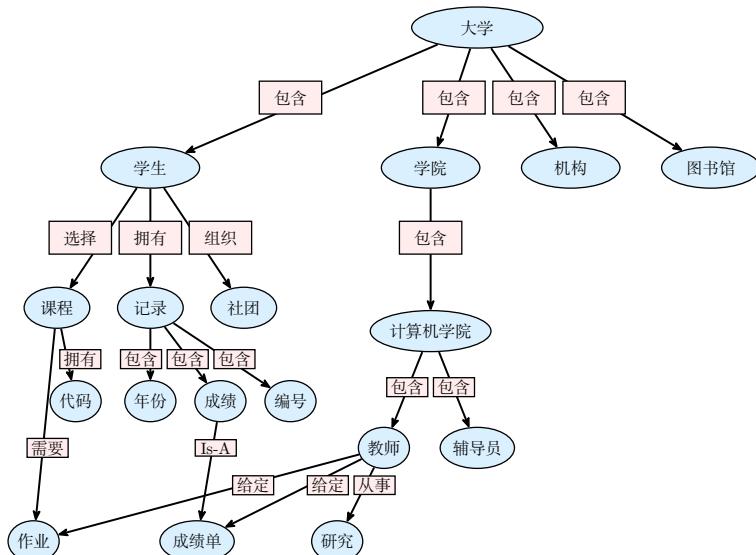


图 4.1 “大学”的语言网表示样例

语义网除了可以描述事物间包括类属关系、聚集关系、时间关系、位置关系、推论关系等多种复杂语义关系外，还可以通过增加节点的方法表示合取、析取、蕴含等语义表示中常用的连接词。例如，句子“如果明天下雨，就去看电影或者唱歌”的语义网表示如图4.2所示。

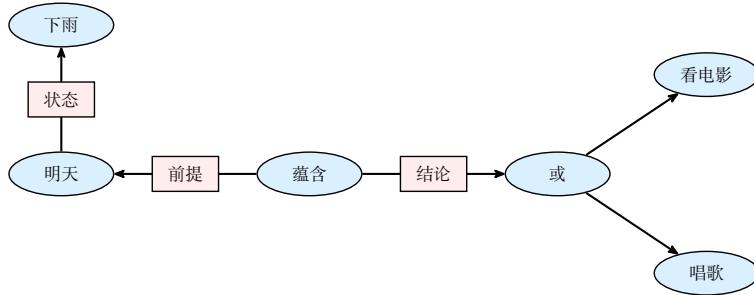


图 4.2 “如果明天下雨，就去看电影或者唱歌” 的语言网表示样例

对于比较复杂的语义还能涉及“每一个”、“有一个”等量词，使用语义网进行表示时可以通过引入分区技术进行实现<sup>[171]</sup>。其基本思想是将复杂的语义划分为若干个子语义，每个子语义采用语义网进行表示。若干个子网络合并构成更大的网络。语义网可以逐层嵌套，子网络之间也可以采用弧线进行连接。例如：“所有的学生都完成了一个课程设计”的语义网表示如图4.3所示。图中节点“s”、“r”、“p”构成了子语义网，其中“s”是全称变量、“r”和“p”是存在变量。节点“g”是表示这个子网络，由弧线“F”指向其所代表的子网络结构。

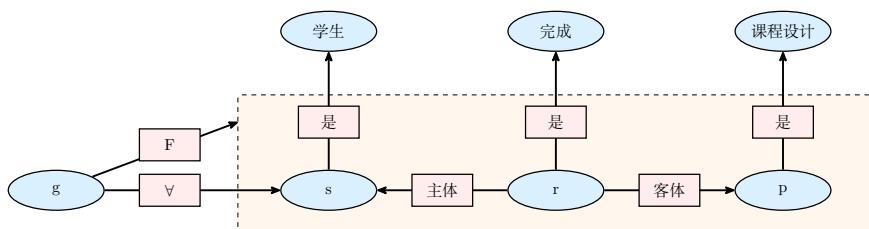


图 4.3 “所有的学生都完成了课程设计” 的语言网表示样例

语义网可以较好的把事物的属性以及事物之间的各种语义联系显式的进行表示，也可以比较容易的实现语义检索。但是，由于语义网没有公认的形式表示体系，所表达的语义需要依赖分析算法对其进行解释，表示形式的不唯一又进一步增加了其处理的复杂性。

## 4.3 分布式表示

分布式表示（Distributed Representation）旨在将文本表示为低维空间下稠密的向量，并在低维表示空间中利用表示向量之间的计算关系，体现文本间的语义关联。现代的文本表示学习，很大程度上受到 Salmon 等人在 1975 年提出的向量空间模型（Vector Space Model, VSM）<sup>[172]</sup> 的影响。这项工作结合信息检索领域的具体应用，阐述了将单词和篇章表示为向量的思想。一方面，对文本的处理可以直观地映射到向量空间，体现为对文本向量的加法、减法、距离度量等操作；另一方面，可以将量化的文本作为输入，从而直接将统计学习与机器学习算法应用在自然语言处理应用上。结合语言模型的相关理论，量化文本的思想在统计学习时期就已在自然语言处理系统中广泛应用。

文本分布式表示进一步提升了量化文本的实用性，使文本表示模块成为自然语言处理系统必不可少的一部分。在分布式表示提出之前，许多自然语言处理算法采用独热表示（One-hot Representation），其中每个维度表示某个单词是否在该文中出现。独热表示的维度和词表的大小一致，存在表示稀疏性的问题，而且无法表示单词之间的语义相似度。分布式表示通过将文本表示为低维空间下稠密的向量，有效地解决了这一问题。当应用在下游任务时，文本分布式表示也体现出良好的泛化能力，而且能有效地编码任务所需要的语法和语义信息<sup>[173, 174]</sup>。因此，文本分布式表示作为模型的基本输入，已经被广泛应用于自然语言处理领域的各种任务。

早期的分布式表示方法聚焦在词汇的表示向量构建上。随着自然语言处理技术应用领域的拓展，为了适配多样化的任务需求，句子、篇章级别的分布式表示方法也逐渐广泛应用。本节针对不同级别的语言粒度，分别介绍单词、句子和篇章级别分布式表示方法。

### 4.3.1 单词分布式表示

单词分布式表示（Word Distributed Representation）通过将单词表示为定长低维稠密向量，在向量空间建构单词之间的语义关系。形式上，单词分布式表示的目标是建立单词嵌入矩阵  $\mathbf{W} \in \mathbb{R}^{|V|*d}$ ，其中矩阵的每一行对应一个单词，为单词的向量表示，即词向量。

例如：“计算机”表示为  $[0.16, 0.19, -0.28, \dots, 0.87]$

“电脑” 表示为  $[0.20, 0.17, -0.21, \dots, 0.97]$

“冰激凌” 表示为  $[-0.90, 0.72, 0.65, \dots, 0.06]$

相比于独热表示，分布式表示可以编码不同单词之间的语义关联。如上例中，如果采用独热表示，“计算机”与“电脑”以及“计算机”与“冰激凌”之间的相似度都相同。但是采用分布式表示可以使得“计算机”和“电脑”在大多数维度上相近，这样“计算机”和“电脑”的向量之间的距离可以远小于“计算机”和“冰激凌”之间的距离。

根据单词分布式表示的目标，即在向量空间建构单词之间的语义关联，使含义相近的单词具有相似的向量表示。这自然地引出了两个问题：(1) 如何衡量单词语义的相近；(2) 如何衡量表示的相似。针对第一个问题，大部分单词分布式表示方法遵从分布式假设，即出现在相同上下文

中的单词往往具有相似的语义<sup>[175]</sup>。在分布式假设的基础上，这些方法侧重于还原单词之间的共现关系，即为频繁出现在相同上下文中的词语之间赋予较高的表示相似度。针对第二个问题，根据下游应用场景的不同，可以根据表示向量的余弦相似度、L2 范数距离等方式衡量表示向量的相似性。为了实现上述的目标，早期的词向量模型通过统计方法，根据文本中词级别信息的统计数据构建低维稠密词向量。目前更多的方式是利用机器学习方法，通过对大量无标注文本进行自监督学习，使用机器学习模型来学习词向量。本节将分别介绍基于统计和机器学习的词向量模型。

### 1. 基于共现矩阵奇异值分解的词向量模型

在分布式假设下，希望单词之间的相似度体现为两个词出现在相同上下文的频率。因此可以采用针对共现矩阵 (Co-occurrence Matrix) 的矩阵分解方法。隐式语义分析 (Latent Semantic Analysis, LSA) 模型<sup>[176]</sup> 采用奇异值分解方法 (Singular Value Decomposition, SVD)，将单词文档共现矩阵 (Term-Document Co-occurrence Matrix) 或单词上下文共现矩阵 (Window based Co-Occurrence Matrix) 转换为单词向量表示。这里以单词上下文共现矩阵为例，介绍基于奇异值分解的词向量模型。

共现矩阵  $A \in \mathbb{R}^{|\mathbb{V}| \times |\mathbb{V}|}$ ,  $A_{ij}$  表示词表  $\mathbb{V}$  中下标为  $i$  和  $j$  的单词出现在相同上下文中的次数。根据语料库，对于句子中的单词，将其向前、向后各  $n$  个单词的范围作为该单词的上下文范围，称为该单词大小为  $n$  的上下文窗口。对于每个句子，取其中每个单词大小为  $n$  的上下文窗口，对窗口范围内的每个单词与该单词进行共现计数。对于语料库整体进行共现次数统计，获得共现矩阵。具体过程如算法4.1所示。

---

#### 代码 4.1: 共现矩阵统计算法

---

```

输入: 训练语料库  $D$ , 上下文窗口大小  $n$ 
输出: 词表  $\mathbb{V}$ , 共现矩阵  $A$ 
 $\mathbb{V} = \text{unique}(D)$  // 由语料库统计出现的词表
 $A_{ij} = 0, \forall i, j \in [1, \dots, |\mathbb{V}|]$  // 初始化共现矩阵
foreach  $s \in D$  do
     $w_1, \dots, w_N = s$  // 取训练语料库中的句子，句子由  $N$  个词构成
     $id_i = \mathbb{V}.\text{index}(w_i), \forall i \in [1, \dots, N]$  // 将句子中的单词转化为词表下标
    for  $i = 1$  to  $N$  do
        for  $j = \max(1, i - n)$  to  $\min(N, i + n), j \neq i$  do
             $A_{id_i, id_j} = A_{id_i, id_j} + 1$  // 共现计数
        end
    end
end
return  $\mathbb{V}, A$ 

```

---

共现矩阵的每一行可以自然地当做对应词的向量表示，因为它的不同维度表示和各个单词的共现次数，共现矩阵提供的向量表示可以体现词语之间的相似度。然而，共现矩阵提供的词向量表示维度和词表大小相同，依然面临表示稀疏性的问题。奇异值分解方法将共现矩阵分解为低阶近似矩阵，从而为每个单词提供低维表示。对于矩阵  $A \in \mathbb{R}^{m \times n}$ ，通过奇异值分解可将其分解为  $A = U\Sigma V^T$ ，其中  $U \in \mathbb{R}^{m \times m}$  和  $V \in \mathbb{R}^{n \times n}$  是酉矩阵， $\Sigma \in \mathbb{R}^{m \times n}$  在主对角线之外的元素为 0。特征向量的顺序均按特征值由大及小顺序排列。矩阵  $\Sigma$  对角线上的奇异值元素均为非负实数，且按由大及小顺序排列。

一般情况下，矩阵奇异值的大小分布极不均匀，前 10% 奇异值之和可以占到全部奇异值之和的 90% 以上。因此，可以在奇异值矩阵中只保留最大的  $d$  个奇异值，同时在两侧矩阵中仅保留对应的分量，对矩阵  $A$  进行低秩近似。应用上述方法，对于共现矩阵进行数据压缩，就可以获得单词的低维稠密表示  $W \in \mathbb{R}^{|V| \times d}$ ，目的是使  $W_i W_j^T$  能够近似地还原词表中下标为  $i, j$  的单词在相同上下文中出现的频率。词表示矩阵  $W$  的计算如算法4.2所示。

---

#### 代码 4.2: 基于共现矩阵奇异值分解的词向量模型<sup>[177]</sup>

---

```

输入: 共现矩阵  $A \in \mathbb{R}^{|V| \times |V|}$ , 嵌入维度  $d$ 
输出: 词表示矩阵  $W \in \mathbb{R}^{|V| \times d}$ 

 $U, \Sigma, V = \text{SVD}(A)$  // 由  $A$  是对称方阵, 可知  $U = V$ ,  $\Sigma$  是对角方阵
 $U, \Sigma = U_{:, :d}, \Sigma_{:d, :d}$  // 特征降维, 保留最重要的  $d$  维,  $A \approx U\Sigma U^T$ 
 $diag(\sigma_1, \dots, \sigma_d) = \Sigma$  //  $\Sigma$  包含最重要的  $d$  个奇异值  $\{\sigma_i\}_{i=1}^d$ 
 $\sqrt{\Sigma} = diag(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_d})$  //  $\sqrt{\Sigma}\sqrt{\Sigma}^T = \Sigma$ 
 $W = U\sqrt{\Sigma}$  //  $WW^T = U\Sigma U^T \approx A$ 
return  $W$ 

```

---

相对于词语的独热表示，基于共现矩阵奇异值分解的词向量模型初步解决了表示稀疏性的问题，且可以在一定程度上体现词语之间的相似度。

## 2. 基于上下文单词预测词向量模型

文献[178]中提出了大幅度简化以往的神经网络语言模型(Neural Probabilistic Language Model, NPLM)的Word2vec方法，去除了非线性隐藏层，使用自监督的方式从大量无监督文本训练词表示模型。构建了两个非常简单的神经网络模型结构：连续词袋模型(Continuous Bag Of Words, CBOW)和跳字模型(Skip-Gram, SG)，用于学习单词分布式表示。其框架如图4.4和图4.5所示。它们分别基于不同的假设，以基于条件概率的方式训练词表示模型。另外，针对梯度反向传播计算量过大的问题，采用负采样(Negative Sampling)和层次Softmax(Hierarchical Softmax)两种近似训练方法。

Skip-Gram 模型的基本假设是文本中的词可以预测其上下文窗口内的其他词。如图4.4所示，在

句子“我稍后回答这个问题”中，对于中心词“回答”，当以大小为 2 的上下文窗口预测时，模型考虑上下文窗口内词语“我”，“稍后”。“这个”，“问题”在中心词为“回答”条件下的出现概率。Skip-Gram 模型通过最大化上下文词在给定中心词时的条件概率，来进行模型的训练。

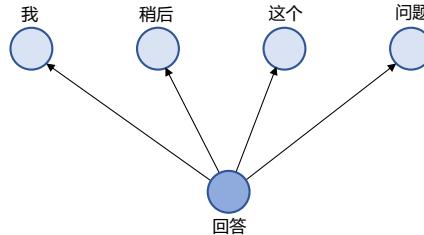


图 4.4 Skip-Gram 模型对上下文词的示例

Skip-Gram 模型以负对数概率形式的损失函数作为优化目标，形式化表示为：

$$\mathcal{L}(w_1 \dots w_T) = - \sum_{t=1}^T \sum_{-n \leq i \leq n, i \neq 0} \log P(w_{t+i} | w_t) \quad (4.1)$$

其中， $w_t$  是位置  $t$  的中心词， $w_{t+i}$  是上下文窗口内每个位置的上下文词， $T$  是文本序列的长度。 $P(w_{t+i} | w_t)$  条件概率通过通过训练上下文词和中心词的词嵌入参数矩阵，来近似估计上述条件概率。具体来说，Skip-Gram 包括  $\mathbf{U} \in \mathbb{R}^{|\mathcal{V}| \times d}$  和  $\mathbf{V} \in \mathbb{R}^{|\mathcal{V}| \times d}$  两个词嵌入矩阵，分别表示词表中每个单词作为上下文词和中心词时的词向量。Skip-Gram 模型通过上下文词和中心词向量的相似度估计上下文词的出现概率，具体公式如下所示：

$$P(w_o | w_c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^T \mathbf{v}_c)} \quad (4.2)$$

其中， $w_o$  和  $w_c$  分别表示特定的上下文词和中心词， $\mathbf{u}_o$  是  $w_o$  用作上下文词的表示， $\mathbf{v}_c$  是  $w_c$  用作中心词的表示， $\mathbf{u}_i$  是词表中每个词用作上下文词的表示。在优化上述目标函数后，Skip-Gram 模型通常采用训练好的中心词表示作为最终的词表示。

CBoW 模型则假设文本中的词可以通过其在文本中的上下文词推导出来。如图4.5所示，在句子“我稍后回答这个问题”中，当以大小为 2 的上下文窗口预测中心词“回答”时，模型仅考虑上下文词“我”，“稍后”，“这个”，“问题”，以此推断中心词的出现概率。CBoW 模型通过最大化中心词在上下文中出现的条件概率，来进行模型的训练。

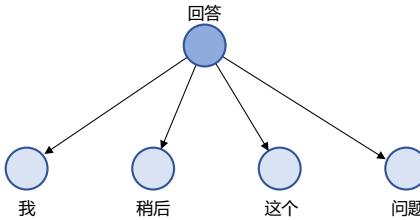


图 4.5 CBoW 模型对中心词的预测示例

CBoW 模型也是以负对数概率形式的损失函数作为优化目标:

$$\mathcal{L}(w_1 \dots w_T) = - \sum_{t=1}^T \log P(w_t | w_{t-n}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+n}) \quad (4.3)$$

CBoW 模型也通过学习上下文词向量矩阵和中心词词向量矩阵来优化目标函数。通常使用与 Skip-Gram 相反的记号, 即用  $\mathbf{U} \in \mathbb{R}^{|\mathbb{V}| \times d}$  表示中心词词向量矩阵,  $\mathbf{V} \in \mathbb{R}^{|\mathbb{V}| \times d}$  表示上下文词向量矩阵。根据上下文词生成中心词的条件概率具体计算公式如下所示:

$$\mathbf{v}_o = \frac{1}{2m} \sum_{-n \leq i \leq n, i \neq 0} \mathbf{v}_{c+i} \quad (4.4)$$

$$P(w_c | w_{c-n}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+n}) = \frac{\exp(\mathbf{u}_c^T \mathbf{v}_o)}{\sum_{i \in \mathbb{V}} \exp(\mathbf{u}_i^T \mathbf{v}_o)} \quad (4.5)$$

其中,  $\mathbf{v}_o$  是平均的上下文词向量, 用于计算和中心词的相似度;  $\mathbf{u}_c$  是  $w_c$  用作中心词的表示,  $\mathbf{u}_i$  是词表中每个词用作中心词的表示。

在实际应用中, 由于词表内通常包含数万甚至数十万单词, Skip-Gram 和 CBoW 模型在基于 Softmax 计算上下文词和中心词的出现概率进行梯度更新时, 会产生非常大规模的计算开销。因此, 通常使用负采样或者层次 Softmax 的方法降低计算开销。下面我们结合 Skip-Gram 模型介绍这上述两种方法。

如 Skip-Gram 的目标函数公式4.1所示, 优化的最终目标是最大化上下文词的条件概率  $P(w_{t+i} | w_t)$ , 即  $P(w_o | w_c)$ 。结合公式4.2来看, 这个损失函数使正确的上下文词  $w_o$  和中心词  $w_c$  之间的相似度  $\mathbf{u}_o^T \mathbf{v}_c$  尽量大, 同时对于词表中每个不在上下文窗口内的词, 使其和中心词的相似度  $\mathbf{u}_i^T \mathbf{v}_c$  尽量小。后者针对词表中每个词进行计算, 因此引入大量的计算开销。针对上述问题, 负采样 (Negative Sampling) 将目标函数中全体词表范围的相似度计算修正为目标词和  $K$  个负例的相似度计算, 其中  $K$  是远小于词表大小的超参数。通过这种方式, 使得训练的计算开销与词表大小无关, 而只与超参数  $K$  相关。

具体来说，首先将词向量的相似度和词出现在上下文窗口的概率相关联：

$$P(D = 1|w_o, w_c) = \sigma(\mathbf{u}_o^T \mathbf{v}_c) \quad (4.6)$$

$$P(D = 0|w_o, w_c) = 1 - P(D = 1|w_o, w_c) \quad (4.7)$$

其中， $D = 1$  或  $D = 0$  表示这组上下文词-中心词是否来自训练语料，即训练语料中单词  $w_o$  是否出现在  $w_c$  的上下文窗口中。 $\sigma$  表示 Sigmoid 函数， $\mathbf{u}_o$  是  $w_o$  用作上下文词的表示， $\mathbf{v}_c$  是  $w_c$  用作中心词的表示。

在优化目标方面，对于每组上下文词-中心词，原始的 Skip-Gram 模型只优化  $P(D = 1|w_o, w_c)$  的正样本，而负采样下的模型从原始词分布  $P(w)$  中采样  $K$  个未出现在上下文窗口中的词  $w_1, \dots, w_K$  作为负样本，并额外将  $P(D = 0|w_i, w_c)$  作为优化目标，形式化表示为：

$$P(w_o|w_c) = P(D = 1|w_o, w_c) \prod_{i=1}^K P(D = 0|w_i, w_c) \quad (4.8)$$

将上述公式代替公式4.2中的概率项估计，并代入公式4.1，可以得到负采样下的目标函数。

相比于负采样，层次 Softmax 则以另外一种角度为上下文词的出现概率  $P(w_o|w_c)$  建立近似训练方法。如图4.6所示，树的每个叶子节点表示词表中的一个词语。用每个分叉节点  $n$  的上下文表示向量  $\mathbf{u}_n$  建模其左、右子树上每个词语出现在上下文窗口中的概率之和。

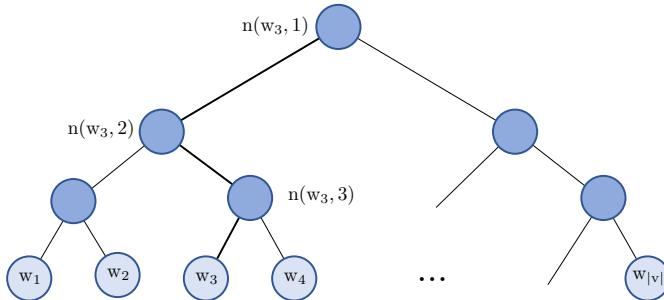


图 4.6 层次 softmax 将词表组织成二叉树结构，树的每个叶子节点代表词表中的一个词语

具体地，可以使用下述公式对此概率进行估计：

$$\sum_{w \in \text{lSub}(n)} \frac{P(w|w_c)}{\sum_{w \in n} P(w|w_c)} = \sigma(\mathbf{u}_n^T \mathbf{v}_c) \quad (4.9)$$

其中， $\text{lSub}(n)$  表示对应的叶子节点位于  $n$  的左子树下的全体单词集合。

由于  $\sigma(x) + \sigma(-x) = 1$ , 可以得到:

$$\sum_{w \in \text{rSub}(n)} \frac{P(w|w_c)}{\sum_{w \in n} P(w|w_c)} = 1 - \sigma(\mathbf{u}_n^T \mathbf{v}_c) = \sigma(-\mathbf{u}_n^T \mathbf{v}_c) \quad (4.10)$$

类似地,  $\text{rSub}(n)$  表示对应的叶子节点位于  $n$  的右子树下的全体单词集合。

在公式4.9和4.10的基础上, 对于词表中的词  $w$ , 用  $L(w)$  表示从树的根节点到词  $w$  对应的叶子节点的路径, 其中包括从根节点到叶子节点的父节点的全部非终节点, 但不包括叶子节点本身。记  $n(w, i)$  是路径  $L(w)$  上的第  $i$  个节点, 其上下文表示向量记为  $\mathbf{u}_i$ , 则词  $w$  出现在  $w_c$  的上下文窗口中的概率估计为:

$$P(w|w_c) = \prod_{i=1}^{|L(w)|} \sigma(\mathbf{u}_i^T \mathbf{v}_c) \quad (4.11)$$

可以发现, 对于任意的中心词  $w_c$ , 层次 Softmax 保证词表中所有词出现在中心词上下文窗口特定位置的条件概率之和为 1, 即  $\sum_{w \in \mathbb{V}} P(w|w_c) = 1$ 。基于层次 softmax 计算词表示模型, 计算开销和  $|L(w)|$  的平均值呈线性关系。当使用满二叉树结构容纳词表时,  $|L(w)|$  具有  $\mathcal{O}(\log_2 |\mathbb{V}|)$  的上界, 所以层次 Softmax 也可以显著降低 Skip-Gram 模型的计算复杂度。

### 3. 全局向量 (GloVe) 模型

Skip-Gram 和 CBOW 模型利用每个单词的上下文窗口信息作为监督信号, 自监督地对语料库进行学习, 而 LSA 模型则基于词共现矩阵通过矩阵分解得到, 全局统计信息和局部信息都对词表示学习提供有效信息。全局向量 (Global Vectors for Word Representation, GloVe) 模型<sup>[179]</sup> 则结合了上述模型的思想, 从共现概率的角度分析并改进了 Skip-Gram 模型, 即使用文本中局部的上下文信息, 又对语料库的全局共现统计数据加以利用。

GloVe 模型基于上下文窗口共现矩阵的统计, 即对语料库中特定中心词-上下文词对的出现次数的统计。在算法4.1所述的共现计数方法基础上, GloVe 模型中的共现矩阵进一步地考虑了中心词和上下文词之间的距离, 使相距更近的中心词-上下文词对于共现次数起到更大的贡献。记  $w_i, w_j$  为词表中下标为  $i, j$  的单词, 在它们的每次共现中, 记  $d(w_i, w_j)$  为单词  $w_i, w_j$  之间的距离。GloVe 模型中的共现矩阵将词与词之间的共现次数按共现距离的倒数进行加权, 即  $C_{ij} = \sum d^{-1}(w_i, w_j)$ 。由共现矩阵可以得到单词  $w_j$  出现在单词  $w_i$  上下文的共现概率为  $p_{ij} = P(w_j|w_i) = \frac{C_{ij}}{\sum_{j=1}^{|\mathbb{V}|} C_{ij}}$ 。

GloVe 模型的损失函数形式与上节介绍的 Skip-Gram 模型相似, 同样以还原共现频率  $p_{ij}$  为目标, 并在其基础上进行改进。记  $\hat{p}_{ij} = P(w_j|w_i)$  为  $w_j$  出现在  $w_i$  上下文范围内的预测概率, 即  $p_{ij}$  的预测值。在共现矩阵统计的视角下, Skip-Gram 模型使用统计的共现频率作为监督信号, 通过估计共现概率优化词向量。据此, Skip-Gram 模型的损失函数可以重新表示如下:

$$\mathcal{L} = - \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} C_{ij} \log \hat{p}_{ij} \propto - \sum_{i=1}^{|V|} x_i \sum_{j=1}^{|V|} p_{ij} \log \hat{p}_{ij} \quad (4.12)$$

其中

$$\hat{p}_{ij} = \frac{\exp(\mathbf{u}_j^T \mathbf{v}_i)}{\sum_{j=1}^{|V|} \exp(\mathbf{u}_j^T \mathbf{v}_i)} \quad (4.13)$$

$x_i$  是训练语料中  $w_i$  作为中心词出现的次数。

通过公式4.12，可以将 Skip-Gram 模型理解为使用交叉熵损失，优化词共现条件概率分布的过程。GloVe 模型对 Skip-Gram 模型进行了如下改进。首先，GloVe 模型用平方损失代替 Skip-Gram 模型中的交叉熵损失，并使用变量  $p'_{ij} = C_{ij}$  和  $\hat{p}'_{ij} = \exp(\mathbf{u}_j^T \mathbf{v}_i)$  代替原来的概率分布。另外，为每个单词  $w_i$  引入可训练的中心词偏置项  $b_i$  以及上下文词偏置项  $c_j$  对训练目标进行校正。在对数形式下，平方损失项如下所示：

$$(\log p'_{ij} - \log \hat{p}'_{ij})^2 = (\mathbf{u}_j^T \mathbf{v}_i - \log C_{ij} + b_i + c_j)^2 \quad (4.14)$$

此外，GloVe 模型使用  $h_{ij} = h(C_{ij})$  作为每个损失项的权重，建模单词  $w_i$  与  $w_j$  的相关度。对于  $h_{ij}$ ，目标是为共现频率较高的词赋予较高的权重，所以  $h$  是  $C_{ij}$  的非递减函数。此外，不希望这个权重随着共现频率无限地增大，因此当函数值到达界限之后不应继续增加。另外，对于从未共现的单词  $w_i$  与  $w_j$ ，应有  $h_{ij} = 0$ ，表示两者之间不存在关联。为达到上述目标，使用如下分段函数来建模每个损失项的权重：

$$h(c) = \begin{cases} (c/c_{max})^\alpha, & 0 \leq c < c_{max} \\ 1, & c \geq c_{max} \end{cases} \quad (4.15)$$

其中， $c_{max}$  和  $\alpha$  是预设的超参数。

综合上述各项，GloVe 模型最终的损失函数如下：

$$\mathcal{L} = - \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} h_{ij} (\mathbf{u}_j^T \mathbf{v}_i - \log C_{ij} + b_i + c_j)^2 \quad (4.16)$$

在训练过程中，为了提高训练效率，可以省略任意  $h_{ij} = 0$  的损失项，例如在每步优化时随机抽取一批次  $C_{ij} > 0$  的词对进行梯度更新。

考虑到共现矩阵为对称矩阵，即  $C_{ij} = C_{ji}$ ，在 GloVe 模型中优化得到的中心词向量和上下文词向量理论上是相等的。在实际应用中，由于权重的随机初始化不同，同一个词最终得到的中心词向量和上下文词向量可能不相等。GloVe 通常使用两个向量的和作为输出的词向量。

#### 4. 基于字节对编码的子词表示模型

本章前几节所介绍的词表示模型都依赖预先确定的词表  $V$ ，在编码输入词序列时，这些词表示模型只能处理词表中存在的词。因此，在使用中，如果遇到不在词表中的未登录词，模型无法为其生成对应的表示，只能给予这些未登录词一个默认的通用表示。通常的处理方式是，词表示模型会预先在词表中加入一个默认的“[UNK]”（unknown）标识，表示未知词，并在训练的过程中将 [UNK] 的向量作为词表示矩阵的一部分一起训练，通过引入某些相应机制来更新 [UNK] 向量的参数。在使用时，对于全部的未登录词，都使用 [UNK] 的向量作为这些词的表示向量。此外，基于固定词表的词表示模型对词表大小的选择比较敏感。当词表大小过小时，未登录词的比例较高，影响模型性能。而当词表大小过大时，大量低频词出现在词表中，而这些词的词向量很难得到充分学习。理想模式下，词表示模型应能覆盖绝大部分的输入词，并避免词表过大所造成的数据稀疏问题。

为了缓解未登录词问题，一些工作通过利用亚词级别的信息构造词表示向量。一种直接的解决思路是为输入建立字符级别表示，并通过字符向量的组合来获得每个单词的表示，以解决数据稀疏问题。然而，单词中的词根、词缀等构词模式往往跨越多个字符，基于字符表示的方法很难学习跨度较大的模式。为了充分学习这些构词模式，子词表示模型提出了子词（Subword）的概念，试图缓解上文介绍的未登录词问题。子词表示模型会维护一个子词词表，其中既存在完整的单词，也存在形如“c”，“re”，“ing”等单词部分信息，称为子词。子词表示模型对词表中的每个子词计算一个定长向量表示，供下游模型使用。对于输入的词序列，子词表示模型将每个词拆分为词表内的子词。例如，将单词“reborn”拆分为“re”和“born”。模型随后查询每个子词的表示，将输入重新组成为子词表示序列。当下游模型需要计算一个单词或词组的表示时，可以将对应范围内的子词表示合成为需要的表示。因此，子词表示模型能够较好地解决自然语言处理系统中未登录词的问题。

字节对编码模型（Byte Pair Encoding, BPE）<sup>[180]</sup>是一种常见的子词表示模型。该模型所采用的词表包含最常见的单词以及高频出现的子词。在使用中，常见词通常本身位于 BPE 词表中，而罕见词通常能被分解为若干个包含在 BPE 词表中的子词，从而大幅度降低未登录词的比例。BPE 算法包括两个部分：(1) 子词词表的确定；(2) 全词切分为子词以及子词合并为全词的方法。

BPE 子词词表的计算过程如图4.7所示。首先确定语料库中全词的词表和词频，然后将每个单词切分为单个字符的序列，并在序列最后添加符号“</w>”作为单词结尾的标识。比如单词“low”被切分为序列“l\_o\_w\_</w>”。所切分出的序列元素称为字节，即每个单词都切分为字节的序列。之后，按照每个字节序列的相邻字节对和单词的词频，统计每个相邻字节对的出现频率，合并出现频率最高的字节对，将其作为新的子词加入词表，并将全部单词中的该字节对合并为新的单一字节。如图4.7所示，在第一次迭代时，出现频率最高的字节对是 (e,s)，故将“es”作为子词加入词表，并将全部序列中相邻的 (e,s) 字节对合并为 es 字节。重复这一步骤，直至 BPE 子词词表的大小达到指定的预设值，或没有可合并的字节对为止。

在子词词表确定之后，对于输入词序列中未在词表中的全词进行切分，BPE 算法对词表中的

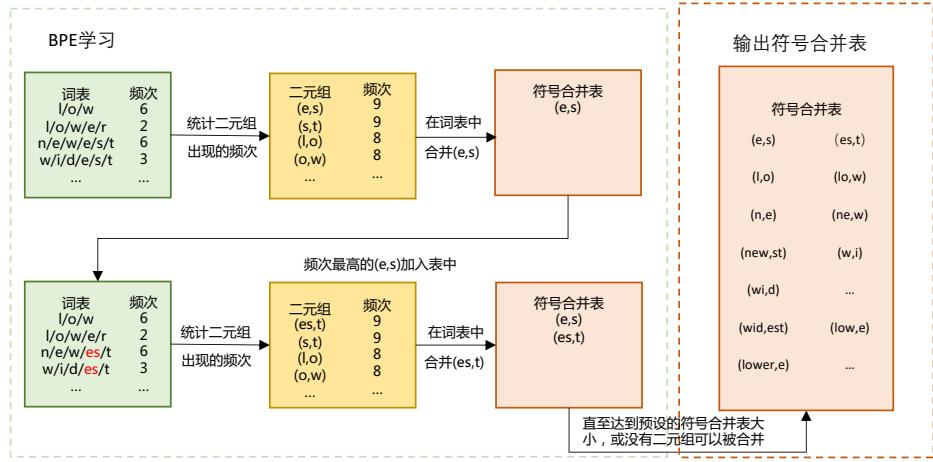


图 4.7 BPE 模型中子词词表的计算过程

子词按从长到短的顺序进行遍历，用每一个子词和当前序列中的全词或未完全切分为子词的部分进行匹配，将其切分为该子词和剩余部分的序列。如图4.8所示，对于全词“lowest</w>”，首先通过匹配子词“est</w>”将其切分为“low”，“est</w>”的序列，再通过匹配子词“low”，确定其最终切分结果为“low”，“est</w>”的序列。通过这样的过程，BPE 尽量将词序列中的词切分成已知的子词。



样例1: lower<e> → lower<e> → lower<e> → lower<e> → lower<e>

样例2: lowest<e> → lowest<e> → lowest<e> → lowest<e> → lowest<e> → lowest<e>

(b) 合并样例

图 4.8 BPE 模型子词切分和合并示例

在遍历子词词表后，对于切分得到的子词序列，为每个子词查询子词表示，构成子词表示序列。若出现未登录子词，即未出现在 BPE 词表中的子词，则采取和未登录词类似的方式，为其赋予相同的表示，最终获得输入的子词表示序列。

对于使用了子词表示模型的自然语言处理系统，比如机器翻译系统，其输出序列也是子词序列。如图4.8所示，对于原始输出，根据终结符 </w> 的位置确定每个单词的范围，合并范围内的

子词，将输出重新组合为词序列，作为最终的结果。

### 5. 单词分布式表示的评价与应用

单词分布式表示模型的定量评估方法主要分为内部（Intrinsic Evaluation）和外部（Extrinsic Evaluation）方法。内部评估方法通常基于一个特殊设计的辅助任务，这个辅助任务探测词向量应该具有的某种性质，如词义相关性、类比性等，并最终返回一个分数，来表示词向量的好坏，从而帮助我们理解词向量模型的特点。外部评估方法通常基于一个实际应用任务，通过将词向量作为该任务的输入表示，比较不同词向量模型在该任务上的性能，来选择适合于该任务的词向量模型。在评价词向量模型的综合性能时，通常会使用内部评估方法。一方面，内部评估方法所使用的辅助任务也比一般的应用任务简单，而且计算速度快。另一方面，外部评估方法除了词向量模型外，还会涉及下游任务的模型。当系统整体表现达不到预期时，问题可能来自于词向量模型，也可能来自于任务模型或两者之间的交互，因此不能很好地指导词向量模型的选择与改进。本节介绍两种常用的内部评估方法，即词义相关性和词语类比性。

词义相关性任务通过探索词向量对词义相关性的表达能力，来评价词向量的质量。理想状态下，词向量应该稠密、连续地分布在低维语义空间上，所以应该存在一种形式简洁、易于计算的相似度度量，使得词向量之间的相似度可以反映词语之间的词义相关性。一般地，对于单词  $w_i, w_j$  及其词向量  $\mathbf{v}_i, \mathbf{v}_j$ ，简单地使用余弦相似度作为词义相似性的度量：

$$\text{sim}(w_i, w_j) = \cos(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \quad (4.17)$$

通过直接将词义相似度作为目标，可以定量衡量词向量模型的性能。常用的英文词义相似度评测基准包括 WordSim-353<sup>[181]</sup> 和 SimLex-999<sup>[182]</sup>，他们分别包含 353 和 999 个英文词对，每个词对包括一个位于 [0, 10] 区间内的相似度分数，如表 4.5 所示。针对词向量模型的评测，由词向量计算的语义相似度与标注值之间的 Spearman 或 Pearson 相关系数可以作为词向量模型的评价。

表 4.5 语义相似度评测基准样本示例

单词 1	单词 2	相似度
dirty	narrow	0.30
student	pupil	9.40
win	dominate	5.68
smart	dumb	0.60
attention	awareness	8.73
leave	enter	1.38

在应用中，对于给定的单词  $w$ ，可以通过词向量的余弦相似度，在词表中检索意义最为接近的词语，即  $w^* = \arg \max_{w' \in \mathbb{V}} \text{sim}(w, w')$ 。另外，词向量还可以应用于类比性任务。例如，在由 (man,

woman) 词对确定的类比关系下, 可以为单词 son 检索类比词 daughter, 它们满足 man 之于 woman, 相当于 son 之于 daughter 的类比关系。一般地, 对于形如“ $w_a$  之于  $w_b$ , 相当于  $w_c$  之于  $?$ ”的问题, 可以通过以下公式检索最恰当的类比词:

$$w^* = \arg \min_{w' \in \mathbb{V}} \cos(\mathbf{v}^*, \mathbf{v}_c + \mathbf{v}_b - \mathbf{v}_a) \quad (4.18)$$

其中  $\mathbf{v}^*, \mathbf{v}_a, \mathbf{v}_b, \mathbf{v}_c$  分别是  $w^*, w_a, w_b, w_c$  对应的词向量。

### 4.3.2 句子分布式表示

句子分布式表示主要用于句子级别的任务, 如情感分析、文本推理、语义匹配等。对于句子级别表示的构建, 不但要考虑句子中所包含单词的语义, 也要考虑句子内部词之间的关系, 即词的共现信息和句子语义之间的联系。还要考虑句子和句子之间隐含的语义相似性, 以及其他语义关系。这些性质对于句子级别的下游应用任务都很重要。本节将介绍两种句子级分布式表示算法 Skip-Thought 和 Sent2Vec 模型。

#### 1. Skip-Thought 句子表示模型

Skip-Thought 模型<sup>[183]</sup> 的目的主要是建模句子与句子之间的上下文语义关系, 从而构建句子表示模型。Skip-Thought 模型借鉴了 Skip-Gram 模型的思想, 认为可以基于一个句子预测出其上下文的句子, 并以此作为监督信号, 学习句子之间的语义关系, 得到句子表示模型。Skip-Thought 模型结构如图 4.9 所示, 包括一个编码器和两个结构相同的解码器, 输入当前位置的句子  $s^i = w_1^i, \dots, w_N^i$ , 编码器将输入句转化为向量表示, 而两个解码器分别预测该句在上下文中的前一个句子  $s^{i-1}$  和后一个句子  $s^{i+1}$ 。

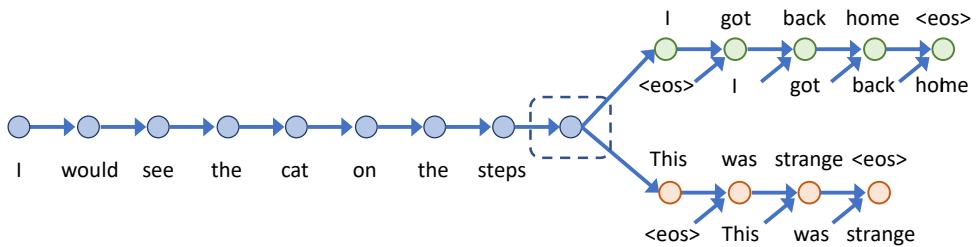


图 4.9 Skip-Thought 模型结构图<sup>[183]</sup>

在编码器方面, Skip-Thought 模型使用一个 GRU 网络编码输入  $s^i$ , 在每个时刻  $t$  生成表示  $\mathbf{h}_t^i$ :

$$\mathbf{h}_1^i, \dots, \mathbf{h}_N^i = \text{Encoder}(\mathbf{x}_1^i, \dots, \mathbf{x}_N^i), t \in [1, T] \quad (4.19)$$

其中  $\mathbf{x}_1^i, \dots, \mathbf{x}_N^i$  是单词  $w_1^t, \dots, w_N^t$  的独热表示。在网络中所使用的 GRU 单元的结构计算公式如下所示：

$$\begin{aligned}\mathbf{h}_t &= (1 - z_t) \odot \mathbf{h}_{t-1} + z_t \odot \tilde{\mathbf{h}}_t \\ z_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z) \\ \tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{r}_t \odot (\mathbf{U}_h \mathbf{h}_{t-1}) + \mathbf{b}_h) \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r)\end{aligned}\quad (4.20)$$

解码器的结构对 GRU 进行了部分修改，取编码器在最后一个时刻的输出  $\mathbf{h}_N^i$  作为输入句子的表示  $\mathbf{h}_t$ ，加入到网络输入中。解码器在  $t$  时刻进行的迭代运算如下：

$$\begin{aligned}\mathbf{h}_t &= (1 - z_t) \odot \mathbf{h}_{t-1} + z_t \odot \tilde{\mathbf{h}}_t \\ z_t &= \sigma(\mathbf{W}_z \mathbf{y}_{t-1} + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{C}_z \mathbf{h}^i) \\ \tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \mathbf{y}_{t-1} + \mathbf{r}_t \odot (\mathbf{U}_h \mathbf{h}_{t-1}) + \mathbf{C}_h \mathbf{h}^i) \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{y}_{t-1} + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{C}_r \mathbf{h}^i)\end{aligned}\quad (4.21)$$

其中，修改的 GRU 单元接受三项输入：上一时刻的输出状态  $\mathbf{h}_{t-1}$ ；上一时刻输出的单词对应的词表示  $\mathbf{y}_{t-1}$ ，在自回归解码器中作为下一时刻的输入；输入句子  $s_i$  的表示向量  $\mathbf{h}^i$ 。GRU 单元的输出是  $\mathbf{h}_t$ ，即在时刻  $t$  对输出词语的预测。

给定输出状态后，解码器在每个时刻预测  $s^{i-1}$  和  $s^{i+1}$  在当前位置的单词。以解码  $s^{i+1}$  句子的解码器为例，给定输出表示  $\mathbf{h}_t^{i+1}$ ，对输出词语  $w_t^{i+1}$  的预测概率满足：

$$P(w_t^{i+1} | w_{<t}^{i+1}, \mathbf{h}^i) \propto \exp(\mathbf{v}_{w_t^{i+1}} \mathbf{h}_t^{i+1}) \quad (4.22)$$

其中  $\mathbf{v}_{w_t^{i+1}}$  是单词  $w_{i+1}^t$  的词向量， $\mathbf{h}_t^{i+1}$  是负责编码  $s^{i+1}$  的编码器在  $t$  时刻的输出。

模型的训练目标和语言模型相同，即句子  $s^{i-1}$  和  $s^{i+1}$  的预测概率最大化：

$$\mathcal{L}(s^i, s^{i+1}, s^{i-1}) = \sum_t \log P(w_t^{i+1} | w_{<t}^{i+1}, \mathbf{h}^i) + \sum_t \log P(w_t^{i-1} | w_{<t}^{i-1}, \mathbf{h}^i) \quad (4.23)$$

模型训练完成后，使用最终得到的模型编码器作为句子的表示模型。这个编码器可以为每个句子生成定长向量，且这个向量具有反映句子之间上下文相关关系的性质。

## 2. Sent2Vec 句子表示模型

为了捕捉句子的语义，需要通过整个句子的全局信息，学习句子中不同词语的关联关系。针对此问题，Sent2Vec 模型<sup>[184]</sup> 将 CBoW 模型的基于上下文窗口的学习机制扩展到整个句子的范围上，引入词级别的  $n$  元语法 ( $n$ -gram)<sup>①</sup> 特征提升句子中单词顺序的学习能力，从而更好地捕获上

<sup>①</sup>  $n$  元语法详细内容参考本书第6.2节中相关内容

下文语义。Sent2Vec 模型将句子中所有单词和所有  $n$  元语法单元的表示向量均值作为句子的表示：

$$\mathbf{v}_s = \frac{1}{|R(s)|} \sum_{w \in R(s)} \mathbf{v}_w \quad (4.24)$$

其中， $\mathbf{v}_s$  是句子  $s$  的表示， $R(s)$  是句子中全部单词及全部  $n$  元语法单元集合， $\mathbf{v}_w$  是  $R(s)$  的元素  $w$  的上下文词表示， $w$  可能是一个单词或一个  $n$  元语法单元。因此，Sent2Vec 词表和对应的嵌入矩阵同时包含单词和  $n$  元语法单元。

Sent2Vec 的训练目标和 CBoW 类似，通过优化中心词和上下文的相似性量度对文本向量进行自监督训练。具体而言，模型最大化中心词表示和除去该词后其余上下文表示的相似度。同时，Sent2Vec 也采用了负采样的技术，以降低计算成本。对于句子  $s$  和其中的单词  $w$ ，负采样词集合  $N(w)$  在除  $w$  外的词表上通过多项式分布采样得到，其中，记  $f_w$  为单词  $w$  的原始词频，单词  $w$  的采样概率为  $q_n(w) = \frac{\sqrt{f_w}}{\sum_{w \in \mathcal{V}} \sqrt{f_w}}$ 。Sent2Vec 的损失函数如下所示：

$$\mathcal{L}(w, s) = \ell(\mathbf{u}_w^T \mathbf{v}_{s \setminus \{w\}}) + \sum_{w' \in N(w)} \ell(-\mathbf{u}_{w'}^T \mathbf{v}_{s \setminus \{w\}}) \quad (4.25)$$

其中， $\ell(x) = \log(1 + e^{-x})$  是 Logistic 函数， $\mathbf{u}_w$  是单词  $w$  的中心词表示， $\mathbf{v}_{s \setminus \{w\}}$  是单词  $w$  上下文的表示， $w'$  是负采样词，来自负采样词集  $N(w)$ 。

在整个语料集上训练时，为了避免模型对高频词的倾向性，采用下采样（Subsampling）的方式使模型对单词词频脱敏。对于每个形如  $(w, s)$  的训练样本，以  $1 - q_p(w)$  的概率丢弃这个样本，其中  $q_p(w) = \min\{1, \sqrt{t/f_w} + t/f_w\}$ 。Sent2Vec 在整个语料库上训练的损失函数为：

$$\mathcal{L}(\mathcal{D}) = \sum_{s \in \mathcal{D}} \sum_{w \in s} k(w, s) \mathcal{L}(w, s) \quad (4.26)$$

其中  $k(w, s) \in \{0, 1\}$  是概率为  $q_p(w)$  的伯努利分布在样本  $(w, s)$  上的取样结果。

### 4.3.3 篇章分布式表示

在自然语言处理和信息检索领域，部分任务会要求模型学习并表示文档级别的特征，如文档检索、文档去重、文档级情感分析、主题识别等任务。相对一般的自然语言处理任务，这类任务不需要模型精确地捕获细粒度的词句信息，但需要模型建模文档的主题、包含的关键词等信息。编码这些信息成为了文档分布式表示的关键点。本节中将重点介绍词频-逆文档频率（TF-IDF）和 fastText 两种篇章分布式表示方法。

#### 1. 词频-逆文档频率篇章表示方法

词频-逆文档频率（TF-IDF）用来评估在特定文档中词的重要程度，其基本假设是文档中词重要程度随其在文档中出现的频率增加，同时也会随其在整个语料库中出现的频率而下降。对于上

上述假设可以从以下角度进行理解，如果一个词在特定文档的出现频率高，则说明这个词与该文档的主题具有比较强的相关关系，因此该词相对于该文档的重要性应该较高；但是，如果一个词语在整个文档集合很多文档上都出现了，那么说明该词是常见词语，其区分性不好，因此其重要程度应该较低。在信息检索领域，TF-IDF 常被用于衡量用户查询和文档之间的相似性。

TF-IDF 方法包括词汇频率 (Term Frequency, TF) 和逆文档频率 (IDF, Inverse Document Frequency) 两个部分。具体地，词汇频率主要用来衡量词汇在特定文档中的重要程度，建模文档中的关键词。而逆向文件频率用来衡量词汇在普遍情况下的常见性，用于去除常见词对文档关键词建模的影响。对于文档  $d$  中的词项  $t$ ， $\text{TF}(t, d)$  表示词项在文档中的出现频率，具体计算方法如下所示：

$$\text{TF}(t, d) = \frac{\text{COUNT}(t, d)}{\sum_{t' \in \mathcal{V}} \text{COUNT}(t', d)} \quad (4.27)$$

其中， $\text{COUNT}(t, d)$  是文档  $d$  中的词项  $t$  的出现次数，分母的和式表示文档中的总词数。

对于词项  $t$  的逆向文件频率  $\text{IDF}(t)$ ，计算包含该词项的文档占全体文档的比例， $\text{IDF}(t)$  与之呈对数反相关关系：

$$\text{IDF}(t) = \log \frac{|\mathcal{D}|}{\sum_{d \in \mathcal{D}} \mathbf{1}^{t \in d}} \quad (4.28)$$

其中， $|\mathcal{D}|$  表示文档总数，分母的和式表示包含词项  $t$  的文档数。在计算未登录词项的逆向文件频率时，为了防止分母为零，通常采用平滑化方法，用  $\sum_{d \in \mathcal{D}} \mathbf{1}^{t \in d} + 1$  替代  $\sum_{d \in \mathcal{D}} \mathbf{1}^{t \in d}$  作为分母。

词项  $t$  在文档  $d$  中的 TF-IDF 表示为词汇频率和逆向文件频率的乘积：

$$\text{TF-IDF}(t, d) = \text{TF}(t, d)\text{IDF}(t) \quad (4.29)$$

文档的表示向量由词表中每个词项在文档中的 TF-IDF 值构成，每个维度对应一个词项。这种表示向量的特点是对关键词信息的反映。对于每个文档，具有较高的 TF-IDF 的词项是在整个文档集合中出现频率较低，但在本文档中出现频率较高的词项。因此，基于 TF-IDF 的表示向量筛选文档中的高频词，过滤掉其中的常见词，保留反映文档主题、主要内容的关键词。在信息检索系统中，对于用户输入的检索关键词，可以容易地通过比较文档中词项的 TF-IDF，来返回与检索词相关性高的文档。

## 2. fastText 篇章表示模型

fastText 模型<sup>[185, 186]</sup> 旨在高效训练文本表示模型，因其良好的性能和效率，而被广泛地使用在文本分类任务上。在构建词级别表示时，fastText 会利用字符 n-gram 特征，以更好地表示罕见词和未登录词。具体地，fastText 在单词的开头和末尾添加表示前缀和后缀的字符“<”和“>”，然后通过滑动窗口的方式获得该词固定长度的所有子词。

例如：单词“where”长度为 3 的子词包括：“<wh”，“whe”，“her”，“ere”，“re>”  
基于对训练语料的统计，fastText 为全体单词长度不小于  $n$  的子词建立词表和对应的子词表示矩

阵，并使用这些子词辅助构建词表示。例如， $n=3$  时，使用单词“where”的长度为 3 到 7<sup>①</sup>的全体子词（包括该词自身）辅助构建该词的表示。

fastText 直接使用子词向量的和作为对应单词的词向量：

$$\mathbf{v}_w = \sum_{s \in \mathbb{V}_s(w)} \mathbf{z}_s \quad (4.30)$$

其中， $\mathbf{v}_w$  是单词  $w$  的词向量， $\mathbb{V}_s(w)$  是子词词表中  $w$  的全体子词， $\mathbf{z}_s$  是子词  $s$  的表示向量。

fastText 通常使用 Skip-Gram 模型的训练方式得到预训练的词级别表示，其中中心词的向量使用形如公式4.30的方式构建，而其余部分和 Skip-Gram 模型保持一致。在预训练中，fastText 采用 4.3.1 节所述的层次 Softmax 方法提升训练效率。

在构建文档表示时，fastText 首先基于上述方式计算每个词的表示向量，再将其进行平均，得到句表示向量。模型结构如图4.10所示。在将 fastText 句向量应用于文本分类任务时，通常以对数概率作为优化目标：

$$\mathcal{L}(x = w_1, \dots, w_N, y) = -y \cdot \log(\text{softmax}(\mathbf{W} \cdot \mathbf{v})) \quad (4.31)$$

$$\mathbf{v} = \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i \quad (4.32)$$

其中， $\mathbf{v}_i$  是单词  $w_i$  的词向量， $\mathbf{v}$  是文档的表示向量， $\mathbf{W}$  是可训练的线性映射层，将样本表示映射为预测分布。

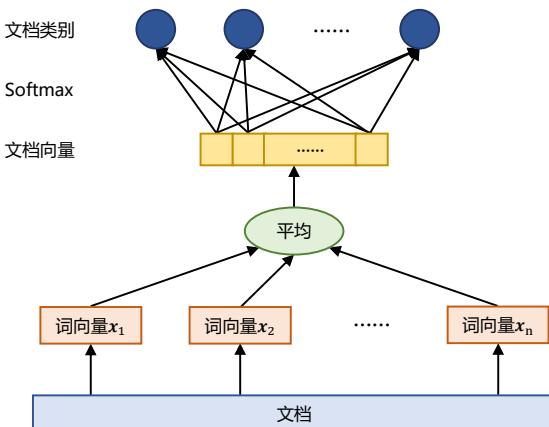


图 4.10 fastText 模型结构图

<sup>①</sup> 单词前后增加了“<”和“>”符号，因此单词 where 的长度为 7

## 4.4 词义消歧

词义消歧（Word Sense Disambiguation, WSD）是指确定一个多义词在给定的上下文中的具体含义。根据本章第4.1.1节词汇语义学相关介绍，我们可以知道语言中一词多义现象十分普遍。例如，“水分”既可以表示物体内所含的水，也可以表示某些情况中夹杂的不真实的成分，可以使用“水分<sup>1</sup>”和“水分<sup>2</sup>”分别表示两个含义，以下句子为例：

- (1) 葡萄糖液可用来供给水分。单词义项：水分<sup>1</sup>
- (2) 这个报导有些水分，需要核实。单词义项：水分<sup>2</sup>

词义消歧任务核心就是根据词语所处的句子或者篇章，确定该词在当前环境下的确切含义。上例中句子(1)和句子(2)中的“水分”分别对应两个不同的义项。该任务对于机器翻译、语义理解、对话系统、阅读理解等任务具有十分重要的作用。本节将介绍基于目标词上下文、基于词义释义匹配以及基于词义知识增强预训练等三类词义消歧方法。

### 4.4.1 基于目标词上下文的词义消歧方法

对于待消歧的目标词，词义消歧方法通常采用有监督分类方法，将词语的每个词义项作为候选词义，通过估计待消歧词义的概率分布从而完成目标词的词义消歧。基于目标词上下文的词义消歧方法利用待消歧目标词的上下文进行训练，预测上下文中目标词属于每个候选词义的条件概率。自然语言处理中常用的统计机器学习方法和深度学习算法，均可用于构建基于目标词上下文的词义消歧方法。本节以基于朴素贝叶斯分类器、上下文向量表示的词义消歧方法为例，介绍基于统计机器学习及深度学习的词义消歧方法。

#### 1. 基于朴素贝叶斯分类器的消歧方法

使用  $w$  表示待消歧的目标词， $c$  表示目标词所处的句子， $\{s_i\}_{i=1}^N$  为目标词的候选词义集合。由于待消解的目标词词义仅与其所处的上下文语境有关，因此可以通过估计条件概率  $P(s_i|c)$  来预测目标词  $w$  的词义<sup>[187]</sup>。

给定句子  $c = \{w_k\}_{k=1}^K$ ，可以将条件概率  $P(s_i|c)$  通过贝叶斯公式转换为：

$$P(s_i|c) = \frac{P(c|s_i)P(s_i)}{P(c)} \propto P(c|s_i)P(s_i) \quad (4.33)$$

其中， $P(c|s_i)$  的计算方式与语言模型类似。随着上下文长度的增加，上下文  $c$  的数量指数级增长，该概率值难以估计。因此，需要引入单词独立假设，近似地将概率估算为上下文中每个单词的独立出现概率：

$$P(c|s_i) = \prod_{k=1}^K P(w_k|s_i) \quad (4.34)$$

通过公式4.33和公式4.34，词义分类可以根据如下公式选择最大条件概率的词义：

$$\hat{s} = \arg \max_{s_i} P(s_i|c) = \arg \max_{s_i} P(s_i) \prod_{k=1}^K P(w_k|s_i) \quad (4.35)$$

$P(s_i)$  和  $P(w_k|s_i)$  可以通过训练语料利用最大似然估计得到：

$$P(w_k|s_i) = \frac{\text{COUNT}(w_k, s_i)}{\text{COUNT}(s_i)} \quad (4.36)$$

$$P(s_i) = \frac{\text{COUNT}(s_i)}{\text{COUNT}(w)} \quad (4.37)$$

其中， $\text{COUNT}(w_k, s_i)$  是训练语料中目标词  $w$  以语义  $s_i$  在上下文中出现的次数； $\text{COUNT}(s_i)$  是训练语料中语义  $s_i$  出现的总次数； $\text{COUNT}(w)$  是训练语料中目标词  $w$  出现的总次数。

在实际算法实现中，为了提升计算精度，概率值通常采用对数形式参与计算。具体的算法如4.3所示：

---

#### 代码 4.3: 基于朴素贝叶斯分类器的词义消歧算法

---

输入：训练数据  $D$ ，语句  $c$ ，目标词  $w$ ，候选语义  $\{s_i\}_{i=1}^N$

输出：预测的词义  $\hat{s}$

```

for  $i = 1$  to  $N$  do
    for  $k = 1$  to  $K$  do
         $| P(w_k|s_i) = \text{COUNT}(w_k, s_i) / \text{COUNT}(s_i) \quad //$  根据训练数据计算单词出现概率
    end
     $| P(s_i) = \text{COUNT}(s_i) / \text{COUNT}(w) \quad //$  根据训练数据计算语义出现概率
end
 $\hat{s} = \arg \max_{s_i} \log P(s_i) \sum_{k=1}^K \log P(w_k|s_i) \quad //$  选择出现概率最大的词义
return  $\hat{s}$ 

```

---

## 2. 基于上下文向量表示的消歧方法

本章第 4.3 节分布式表示中介绍了句子和短语分布式向量表示算法，可以看到深度神经网络算法可以很好地对句子和短语的语义进行表示。词义消歧任务的核心就是根据上下文信息判断当前单词的词义，因此，也可以利用目标词上下文的分布式表示，建模目标词上下文语义，并基于上下文向量表示构建词义消歧算法。

基于上下文向量表示的最近邻方法<sup>[28, 188]</sup> 将词义消歧任务形式化为词义表示和上下文表示的相似度学习问题。如图4.11所示，该方法的主要过程是根据词义在训练集样本中出现的上下文，学习词义的表示。根据语料库训练目标词的上下文表示和候选词义表示的相似度计算模型。在应用

时，将每个候选词义的表示和目标词的上下文表示进行比较，选择相似度最高的词义。针对未在词义消歧语料库中出现的词义，可以进一步使用相似的语义集合确定其词义表示。

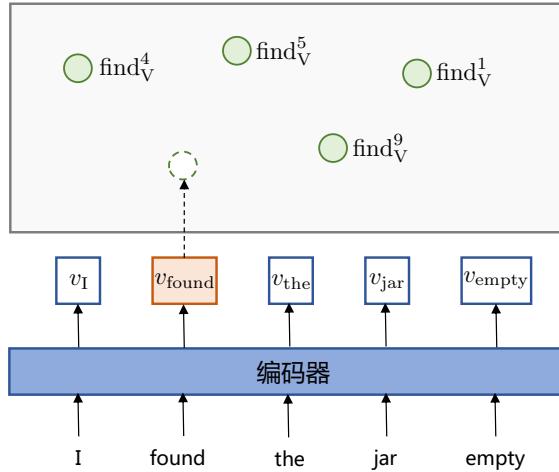


图 4.11 基于上下文向量表示的最近邻模型结构图<sup>[188]</sup>

在词义编码部分，首先考虑在词义消歧语料库中存在标注的语义。对于每一个标注词义，在训练集中抽取全体包含该词义标注的样本。随后，通过预训练上下文表示模型，计算词义对应的目标词在样本上下文中的表示。最后，以目标词表示的平均值作为词义的表示。具体地，对于词义  $s$ ，若该词义在词义消歧语料库出现过，则根据如下公式计算该词义表示  $v$ ：

$$v_0, v_1, \dots, v_w, \dots, v_n = \text{Encoder}(c = c_0, c_1, \dots, w, \dots, c_n) \quad (4.38)$$

$$v = \frac{1}{|C(s)|} \sum_{c \in C(s)} v_w \quad (4.39)$$

其中， $C(s)$  为全体标记词义为  $s$  的样本集合，Encoder 代表使用预训练语言模型初始化的编码器，如 ELMo、BERT 等。在 BERT 模型中，对于分解成多个标识位的单词，采用每个标识位输出的平均值作为单词的表示。 $w$  是词义在样本中对应的目标词，在不同的样本中，目标词不必相同。 $v_w$  是目标词  $w$  在上下文  $c$  中的表示， $v$  是词义  $s$  的上下文表示。

针对未在词义消歧语料库中出现的词义，可以采用 LMMS<sup>[189]</sup> 方法，利用 WordNet 中标注的同义词、上位词和词性标注（Lexname）等语义关系信息，寻找与目标词义相似或相关的词义，再以这些词义表示的平均值作为该词义的表示。具体地，以同义词关系为例，对于待确定表示的词

义  $s$ , 记  $S(s)$  为  $s$  的同义语义集合。若  $S(s)$  不是空集,  $s$  的语义表示为  $S(s)$  中同义语义的平均表示:

$$\mathbf{v} = \frac{1}{|S(s)|} \sum_{s \in S(s)} \mathbf{v}_s, \text{if } |S(s)| > 0 \quad (4.40)$$

当同义语义缺失时, 可依次使用相同上位的语义或相同词性的语义作为近义语义集合, 利用相似的方式计算目标语义的表示, 具体计算公式如下所示:

$$\mathbf{v} = \frac{1}{|H(s)|} \sum_{s \in H(s)} \mathbf{v}_s, \text{if } |H(s)| > 0 \quad (4.41)$$

$$\mathbf{v} = \frac{1}{|L(s)|} \sum_{s \in L(s)} \mathbf{v}_s, \text{if } |L(s)| > 0 \quad (4.42)$$

其中  $H(s)$  为  $s$  的同上位语义集合, 包含与  $s$  属于同类概念的语义。 $L(s)$  为  $s$  的同词性语义集合, 包含与  $s$  词性相同的全体语义。在结合语义关系之后, 可以全面覆盖 WordNet 中的词义, 从而解决候选词义未在训练集出现的问题。

在构建了所有词义的向量表示后, 对于每一条输入的待进行词义消歧的样本, 首先基于语言模型计算目标词的上下文表示, 在此基础上, 计算上下文表示与全体候选词义表示的点积相似度, 选择相似度最大的语义做为分类结果, 具体计算公式如下所示:

$$\hat{s} = \arg \max_s (\mathbf{v}_w \cdot \mathbf{v}(s)) \quad (4.43)$$

其中,  $\mathbf{v}_w$  为目标词的上下文表示, 与词义表示映射到相同维度;  $\mathbf{v}(s)$  为候选词义  $s$  的表示。

#### 4.4.2 基于词义释义匹配的词义消歧方法

以知网 (HowNet)<sup>[165]</sup>、WordNet<sup>[36]</sup> 等为代表的词汇知识资源中不仅包含了词义之间的关系, 还包含了词义的解释信息。

例如: WordNet 3.1 中对“table”给出了如下词义解释:

table<sup>1</sup>: a set of data arranged in rows and columns

table<sup>2</sup>: a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs

table<sup>3</sup>: a piece of furniture with tableware for a meal laid out on it

table<sup>4</sup>: flat tableland with steep edges

table<sup>5</sup>: a company of people assembled at a table for a meal or game

table<sup>6</sup>: food or meals in general

这些释义与目标词上下文之间存在着非常强的联系。比如, table<sup>1</sup> 所对应的“表格”含义, 其上下文

更多的对应的设计、制作、数据等词汇。而 table<sup>2</sup> 所对应的“桌子”含义，其上下文更多的对应的椅子、沙发等词汇。因此，也可以将词义消歧问题转化为目标词上下文和词义释义之间的语义匹配问题。对于待消歧的目标词  $w$  所在的上下文句子  $c$ ，以及候选词义  $s$ ，构建相似度度量函数，建模目标词上下文和候选词义的匹配度。在应用时，根据目标词和每个候选语义的相似度得分，来确定目标词义的预测分布  $\phi(w|c, s)$ 。本节将以基于特征式和交互式匹配的两类消歧方法为例，介绍基于词义释义匹配的词义消歧方法。

### 1. 基于特征式匹配的消歧方法

BEM 模型<sup>[190]</sup> 通过分布式向量表示匹配方式学习目标词上下文和词义释义的相关性。BEM 模型结构如图4.12所示，主要包含上下文编码器和词义编码器两个组成部分。上下文编码器  $T_c$  对输入的目标词及其上下文进行编码，计算目标词上下文的分布式表示。词义编码器  $T_g$  对输入的词义释义文本进行编码，将输入词义和上下文表示在同一表示空间内。通过建立上下文语义表示和候选词义表示的相似度计算模型，来完成词义消歧任务。

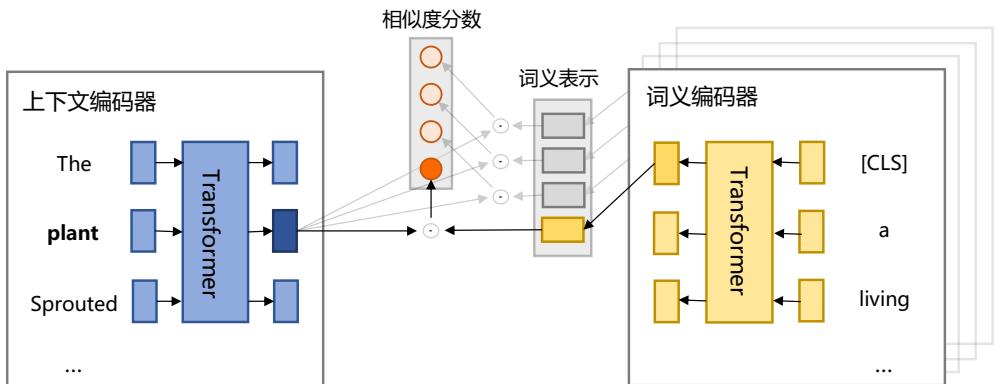


图 4.12 BEM 模型结构图<sup>[190]</sup>

BEM 模型结构的上下文编码器  $T_c$  和词义编码器  $T_g$  都采用基于 BERT 的架构。针对目标词上下文表示的计算，使用  $w$  表示待消歧的目标词， $c = c_0, c_1, \dots, w, \dots, c_n$  表示目标词所处的上下文，上下文编码器  $T_c$  将序列中的每一个单词构建对应的上下文表示，并取目标词位置的输出为目标词的上下文表示，具体计算公式如下所示：

$$\mathbf{v}_{cls}, \mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_w, \dots, \mathbf{v}_n, \mathbf{v}_{sep} = T_c([CLS], c_0, c_1, \dots, w, \dots, c_n, [SEP]) \quad (4.44)$$

其中  $\mathbf{v}_w$  是目标词  $w$  在句子中的上下文表示。对于分解成多个标识位的单词，采用每个标识位输出的平均值作为该单词的表示。

分别针对候选词义表示的计算，候选词义  $s$  的词义释义为  $g_s = g_0, g_1, \dots, g_m$ ，在词义释义序列的首尾分别添加 [CLS] 及 [SEP] 标识，输入词义编码器  $T_g$ ，取 [CLS] 位置的输出作为词义的表示。计算过程如下所示：

$$\mathbf{v}_{cls}^g, \mathbf{v}_0^g, \mathbf{v}_1^g, \dots, \mathbf{v}_m^g, \mathbf{v}_{sep}^g = T_g([CLS], g_0, g_1, \dots, g_m, [SEP]) \quad (4.45)$$

其中，记  $\mathbf{v}_s = \mathbf{v}_{cls}^g$  为候选词义  $s$  的表示。

对于上下文  $c$  中待消歧的目标词  $w$ ，以及候选词义  $s$ ，它们的相似度由如下公式计算得到：

$$\phi(w|c, s) = \mathbf{v}_w \cdot \mathbf{v}_s \quad (4.46)$$

其中  $\mathbf{v}_w$  是目标词的表示向量， $\mathbf{v}_s$  是词义  $s$  的表示向量。

在模型训练过程中，对于待消歧的目标词  $w$ ，取该目标词在句子中的表示与全体候选词义的表示进行相似度计算，以相似度作为预测词义的对数概率分布，优化交叉熵损失函数，具体计算公式如下：

$$\mathcal{L}(w; c) = -\phi(w|c, s^+) + \log \sum \exp(\phi(w|c, s^-)) \quad (4.47)$$

其中， $s^+$  表示正确的候选词义， $s^-$  表示其余的候选词义。

在使用模型进行词义消歧时，与目标词具有最大相似度的词义将被预测为目标词的词义，即：

$$\hat{s} = \arg \max_s \phi(w|c, s) \quad (4.48)$$

## 2. 基于交互式匹配的消歧方法

基于深度神经网络的交互式方法在语义匹配任务上取得了不错的效果。在交互式匹配中，待判断文本对以特定的方式拼接在一起输入模型，然后以与单文本相同的方式进行处理。交互式匹配的优点是只使用一个编码器进行匹配任务，大大减小了训练参数的规模。此外，交互式匹配可以充分利用词粒度的信息，参考输入的一对文本中的每个单词，进行充分的比较，从而实现更好的学习效果。交互式匹配方法也可以应用于上下文和词义释义的相似度学习中。

GlossBERT<sup>[191]</sup> 使用交互式匹配方法，通过对预训练模型 BERT 进行微调，实现上下文和词义释义的相似度计算。GlossBERT 以 BERT 双句分类的方式，将目标词所处的上下文句子和词义释义组合为输入，以是否匹配作为二分类标签，构造分类模型的微调样本，通过这些样本进行模型的微调。模型通过微调后，对于待消歧的目标词和候选词义，将目标词上下文和每一个候选词义组合成输入，通过模型计算语义匹配的置信度，根据置信度选取预测词义。

对于微调样本的构造，表4.6通过一个示例样本展示了 GlossBERT 微调样本的格式。该句以“research”作为消歧目标词时，目标词具有“research%1:04:00::”等 4 个词义。GlossBERT 按照输入模板将上下文句子和各个词义的释义分别拼接为 BERT 的输入形式，并在输入词序列中使用双引号

(“)、冒号 (:) 等标点符号标出目标词的位置。对于和目标词匹配的词义，GlossBERT 为微调样本赋予“Yes”类标签，否则赋予“No”类标签。

表 4.6 GlossBERT 的微调样本构造，目标词用斜体表示。

原句：Your <i>research</i> stopped when a convenient assertion could be made .		
微调样本	标签	原始词义
[CLS] Your ”research” ... [SEP] research: systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your ”research” ... [SEP] research: a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your ”research” ... [SEP] research: inquire into [SEP]	No	research%2:31:00::
[CLS] Your ”research” ... [SEP] research: attempt to find out in a ... [SEP]	No	research%2:32:00::

整个过程可以形式化地表示为：给定待消歧对目标词  $w$  及其所处的上下文句子  $c$ ，对于  $w$  的每个候选词义  $s_i$  的释义  $g^i = g_0^i, \dots, g_m^i$ ，将句子和词义组合为如下的输入形式：

$$\begin{aligned} x &= f(w, c, s_i) \\ &= [CLS], c_0, c_1, \dots, ", w, ", \dots, c_n, [SEP], w, :, g_0^i, \dots, g_m^i, [SEP] \end{aligned} \quad (4.49)$$

其中，[CLS] 和 [SEP] 是 BERT 使用的特殊标识，双引号 (‘)、冒号 (:) 是用来标出目标词位置的特殊标识。如果目标词实际词义和词义释义匹配，则标注正例标签，否则标注负例标签。因此，对于样本中每个待消歧的目标词，可以构造  $N$  个分类样本，其中  $N$  是目标词候选词义的个数。

GlossBERT 的模型结构如图4.13所示，其中，BERT 分类层包含一个 BERT 编码层和一个线性分类层。如公式4.50所示，GlossBERT 根据训练集中每个样本的每个目标词所构造的分类样本，使用 BERT 编码层在 [CLS] 位置的输出作为分类判据，通过下游的线性分类层进行语义是否匹配的二元分类，分类层计算公式如下所示：

$$\hat{y} = \mathbf{W} \cdot \text{BERT}(x)[0] \quad (4.50)$$

其中  $\mathbf{W}$  为线性分类层的权重。

GlossBERT 根据交叉熵损失函数微调 BERT 编码层及线性输出层的全部权重：

$$\mathcal{L}(x, y) = \text{CrossEntropy}(\hat{y}, y) \quad (4.51)$$

当使用模型进行词义消歧时，将目标词构造对应的分类样本，输入模型对其进行预测。预测正例标签置信度最高的候选词义将被预测为目标词的词义。

$$\hat{s} = \arg \max_s P(\hat{y}(w, c, s) = 1) \quad (4.52)$$

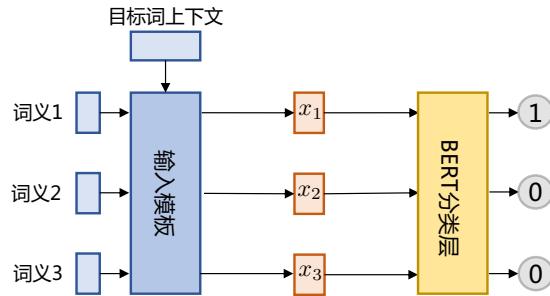


图 4.13 GlossBERT 模型结构图

#### 4.4.3 基于词义知识增强预训练的消歧方法

基于预训练语言模型的方法在词义消歧任务中取得了不错的结果，为了使得预训练语言模型更好地适应词义消歧任务，可以通过设计词义级别的预训练任务，使得预训练模型融合知识库中所包含词义信息。然而，预训练模型需要大规模的有监督数据才能对模型参数进行有效训练。但是，目前缺乏标注了词义的大规模数据用于支持模型预训练。

SenseBERT 模型<sup>[192]</sup>，针对缺失语义监督数据问题，在 BERT 的预训练中添加了一个掩码词义预测任务作为辅助任务。SenseBERT 利用 WordNet 所包含的超义（Supersense）信息作为弱监督信号。WordNet 将所有义项归纳为多个类别，这些类型称之为超义。例如，针对名词有 26 个超义，包括：BODY、LOCATION、PLANT 等。模型在预训练中不但预测掩码位置单词的词形，还预测缺失单词的超义。通过这种方式引入语义级别的监督信息，从而提升预训练模型在词义消歧任务上的效果。

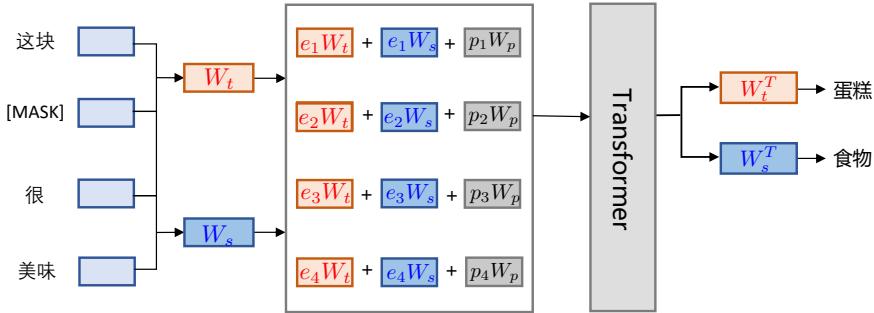
SenseBERT 的模型结构如4.14所示，在输入表示部分增加了语义嵌入模块，用来建构单词的词义信息。在 SenseBERT 中，对于输入序列 \$w\_1, \dots, w\_n\$，每个标识位的输入表示 \$v\_t\$ 由词形嵌入、词义嵌入和位置嵌入的和组成：

$$v_t = e_t(\mathbf{W}_t + \mathbf{W}_s) + p_t \mathbf{W}_p \quad (4.53)$$

其中，\$e\_t \in \mathbb{R}^{|\mathbb{V}|}\$ 是第 \$t\$ 个位置上标识的独热向量表示，\$\mathbf{W}\_t \in \mathbb{R}^{|\mathbb{V}| \times d\_t}\$ 是词嵌入矩阵，\$\mathbf{W}\_s \in \mathbb{R}^{|\mathbb{V}| \times d\_s}\$ 是语义嵌入矩阵，\$|\mathbb{V}|\$，\$d\_t\$ 和 \$d\_s\$ 分别表示词表大小、词嵌入维度和语义嵌入维度，其中语义嵌入维度为超义类型的个数；\$p\_t \in \mathbb{R}^N\$ 是标识所在位置的独热向量表示，\$\mathbf{W}\_p \in \mathbb{R}^{N \times d\_p}\$ 是位置嵌入矩阵，\$N\$ 和 \$d\_p\$ 分别表示位置长度上限和位置嵌入维度。

SenseBERT 使用和 BERT 相同的 Transformer 编码层作为输入的编码结构：

$$o_1, \dots, o_n = \text{Transformer-Encoder}(v_1, \dots, v_n) \quad (4.54)$$

图 4.14 SenseBERT 模型结构图<sup>[192]</sup>

在预训练任务方面，SenseBERT 包括掩码单词预测和掩码语义预测两个任务。通过与词嵌入、语义嵌入矩阵的比较，模型计算每一个掩码位置的单词预测分布和语义预测分布，并将其与实际标签比对。具体计算公式如下所示：

$$P(\hat{w}_t | context) = \text{Softmax}(\mathbf{W}_t^T \mathbf{o}_t) \quad (4.55)$$

$$P(\hat{s}_t | context) = \text{Softmax}(\mathbf{W}_s^T \mathbf{o}_t) \quad (4.56)$$

在掩码单词预测任务上，如公式4.57所示，模型约束单词预测分布和实际单词的交叉熵损失。根据 WordNet 统计实际单词全体词义所属的超义，单词  $w$  的超义集合记为  $A(w)$ ，即单词  $w$  的可取义项，作为掩码语义预测的弱监督信号。

$$\mathcal{L}_{LM}(w_t) = -\log P(\hat{w}_t | context) \quad (4.57)$$

在掩码语义预测的优化中，要求模型预测与超义集合中的语义吻合即可。如公式4.58所示，损失项  $\mathcal{L}_{SLM}^{allowed}(w_t)$  优化模型预测掩码位置语义  $\hat{s}_t$  属于可取义项集合  $A(w_t)$  的概率。考虑到模型在训练过程中可能会出现过拟合特定超义项的情况，SenseBERT 引入了正则项  $\mathcal{L}_{SLM}^{reg}(w_t)$ 。如公式4.59所示，此损失项对  $A(w_t)$  中每个超义项的模型预测置信度的平均值进行约束，保证模型对全体可取义项都有一定的预测置信度，防止出现模型对单一义项预测置信度过高，预测分布退化的情况。

$$\mathcal{L}_{SLM}^{allowed}(w_t) = - \sum_{s_t \in A(w_t)} \log P(\hat{s}_t | context) \quad (4.58)$$

$$\mathcal{L}_{SLM}^{reg}(w_t) = - \sum_{s_t \in A(w_t)} \frac{1}{|A(w_t)|} \log P(\hat{s}_t | context) \quad (4.59)$$

模型在预训练阶段的整体损失函数为掩码单词预测和掩码语义预测目标的加权求和值：

$$\mathcal{L} = \sum_{w_t=[MASK]} \lambda_1 \mathcal{L}_{LM}(w_t) + \lambda_2 \mathcal{L}_{SLM}^{allowed}(w_t) + \lambda_3 \mathcal{L}_{SLM}^{reg}(w_t) \quad (4.60)$$

由于掩码词义预测只对被掩盖的全词有意义, SenseBERT 在掩码标识为子词时, 只优化掩码单词预测损失, 不训练掩码语义预测任务。为了进一步增强对词义知识的学习, SenseBERT 在 BERT 的原始词表中增加了维基百科中的高频词, 使掩码全词的比例更大, 有助于学习低频词义。

#### 4.4.4 词义消歧评价方法

词义消歧评测的主要指标通常采用机器学习算法评测常用指标, 包括词义消歧的精确率 (Precision), 召回率 (Recall), 和 F 值 (F-Score) 得分。在词义消歧任务中的计算方式如下:

$$\text{精确率 (P)} = \frac{\text{算法输出的正确标记个数}}{\text{算法输出的全部标记个数}} \times 100\% \quad (4.61)$$

$$\text{召回率 (R)} = \frac{\text{算法输出的正确标记个数}}{\text{测试集合中全部标记个数}} \times 100\% \quad (4.62)$$

$$\text{F 值 (F1)} = \frac{2 \times P \times R}{P + R} \quad (4.63)$$

在部分基于机器学习的方法中, 也会使用精度和覆盖率 (Coverage) 作为词义消歧系统的评测指标<sup>[193]</sup>。覆盖率的计算方式如4.64所示:

$$\text{覆盖率} = \frac{\text{算法输出的全部标记个数}}{\text{测试集合中全部标记个数}} \times 100\% \quad (4.64)$$

#### 4.4.5 词义消歧语料库

1997 年以来, 国际计算语言学联合会 (ACL) 的词法研究小组 (SIGLEX) 开始组织关于词义消歧的公共评测任务 SensEval, 发布了词义消歧任务的训练集和测试集, 用于评测词义消歧任务的性能。SIGLEX 同样也支持了 SemEval 研讨会的举办, 在每年的研讨会上发布一部分有挑战性的共享任务, 建立高质量的注释数据集。这些大规模词义消歧标注语料库和竞赛, 极大的推动了自有监督机器学习算法大规模应用于词义消歧。

词义消歧语料库主要包含两种类型: 义项分类和义项相同判断。义项分类语料库针对目标词语, 构建包含该目标词句子, 并对句子中目标词所属的语义项进行标注。义项相同判断语料库则针对目标词给出两个包含该词语的句子, 并依据在两个句子中目标词的词义是否相同给出分类标签。本节将分别介绍上述两种常用词义消歧系统训练的大规模语料库。

## 1. 词义消歧项分类标注语料库

SemCor<sup>[194]</sup> 是迄今为止最大的手工标注词义的语料库之一。目前的词义消歧算法研究中，大多数有监督方法均使用 SemCor 语料库进行训练。SemCor 是基于 WordNet 词义进行标注的语料库。SemCor 3.0 版本包含 352 个文档和 22 万余条手动语义注释，其原始语料从布朗（Brown）语料库获取，经过筛选后，参考 WordNet 1.4 的词义清单进行词义标记。SemCor 分为三部分，其中 Brown1, Brown2 分别包含 103 个和 83 个文件，标记了所有开放类型的词语；Brownv 包含 166 个文件，只标注了动词。在这些文件中，对于文中的每一个句子或段落，标注其中每个单词的词性，并对名词、动词、形容词和副词标记来自 WordNet 的义项和语义。表4.7展示了 SemCor 的一个标注数据样例，在句中标出了开放域词汇，如“got... up”；包含多个词汇的表达，如“on their feet”；以及命名实体，如 “Kim” 以实体类型进行标注<sup>[195]</sup>。

表 4.7 SemCor 语料库的示例样本标注

例句： Kim <sub>a</sub> got <sub>b</sub> slowly <sub>c</sub> up <sub>b</sub> , the children <sub>d</sub> were <sub>e</sub> already <sub>f</sub> on <sub>g</sub> their <sub>g</sub> feet <sub>g</sub> .			
编号	范围	对应词	词义
a	Kim	Kim	<b>org</b>
b	got, up	get_up	get_up_4
c	slowly	slowly	slowly_1
d	children	child	child_1
e	were	be	be_3
f	already	already	already_1
g	on their feet	on_one's_feet	<b>no_tag</b>

OMSTI (One Million Sense-Tagged Instances)<sup>[196]</sup> 是自动标注的语料库，也常用于词义消歧系统的训练。OMSTI 使用 WordNet 3.0 的词义进行注释，它是通过在大型英汉平行语料库（MultiUN 语料库）上使用基于对齐的词义消歧方法自动构建的。OMSTI 中包含针对 22437 个单词，由 85 万句子组成的 113 万训练样例。虽然 OMSTI 是自动标注的，但其具有比 SemCor 更大的规模，且包含更丰富的歧义情况，很多工作的实验说明将 OMSTI 应用于词义消歧算法训练，可以有效提升算法效果。

WSDEvaL<sup>[197]</sup> 是统一词义消歧基准评测框架，将不同时期构建的采用不同词义注释构建的评测基准语料统一使用 WordNet 3.0 词义进行注释。WSDEvaL 包含 5 个来自 SensEval 和 SemEval 的测试基准语料，具体如下：

- SensEval-2 使用 WordNet 1.7 进行标注，包含 2282 个词语义标注。
- SensEval-3 Task 1 使用 WordNet 1.7.1，包含来自社论，新闻和小说领域的 3 个文档，包含 1850 个词语义标注。
- SemEval-07 Task 17 使用 WordNet 2.1，包含 455 个词语义标注。

- SemEval-13 Task 12 使用 WordNet 3.0 的标注，包含来自多个领域的 13 个文档和名词上的 1644 个词语义标注。
- SemEval-15 Task 13 使用 WordNet 3.0 的标注，包含生物医学、数学/计算和社会问题题材的 4 个文档和 1022 个词语义标注。

## 2. 词义消歧义项相同判断标注语料库

WiC (Word in Context) 数据集是一个由专家标注的词义消歧数据集，每个样本对同一个目标词给出两个包含该词语的句子，并依据在两个句子中目标词的词义是否相同，给出 T 或 F 的分类标签。表4.8给出了 WiC 数据集中的一些示例样本。在 WiC 的构建过程中，句子主要来自 WordNet 中的例句语料，并使用了一部分来自 VerbNet 和 Wiktionary 的语料资源。在这些句子中，只取具有多种含义、并出现在不同句子中的名词和动词作为目标词。为了筛选高质量的验证集和测试集，标注者要求包含相同目标词的样本不超过 3 个，且样本之间没有重复的上下文句子。在此基础上，尽量保证类别平衡、以及目标词的多样性。表4.9给出了 WiC 数据集的统计数据，包括样本个数、目标词个数和目标词中名词、动词的占比。

表 4.8 WiC 数据集的示例样 ( 目标词用斜体表示，同义/非同义的语句对标记为 T/F )

语句对	标签
There is a lot of trash on the <i>bed</i> of the river I keep a glass of water next to my <i>bed</i> when I sleep	F
<i>Justify</i> the margins The end <i>justifies</i> the means	F
Air pollution Open a window and let in some <i>air</i>	T
The expanded <i>window</i> will give us time to catch the thieves You have a two-hour <i>window</i> of clear weather to finish working on the lawn	T

表 4.9 WiC 数据集样本统计

数据划分	样本个数	名词比例	动词比例	目标词数
训练集	5,428	49%	51%	1,256
验证集	638	62%	38%	599
测试集	1,400	59%	41%	1,184

与 SemEval 和 SensEval 评测基准相比，WiC 不依赖于 WordNet 等外部数据库的词义信息，而且作为分类数据集，具有较为简单的任务形式。简洁的任务形式促进了模型的灵活性和丰富性，而不依赖于外部信息的特质使 WiC 支持低资源场景下的评测，而不强制假设完整训练数据的可用性。因为 WiC 的简洁性、独立性，以及其对模型语义理解能力较高的要求，SuperGLUE 系列基准

测试任务收录 WiC 任务作为词义消歧方面的评测基准。

在 WiC 的基础上，针对词义消歧系统对专用领域词义的建模，WiC-TSV（Word in Context - Target Sense Verification）对 WiC 的语料筛选和任务形式进行了改进，形成了新的跨越多个领域的词义消歧评测基准。与 WiC 不同，WiC-TSV 中的每个样本仅包含一个句子，其中标出待消歧的目标词。同时，样本包含该单词的一个预期词义，根据目标词在上下文中的实际词义是否和给出的词义相符，标记 T 或 F 的分类标签。另外，样本中还会给出单词词义的上位词，保证预期词义和实际词义在领域上是接近的。表4.10给出了 WiC 数据集中的一些示例样本。

表 4.10 WiC-TSV 数据集的示例样本（目标词用斜体表示，语义匹配/不匹配的样本标记为 T/F）

数据划分	语句	预期词义	上位词	标签
WNT/WKT	<i>Smoking</i> is permitted .	the act of smoking tobacco or other substances	breathing, respiration, ventilation	T
	All that work went down the <i>sewer</i> .	someone who sews	needleworker	F
CTL	We started the evening with <i>Bellini</i> , made with fresh , Niagara peaches .	A Bellini cocktail is a mixture of Prosecco sparkling wine and peach purée	cocktail	T
	After a morning 's work I went off to see the <i>Bellini</i> retrospective at the Quirinale .	A Bellini cocktail is a mixture of Prosecco sparkling wine and peach purée	cocktail	F
MSH	Italy now reports the second highest number of <i>corona</i> cases worldwide .	A viral disorder characterized by high fever, ... and other symptoms of a viral pneumonia.	viral pneumonia; coronavirus infection	T
	<i>Corona</i> Labs is happy to announce the general availability of the public beta of Android 64-bit Corona builds .	A viral disorder characterized by high fever, ... and other symptoms of a viral pneumonia.	viral pneumonia; coronavirus infection	F
CPS	pandas is an open source data analysis and manipulation tool built on top of the <i>Python</i> programming language .	Python is an interpreted, high-level, general-purpose programming language	programming language	T
	The present paper compares the recently studied <i>pythons</i> with those examined 20 years ago , and uses the combined dataset to assess the ecological sustainability .	Python is an interpreted, high-level, general-purpose programming language	programming language	F

WiC-TSV 的原始语料同样来自 WordNet 和 Wiktionary，构造的训练和验证集包含通用域的样本，而测试集包含通用域和专用域的样本。具体而言，WNT/WKT 测试集包含通用域的样本，而 Cocktails (CTL)，Medical Subjects (MSH) 和 Computer Science (CPS) 测试集分别包括酒饮、医疗和计算机科学领域的测试样本。

## 4.5 语义角色标注

语义角色标注（Semantic Role Labeling, SRL）是一种浅层语义分析技术，目标是分析句子的谓词-论元结构，揭示句子中概念范畴之间的语义关系。语义角色标注的主要语言学理论来源于题元理论（Thematic Theory）、格语法（Case Grammar）以及配价理论（Valency Theory）等句子语义理论等。题元理论认为句子以谓语为中心，谓语决定了句子的基本结构。论元（Argument）是谓语所涉及的对象，担任了施事、客体、受事、地点或命题等不同的题元角色。语义角色标注任务核心是识别句子中谓语的论元，并确定论元的题元角色。

例如：[中国成飞公司]<sub>A0</sub>[正在]<sub>AM-TMP</sub>[制造]<sub>V</sub>[民用飞机]<sub>A1</sub>。

“制造”为谓词（V），代表了一个事件的核心行为；“中国成飞公司”和“民用飞机”为动作的施事者（A0）和受事者（A1）。

在语义角色标注任务研究早期，相关算法往往依赖句子的句法结构。近年来，得益于机器学习和深度学习方法的不断发展，不依赖句法结构信息的语义角色标注方法研究也逐渐兴起。语义角色标注算法虽然有很多类型，但是其基本基本流程都主要由论元识别和论元标注组成。基于句法分析的语义角色标注算法还需要先对句子进行句法分析。论元识别的目标是从句子识别所有由连续几个单词组成的论元。由于如果将句子中所有的连续单词片段都作为论元候选，其数量会过于庞大，因此早期的方法在进行论元识别前，通常还会引入基于规则的候选论元过滤方法，利用句法分析结果构造启发式规则对候选项进行大幅度删减。论元标注则是对论元和谓词之间的关系类型进行标注。论元识别和论元分类通常采用有监督机器学习算法，将上述任务转换为分类问题。两个任务之间可以采用流水线结构，也可以采用联合学习的方法。

本节将介绍常见的基于句法树和基于深度神经网络的语义角色标注算法。

### 4.5.1 基于句法树的语义角色标注方法

句法结构主要有成分结构和依存结构两大类。因此，依赖句法结构的语义角色标注算法可以进一步细分为：基于成分结构的语义角色标注（Span-Based SRL）和基于依存形式的语义角色标注（Dependency-Based SRL）。本节将针对上述两类方法分别进行介绍。

#### 1. 基于成分句法树的语义角色标注方法

在基于成分结构的语义角色标注中，模型基于句子的成分句法分析结果，对句中论元短语对应的跨度进行语义成分标注。例如，针对句子“中国成飞公司正在制造民用飞机”，可以得到如图4.15所示成分句法分析结果，模型的目标是根据句子的句法成分标注，将名词短语“中国成飞公司”，副词“正在”和名词短语“民用飞机”识别为谓词“制造”的施事者（A0）、受事者（A1），以及表示时间关系的修饰语（AM-TMP）。

基于成分句法树的语义角色标注方法，通过对成分句法树进行剪枝，从句子中初步识别候选论元，供后续的论元识别、论元标注步骤使用。该方法的主要思想是考察句子中与谓词短语并列的

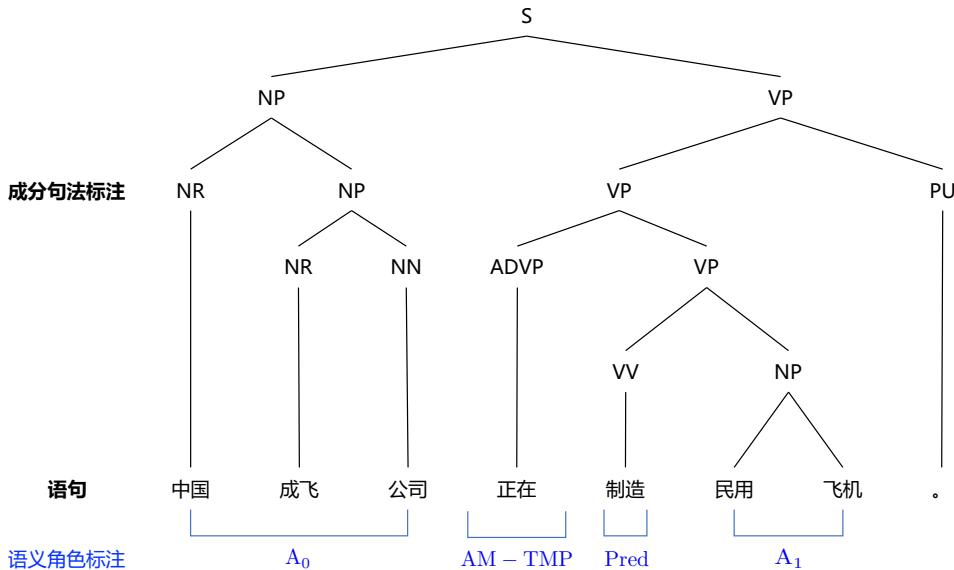


图 4.15 基于成分结构的语义角色标注

成分，筛选符合条件的句子成分作为候选论元。具体而言，此方法从成分句法树的谓词节点开始，考察该节点的每个兄弟节点；如果兄弟节点和该节点在句法结构上不是并列关系，则将兄弟节点加入候选论元集合；如果兄弟节点是介词短语（PP），则将兄弟节点的全体子节点加入候选论元集合。依次对谓词节点的父节点等每个祖先节点执行上述过程，直至到达根节点为止。以图4.16为例，自谓语（VV）“制造”开始，此方法逐次考察包含此谓语的谓语短语（VP），即“制造民用飞机”等成分。在此过程中，此方法将“民用飞机”、“正在”、“。”和“中国成飞公司”加入候选论元集合，并过滤掉大量不可能是论元的成分。

在上述筛选过程后，训练分类模型从候选论元集合中识别真正的论元，并标注论元类型。在此过程中，通常需要为分类器构造有效的特征，常用特征可以分为以下类别<sup>[38]</sup>：

- 谓词及相关特征：谓词，谓词的语态，或论元和谓词出现的前后关系等。
- 论元的词特征：论元的中心词及其词性，以及头尾单词等。
- 基于成分句法标注的特征：论元的成分类型，树中论元到谓词的路径，成分的父亲、兄弟节点类型等。

在上述特征的基础上，可以利用最大熵分类器、SVM、感知机等有监督机器学习方法构建语义角色标注算法。

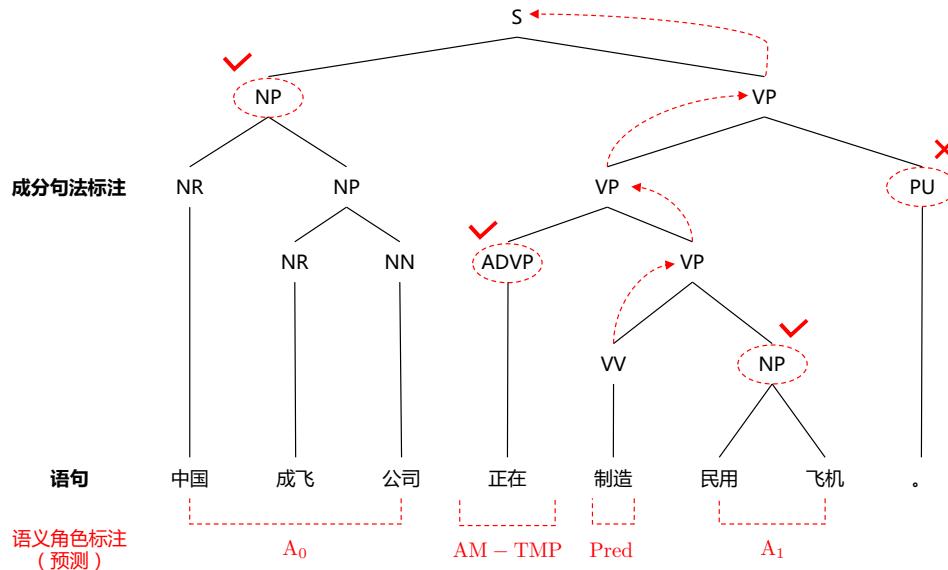


图 4.16 基于成分句法树的语义角色标注方法

## 2. 基于依存关系树的语义角色标注方法

基于依存的语义角色标注算法根据句子依存树进行语义角色标注。如图4.17所示，给定句子“中国成飞公司正在制造民用飞机”及其依存句法树，样本中标注谓词-论元关系，表示谓词与论元的中心词之间的语法关系。在依存句法树中，每个论元自身内部的语法结构由依存关系展示，而论元和谓词的语法关系体现为论元中心词和谓词之间的依存关系。对于论元“民用飞机”和谓词“制造”，两者之间的语法关系通过从“制造”指向“飞机”的边，体现为宾语（OBJ）关系。

由于依存形式的语义角色标注把语义角色表示成谓词和论元中心词之间的语义关系，和依存句法标注完全对应，所以可以采用类似上节所述的剪枝方法，识别句子中潜在的论元。基于依存句法树的语义角色标注方法将上节所述的候选论元筛选过程迁移到依存句法树上。首先，从谓词节点开始，将当前节点的全体子节点加入候选论元集合；然后将当前节点的父节点作为当前节点，重复上述过程，逐次考察谓词节点的祖先节点；至当前节点作为句子的根节点为止。如图4.18所示，谓语“制造”分别指向“公司”、“正在”、“飞机”和句尾的标点符号，对应的论元“中国成飞公司”、“正在”和“民用飞机”将被识别为候选论元。

针对后续的论元识别、论元标注阶段，基于依存句法树的语义角色标注方法将其建模为判断谓词和论元中心词之间语义关系的任务，并建立分类模型来解决。在此过程中常用的分类特征包括以下几类<sup>[38]</sup>：

- 谓词及相关特征：谓词，谓词的词根、词义、词性、语态，或论元和谓词出现的前后关系等。

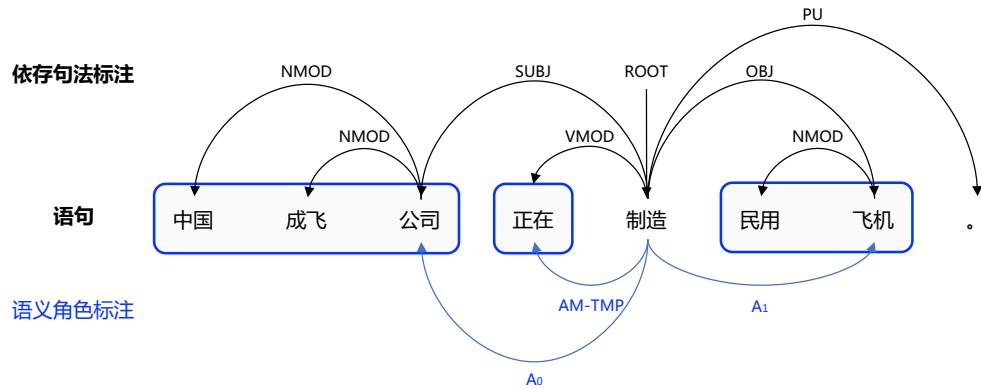


图 4.17 基于依存句法的语义角色标注

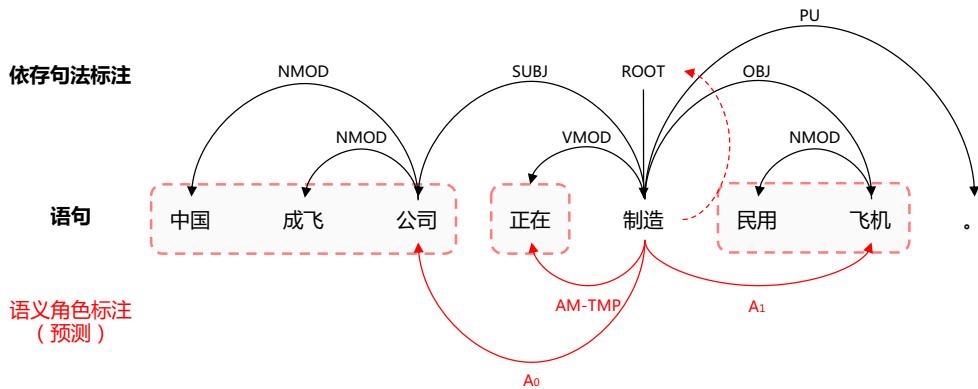


图 4.18 基于依存句法树的语义角色标注方法

- 论元的词特征：论元的中心词及其词性，以及头尾单词等。
- 基于成分句法标注的特征：树中论元中心词到谓词的路径，谓词与其父节点的依存关系，以及其父节点的相关信息；谓词与其子节点的依存关系；候选论元中心词的子节点、兄弟节点相关信息等。

同样，在上述特征的基础上，可以利用最大熵分类器、SVM、感知机等有监督机器学习方法构建语义角色标注算法。

#### 4.5.2 基于深度神经网络的语义角色标注

在深度学习应用到自然语言处理领域后，由于模型能够自动学习到多层次的特征表示，不依赖句法标注、而直接对文本进行表示学习的方法也能达到较好的效果，并逐渐得到重视。针对语

义角色标注而言，可以用 BIO 标注方案表示论元标签，从而可以直接利用通用的序列标注模型来解决；也可以以跨度标注句子中的论元短语位置，采用基于跨度预测的方法。由于跨度预测模型显式地建模了句子中短语级别的语义，模型可以更好地学习论元短语的长程邻接关系。因此，大量语义角色标注系统采用跨度预测的形式进行语义角色标注，并取得良好的效果。此外，由于句法结构信息为语义角色标注任务提供了丰富的语言学信息，因此可有效利用句法树结构的图神经网络也在该任务上取得了不错的成绩。本节将分别介绍上述两类深度学习方法。

### 1. 基于跨度的语义角色标注方法

文献 [198] 中通过端到端模型构建了跨度预测模型，为语句中的每个词和跨度构造表示，实现同时识别句子中的谓词和论元并判断它们之间关系的效果。

该模型分为两个部分，词和跨度表示的构建以及谓词-论元的联合抽取。模型结构如图4.19表示，跨度预测模型包括一个编码器，用于为输入序列的每个标记构建表示。在文献 [198]，使用双向 LSTM 作为文本序列的编码器，编码器在每个位置的输入是单词的 GloVe 向量，以及通过 CharCNN 提取的字符级特征。针对每个单词  $w_i$ ，其嵌入表示为  $x_i = [\text{WORDEMB}(w_i); \text{CHARCNN}(w_i)]$

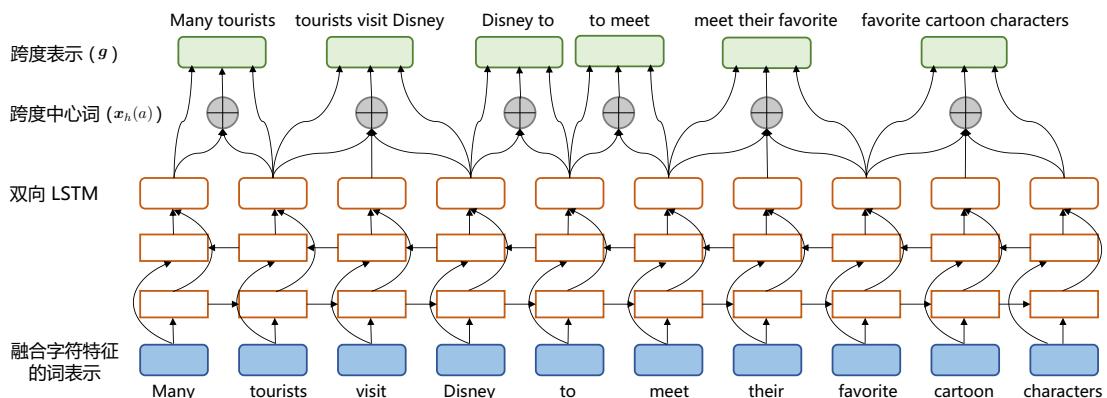


图 4.19 基于跨度预测的语义角色标注模型中跨度表示神经网络架构<sup>[198]</sup>

编码过程可以形式化的表示为：

$$\mathbf{v}_1, \dots, \mathbf{v}_n = \text{Encoder}(w_1, \dots, w_n) \quad (4.65)$$

其中  $s = w_1, \dots, w_n$  是输入序列， $\mathbf{v}_1, \dots, \mathbf{v}_n$  是对应位置的表示。

模型的词表示直接使用编码器在对应单词位置的输出，若潜在的谓词  $p = w_i$  是句子中的第  $i$  个单词，则  $g(p) = \mathbf{v}_i$ 。跨度表示通过四部分构成，以  $a = (w_i, \dots, w_j)$  为例，分别是跨度内头尾

标记的表示  $\mathbf{v}_i, \mathbf{v}_j$ 、中心词特征  $\mathbf{x}^h(a)$  和跨度范围特征  $\phi(a)$ ，公式如下所示：

$$g(a) = [\mathbf{v}_i, \mathbf{v}_j, \mathbf{x}^h(a), \phi(a)] \quad (4.66)$$

其中，中心词特征使用注意力机制建模了跨度内词语的重要程度，捕获句子中重要词在跨度内出现的信息。对于跨度  $a = (w_i, \dots, w_j)$ ，中心词表示的计算如下所示：

$$e(a) = \text{softmax}(\mathbf{W}^e[\mathbf{v}_i; \dots; \mathbf{v}_j]) \quad (4.67)$$

$$\mathbf{x}^h(a) = [\mathbf{x}_i; \dots; \mathbf{x}_j] e(s)^T \quad (4.68)$$

其中， $\mathbf{W}^e$  是可学习参数，用于衡量单词的重要性； $e(s)$  是句子  $s$  中每个词的注意力分数， $\mathbf{x}_i, \dots, \mathbf{x}_j$  是单词  $w_i, \dots, w_j$  经过编码器之前的原始表示，由它们计算中心词特征。

跨度范围特征  $\phi(a)$  仅和跨度的长度有关，通常按照跨度的出现频率将跨度长度分为若干个范围，为每个范围训练一个表示向量作为范围特征。在一般的语义角色标注任务中，可以取 [1, 2, 3, 4, 5-7, 8-15, 16-31, 32-63, 64+] 作为跨度长度的区分范围。

在谓词-论元联合抽取部分，通过计算谓词-论元匹配分数  $\Phi(l, a, p)$  来识别谓词和论元，以及论元相对谓词的语义角色，其神经网络架构如图4.20所示。谓词-论元匹配分数主要通过组合谓词和论元的表示来计算：

$$\Phi_a(a) = \mathbf{W}^a \text{MLP}^a(g(a)) \quad (4.69)$$

$$\Phi_p(p) = \mathbf{W}^p \text{MLP}^p(g(p)) \quad (4.70)$$

$$\Phi_{rel}(l, a, p) = \mathbf{W}_l^r \text{MLP}^r([g(a); g(p)]) \quad (4.71)$$

$$\Phi(l, a, p) = \Phi_a(a) + \Phi_p(p) + \Phi_{rel}(l, a, p) \quad (4.72)$$

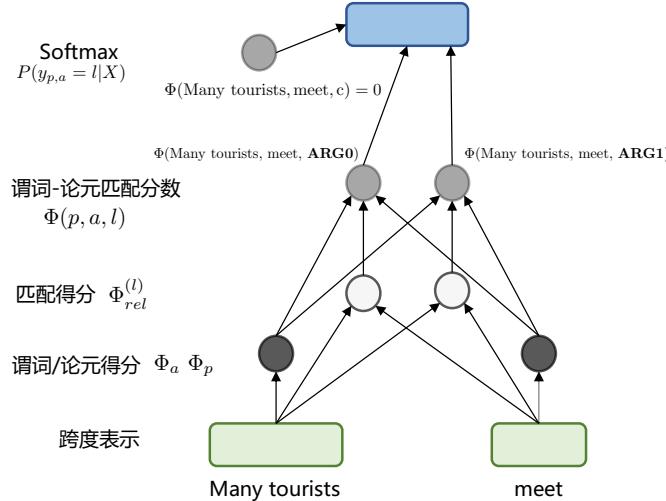
在句子长度为  $n$  时，潜在谓词和论元的个数分别为  $\mathcal{O}(n)$  和  $\mathcal{O}(n^2)$  个，如果要确定每个候选谓词和候选论元之间的语义角色关系，涉及的计算量为  $\mathcal{O}(n^3|\mathcal{L}|)$ ，其中  $\mathcal{L}$  是标签集合。因此，使用束剪枝 (Beam pruning) 方法降低计算量，在计算谓词-论元匹配分数时，首先考察  $\Phi_a(a)$  和  $\Phi_p(p)$ ，只保留分数较高的  $\lambda_a n$  个论元和  $\lambda_p n$  个谓词进行语义关系标签的计算，从而将计算量降至  $\mathcal{O}(n^2|\mathcal{L}|)$ 。

对于在剪枝中保留下来的谓词-论元对，计算匹配分数后，根据匹配分数的 Softmax 预测语义标签。在训练时，以负对数概率作为目标函数，最大化标准答案的出现概率。

$$P(y_{p,a} = l | X) = \text{softmax}(\Phi(l, a, p)) \quad (4.73)$$

$$P(Y|X) = \prod_{p,a} P(y_{p,a}|X) \quad (4.74)$$

$$\mathcal{L} = -\log P(Y^*|X) \quad (4.75)$$

图 4.20 基于跨域预测的语义角色标注模型中谓词-论元联合抽取神经网络架构<sup>[198]</sup>

## 2. 基于图卷积神经网络融合句法树的语义角色标注方法

语义角色标注与语法树有紧密的联系，如果能够在深度学习模型中融入语法树结构，可以提供更丰富的信息。文献 [199] 提出了一种基于图网络的方法，可以有效融入依存句法树信息。该方法设计了一种结合门控机制的图网络，用于输入语句和句法树的编码。模型结构如图4.21所示。在传统的双向 LSTM 编码的基础上，该模型将 LSTM 编码层的输出接入一个图网络，通过在图网络中嵌入依存关系树，使语句编码获得语法信息。

具体地，记  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  为句法树对应的图，其中节点集合  $\mathcal{V}$  由句子中的每个词语组成，边集合  $\mathcal{E}$  对应句法树中的依存弧。若边  $(u, v) \in \mathcal{E}$ ，记  $L(u, v)$  为边  $(u, v)$  的类别标签， $D(u, v)$  为边  $(u, v)$  的方向标签。方向标签的取值有三种，分别对应自指关系，以及依存句法树中的依存关系（双向）。若在依存关系树中存在由  $u$  指向  $v$  的弧，则  $(u, v) \in \mathcal{E}$  且  $D(u, v) = 0$ ， $(v, u) \in \mathcal{E}$  且  $D(v, u) = 1$ 。另外，对于  $\forall v \in \mathcal{V}$ ， $(v, v) \in \mathcal{E}$  且  $D(v, v) = 2$ 。类别标签包括方向信息和依存关系类型。易发现， $L(u, v)$  和  $D(u, v)$  分别对应边的细粒度和粗粒度类型标签。

在图网络的第  $k$  层，对于每个节点  $v$ ，通过  $v$  的邻接节点来更新  $v$  的表示向量：

$$\mathbf{h}_v^{(k+1)} = \text{ReLU} \left( \sum_{u \in N(v)} \mathbf{g}_{v,u}^{(k)} (\mathbf{W}_{D(u,v)}^{(k)} \mathbf{h}_u^{(k)} + \mathbf{b}_{L(u,v)}^{(k)}) \right) \quad (4.76)$$

其中， $\mathbf{h}_v^{(k+1)}$  是节点  $v$  在第  $k+1$  层的表示， $N(v)$  是和  $v$  邻接的节点集合， $\mathbf{g}_{v,u}$  是衡量  $u$  对  $v$  的重要性权重， $\mathbf{W}, \mathbf{b}$  是线性层的可学习权重，前者仅利用边的方向标签关系，从而大幅度减少了模

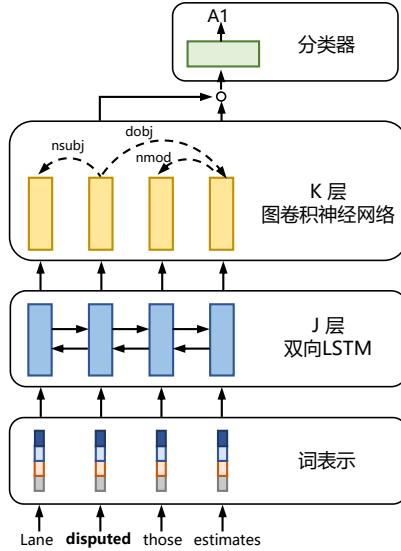


图 4.21 融合句法树的语义角色标注神经网络模型<sup>[199]</sup>

型的参数规模。

门控机制通过为节点  $v$  的每个邻接节点  $u$  赋予权重，用来排除句法树中和谓词-论元关系无关的依存弧，也起到对依存句法分析结果进行去噪的效果。 $v$  对  $u$  的重要性权重按照如下公式进行计算：

$$\mathbf{g}_{u,v}^{(k)} = \sigma \left( \mathbf{W}'_{D(u,v)} \mathbf{h}_u^{(k)} + \mathbf{b}'_{L(u,v)} \right) \quad (4.77)$$

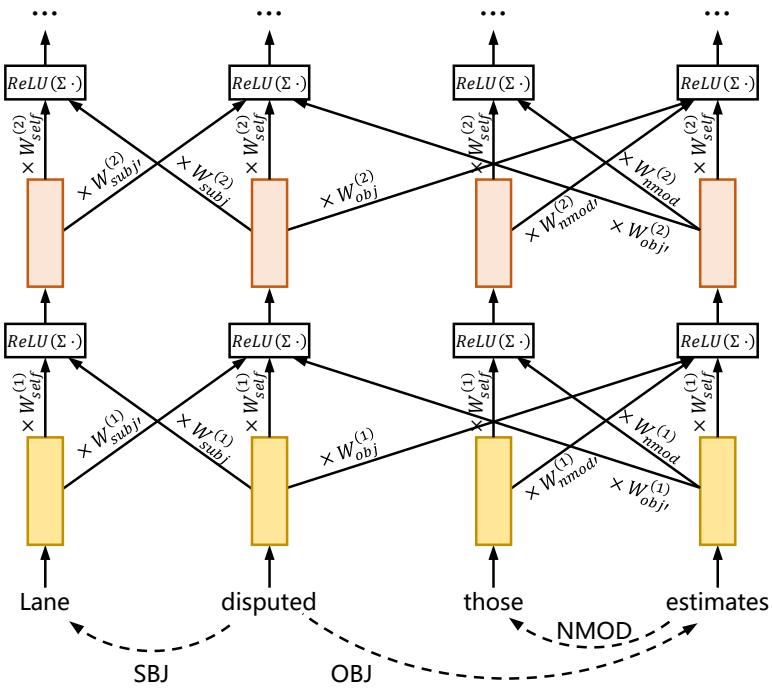
其中， $\mathbf{W}', \mathbf{b}'$  是线性层的可学习权重。整体网络结构如图4.22所示。

在模型的整体训练过程中，输入序列首先经过双向 LSTM 编码，将编码结果作为图网络的节点初始表示和句法树信息同时输入图网络。以图网络的输出作为句子中每个词的表示，该模型将数据标签以序列标注的形式进行编码，每个单词的类标签分布概率预测如下所示：

$$p(r|\mathbf{t}_i, \mathbf{t}_p, l) \propto \exp(\mathbf{W}_{l,r}(\mathbf{t}_i \circ \mathbf{t}_p)) \quad (4.78)$$

$$\mathbf{W}_{l,r} = \text{ReLU}(\mathbf{U}(\mathbf{q}_l \circ \mathbf{q}_r)) \quad (4.79)$$

其中  $\mathbf{t}_i, \mathbf{t}_p$  是句子中第  $i$  个单词和谓词的输出表示， $r$  是语义角色标签， $l$  是谓词的词干， $\mathbf{U}, \mathbf{q}$  是可学习的权重。

图 4.22 基于图网络的句法信息融合<sup>[199]</sup>

### 4.5.3 语义角色标注评价方法

语义角色标注的评价指标包括精确率 (Precision, P)、召回率 (Recall, R) 和 F 值 (F-Score) 得分。在预测结果中，仅当论元范围和类型均预测正确时，才视为该论元预测正确，计为真正例 (True-Positives, TP)。样本中标注，但未被模型预测正确的论元计为假反例 (False-Negatives, FN)。预测结果中出现，但无法与样本标注对应的论元预测结果计为假正例 (False-Positives, FP)。语义角色标注的精确率为预测正确的论元数和预测结果中总的论元数之比值，即  $P = TP / (TP + FP)$ 。召回率为预测正确的论元数和样本中标记的论元数之比值，即  $R = TP / (TP + FN)$ 。F1 得分为准确率、召回率的调和平均值， $F = 2PR / (P + R)$ 。

### 4.5.4 语义角色标注语料库

语义角色标注所采用的语义理论不同，比如题元理论中称为论元或题元，而格语法中则称为语义格，因此形成了多种不同的标注数据库，主要包含三个类型：FrameNet<sup>[167]</sup>、PropBank<sup>[200]</sup>、NomBank<sup>[201]</sup>。本节中将首先介绍上述三类语义角色标注语料库，在此基础上介绍基于成分句法和依存句法的语义角色标注评测集合。

## 1. 语义角色标注语料库

框架网络（FrameNet）根据框架语义学理论，对英国国家语料库进行标注。框架语义学理论提出了语义框架的概念，认为词语的语义反映为人类理解语言时在大脑中激活的认知结构，即语义框架。每个词语对应的语义框架由不同类型、数量的框架元素组成，用来体现和区分词语的语义与功能。在 FrameNet 中，语义标注主要以框架的形式描述谓词的语义，并试图描述框架之间的关系。除主流的动词谓词外，FrameNet 也包括部分名词以及形容词谓词标注。FrameNet 目前包括超过 200,000 个人工标注的句子，在句子中标注目标谓词的语义框架、语义角色，以及每个语义角色在句法层面的短语类型和句法功能。此外，FrameNet 目前包含超过 1200 个语义框架，涵盖了动词及部分形容词和名词。FrameNet 的发布机构认为，基于 FrameNet 的语义角色标注能够在信息提取、机器翻译、事件识别、情感分析等领域起到帮助。图4.23给出了 FrameNet 的数据样例。

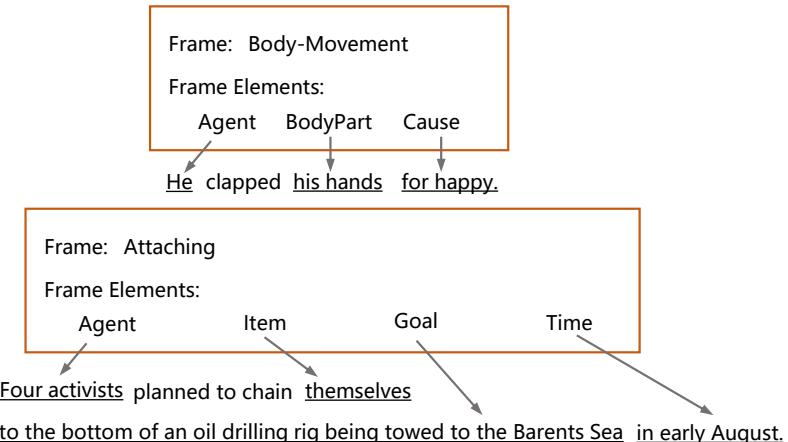


图 4.23 FrameNet 数据样例

命题库（Propositional Bank, PropBank）是论元角色语义知识库，针对动词性谓词，在语料中标注所在句的谓词-论元（predicate-argument）信息。PropBank 是基于英文宾州树库（Penn Tree Bank）标注，在英文宾州树库的句法结构标注基础上，标注谓词的论元及其语义角色。PropBank 对英文宾州树库中的 WSJ（华尔街日报标注）语料和一部分 BROWN（布朗语料库标注）语料进行了标注。

PropBank 定义了四大类的语义角色<sup>[202]</sup>：

- 核心语义角色：标记为 A0-A5 六种，如 A0 通常表示动作的施事者，A1 一般表示动作的影响，A2-A5 根据谓语动词有不同的语义含义。
- 修饰作用的附加语义角色：其角色标签以 AM 开头，常见的 15 种有：ADV（附加的，默认

标记)、BNE (受益人)、CND (条件)、DIR (方向)、DGR (程度)、EXT (扩展)、FRQ (频率)、LOC (地点)、MNR (方式)、PRP (目的或原因)、TMP (时间)、TPC (主题)、CRD (并列参数)、PRD (谓语动词)、PSR (持有者) 和 PSE (被持有)。

- 参考语义角色：其角色标签以 R 开头，表示出现在句子中的其他论元。
- 动词：其角色标签为 V，表示句子中的谓语动词。

名词命题库 (NomBank) 关注并侧重名词性谓词的语义角色标注，对 PropBank 的涵盖范围进行了补充。在标注句子中的谓词时，PropBank 是基于动词词典进行标注的，只考虑动词性谓词，而未涉及语料中的名词性谓词。针对这项不足，NomBank 对语料中的名词性谓词进行语义角色标注。其主要语料资源同样来自《华尔街日报》，研究人员对其中名词性的谓词对应的论元角色进行人工标注。在 NomBank 的标注过程中，标注者尽可能地使角色定义在词性之间保持一致。例如，NomBank 在名词“decision”的标注中依然使用 PropBank 对动词“decide”的标注框架文件。图4.24给出了 NomBank 的数据样本示例。在名词短语“John’ s replacement Ben”和“Ben’ s replacement of John”中，名词 replacement 是谓词，Ben 是 Arg0，表示替代者；John 是 Arg1，表示被替代者。

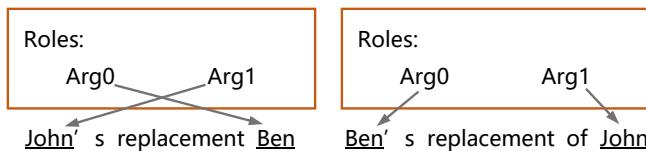


图 4.24 NomBank 数据样例

NomBank 对名词性谓词的语义解析是对该领域语言资源的一项重要补充。根据标注者的统计，对于 NomBank 所标注的伴随论元结构的名词，其中约有一半是名词化的，或者具有类似名词化的性质。例如，aggression 和 agenda 具有类似动词 destroy 和 schedule 的论元结构。NomBank 的标注工作揭示了名词论元结构的一系列语言现象，包括支持动词结构、跨系论元、和括号内的 PP 结构等。由于名词性谓词的论元可以出现在以该名词为首的 NP 之外，这些语言现象都具有较高的研究价值。

## 2. 基于成分结构的语义角色标注评测

在基于成分结构的语义角色标注中，常用的评测基准是 CoNLL05 和 CoNLL12 数据集，其标注格式如4.5.1节图4.15所示。CoNLL05 数据集主要由英文宾州树库的 WSJ 部分组成，语料来自《华尔街日报》的新闻报道，数据标注包括宾州树库的句法树标注，以及从 PropBank 中提取的谓词-参数结构信息。在该基准中，遵循句法解析中使用的标准分区，WSJ 语料的第 02-21 节用作训练集，第 24 节用作验证集，第 23 节用作测试集。此外，CoNLL05 的测试集还包括一个领域外数据集，语料来自 Brown 语料库的 ck01-03 三个部分，类型为小说文本，用于测试模型对领域外样

本的泛化能力。

在 OntoNotes 语料库提出后,由于其规模大并且涉及多个领域,还具有多种语言等特点,CoNLL 基于该语料库提出了语义角色标注的 CoNLL12 基准。CoNLL12 的标注形式和标注方式和 CoNLL05 接近,主要的改进体现在更大的规模和更广泛的语料来源上。另外,CoNLL12 包含多种语言的评测基准。表4.11给出了 CoNLL05 及 CoNLL12 评测基准的统计数据,包括 CoNLL12 的中文评测基准<sup>[203]</sup>。

表 4.11 CoNLL05、CoNLL12 评测基准的统计数据

数据集 数据划分	CoNLL 2005				CoNLL 2012 ( 英文 )			CoNLL 2012 ( 中文 )		
	训练集	验证集	测试集	Brown	训练集	验证集	测试集	训练集	验证集	测试集
句子个数	39.8k	1.3k	2.4k	0.4k	75.2k	9.6k	9.5k	36.5k	6.1k	4.5k
谓词个数	90.8k	3.2k	5.3k	0.8k	188.9k	23.9k	24.5k	117.1k	16.6k	15.0k
论元个数	333.7k	11.7k	19.6k	3.0k	622.5k	78.1k	80.2k	365.3k	51.0k	46.7k

### 3. 基于依存结构的语义角色标注评测

CoNLL09 是目前常用于依存形式语义角色标注的评测基准。CoNLL09 数据集包含来自不同语系的加泰罗尼亚语、汉语、捷克语、英语、德语、日语和西班牙语等 7 种语言,用于评估系统在给定谓词的情况下,进行谓词消歧、论元识别和论元分类的能力。CoNLL09 的英语部分主要基于 PropBank 和 NomBank 的标注,包含动词性和名词性谓语,训练集和验证集分别包含 39.3k 和 2.4k 个句子。汉语部分主要基于汉语树库和汉语命题库的标注,训练集和验证集分别包含 22.3k 和 2.6k 个句子,具体的统计数据如表4.12所示<sup>[204, 205]</sup>。由于语言内容的丰富性,CoNLL09 经常用于评估语义角色标注模型的多语言扩展能力。另外,由于 CoNLL09 在捷克语、英语和德语上准备了领域外测试集,其也经常用于测试语义角色标注模型的泛化性能。

表 4.12 CoNLL09 评测基准的统计数据

数据集	训练集句数	训练集词数	平均句长	谓词比例	验证集句数	验证集词数
CoNLL 2009 ( 英文 )	39.3k	958.2k	24.4	18.7%	2.4k	57.7k
CoNLL 2009 ( 中文 )	22.3k	609.1k	27.3	16.9%	2.6k	73.2k

## 4.6 延伸阅读

本章首先介绍了语义学的基本概念和研究范畴,随后介绍了自然语言的语义表示方法,并介绍了自然语言处理中应用语义分析的词义消歧和语义角色标注任务。近年来,随着深度学习的应用和语言资源的逐步完善,语义表示和语义分析获得了很大的发展。然而,该领域依然存在问题和挑战。一方面,大部分的语义分析任务还没有彻底得到解决,尤其是在资源受限的情境下,模

型的表现距离人类有明显差距。另一方面，大规模预训练模型在自然语言领域取得了巨大的成功，如何通过融合语义分析的知识，预训练语言模型的性能可以得到进一步地改进。本节中将针对上述两个方面当前工作进行简单介绍。

在自然语言处理的实际应用场景中，数据往往是稀缺的，需要利用额外的数据和模型来提升任务效果。一系列工作通过少样本/零样本学习和迁移学习等方法，提升模型在低资源情况下的表现。在少样本/零样本情境下，文献<sup>[206]</sup> 通过结合上下文语法结构信息进行罕见词词义学习。文献<sup>[207]</sup> 针对零样本的语义角色标注任务，提出基于论元识别-角色聚类范的神经网络模型。在跨语言语义分析情境下，文献<sup>[208]</sup> 基于高资源语言向低资源语言迁移的方式，利用依存关系标注来提升低资源语言上语义角色标注的性能。文献<sup>[209]</sup> 针对跨语言语义角色标注任务，为全体语言建立了统一的模型编码器和训练目标，在低资源语言上获得良好的效果。

随着硬件计算资源和互联网大数据的发展，大规模预训练模型在自然语言领域取得了巨大的成功，以丰富的方式被应用在各种各样的具体任务和实际场景中。通过融合语义分析的知识，预训练语言模型的性能可以得到进一步地改进。部分工作针对预训练语言模型的原始语法分析能力进行了讨论<sup>[210-214]</sup>。在此基础上，一些工作修改模型的原始结构，使其显式地接收语法结构输入并进行学习<sup>[215]</sup>；另外一些工作让模型在语法结构任务上进行自适应地预训练，从而隐式地提升模型的语法分析能力<sup>[216, 217]</sup>。针对跨语言的语义分析，文献<sup>[218]</sup> 基于跨语言的依存关系标注<sup>[219]</sup> 提出一个多任务框架，用来增强多语语言模型的性能。

## 4.7 习题

- (1) 词汇语义关系在自然语言处理的下游任务中有哪些应用？试结合 WordNet 举例说明。
- (2) 如何在深度学习模型中融合句子的谓词逻辑表示式？
- (3) 语言的分布式表示和语言模型之间有什么区别和联系？
- (4) 除了词义相似性和词的类比性之外，词向量还能体现哪些单词的语义性质？试结合 t-SNE 可视化分析举例说明。
- (5) 如何在对话系统中应用词义消歧方法，提升综合性能？
- (6) 试分析本章介绍的语义角色标注模型在少样本、资源受限场景下的弊端，并举例说明少样本场景下语义角色标注的方法。

## 5. 篇章分析

---

到目前为止，本书中讨论的都是词语或句子层面的语言现象。然而，语言通常并不是由独立无关的句子组成，而是由搭配在一起具有一定结构的连贯的句子集合组成。我们将这样的句子集合称为篇章（Discourse）。篇章分析的目的是从整体上理解篇章，其中最重要的是对篇章的连贯性（Coherence）和衔接性（Cohesion）进行分析。连贯性是将真正的篇章区别于无关、随机的句子集合的重要性质，而衔接性帮助我们分析和理解篇章的结构，包括其名词、代词之间的指代关系等。篇章分析在自然语言处理中具有非常重要的作用是摘要生成、阅读理解等篇章级别任务的必要环节。

本章首先介绍篇章分析的基本概念，在此基础上介绍篇章分析的三个子任务：话语分割、话语分析和指代消解的主要算法和语料库。

### 5.1 篇章理论概述

篇章语言学是在二十世纪五十年代以后发展起来的一门新兴学科。传统语言学通常以句子本身及其组成部分为研究对象。但是随着语言学研究的进展，人们发现句子在不同的上下文和语境中也可以有不同的意义或者具有不同交际功能，多个合乎句法的句子也并不是随意堆在一起就能构成一个合格的语篇。越来越多的语言学家开始认识到语言研究应该超越句子层次，句子的组合受到语法以外的规则的制约。Harris (1952) 在《Discourse Analysis》一书中首次提出了“话语分析”这一术语。W. Weinrich 在 1967 年首次提出了“篇章语言学”这一概念，认为任何语言学研究都应该以语篇为描述框架。

篇章（Discourse）也称语篇，是指由一系列连续的语段或句子组成的整体，是语言运用或交际的基本单位。篇章的形式是多种多样的，既包含新闻、小说、论文、报告又包含警示标语、交通标识等。像“禁止通行”这样的警示语以及“停！”之类的有意义的语言单位都可以看作是一个篇章。但是，并不是所有大于句子的单位都可以组成一个合格的篇章。Beaugrande 和 Dressler 在 1981 年所著的《Introduction to Text Linguistics》<sup>[220]</sup> 一书中指出，一个合格的篇章需要满足七个标准：衔接（Cohesion）、连贯（Coherence）、意图性（Intentionality）、可接受性（Acceptability）、信息性（Informativity）、情景性（Situationality）和互文性（Intertextuality）。由此，可以看到篇章与孤立句子的主要区别在于：篇章是由句子组成的前后连贯的、有主题的统一整体。篇章所呈现的特定结

构，不仅包含音、词和句法等表层结构上，也体现在语义连贯的深层结构上。需要注意的是，由于篇章的类型和特点千差万别，一个合格的篇章并不一定完全满足上述所有七个标准。

本节中针对篇章分析中最重要三个方面：衔接、连贯和组织对篇章语言学理论进行简略介绍。

### 5.1.1 篇章的衔接

衔接 (Cohesion) 是指篇章中的某一语言成分需要依赖另一语言成分进行解释<sup>[221]</sup>。衔接是一种语义关系，使得篇章各组成部分在语义上相互联系，关系紧凑。衔接也被作为语篇连贯性的必要条件之一，在语篇中体现为词汇衔接和语法衔接。词汇衔接包括重述 (Reiteration)、搭配 (Collocation) 等衔接手段。语法衔接包括照应 (Reference)、替代 (Substitution)、省略 (Ellipsis)、连接 (Conjunction) 等衔接手段。

#### 1. 词汇衔接

重述关系是通过词的重复、同义词或近义词、反义词、上下位词等词汇手段形成的篇章衔接关系。

例如：

- (1) 苏州园林 据说有一百多处，我到过的不过十多处。其他地方的园林 我也到过一些。倘若要我说说总的印象，我觉得苏州园林是我国各地园林的标本，各地园林或多或少都受到苏州园林的影响。因此，谁如果要鉴赏我国的园林，苏州园林就不该错过。
- (2) 那棵树立在那条路边上已经很久很久了。当那路还只是一条泥泞的小径时，它就立在那里；当路上驶过第一辆汽车之前，它就立在那里；当这一带只有稀稀落落几处老式平房时，它就立在那里。

在上述例子 (1) 中“苏州园林”出现了四次，“我国的园林”、“各地园林”等则通过上下位关系进行衔接。例子 (2) 中，“它就立在那里”出现了三次，与“已经很久很久了”也形成了近义重述关系。

搭配关系是指词的共现关系，包括一个词组或者一个句子内部的词之间的组合关系，也包括句子间或段落间的词的习惯性共现。

例如：

有一天早上，撒了三次网，什么都没捞着，他很不高兴。第四次把网拉拢来的时候，他觉得太重了，简直拉不动。他就脱了衣服跳下水去，把网拖上岸来。打开网一看，发现网里有一个胆形的黄铜瓶，瓶口用锡封着，锡上盖着所罗门的印。

上述例子中，围绕“撒网”展开，形成了一个与撒网打鱼相关的动词链，“撒-捞-拉-拖”够成了词汇衔接。在同一个篇章中，词汇之间通过语义联想构成衔接关系。

#### 2. 语法衔接

照应是指篇章中一个语言成分与另一可以与之相互解释的成分之间的关系，即一个成分作为另一个成分的参照点。

例如：

那只最后从蛋壳里爬出来的小鸭是那么丑陋，他处处挨啄，被排挤，被讪笑，不仅在鸭群中是如此，连在鸡群中也是这样。

这里代词“他”的确切含义是由它所指的对象决定的。本例中“他”是指“最后从蛋壳里爬出来的小鸭”。

照应性（Phoricity）是语言交际过程中一个普遍现象，用来指代篇章中的实体、概念或事件。照应可以分为两种：外指（Exophora）和内指（Endophora）。外指照应是指篇章中的某个成分的参照点不在篇章本身，而是在语境中。内指照应是指语言成分的参照点在篇章上下文中。内指照应又可以进一步细分为回指（Anaphora）和下指（Cataphora）。回指照应是指所指对象位于上文；下指照应是指所指对象位于下文。在词汇语法层面，照应还可以分为人称照应、指示照应和比较照应，分别是指用人称代词、指示代词以及表示比较的形容词副词所表示的照应关系。照应在篇章中具有重要的作用，可以使篇章在结构上更加紧凑，同时在修辞上达到言简意赅的效果。

替代是指用替代形式来取代上文中的某一成分。在篇章中，由于替代形式的意义必须通过所替代的成分才能获取，因而替代起到了衔接的作用，使得替代成分和替代对象所属句子紧密连接。从语法和修辞角度，替代也是避免重复的一种重要语言手段。替代可以进一步细分为名词性替代（Nominal Substitution）、动词性替代（Verbal Substitution）和小句性替代（Clausal Substitution）。

例如：

各式各样的球鞋像装在万花筒里，在她面前转开了：白色的，蓝色的，高筒的，矮帮的，白色带红边的，白色带蓝边的。

上例中“白色的，蓝色的，高筒的，矮帮的，白色带红边的，白色带蓝边”与上一句中“球鞋”构成了紧密的衔接关系，“的”替代了上文中的“球鞋”，属于名词性替代。汉语中“做”和“干”经常用于动词性的替代。“这样”、“这么”等经常用于小句性替代。

省略是指将语言结构中某个成分在句子中去除。虽然省略结构在语法层面不完整，但是并不是不可理解的，并且表达更加精炼。由于省略成分需要从上下文中获取，也使得省略成为了常见的语法衔接手段。

例如：

雨是最寻常的，一下就是三两天。可别恼。看，雨像牛毛，雨像花针，雨像细丝，密密地斜织着，人家屋顶上全笼着一层薄烟。

上例中“像牛毛，像花针，像细丝”前都省略了“雨”，但是很容易理解并且语篇前后衔接，结构紧凑。省略也可以分为名词性省略（Nominal Ellipsis）、动词性省略（Verbal Ellipsis）以及小句性省略（Clausal Ellipsis）。

连接是通过连接成分体现篇章中逻辑关系。从逻辑语义关系类型上，可以细分为三大类：详述（Elaboration）、延伸（Extension）和增强（Enhancement）。详述是对上文内容进一步说明、评论或解释，主要包括同位语和阐明两种情况。延伸是从正面或反面增加新的陈述，包括添加、转折、变换等类型。增强则是指补充额外必要信息，达到加强语义并使其更加完整，包括时空、方式、因

果与条件、话题等条件。

例如：

不必说碧绿的菜畦，光滑的石井栏，高大的皂荚树，紫红的桑葚；也不必说鸣蝉在树叶里长吟，肥胖的黄蜂伏在菜花上，轻捷的叫天子（云雀）忽然从草间直窜向云霄里去了。单是周围的短短的泥墙根一带，就有无限趣味。

上例中通过“不必说”、“也不必说”、“单是”层层递进，使得句子之间紧密连接。

### 5.1.2 篇章的连贯

连贯（Coherence）是指篇章在语义、功能和心理上构成一个整体，围绕同一个主题或意图展开<sup>[222]</sup>。连贯性（Coherent）是衡量篇章质量的重要指标，只有连贯的句子集合才能够形成篇章。这也是篇章与无关的、随机的句子集合区分开的最主要因素。篇章应该同时具有局部连贯性（Local Coherent）和整体连贯性（Global Coherent）。局部连贯性是在微观层面，篇章中前后相连的命题在语义上的联系。整体连贯性是在宏观层面，篇章中的所有命题与篇章主题之间的联系。

篇章局部连贯通常是通过话语序列的语义结构实现。一般来说，话语序列的语义结构表现可以分为外延的（Extensional）和内涵的（Intensional）两种类型。外延的语义结构表示话语序列所表达的事态与真实世界的排序顺序相对应；内涵的语义结构表示话语序列所表达的事态在真实世界中找不到对应。

例如：

- (1) 他点了一份外卖。
- (2) 外卖很快就送到了。

上例中，(1) 表示 (2) 的条件，并且与真实世界中事件存在对应关系，属于外延语义结构。为了达到话语序列在语义上的连贯性，需要与现实世界中的自然顺序相对应，也就是说如果把 (1) 和 (2) 的顺序颠倒过来，那么该话语序列就不再具有语义上的连贯性。

除了现实世界中自然顺序的限制，话语序列的语义结构还受到人们普遍认知规律的制约。人们认识和描述客观世界时通常遵循从一般到特殊、从整体到局部、从大到小，从集合到子集的认知模式。例如：

例如：

单是周围的短短的泥墙根一带，就有无限趣味。油蛉在这里低唱，蟋蟀们在这里弹琴。翻开断砖来，有时会遇见蜈蚣；还有斑蝥，倘若用手指按住它的脊梁，便会啪的一声，从后窍喷出一阵烟雾。

上例符合整体到局部的排列顺序，从对百草园的“无限趣味”开始，再以局部的细节展开，详细描写了由“油蛉”、“蟋蟀”、“斑蝥”所带来的乐趣。这样的顺序与人们一般的认知规律和感知顺序相符合，从而也就更容易让人接受。

局部连贯说明篇章中相邻句子存在联系，但是仅有局部连贯是不够的，篇章在整体上还需要

围绕一个主题展开，既需要具有整体连贯性。整体连贯性对篇章中句子之间的联系施加宏观制约。

例如：

- (1) 对于一个在北平住惯的人，像我，冬天要是不刮风，便觉得是奇迹；济南的冬天是没有风声的。对于一个刚由伦敦回来的人，像我，冬天要能看得见日光，便觉得是怪事；济南的冬天是响晴的。自然，在热带的地方，日光永远是那么毒，响亮的天气，反有点儿叫人害怕。可是，在北方的冬天，而能有温晴的天气，济南真得算个宝地。
- (2) 母亲还从来没有一次给我这么多钱。我也从来没有向母亲一次要过这么多钱。我来到母亲工作的地方，呆呆地将那些母亲扫视一遍，却没有发现我的母亲。背直起来了，我的母亲。转过身来了，我的母亲。褐色的口罩上方，一对眼神疲惫的眼睛吃惊地望着我，我的母亲。

上例中，(1) 在微观层面上前后承接，句子之间的逻辑关系清晰，宏观层面围绕“济南的冬天”主题开展。例(2) 的局部上是连贯的，但是宏观层面上缺乏主题，从而缺乏整体连贯。

篇章的连贯是一个复杂的现象，有些现象不能完全从语义角度解释话语序列的连贯性，还需要从语用角度以及认知角度进行讨论。围绕语篇的连贯也有很多理论和方法，包括从关联理论角度对微观层面连贯性研究，利用修辞结构理论（Rhetorical Structure Theory）进行语篇连贯性研究，运用图式理论（Schema Theory）的连贯性研究，基于语篇策略（Discourse Strategy）的连贯性研究等。研究篇章的连贯性是自然语言处理中的重要问题，对文本摘要、阅读理解、机器翻译等篇章级别的任务都具有重要的作用。

### 5.1.3 篇章的结构

篇章同时具有线性结构和等级结构。篇章中的句子按照一定的线性规则排列在一起，因此篇章是线性的。同时，句子的组合可以构成更大的语言单位，因此篇章又是具有等级结构的。

例如：

[1] 没有春节不是流动的，也没有春节不是走动的。[2] 这是以往中国人过春节的常态，热热闹闹、走亲串户、朋友相聚，动起来的春节被视为祥和、欢乐的时节。

[2] 然而，这个春节，真的不一样。[3] 一个现实原因就是，新型冠状病毒引发的疫情还在持续，全国人民为此揪心。[4] 应该以什么样的状态与心态，过好这个春节，值得我们细细思量。[5] 春节的流动、拜年的走动、庙会的人头攒动，这些人们已经习惯了的过年方式，在这些日子里恐怕需要改一改了。

[6] 此时，“动”的年节莫若“静”的岁月。[7] 人们越是大规模流动，越是大范围聚集，越容易增加疾病传染的概率。[8] 走动起来还是宅上一宅，理性人不难看透其中的得失，既为人也为己。[9] 事实上，不走动也能过好年。[11] 技术发达了，信息拜年、视频祝福、在线聚会，都不失为一种时尚，那些以往通过面对面完成的新春祝福，借助云端就能迅速直抵耳畔、身边，过年礼仪一样也缺不了。

[12] 此时，“动”的脚步莫若“静”的心意。[13] 在抗击疫情的最前沿，各条战线上的“勇士”都已经动起来了，他们为了更多人的生命安全，以这样一种方式过了个“动”的年，是真正的大无畏。[14] 相反，对普通人来说，如无特殊情况，宜静不宜动，什么自驾跨城回家、什么一定上门拜年、什么提前安排好的聚会等等，都不妨在冷静且理性地审视下做个宅男宅女，不远行、不扎堆、少聚会。[15] 现在，最好的祝福是以你我的安全距离为彼此送上健康祝福，最大的心意是以你我的实际行动护佑早日战胜疫情。

上例中，句子[1][2]组成了引论，句子[2]-[5]给出了论点，句子[6]-[11]组成了第一个分论点，其中句子[6]是提出分论点，句子[7]-[11]是论据，句子[12]-[15]组成了第二个分论点，句子[12]是提出分论点，句子[13]-[15]是论据。

本节中将介绍三种常见的篇章结构的表示方法：超级结构（Superstructure）、修辞结构理论（Rhetorical Structure Theory）和语篇模式（Textual Pattern）。

### 1. 篇章超级结构

篇章超级结构（Superstructure）是采用规约化图式结构来表示篇章宏观内容组织形式的一种形式结构。只涉及篇章内容的组织方法，与篇章所表达的具体内容没有直接关系。不同类型的语篇往往具有不同的超级结构。比如，科技论文通常包含标题、摘要、引言、相关工作、方法介绍、实验、结论、参考文献等成分组成。新闻报道一般由概述、故事和结局等结构要素组成。概述包括标题和导语，故事包括情节和背景，结局包括评论和结论<sup>[223]</sup>。新闻篇章的超级结构如图5.1所示。

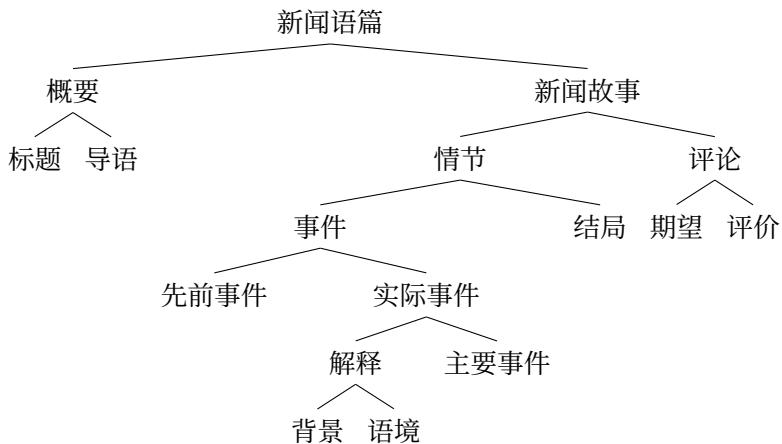


图 5.1 新闻篇章超级结构示例<sup>[223]</sup>

篇章超级结构提供了组织相关类型语篇的基础纲要和架构。在具体的篇章中，结构也具有一定灵活性，并不是所有的结构成分都要存在，结构成分的位置也是不固定的。

## 2. 修辞结构理论

修辞结构理论 (Rhetorical Structure Theory, RST) 是 Mann 和 Thompson 于 1987 年提出的一种通过描述篇章各个组成部分之间的修辞关系来分析篇章结构的理论<sup>[224]</sup>。修辞结构理论将修辞关系定义在两个或多个文本单元 (Text Span) 之间。文本单元又称基本篇章单元 (Elementary Discourse Unit, EDU)，有两种主要类型：核心 (Nucleus) 和辅助 (Satellite)。核心单元是篇章中最重要的部分，表达作者的核心意图，并且具有相对完整的语义，能够独立解释。辅助单元则较少表达作者的核心意图，用于传达支撑其他信息，补充说明核心单元，通常只有在与核心单元关联时才能够被解释。修辞关系通常定义在核心单元和辅助单元之间，也有少部分修辞关系定义在两个或多个核心单元之间。

例如：[1] 这个草莓真的好吃，[2] 我吃了一大盆。

在上例中，小句 [1] 是一个陈述或者判断，小句 [2] 则为这一判断提供了证据。小句 [1] 是核心单元，小句 [2] 是辅助单元，两个小句之间构成证据关系 (Evidence)。该句可以使用如图5.2所示的“证据”关系的图示表示。

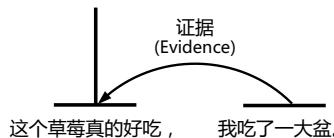


图 5.2 “证据”关系图示样例

根据修辞结构理论，篇章中存在多种多样的修辞结构关系，并且随着研究的不断深入，研究人员也在不断地对修辞关系进行补充。根据文献 [224] 的定义，篇章中的修辞关系主要包括两种类型：(1) 不对称性的核心-辅助关系 (Nucleus-Satellite Relation)，也称单核关系；(2) 无主次之分的多核心关系 (Multinuclear Relation)。图5.3中给出了修辞结构理论中五种图示类型。竖线指示出核心单元，弧线连接具有关系的单元，关系的名称标注在连线上。环境 (Circumstance) 关系是单核心关系，弧线箭头指向核心单元。对比 (Contrast) 关系总是两个核心单元。序列 (Sequence) 关系则可以具有多个连续的单元，相邻的两个单元之间构成序列关系。联合 (Joint) 关系也可以具有多个单元，这些单元一起构成该关系。

篇章修辞关系中绝大多数都是核心-辅助关系，包括：对立 (Antithesis)，动机 (Motivation)、背景 (Background)、析取 (Otherwise)、意图 (Purpose)、总结 (Summary)、评价 (Evaluation)、证据 (Evidence)、使能 (Enablement) 等。无主次修辞关系主要包含对比 (Contrast)、联合 (Joint)、列举 (List) 和序列 (Sequence)。此外，篇章的修辞结构在总体上表现为等级结构，连贯的篇章可以由不同层次的修辞关系组织成层次结构，从而形成一个修辞关系树。图5.4给出了根据修辞结构理论构成的一个修辞结构关系树样例。

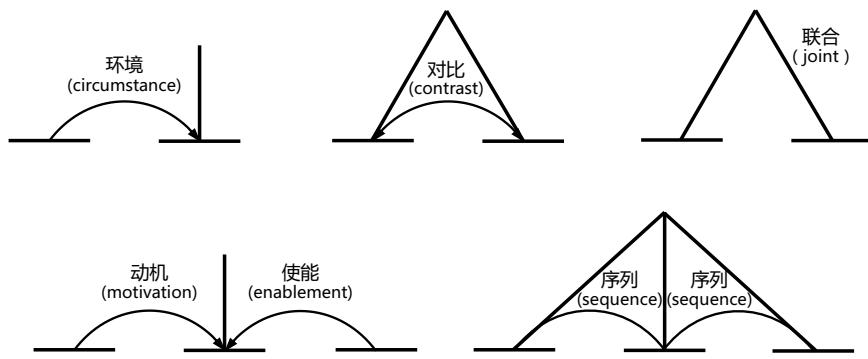


图 5.3 修辞结构理论中关系图示类型

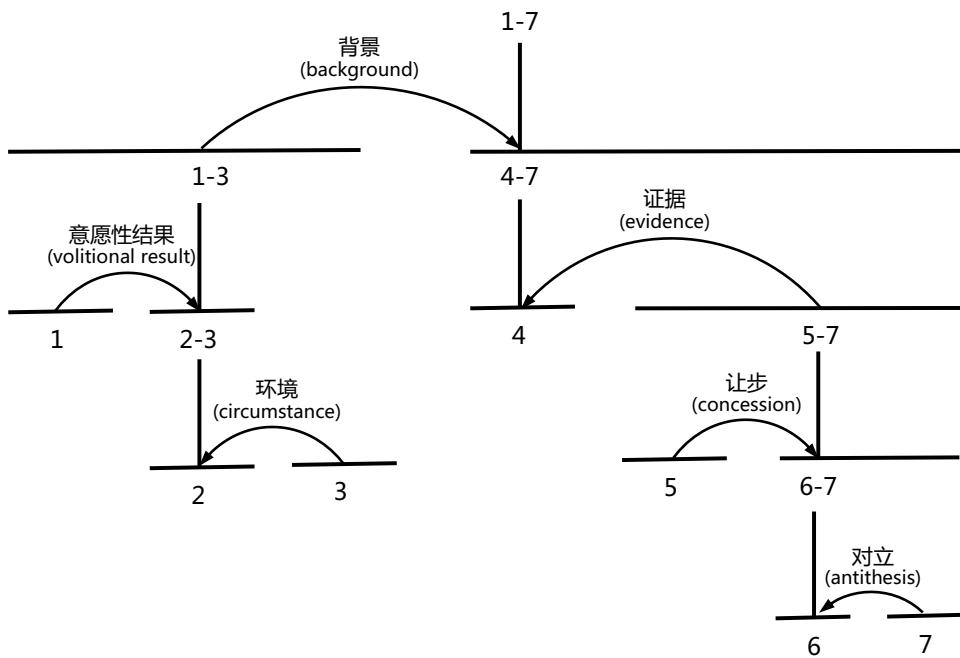


图 5.4 修辞结构关系树样例

### 3. 语篇模式

语篇模式（Textual Pattern）是指人们长期积累并根据经验形成的一些程式化的语篇组织形式或策略<sup>[225]</sup>。语篇模式是在一定的文化中形成的，因此往往带有不同文化积淀的内涵和文化规约性。

语言学家总结出了“问题-解决”(Problem-Solution)、“概括-具体”(General-Specific)等英语中常见的语篇模式。语篇模式与小句关系之间存在着密切的联系，小句通过组合形成逻辑序列关系或匹配关系，通过这些关系小句又组合为更大的语篇单位。语篇模式与具体的篇章内容通常没有直接的联系，但是每种语篇模式通常都具有特定的词汇标记。

以问题-解决模式为例，该模式通常由四个部分构成：情景、问题、反应、评价。这个过程也符合人们通常的认知模式。

例如：

[1] 长征五号遥三运载火箭 27 日晚在海南文昌一飞冲天，将实践二十号卫星成功送入太空预定轨道。[2] “胖五”也以实际行动，诠释着中国俗话所说“哪里跌倒，就要从哪里爬起来”的坚持与坚韧。[3] 2017 年 7 月长征五号遥二火箭因发动机故障发射失利。[4] 科研人员历经两年多的艰苦攻关、连续奋战，进行大量地面试验，完成遥二失利故障归零和遥三火箭各项工作，还采取一系列改进优化措施，切实提升火箭飞行任务可靠性。[5] 长征五号遥三火箭在此背景下成功发射，对研制团队直面挑战、发现问题、解决问题的心理能力建设也是一次巨大考验，也为航天人才特别是青年人才树立起不怕失败、敢于挑战、勇于拼搏的榜样力量。

上例中 [1][2] 句描述了情景，句子 [3] 说明了问题，[4] 句给了反应和解决问题的方法，最后句子 [5] 给出了评价。语篇模式如图 5.5 所示：

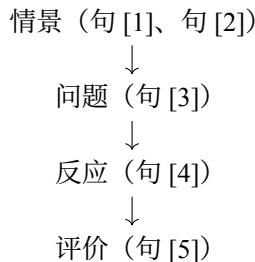
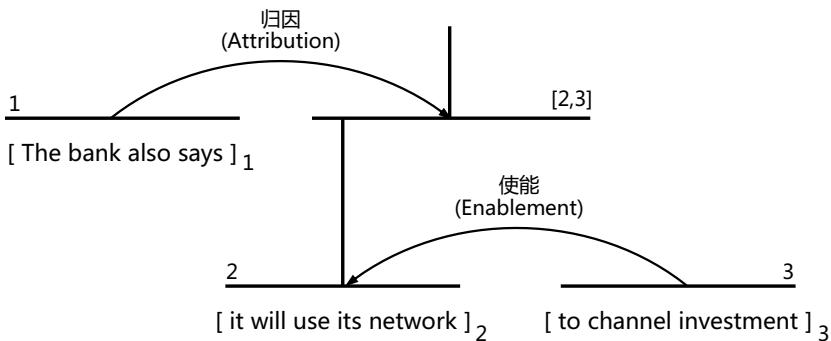


图 5.5 《述评：长征五号“王者归来”意味着什么》语篇模式示例

## 5.2 话语分割

根据修辞结构理论，篇章修辞关系定义在两个或多个基本篇章单元(EDU)之间。话语分割(Discourse Segmentation)的目标就是将篇章分割为基本篇章单元，从而实现后续的篇章分析任务。话语分割任务通常被形式化为序列标注任务或者单词级别的二分类任务，对每个单词位置输出预测其是否为一个基本篇章单元的边界。如图 5.6 所示，句子“The bank also says it will use its network to channel the investments”由三个基本篇章单元组成。本节将分别介绍两种话语分割算法：基于词汇句法树的统计话语分割和基于循环神经网络的话语分割方法。

图 5.6 话语分割样例<sup>[226]</sup>

### 5.2.1 基于词汇句法树的统计话语分割

SynDS<sup>[226]</sup> 算法采用基于句法树的统计模型估计句子中每个词作为分界点的概率。具体来说，给定句子  $s = w_1 w_2 \dots, w_n$ ，首先使用句法分析工具得到该句子的句法树  $t$ ，随后对句子中的每个词  $w_i$ ，使用最大似然估计的方法学习其作为分界点的概率  $P(b_i|w_i, t)$ ，其中  $b_i \in \{0, 1\}$ 。0 表示为非边界，1 表示为边界。由于句子间的分界点较易得到，SynDS 重点关注句子内部的分界，因此将句子间的分界设为  $P(b_n = 1|w_n, t) = 1$ 。

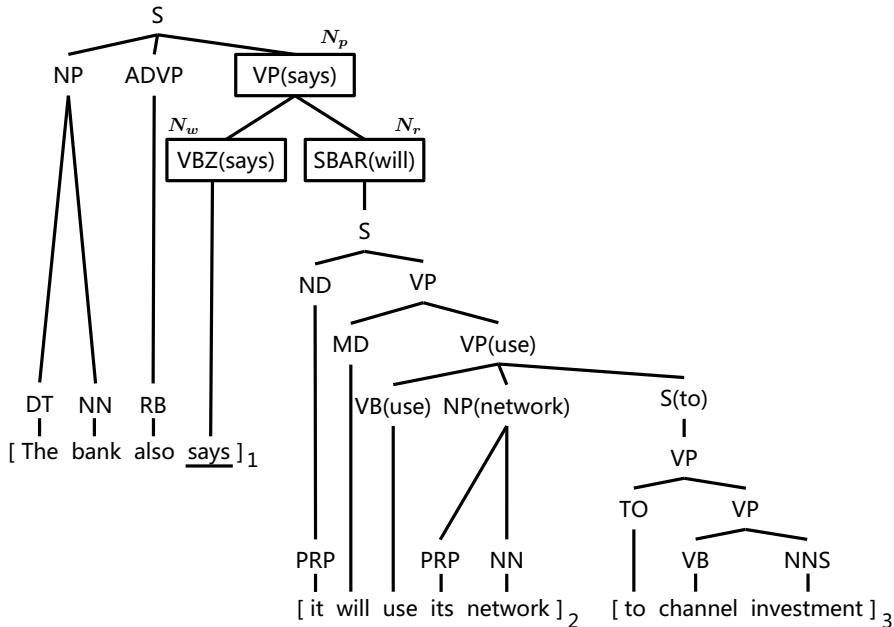
为了同时使用词汇及句法特征估计基本篇章单元分界，SynDS 算法使用文献 [227] 中提出词汇中心（Lexical Head）映射规则将词汇成分引入句法树。对于每个词  $w$  来说，SynDS 算法关注其包含一个右兄弟节点的最高父节点，使用其构建的特征决定当前词是否作为分界词。具体来说，将词  $w$  对应的带词汇信息的节点记为  $N_w$ ，使用的特征包含  $N_w$  本身、其父节点  $N_p$  及其兄弟节点。例如，针对图5.7的例子，SynDS 在判断词“says”是否为分界词时，使用该词本身对应的节点  $N_w = \text{VBZ}(\text{says})$ 、其父节点  $N_p = \text{VP}(\text{says})$  及其右兄弟节点  $N_r = \text{SBAR}(\text{will})$  作为特征。

为了训练分类模型，使用 RST-DT<sup>[228]</sup> 语料的统计量估计每个词作为分界词的似然概率：

$$P(b|w, t) \simeq \frac{\text{Cnt}(N_p \rightarrow \dots N_w \uparrow N_r \dots)}{\text{Cnt}(N_p \rightarrow \dots N_w N_r \dots)} \quad (5.1)$$

其中，分子表示规则  $\text{Cnt}(N_p \rightarrow \dots N_w N_r \dots)$  的在语料中出现且  $w$  为分界词的次数（↑表示该处为分界），分母表示该规则在语料中出现的总次数。

当通过统计模型获得每个词作为分界词的概率后，SynDS 算法对于给定一个句法树  $t$ ，当  $P(b = 1|w, t) > 0.5$  时，选择  $w$  作为分界词（即在  $w$  后插入分界）。

图 5.7 基于词汇句法树的统计话语分割样例<sup>[228]</sup>

### 5.2.2 基于循环神经网络的话语分割

话语分割任务还可以转换为序列标注问题<sup>[229]</sup>，给定一个输入句子  $x = \{x_t\}_{t=1}^n$ ，其输出  $y = \{y_t\}_{t=1}^n$  中每个  $y_t$  表示第  $t$  个词是否为一个基本单元的开头，如果是，则  $y_t = 1$ ，否则  $y_t = 0$ 。可以采用基于 BiLSTM-CRF 模型实现这一任务。

将每个词  $x_t$  表示为一个向量  $e_t$ ，然后使用一个 Bi-LSTM 网络建模序列中每个词的表示：

$$\mathbf{h}_t = \text{BiLSTM}(\mathbf{h}_{t-1}, e_t) \quad (5.2)$$

其中  $\mathbf{h}_t = [\mathbf{h}_t^f, \mathbf{h}_t^b]$  为  $t$  位置正向 LSTM 及反向 LSTM 编码得到的隐向量表示的串联。

在获得每个词的隐向量表示后，为了更好地利用序列信息，使用条件随机场层进行输出解码。给定一个句子  $x$  的输出隐向量序列  $\mathbf{h} = \{\mathbf{h}_t\}_{t=1}^n$ ，该句子预测为序列  $y$  的概率为：

$$P(y | \mathbf{h}; \mathbf{W}; \mathbf{b}) = \frac{\prod_{i=1}^n \psi_i(y_{i-1}, y_i, \mathbf{h})}{\sum_{y' \in \mathcal{Y}} \prod_{i=1}^n \psi_i(y'_{i-1}, y'_i, \mathbf{h})} \quad (5.3)$$

其中， $\mathcal{Y}$  表示所有可能的输出标签序列， $\phi_i(y'_{i-1}, y'_i, \mathbf{h}) = \exp(\mathbf{w}^T \mathbf{h}_i + b)$  为势函数， $\mathbf{w}$  和  $b$  为和标签对  $(y'_{i-1}, y'_i)$  相关的权重和偏置。在训练时，模型最大化输出正确标签序列的似然概率。在解

码时，模型计算产生最大似然概率的标签序列（关于条件随机场的具体介绍可以参见2.3.3章节）：

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} P(\mathbf{y} | \mathbf{h}; \mathbf{W}; \mathbf{b}) \quad (5.4)$$

由于用于话语分割训练的语料通常较小，使用上述 BiLSTM 模型在该语料上训练难以取得理想地效果。因此，可以使用预训练词向量将更大语料上获取的知识迁移到话语分割任务上。例如使用 ELMo<sup>[230]</sup> 词向量建模输入序列：

$$\mathbf{r}_t = \gamma^{LM} \sum_{l=0}^3 s_l^{LM} \mathbf{h}_{t,l}^{LM} \quad (5.5)$$

其中， $s^{LM}$  为归一化权重，对 ELMo 词向量的三个组成部分进行加权， $\gamma^{LM}$  为整个 ELMo 词向量的权重。随后， $\mathbf{r}_t$  被拼接到词向量  $e_t$  上作为模型的输入。

此外，由于许多 EDU 边界的预测需要长距离信息，而 LSTM 模型较难处理长距离依赖，还可以考虑使用自适应机制 (Restricted Self-attention) 加强句子的长距离依赖建模。模型架构如图5.8所示。

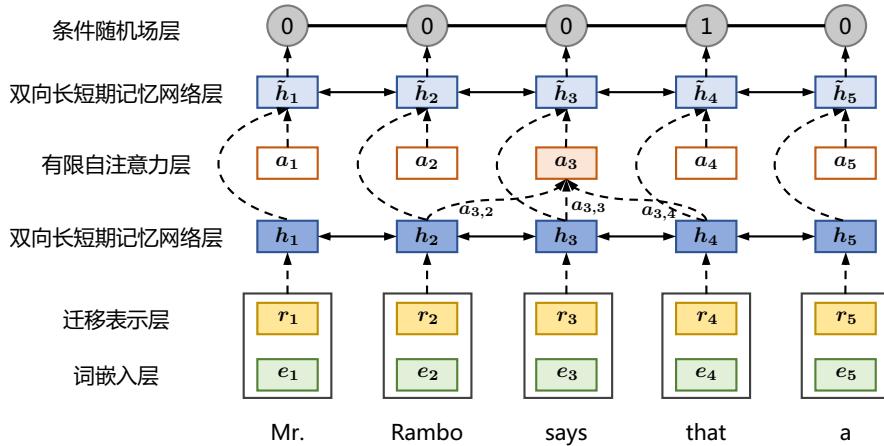


图 5.8 基于循环神经网络的话语分割<sup>[229]</sup>

但是并不是所依赖的距离越长越好，模型关注过长距离的信息可能会引入噪音，不利于预测。考虑 EDU 边界识别任务特性，即通常只需要使用相邻 EDU 的信息。因此，使用距离限制的自适应机制，只使用邻近的信息来预测。首先计算当前词  $x_i$  和一个窗口内的相邻词  $x_j$  之间的相似度：

$$s_{i,j} = \mathbf{W}_{attn}^T [\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_i \odot \mathbf{h}_j] \quad (5.6)$$

随后，每个词的注意力向量由相邻词加权获得：

$$\alpha_{i,j} = \frac{e^{s_{i,j}}}{\sum_{k=-K}^K e^{s_{i,i+k}}} \quad (5.7)$$

$$\mathbf{a}_i = \sum_{j=-K}^K \alpha_{i,i+k} \mathbf{h}_{i+k} \quad (5.8)$$

其中， $K$  为使用的窗口大小。该注意力向量随后和  $\mathbf{h}$  一起被输入另一层 BiLSTM，并输出  $\tilde{\mathbf{h}}$  作为 CRF 的输入：

$$\tilde{\mathbf{h}}_t = \text{BiLSTM}(\tilde{\mathbf{h}}_{t-1}, [\mathbf{h}_t, \mathbf{a}_t]) \quad (5.9)$$

最后，利用序列标注模型 CRF 给出输出。

## 5.3 篇章结构分析

篇章结构分析的目标是分析篇章单元之间存在的连贯关系，从而服务于下游任务。现有的篇章分析工作基于不同的篇章分析标注框架，主要可以分为两大类：基于词汇的浅层篇章分析及基于语义或意图关系的完整篇章分析。前者的代表性框架为文献 [231] 所提出的 Penn Discourse Treebank (PDTB) 标注框架；后者的代表性框架为基于修辞结构理论的 RST Discourse TreeBank (RST-DT) 标注框架<sup>[228]</sup>。本节将介绍基于这两种代表性标注框架的篇章分析方法。

### 5.3.1 修辞结构篇章分析

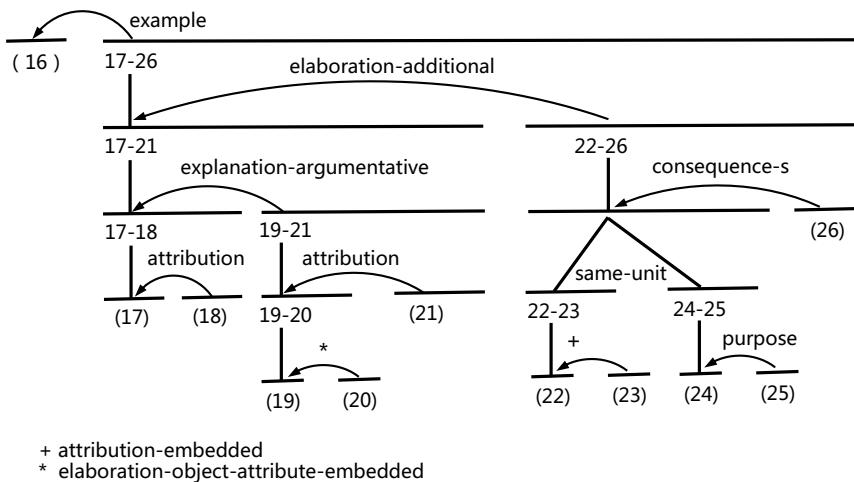
RST-DT<sup>[228]</sup> 是篇章分析中的代表性标注框架，其标注基于修辞结构理论，将一个完整的篇章标注成由基本篇章单元组成的层次树状结构。其中，树的节点关系由相邻篇章单元之间的关系构成。在标注时，一个完整的篇章首先被切分成不相交的基本篇章单元 (EDU)，随后，基于修辞关系理论，相邻的基本篇章单元被连接并标注为 78 种修辞关系的一种。篇章整体最终被标注为层次化的树状结构。图5.9中展示了一个篇章的 RST-DT 标注样例。

#### 1. 基于 SVM 分类器的 RST 篇章分析

HILDA (HIgh-Level Discourse Analyzer)<sup>[232]</sup> 是基于 SVM 分类器的 RST 篇章分析算法，将修辞结构树定义为二叉树结构，并定义了建立一个有效修辞结构树 (valid RS-tree)  $T$  的两项规则：

- (1)  $T$  的所有叶子结点均为 EDU (单个 EDU 也可以构成一个修辞结构树)。
- (2)  $T$  的所有非叶子结点被标注为篇章关系集合中的一种关系 ( $R_i \in \mathcal{R}$ )。

基于这一修辞结构树的定义，HILDA 采用了基于贪心原则的流水线方法，使用两个支持向量机分类器对 EDU 之间是否存在关系以及存在何种关系分别进行分类。具体来说，HILDA 定义的两个分类器：

图 5.9 RST-DT 标注样例<sup>[228]</sup>

- 结构分类器  $\text{Struct}(l_i, l_j)$ : 用于判断篇章结构的二元分类器，即判断两个有效修辞结构树之间是否存在修辞关系，分类目标为 0 和 1，0 表示没有关系，1 表示有关系。
- 类型分类器  $\text{Label}(l_i, l_j)$ : 用于判断修辞关系类型及核类型的多元分类器，分类目标为篇章关系集合  $\mathcal{R} = \{R_1, \dots, R_n\}$ ,  $R_i = \langle RR_i, Left_i, Right_i \rangle$  定义为由两个有效修辞结构树的核类型及其之间的修辞关系类型组成的三元组，其中  $RR_i \in \{\text{ATTRIBUTION}, \text{CAUSE}, \dots\}$  为 HILDA 所使用的修辞关系集合中的一种关系； $Left_i, Right_i \in \{\text{Nucleurs}, \text{Satellites}\}$  为两个修辞结构树的核类型。

基于上述两个分类器，HILDA 的算法流程如算法5.1 所示。对输入文本，首先创建一个包含所有 EDU 的列表，其中 EDU 的排序按照从左到右的阅读顺序。当列表元素数目大于 1 时，使用结构分类器计算列表中所有相邻单元的结构预测分数。基于贪心原则，取出所有结构预测中分数最高的一组，使用类型分类器预测其修辞关系类型及核类型，并基于预测结果建立一个新的子树。得到新的子树后，分别重新计算该子树与相邻单元之间的结构预测分数，并将列表中对应单元替换为新的子树。重复上述过程直到列表元素数目等于 1，则留在列表中的元素即为输出的篇章树。

为了能够更准确地对篇章结构及修辞关系进行分类，HILDA 构建了多种类型的特征作为 SVM 分类器的输入，包括 N-gram 特征、句法结构特征、POS 特征等。文献 [233] 介绍了在 HILDA 的基础上，通过构建上下文等更丰富的语言特征进一步提升了性能的方法。

## 2. 基于递归神经网络的 RST 篇章分析

文献 [234] 提出了基于递归神经网络的 RST 篇章分析算法 RNN-RST。与 HILDA 算法相似，RNN-RST 也是通过训练两个分类器，即结构分类器及修辞关系类型分类器构造修辞结构树，但用

---

**代码 5.1: HILDA 算法**

---

```

输入: EDU 列表  $E = \langle e_1, e_2, \dots \rangle$ 
输出: 篇章树  $FinalTree$ 

// 初始化
 $L \leftarrow E$  ;
foreach  $(l_i, l_{i+1})$  in  $L$  do
    Scores[i]  $\leftarrow$  Struct( $l_i, l_{i+1}$ )a; // 使用结构分类器计算列表中所有相邻单元的结构预测
    分数 ;
end

// 解码过程
while  $|L| > 1$  do
     $i \leftarrow \arg \max(\text{Scores})$ ;           // 取出结构预测分数最高的一组相邻单元 ;
    NewLabel  $\leftarrow$  Label( $l_i, l_{i+1}$ );           // 预测修辞关系类别 ;
    NewSubTree  $\leftarrow$  CreateTree( $l_i, l_{i+1}, NewLabel$ )b;           // 建立新的子树 ;
    // 更新结构预测列表
    Scores[i - 1]  $\leftarrow$  Struct( $l_{i-1}$ , NewSubTree) ;
    Scores[i + 2]  $\leftarrow$  Struct(NewSubTree,  $l_{i+2}$ ) ;
    Delete(Scores[i]) ;
    Delete(Scores[i + 1]) ;
     $L \leftarrow [l_0, \dots, l_{i-1}, NewSubTree, l_{i+2}, \dots]$ ;           // 更新子树列表 ;
end

FinalTree  $\leftarrow l_0$  ;
return FinalTree

```

---

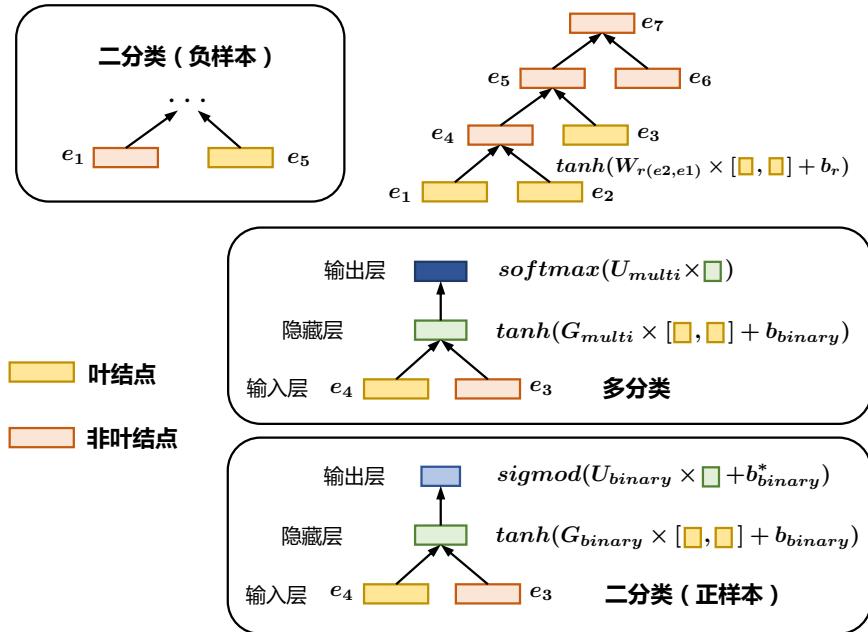
于分类的特征则使用递归神经网络进行计算。算法的整体结构如图5.10所示。

具体来说，对于每个给定句子  $S = w_1 w_2 \dots w_{n_s}$ ，其中  $w_i$  代表句子中的第  $i$  个词， $n_s$  代表句子  $S$  中词的总数。每个词映射为一个词向量  $e_w \in \mathbb{R}^K$ ，其中  $K$  为词向量的维度。对于给定句子，计算它的特征向量  $h_s \in \mathbb{R}^K$ 。

首先，使用句法分析工具对给定句子进行分析，分析得到的每个子句被视为一个基本篇章单元 EDU。基于得到的句法树，对于其中的每个父节点  $p$  及它的两个子节点  $c_1$  和  $c_2$ （分别对应向量表示  $h_{c_1}$  和  $h_{c_2}$ ），该父节点的表示计算如下：

$$h_p = f(\mathbf{W} \cdot [h_{c_1}, h_{c_2}] + b) \quad (5.10)$$

其中  $[h_{c_1}, h_{c_2}]$  为子节点向量表示  $h_{c_1}$  和  $h_{c_2}$  的拼接， $\mathbf{W}$  为一个  $K \times 2K$  的矩阵， $b$  为  $1 \times K$  的偏移向量。 $f(\cdot)$  为  $\tanh$  激活函数。对于整个句法树，递归神经网络由下至上递归计算每个父节点

图 5.10 基于递归神经网络的 RST 篇章分析算法神经网络结构图<sup>[234]</sup>

的表示，直到获得该句的根结点的表示，并将该表示作为该句子的向量表示  $h_s$ 。

RNN-RST 同样使用结构分类器和修辞关系类型分类器分别对 EDU 之间的结构和关系进行分类。结构分类器使用单层卷积神经网络编码，并使用单层线性分类器投影至  $[0,1]$  输出空间进行分类：

$$\begin{aligned} L_{(e_i, e_j)}^{binary} &= f(\mathbf{G}_{binary} * [h_{e_i}, h_{e_j}] + b_{binary}) \\ P[t_{binary}(e_i, e_j) = 1] &= g(\mathbf{U}_{binary} \cdot L_{(e_i, e_j)}^{binary} + b_{binary}^*) \end{aligned} \quad (5.11)$$

其中， $\mathbf{G}_{binary}$  是一个  $N_{binary} \times 2K$  的卷积矩阵， $b_{binary}$  是偏移向量， $f(\cdot)$  为  $\tanh$  激活函数； $\mathbf{U}_{binary}$  是一个  $N_{binary} \times 1$  的向量， $b_{binary}^*$  表示偏移值， $g(\cdot)$  为  $\text{sigmoid}$  激活函数。 $t_{binary}(e_i, e_j) = 1$  代表两个 EDU  $e_i$  和  $e_j$  之间存在依存关系。

当  $t_{binary}(e_i, e_j)$  的预测值为 1 时，接着使用一个多元分类器预测其修辞关系类型，预测的关系类型表示为  $r(e_i, e_j)$ 。修辞关系类型分类器的结构和结构分类器类似，但使用  $\text{Softmax}$  激活函数进行输出：

$$L_{(e_i, e_j)}^{multi} = f(\mathbf{G}_{multi} * [h_{e_i}, h_{e_j}] + b_{multi}) \quad (5.12)$$

$$S_{(e_1, e_2)} = U_{multi} \cdot L_{(e_i, e_j) multi} \quad (5.13)$$

$$P_{(e_1, e_2)}(i) = \frac{\exp(S_{(e_1, e_2)}(i))}{\sum_k \exp(S_{(e_1, e_2)})(k)} \quad (5.14)$$

其中， $G_{multi}$  是一个  $N_{multi} \times 2K$  的矩阵， $b_{multi}$  是偏移向量， $f(\cdot)$  为  $\tanh$  激活函数。 $U_{multi}$  是一个  $N_r \times 2K$  的矩阵， $P_{(e_1, e_2)}$  中的第  $i$  个元素代表  $e_i$  和  $e_j$  之间存在第  $i$  种关系的概率。需要注意的是，二元分类器及多元分类器在训练时为分别训练。

上述的分类计算基于每个节点的表示进行。然而，公式5.11只能获得每个句子即叶子节点的表示，而无法获得非叶子节点的表示。对于非叶子节点，即由 EDU 构成的子树，该方法使用递归神经网络进一步由下至上计算其父节点。具体地，对于给定子节点表示  $h_{e_i}$  和  $h_{e_j}$  及其标注类型  $r(e_i, e_j)$ ，其父节点表示  $h_p$  的计算如下：

$$h_p = f(W_{r(e_i, e_j)} \cdot [h_{e_i}, h_{e_j}] + b_{r(e_i, e_j)}) \quad (5.15)$$

其中， $W_{r(e_i, e_j)}$  为  $r(e_i, e_j)$  所对应关系类型的  $K \times 2K$  参数矩阵， $b_{r(e_i, e_j)}$  为该关系类型对应的偏移向量， $f(\cdot)$  为  $\tanh$  激活函数。

利用标注语料集合，可以分别构造上述两个分类器的训练数据，并利用交叉熵损失函数进行模型参数训练。在训练完成后可以采用类似用于句法分析的 CKY 动态规划方法，对于给定的篇章进行修辞结构树构建。对于由  $n$  个 EDU 组成的篇章，可以构建  $N_r \times n \times n$  组成的动态规划表  $Pr$ ， $N_r$  表示关系类型数量， $Pr$  表中每个单元格  $Pr[r, i, j]$  表示从片段从第  $i$  个 EDU 到第  $j$  个 EDU 中具有关系  $r$  的概率，其计算过程如下：

$$\begin{aligned} Pr[r, i, j] &= \max_{r_1, r_2, k} Pr[r_1, i, k] \cdot Pr[r_2, k, j] \\ &\quad \times P(t_{binary}(e_{[i, k]}, e_{[k, j]})) = 1 \\ &\quad \times P(r(e_{[i, k]}, e_{[k, j]})) = 1 \end{aligned} \quad (5.16)$$

### 5.3.2 浅层篇章分析

Penn Discourse Treebank (PDTB) 是基于词汇化树型连接语法 (Discourse Lexical Tree Adjunct Grammar, D-LTAG) 理论<sup>[235]</sup> 构建的篇章分析标注框架，是篇章分析中的另一代表性框架。PDTB 以篇章内相邻或者跨度在一定范围内的片段，以连接词为核心，对片段间关系进行标注。相较于修辞结构理论将整个篇章构建为树结构而言，PDTB 则针对两个片段之间的关系，因此也可以称为浅层篇章分析。每个篇章关系由两个论据 (Argument) 及其之间的关系组成，两个论据分别标注为 Arg1 和 Arg2。在相邻句子构成的关系中，Arg1 和 Arg2 则反映论据之间的线性顺序，其中 Arg1 在 Arg2 之前。根据连接词是否显式存在，PDTB 标注的关系可分为显式篇章关系和隐式篇章

关系两类。

显式篇章关系（Explicit Discourse Relation）由显式连接词定义，通过显式连接词连接 Arg1 和 Arg2。在显式关系中，Arg2 一般为句法上关联的论据，Arg1 则为另一个论据，显式连接词由三种语法连接词产生：

- 从属连词，如 because, when 等
- 并列连词，如 and, or 等
- 语篇副词，如 for example, instead 等

除此之外，还有一些带修饰或联合形式的连接词（如“only because”，“if and when”等）以及小部分并列连接词（如“either..or”，“on the one hand..on the other hand”等）。以下是一些基于显式关系定义的标注样例<sup>[231]</sup>（下划线标注显式连接词，斜体标注 Arg1，粗体标注 Arg2）：

- (1) *Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products –although weaker foreign currencies reduced the company's earnings.*
- (2) *Most oil companies, when they set exploration and production budgets for this year, forecast revenue of \$15 for each barrel of crude produced.*

隐式篇章关系（Implicit Discourse Relation）则是除显式篇章关系以外，需要靠读者通过推断判断的篇章关系。例如下面的例句：

- (1) *But a few funds have taken other defensive steps. Some have raised their cash positions to record levels. Implicit = BECAUSE High cash positions help buffer a fund when the market falls.*

虽然没有显式连接词，但读者能够通过论据之间的语义判断出其之间表达的因果关系。在 PDTB 中，这样的隐式关系通常通过标注者插入一个连接词进行标注（例如上面例子中的 BECAUSE 被插入以表示因果关系）。而当隐式关系无法使用一个隐式连接词进行标注时，则构成三种特殊的隐式关系：

- AltLex 表示语篇关系已经由非连接词的词汇表达，额外插入连接词会构成冗余的情况
- EntRel 表示句子之间只存在基于实体的连贯关系的情况
- NoRel 表示句子之间不存在任何篇章关系或基于实体的连贯关系的情况

下面三个例子分别为 AltLex、EntRel、NoRel 的样例：

- (1) *Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject. Implicit = AltLex Mayhap this metaphorical connection made the BPC Fine Arts Committee think she had a literal green thumb.*
- (2) *Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. Implicit = EntRel Mr. Milgrim succeeds David Berman, who resigned last month.*
- (3) *Jacobs is an international engineering and construction concern. Implicit = NoRel Total capital investment at the site could be as much as \$400 million, according to Intel.*

由于一个连接词在不同的篇章中可能表达不同的语义关系，PDTB 中为显式关系、隐式关系和 AltLex 关系提供了三级语义标注（Sense Tag）：CLASS, TYPE, SUBTYPE。其中，最高层标注（CLASS）包含四个主要语义类别：TEMPORAL, CONTINGENCY, COMPARISON, EXPANSION。对于每个语义类别，使用 TYPE 进一步标注其语义。第三级语义标签 SUBTYPE 则具体化每个论据的语义贡献。图5.11给出了 PDTB 的三级语义标签。

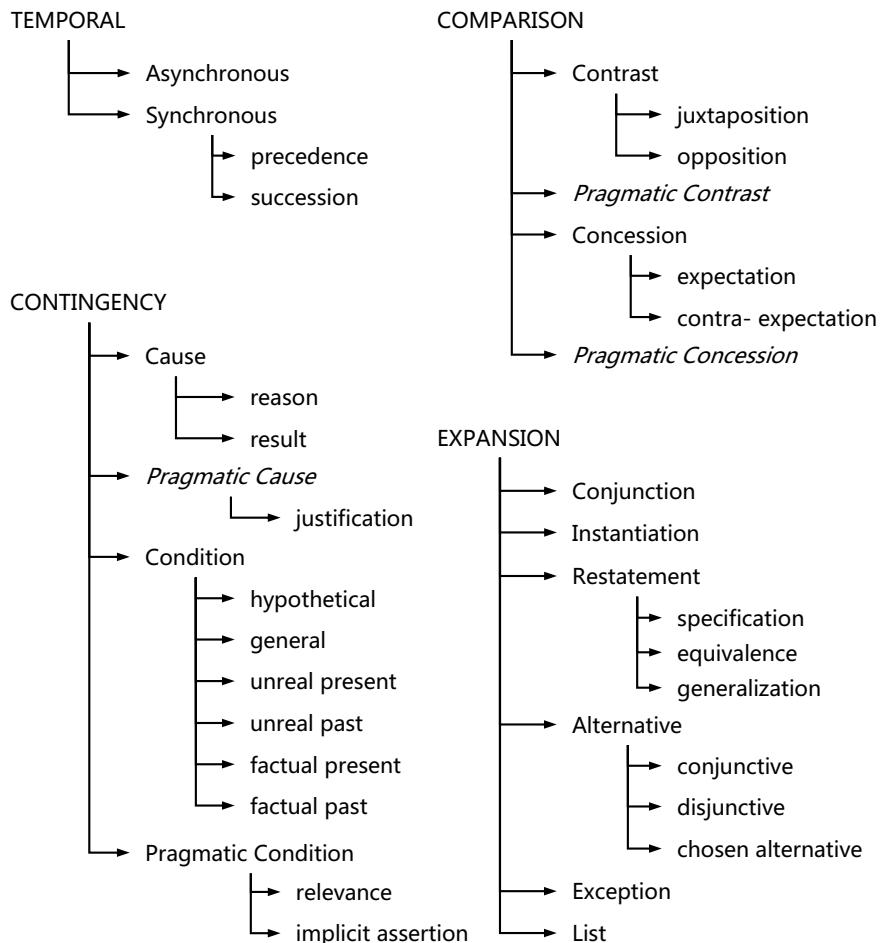


图 5.11 PDTB 标注的三级语义标签

基于上述 PDTB 标注的篇章分析工作一般可以划分为显式篇章分析（Explicit Discourse Pars-

ing) 和隐式篇章分析 (Implicit Discourse Parsing) 两类。其中，显式篇章分析关注篇章的显式关系识别，包括显式连接词检测、论据标注等子任务；隐式篇章分析则更关注给定句子对之间的隐式关系分类。本节将分别介绍显式篇章分析和隐式篇章分析的代表性工作。

### 1. 基于句法特征构建的显式篇章分析

由于部分显式连接词在不同语境下具有不同语义，显式篇章分析的重点在于对显式连接词进行消歧，并将每一连接词分类为 PDTB 标注的四个一级语义类别 (TEMPORAL、CONTINGENCY、COMPARISON、EXPANSION)。

例如：下述两个包含 since 的句子

- (1) Guangzhou has a wide water area with many rivers and water systems since it is located in the water-rich area of southern China.
- (1) She has been living in Beijing since she graduated from Fudan University.

显式连接词“Since”在句子 (1) 中表示因果关系，为 CONTINGENCY 语义类别；在句子 (2) 中则为 TEMPORAL 语义类别。显式篇章分析关注的主要问题即为将连接篇章中话语的显式连接词正确划分为其所属的语义类别，这一任务通常以文本分类的方式实现。

文献 [236] 使用最大熵分类器，通过利用句法特征对显式篇章关系进行分类。基于标准 Penn Treebank 句法分析标注<sup>[237]</sup>，构建了多种句法特征对显式连接词的语义进行消歧，所构建的句法特征包括：

- 自身类别 (Self Category)：子树包含且仅包含该显式连接词的最高父节点。对于单个单词构成的显式连接词，其特征为该词自身的 POS 标注；对于多个单词构成的显式连接词则不然。例如，*in addition* 的成分句法标注为 (PP (IN In) (NP (NN addition ))), 其自身类别则为 PP (Prepositional Phrase)。
- 父节点类别 (Parent Category)：自身类别的最近父节点的类别。
- 左兄弟节点类别 (Left Sibling Category)：离自身类别最近的左兄弟节点类别。如果左兄弟节点不存在，则其特征为 “None”。
- 右兄弟节点类别 (Right Sibling Category)：离自身类别最近的右兄弟节点类别。文献 [236] 认为，由于英语是右分支结构，句子的依赖一般出现在其头部之后，因此右兄弟节点类别一般包含该显式连接词的依赖，从而显得尤为重要。对于表达篇章关系的显式连接词，其依赖一般为一个从句。例如，句子 “*After I went to the store, I went home*” 可以通过右兄弟节点类别显示其表达的篇章关系，从而和句子 “*After May, I will go on vacation*” 区分开。除了右兄弟节点类别以外，作者还增加了两项特征以进一步利用右兄弟节点的信息，提升消歧效果：包含一个 VP 的右兄弟节点 (Right Sibling Contains a VP) 和包含一个 Trace 的右兄弟节点 (Right Sibling Contains a Trace)。

文献 [236] 中给出的实验结果表明，通过构建句法特征，对显式连接词进行分类能够达到 94.15% 的准确率。

## 2. 基于循环神经网络语言模型的隐式篇章分析

由于传统机器学习方法在显式篇章分析任务上已经能够达到较高的准确率<sup>[238]</sup>, 后续基于 PDTB 的篇章分析工作更多地关注隐式篇章分析, 相关的工作包括基于前馈网络<sup>[239]</sup>、基于浅层卷积神经网络<sup>[240]</sup>、基于循环神经网络语言模型<sup>[241]</sup> 及基于预训练语言模型<sup>[242]</sup> 的隐式篇章分析等。本节中将介绍基于循环神经网络语言模型 (RNN 语言模型) 的隐式篇章分析算法 DRLM<sup>[241]</sup>。

DRLM 算法<sup>[241]</sup> 使用包含隐变量的循环神经网络语言模型建模隐式篇章分析算法, 整个过程建模为两阶段生成过程。首先, 句子  $t - 1$  和句子  $t$  之间的隐式篇章关系  $z_t$  由句子  $t - 1$  的信息建模。在此基础上, 句子  $x_t$  根据句子  $x_{t-1}$  和  $z_t$  生成。DRLM 的网络结构如图5.12所示。

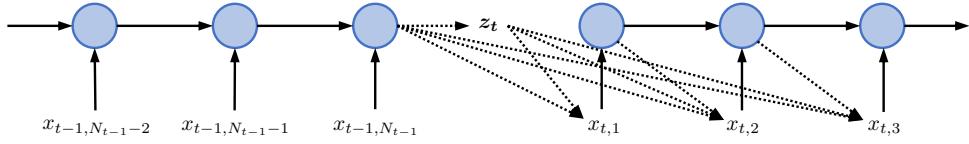


图 5.12 DRLM 算法包含隐变量结构图<sup>[241]</sup>

给定输入句子  $\mathbf{x}_t = \{x_{t,n}\}_{n \in \{1 \dots N_t\}}$ , 其中  $t$  表示该句为篇章中的第  $t$  个句子,  $N_t$  为句子  $t$  的长度。基于链式法则, RNN 语言模型将该句出现的概率转化为每个词出现的条件概率的乘积:

$$p(\mathbf{x}_t) = \prod_{n=1}^{N_t} p(x_{t,n} | \mathbf{x}_{t,<n}) \quad (5.17)$$

在每个时刻  $n$ , RNN 语言模型的输入为包含所有历史信息的上一刻输出隐变量  $\mathbf{h}_{t,n-1}$  和当前词的词向量  $\mathbf{X}_{x_{t,n}}$ , 其预测下一个词的条件概率为:

$$\begin{aligned} \mathbf{h}_{t,n} &= f(\mathbf{X}_{x_{t,n}}, \mathbf{h}_{t,n-1}) \\ p(x_{t,n} | \mathbf{x}_{t,<n}) &= \text{softmax}(\mathbf{W}_o \mathbf{h}_{t,n-1} + \mathbf{b}_o) \end{aligned} \quad (5.18)$$

其中,  $\mathbf{W}_o \in \mathbb{R}^{V \times K}$  为输出层权重,  $\mathbf{b}_o \in \mathbb{R}^V$  为输出层偏移量。

由于篇章分析需要对包含多句话的长文本进行语言模型建模, 而 RNN 语言模型难以处理长距离依赖关系, DRLM 使用了基于文档的语言模型<sup>[243]</sup>。具体来说, 文档中第  $t$  个句子的第  $n$  步输出的条件概率为:

$$p(x_{t,n} | \mathbf{x}_{t,<n}, \mathbf{x}_{<t}) = \text{softmax}(\mathbf{W}_o \mathbf{h}_{t,n-1} + \mathbf{W}_c \mathbf{c}_{t-1} + \mathbf{b}_o) \quad (5.19)$$

其中,  $\mathbf{c}_{t-1}$  为句子  $t - 1$  的上下文信息, 此处设为上一个句子的最后一步输出的隐向量。

基于上述 RNN 语言模型, DRLM 将篇章关系  $z_t$  作为隐变量引入, 并设计了两步语言模型生成过程, 如图5.12所示。第一步, 使用句子  $t - 1$  的上下文信息  $\mathbf{c}_{t-1}$  生成句子  $t - 1$  和句子  $t$  之间的篇章关系:

$$p(z_t | \mathbf{x}_{t-1}) = \text{Softmax}(\mathbf{U}\mathbf{c}_{t-1} + \mathbf{b}) \quad (5.20)$$

其中,  $z_t$  是一个表示句子间篇章关系的随机变量。

第二步, 基于句子  $\mathbf{x}_{t-1}$  及上一步生成的篇章关系  $z_t$ , 生成句子  $\mathbf{x}_t$ :

$$p(\mathbf{x}_t | z_t, \mathbf{x}_{t-1}) = \prod_n^{N_t} p(x_{t,n} | \mathbf{x}_{t,< n}, \mathbf{x}_{t-1}, z_t) \quad (5.21)$$

其中, 引入篇章关系隐向量的输出条件概率计算为:

$$p(x_{t,n} | \mathbf{x}_{t,< n}, \mathbf{x}_{t-1}, z_t) = g(\mathbf{W}_o^{(z_t)} \mathbf{h}_{t,n} + \mathbf{W}_c^{(z_t)} \mathbf{c}_{t-1} + \mathbf{b}_o^{(z_t)}) \quad (5.22)$$

$\mathbf{W}_o^{(z_t)} \mathbf{h}_{t,n}$ 、 $\mathbf{W}_c^{(z_t)} \mathbf{c}_{t-1}$  及  $\mathbf{b}_o^{(z_t)}$  为由篇章关系决定的参数, 不同的篇章关系通过不同的权重关注特征空间中的不同部分特征。

最后, 文本及篇章关系的联合概率为:

$$p(\mathbf{x}_{1:T}, z_{1:T}) = \prod_t^T p(\mathbf{x}_t | z_t, \mathbf{x}_{t-1}) \times p(\mathbf{x}_t | z_t, \mathbf{x}_{t-1}) \quad (5.23)$$

在训练阶段, DRLM 可以使用两种目标函数进行训练: 联合似然目标函数和条件目标函数。其中, 联合似然目标函数的损失函数计算为:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_t^T \log p(z_t | \mathbf{x}_{t-1}) + \sum_n^{N_t} \log p(x_{t,n} | \mathbf{x}_{t,< n}, \mathbf{x}_{t-1}, z_t) \quad (5.24)$$

其中  $\boldsymbol{\theta}$  表示模型参数。使用上述联合似然目标函数能够同时优化模型的篇章关系预测能力及语言模型能力, 可以视为一种多任务学习。然而, 在实际实现时, 由于词的数量比句子的数量更多, 使用这一目标对模型语言模型能力的优化占主导地位。

因此, DRLM 的条件目标函数的损失函数为:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_t^T \log p(z_t | \mathbf{x}_{t-1}) + \log p(\mathbf{x}_t | z_t, \mathbf{x}_{t-1}) - \log \sum_{z'} p(z' | \mathbf{x}_{t-1}) \times p(\mathbf{x}_t | z', \mathbf{x}_{t-1}) \quad (5.25)$$

该式的前两项与公式5.24一致, 但第三项计算了所有可能的  $z'$  生成句子  $\mathbf{x}_t$  的损失。这项损失的加入使得目标函数倾向于优化篇章关系相关的条件似然损失, 而将语言模型任务视为一个辅助任务。

在推断时，通过贝叶斯公式计算句子  $t - 1$  和句子  $t$  之间为关系  $z_t$  的条件概率：

$$p(z_t | \mathbf{x}_t, \mathbf{x}_{t-1}) = \frac{p(\mathbf{x}_t | z_t, \mathbf{x}_{t-1}) \times p(z_t | \mathbf{x}_{t-1})}{\sum_{z'} p(\mathbf{x}_t | z', \mathbf{x}_{t-1}) \times p(z' | \mathbf{x}_{t-1})} \quad (5.26)$$

其中的每项概率可以根据公式5.20和公式5.21进行计算。

## 5.4 指代消解

在本章第 5.1 节篇章衔接的概要介绍中，提到了篇章中的指代现象。该现象是体现篇章衔接性的重要组成部分。虽然指代现象并不影响人类阅读和理解篇章，甚至还起到了避免重复以及提高语言效率的作用。但是指代对于一些自然语言处理任务却有一定的影响，需要明确不同表述之间的指代关系。在本节中，我们将介绍篇章分析的一重要子任务：指代消解 (Coreference Resolution)。指代消解旨在将同一实体 (Entity) 在篇章中出现的不同表述 (Mention, 也称提及) 划分到同一等价类 (或称表述类) 中。其中，实体指某一客观存在的事物；表述则为指代某一实体的在篇章中不同描述。指代消解任务通常关注两种指代类型：共指 (Coreference) 和回指 (Anaphora)。共指表示两个表述指向真实世界中的同一实体。

例如：上海的卖腌腊的店铺里也卖咸鸭蛋，必用纸条特别标明：“高邮咸蛋”。

上例中，“咸鸭蛋”和“咸蛋”指代真实世界中的统一实体，因此为共指关系。

回指表示当前表述指向上文出现的另一表述，通常将指代上文的表述称为照应词 (Anaphor)，将照应词指代的上文表述称为先行词 (Antecedent)。

例如：我围着火炉，烤热漫长一生的一个时刻。我知道这一时刻之外，我其余的岁月，我的亲人们的岁月，远在屋外的大雪中，被寒风吹彻。

上例中，“这一时刻”指代上文的“一个时刻”，为回指关系。其中“这一时刻”为照应词，“一个时刻”为先行词。

指代消解任务将语篇中所有表示同一实体的指代分配到同一等价类中，并给出每一语篇中的所有等价类。

例如：其间有一个十一二岁的少年 [1]，项带银圈，手捏一柄钢叉，向一匹猹 [2] 用力地刺去。那猹 [2] 却将身一扭，反从他 [1] 的胯下逃走了。

上例中，“一个十一二岁的少年”和“他”指代同一实体，属于同一等价类；“一匹猹”和“那猹”指代同一实体，属于同一等价类。指代消解任务的目标是发现文中的等价类 [1] 和 [2]。

指代消解任务一般可分为两个步骤：表述发现 (Mention Detection) 和指代消解 (Coreference Resolution)。其中，表述发现也称提及发现，旨在找出句子中所有可能存在指代关系的名词表述，一般包含人称代词（“你”、“我”、“他”等）、命名实体（人名、地名等）及一些名词短语（“那只猫”、“右边的女士”）等。表述发现的方法一般更注重将所有的表述找出，即更注重提升召回率，并在之后对无关的表述进行过滤。指代消解旨在对表述同一实体的表述聚合在一起，是这一任务的核心。

心，也是最具挑战的步骤。

指代消解的方法主要包括基于表述对（Mention Pair）、基于表述排序（Mention Ranking）和基于实体等三种方法。其中，基于表述对的方法使用一个二分类分类器对每一对表述是否为指代关系进行判断。基于表述排序的方法针对给定指代，通过计算其与每一表述组成的表述对的分数，对相关表述进行排序，以确定其指代关系。基于聚类的方法对文本中所有表述进行聚类，每个类被认为指代同一实体名词。本节将对上述三类方法分别进行介绍。<sup>231</sup>

### 5.4.1 基于表述对的指代消解

基于表述对的指代消解算法是将该任务转换为二分类问题，分别对每个表述与其所有先行词所构成的表述对是否构成指代关系进行分类。

例如：对如下句子进行指代消解

长妈妈，已经说过，是一个一向带领着我的女工，说得阔气一点，就是我的保姆。我的母亲和许多别的人都这样称呼她，似乎略带些客气的意思。

对于所选表述“她”，基于表述对的指代消解算法需要分别计算“她”和其所有先行词构成的表述对是否为指代关系进行分类。在这一例子中，“长妈妈”和“她”为正确的指代关系，“我的母亲”这一先行词与“她”不为指代关系，指代消解算法目标就是对上述所有表对是否为指代关系进行正确分类。上述过程如图5.13所示。

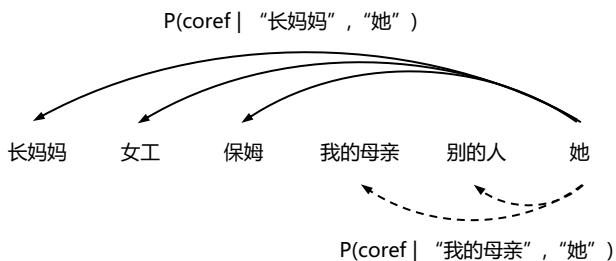


图 5.13 基于表述对的指代消解示例

在训练用于表述对分类的二分类器时，尽管训练语料中提供了表述之间的所有等价类标注，但由于相同指代的表述之间并没有直接的指代关系，将属于同一等价类的所有表述都视为正类训练可能不能取得很好的效果<sup>[244]</sup>。因此，对于一个表述  $m$ ，经典的策略是选择和当前表述存在指代关系的距离最近的先行词  $a$  与  $m$  构成的表述对  $(a, m)$  作为正样本，而选取所有和  $m$  不属于同一等价类的先行词  $b$  构成的表述对  $(b, m)$  作为负样本<sup>[245]</sup>。

基于一个训练得到的用于表述对分类的二分类器，对每个测试文本的指代消解推理可以视为一个构造消解图（Coreference Graph）的过程（也可视为聚类过程）：每个表述为图中的一个节点，当分类器预测一个对表述之间有指代关系时，则为这两个节点之间添加一条有向边。由此种方式

构成的图，能够表示每两个表述之间是否互为指代。同时通过所有连接构成的传递闭包（Transitive Closure）我们能够找出所有等价类<sup>[244]</sup>。为了实现这一目标，基于表述对的指代消解算法通常使用分类器为每个表述选择一个与其为指代关系的先行词，并将该表述与该先行词连接。对于该先行词的选择策略，基于最近原则（Closest-First）的算法<sup>[245]</sup>从后向前依次计算所有先行词与该表述构成的表述对为指代关系的分数，并选择第一个大于阈值的先行词。而基于最优原则（Best-First）的算法<sup>[246]</sup>则计算所有先行词与该表述构成的表述对为指代关系的分数，并选出分数最高的先行词进行连接。

基于上述训练及推断框架，基于表述对的指代消解系统通常使用不同的模型作为判断表述对指代关系的二分类器。早期方法使用手工构造特征训练分类器<sup>[244]</sup>，近年来基于深度神经网络的方法则使用神经网络学习用于分类的特征<sup>[247]</sup>。

### 1. 基于特征工程的表述对指代解

文献 [244] 提出了基于多类特征及感知器分类的表述对指代消解系统 Feature-pair。其构造的特征主要包括两个方面：表述特征及表述对特征。其中，表述特征包括表述类型特征，例如其是否为专有名词（Proper Noun）、普通名词（Common Noun）或代词（Pronoun）等；表述对特征包括表述对的字符串关系特征（如一个字符串是否为另一个的子串）、语义相符性特征（例如性别、数字是否相符）、相对位置特征、实体类别特征等。具体特征描述如下：

- 表述类型特征：表述所属的类型，为专有名词（Proper Noun）、普通名词（Common Noun）或代词（Pronoun）。
- 字符串关系特征：两个字符串之间是否存在一些共有特征，例如一个字符串是另一个字符串的子串等。
- 语义特征：包括两个名词之间的性别是否相符、数字是否相符；两个名词是否为近义词、反义词、或上位词等。
- 相对位置特征：两个表述之间的位置关系，例如将距离转化为二元特征 ( $[distance \leq i]$ ,  $i$  包括所有间隔值)、两个表述是否属于同一个句子等。
- 可学习特征：基于可学习分类器得到的特征，例如使用分类器判断两个由同一修饰语修饰的表述，其修饰语之间是否存在指代关系。
- 修饰语对齐特征：两个上位词相同的修饰语之间存在的关系，例如是否为子串、近义词、反义词等。
- 记忆特征：选取一些常常构成指代关系的表述对构造特征（例如“the queen”和“Elizabeth II”），供模型记忆学习。
- 预测实体类别特征：基于模版匹配预测实体所属的实体类别（人名、地名、机构名等），并基于预测类别构造两个表述之间实体类别是否匹配或是否相交等特征。

更具体特征分类可参考文献 [244]。

基于上述构造的特征，Feature-pair 算法使用感知器对每一指代对是否构成指代关系进行分类。

训练时, Feature-pair 基于最近原则, 选择当前表述  $m$  与其所指代的距离最近的先行词  $a$  所构成的表述对  $(a, m)$  作为正样本; 选取在  $m$  之前所有和  $m$  不属于同一等价类的表述作为负样本。测试时, Feature-pair 基于最优原则, 每次选择与  $m$  构成分数最高的表述对的先行词:

$$a = \arg \max_{b \in B_m} (\text{PC}(b, m)) \quad (5.27)$$

其中  $\text{PC}(\cdot)$  为表述对分数计算函数, 当  $\text{PC}(a, m)$  大于某一预定义阈值时, 则将  $a$  和  $m$  链接为同一等价类。

## 2. 基于神经网络的表述对指代消解

文献 [247] 构造了基于前馈神经网络的表述对指代消解分类器 Feedforward-pair, 其结构如图5.14所示。其中, 输入层将输入词映射到输入特征空间, 其输入特征由表述及表述相关词的词嵌入向量及一些其他特征构成。表述相关词包括表述的依赖词、句法树中的父节点、表述的第一个词、表述的第二个词、表述之前的两个词及之后的两个词等等; 其他特征包括表述的类型特征、位置特征、文档类型特征等。

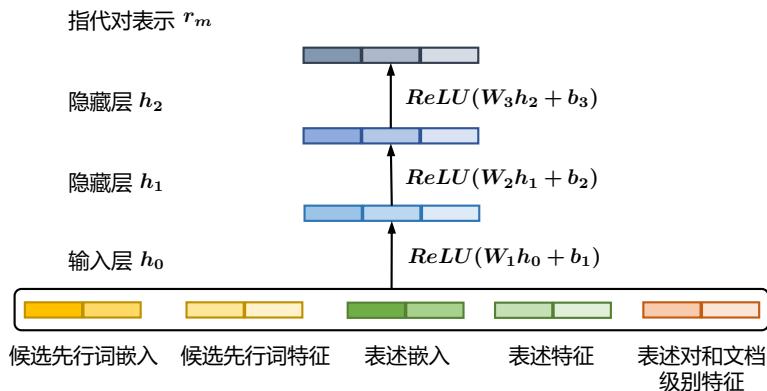


图 5.14 基于神经网络的表述对指代消解<sup>[247]</sup>

基于构建的输入特征, 使用前馈网络和 ReLU 激活层构造三层前馈神经网络:

$$\mathbf{h}_i(a, m) = \max(0, \mathbf{W}_i \mathbf{h}_{i-1}(a, m) + \mathbf{b}_i) \quad (5.28)$$

其中,  $\mathbf{h}_i(a, m)$  为输入表述对  $(a, m)$  的第  $i$  层输出特征。 $\mathbf{W}_i$  为第  $i$  层的权重矩阵,  $\mathbf{b}_i$  为第  $i$  层的偏置向量。最后, 模型得到第三层的输出特征  $\mathbf{h}_2(a, m)$  用于分类。

### 5.4.2 基于表述排序的指代消解

基于表述对的指代消解基于二分类器分别对每个先行词和当前指代构成的表述对进行预测。这种做法对于不同先行词的预测是相互独立的，只能判断每个先行词相对当前指代的合理程度，而无法直接通过比较判断哪个先行词是最正确的<sup>[248]</sup>。为了解决这一问题，研究人员提出了基于表述排序的指代消解算法，如图5.15所示。其基本思路是使用一个多分类器，基于多个先行词候选计算出分数最高的先行词。

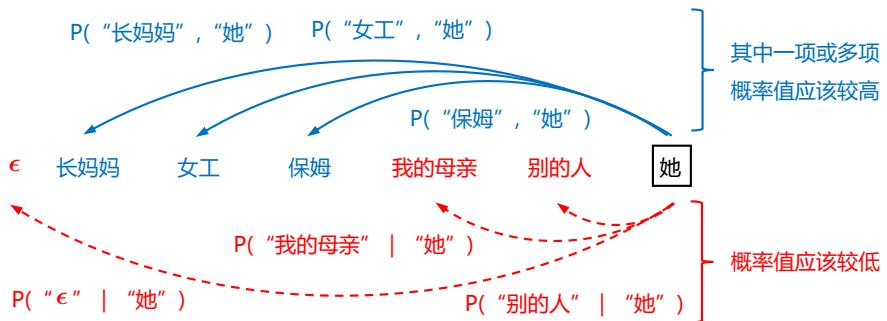


图 5.15 基于表述排序的指代消解示例

#### 1. 基于特征工程和最大熵分类器的表述排序指代消解

文献 [248] 介绍了基于表述排序的指代消解算法 RK，该方法将指代消解任务从基于指代对二分类器的多步推理（先分别计算各指代对的分数，再基于某种策略选出最高分的先行词），转化为同时计算并比较所有先行词候选的单步推理过程。具体来说，对每个先行词候选  $\alpha_i$ ，模型计算其为当前指代词  $\pi$  的被指代先行词的条件概率  $P_r(\alpha_i | \pi)$ ，从而对于每个指代词  $\pi$ ，通过比较多个候选先行词的条件概率即可以选出最可能和  $\pi$  构成指代关系的先行词：

$$P_r(\alpha_i | \pi) = \frac{\exp(\sum_{i=1}^n \lambda_i f_i(\pi, \alpha_i))}{\sum_k \exp(\sum_{i=1}^n \lambda_i f_i(\pi, \alpha_k))} \quad (5.29)$$

RK 算法通过最大熵分类器建模这一条件概率，其使用的特征包含三个类别：(1) 照应词特征，即描述待分类表述的特征，包括其代词类型特征、大小写特征等；(2) 候选先行词特征，即描述候选先行词的特征，包括其词性特征、其左右相关词的词性特征等；(3) 关系特征，即描述两个表述之间关系的特征，包括两个词之间的距离、两个词的语义相符性特征等。其使用的部分特征可见表5.1。

在训练时，RK 算法同样需要根据给定标注数据，为每个表述选取正负样本进行训练。对于任意表述  $\pi$ ，其正样本选取和  $\pi$  存在指代关系的距离最近的一个先行词。其负样本选取策略是：在

表 5.1 RK 算法使用的部分特征<sup>[248]</sup>

代词特征	
PERS_PRO	如果 $\pi$ 是人称代词则为 T, 否则为 F
POSS_PRO	如果 $\pi$ 是所有格代名词则为 T, 否则为 F
THIRD_PERS_PRO	如果 $\pi$ 是第三人称代词则为 T, 否则为 F
SPEECH_PRO	T 如果 $\pi$ 是第一或第二人称代词则为 T, 否则为 F
PRO_FORM	T 如果 $\pi$ 是小写字母组成的代词则为 T, 否则为 F
候选先行词特征	
ANTE_WD_LEN	$\alpha$ 中单词的数量
PRON_ANTE	如果 $\alpha$ 是代词则为 T, 否则为 F
PN_ANTE	如果 $\alpha$ 是专有名词则为 T, 否则为 F
INDEF_ANTE	如果 $\alpha$ 是无定名词短语 (indefinite NP) 则为 T, 否则为 F
DEF_ANTE	如果 $\alpha$ 是有定名词短语 (definite NP) 则为 T, 否则为 F
关系特征	
S_DIST	$\pi$ 和 $\alpha$ 之间句子数量的分桶值 (Binned values)
NP_DIST	$\pi$ 和 $\alpha$ 之间表述 (Mention) 数量的分桶值
NUM_AGR	如果 $\pi$ 和 $\alpha$ 在单复数形式上符合则为 T, 否则为 F
	如果 $\pi$ 或者 $\alpha$ 的单数复数形式不能确定则为 UNK
GEN_AGR	如果 $\pi$ 和 $\alpha$ 性别一致则为 T, 否则为 F
	如果 $\pi$ 或者 $\alpha$ 的性别不能确定则为 UNK

选定正样本后，以正样本为中心，选取窗口大小为 4 个句子内的所有和  $\pi$  不具有指代关系的表述作为负样本，其中，4 个句子包括  $\pi$  所在的句子、 $\pi$  所在句的前一个句子、 $\pi$  所在句的后两个句子。在训练过程中，模型需要最大化正样本的条件概率  $P_r(\alpha_i | \pi)$ ，而负样本则作为分子中的项被计算在损失函数中。

在测试时，考虑到大部分指代为局部指代，并为了节约测试时间，RK 算法只选取指代词  $\pi$  所在的句子及所在句之前的 3 个句子内的表述作为候选。模型基于所有候选词，计算每个词被选为和  $\pi$  构成指代关系的先行词的概率并选取概率最高的作为输出结果。

## 2. 基于循环神经网络的端到端的表述排序指代消解

E2E-COREF<sup>[249]</sup> 是端到端的指代消解模型，同样采用基于表述排序的方法实现。E2E-COREF 在训练时同时学习判断每个片段 (Span) 是否为实体表述并优化对实体表述的指代聚类。

具体来说，对于每个片段  $i$ ，模型的目标是在所有候选先行词中选出一个其指代的先行词  $y_i$ 。其中，先行词的候选集合为  $\mathcal{Y}(i) = \{\epsilon, 1, \dots, i - 1\}$ ，包括一个虚先行词  $\epsilon$  及所有在  $i$  之前的片段（需要注意的是，这里的片段可能不是实体表述）。当模型选择虚先行词  $\epsilon$  作为输出时，可能对应

两种情况：(1) 该片段  $i$  不是实体表述；(2) 该片段  $i$  是实体表述，但不指代在其之前的任一个片段（例如，可能是该实体在文中的第一次提及）。由此，根据对每个片段得到的先行词预测，可以构建出整个文本中的指代集合。

与 RK 算法相似，E2E-COREF 对每个候选先行词，计算其和片段  $i$  为指代关系的条件概率：

$$P(y_i) = \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))} \quad (5.30)$$

其中  $s(i, j)$  是表示片段  $i$  和片段  $j$  之间存在指代关系的分数，这一分数与三个因素相关： $s_m(i)$ ：片段  $i$  是否为实体表述； $s_m(j)$ ：片段  $j$  是否为实体表述； $s_a(i, j)$ ：片段  $j$  是否为  $i$  的先行词：

$$s(i, j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \end{cases} \quad (5.31)$$

通过将虚先行词的分数设为 0，当模型预测任意非虚先行词的分数为正时，则可以选出分数最高的先行词预测；当模型预测所有非虚先行词的分数都为负时，则输出虚先行词。接下来我们将分别介绍模型如何计算以上三个分数。

针对片段表示编码表示，E2E-COREF 首先基于双向 LSTM 网络和注意力机制编码片段表示，其网络结构如图 5.16 所示。对于每个输入句子，其输入表示  $\mathbf{x}_1, \dots, \mathbf{x}_T$  由预训练词向量及对字符的 1 维卷积组成。模型首先基于双向 LSTM 网络得到每个词的上下文表示  $\mathbf{x}_t^* = [\mathbf{h}_{t,1}, \mathbf{h}_{t,-1}]$ ，其中  $\mathbf{x}_t^*$  是 LSTM 的前向输出表示  $\mathbf{h}_{t,1}$  和反向输出表示  $\mathbf{h}_{t,-1}$  的拼接。基于词表示  $\mathbf{x}_t^*$ ，E2E-COREF 通过注意力机制判断片段中每个词的重要性，找出可能的关键词并给予更高权重，适用加权计算得到关键片段表示  $\hat{\mathbf{x}}_i$ ：

$$\alpha_t = \mathbf{w}_\alpha \cdot \text{FFNN}_\alpha(\mathbf{x}_t^*) \quad (5.32)$$

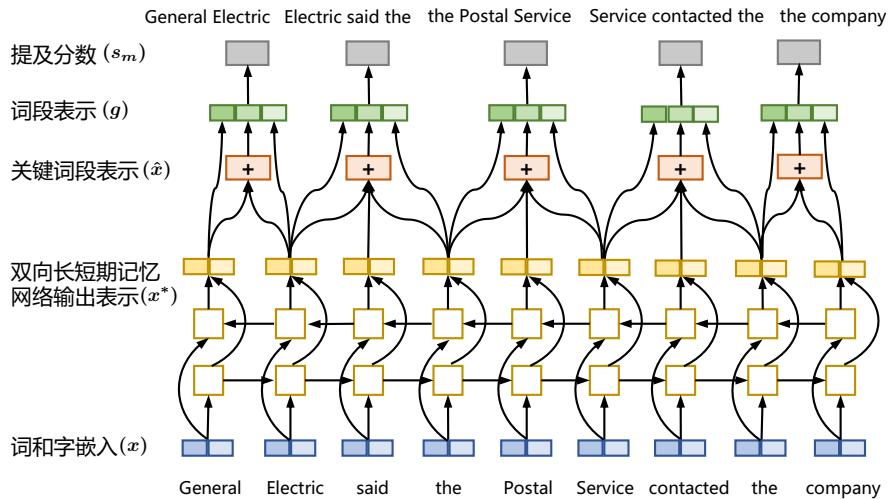
$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)} \quad (5.33)$$

$$\hat{\mathbf{x}}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot \mathbf{x}_t \quad (5.34)$$

其中，FFNN 表示前馈神经网络。

对于每个片段  $i$ ，其片段表示由其首尾词表示、关键片段表示及一个表示片段长度的特征向量  $\phi(i)$  的拼接构成：

$$\mathbf{g}_i = [\mathbf{x}_{\text{START}(i)}^*, \mathbf{x}_{\text{END}(i)}^*, \hat{\mathbf{x}}_i, \phi(i)] \quad (5.35)$$

图 5.16 E2E-COREF 基于双向 LSTM 的片段表示编码<sup>[249]</sup>

针对分数计算，E2E-COREF 基于上述得到的片段表示，使用前馈神经网络计算公式5.31中的各项分数，其网络结构如图5.17所示。分数具体计算公式如下：

$$\begin{aligned} s_m(i) &= \mathbf{w}_m \cdot \text{FFNN}_m(\mathbf{g}_i) \\ s_a(i, j) &= \mathbf{w}_a \cdot \text{FFNN}_a([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \circ \mathbf{g}_j, \phi(i, j)]) \end{aligned} \quad (5.36)$$

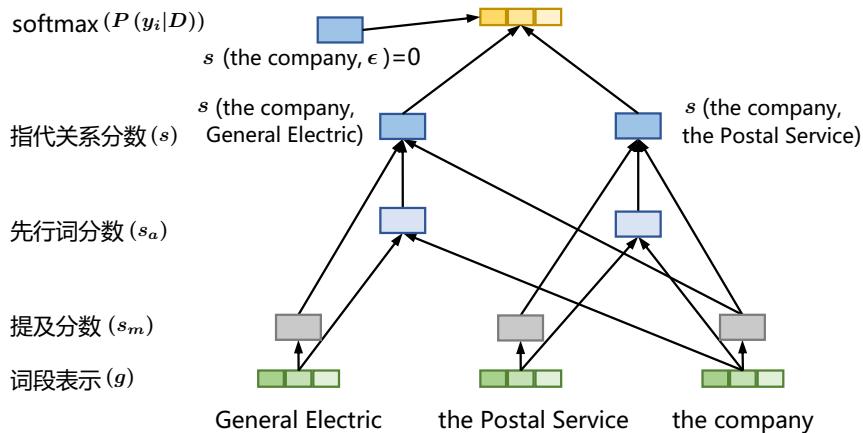
其中 · 表示点积运算，◦ 表示逐项乘积运算，[·] 表示向量拼接。可以看到，片段的实体表述分数  $s_m(i)$  与片段表示相关，而片段对的指代关系预测分数  $s_a(i, j)$  和两个片段表示、两个片段表示的乘积、及一个与话语主体、文本类别（从元数据中获得）和片段对距离相关的特征表示  $\phi(i, j)$  相关。

在训练阶段，对于每个片段  $i$ ，E2E-COREF 希望最大化与其具有正确指代关系的所有片段的条件概率  $P(y_i)$ 。假设  $\text{GOLD}(i)$  为与片段  $i$  的具有正确指代关系的片段集合， $\mathcal{Y}(i)$  为我们前面定义的先行词候选集合，则基于交叉熵损失，模型希望最小化的损失可以表示为：

$$-\sum_{i=2}^N \log \sum_{\hat{y} \in \mathcal{Y}(i) \cap \text{GOLD}(i)} P(\hat{y}) \quad (5.37)$$

需要注意的是，如果片段  $i$  不和任何片段构成指代关系，则其正确指代片段集合  $\text{GOLD}(i) = \{\epsilon\}$ 。

在测试时，考虑到对所有片段组合计算的效率问题，E2E-COREF 制定了以下策略：(1) 只考虑词数小于等于  $L$  的片段；(2) 在计算得到每个片段的实体表述分数  $s_m(i)$  后，只保留分数最高的  $\gamma T$  个片段进行后续计算；(3) 对于每个保留的片段，只考虑其最近的  $K$  个先行词候选进行片段对的分数计算。基于上述策略和公式5.30，E2E-COREF 能够解码得到最可能的指代关系分布。

图 5.17 E2E-COREF 的分数计算<sup>[249]</sup>

### 5.4.3 基于实体的指代消解

基于表述对和基于表述排序的两种指代消解算法旨在将一个表述与其所指代的一个表述相对应，通常只关注局部的指代信息。基于实体的指代消解则认为将单个表述归类至其指代的实体（通常对应一个表述的等价类）能利用实体级别的全局信息，因此能更好地实现指代消解任务。

基于实体的指代消解和基于表述的指代消解算法相似，区别在于基于表述的方法将当前表述分配到一个先行的表述，而基于实体的方法将当前表述分配到先行的实体（表述等价类）上。基于实体的指代消解同样可以分为基于实体-表述的方法和基于实体排序的方法。在本节中，我们将介绍基于 SVM 的实体-表述指代消解和基于循环神经网络的实体排序指代消解。

#### 1. 基于 SVM 分类器的实体-表述指代消解

Entity-mention-coref<sup>[250]</sup> 是一种基于实体-表述的指代消解算法。基于实体-表述的指代消解算法和本章第 5.4.1 节中所介绍的基于表述对的指代消解算法相似，都是通过训练二分类器对每个表述进行分类。不同之处在于，对于任意表述  $m_i$ ，基于表述对的指代消解算法仅关注其是否和某一候选先行词  $m_j$  为指代关系，并训练一个二分类器计算  $m_i$  和  $m_j$  为指代关系的分数  $s(m_j, m_i)$ 。而基于实体-表述的指代消解算法则关注表述  $m_i$  是否属于某个表述类  $c_j$ ，该表述类由指代同一实体的所有先行词构成。同样地，Entity-mention-coref 通过训练一个 SVM 二分类器计算  $m_i$  属于表述类  $c_j$  的分数  $s(c_j, m_i)$ ，其中  $c_j \in \mathcal{C}(i)$ ， $\mathcal{C}(i)$  代表在  $m_i$  之前的所有实体（表述类）的集合。

为了更好地表示输入数据，Entity-mention-coref 中所构建的特征包含两类：(1) 描述待分类表述  $m_i$  的表述级别特征：与第 5.4.1 节基于表述对的指代消解方法所构建的表述级别特征相似，包括其词性特征等；(2) 实体级别的特征：用于描述待分类表述  $m_i$  和候选实体  $c_j$  的关系，包括实体级别的性别、数字、语意相符程度，实体级别的距离特征，实体级别的字符串关系特征等。部

分构建的特征如表5.2所示。

表 5.2 Entity-mention-coref 使用的部分特征

描述待分类表述 $m_i$ 的表述级别特征	
NUMBER_2	SINGULAR 或 PLURAL, 根据词典确定
GENDER_2	MALE, FEMALE, NEUTER, or UNKNOWN, 根据常见人名列表确认
PRONOUN_2	如果 $m_k$ 是代词则为 Y, 否则为 N
待分类表述 $m_i$ 和其候选先行词 $m_k$ 的关系特征	
HEAD_MATCH	如果表述具有相同的中心名词则为 C, 否则为 I
STR_MATCH	如果表述具有相同的字符串则为 C, 否则为 I
SUBSTR_MATCH	如果一个表述是另一个的子字符串则为 C, 否则为 I
PRO_STR_MATCH	如果两个表示都是代词并且相同则为 C, 否则为 I
描述待分类表述 $m_i$ 和其候选先行词 $m_k$ 的关系的额外特征	
NUMBER'	$m_j$ 和 $m_k$ 的 NUMBER_2 特征合并
GENDER'	$m_j$ 和 $m_k$ 的 GENDER_2 特征合并
PRONOUN'	$m_j$ 和 $m_k$ 的 PRONOUN_2 特征合并

在训练时, 对于每个表述  $m_i$ , Entity-mention-coref 使用  $m_i$  之前的实体分别组成正、负训练样本, 其中正样本由  $m_i$  和其所属的实体  $c_j$  组成, 负样本由  $m_j$  到其所属实体距离  $m_i$  最近的先行词  $m_j$  (也即实体  $c_j$  的最后一个表述) 之间的表述和  $m_i$  组成。

例如: [Barack Obama]<sup>1</sup><sub>1</sub> nominated [Hillary Rodham Clinton]<sup>2</sup><sub>2</sub> as [[his]<sup>1</sup><sub>3</sub> secretary of state]<sup>3</sup><sub>4</sub> on [Monday]<sup>4</sup><sub>5</sub>. [He]<sup>1</sup><sub>6</sub>...

其中, 每个表述的下标代表其出现的次序, 上标代表其所属的实体。当对表述 “He” 进行分类时, 将产生三条训练样本: I({Monday},He), I({secretary of state},He) 和 I({Barack Obama, his},He)。其中, 前两条样本为负样本, 最后一条为正样本。

在测试时, 同样考虑  $m_i$  之前出现的所有实体, 并基于第5.4.2节中描述的最近原则, 将  $m_i$  分配至与其距离最近的被分类器判别为存在指代关系的实体。相反, 如果  $m_i$  和之前的所有实体都不存在指代关系, 则其将被认为是一个新的实体。

## 2. 基于循环神经网络的实体排序指代消解

Global-rank<sup>[251]</sup> 是基于循环神经网络的实体排序指代消解模型。基于实体排序的指代消解方法和本章第5.4.2节中所介绍的基于表述排序的方法相似。具体来说, 对于当前待分类表述  $x_n$ , 模型通过对其之前出现的所有实体 (表述类)  $\{X^{(m)}\}_{m=1}^M$  计算分数并排序, 选出  $x_n$  所属的表述类。

由于 Global-rank 是基于表述类的算法, 首先定义表述类的相关符号:  $X^{(m)}$  为第  $m$  个表述类,  $X_j^{(m)}$  为第  $m$  个表述类中的第  $j$  个表述, 其中表述的排序由其在文档中出现的顺序确定。由

于一个有效的文档表述聚类会将每个表述都分入一个确定的表述类，定义一组表述-表述类映射  $\mathbf{z} \in \{1, \dots, M\}^N$ ，当  $x_n$  属于第  $m$  个表述类时  $z_n = m$ 。

在本章第5.4.2节中，我们介绍了基于表述排序的指代消解，其基本思路是对每个  $x_n$  和候选先行词  $y \in \mathcal{Y}(x_n)$  计算其为指代关系的分数  $f(x_n, y)$ ，其中  $\mathcal{Y}(x_n) = \{1, \dots, n-1, \epsilon\}$  为  $x_n$  的候选先行词集合， $\epsilon$  为虚先行词，表示当前表述为该实体在文中的第一次提及。这样可以使得基于表述排序的算法能够高效计算和解码。然而，这种方式使用单个指代代表整个表述类，只能计算局部的表述分数，无法利用整个表述类的全局信息。因此，Global-rank 在局部表述分数  $f(x_n, y)$  的基础上，增添了一项全局实体-表述分数  $g(x_n, y, \mathbf{z}_{1:n-1})$  的计算，其中  $\mathbf{z}_{1:n-1}$  代表当前表述  $x_n$  之前所有表述所属的表述类的映射。具体实现时，为了保留基于表述排序的方法的优势，并简化测试过程，Global-rank 在解码时同时计算局部和全局的分数，并寻找使该分数最大的表述类分配方式：

$$\arg \max_{y_1, \dots, y_N} \sum_{n=1}^N f(x_n, y_n) + g(x_n, y_n, \mathbf{z}_{1:n-1}) \quad (5.38)$$

接下来我们将分别介绍局部表述分数  $f(x_n, y)$  和全局实体-表述分数  $g(x_n, y, \mathbf{z}_{1:n-1})$  的计算。

局部表述分数计算和第5.4.2节中介绍的表述排序分数计算相似，Global-rank 首先基于文献[252]中定义的特征映射  $\phi_a(x_n) : \mathcal{X} \rightarrow \{0, 1\}^F$  和  $\phi_p(x_n, y) : (\mathcal{X}, \mathcal{X}) \rightarrow \{0, 1\}^F$  将表述  $x_n$  和表述对  $(x_n, y)$  分别表示成表述和表述对级别的特征向量，接着使用非线性特征映射  $\mathbf{h}_a$  和  $\mathbf{h}_p$  将该特征向量映射到连续的特征空间：

$$\begin{aligned} \mathbf{h}_a(x_n) &= \tanh(\mathbf{W}_a \phi_a(x_n) + \mathbf{b}_a) \\ \mathbf{h}_p(x_n, y) &= \tanh(\mathbf{W}_p \phi_a(x_n, y) + \mathbf{b}_p) \end{aligned} \quad (5.39)$$

则局部表述分数计算定义如下：

$$f(x_n, y) = \begin{cases} \mathbf{u}^\top [\mathbf{h}_a(x_n), \mathbf{h}_p(x_n, y)] + u_0 & \text{if } y \neq \epsilon \\ \mathbf{v}^\top \mathbf{h}_a(x_n) + v_0 & \text{if } y = \epsilon \end{cases} \quad (5.40)$$

在计算全局的实体-表述分数之前，首先需要对实体表述类级别的表示进行计算。Global-rank 使用 RNN 对每个表述类所包含的表述进行依次编码，从而计算该实体表述类的整体表示。具体来说，首先将表述表示成和公式5.39相似的形式：

$$\mathbf{h}_c(x_n) = \tanh(\mathbf{W}_c \phi_a(x_n) + \mathbf{b}_c) \quad (5.41)$$

其中  $\mathbf{W}_c$  和  $\mathbf{b}_c$  为全局分数计算所对应的表示参数。

接着，对于表述类  $m$  中的第  $j$  个表述，基于 RNN 的表示计算可以表示成：

$$\mathbf{h}_j^{(m)} \leftarrow \text{RNN}(\mathbf{h}_c(X_j^{(m)}), \mathbf{h}_{j-1}^{(m)}; \boldsymbol{\theta}) \quad (5.42)$$

在对全局实体-表述分数进行计算时，使用  $\mathbf{h}_{<n}^{(z_y)}$  表示表述类  $z_y$  在对  $x_n$  之前所有的表述依次编码后得到的表述类表示，并将全局实体-表述分数定义成和局部表述分数计算相似的形式：

$$g(x_n, y, \mathbf{z}_{1:n-1}) = \begin{cases} \mathbf{h}_c(x_n)^\top \mathbf{h}_{<n}^{(z_y)} & \text{if } y \neq \epsilon \\ \text{NA}(x_n) & \text{if } y = \epsilon \end{cases} \quad (5.43)$$

其中，使用 NA 函数对该表述为首次出现的情况进行了分别计算：

$$\text{NA}(x_n) = \mathbf{q}^\top \tanh(\mathbf{W}_s[\phi_a(x_n), \sum_{m=1}^M \mathbf{h}_{<n}^{(m)}] + \mathbf{b}_s) \quad (5.44)$$

在训练时，每个表述所属的实体由训练集的标注所给定，但每个表述所指代的先行词可以有多个。因此，Global-rank 对每个表述，选取使其分数最高的先行词作为隐式目标先行词，并定义了最大间隔的目标函数：

$$\sum_{n=1}^N \max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y})(1 + f(x_n, \hat{y}) + g(x_n, \hat{y}, \mathbf{z}^{(o)}) - f(x_n, y_n^l), g(x_n, y_n^l, \mathbf{z}^{(o)})) \quad (5.45)$$

其中，当  $x_n$  为前指项时，隐式目标先行词定义为使局部和全局分数之和最高的先行词，否则为  $\epsilon$ ：

$$y_n^l = \arg \max_{y \in \mathcal{Y}(x_n): z_y^{(o)} = z_n^{(o)}} f(x_n, y) + g(x_n, y, \mathbf{z}^{(o)}) \quad (5.46)$$

$\Delta(x_n, \hat{y})$  表示针对文献 [253] 定义的不同错误类型“假链接（False link）”、“假新实体（False new）”、“错误链接（Wrong link）”定义不同的损失权重  $(\alpha_1, \alpha_2, \alpha_3)$ ，其中，“假链接”表示预测当前表述为前指，实际为新实体的错误；“假新实体”表示预测表述为新实体，实际为前指；“错误链接”为预测了错误的指代先行词。

Global-rank 的解码算法如算法5.2所示。对于每个待分类表述  $x_n$ ，首先计算得到使得局部和全局分数之和最高的先行词  $y^*$ ，并得到  $y^*$  所属的表述类  $m$ 。当  $y^* = \epsilon$  时，表示当前表述不属于任何表述类，则建立一个新的表述类。接着，更新表述类  $m$ ，表述类映射及表述类  $m$  的表示，以用于下次解码计算。对所有表述  $x_n$  遍历分类后，返还所有表述类。

---

代码 5.2: Global-rank 解码算法

---

```

输入: 待解码文本序列  $(x_1, \dots, x_N)$ 
输出: 实体表述类  $X^{(1)}, \dots, X^{(M)}$ 
// 初始化
foreach  $X^{(i)}$  do
|  $X^{(i)} \leftarrow []$ ; // 初始化每个表述类  $X^{(i)}$  为空列表;
end
foreach  $\mathbf{h}_0^{(i)}$  do
|  $\mathbf{h}_0^{(i)} \leftarrow \mathbf{0} \in \mathbb{R}^D$ ; // 初始化每个表述类的 RNN 初始输入  $\mathbf{h}_0^{(0)}$  为 0 向量;
end
 $z \leftarrow \mathbf{0}$ ; // 初始化表述类映射向量;
 $M \leftarrow 0$ ; // 初始化表述类计数;
// 解码过程
for  $n = 2$  to  $N$  do
|  $y^* \leftarrow \arg \max_{y \in \mathcal{Y}(x_n)} f(x_n, y) + g(x_n, y, z_{1:n-1})$ ; // 计算使局部和全局分数之和最高的先行
词;
|  $m \leftarrow z_{y^*}$ ; // 得到  $y^*$  所属的表述类;
| if  $y^* = \epsilon$  then
| | // 建立一个新的表述类
| |  $M \leftarrow M + 1$ ;
| |  $m \leftarrow M$ ;
| end
|  $X^{(m)} \leftarrow X^{(m)} + [x_n]$ ;
|  $z_n \leftarrow m$ ;
|  $\mathbf{h}^{(m)} \leftarrow \text{RNN}(\mathbf{h}_c(x_n), \mathbf{h}^{(m)})$ ; // 更新表述类  $m$  的表示;
end
return  $X^{(1)}, \dots, X^{(M)}$ 

```

---

## 5.5 延伸阅读

在本章中，我们介绍了话语分割、篇章结构分析和指代消解三个篇章分析任务。其中，话语分割和篇章结构分析帮助我们理解和分析篇章的结构，指代消解任务则帮助我们理解篇章的衔接，这些任务同时也直接影响许多下游自然语言处理应用。近年来，随着神经网络的发展，话语分割任务已经能够取得接近人类的准确率<sup>[254]</sup>，因此研究人员更多地将注意力转移到基于话语分割的篇章结构分析任务。而指代消解作为自然语言处理中另一重要任务，也长期受到研究人员的关注。

在 RST 篇章分析方面，现有的工作主要可以分为两类：基于转移（Transition-based）算法的

RST 篇章分析 [255–260]，和基于图（chart parsing）的 RST 篇章分析 [234, 261, 262]。为了进一步优化 RST 篇章分析算法，近年来的工作从不同角度提出了改进。一些研究工作引入联合建模篇章结构和篇章关系的方式，增加篇章结构和篇章关系建模的信息交互，从而优化篇章分析树 [257, 262]；为了利用更多全局信息，一些研究工作引入了自顶向下（top-down）的篇章分析方式 [263–265]，另外一些工作通过引入集束搜索（beam search）算法 [266, 267]、对抗训练 [268]、指针网络（pointer network） [269] 优化全局依赖。近年来，一些研究工作通过预训练语言模型优化 EDU 表示 [270–272]。

在浅层篇章分析方面，由于显式篇章分析任务已经取得较高的准确率 [273]，现有研究工作主要集中在针对隐式篇章分析的提升。除了本章介绍的算法外，为了解决标注数据稀缺的问题，一些研究工作利用数据中的显式关系实现数据增强，以提升隐式篇章分析性能，例如利用显示关系构建隐式篇章关系的弱监督数据 [274]、利用显示关系学习篇章关系特有词表示 [275]、基于多任务学习利用显式关系信息 [276] 等。一些研究工作探索更好地利用连接词信息以提升性能，例如预测显式连接词并将预测结果作为信息引入隐式关系预测 [277]、利用对抗网络引入连接词信息 [278] 等。还有一些研究工作通过优化文本表示角度提升隐式篇章分析性能，例如引入预训练词向量 [279]、引入多粒度文本表示 [280]、利用预训练语言模型增强表示 [281] 等。

在指代消解方面，除了本章所介绍的研究工作外，还有一些研究工作从如何利用外部知识或相关任务实现知识增强 [282]、利用图神经网络改进表示编码 [283]、探索更有效的指代消解新范式 [284]、如何更好地利用实体的全局信息 [285] 等方面展开了深入研究。一些研究工作针对指代消解系统在实际应用时的泛化能力，从引入语言学特征 [286]、引入对抗训练 [287] 等方面实现改进。随着预训练语言模型的兴起，许多研究工作将预训练语言模型融入指代消解任务 [288]，并针对片段（span）级别的指代消解方法展开了深入研究 [289, 290]。还有一些研究工作关注指代消解在实际应用中的效率问题，针对在线文本处理中涉及的内存拓展问题，通过引入记忆网络维护固定内存 [291]、引入增量式聚类算法 [292]、在内存中维护少量有效实体 [293] 等方式实现改进，同时使模型能够实现针对更长文档的指代消解 [293]。

## 5.6 习题

- (1) 如何判断一段文本是一个篇章还是孤立句子的集合？
- (2) 试举例说明照应的几种情况。
- (3) 试举例说明什么是篇章的局部连贯性和整体连贯性，并说明它们之间的区别。
- (4) 除了本章介绍的基于循环神经网络的话语分割系统外，还有什么网络可以实现序列标注范式的话语分割？
- (5) 基于修辞结构理论的篇章结构分析和基于词汇化树形连接语法理论的篇章结构分析有何不同？二者各有什么难点？
- (6) 试比较基于表述对、基于表述排序及基于实体的三种指代消解系统的优缺点。

## 6. 语言模型

---

语言模型目标是建模自然语言的概率分布，在自然语言处理研究中具有重要的作用，是机器翻译、语音识别、输入法、句法分析等任务的支撑。语言模型是自然语言处理基础任务之一，大量的研究从  $n$  元语言模型 ( $n$ -gram Language Models) 和神经语言模型 (Neural Language Models) 等不同角度开展了系列工作。由于语言模型可以为自然语言的表示学习提供天然的自监督训练目标，近年来，预训练语言模型 (Pre-trained Language Models) 做为通用的基于深度神经网络的自然语言处理算法的基础工具，受到越来越多的重视。大规模的预训练语言模型对于提升各类自然语言处理任务的效果起到了重要作用。

本章首先介绍语言模型的基本概念，在此基础上介绍  $n$  元语言模型、神经网络语言模型以及预训练语言模型。

### 6.1 语言模型概述

语言模型 (Language Model, LM) 目标是构建词序列  $w_1 w_2 \dots w_m$  的概率分布  $P(w_1 w_2 \dots w_m)$ ，即计算给定的词序列  $w_1 w_2 \dots w_m$  作为一个句子出现的可能性大小。词汇表  $\mathbb{V}$  上的语言模型由函数  $P(w_1 w_2 \dots w_m)$  表示，对于任意词串  $w_1 w_2 \dots w_m \in \mathbb{V}^+$ ，则有  $P(w_1 w_2 \dots w_m) \geq 0$ ，并且对于所有词串，函数  $P(w_1 w_2 \dots w_m)$  满足归一化条件  $\sum_{w_1 w_2 \dots w_m \in \mathbb{V}^+} P(w_1 w_2 \dots w_m) = 1$ 。 $P(w_1 w_2 \dots w_m)$  是定义在  $\mathbb{V}^+$  上的概率分布。

由于联合概率  $P(w_1 w_2 \dots w_m)$  的参数量十分巨大，直接计算  $P(w_1 w_2 \dots w_m)$  非常困难。如果把  $w_1 w_2 \dots w_m$  看作一个变量，那么它具有  $|\mathbb{V}|^m$  种可能，其中  $m$  代表句子的长度， $|\mathbb{V}|$  表示词表中单词的数量。按照《现代汉语词典（第七版）》包含 7 万词条，句子长度按照 20 个词计算，模型参数量达到  $7.9792 \times 10^{96}$  的天文数字。中文的书面语中超过 100 个单词的句子也并不罕见，如果要将所有可能都纳入考虑，模型的复杂度还会进一步急剧增加，无法进行存储和计算。

为了减少  $P(w_1 w_2 \dots w_m)$  模型参数量，可以利用句子序列通常情况下从左至右的生成过程进

行分解，使用链式法则得到：

$$\begin{aligned} P(w_1 w_2 \dots w_m) &= P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \cdots P(w_m | w_1 w_2 \dots w_{m-1}) \\ &= \prod_{i=1}^m P(w_i | w_1 w_2 \cdots w_{i-1}) \end{aligned} \quad (6.1)$$

由此， $w_1 w_2 \dots w_m$  的生成过程可以看作单词逐个生成的过程。首先生成  $w_1$ ，之后根据  $w_1$  生成  $w_2$ ，再根据  $w_1$  和  $w_2$  生成  $w_3$ ，以此类推，根据前  $m-1$  个单词生成最后一个单词  $w_m$ 。通过上述过程将联合概率  $P(w_1 w_2 \dots w_m)$  转换为了多个条件概率的乘积。

例如：对于句子“把努力变成一种习惯”的概率计算，使用公式6.1可以转化为：

$$\begin{aligned} P(\text{把 努力 变成 一 种 习 惯}) &= P(\text{把}) \times P(\text{努力}|\text{把}) \times P(\text{变 成}|\text{把 努 力}) \times \\ &\quad P(\text{一 种}|\text{把 努 力 变 成}) \times P(\text{习 惯}|\text{把 努 力 变 成 一 种}) \end{aligned} \quad (6.2)$$

但是，仅通过上述过程模型的参数量依然没有下降， $P(w_m | w_1 w_2 \dots w_{m-1})$  的参数量依然是天文数字。然而基于上述转换，我们可以进一步的对模型进行简化， $n$  元语言模型就是其中一种常见的简化方法。本章接下来的章节将对如何估计  $n$  元语言模型参数值以及模型平滑技术进行详细介绍。

由于高阶  $n$  元语言模型还是会面临十分严重的数据稀疏问题，并且单词的离散表示也忽略了单词之间的相似性。因此，基于分布式表示和神经网络的语言模型逐渐成为了研究的热点。Bengio 等人在 2000 年提出了使用前馈神经网络对  $P(w_i | w_{i-n+1} \dots w_{i-1})$  进行估计的语言模型<sup>[173]</sup>。此后，循环神经网络<sup>[22]</sup>、卷积神经网络<sup>[294]</sup>、端到端记忆网络<sup>[295]</sup>等神经网络方法都成功应用于语言模型建模。相较于  $n$  元语言模型，神经网络方法可以在一定程度上避免数据稀疏问题，有些模型还可以避免对历史长度的限制，从而更好的建模长距离依赖关系。在本章中，我们也将对常见的基于神经网络的语言模型进行介绍。

语言模型的训练过程虽然采用的有监督方法，但是由于训练目标可以通过原始文本直接获得，从而使得模型的训练仅需要大规模无标注文本即可。语言模型也成为了典型的自监督学习（Self-supervised Learning）任务。互联网的快速发展，使得大规模无标注文本非常容易获取，因此训练超大规模的基于神经网络的语言模型成为了可能。2018 年艾伦人工智能研究所（Allen Institute for AI）Peters 等人提出了使用大规模语料，利用语言模型任务获取单词更好的表示的方法 ELMo<sup>[28]</sup>，在多个自然语言处理任务上得到了很好的效果。此后，谷歌公司 Devlin 等人在 2018 年提出了基于 Transformer 模型和掩码语言模型的方法 BERT<sup>[29]</sup>，在包括阅读理解、语义匹配等在内的多个自然语言处理任务中取得了更大幅度的提升，开启了大规模预训练语言模型研究热潮。2021 年谷歌开发的 Switch Transformer 模型参数量首次超过万亿。此后不久，北京智源研究院所就发布参数量超过 1.75 万亿的预训练模型“悟道 2.0”。本章也将介绍采用单向、双向、掩码语言模型的常见预训练方法。

## 6.2 n 元语言模型

语言模型通常用于反映一个句子出现的可能性，给定由单词序列  $w_1 w_2 \dots w_n$  组成的句子  $S$ ，可以利用语言的特性，使用链式法分解则得到：

$$P(S) = \prod_{i=1}^n P(w_i | w_1 w_2 \dots w_{i-1}) \quad (6.3)$$

其中，词  $w_i$  出现的概率受它前面的  $i-1$  个词  $w_1 w_2 \dots w_{i-1}$  影响，我们将这  $i-1$  个词  $w_1 w_2 \dots w_{i-1}$  称之为词  $w_i$  的历史。如果历史单词有  $i-1$  个，那么可能的单词组合就有  $|V|^{i-1}$  种，其中  $V$  表示单词词表， $|V|$  表示词表的大小。为了简化起见，使用  $w_1^{i-1}$  表示  $w_1 w_2 \dots w_{i-1}$ 。最简单的根据语料库对  $P(w_i | w_1 w_2 \dots w_{i-1})$  进行估计的方法是基于词序列在语料中出现次数（也称为频次）的方法。

$$P(w_i | w_1 w_2 \dots w_{i-1}) = \frac{C(w_1 w_2 \dots w_{i-1} w_i)}{C(w_1 w_2 \dots w_{i-1})} \quad (6.4)$$

其中， $C(\cdot)$  表示在语料库中词序列在语料库中出现次数。这种方法称为最大似然估计（Maximum Likelihood Estimation, MLE）。随着历史单词数量的增长，这种建模方式所需的数据量会指数级增长，这一现象称为维数灾难（Curse of Dimensionality）。并且，随着历史单词数量增多，绝大多数的历史并不会在训练数据中出现，这也意味着  $P(w_i | w_1 w_2 \dots w_{i-1})$  就很可能为 0，使得概率估计失去了意义。

为了解决上述问题，可以进一步假设任意单词  $w_i$  出现的概率只与过去  $n-1$  个词相关，即：

$$\begin{aligned} P(w_i | w_1 w_2 \dots w_{i-1}) &= P(w_i | w_{i-(n-1)} w_{i-(n-2)} \dots w_{i-1}) \\ P(w_i | w_1^{i-1}) &= P(w_i | w_{i-n+1}^{i-1}) \end{aligned} \quad (6.5)$$

满足上述条件的模型被称为  $n$  元语法或  $n$  元文法( $n$ -gram) 模型。其中  $n$ -gram 表示  $n$  个连续单词构成的单元，也被称为  $n$  元语法单元。 $n$  的取值越大，其历史信息越完整，但参数量也会随之增大。实际应用中， $n$  的取值通常小于等于 3。当  $n=1$  时，每个词  $w_i$  的概率独立于历史，称为一元语法（Unigram）。当  $n=2$  时，词  $w_i$  只依赖前一个词  $w_{i-1}$ ，称为二元语法（Bigram），又被称为一阶马尔可夫链。当  $n=3$  时，词  $w_i$  只依赖于前两个历史词  $w_{i-1}$  和  $w_{i-2}$ ，称为三元语法（Trigram），又被称为二阶马尔可夫链。

以二元语法为例，一个词的概率只依赖于前一个词，则句子  $S$  的出现概率可以表示为：

$$P(S) = \prod_{i=1}^n P(w_i | w_{i-1}) \quad (6.6)$$

为了使  $i < 2$  时上式也成立，通常在句子开头加上句首标识  $\langle \text{BOS} \rangle$ ，使  $w_0$  为  $\langle \text{BOS} \rangle$ 。此外，句

子结尾也会添加句尾标记 <EOS>。我们还是以计算句子“把努力变成一种习惯”的概率为例，其计算可以转化为：

$$\begin{aligned} P(\text{<BOS>} \text{ 把 努力 变成 一 种 习 惯 } \text{ <EOS>}) = & P(\text{<BOS>}|\text{把}) \times P(\text{努力}|\text{把}) \times P(\text{变成}|\text{努力}) \times \\ & P(\text{一 种}|\text{变成}) \times P(\text{习 惯}|\text{一 种}) \end{aligned} \quad (6.7)$$

对比公式6.2和公式6.7，可以看到语言模型计算通过  $n$  元语法假设进行了大幅度的简化。

尽管  $n$  元语言模型能缓解句子概率为 0 的问题，但语言是由人和时代创造的，具备无穷的可能性，再庞大的训练语料也无法覆盖所有的  $n$ -gram，而训练语料中的零频率并不代表零概率。因此，需要使用平滑技术（Smoothing）来解决这一问题，对所有可能出现的字符串都分配一个非零的概率值，从而避免零概率问题。平滑是指为了产生更合理的概率，对最大似然估计进行调整的一类方法，也称为数据平滑（Data Smoothing）。平滑处理的基本思想是提高低概率，降低高概率，使整体的概率分布趋于均匀。本节将介绍三种常用的平滑技术。

### 6.2.1 加法平滑

G.J.Lidstone, W.E.Johnson 和 H.Jeffrey 提出的加法平滑（Additive Smoothing）是实际运用中最常用的平滑技术之一。其思想是假设事件出现的次数比实际出现的次数多  $\delta$  次。以二元语法模型为例，其平滑后的条件概率为：

$$P(w_i|w_{i-1}) = \frac{\delta + C(w_{i-1}w_i)}{\sum_{w_i} \delta + C(w_{i-1}w_i)} = \frac{\delta + C(w_{i-1}w_i)}{\delta |\mathbb{V}| + C(w_{i-1})} \quad (6.8)$$

其中  $0 \leq \delta \leq 1$ ， $V$  是训练语料中所有单词的集合， $C(w_i)$  表示单词  $w_i$  出现的次数， $C(w_{i-1}w_i)$  代表词  $w_{i-1}$  和词  $w_i$  同时出现次数。可以进一步将公式6.8拓展到  $n$  元语言模型上：

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{\delta + C(w_{i-n+1}^i)}{\delta |\mathbb{V}| + \sum_{w_i} C(w_{i-n+1}^i)} \quad (6.9)$$

当  $\delta = 1$  时，该方法又称为加一平滑。此外，针对所有不在词表  $V$  中的单词，可以统一映射为一个特定的词汇，从而保证所有情况下都不存在非零概率。

### 6.2.2 古德-图灵估计法

古德-图灵估计法（Good-Turing Estimate）<sup>[296]</sup> 是 1953 年由 I.J.Good 基于图灵（Turning）的方法提出的，是多种平滑技术的核心。该方法基于的核心思想是将一部分已知事件的概率分配给未见的事件。对于  $n$  元语言模型来说，降低出现次数多的  $n$ -gram，同时将剩余概率分配给未出现的

$n$ -gram。具体来说，对于任意一个出现了  $r$  次的  $n$  元语法，按照如下公式修改为出现了  $r^*$  次：

$$r^* = (r + 1) \frac{n_r + 1}{n_r} \quad (6.10)$$

其中， $n_r$  代表有  $n_r$  个  $n$ -gram 在训练语料中出现了  $r$  次。对其进行归一化后，即可得到出现  $r$  次的  $n$  元语法概率：

$$p_r = \frac{r^*}{N} \quad (6.11)$$

其中， $N = \sum_{r=0}^{\infty} n_r r^*$ ，即  $N$  为分布中最初的计数，样本中所有事件的概率之和为：

$$\sum_{r>0} n_r p_r = 1 - \frac{n_1}{N} < 1 \quad (6.12)$$

对于  $r = 0$  的未见事件，有  $\frac{n_1}{N}$  的概率余量可以用于分配。

表6.1给出了一个使用 Good-Turing 方法进行估计的样例。通过语料库统计 Bigram 的出现次数，并通过公式6.10修正后得到的  $r^*$  以及通过公式6.11得到修正后的概率  $p_r$ 。对于没有出现过的 Bigram 的概率总和为： $p_0 = \frac{n_1}{N}$ 。可以根据词表  $\mathbb{V}$ ，估计未出现的 Bigram 的总数  $n_0 = |\mathbb{V}|^2 - \sum_{r>0} n_r$ 。将  $p_0$  根据  $n_0$  均分，可以得到未出现过得 Bigram 的概率值。

表 6.1 Good-Turing 估计方法样例

$r$	$n_r$	$r^*$	$p_r$
1	3286	0.448	$5.220 \times 10^{-5}$
2	736	1.25	$1.454 \times 10^{-4}$
3	306	2.25	$2.620 \times 10^{-4}$
4	172	3.17	$3.693 \times 10^{-4}$
5	109	4.18	$4.875 \times 10^{-4}$
6	76	5.53	$6.440 \times 10^{-4}$
7	60	5.73	$6.681 \times 10^{-4}$
8	43	5.86	$6.830 \times 10^{-4}$
9	28	7.14	$8.324 \times 10^{-4}$
10	20	—	—

古德-图灵估计法的缺点是其无法用于估计  $n_r = 0$  的  $n$  元语法概率，并且其不能用于高阶语言模型和低阶语言模型的结合，而高阶与低阶模型的结合通常能带来更好的平滑效果。但古德-图灵方法的思想简单普适，因此其往往是作为一种基本方法与其他的平滑方法结合。

### 6.2.3 Katz 平滑

Katz 平滑是 1987 年由 S. M. Katz 所提出的后备（back-off）平滑方法<sup>[297]</sup>，其在古德-图灵估计

法的基础上引入了高阶模型与低阶模型的结合。Katz 平滑法的基本思想是将因减值获得的概率余量根据低阶模型的分布分配给未见事件，而不是进行平均分配，从而令低概率事件有更合理的概率分布。Katz 平滑法的做法是，当事件在样本中出现的频次大于某一数值  $k$  时，运用最大似然估计法，通过减值来估计其概率值；而当事件的频次小于  $k$  值时，使用低阶的语法模型作为代替高阶语法模型的后备。

下面以二元语法模型为例说明 Katz 平滑方法的实现方法。对于一个出现次数为  $r$  的二元语法  $w_{i-1}^i$ ，用下列公式对其次数进行修正：

$$C_{katz}(w_{i-1}^i) = \begin{cases} d_r \frac{C(w_{i-1} w_i)}{C(w_{i-1})}, & r > 0 \\ a(w_{i-1}) P_{ML}(w_i), & r = 0 \end{cases} \quad (6.13)$$

其中  $d_r \approx \frac{r^*}{r}$  是由古德-图灵估计法预测的折扣率。可以看出，所有具有非零计数  $r$  的二元语法都根据折扣率  $d_r$  被减值了。从非零计数中减去的计数量，根据低一阶的分布，即一元语法模型，被分配给了计数为零的二元语法。式中  $P_{ML}(w_i)$  为  $w_i$  的最大似然估计概率， $a(w_{i-1})$  使分布中总计数保持不变，即  $\sum_{w_i} C_{katz}(w_{i-1}^i) = \sum_{w_i} C(w_{i-1} w_i)$ 。 $a(w_{i-1})$  的值通常按照如下公式估计：

$$a(w_{i-1}) = \frac{1 - \sum_{w_i: C(w_{i-1} w_i) > 0} P_{katz}(w_i | w_{i-1})}{\sum_{w_i: C(w_{i-1}^i) = 0} P_{ML}(w_i)} = \frac{1 - \sum_{w_i: C(w_{i-1}^i) > 0} P_{katz}(w_i | w_{i-1})}{1 - \sum_{w_i: C(w_{i-1}^i) > 0} P_{ML}(w_i)} \quad (6.14)$$

根据修正的计数计算概率  $P_{katz}(w_i | w_{i-1})$ ，需要按照如下公式进行归一化：

$$P_{katz}(w_i | w_{i-1}) = \frac{C_{katz}(w_{i-1}^i)}{\sum_{w_i} C_{katz}(w_{i-1}^i)} \quad (6.15)$$

折扣率  $d_r$  需要满足两个约束条件：(1) 保证总折扣量和古德-图灵估计得到的减值量成比例，即保证对于常数  $\mu, r \in \{1, 2, \dots, k\}$ ，如以下公式所示：

$$1 - d_r = \mu \left(1 - \frac{r^*}{r}\right) \quad (6.16)$$

(2) 保证二元语法分布中被折扣的计数总量等于古德-图灵估计得到的次数为零的 Bi-gram 总数  $n_0 0^* = n_0 \frac{n_1}{n_0} = n_1$ ，相当于：

$$\sum_{r=1}^k n_r (1 - d_r) r = n_1 \quad (6.17)$$

上述公式6.16和公式6.17的唯一解为：

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \quad (6.18)$$

在二元语法的基础上可以将 Katz 平滑算法拓展到高阶  $n$  元语法模型。类似于公式6.13，可以根据一元语法模型定义二元语法模型，Katz 的  $n$  元语法模型可以根据  $n-1$  元语法模型定义：

$$P(w_i|w_{i-n+1}^{i-1}) = \begin{cases} P_{GT}(w_i|w_{i-n+1}^{i-1}), & C(w_{i-n+1}^i) > 0 \\ a(w_{i-n+1}^{i-1})P_{GT}(w_i|w_{i-n+2}^{i-1}), & C(w_{i-n+1}^i) = 0 \text{ 并且 } C(w_{i-n+1}^{i-1}) > 0 \\ P_{BF}(w_i|w_{i-n+2}^{i-1}), & C(w_{i-n+1}^{i-1}) = 0 \end{cases} \quad (6.19)$$

其中， $P_{BF}$  和  $P_{GT}$  分别代表后备法和吉德-图灵估计法计算得到的概率值。 $a(w_{i-n+1}^{i-1})$  定义为：

$$\begin{aligned} a(w_{i-n+1}^{i-1}) &= \frac{1 - \sum_{w_i: C(w_{i-n+1}^i) > 0} P_{GT}(w_i|w_{i-n+1}^{i-1})}{\sum_{w_i: \{C(w_{i-n+1}^i) = 0 \& C(w_{i-n+1}^{i-1}) > 0\}} P_{GT}(w_i|w_{i-n+2}^{i-1})} \\ &= \frac{1 - \sum_{w_i: C(w_{i-n+1}^i) > 0} P_{GT}(w_i|w_{i-n+1}^{i-1})}{1 - \sum_{w_i: C(w_{i-n+1}^i) > 0} P_{GT}(w_i|w_{i-n+2}^{i-1})} \end{aligned} \quad (6.20)$$

满足以下约束：

$$\sum_{w_i: \{C(w_{i-n+1}^i) = 0 \& C(w_{i-n+1}^{i-1}) > 0\}} P_{BF}(w_i|w_{i-n+1}^{i-1}) + \sum_{w_i: C(w_{i-n+1}^i) > 0} P_{BF}(w_i|w_{i-n+1}^{i-1}) = 1 \quad (6.21)$$

## 6.2.4 平滑方法总结

除了我们在上述章节介绍的平滑算法之外，研究人员们提出了有很多针对语言模型的平滑算法，包括 Jelinek-Mercer 平滑算法<sup>[298]</sup>、Witten-Bell 平滑算法<sup>[299]</sup>、Kneser-Ney 平滑算法<sup>[300]</sup>等。这些方法的核心思想大都可以归纳为：如果  $n$ -gram 存在则使用其本身计数，如果不存在再退后到低阶分布。可以用如下公式表示：

$$P_{\text{smooth}}(w_i|w_{i-n+1:i-1}) = \begin{cases} \alpha(w_i|w_{i-n+1}^{i-1}), & C(w_{i-n+1}^i) > 0 \\ \gamma(w_{i-n+1}^{i-1})P_{\text{smooth}}(w_i|w_{i-n+2}^{i-1}), & C(w_{i-n+1}^i) = 0 \end{cases} \quad (6.22)$$

在此基础上，一些平滑算法采用高阶和低阶  $n$  元语法模型的线性插值的方法，融合高阶和低

阶语法的估计，如以下公式所示：

$$P_{\text{smooth}}(w_i|w_{i-n+1}^{i-1}) = \lambda P_{\text{ML}}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda) P_{\text{smooth}}(w_i|w_{i-n+2}^{i-1}) \quad (6.23)$$

$n$  语法模型整体上来看与训练语料规模和模型的阶数有较大的关系，不同的平滑算法在不同情况下的表现有较大的差距。平滑算法虽然较好的解决了零概率问题，但是基于稀疏表示的  $n$  元语言模型仍然有三个较为明显的缺点：(1) 无法建模长度超过  $n$  的上下文；(2) 依赖人工设计规则的平滑技术；(3) 当  $n$  增大时，数据的稀疏性随之增大，模型的参数量更是指数级增加，并且模型受到数据稀疏问题的影响，其参数难以被准确的学习。

## 6.3 神经网络语言模型

随着深度神经网络的发展，利用神经网络的语言模型展现出了比  $n$  元语言模型更强学习能力。神经网络先进的结构使其能有效的建模长距离上下文依赖，以词向量（Word Embedding）为代表的分布式表示的语言模型深刻地影响了自然语言处理领域的其他模型与应用的变革<sup>①</sup>。因此， $n$  元语言模型几乎已经被神经网络的语言模型所替代。本节将介绍如何使用经典的前馈神经网络和循环神经网络来建模语言模型。

### 6.3.1 前馈神经网络语言模型

给定历史单词序列  $w_1 w_2 \dots w_{i-1}$ ，神经网络语言模型的目标是根据历史单词对下一时刻词进行预测。与传统  $n$  元语言模型类似，前馈神经网络语言模型<sup>[173]</sup>沿用了马尔可夫假设，即下一时刻的词只与过去  $n-1$  个词相关，其目标可以表示为输入历史单词  $w_{(i-n+1)} \dots w_{i-1}$ ，输出词  $w_i$  在词表  $\mathbb{V}$  上的概率分布，即估计条件概率  $P(w_i|w_{(i-n+1)}^{i-1})$ 。

前馈神经网络由三部分组成，如图6.1所示，分别为输入层、隐藏层和输出层。历史词序列首先经过输入层被转换为离散的独热编码，随后每个词的独热编码被映射为一个低维稠密的实数向量；隐藏层对词向量层的输出进行编码，进行多次线性变换与非线性映射；最后，隐藏层向量经过输出层被映射到词表空间，再利用 Softmax 函数得到其词表上的概率分布。

输入层的目标是将由文本组成的词序列转化为模型可接受的低维稠密向量。在具体实现中，模型首先根据每个词在词表  $\mathbb{V}$  中的位置，将历史词序列  $w_{(i-n+1)}, \dots, w_{i-1}$  转化为对应的独热编码（One-Hot Encoding），再将每个词的独热编码映射到一个低维稠密的实数向量。该映射可以视作是根据一个查找表，获取每个词特有的词向量的过程：

$$\mathbf{v} = [\mathbf{v}_{(i-n+1)}, \dots, \mathbf{v}_{i-1}] \quad (6.24)$$

其中  $\mathbf{v}_{i-1} \in \mathbb{R}^d$  代表词  $w_{i-1}$  所对应的词向量， $d$  代表词向量的维度， $\mathbf{v} \in \mathbb{R}^{(n-1) \times d}$  代表将所有历

---

<sup>①</sup> 详见本书第 4 章分布式表示

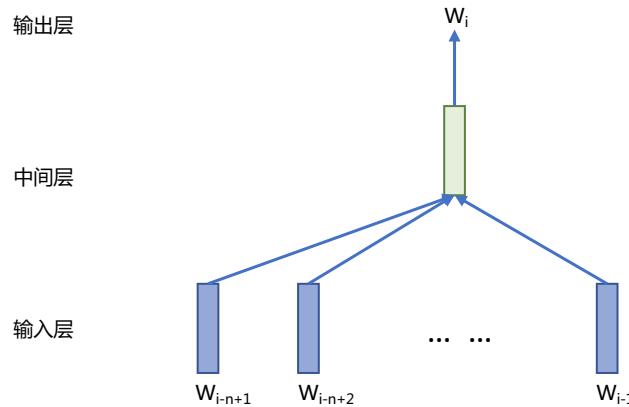


图 6.1 基于前馈神经网络的语言模型结构图

史词向量拼接后的结果。

隐藏层的目标是对词向量  $v$  进行线性变换与非线性映射。具体计算过程可以使用如下公式表示：

$$\mathbf{h} = f(\mathbf{W}^{hid}\mathbf{v} + \mathbf{b}^{hid}) \quad (6.25)$$

其中，隐藏层由线性变换矩阵  $\mathbf{W}^{hid} \in \mathbb{R}^{m \times (n-1)d}$ 、偏置项  $\mathbf{b}^{hid} \in \mathbb{R}^m$  组成， $m$  为隐藏层维度， $f$  为非线性激活函数，常见激活函数的有 Sigmoid、tanh 和 ReLU 等。

输出层的目标是基于隐藏层向量  $\mathbf{h}$  得到词表空间  $\mathbb{V}$  上的概率分布。输出层的计算可以用如下公式表示：

$$\mathbf{y} = \text{Softmax}(\mathbf{W}^{out}\mathbf{h} + \mathbf{b}^{out}) \quad (6.26)$$

其中， $\mathbf{W}^{out} \in \mathbb{R}^{|\mathbb{V}| \times m}$  是输出层的线性变换矩阵， $\mathbf{b}^{out}$  为偏置项， $|\mathbb{V}|$  为词表大小。

上述前馈神经网络语言模型的总参数量为  $|\mathbb{V}| \times d + m \times (n-1)d + m + |\mathbb{V}| \times m + |\mathbb{V}|$ ，即  $|\mathbb{V}|(d+m+1) + m((n-1)d+1)$ 。词向量维度  $d$ ，隐藏层维度  $m$  和历史词长度  $n-1$  可以根据实际需求进行调整。可以看出，词表大小  $|\mathbb{V}|$  和历史词长度  $n-1$  的增大并不会显著增加前馈神经网络语言模型的总参数量，而是维持着线性增长的关系，这也是神经网络模型优于  $n$  元语言模型重要方面。

### 6.3.2 循环神经网络语言模型

在实际场景下，固定长度的历史词并不是总能提供充分的信息，对于信息较为复杂的长文本，模型需要依赖较长的历史才能做出准确预测。

例如：与小明一起旅行游玩总是充满了惊喜，你永远不会知道他将要带你到哪里去。  
模型需要获取“小明”这一信息才能对“他”进行准确的预测。而这两个词语之间的距离接近 15 个单

词，如果采用前馈神经网络，需要的历史词长度过长。

循环神经网络（Recurrent Neural Network, RNN）<sup>[22]</sup> 常用于处理序列结构的数据，其特点是上一时刻的模型隐藏层状态会作为当前时刻模型的输入，每一时刻的隐藏层状态都会维护所有过去词的信息。循环神经网络语言模型不再基于马尔可夫假设，每个时刻的单词都会考虑到过去所有时刻的单词，词之间的依赖通过隐藏层状态来获取，这刚好解决了语言模型需要动态依赖的问题。与前馈神经网络语言模型类似，循环神经网络语言模型由三部分组成：输入层、隐藏层和输出层，其结构如图6.2所示。

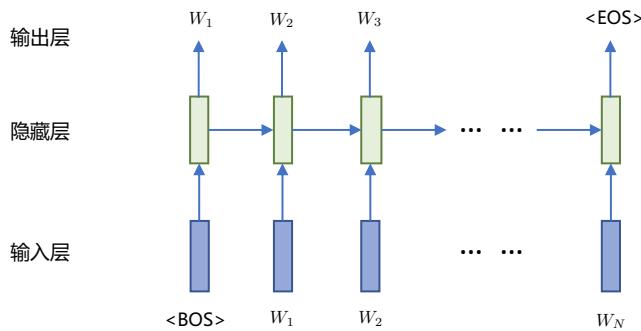


图 6.2 循环神经网络的结构

循环神经网络语言模型不限制历史词长度，而是使用整个历史序列。给定历史序列  $w_1, \dots, w_{i-1}$ ，第  $i$  时刻语言模型的目标是预测第  $i$  个词  $w_i$ ，此时循环神经网络语言模型的输入由两部分组成，前一个词  $w_{i-1}$  的词向量以及包含所有历史词信息的  $i-1$  时刻隐藏层输出  $\mathbf{h}_{i-1}$ ：

$$\mathbf{x}_i = [\mathbf{v}_{i-1}; \mathbf{h}_{i-1}] \quad (6.27)$$

其中  $\mathbf{x}_i \in \mathbb{R}^{d+m}$ ,  $\mathbf{v}_{i-1} \in \mathbb{R}^d$  代表词  $w_{i-1}$  所对应的词向量,  $d$  为词向量维度  $\mathbf{h}_{i-1} \in \mathbb{R}^m$  代表前一时刻模型的隐藏层输出,  $m$  为隐藏层维度。特别的，对于第 1 个词  $w_1$ ，由于其没有历史时刻信息，通常使用一个随机初始化向量或  $\mathbf{0}$  向量  $\mathbf{h}_0$  作为初始隐藏层向量。

循环神经网络语言模型隐藏层的目标是进行线性变化与非线性激活，隐藏层的计算可以用如下公式表示：

$$\mathbf{h}_i = f(\mathbf{W}^{hid} \mathbf{x}_i + \mathbf{b}^{hid}) \quad (6.28)$$

其中， $\mathbf{W}^{hid} \in \mathbb{R}^{m \times (d+m)}$ ,  $\mathbf{b}^{hid} \in \mathbb{R}^m$ 。其中，因为  $\mathbf{x}_i$  可以被分解为两部分， $\mathbf{W}^{hid}$  也可以分解为  $\mathbf{W}^{hid} = [\mathbf{U}; \mathbf{V}]$ ,  $\mathbf{U} \in \mathbb{R}^{m \times d}$  是词向量  $\mathbf{v}_{i-1}$  的权重,  $\mathbf{V} \in \mathbb{R}^{m \times m}$  是隐藏层输出  $\mathbf{h}_{i-1}$  的权重。将

其拆分开更能体现循环神经网络递归的特点：

$$\mathbf{h}_i = f(\mathbf{U}\mathbf{v}_{i-1} + \mathbf{V}\mathbf{h}_{i-1} + \mathbf{b}^{hid}) \quad (6.29)$$

循环神经网络语言模型输出层的目标是基于隐藏层状态  $\mathbf{h}_i$  预测词表  $\mathbb{V}$  上的概率分布，输出层的计算可以用如下公式表示：

$$\mathbf{y}_i = \text{Softmax}(\mathbf{W}^{out}\mathbf{h}_i + \mathbf{b}^{out}) \quad (6.30)$$

其中， $\mathbf{W}^{out} \in \mathbb{R}^{|\mathbb{V}| \times m}$ 。

本节只介绍了最基本的循环神经网络，隐藏层的结构较为简单。在处理长序列时，训练这样的循环神经网络可能会遇到梯度消失或梯度爆炸问题，导致无法进行有效的训练。一种解决方案是在反向传播的过程中按长度对梯度进行截断，但这一做法会损害模型建模长距离依赖的能力。另一种做法是使用如 LSTM<sup>[23]</sup> 等具备门控机制的循环神经网络，这类循环神经网络语言模型往往能实现更稳定的训练和更好的性能。

## 6.4 预训练语言模型

受到计算机视觉领域采用 ImageNet<sup>[30]</sup> 对模型进行一次预选训练，使得模型可以通过海量图像充分学习如何提取特征，然后再根据任务目标进行模型精调的范式影响，自然语言处理领域基于预训练语言模型的方法也逐渐成为主流。以 ELMo<sup>[28]</sup> 为代表的动态词向量模型开启了语言模型预训练的大门，此后以 GPT<sup>[30]</sup> 和 BERT<sup>[29]</sup> 为代表的基于 Transformer 的大规模预训练语言模型的出现，使得自然语言处理全面进入了预训练微调范式新时代。利用丰富的训练语料、自监督的预训练任务以及 Transformer 等深度神经网络结构，使预训练语言模型具备了通用且强大的自然语言表示能力，能够有效地学习到词汇、语法和语义信息。将预训练模型应用于下游任务时，不需要了解太多的任务细节，不需要设计特定的神经网络结构，只需要“微调”预训练模型，即使用具体任务的标注数据在预训练语言模型上进行监督训练，就可以取得显著的性能提升。

本节中，我们将首先介绍以 ELMo 为代表的动态词向量方法，在此基础上介绍基于 Transformer 结构的 BERT 预训练语言模型和以 GPT 和 BART 为代表的生成式预训练模型。

### 6.4.1 动态词向量算法 ELMo

如第 4 章所介绍的单词分布式表示所述，词向量主要利用语料库中词之间的共现信息，学习词语的向量表示。因此，根据给定的语料库所学习的到的词向量是恒定不变的，可以认为是“静态”的，不跟随上下文发生变化。然而，自然语言中词语往往具有多种语义，在不同的上下文或语境下会具有不同的语义。针对该问题，研究人员们提出了动态词向量（Dynamic Word Embedding），也称为上下文相关的词向量（Contextualized Word Embedding）方法，一个词语的向量通过其所在的上下

文计算获得，跟随上下文动态变化。

文献 [28] 提出了深度上下文相关词向量并介绍了双向预训练语言模型 ELMo (Embeddings from Language Models)。双向语言模型是从两个方向进行语言模型建模：从左到右前向建模和从右到左后向建模。双向建模带来了更好的上下文表示，文本中的每个词能同时利用其左右两侧文本的信息。ELMo 的神经网络结构如图6.3所示，主要包含输入层，编码层和输出层三个部分。

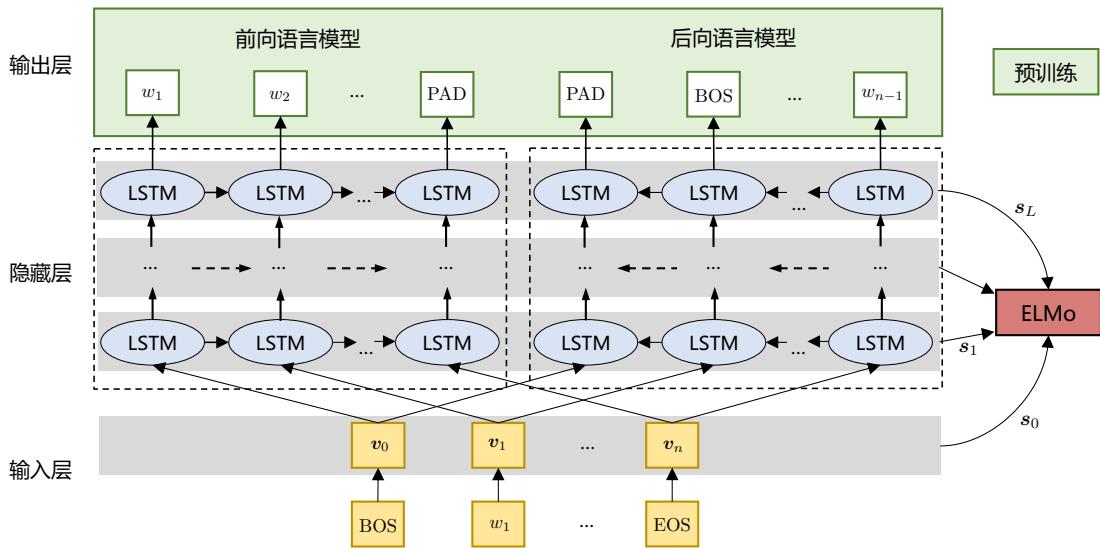


图 6.3 双向预训练语言模型 ELMo 神经网络结构<sup>[28]</sup>

输入层为了减少词不在词表中 (Out-of-Vocabulary) 的情况，对输入文本进行字符级别的编码。具体来说，输入文本中的每个词  $w_i$  视作由字符序列  $c_{i_1}c_{i_2}\dots c_{i_m}$  组成，每个字符  $c_{i_j}$  通过字符嵌入层转化为向量  $v_{c_{i_j}}$ ：

$$v_{c_{i_j}} = \mathbf{W}^{\text{char}} e_{c_{i_j}} \quad (6.31)$$

其中， $\mathbf{W}^{\text{char}} \in \mathbb{R}^{d^{\text{char}} \times |\mathbb{V}^{\text{char}}|}$  为字符嵌入矩阵、 $\mathbb{V}^{\text{char}}$  表示字符库、 $d^{\text{char}}$  表示字符向量维度、 $e_{c_{i_j}}$  表示字符  $c_{i_j}$  的独热向量。得到词  $w_i$  的字符向量表示  $v_{c_{i_1}}, v_{c_{i_2}}, \dots, v_{c_{i_m}}$  后，ELMo 模型使用卷积神经网络对字符级的表示进行语义组合，通过调整卷积神经网络的卷积核与通道数，可以得到不同粒度的字符级上下文信息。随后，在每个位置的卷积输出上使用池化层，得到词  $w_i$  的词级别表示  $\hat{v}_i$ 。在得到卷积神经网络的输出  $\hat{v}_i$  后，为了避免梯度消失或爆炸，模型使用 Highway 网络对  $\hat{v}_i$  进一步转换：

$$v_i = g \cdot \hat{v}_i + (1 - g) \cdot \text{ReLU}(\mathbf{W} \hat{v}_i + b) \quad (6.32)$$

其中， $\mathbf{g}$  为门控向量，以卷积神经网络输出  $\hat{\mathbf{v}}_i$  为输入：

$$\mathbf{g} = \sigma(\mathbf{W}^g \hat{\mathbf{v}}_i + \mathbf{b}^g) \quad (6.33)$$

其中， $\mathbf{W}^g$  为线性转换矩阵、 $\mathbf{b}^g$  为偏置。得到了每个词上下文无关的词向量后，接下来 ELMo 的编码层将从两个方向对词向量进一步编码。

ELMo 使用了两个独立的编码器分别对前向和后向进行语言模型建模，在进行预训练时，分别取最高层的正向和反向 LSTM 输出  $\vec{\mathbf{h}}_{i,L}$  和  $\overleftarrow{\mathbf{h}}_{i,L}$  预测下一时刻的词。对于给定的一段文本  $w_1 w_2 \dots w_n$  而言，前向语言模型在  $t$  时刻的目标词为  $w_{t+1}$ ，而后向语言模型的目标词则为  $w_{t-1}$ 。采用前向和后向语言模型的建模过程可以表示为：

$$\begin{aligned} P_{forward}(w_1 w_2 \dots w_n) &= \prod_{i=1}^n P(\mathbf{w}_i | w_{1:i-1}; \theta_f) \\ P_{backward}(w_1 w_2 \dots w_n) &= \prod_{i=1}^n P(\mathbf{w}_i | w_{i+1:n}; \theta_b) \end{aligned} \quad (6.34)$$

其中， $\theta_f$  和  $\theta_b$  分别代表了代表前向和后向 LSTM 模型的参数。特别需要注意的是双向模型共享输出层的参数。

ELMo 算法的编码层采用了多层次双向 LSTM 结构，通常认为，模型低层能捕捉语法等基础特征，高层能捕捉语义语境等更深层次的语言特征，双向的 LSTM 能保证在编码过程中每个位置都能获得该位置过去和未来位置的词信息。对于词  $w_i$  来说，一个  $L$  层的 ELMo 模型会产生  $2L + 1$  向量表示：

$$R_i = \{\mathbf{v}_i, \vec{\mathbf{h}}_{i,j}, \overleftarrow{\mathbf{h}}_{i,j} | j = 1, \dots, L\} \quad (6.35)$$

其中， $\mathbf{v}_i$  代表输入层得到的上下文无关词向量， $\vec{\mathbf{h}}_{i,j}$  代表第  $j$  层前向 LSTM 编码得到的特征， $\overleftarrow{\mathbf{h}}_{i,j}$  代表第  $j$  层后向 LSTM 编码得到的特征。对于每层得到的两个方向的特征，ELMo 将其拼接起来得到  $\mathbf{h}_{i,j}^{LM} = [\overleftarrow{\mathbf{h}}_{i,j}; \vec{\mathbf{h}}_{i,j}]$ 。在进行下游任务时，ELMo 将  $R_i$  中的所有向量整合成一个向量，整合的方式由任务而定，最简单的情况是直接使用最后一层的表示  $\mathbf{h}_{i,j}^{LM}$ 。因为每层 LSTM 学习到的信息不相同，对于不同任务来说，每层特征的重要性也不尽相同，因此更普遍的做法是根据任务所需信息，对每层的特征进行加权得到词  $w_i$  的对应的 ELMo 向量，其计算过程可以表示为：

$$\text{ELMo}_i^{\text{task}} = \gamma^{\text{task}} \sum_{j=0}^L s_j^{\text{task}} \mathbf{h}_{i,j}^{LM} \quad (6.36)$$

其中  $\gamma^{\text{task}}$  是整体的缩放系数， $s_j^{\text{task}}$  是每层的权重系数，反映每一层向量对于目标任务的重要性。在执行下游任务时，一般将  $\mathbf{v}_i$  和  $\text{ELMo}_i^{\text{task}}$  拼接起来作为词  $w_i$  的最终表示向量进行分类，使用  $\gamma^{\text{task}}$  对 ELMo 向量进行适当缩放。 $s_j^{\text{task}}$  则通常在下游任务的训练过程中学习得到。而 ELMo 模型的中

编码器参数在下游任务训练时则被“冻结”，不参与更新。

### 6.4.2 生成式预训练语言模型 GPT

OpenAI 公司在 2018 年提出的 GPT (Generative Pre-Training) [30] 模型是典型的生成式预训练语言模型之一。GPT-2 模型结构如图 6.4 所示，由多层 Transformer 组成的单向语言模型，主要可以分为输入层，编码层和输出层三部分。本节将介绍 GPT-2 模型结构以及单向语言模型的预训练过程和判别式任务精调。

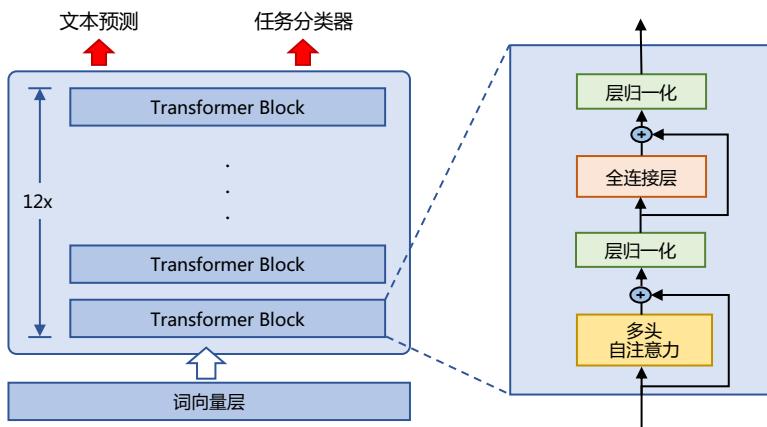


图 6.4 GPT-2 预训练语言模型结构

#### 1. 无监督预训练

GPT 采用生成式预训练方法，单向意味着模型只能从左到右或从右到左对文本序列建模，所采用的 Transformer 结构<sup>①</sup> 保证了输入文本每个位置只能依赖过去时刻的信息。

给定文本序列  $w = w_1 w_2 \dots w_n$ ，GPT-2 首先在输入层中将其映射为稠密的向量：

$$\mathbf{v}_i = \mathbf{v}_i^t + \mathbf{v}_i^p \quad (6.37)$$

其中， $\mathbf{v}_i^t$  是词  $w_i$  的词向量， $\mathbf{v}_i^p$  是词  $w_i$  的位置向量， $\mathbf{v}_i$  为第  $i$  个位置的单词经过模型输入层（第 0 层）后的输出。GPT-2 模型的输入层与前文中介绍的神经网络语言模型的不同之处在于其需要添加位置向量，这是 Transformer 结构自身无法感知位置导致的，因此需要来自输入层的额外位置信息。

经过输入层编码，模型得到表示向量序列  $\mathbf{v} = \mathbf{v}_1 \dots \mathbf{v}_n$ ，随后将  $\mathbf{v}$  送入模型编码层。编码层由

<sup>①</sup> Transformer 解码器的具体结构请参考第 8 章 8.3.3 节。

$L$  个 Transformer 模块组成，在自注意力机制的作用下，每一层的每个表示向量都会包含之前位置表示向量的信息，使每个表示向量都具备丰富的上下文信息，并且经过多层解码后，GPT-2 能得到每个单词层次化的组合式表示，其计算过程表示如下：

$$\mathbf{h}^{(L)} = \text{Transformer-Block}^{(L)}(\mathbf{h}_i^{(0)}) \quad (6.38)$$

其中  $\mathbf{h}^{(L)} \in \mathbb{R}^{d \times n}$  表示第  $L$  层的表示向量序列， $n$  为序列长度， $d$  为模型隐藏层维度， $L$  为模型总层数。

GPT-2 模型的输出层基于最后一层的表示  $\mathbf{h}^{(L)}$ ，预测每个位置上的条件概率，其计算过程可以表示为：

$$P(w_i|w_1, \dots, w_{i-1}) = \text{Softmax}(\mathbf{W}^e \mathbf{h}_i^{(L)} + \mathbf{b}^{out}) \quad (6.39)$$

其中， $\mathbf{W}^e \in \mathbb{R}^{|\mathbb{V}| \times d}$  为词向量矩阵， $|\mathbb{V}|$  为词表大小。

单向语言模型是按照阅读顺序输入文本序列  $w$ ，用常规语言模型目标优化  $w$  的最大似然估计，使之能根据输入历史序列对当前词能做出准确的预测：

$$\mathcal{L}^{\text{PT}}(w) = - \sum_{i=1}^n \log P(w_i|w_0 \dots w_{i-1}; \theta) \quad (6.40)$$

其中  $\theta$  代表模型参数。也可以基于马尔可夫假设，只使用部分过去词进行训练。预训练时通常使用随机梯度下降法进行反向传播优化该似然函数。

## 2. 有监督下游任务精调

通过无监督语言模型预训练，使得 GPT 模型具备了一定的通用语义表示能力。根据下游任务精调 (Fine-tuning) 的目的是在通用语义表示基础上，根据下游任务的特性进行适配。下游任务通常需要利用有标注数据集进行训练，数据集合使用  $\mathbb{D}$  进行表示，每个样例输入长度为  $n$  的文本序列  $x = x_1 x_2 \dots x_n$  和对应的标签  $y$  构成。

首先将文本序列  $x$  输入 GPT 模型，获得最后一层的最后一个词所对应的隐藏层输出  $\mathbf{h}_n^{(L)}$ ，在此基础上通过全连接层变换结合 Softmax 函数，得到标签预测结果。

$$P(y|x_1 \dots x_n) = \text{Softmax}(\mathbf{h}^{(L)} \mathbf{W}^y) \quad (6.41)$$

其中  $\mathbf{W}^y \in \mathbb{R}^{d \times k}$  为全连接层参数， $k$  为标签个数。通过对整个标注数据集  $\mathbb{D}$  优化如下损失函数精调下游任务：

$$\mathcal{L}^{\text{FT}}(\mathbb{D}) = \sum_{(x,y)} \log P(y|x_1 \dots x_n) \quad (6.42)$$

下游任务在精调过程中，针对任务目标进行优化，很容易使得模型对预训练阶段所学习到的通

用语义知识表示遗忘,从而损失模型的通用性和泛化能力,造成灾难性遗忘(Catastrophic Forgetting)问题。因此,通常会采用混合预训练任务损失和下游精调损失的方法来缓解上述问题。在实际应用中,通常采用如下公式进行下游任务精调:

$$\mathcal{L} = \mathcal{L}^{\text{FT}}(\mathbb{D}) + \lambda \mathcal{L}^{\text{PT}}(\mathbb{D}) \quad (6.43)$$

其中  $\lambda$  取值为 [0,1], 用于调节预训练任务损失占比。

### 6.4.3 掩码预训练语言模型 BERT

2018 年, Devlin 等人提出了掩码预训练语言模型 BERT<sup>[29]</sup> (Bidirectional Encoder Representation from Transformers)。BERT 利用掩码机制构造了基于上下文预测中间词的预训练任务,相较于传统的语言模型建模方法, BERT 能进一步挖掘上下文所带来的丰富语义。BERT 所采用的神经结构如图6.5所示,其由多层 Transformer 编码器组成,这意味着在编码过程中,每个位置都能获得所有位置的信息,而不仅仅是历史位置的信息。BERT 同样由输入层,编码层和输出层三部分组成。编码层由多层 Transformer 编码器组成。在预训练时,模型的最后有两个输出层 MLM 和 NSP, 分别对应了两个不同的预训练任务: 掩码语言模型 (Masked Language Modeling, MLM) 和下一句预测 (Next Sentence Prediction, NSP)。

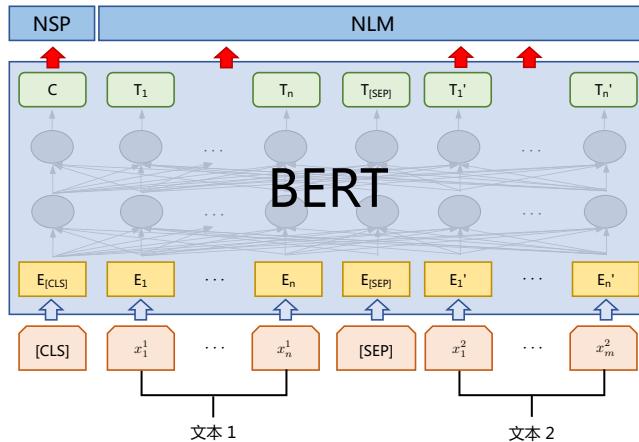


图 6.5 掩码预训练语言模型 BERT 神经网络结构<sup>[29]</sup>

需要注意的是,掩码语言模型的训练对于输入形式没有要求,可以是一句话也可以一段文本,甚至可以是整个篇章,但是下一句预测则需要输入为两个句子,因此 BERT 在预训练阶段的输入形式统一为两段文字的拼接,这与其他预训练模型相比有较大区别。

## 1. 模型结构

BERT 输入层采用了 WordPiece 分词，根据词频，决定是否将一个完整的词切分为多个子词（例如：单词 highest 可以被切分为 high 和 #est 两个子词）以缓解 OOV 问题。对输入文本进行分词后，BERT 的输入表示由三部分组成：词嵌入 (Token Embedding)、段嵌入 (Segment Embedding) 和位置嵌入 (Position Embedding)。每个词的输入表示  $\mathbf{v}$  可以表示为：

$$\mathbf{v} = \mathbf{v}^t + \mathbf{v}^s + \mathbf{v}^p$$

其中， $\mathbf{v}^t$  代表词嵌入； $\mathbf{v}^s$  代表段嵌入； $\mathbf{v}^p$  代表位置嵌入；三种嵌入维度均为  $e$ 。

词嵌入用来将词转换为实值向量表示。完成分词后，切分完的子词通过词嵌入矩阵转化为词嵌入表示，假设子词对应的独热向量表示为  $\mathbf{e}^t \in \mathbb{R}^{N \times |\mathbb{V}|}$ ，其对应的词嵌入  $\mathbf{v}_t$  为：

$$\mathbf{v}^t = \mathbf{e}^t \mathbf{W}^t \quad (6.44)$$

其中， $\mathbf{W}^t \in \mathbb{R}^{|\mathbb{V}| \times e}$  表示词嵌入矩阵； $|\mathbb{V}|$  表示词表大小； $e$  表示词嵌入维度。

段嵌入用于区分不同词所属的段落 (Segment)，同一个段落中所有词的段嵌入相同。每个段落有其特有的段编码 (Segment Encoding)，段编码从 0 开始计数。通过段嵌入矩阵  $\mathbf{W}^s$  将独热段编码  $\mathbf{e}^s$  转化为段嵌入  $\mathbf{v}^s$ ：

$$\mathbf{v}^s = \mathbf{e}^s \mathbf{W}^s \quad (6.45)$$

其中， $\mathbf{W}^s \in \mathbb{R}^{|\mathbb{S}| \times e}$  表示段嵌入矩阵； $|\mathbb{S}|$  表示段落数量； $e$  表示段嵌入维度。

位置嵌入用于表示不同词的绝对位置。将输入序列中每个词从左到右编号后，每个词都获得位置独热编码  $\mathbf{e}^p$ ，通过可训练的位置嵌入矩阵  $\mathbf{W}^p$  即可得到位置向量  $\mathbf{v}^p$ ：

$$\mathbf{v}^p = \mathbf{e}^p \mathbf{W}^p \quad (6.46)$$

其中， $\mathbf{W}^p \in \mathbb{R}^{N \times e}$  表示位置嵌入矩阵； $N$  表示位置长度上限； $e$  表示位置嵌入维度。

BERT 的编码层采用多层 Transformer 结构，使用  $L$  表示所采用的层数， $H$  表示每层的隐藏单元数， $A$  是指自注意力头数量。在文献 [29] 给出了两种不同的参数设置，BERT<sub>BASE</sub> 使用  $L = 12$ ,  $H = 768$ ,  $A = 12$ ，总参数量为 110M，BERT<sub>LARGE</sub> 使用  $L = 24$ ,  $H = 1024$ ,  $A = 16$ ，总参数量为 340M。需要注意的是，与 GPT 中 Transformer 结构所采用的约束自注意力 (Constrained Self-Attention) 仅关注当前单元左侧上下文不同，BERT 采用的 Transformer 结构使用了双向多头自注意机制，不仅关注当前单元左侧上下文情况，也会关注右侧上下文。

## 2. 预训练任务

不同于传统的自回归语言建模方法，BERT 使用去噪自编码 (Auto-Encoding) 的方法进行预训练。接下来将详细介绍 BERT 所采用的预训练任务。

**掩码语言建模：**传统的语言模型只能顺序或逆序进行建模，这意味着除了当前词本身外，每个词的表示只能利用词左侧（顺序）或右侧（逆序）的词信息。但对于大部分下游任务来说，单向的信息是不充分的，因此同时利用两个方向的信息能带来更好的词表示，双向语言模型 ELMo 使用了顺序和逆序两个语言模型来解决这一问题。为了更好的利用上下文信息，让当前时刻的词表示同时编码“过去”和“未来”的文本，BERT 采用了一种类似于完形填空的任务，即掩码语言建模。在预训练时，随机将输入文本的部分单词掩盖（Mask），让模型预测被掩盖的单词，从而让模型具备根据上下文还原被掩盖的词的能力。

在 BERT 的预训练过程中，输入文本中 15% 的子词会被掩盖。具体来说，模型将被掩盖位置的词替换为特殊字符“[MASK]”，代表模型需要还原该位置的词。但在执行下游任务时，[MASK] 字符并不会出现，这导致预训练任务和下游任务不一致。因此，在进行掩盖时，并不总是直接将词替换为 [MASK]，而是根据概率从三种操作中选择一种：(1) 80% 的概率替换为 [MASK]；(2) 10% 的概率替换为词表中任意词；(3) 10% 的概率不进行替换。

针对该掩码语言模型任务，使用  $x_1x_2\dots x_n$  表示原始文本，在经过上述掩码替换后得到输入为  $x'_1x'_2\dots x'_n$ 。对掩码替换后的输入按照 BERT 框架输入层处理后，得到 BERT 的输入表示  $v$ ：

$$X = [\text{CLS}]x'_1x'_2\dots x'_n[\text{SEP}] \quad (6.47)$$

$$v = \text{InputRepresentation}(X) \quad (6.48)$$

在编码层，对于输入表示  $v$  经过  $L$  层 Transformer，根据双向自注意力机制充分学习到文本中词语之间的联系，可以得到每个隐藏层输出以及最后的输出：

$$\mathbf{h}^{(l)} = \text{Transformer-Block}(\mathbf{h}^{(l-1)}) \quad l \in 1, 2, \dots, L \quad (6.49)$$

其中  $\mathbf{h}^{(l)} \in \mathbb{R}^{N \times d}$  表示第  $l$  层 Transformer 的隐藏层输出， $d$  表示隐藏层维度， $N$  为输入的最大序列长度， $\mathbf{h}^{(0)} = v$  表示输入。为了简化标记，可以还可以省略中间层，使用如下公式表示最终输出：

$$\mathbf{h} = \text{Transformer}(v) \quad (6.50)$$

其中  $\mathbf{h} = \mathbf{h}^{(L)}$ ，即模型最后一层的输出，得到最终上下文语义表示  $\mathbf{h} \in \mathbb{R}^{N \times d}$ 。

根据对于原始文本进行的掩码情况，得到掩盖位置的下标集合  $\mathbb{M} = \{m_1, m_2, \dots, m_k\}$ ， $k$  表示掩码数量。BERT 模型输出层，首先根据集合  $\mathbb{M}$  中元素下标，从隐藏层得到的上下文语义表示  $\mathbf{h}$  中抽取对应的表示  $\mathbf{h}_{m_i}$ 。在此基础上，利用公式 6.44 中所给出的词向量矩阵  $\mathbf{W}^t \in \mathbb{R}^{V \times e}$  将其映射到词空间表示，并通过如下公式计算对应词表上的概率分布  $P_i$ ：

$$P_i = \text{Softmax}(\mathbf{h}_{m_i} \mathbf{W}^{t \top} + \mathbf{b}^0) \quad (6.51)$$

其中  $\mathbf{b}^0 \in \mathbb{R}^V$  表示全连接层偏置。最后利用  $P_i$  与原始单词独热向量表示之间的交叉熵损失学习模型参数。

**下一句预测：**通过掩码语言建模，BERT 能够根据上下文还原掩码单词，从而具备构建对文本的语义表示能力。然而，对于阅读理解、语言推断等需要输入两段文本的任务来说，模型尚不具备判断两段文本关系的能力。因此，为了学习到两段文本间的关联，BERT 引入了第二个预训练任务：下一句预测（NSP）。

故名思义，下一句预测的任务目标是预测两段文本是否构成上下句的关系。具体来说，对于句子 A 和句子 B，若语料中这两个句子相邻，则构成正样本，若不相邻，则构成负样本。在预训练时，一个给定的句子对，有 50% 的概率将其中一句替换成来自其他段落的句子。这样可以将训练样本的正负例比例控制在 1:1。

该预训练任务与掩码语言模型任务非常类似，主要区别在于输出层。在输入层，对于给定的经过掩码处理的句子对  $x^{(1)} = x_1^{(1)}x_2^{(1)}\dots x_n^{(1)}$  和  $x^{(2)} = x_1^{(2)}x_2^{(2)}\dots x_m^{(2)}$ ，经过如下处理得到 BERT 的输入表示  $v$ ：

$$X = [\text{CLS}]x_1^{(1)}x_2^{(1)}\dots x_n^{(1)}[\text{SEP}]x_1^{(2)}x_2^{(2)}\dots x_m^{(2)}[\text{SEP}] \quad (6.52)$$

$$\mathbf{v} = \text{InputRepresentation}(X) \quad (6.53)$$

在 BERT 编码层，与掩码语言模型一样，通过  $L$  层 Transformer 编码，可以充分学习文本每个单词之间的关联，并最终得到文本语义表示：

$$\mathbf{h} = \text{Transformer}(\mathbf{v}) \quad (6.54)$$

下一句预测任务的输出层目标是判断输入文本  $x^{(2)}$  是否是  $x^{(1)}$  的下一个句子，可以转化为二分类问题。在该任务中，BERT 使用输入文本的开头添加 [CLS] 所对应的表示  $\mathbf{h}_{[\text{CLS}]}$  进行分类预测。使用全连接层预测输入文本的分类概率  $P \in \mathbb{R}^2$ ：

$$P = \text{Softmax}(\mathbf{h}_{[\text{CLS}]} \mathbf{W}^p + \mathbf{b}^o) \quad (6.55)$$

其中， $\mathbf{W}^p \in \mathbb{R}^{d \times 2}$  为全连接层权重； $\mathbf{b}^o$  表示全连接层偏置。根据分类概率  $P$  与真实分类标签之间的交叉熵损失，学习模型参数。

#### 6.4.4 序列到序列预训练语言模型 BART

在之前的章节中，我们介绍了适合自然语言生成的自回归式单向预训练语言模型 GPT 和适合自然语言理解任务的掩码预训练语言模型 BERT，自回归的 GPT 缺乏了上下文语境信息，BERT 虽然能利用上下文信息，但其预训练任务使其在自然语言生成任务上表现不佳。本节中，我们介绍一种符合自然语言生成任务需求的预训练模型 BART (Bidirectional and Auto-Regressive Transformers)

[302]。BART 兼具上下文语境信息的编码器和自回归特性的解码器，配合上针对自然语言生成制定的预训练任务，使其格外契合生成任务的场景。

BART 模型也是使用基于 Transformer 的序列到序列结构，相较于标准的 Transformer，BART 选择了 GeLU 而不是 ReLU 作为激活函数，并且使用了正态分布  $N(0, 0.02)$  进行初始化。Transformer 编码器具备双向编码上下文信息的能力，单向的 Transformer 解码器又满足生成任务的需求。BART 模型的基本结构如图6.6所示，结合了双向 Transformer 编码器以及单向的自回归解码器。

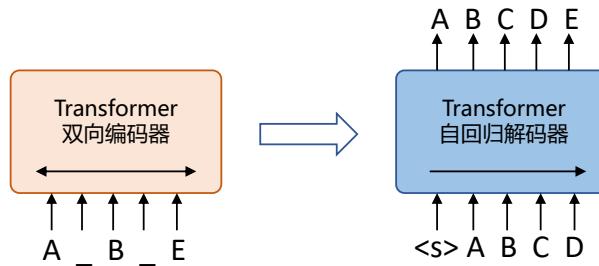


图 6.6 BART 的神经网络结构<sup>[302]</sup>

### 1. 预训练任务

BART 的预训练过程采用的是对含有噪声的输入文本进行去噪重构方法，属于去噪自编码器 (Denoising Autoencoder)。BART 使用双向编码对引入噪声的文本进行编码。然后，单向的自回归解码器通过自回归方式顺序重构原始文本。编码器最后一层隐藏层表示参与解码器每一层的计算。BART 的预测过程与 BERT 独立预测掩码位置的词有很大不同。因此，BART 的预训练任务主要关注如何引入噪声。BART 模型使用了五种方式在输入文本上引入噪音：

- **单词掩码 (Token Masking)**：随机从输入文本中选择一些单词，将其替换为掩码 ([MASK]) 标记，类似于 BERT。该噪音需要模型具备预测单个单词的能力。
- **单词删除 (Token Deletion)**：随机从输入文本中删除一部分单词。该噪音除了需要模型预测单个单词的能力，还需要模型能定位缺失单词的位置。
- **文本填充 (Text Infilling)**：随机将输入文本中多处连续的单词（称作文本片段）替换为一个掩码标记。文本片段的长度服从  $\lambda = 3$  的泊松分布。当文本片段长度为 0 时，相当于插入一个掩码标记。该噪音需要模型能识别一个文本片段有多长，并具备预测缺失片段的能力。
- **句子排列变换 (Sentence Permutation)**：对于一个完整的句子，根据句号将其分割为多个子句，随机打乱子句的顺序。该噪音需要模型能一定程度上理解输入文本的语义，具备推理前后句关系的能力。
- **文档旋转变换 (Document Rotation)**：随机选择输入文本中的一个单词，以该单词作为文档

的开头，并旋转文档。该噪音需要模型具备找到原始文本开头的能力。

图6.7给出了各种加噪方案的示例，输入加噪过程可以对这些方式进行组合使用。

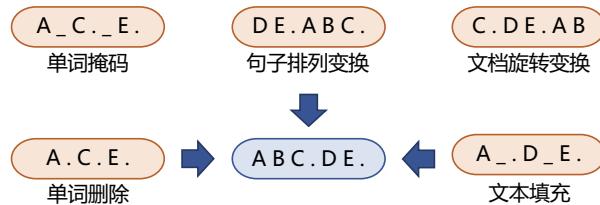


图 6.7 BART 各类型加噪方式示例

可以看到，BART 的预训练时包含单词、句子和文档多种级别的任务，除了上述噪音之外，其他任意形式的文本噪声也是适用的。实验表明，使用文本填充任务能在下游任务上普遍取得性能提升，在文本填充噪音的基础上添加句子级别的去噪任务还能带来小幅提升。另外，尽管 BART 的预训练任务主要是为自然语言生成任务设计，但是它在一些自然语言理解任务上也展现出了不错的性能。

## 2. 模型精调

BART 预训练模型具备文本表示和生成能力，因此不仅适用于文本理解任务，也适用于文本生成任务，但是用于不同类型任务时，其精调方式有所不同。

对于序列分类任务，BART 模型的编码器和解码器的输入相同，但是将解码器最终时刻的隐藏层状态作为输入文本的语义向量表示，并利用线性分类器进行标签预测。利用标注数据和模型输出结果对模型参数进行调整。整个过程与 BERT 模型类似，在句子末尾添加特殊标记，利用该位置所对应的隐藏层状态表示文本。

对于生成式的任务，比如生成式文本摘要 (Abstractive Summarization)、生成式问答 (Abstractive Question Answering) 等任务，精调时模型输入为任务所给定的输入文本，解码器所产生的文本与任务的目标文本构成学习目标。

对于机器翻译任务，由于其输入文本和输出文本是两种不同的语言，使用的不同词汇集合，因此不能采用与生成式任务相同的方法。为了解决上述问题，研究人员们提出了将 BART 模型的输入层前增加小型 Transformer 编码器，将源语言文本映射到目标语言的输入表示空间。同时，为了解决两段模型训练过程不匹配的问题，采取分阶段的训练方法。详细过程可以参见文献 [302]。

### 6.4.5 预训练语言模型的应用

在预训练阶段，大规模的数据使预训练语言模型有效地学习到了语言的通用语义表示，微调 (Fine-tuning) 则是利用预训练语言模型的主要范式，其目的是基于学习到的通用表示，针对目标领域的特性对模型进行调整，使其更适合下游任务。相较于深入了解下游任务的特有知识，为其

精心设计特别的模型，预训练模型只用转换下游任务的输入输出形式后进行微调，即可获得相当有竞争力的性能。本节将以 BERT 为例，针对三种经典的自然语言处理下游任务介绍如何微调预训练模型。

### 1. 单句文本分类

单句文本分类是自然语言处理中最为常见任务之一，其目的是判断一段文本所属的类别。例如，判断一段电影评价的情感倾向是正面还是负面，判断一篇新闻所属的类别。使用 BERT 进行单句文本分类任务时，对于将要进行单句文本分类的句子，BERT 首先使用 WordPiece 进行分词，得到分词后的句子  $w_1 w_2 \dots w_n$ ，分别添加特殊字符 [CLS] 和 [SEP] 到句首和句尾，再经过输入层将其转换为 BERT 编码层所需的输入表示，其过程可以表述如下：

$$w = [\text{CLS}] w_1 w_2 \dots w_n, [\text{SEP}] \quad (6.56)$$

$$\mathbf{v} = \text{InputLayer}(w) \quad (6.57)$$

随后，输入表示进入编码层，经过多层 Transformer 编码，每个位置的表示都通过自注意力机制进行充分交互，在最后一层得到具备丰富上下文信息的表示  $\mathbf{h}$ 。和预训练阶段时使用 [CLS] 进行 NSP 任务类似，在进行文本分类时，模型使用 [CLS] 位置的隐藏层表示  $\mathbf{h}_{[\text{CLS}]}$  进行预测。在编码层之后，模型通过一个全连接层预测输入文本对应的类别。其过程可以表述如下：

$$\mathbf{H} = \text{BERT}(\mathbf{V}) \quad (6.58)$$

$$P = \text{Softmax}(\mathbf{h}_{[\text{CLS}]} \mathbf{W} + \mathbf{b}) \quad (6.59)$$

其中， $\mathbf{W} \in \mathbb{R}^{d \times K}$  和  $\mathbf{b} \in \mathbb{R}^K$  分别为全连接层的权重和偏置， $K$  为类别总数。

在得到概率  $P$  后，若为训练阶段，则可以计算  $P$  与真实标签间的交叉熵对模型参数进行训练。若为预测阶段，则可以取概率最高的一项作为输入文本的类别。

### 2. 句子对分类

句子对分类需要预测一对有关联的句子的类别，例如判断一个句子的意思是否蕴含在另一个句子之中。句子对分类与单句分类的区别在于处理的输入不同，BERT 处理这两个任务时，也主要在输入上有所区别。对于分词后的两个句子  $w_1^{(1)}, w_2^{(1)}, \dots, w_n^{(1)}$  和  $w_1^{(2)}, w_2^{(2)}, \dots, w_m^{(2)}$ ，BERT 使用 [SEP] 作为分隔符，将两个句子拼接到一起，在输入层转换为输入表示。

$$w = [\text{CLS}] w_1^{(1)} w_2^{(1)} \dots w_n^{(1)} [\text{SEP}] w_1^{(2)} w_2^{(2)} \dots w_m^{(2)} [\text{SEP}] \quad (6.60)$$

$$\mathbf{v} = \text{InputLayer}(w) \quad (6.61)$$

其中， $\mathbf{v} \in \mathbb{R}^{(n+m+3) \times d}$ ， $d$  为模型隐藏层维度， $n$  和  $m$  分别代表第一个句子和第二个句子的长度。得到输入表示后，剩下的流程与单句文本分类一致，此处不再赘述。

### 3. 序列标注

序列标注任务需要解决的是字符级别的分类问题，其应用范围非常广泛，可用于分词，词性标注和命名实体识别等自然语言处理基础任务。以命名实体识别为例，在用序列标注的形式完成该任务时，需要对输入文本中的每一个词预测一个相应的标签，再根据整个序列的标签抽取出句子中的实体词。

传统的序列标注方法通常以词为输入的最小粒度，而在使用 BERT 等预训练模型时，通常会使用分词器将词分割为更小粒度的子词，这会破坏序列标注中词和标签一对一的关系。为了处理这种情况，可以让一个词的所有子词都保持原标签，或者只让第一个子词参与训练，预测时也只考虑第一个子词的预测结果。在完成分词后，将输入序列送入输入层转化为词向量，再将词向量送入预训练模型得到最终的隐藏层表示，其过程可以表示如下：

$$\mathbf{w} = [\text{CLS}] w_1^{(1)} w_2^{(1)} \dots w_n^{(1)} [\text{SEP}] \quad (6.62)$$

$$\mathbf{v} = \text{InputLayer}(\mathbf{w}) \quad (6.63)$$

$$\mathbf{h} = \text{BERT}(\mathbf{v}) \quad (6.64)$$

其中， $\mathbf{v} \in \mathbb{R}^{(n+2) \times d}$  为模型输入层的输出， $\mathbf{h} \in \mathbb{R}^{(n+2) \times d}$  为预训练模型最后一层的隐藏层表示， $n$  为分词后的序列长度， $d$  为模型隐藏层维度。

在得到了隐藏层表示后，需要使用一个分类器对预测每个词在标签集上的概率分布：

$$\mathbf{P} = \text{Softmax}(\mathbf{h}_i \mathbf{W} + \mathbf{b}) \quad (6.65)$$

其中  $\mathbf{h}_i$  是隐藏层表示  $\mathbf{h}$  在第  $i$  时刻的分量， $i \in \{1, \dots, n\}$ 。得到概率分布后，可以使用交叉熵损失学习模型参数。除此以外，还可以使用条件随机场等方法进一步提升序列标注性能，感兴趣的读者可以参考第 7 章信息抽取对序列标注任务作进一步了解。

## 6.5 大规模语言模型

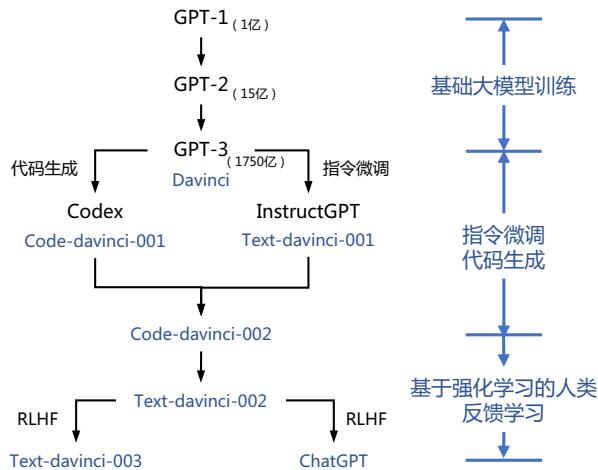
自 2020 年 Open AI 发布了包含 1750 亿参数的生成式大规模预训练语言模型 GPT 3 (Generative Pre-trained Transformer 3)<sup>[40]</sup> 以来，包含 Google、Meta、百度、智源等公司和研究机构都纷纷发布了包括 PaLM<sup>[41]</sup>、LaMDA<sup>[303]</sup>、T0<sup>[304]</sup> 等为代表的不同大规模语言模型 (Large Language Model, LLM)，也称大模型。大模型在文本生成、少样本学习、零样本学习、推理任务等方面取得了非常大的进展。表 6.2 给出了截止 2023 年 1 月典型大规模语言模型的基本情况。我们可以看到从 2022 年开始大模型呈现爆发式的增长，各大公司和研究机构都在发布各种不同类型的大模型。

表 6.2 典型大规模语言模型汇总

模型名称	参数量	训练单词数	研发机构	发布时间
ChatGPT	1750 亿	3000 亿	OpenAI	2022 年 11 月
Galactica	1200 亿	4500 亿	Meta AI	2022 年 11 月
BLOOMZ	1760 亿	3660 亿	BigScience	2022 年 11 月
U-PaLM	5400 亿	7800 亿	Google Research	2022 年 10 月
CodeGeeX	130 亿	8500 亿	清华大学	2022 年 9 月
PaLM	5400 亿	7800 亿	Google Research	2022 年 4 月
ERNIE 3.0 Titan	2600 亿	—	Baidu	2021 年 12 月
FLAN	1370 亿	—	Google	2021 年 9 月
GPT-3	1750 亿	3000 亿	OpenAI	2020 年 5 月
T5	110 亿	340 亿	Google	2019 年 10 月
RoBERTa	3.55 亿	22000 亿	Meta AI	2019 年 7 月
GPT-2	15 亿	100 亿	OpenAI	2019 年 2 月
BERT	3 亿	1370 亿	Google	2018 年 10 月
GPT-1	1 亿	—	OpenAI	2018 年 6 月

2022 年 11 月 ChatGPT (Chat Generative Pre-trained Transformer) 自发布起就引起了极大的关注，5 天内注册用户超 100 万，在系统推出仅两个月后，月活跃用户估计已达 1 亿，并与 Bing 深度搜索结合构造了对话式搜索新范式。ChatGPT 允许用户使用自然语言与系统交互，便可实现包括问答、分类、摘要、翻译、聊天等从理解到生成的各种任务。在很多自然语言理解的开放领域识别结果上都达到了非常好的效果，甚至在一些任务上超过了针对特定任务设计并且使用有监督数据进行训练的模型。ChatGPT 的生成能力也非常优秀，针对用户提出的各种各样的问题，大多数情况下都可以生成出语言通畅、有一定逻辑并且多样化的长文本。ChatGPT 的整个发展和技术演进过程如图 6.8 所示。图中黑色字表示论文相关介绍中的名字，蓝色字表示 Open AI 的 API 中的模型名称。整个过程我们可以看到大体上可以分为三个主要阶段：第一个阶段是基础大模型训练阶段，该阶段主要完成长距离语言模型的预训练；第二阶段是指令微调 (Instruct Tuning) 和代码生成训练阶段，通过给定指令进行微调的方式使得模型具备完成各类任务的能力，通过代码预训练使得模型具备代码生成的能力；第三个阶段是加入更多人工提示词，并利用基于强化学习的方式，使得模型输出更贴合人类需求。

本章中，我们以 ChatGPT 为例，介绍大模型训练三个基本阶段：基础大模型训练、指令微调以及人类反馈。需要特别说明的是，由于在本书写作阶段，包括 ChatGPT 在内的绝大部分大模型的技术细节还没有完全公开，一些已经公开的研究内容和方法也仍然需要更多时间进行验证，阅读该部分内容需要大家更多的独立思考和批判精神，并结合当前的研究进行理解。

图 6.8 ChatGPT 发展历程<sup>[305]</sup>

### 6.5.1 基础大模型训练

文献 [40] 介绍了 GPT-3 模型的训练过程，包括模型架构、训练数据组成、训练过程以及评估方法。由于 GPT-3 并没有开放源代码，根据论文直接重现整个训练过程不容易，因此文献 [306] 介绍了根据 GPT-3 的描述复现的过程，并构造开源了系统 OPT (Open Pre-trained Transformer Language Models)。

在模型架构方面不论是 GPT-3 还是 OPT 所采用的模型结构都与我们在本章第 6.4.2 所介绍的 GPT-2 模型一样，都采用由多层 Transformer 组成的单向语言模型，采用自回归方式从左到右对文本序列建模。但是针对不同的规模的参数量要求，其所使用的层数、自注意力头数、嵌入表示维度大小等具体参数各不相同。OPT 给出了 8 种模型参数的细节，如表 6.3 所示。采用 AdamW 优化器进行优化，其参数  $(\beta_1, \beta_2)$  设置为  $(0.9, 0.95)$ 。其他参数细节可以参考文献 [306]。

在预训练语料集方面，根据文献 [40] 中的报道，GPT-3 中通过主要包含经过过滤的 Common Crawl 数据集<sup>[307]</sup>、WebText2、Books1、Books2 以及英文 Wikipedia 等数据集合。其中 CommonCrawl 的原始数据有 45TB，进行过滤后仅保留了 570GB 的数据。通过子词方式对上述语料进行切分，大于一共包含 5000 亿字词。为了保证模型使用更多高质量数据进行训练，在 GPT-3 训练时，赋予不同的语料来源的不通过的权重采样。在完成 3000 亿字词训练时，英文 Wikipedia 的语料平均训练轮数为 3.4 次，而 Common Crawl 和 Books 2 仅有 0.44 次和 0.43 次。由于 Common Crawl 数据集合的过滤过程繁琐复杂，OPT 则采用了混合 RoBERTa<sup>[308]</sup>、Pile<sup>[309]</sup> 和 PushShift.io Reddit<sup>[310]</sup> 数据的方法。由于这些数据集合中包含的绝大部分都是英文数据，因此 OPT 也从 Common Crawl 数据集中抽取了部分非英文数据加入训练语料。

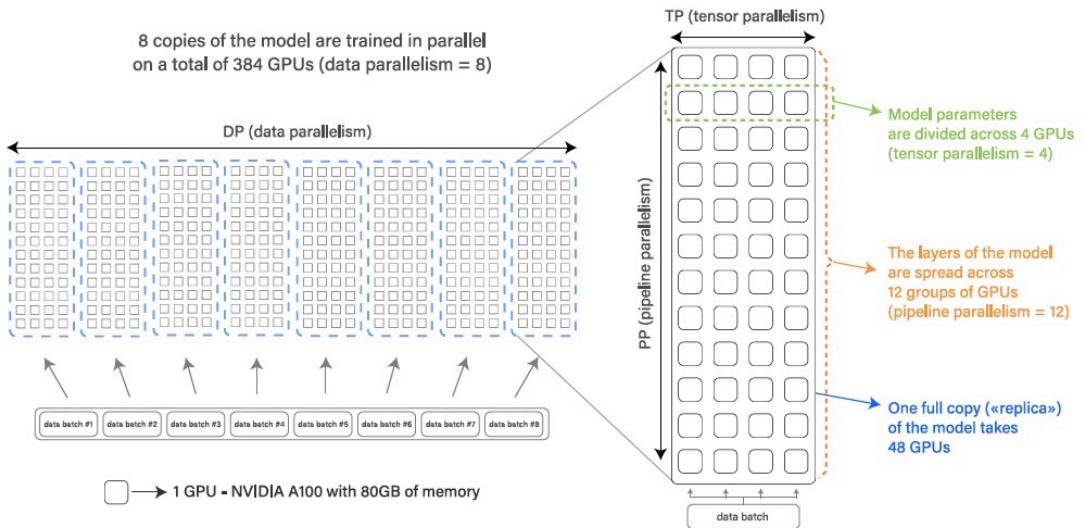
表 6.3 OPT 不同模型规模下的具体参数细节<sup>[306]</sup>

参数规模	层数	自注意力头数	嵌入向量维度	学习率	全局批次大小
125M	12	12	768	6.0e-4	50 万
350M	24	16	1024	3.0e-4	50 万
1.3B	24	32	2048	2.0e-4	100 万
2.7B	32	32	2560	1.6e-4	100 万
6.7B	32	32	4096	1.2e-4	200 万
13B	40	40	5120	1.0e-4	400 万
30B	48	56	7168	1.0e-4	400 万
66B	64	72	9216	0.8e-4	200 万
175B	96	96	12288	1.2e-4	200 万

由于模型参数量和所使用的数据量都非常巨大，普通的服务器单机无法完成训练过程，因此通常采用分布式架构完成训练。GPT-3 和 OPT 中没有对这个部分给出详细的描述。文献 [40] GPT-3 仅介绍了训练过程全部使用 NVIDIA V100 GPU，文献 [306] 介绍了 OPT 使用了 992 块 NVIDIA A100 80G GPU，采用全分片数据并行（Fully Shared Data Parallel）<sup>[311]</sup> 以及 Megatron-LM 张量并行（Tensor Parallelism）<sup>[312]</sup>，整体训练时间将近 2 个月。BLOOM<sup>[313]</sup> 则公开了更多在硬件和所采用的系统架构方面的细节。该模型的训练一共花费 3.5 个月，使用 48 个计算节点，每个节点包含 8 块 NVIDIA A100 80G GPU（总计 384GPU）。节点内容包含 4 NVLink 用于节点内部 GPU 之间通信，节点之间采用四个 Omni-Path 100 Gbps 网卡构建的增强 8 维超立方体全局拓扑网络通信。

BLOOM 使用 Megatron-DeepSpeed<sup>[314]</sup> 框架进行训练，主要包含两个部分：Megatron-LM 提供张量并行能力和数据加载原语；DeepSpeed<sup>[315]</sup> 提供 ZeRO 优化器、模型流水线以及常规的分布式训练组件。通过这种方式可以实现数据、张量和流水线三维并行，如图6.9所示。数据并行（Data Parallelism）将模型构建多个副本，每个副本放置在不同的设备上，并分别针对一部分数据并行进行训练，在每个训练步结束时同步副本间数据。张量并行（Tensor Parallelism）将模型的单个层划分到不同设备中，这样可以避免将所有激活或梯度张量都放置在一个 GPU 上，这种方法也称为水平并行或层内模型并行。流水线并行（Pipeline Parallelism）将模型不同层放置在多个 GPU 中，每个 GPU 中仅包含部分的层，这种方法也称为垂直并行。ZeRO（Zero Redundancy Optimizer）优化器<sup>[316]</sup> 允许不同的进程只保存一小部分数据（训练步骤所需的参数、梯度和优化器状态）。通过上述四个步骤可以实现数百个 GPU 的高效并行计算。

基础大模型构建了长文本的建模能力，使得模型具有语言生成能力，根据输入的提示词（Prompt），模型可以生成文本补全句子。也有部分研究人员认为，语言模型建模过程中也隐含的构建了包括事实性知识（Factual Knowledge）和常识知识（Commonsense）在内的世界知识（World Knowledge）。

图 6.9 BLOOM 并行结构<sup>[313]</sup>

### 6.5.2 指令微调

以 BERT 为代表的预训练语言模型需要根据任务数据进行微调（Fine-tuning），这种范式可以应用于参数量在几百万到几亿规模的预训练模型。但是针对数十亿甚至是数百亿规模的大模型，针对每个任务都进行微调的计算开销和时间成本几乎都是不可接受的。因此，研究人员们提出了指令微调（Instruction Finetuning）<sup>[42]</sup> 方案，将大量各类型任务，统一为生成式自然语言理解框架，并构造训练语料进行微调。

例如，可以将情感倾向分析任务，通过如下指令，将贬义和褒义的分类问题转换到生成式自然语言理解框架：

For each snippet of text, label the sentiment of the text as positive or negative.

Text: this film seems thirsty for reflection, itself taking on adolescent qualities.

Label: [positive / negative]

利用有标注数据集合，再结合上述指令模板，就可以生成大量用于微调的训练数据。利用这些训练数据，就可以在一个大模型中同时训练大量不同的任务。当前的研究工作表明，这种训练方法可以使得模型具有很好的任务泛化能力<sup>[317]</sup>，很多没有出现在指令微调训练语料中的任务也可以很好的完成，在零样本和少样本的情况下获得非常好的任务性能。

FLAN-T5<sup>[42]</sup> 中通过混合之前的 Muffin<sup>[317]</sup>、T0-SF<sup>[304]</sup>、NIV2<sup>[318]</sup> 以及思维链（Chain-of-thought, CoT）混合微调等工作，使用 473 个数据集合，针对 146 个任务类型，构造了 1836 个任务。数据

集合包括 SQuAD、MNLI、CoNLL2003 等。任务类型包括阅读理解、问题生成、常识推理、对话上下文生成、完型填空、命名实体识别、文本分类等。任务是指<数据集, 任务类型>的组合, 比如 SQuAD 既可以构造阅读理解任务, 也可以用来构造问题生成任务。Muffin、T0-SF、NIV2 是之前类似研究中所构造的任务, FLAN-T5 中直接进行了使用。CoT 混合微调模型任务, 则与前面三种不同, 通过使用之前的任务, 但是在任务的指令中增加思维链, 以及在目标结果中增加思维链, 构造新的任务, 试图用于提升模型的在未知任务上的推理能力。FLAN-T5 中使用构造 9 个数据集合用于思维链任务, 包括数学推理、多跳推理以及自然语言推断等。图6.10给出了指令和训练目标中添加思维链的样例。

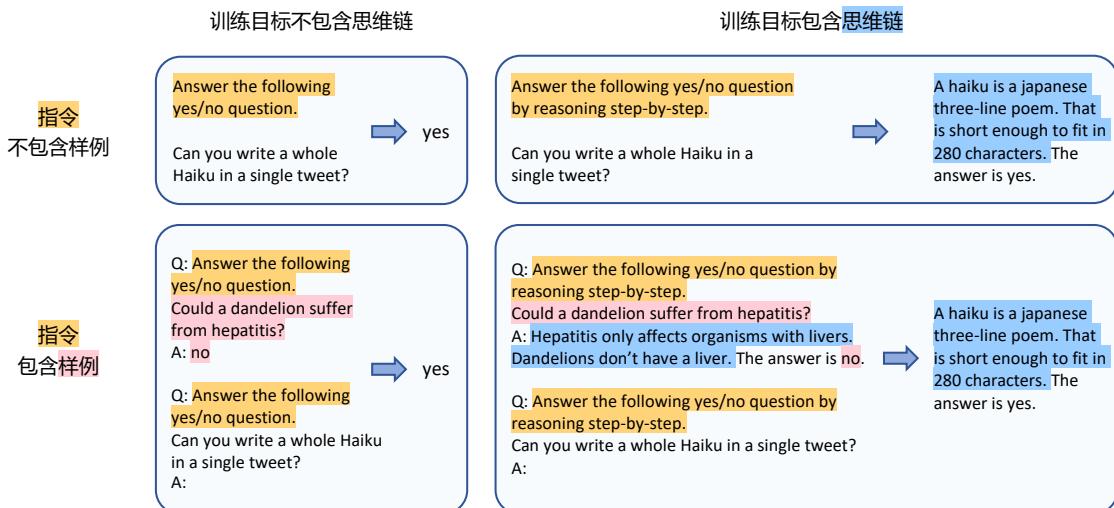


图 6.10 指令和训练目标中思维链示例<sup>[42]</sup>

通过指令微调, 大模型学习到了如何响应人类指令, 而是可以根据指令直接能够生成合理的答案。基础大模型 GPT-3 在处理任务上通常仅能生成一些句子, 需要后续模块根据句子内容再提取答案。由于指令微调阶段训练了非常多的任务, 大模型任务能力可以泛化到之前没有见过的任务上, 这使得模型初步具备了回答人们提出的任何指令的可能。这种能力对于大模型来说至关重要, 使其可以在开放领域有很好的表现。思维链以及代码生成的引入, 又在一定程度上提升了大模型的推理能力。指令微调使得大模型在处理任务的能力上有了质的飞跃。

### 6.5.3 人类反馈

经过指令微调后的模型, 虽然在开放领域任务能力表现优异, 但是模型输出的结果通常是简单答案, 这与人类的回答相差很大。因此需要进一步优化模型, 使其可以生成更加贴近人类习惯的

文本内容。但是，由于自然语言处理语料集合中所包含的答案，通常都是简短的标签，基本没有类似人类回答的有监督数据。如果全部使用人工，将自然语言处理任务数据集进行改造，构造类似人工回答的训练语料，所需要的规模过于庞大，时间成本和人工成本都过于高昂。因此，Open AI 提出了使用基于人类反馈的强化学习方法 (Reinforcement Learning from Human Feedback, RLHF)，从而大幅度降低了数据集构建成本，但是达到了非常好的效果。

在文献 [319] 中，使用的 RLHF 方法与风格续写<sup>[320]</sup> 以及文本摘要<sup>[321]</sup> 中所使用的方法非常类似。首先，仍然需要收集一定数量的用户输入以及期望的系统输出。Open AI 团队针对初始数据收集，制定了严格的标准和规范，雇佣了 40 人的团队完成该项工作。再结合通过线上 API 所收集到的高质量数据，一共构造了约 11.28 万标注集合。基于上述标注数据和初始大模型，整个 RLHF 算法训练过程如图6.11所示。

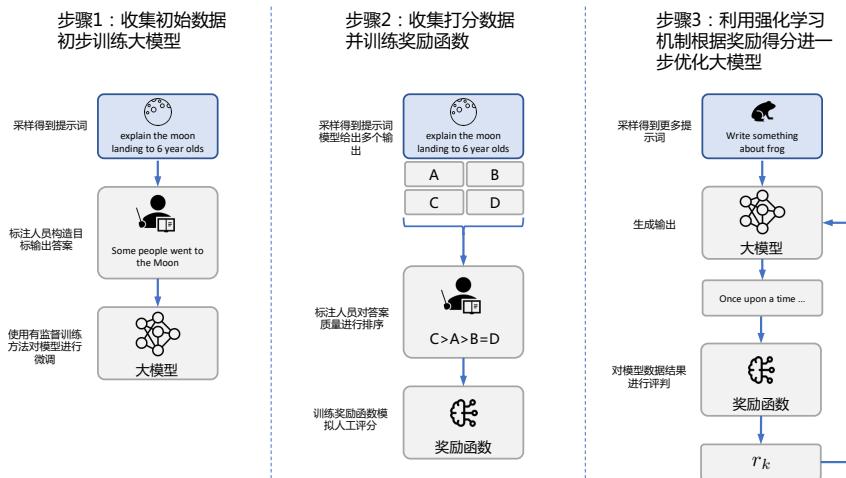


图 6.11 基于人类反馈的强化学习方法基本步骤图<sup>[319]</sup>

RLHF 算法主要分为如下三个步骤：

- (1) 收集初始数据初步训练大模型，从得的指令集合中采样部分数据，对初始的大模型进行有监督微调；
- (2) 收集打分数据并训练奖励 (Reward) 函数，使用无标注的指令数据，收集模型的多个输出结果，评价人员根据模型的输出结果进行对比评价，确定模型输出的结果排序，并利用该评分训练奖励函数，使其未来可以对模型输出的优劣进行判断；
- (3) 使用更多的指令数据，利用奖励函数输出的得分，利用强化学习机制根据奖励得分进一步优化大模型。

## 6.6 语言模型评价方法

语言模型最直接的测评方法就是使用模型计算测试集的概率，或者利用交叉熵（Cross-entropy）和困惑度（Perplexity）等派生测度。

对于一个平滑过的概率  $P(w_i|w_{i-n+1}^{i-1})$  的  $n$  元语法模型，可以用下列公式计算句子  $P(s)$  的概率：

$$P(s) = \prod_{i=1}^n P(w_i|w_{i-n+1}^{i-1}) \quad (6.66)$$

对于由句子  $(s_1, s_2, \dots, s_n)$  组成的测试集  $T$ ，可以通过计算  $T$  中所有句子概率的乘积来得到整个测试集的概率：

$$P(T) = \prod_{i=1}^n P(s_i) \quad (6.67)$$

交叉熵的测度则是利用预测和压缩的关系进行计算。对于  $n$  元模型  $P(w_i|w_{i-n+1}^{i-1})$ ，文本  $s$  的概率为  $P(s)$ ，在数据  $s$  上  $n$  元模型  $P(w_i|w_{i-n+1}^{i-1})$  的交叉熵为：

$$H_p(s) = -\frac{1}{W_s} \log_2 P(s) \quad (6.68)$$

其中， $W_s$  为文本  $s$  的长度，该公式可以解释为：利用压缩算法对  $s$  中的  $W_s$  个词进行编码，每一个编码所需要的平均比特位数。

困惑度的计算可以视为模型分配给测试集中每一个词汇的概率的几何平均值的倒数，它和交叉熵的关系为：

$$PP_s(s) = 2^{H_p(s)} \quad (6.69)$$

交叉熵和困惑度越小，语言模型性能就越好。不同的文本类型其合理的指标范围是不同的，对于英文来说， $n$  元语言模型的困惑度约在 50 到 1000 之间，相应的，交叉熵在 6 到 10 之间。

## 6.7 延伸阅读

随着深度学习的发展，预训练语言模型正逐渐成为自然语言处理的基础模型。预训练语言模型通过设计特定的自监督训练目标，有效地从大量标记和未标记数据中获取知识，并存储到巨大的参数中。通过在特定任务上进行微调，隐藏在巨大参数中的丰富知识可以使各种下游任务受益。尽管预训练-微调范式取得了巨大的成功，但在实际场景中应用该范式依旧面临许多困难与挑战，在本节中，我们探讨限制预训练模型在真实场景中应用的三点困难，并简单介绍一些前沿的解决方案。

(1) 更好的预训练语言模型迁移范式。在下游任务上微调预训练模型时，通常会在预训练模型最后添加任务特有的分类器层，模型的优化目标是基于下游任务的分类任务，而预训练阶段的

进行的是语言模型建模任务。预训练和微调阶段不一致的优化目标为预训练模型的迁移带来了隐形的阻碍，导致其需要更多的训练样本才能使预训练模型“适配”到下游任务上。受到 GPT-3 启发，一种新的预训练模型迁移范式，提示学习（Prompt Learning）<sup>[322]</sup>，逐渐走入研究者的视野。如果说微调是让预训练模型“迁就”下游任务，提示学习则可以看作是让下游任务“迁就”预训练模型。具体来说，提示学习需要将下游任务转化为预训练任务的形式，举一个在掩码预训练语言模型 BERT 上运用提示学习的例子，若对“这部电影剧情拖沓，演员演技差，我很不喜欢。”进行电影情感分类，提示学习会在待分类的句子后面添加模版“这是部 [MASK] 电影。”，让预训练语言模型在 [MASK] 位置预测标签相关词，若预测结果为坏、差、烂等负面情感的词，则这句话的情感为负面。通过将下游任务转化为预训练任务，提示学习减少了与训练阶段和下游任务阶段的差距，让模型能快速完成迁移，因此提示学习在少样本场景下表现出色，能高效地利用预训练模型。提示学习在文本分类<sup>[323–325]</sup>、命名实体识别<sup>[326, 327]</sup>、阅读理解<sup>[328, 329]</sup> 等任务的小规模语料学习上都取得了一定的效果。

(2) 绿色低碳预训练。预训练阶段需要在大量无监督语料上对大模型进行语言模型建模，这一阶段需要大量计算资源支撑。例如，BERT-base 需要在 64 块 TPU 上进行 4 天预训练，对于大部分科研人员和企业来说，训练一个自己的预训练模型所需的代价是非常高昂的，并且模型训练过程中会耗费大量的能源，对经济和环保带来额外的负担。因此，减少预训练阶段的成本是一项重大挑战。文献 [330] 提出了 ELECTRA 算法，基于对抗的思想，使用替換词检测作为预训练任务，只需要 1/4 的预训练计算量，就可以实现和其他预训练模型相似的性能。除此之外，还有一些模型通过数据集合选择来大幅降低预训练模型训练时间<sup>[331]</sup>，根据领域特性进行数据选择<sup>[332]</sup>，利用领域之间关联<sup>[333]</sup>，自动优化超参数选择<sup>[334]</sup> 等方法降低预训练模型的计算消耗。

(3) 高效微调方法。现有的微调范式需要更新预训练模型的所有参数，这意味着对于每个下游任务来说，每次微调都会得到一个不同的模型。而预训练模型的参数量越来越大，微调所有参数需要耗费大量计算资源，存储微调后的模型需要占用大量存储资源，在工业界中，模型是需要部署在服务器上供用户调用的，为每个任务都部署微调后的模型需要占用大量显存资源。为了解决传统微调方法耗费资源过多，使用成本过高的问题，大量研究者开始探究如何进行高效微调<sup>[335–337]</sup>，即冻结模型大部分参数，只更新部分参数来完成下游任务。最近的研究表明，只需要更新预训练模型的 1% 的参数，就可以实现和微调所有参数相似的性能，大大降低了预训练模型的使用成本。

## 6.8 习题

- (1) 试比较不同平滑方法的优缺点。
- (2) 预训练语言模型中常用的子词（Subword）是为了解决什么问题？
- (3) 常见的预训练任务有哪些？这些预训练任务的目的是什么？
- (4) 预训练-微调范式可能存在哪些问题？

## 7. 信息抽取

随着互联网的迅猛发展，大量的信息以电子文档的形式出现，人们面临的不再是信息匮乏，而是严重的信息过载。为了应对信息爆炸带来的挑战，迫切需要一些自动化的工具对大量无结构的文本内容及时准确地进行抽取、过滤、归类和组织，帮助人们在海量内容中迅速找到真正需要的信息。信息抽取就是在这样的需求下应用而生。信息抽取（Information Extraction, IE）的目标就是从非结构化的文本内容中提取特定的信息。信息抽取并不试图对全文进行理解，仅针对任务需求从篇章中抽取特定信息。信息抽取的应用广泛，在阅读理解、机器翻译、知识图谱等任务中都发挥着非常基础和重要的作用。

本章首先介绍信息抽取的基本概念，在此基础上，详细介绍信息抽取的三个主要任务：命名实体识别、关系抽取和事件抽取，并介绍不同场景下信息抽取的难点及主要算法。

### 7.1 信息抽取概述

海量的文本内容提供了人们丰富的信息获取的可能，但是面对如此众多的内容，人们也难以快速从这些文本中快速发现所需的信息。迫切需要自然语言处理算法能够自动化的从这些无结构的文本中发现特定信息。但是通过第4章语义分析的介绍，我们可以知道，通用的句子和篇章的语义表示和理解，目前还远达不到实用的阶段。信息抽取目标不是构建通用的句子或者篇章理解方法，而是针对特定的需求，从自然语言构成的非结构化文本中抽取指定类型的实体、关系、事件等信息，进而形成结构化数据。

如图7.1所示，利用信息抽取算法可以从一段非结构化的新闻文本抽取出公司：“苹果公司”、时间：“北京时间9月13日”、地点：“史蒂夫·乔布斯剧院”等实体信息，以及“蒂姆·库克”与“苹果公司”是“CEO-Of”关系的信息，并且还可以获得整段文本描述的是“发布会”事件。我们还可以从社会媒体中抽取出恐怖事件的详细情况：时间、地点、作案者、受害者、袭击目标、使用的武器等；从经济新闻中抽取出新产品发布情况：公司名、产品名、发布时间、产品性能等；从医疗记录中抽取出症状、诊断记录、检验结果、处方等。使用信息抽取方法所获的信息构成结构化描述，可以直接存入数据库中，供用户查询进一步分析利用。抽取后的数据和事实可以直接向用户显示，也可作为原文检索的索引，或存储到数据库、电子表格中，以便于以后的进一步分析。

实体

实体关系

事件触发词

事件类型	发布会
时间	北京时间9月13日
公司	苹果公司
地点	史蒂夫·乔布斯剧院
人员	蒂姆·库克
产品	Iphone13、Apple TV
实体关系	苹果公司，蒂姆·库克，CEO

图 7.1 非结构化文本信息抽取样例

信息抽取技术属于知识技术中知识发现的范畴，它突破了信息检索中必须由人来阅读、理解、抽取信息的局限性，实现了信息的自动查找、理解和抽取。信息抽取模型可以极大地促进下游自然语言处理任务性能的提高。实体、关系、事件作为文本中重要的语义知识，可以为信息检索、知识图谱、问答系统等提供基础支撑。例如，实体及关系可以改善系统检索文档的相关度，并提高检索系统的召回率和准确率；实体、关系及事件等是知识图谱的基本元素；实体与关系可以支持问答系统对文本中的关键信息做出更准确的分析，给出更精确、更简洁的短语级的答案。因此信息抽取是也自然语言处理任务中重要的研究方向和底层任务。

一般来说，信息抽取系统的处理对象是自然语言文本，尤其是非结构化文本。但从广义上讲，除了电子文本以外，信息抽取系统的处理对象还可以是语音、图像、视频等其他媒体类型的数据。本章只讨论狭义上的信息抽取，即针对自然语言文本的信息抽取。

信息抽取研究开始于 20 世纪 60 年代中期，以纽约大学的 Linguistic String 和耶鲁大学的 FRUMP 这两个项目为代表。直到 20 世纪 80 年代末期，得益于 Message Understanding Conference (MUC) 的长期举办，信息抽取的研究与应用逐步进入繁荣期。从 1987 年到 1998 年，MUC 会议共举行了 7 届，MUC 为信息抽取制定了具体的任务和严密的评测体系。该会议提出了一套完整的基于模板填充机制的信息抽取方案，核心内容包括命名实体识别、共指消解、关系抽取、事件抽取等具体内容。MUC 会议吸引了世界各地的研究者参与其中，从理论和技术上促进了信息抽取的研究成果不断涌现，为信息抽取在 NLP 领域中成为一个独立分支做出了重大贡献。

命名实体 (Named Entity) 是在 1995 年的 MUC-6 信息理解会议中首次提出，其主要的研讨内容为如何从论文、新闻报纸等非结构化文本中抽取公司、国防等相关信息，这些信息包括了现在数据集中常用的人名、地名、机构名等标签。MUC-6 会议评测任务就是自动识别文本中的预定义的实体并进行分类，但当时的研究方法基本上是基于模板规则，比如词汇规则（包括词形、词性）、短语规则等。因为测试语料主题单一且数量较小（30 篇），因此大部分的识别方法都取得了较好的成绩，最高的 F1 值达到 96.42%。除了实体识别任务，MUC 会议还引入三个新的评测任务：共指

关系确定、模板元素填充等。随后的 MUC-7 会议拓展了 MUC-6 的标注规范，并且增加了训练语料的数量。MUC-6 和 MUC-7 中将有待识别的实体指称为“实体的唯一标识符（Unique identifiers of entities）”，目标实体类型分为三类：命名实体、时间、数值，其中命名实体又可细分为：人名、地名、机构名。值得注意的是，MUC-7 中开始出现了隐马尔可夫，最大熵等基于统计机器学习的方法。

继 MUC 会议之后，1999 年至 2008 年美国国家标准技术研究所 (NIST) 组织 Automatic Content Extraction (ACE) 评测会议成为另一个致力于信息抽取研究的重要国际会议。与 MUC 会议相比，ACE 评测不针对某个具体的领域或场景，它采用基于漏报（标准答案中有而系统输出中没有）和误报（标准答案中没有而系统输出中有）的一套评价体系，还对系统跨文档处理（Cross-document Processing）能力进行评测。除 MUC 和 ACE 外，包括 Multilingual Entity Task Evaluation (MET)、Document Understanding Conference (DUC) 等与信息抽取相关的国际学术会议，为信息抽取在不同领域、不同语言中的应用起到了很大的推动作用。

中文的信息抽取研究起步相对英文较晚，由于中文与西方字母型文字的巨大差异，以及中文缺少表示词语边界的分割符号等特殊性，中文信息抽取效果会受到自动分词结果的影响，使得中文信息抽取相较于英文更困难。早期工作主要集中在中文命名实体识别方面，在 MUC-7、MET 等会议的支持下，取得了一定的进步。2006 年 SIGHAN (Special Interest Group of the Association for Computational Linguistics) 将汉语命名实体识别加入 Bakeoff 评测比赛。Bakeoff-2006 遵循了 CoNLL-2002 的标签定义框架，提出并标注了四种命名实体标签，包括人名、地名、机构名和地缘政治实体，并提供了三个中文语料库：MSRA、LDC 和 CITYU。Bakeoff-2007 删除了 LDC 语料库，并将命名实体类型设置为人名、地点和机构名。在 Bakeoff-2006 和 Bakeoff-2007 中，研究者多使用了统计机器学习方法，并且当年最优秀的 NER 方法几乎都使用了条件随机场、最大熵等统计机器学习模型。当前中文信息抽取研究在命名实体识别基础上，在共指消解、关系抽取、事件抽取等众多方面都取得了很多领先的研究成果。虽然当前信息抽取通常还只是面向特定领域开展，通用信息抽取系统仍然亟待研究，但是，近年来信息抽取领域研从理论到应用都有一些新进展。

信息抽取包含命名实体识别 (Named Entity Recognition, NER)、关系抽取 (Relation Extraction, RE)、事件抽取 (Event Extraction)、时间表达式识别 (Temporal Expression)、实体归一化 (Entity Normalization)、模板填充 (Template Filling)、话题检测于跟踪 (Topic Detection and Tracking, TDT) 等任务。由于这些任务是大多数自然语言处理系统所依赖的底层工具，因此自 21 世纪以来，不论在学术界还是工业界都对信息抽取任务的研究给了越来越多的关注。本章主要对命名实体识别、关系抽取以及事件抽取任务和常见算法进行介绍。

## 7.2 命名实体识别

命名实体 (Named Entity) 是指具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。常见的命名实体和样例如表 7.1 所示。命名实体识别 (Named Entity Recognition, NER) 目标

就是从文本中抽取出这些具有特定意义的实体词。命名实体识别一般包含两个步骤，分别是实体边界判断和实体类别判断。其中实体边界判断是为了确定实体字符串在非结构化文本中的开始位置和结束位置，而实体类别的判断则是为了判断该字符串对应的实体类型。

例如：复旦大学始创于 1905 年，原名复旦公学，1917 年定名为复旦大学，位于中国上海，是中国人自主创办的第一所高等院校。

其中“复旦大学”和“复旦公学”是机构名，“1905”和“1917”是时间，“上海”是地名。

表 7.1 常见命名实体以及样例

实体名	标签	举例
人名	PER	[张钹] <sub>PER</sub> 院士：抓住机会、掌握主动发展第三代人工智能
地名	LOC	2021 世界人工智能大会将于 7 月 8 日至 10 日在 [上海] <sub>LOC</sub> 召开
机构名	ORG	[复旦大学] <sub>ORG</sub> 校名取自《尚书大传》之“日月光华，旦复旦兮”
疾病名	DIS	[高血压] <sub>DIS</sub> 已成为影响全球死亡率的第二大危险因素
药品名	DRU	[奥司他韦] <sub>DRU</sub> 是治疗流感的首选药物

命名实体识别算法的主要难度在于处理歧义和未登录词问题。歧义问题是指同一个名称可以指代不同类型的实体。

例如：我们明天在复旦大学见。

复旦大学共有邯郸、枫林、张江、江湾四个校区。

上句中“复旦大学”在不同的上下文中分别作为地名和机构名。在英文中这种现象更加常见，比如，“Harvard”既可以是人名，也可以是机构名，还可以是地名，需要根据上下文对实体的类别进行判断。未登录词问题与中文分词中定义一致，也是指在训练语料中没有出现或者词典当中没有，但是在测试数据中出现的实体。命名实体在语言中通常表现出表达随意、用法复杂、形式多变等特点，未登录词问题相较于中文分词更加严重。

命名实体从表现形式还可以进一步分为两种类型：非嵌套命名实体（Non-nested Named Entities）和嵌套命名实体（Nested Named Entities）。非嵌套命名实体就是普通的命名实体，每个单词只对应一个标签。嵌套命名实体是指实体中存在嵌套的情况，每个的单词可能对应若干个标签。例如，“复旦大学上海医学院”整体上是机构名，但是其中“上海”是地名，而“复旦大学”是机构名。

由于嵌套命名实体识别方法和非嵌套命名实体识别方法有很大不同，基于序列标注的命名实体模型是难以直接有效地处理嵌套命名实体。因此，在本节中，我们将分别对这两种实体类型的识别方法分别进行介绍。

### 7.2.1 非嵌套命名实体识别

非嵌套命名实体（Flat Named Entity）识别通常可以转换为序列标注问题。对给定序列中的每一个元素（Token）标注一个标签。一般来说，序列通常是一个句子，而元素指的是句子中的词语或

者字。标注的标签一般同时能表示实体的边界和类别信息。典型的标注格式是 *BIO* 标签体系，即每将个元素标注为“B-X”、“I-X”或者“O”。其中，“B-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的开头，“I-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的中间位置，“O”表示不属于任何类型。假设需要识别的命名实体包含人名（PER）、地名（LOC）和机构名（ORG），采用 *BIO* 标签体系，对应的标签集合为：{O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG}。图7.2给出了利用该标签体系对于句子中每个词语的分类标签。句子“复旦大学的前身是马相伯于 1902 年在上海创办的震旦学院”，包含四个命名实体：“复旦大学”和“震旦学院”都是组织名，“马相伯”是人名，“上海”是地名。

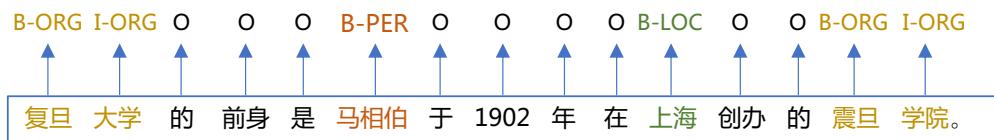


图 7.2 非嵌套命名实体识别示例

在实际应用中，也有一些系统采用更复杂的 *BIOES* 标签体系，在 *BIO* 标签的基础上增加了单字符实体 *S* 和字符实体的结束标识 *E*。同样的，在边界标签基础上需要与实体类别标签融合，对应的标签集合为 {O, B-PER, I-PER, E-PER, S-PER, B-LOC, I-LOC, E-LOC, S-LOC, B-ORG, I-ORG, E-ORG, S-ORG}。具体来说，对于一个长度为 *n* 的句子 *S* = *w<sub>1</sub>w<sub>2</sub>...w<sub>n</sub>*，其中 *w<sub>i</sub>* 表示句子中的第 *i* 个元素。序列标注即给每个元素赋予一个标签，来表示这个元素所在的实体类别和边界信息。一组相邻且类别相同的元素构成的子序列 Span(start,end,type) 就构成了抽取出来的命名实体。通过这种转换，命名实体识别问题就转换为了序列标注问题。在前述章节介绍的包括隐马尔可夫（HMM）、条件随机场（CRFs）、长短时记忆网络（LSTM）等在内的很多算法都可以应用于该任务，这里就不再赘述。

本节将介绍针对命名识别任务的特性进行改进的三种算法：基于半马尔可夫条件随机场和基于 Transformer 方法的命名实体算法，以及融合字典信息的中文命名实体识别算法。

### 1. 基于半马尔可夫条件随机场的命名实体识别

非嵌套的命名实体识别任务通过上述 *BIO* 标签，可以转换为序列标注任务。在自然语言处理算法中通常使用的线性链条件随机场，给定随机变量序列 *X* 的条件下，随机变量序列 *Y* 的条件概率分布  $P(Y|X)$  满足马尔可夫性，即输出标签序列 *y<sub>i</sub>* 仅与其周边的标签 *y<sub>i-1</sub>* 和 *y<sub>i+1</sub>* 相关。在输入序列 *X* 为句子中每个字的情况下，使用线性链条件随机场仅能表示局部依赖的特征。但是，在命名实体识别任务中往往依赖更多的非局部特征。

文献[338]提出了半马尔可夫条件随机场(Semi-Markov Conditional Random Fields,Semi-CRFs)，

从要求每个字的所对应分类标签满足马尔可夫性，放松到仅需由邻接词组成的片段（Segments）间进行满足马尔可夫性即可。将这种模型用于命名体识别任务时，可以更有效、更自由的利用各种有利于识别出命名体片段边界的特征，如实体的长度、与实体相似的已知实体名称等。

在建立一个条件随机场时，首先要定义一个特征函数集，该函数集内的每个特征函数都以文本作为输入，提取的特征作为输出，假设该函数集为：

$$\Phi(x_1 \dots x_n; y_1 \dots y_n) \quad (7.1)$$

其中  $\mathbf{x} = \{x_1 \dots x_n\}$  为  $n$  个字组成的输入文本序列， $\mathbf{y} = \{y_1 \dots y_n\}$  为对应的实体标签序列。条件随机场使用对数线性模型来计算给定观测序列下状态序列的条件概率：

$$P(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}'} \exp(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}'))} \quad (7.2)$$

其中  $\mathbf{y}'$  是所有可能的状态序列， $\mathbf{w}$  是 CRF 模型的参数，可以把它看成是每个特征函数的权重，CRF 模型的训练可以看作是对参数  $\mathbf{w}$  的估计。

当扩展至半马尔可夫条件随机场时，使用片段集合  $s = < s_1, \dots, s_p >$  来表示输入文本  $\mathbf{x}$ ，其中一个片段  $s_j = < t_j, u_j, y_j >$  由起始位置  $t_j$ 、结束位置  $u_j$  和标签  $y_j \in \mathbf{y}$  组成。这里所有片段的长度都为正数，并且完全地覆盖序列  $X$  且没有重叠，也就是说  $t_j$  和  $u_j$  总是满足以下的约束：

$$\begin{aligned} t_1 &= 1, u_p = |\mathbf{x}| \\ 1 \leq t_j &\leq u_j \leq |\mathbf{x}| \\ t_{j+1} &= u_j + 1 \end{aligned}$$

例如：对于输入文本序列“复旦大学创于 1905 年，位于中国上海”，一种可能得片段表示为  $s = < (1, 4, I), (5, 5, O), (6, 6, O), (7, 10, I), (11, 11, O), (12, 12, O), (13, 13, O), (14, 15, I), (16, 17, I) >$ ，对应标签序列  $\mathbf{y} = < I, O, O, I, O, O, O, I, I >$ 。

根据公式7.2，针对片段也可以定义特征函数集合  $\mathbf{g} = < g^1 \dots g^K >$ ，但是要求每个片段特征  $g^k(j, \mathbf{x}, \mathbf{s})$  仅与输入  $\mathbf{x}$ 、当前片段  $s_j$  以及标签  $y_{i-1}$  相关。因此可以将  $g^k(j, \mathbf{x}, \mathbf{s})$  定义为：

$$g^k(j, \mathbf{x}, \mathbf{s}) = g'^k(y_j, y_{j-1}, \mathbf{x}, t_j, u_j) \quad (7.3)$$

使用  $\mathbf{G}(\mathbf{x}, \mathbf{s})$  表示所有  $g^k(j, \mathbf{x}, \mathbf{s})$  集合，Semi-CRF 可以改写为以下的形式：

$$P(\mathbf{s}|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot \mathbf{G}(\mathbf{x}, \mathbf{s}))}{\sum_{\mathbf{s}'} \exp(\mathbf{w} \cdot \mathbf{G}(\mathbf{x}, \mathbf{s}'))} \quad (7.4)$$

Semi-CRF 的推断问题与 CRF 类似，可以定义为给定参数  $\mathbf{w}$  和输入序列  $\mathbf{x}$  的情况下，寻找最

近片段的问题，即  $\arg \max_s P(s|x; w)$ 。根据前述公式的定义，可以进一步转换为：

$$\arg \max_s P(s|x; w) = \arg \max_s w \cdot G(x, s) = \arg \max_s w \cdot \sum_j g(y_j, y_{j-1}, x, t_j, u_j) \quad (7.5)$$

假设  $L$  是一个片段最大的长度， $s_{i:y}$  表示序列从 1 到  $i$  的所有可能片段，并且最后一个片段的标签为  $y$ ，结束位置为  $i$  的集合， $V_{x,g,w}(i,y)$  表示  $w \cdot G(x, s') s' \in s_{i:y}$  的最大值。退关过程也可以使用 Viterbi 算法：

$$V(i, y) = \begin{cases} \max_{y', d=1 \dots L} V(i-d, y') + w \cdot g(y, y', x, i-d+1, i) & \text{if } i > 0 \\ 0 & \text{if } i = 0 \\ -\infty & \text{if } i < 0 \end{cases} \quad (7.6)$$

最好的分割是  $\max_y V(|x|, y)$  对应的路径。

给定训练语料  $\mathbb{D} = (\mathbf{x}_l, \mathbf{s}_l)_{l=1}^N$ ，利用最大似然估计完成参数训练，得到最优参数  $w^*$ ：

$$\mathcal{L}(w) = \sum_l \log P(\mathbf{s}_l | \mathbf{x}_l; w) = \sum_l \left( w \cdot G(\mathbf{x}_l, \mathbf{s}_l) - \log \sum_{s'} \exp(w \cdot G(\mathbf{x}, s')) \right) \quad (7.7)$$

$$w^* = \arg \max_w \mathcal{L}(w) \quad (7.8)$$

## 2. 基于 Transformer 的命名实体识别

由于 Transformer 结构可以很好的并行化，并且具有较好的建模长文本的能力，因此基于 Transformer 结构的深度神经网络被广泛地应用于很多自然语言处理任务，如机器翻译，语言建模，预训练模型等。但是，直接在命名实体识别任务上使用 Transformer 模型往往表现不佳，其主要原因包括：

(1) **位置编码无法捕捉方向信息**：原始 Transformer 结构使用正弦位置编码，这种编码可以捕捉距离信息，但是不能捕捉方向信息。例如，第  $t$  个字符的位置编码表示为：

$$PE_t = \begin{bmatrix} \sin(c_0 t) \\ \cos(c_0 t) \\ \vdots \\ \sin(c_{\frac{d}{2}-1} t) \\ \cos(c_{\frac{d}{2}-1} t) \end{bmatrix} \quad (7.9)$$

其中， $d$  表示位置编码的维度。根据余弦和差公式，对于距离第  $t$  个字符偏移量为  $k$  的字符，两个

位置编码的点积可以写作：

$$PE_t^T PE_{t+k} = \sum_{j=0}^{\frac{d}{2}-1} [\sin(c_j t) \sin(c_j(t+k)) + \cos(c_j t) \cos(c_j(t+k))] = \sum_{j=0}^{\frac{d}{2}-1} \cos(c_j k) \quad (7.10)$$

这个点积反映了两个字符之间的距离。令  $j = t - k$ , 则：

$$PE_t^T PE_{t+k} = PE_j^T PE_{j+k} = PE_{t-k}^T PE_t \quad (7.11)$$

从上面的式子可以看出，对于字符  $t$  偏移量为  $+k$  和  $-k$  的字符，结果是相同的，这意味着正弦位置编码不能捕捉到方向信息。然而，在命名实体识别任务中，词与词之间的相对位置对模型有较大影响。例如，“复旦大学（ORG）位于上海”和“光华楼位于复旦大学（LOC）”两个句子中，“复旦大学”和“位于”处在不同的相对位置可以推断“复旦大学”分别为组织实体和地点实体。

(2) 平滑的注意力分布：Transformer 模型在计算注意力时会对注意力分数进行缩放，得到一个较为平滑的注意力分布，如公式7.12所示：

$$\text{Attn}(K, Q, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (7.12)$$

其中， $d_k$  是向量的维度，缩放因子  $\frac{1}{\sqrt{d_k}}$  的目的是防止较大的维度导致点积的幅度会增大，从而将 Softmax 函数推入梯度极小的区域。但是对命名实体识别任务来说，上下文中少数的词就足够用来判断它的标签，平滑的注意力分布反而可能会引入更多的噪声。

针对上述问题，TENER<sup>[339]</sup> 提出了一种改进版的 Transformer，其模型结构如图7.3所示。

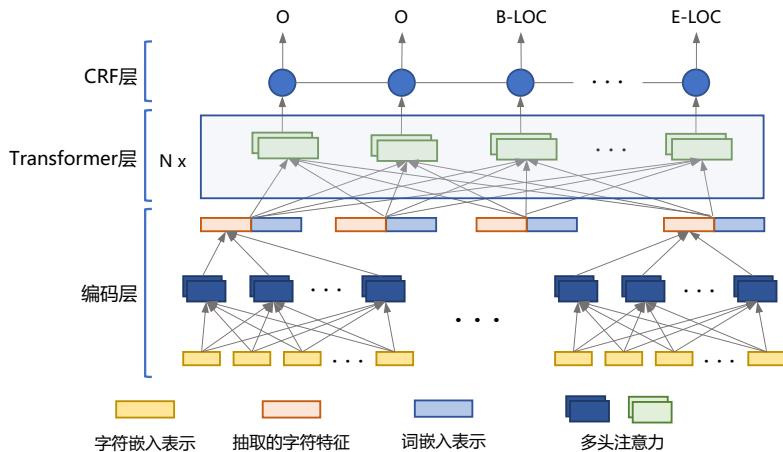


图 7.3 TENER 模型结构<sup>[339]</sup>

TENER 的主要有以下两点改进：(1) TENER 使用了相对位置编码，提升了方向的感知能力。新的相对位置编码为：

$$R_t, R_{-t} = \begin{bmatrix} \sin(c_0 t) \\ \cos(c_0 t) \\ \vdots \\ \sin(c_{\frac{d}{2}-1} t) \\ \cos(c_{\frac{d}{2}-1} t) \end{bmatrix}, \begin{bmatrix} -\sin(c_0 t) \\ \cos(c_0 t) \\ \vdots \\ -\sin(c_{\frac{d}{2}-1} t) \\ \cos(c_{\frac{d}{2}-1} t) \end{bmatrix} \quad (7.13)$$

改进后，对于相同的位移  $t$ ，前向和后向的位置编码  $R_t$  和  $R_{-t}$  不完全相同。此时，注意力机制可以同时捕捉到距离信息和方向信息。(2) TENER 取消了计算注意力分数时的缩放因子，使得产生的注意力分数的分布更加稀疏：

$$R_{t-j} = [\dots \sin(\frac{t-j}{10000^{2i/d_k}}) \cos(\frac{t-j}{10000^{2i/d_k}}) \dots]^T \quad (7.14)$$

$$A_{t,j}^{rel} = Q_t^T K_j + Q_t^T R_{t-j} + K_j^T R_{j-t} + \mathbf{u}^T K_j + \mathbf{v}^T R_{t-j} \quad (7.15)$$

$$\text{Attn}(Q, K, V) = \text{softmax}(A^{rel})V \quad (7.16)$$

通常情况下，一句话只有少数实体，且仅需知道较小部分的上下文就可以判别实体的类别，而不需要关注所有词。因此，这种稀疏的注意力分数分布更适用于命名实体识别任务。

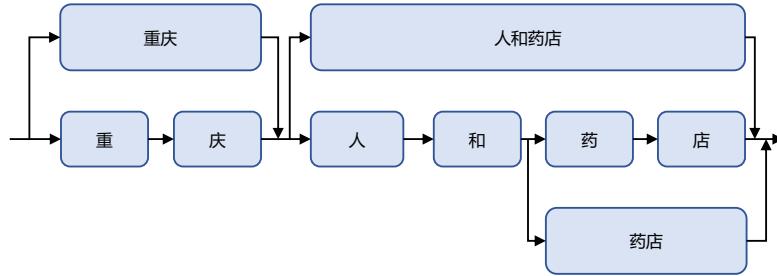
TENER 通过改进 Transformer 的位置编码及自注意力分数计算部分，使得 Transformer 更适用于命名实体识别任务。

### 3. 融合词典知识的栅格网络中文命名实体识别算法

中文命名实体识别与英文命名实体识别相比，由于中文单词边界不确定并且包含复杂的词语组合，使得中文命名实体识别更加困难。中文命名实体可以在中文分词基础上，利用序列标注技术对命名实体进行预测。也可以以中文汉字为基本单元构建序列标注任务。两者各有利弊，在中文分词结果基础上可以有效利用词汇信息，但是分词错误会累计传递。并且命名实体通常是未登录词，实体词在中文分词阶段可能被切分或者与前后词语错误组合，导致命名实体无法识别。而直接采用汉字作为基本单元，虽然可以避免中文分词的错误传递，但是无法利用词汇信息。

文献 [340] 提出了栅格结构长短时记忆网络 (Lattice LSTM) 来编码句子可以匹配所有词语信息。使用词典匹配句子中的单词，可以获得如图7.4所示的栅格 (Lattice) 的结构。栅格结构看作是一个有向无环图，词汇的开始和结束字符决定了其位置。对于输入汉字序列  $c_1, c_2, \dots, c_n$ ，使用字典  $\mathbb{D}$  匹配可以得到以  $b$  开始，以  $e$  结尾的子序列  $w_{b,e}^d$ 。上例中  $w_{1,2}^d$  表示“重庆”， $w_{3,6}^d$  表示“人和药店”。

基于汉字序列的 LSTM 模型中包含四种类型的向量：输入向量 (Input Vectors)、隐藏向量

图 7.4 融合词典知识的栅格结构<sup>[340]</sup>

(Output Hidden Vectors)、单元向量 (Cell Vectors) 和门向量 (Gate Vectors)。输入向量  $\mathbf{x}_j^c = \mathbf{e}^c(c_j)$ , 表示输入汉字  $c_j$  的向量表示。 $\mathbf{c}_j^c$  表示单元向量,  $\mathbf{h}_j^c$  表示隐藏向量。原始 LSTM 门向量包含三个: 输入门  $\mathbf{i}_j^c$ 、遗忘门  $\mathbf{f}_j^c$  和输出门  $\mathbf{o}_j^c$ 。具体的计算公式见第 2 章公式 2.5 至公式 2.10。

Lattice LSTM 在原始 LSTM 的基础上, 引入了词汇单元 (Word Cell)  $\mathbf{c}_{b,e}^w$  表示从句子开始到当前  $w_{b,e}^w$  的信息, 如图7.5所示。词汇单元融合以该字符结束的所有词汇信息, 例如图中“店”融合了“人和药店”和“药店”的信息。单元向量  $\mathbf{c}_j^c$  不仅要考虑字符信息, 还要考虑子序列  $w_{b,e}^w$  的信息。使用  $\mathbf{x}_{b,e}^w = \mathbf{e}^w(w_{b,e}^s)$  表示子序列  $w_{b,e}^w$  的向量表示。词汇单元  $\mathbf{c}_{b,e}^w$  具体计算公式如下:

$$\begin{bmatrix} \mathbf{i}_{b,e}^w \\ \mathbf{f}_{b,e}^w \\ \tilde{\mathbf{c}}_{b,e}^w \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \end{bmatrix} \left( \mathbf{W}^{w\top} \begin{bmatrix} \mathbf{x}_{b,e}^w \\ \mathbf{h}_b^c \end{bmatrix} + \mathbf{b}^w \right) \quad (7.17)$$

$$\mathbf{c}_{b,e}^w = \mathbf{f}_{b,e}^w \odot \mathbf{c}_b^c + \mathbf{i}_{b,e}^w \odot \tilde{\mathbf{c}}_{b,e}^w$$

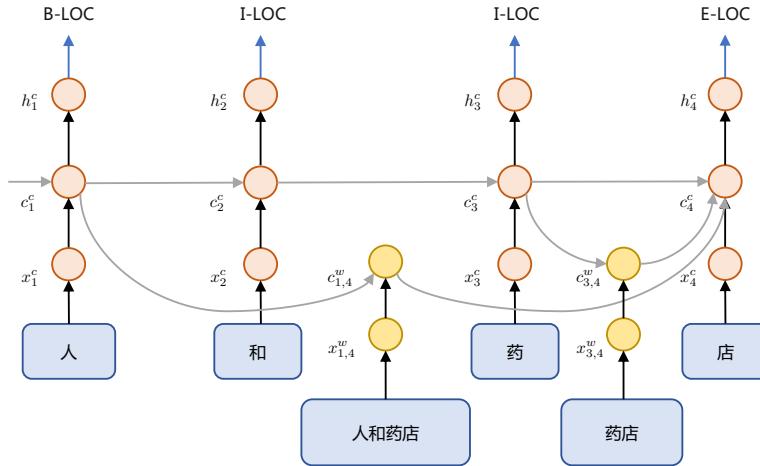
其中,  $\mathbf{i}_{b,e}^w$  和  $\mathbf{f}_{b,e}^w$  分别是输入门和遗忘门。对于词汇单元没有输出门。

由于加入了词汇单元  $\mathbf{c}_{b,e}^w$ , 在计算当前字符的单元向量时  $\mathbf{c}_j^c$ , 会有多条路径的信息流。例如, 对于“店”的单元状态 (cell state) 计算, 即不仅包含它本身的信息 (“店”字本身), 还有对应匹配的词典信息 (“人和药店”, “药店”), 这里引入一个额外的门控单元  $\mathbf{i}_{b,e}^c$  来控制每个词汇  $\mathbf{c}_{b,e}^w$  的权重:

$$\mathbf{i}_{b,e}^c = \sigma \left( \mathbf{W}^{l\top} \begin{bmatrix} \mathbf{x}_e^c \\ \mathbf{c}_{b,e}^w \end{bmatrix} + \mathbf{b}^l \right) \quad (7.18)$$

在此基础上再引入注意力机制, 计算得到当前字符的单元状态:

$$\mathbf{c}_j^c = \sum_{b \in \{b' | w_{b',j}^d \in \mathbb{D}\}} \boldsymbol{\alpha}_{b,j}^c \odot \mathbf{c}_{b,j}^w + \boldsymbol{\alpha}_j^c \odot \tilde{\mathbf{c}}_j^c \quad (7.19)$$

图 7.5 引入词汇单元的栅格结构长短时记忆网络结构图<sup>[340]</sup>

其中  $\alpha_{b,j}^c$  和  $\alpha_j^c$  由  $i_{b,j}^c$  和  $i_j^c$  归一化得到：

$$\alpha_{b,j}^c = \frac{\exp(i_{b,j}^c)}{\exp(i_j^c) + \sum_{b' \in \{b'' | w_{b'',j}^d \in \mathbb{D}\}} \exp(i_{b',j}^c)} \quad (7.20)$$

$$\alpha_j^c = \frac{\exp(i_j^c)}{\exp(i_j^c) + \sum_{b' \in \{b'' | w_{b',j}^d \in \mathbb{D}\}} \exp(i_{b',j}^c)} \quad (7.21)$$

最后，可以在隐藏层  $h_1^c, h_2^c, \dots, h_n^c$  之上添加标准 CRF 层得到最终输出，相关公式可以参考第 2 章公式 2.11、公式 2.12 和公式 2.13。

#### 4. 基于词嵌入融合词典知识的可适应中文命名实体识别算法

基于栅格网络可以有效的融合词典知识，但是 LSTM 的速度较慢，无法很好并行化。文献 [341] 提出了一种基于词嵌入融合词典知识的可适应中文命名实体识别算法 SoftLexicon。通过将句子的所有匹配词典信息融入词嵌入向量，该方法可以适应不同的神经网络结构，并优化词典融合的效率。

具体来说，对于输入汉字序列  $c_1 c_2 \dots c_n$ ，SoftLexicon 方法为每个汉字  $c_i$  的字嵌入向量引入一个词汇特征，从而构建 SoftLexicon 特征作为表示编码层的输入。该特征通过以下三个步骤构建：

- (1) 构建词汇集合：为了保留分词信息，每个字  $c_i$  对应的所有与词典匹配的词汇被划分成四

个集合——“BMES”，对应四种分词标签，该四个词汇集合的构成方式可以形式化为：

$$\begin{aligned} B(c_i) &= \{w_{i,k}, \forall w_{i,k} \in L, i < k \leq n\}, \\ M(c_i) &= \{w_{j,k}, \forall w_{j,k} \in L, 1 \leq j < i < k \leq n\}, \\ E(c_i) &= \{w_{j,i}, \forall w_{j,i} \in L, 1 \leq j < i\}, \\ S(c_i) &= \{c_i, \exists c_i \in L\}. \end{aligned} \quad (7.22)$$

其中， $L$  表示使用的词典。如果一个集合为空，那么一个特殊的词“None”将会被加入这一集中。图7.6展示了这一分类过程的一个例子。

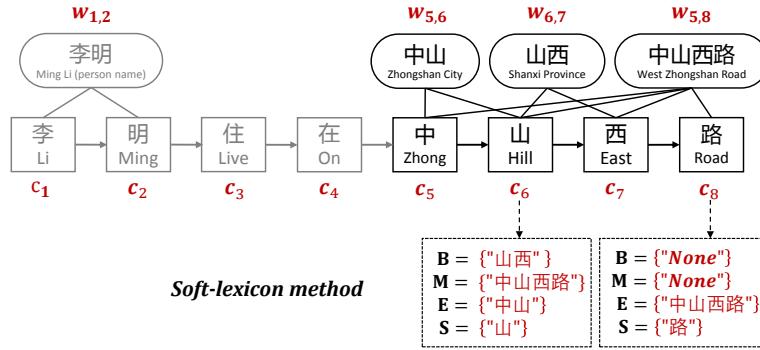


图 7.6 SoftLexicon 方法的四个词汇集合构建示例<sup>[341]</sup>

(2) 压缩词汇集合：完成四个词汇集合的构建后，SoftLexicon 方法使用一种加权算法将每个词汇集合中的词表示压缩成一个固定维度的向量。为了避免动态加权机制可能引入的复杂计算，SoftLexicon 方法使用静态词频数据获取每个词的权重。具体来说，假设  $S$  为一个词汇集合， $e^w$  表示用于查找的词汇嵌入矩阵， $z(w)$  为词典中词  $w$  在语料中出现的频率，加权的词集合  $S$  由如下方式得到：

$$v^s(S) = \frac{4}{Z} \sum_{w \in S} z(w) e^w(w), \quad (7.23)$$

其中，

$$Z = \sum_{w \in \text{BUMEUS}} z(w).$$

表示权重标准化。由于权重标准化在所有四个词汇集合的所有词汇上进行计算，因此能够更全面地比较并计算各词汇的权重。

(3) 将词汇向量与字嵌入向量结合：构建 SoftLexicon 特征的最后一步是将上述得到的四个集

合的向量表示成单一的向量表示，并与字向量结合。为了尽可能保留词信息，SoftLexicon 将四个集合向量直接串联，并串联在字向量表示上：

$$\begin{aligned} \mathbf{e}^s(\text{B}, \text{M}, \text{E}, \text{S}) &= [\mathbf{v}^s(\text{B}); \mathbf{v}^s(\text{M}); \mathbf{v}^s(\text{E}); \mathbf{v}^s(\text{S})], \\ \hat{\mathbf{x}}^c &\leftarrow [\mathbf{x}^c; \mathbf{e}^s(\text{B}, \text{M}, \text{E}, \text{S})]. \end{aligned} \quad (7.24)$$

其中， $\mathbf{v}^s$  表示上述的词汇集合加权函数， $\hat{\mathbf{x}}^c$  则为构建得到的 SoftLexicon 特征。

通过构建 SoftLexicon 特征，SoftLexicon 方法将输入文本匹配的所有词典信息融合入词嵌入向量中作为表示编码层的输入，该方法可以用于不同的表示编码层网络结构，包括 LSTM、CNN、Transformer 等<sup>[341]</sup>。

## 7.2.2 嵌套命名实体识别

**嵌套命名实体** (Nested Named Entity) 是指在实体的内部还存在一个或多个其他实体的类型。比如“北京大学”属于组织机构名实体，同时其中的“北京”又是地名类型的实体；“华为 P50 Pro”属于产品类型的实体，其中“华为”又是公司名类型实体。嵌套命名实体在机构名、生物名词、化学名词等类型中普遍存在，相较于非嵌套识别难度更大。

嵌套命名实体识别问题可以形式化表示为：给定一个序列  $\mathbf{X} = x_1 x_2 \cdots x_n$ ，其中  $x_i$  表示序列的第  $i$  个词或字，该序列对应的标签为  $\mathbf{Y} = y_1 y_2 \cdots y_n$ ，其中  $y_i = \{y_i^1, y_i^2, \dots, y_i^m\}$ ， $m$  为标签嵌套层数。可以看到标签  $y_i$  与非嵌套命名实体识别不同，嵌套命名实体识别的标签  $y_i$  是包含多个标签的集合。HMM、CRF 等传统的序列标注方法只能对每个位置输出一个标签，无法解决多标签问题，因此不能直接应用于嵌套命名实体识别任务。此外嵌套命名实体的标签之间可能存在依赖关系，其嵌套层次数量也不确定，这些都是嵌套命名实体识别需要解决的难点。

解决嵌套命名实体识别的基本方法可以采用基于非嵌套实体识别算法的穷举法<sup>[342]</sup>，将所有可能的单词序列都利用非嵌套实体识别算法进行识别和分类；还可以将原有标签修改为组合形式，将可能共同出现的所有类别进行组合，产生新的标签体系（如：将 B-LOC 与 B-ORG 组合构造 B-LOC|ORG 新标签）。也可以修改原有序列标注算法从单一目标到多目标，利用 KL 散度等做为损失函数进行参数训练等方法。这些基本方法虽然实现相对简单并且直接，但是存在计算消耗、标签量指数增加或者目标学习难度大等问题。本节将介绍几种实现相对复杂，但是结果较好的算法，包括：基于成分句法分析、基于跨度以及基于生成式框架的嵌套式命名实体识别算法。

### 1. 基于成分句法分析嵌套命名实体识别

在本书第 3 章中，我们介绍了成分句法理论，知道了句法范畴之间不是完全对等的，具有一定的层级关系。而嵌套命名实体中也存在着这样的层级关系，因此不仅每个句子可以通过基于上下文无关文法的成分句法分析转化成其对应的成分语法树，该句子中所包含的任意嵌套命名实体也可以转化成特定的树形结构。对比图 7.7 中句子的树形结构不难看出，嵌套命名实体识别任务可以类比成分句法分析任务，参考成分句法分析方法进行嵌套命名实体的识别。

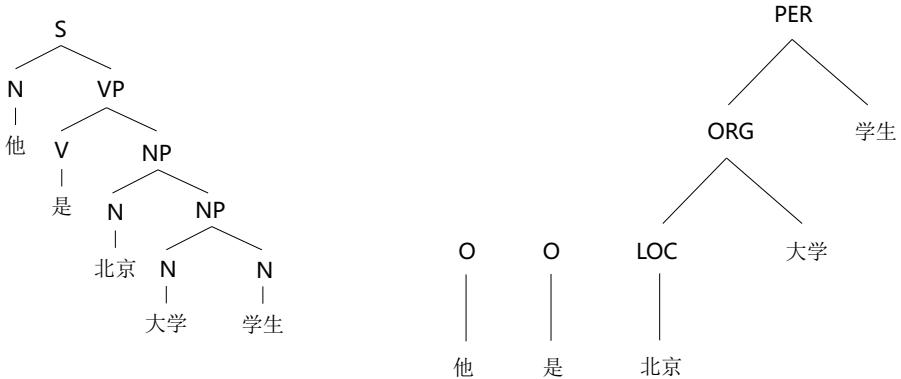


图 7.7 句子“他是北京大学学生”的成分语法树和嵌套命名实体分析

文献 [343] 提出了使用句法分析算法进行嵌套命名实体识别的思路，基于此，本节介绍一种基于状态转移的嵌套命名实体识别的方法，其基本思想与移进-规约成分句法分析类似，从左到右扫描输入的句子，并维护三个部分：堆栈  $S$ 、动作序列  $A$ 、队列  $Q$ 。堆栈  $S$  自底向上存储已规约的实体和待规约的单词，动作序列  $A$  存储规约动作的历史信息，队列  $Q$  存储尚未规约的单词。嵌套实体识别的移进-规约过程包含三种可能的动作：

- 移进 (Shift)：将非空队列  $Q$  最左端的单词移入堆栈  $S$  中；
- 规约-X (Reduce-X)：弹出堆栈  $S$  的两个元素  $s_0$  和  $s_1$ ，将它们合并标注为实体类型  $X$  后重新压入堆栈  $S$  中；
- 一元化-X (Unary-X)：弹出堆栈  $S$  的栈顶元素  $s_0$ ，将它标注为实体类型  $X$  后重新压入  $S$  中。

图 7.8 展示了该方法在处理嵌套命名实体识别任务的示例。其中  $\$$  作为句子的结束符添加在所有句子的末尾。堆栈  $S$ 、动作序列  $A$ 、队列  $Q$  用于表示当前系统的状态，算法通过其当前状态来判断接下来应该执行的动作，由此实现状态的转移，进而完成识别。由于动作只有上面提到的三种，因此可以使用与句法分析相同的框架，将该问题转换为分类问题，输入为当前的系统状态  $[S, A, Q]$ ，输出为下一步的动作。通过正确的动作序列完成嵌套命名实体的识别。

受到句法树生成通常需要引入词性信息的启发，该方法同样将词性标注的信息引入词语的表示中：

$$e_{x_i} = [e_{w_i}, e_{p_i}] \quad (7.25)$$

其中， $e_{w_i}$  是第  $i$  个词的词嵌入， $e_{p_i}$  是第  $i$  个词的词性标注嵌入。接着，该模型用两个长短期记忆网络 (LSTM) 学习任意时刻  $k$  下的历史动作序列  $A = \{a_0, a_1, \dots, a_{k-1}\}$  和队列  $Q = \{x_i, x_{i+1}, \dots, x_n\}$  的表示：

$$A_k = \overleftarrow{LSTM}_a[e_{a_0}, \dots, e_{a_{k-1}}] \quad (7.26)$$

堆栈	动作	队列
Ø	Shift	他 是 北京 大学 学生 \$
他	Shift	是 北京 大学 学生 \$
他 是	Shift	北京 大学 学生 \$
他 是 北京	Unary-LOC	大学 学生 \$
LOC   北京	Shift	大学 学生 \$
LOC 大学   他 是 北京	Reduce-ORG	学生 \$
ORG / \ LOC 大学   他 是 北京	Shift	学生 \$
ORG 学生 / \ LOC 大学   他 是 北京	Reduce-PER	\$
PER / \ ORG 学生 / \ LOC 大学   他 是 北京	Shift	\$
PER / \ ORG 学生 / \ LOC 大学   他 是 北京	Ø	Ø

图 7.8 基于成分句法分析方法识别嵌套命名实体过程示例

$$\mathbf{Q}_k = \overrightarrow{LSTM}_q[\mathbf{e}_{x_i}, \dots, \mathbf{e}_{x_n}] \quad (7.27)$$

由于堆栈 S 中存放的是树形结构，因此可以使用在第 3 章中所介绍的堆栈长短时记忆网络（Stack-LSTM）进行表示：

$$\mathbf{S}_k = \text{Stack-LSTM}[\mathbf{h}_{t_m}, \dots, \mathbf{h}_{t_0}] \quad (7.28)$$

其中  $h_{t_i}$  表示从栈顶开始往下数第  $i$  个树形元素。该树形元素的非叶结点是用循环神经网络（Recursive Neural Network）按照如下公式计算得到：

$$\mathbf{h}_{\text{parent}} = \mathbf{W}_{u,l}^T \mathbf{h}_{\text{child}} + b_{u,l} \quad (7.29)$$

$$\mathbf{h}_{\text{parent}} = \mathbf{W}_{b,l}^T [\mathbf{h}_{l\text{child}}, \mathbf{h}_{r\text{child}}] + b_{b,l} \quad (7.30)$$

其中  $W_{u,l}$  和  $W_{b,l}$  分别是父节点标签为  $l$  时，一元操作 (u) 和二元操作 (b) 对应的权重矩阵， $b$  则是相应的正则化项。其叶子结点计算公式为：

$$\mathbf{h}_{\text{leaf}} = \mathbf{W}_{\text{leaf}}^T [\mathbf{e}_{x_i}, b_k] + b_{\text{leaf}} \quad (7.31)$$

至此，可以将三个部分的表示进行拼接得到任意时刻  $k$  下的系统状态  $\mathbf{P}_k$  的表示：

$$\mathbf{P}_k = [\mathbf{S}_k, \mathbf{A}_k, \mathbf{Q}_k] \quad (7.32)$$

利用训练语料，可以构造各步骤下的系统状态  $\mathbf{P}_k$  以及所对应的动作数据，在此基础上，采用如下损失函数进行模型训练：

$$\mathcal{L}(\boldsymbol{\theta}) = - \sum_i \sum_k \log p(z_{ik}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \quad (7.33)$$

其中  $z_{ik}$  是第  $i$  个句子的第  $k$  个动作，后半部分是 L2 正则项。最后，可以利用已经训练好分类器来判断在系统状态  $\mathbf{P}_k$  下应该执行的下一步动作，从而完成嵌套命名实体识别。

## 2. 基于跨度的嵌套命名实体识别

基于跨度的嵌套命名实体识别（Span-based Nested Named Entity Recognition）<sup>[344]</sup> 方法通过对句子的子序列进行分类来识别嵌套实体。给定一个句子  $\mathbf{X} = x_1 x_2 \dots x_n$ ，其中  $x_i$  表示句子的第  $i$  个词或字，该句子中最大包含  $k$  个词或字的所有连续子序列  $x_i x_{i+1} \dots x_j$  都会进行判定是否为某类实体。基于跨度的方法可以在一定程度上解决基于状态转移方法存在错误传播问题，但是传统的基于跨度的方法通常需要对所有子序列进行分类识别，因此推理效率很低。

针对缺乏跨度边界监督以及推理效率低等问题，文献 [345] 提出了边界增强的基于跨度的嵌套命名实体识别算法 BENSC。该算法将边界检测结合到跨度分类中，通过边界监督信号学习跨度

表示，生成高质量的候选子序列来降低时间复杂度。模型总体结构如图7.9所示，主要包含边界检测和跨度分类两部分。边界检测目标是预测一个词是实体的第一个词还是最后一个词。跨度分类目标是将可能的跨度分类为相应的语义标签。两部分在多任务学习框架下联合训练。

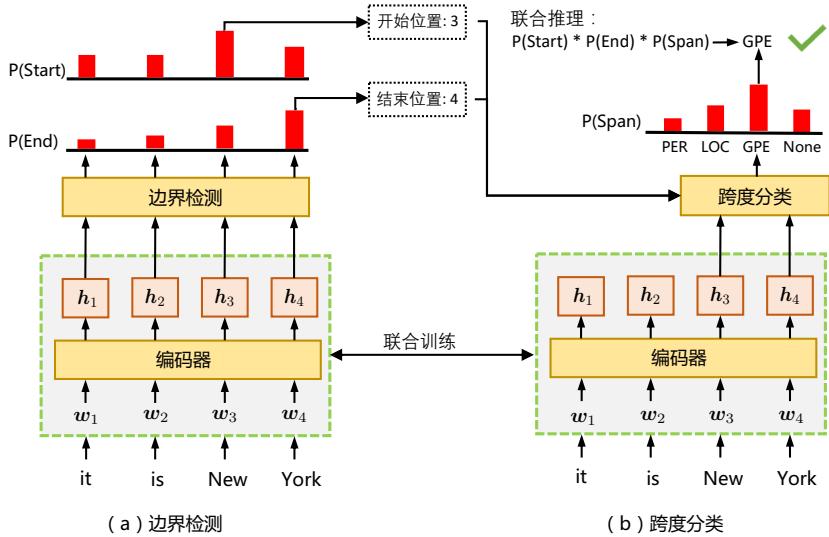


图 7.9 边界增强的基于跨度的嵌套命名实体识别模型结构图<sup>[345]</sup>

具体来说，通过编码器将编码后将文本上下文表示  $\mathbf{h}_i$  输入到多层感知器(Multilayer Perceptron, MLP) 分类器中，并应用 Softmax 函数来获得单词  $w_i$  作为实体的第一个单词的概率  $P_s^i$ ：

$$P_s^i = \text{Softmax}(MLP_{start}(\mathbf{h}_i)) \quad (7.34)$$

同样，可以应用 MLP 分类器与 Softmax 函数来获得单词  $w_i$  是实体的最后一个单词的概率  $P_e^i$ ：

$$P_e^i = \text{Softmax}(MLP_{end}(\mathbf{h}_i)) \quad (7.35)$$

在训练过程中，由于每个句子可能包含多个实体，于是将所有实体的跨度边界标记为正确答案。将边界检测任务的训练目标函数定义为以下两个交叉熵损失的总和，分别检测开始和结束边界：

$$\mathcal{L}_{bd}^s = - \sum_{i=1}^N [y_s^i \log P_s^i + (1 - y_s^i) \log(1 - P_s^i)] \quad (7.36)$$

$$\mathcal{L}_{bdr}^e = - \sum_{i=1}^N [y_e^i \log P_e^i + (1 - y_e^i) \log(1 - P_e^i)] \quad (7.37)$$

$$\mathcal{L}_{bdr} = \mathcal{L}_{bdr}^s + \mathcal{L}_{bdr}^e \quad (7.38)$$

其中  $y_s^i$  和  $y_e^i$  分别表示词  $i$  是否是实体的第一个或者最后一个词。

根据边界检测得到的可能实体的第一个词和最后一个词，使用编码器得到的跨度表示  $\mathbf{v}_{sp}$ ，同样到应用 MLP 分类器与 Softmax 函数：

$$P_{sp} = \text{Softmax}(MLP_{sp}(\mathbf{v}_{sp})) \quad (7.39)$$

$k$  表示语义标签的数量， $y_{sp}^t$  表示跨度  $(w_i, \dots, w_j)$  是否是标签  $t$ ，最小化如下的交叉熵损失函数：

$$\mathcal{L}_{sp} = - \sum_{t=1}^k (y_{sp}^t \log P_{sp}^t + (1 - y_{sp}^t) \log(1 - P_{sp}^t)) \quad (7.40)$$

总体联合训练的损失函数如下：

$$\mathcal{L} = \lambda \mathcal{L}_{bdr} + (1 - \lambda) \mathcal{L}_{sp} \quad (7.41)$$

其中  $\lambda$  是平衡边界检测和跨度分类的超参数。

模型在推断过程中，对于给定的实例  $w_i \dots w_j$ ，首先从边界检测模型中获得开始和结束的边界概率  $P_s^i$  和  $P_e^i$ 。将  $j > i$  且  $P_s^i \cdot P_e^j$  大于预先设定的阈值的区域或子序列作为合法的跨度。最后将合法的跨度输入到跨度分类中得到分类结果。

### 7.2.3 多规范命名实体识别

如前文所述，非嵌套和嵌套的命名实体识别方法只能处理特定标注范式的实体，方法之间不具备普适性及迁移性。但实际应用时可能会遇到多种范式的实体，包括非嵌套命名实体、嵌套命名实体、不连续命名实体（Discontinuous Named Entity）等不同的子任务。针对这些任务需要分别采取序列标注、跨度以及转移等方法分别进行处理，这些算法很难同时处理上述所有子任务。针对上述问题，需要一个能同时处理不同实体范式的统一命名实体识别方法。

文献 [346] 提出了一个面向命名实体识别不同任务类型的统一处理方法，采用序列到序列（Seq2Seq）生成框架，使用指针方式将序列标注任务转化为序列生成任务。同时，将预训练的序列到序列模型 BART 融入框架。首先，以生成式的序列对不同类型的实体进行统一表示。对于给定的由  $n$  个词组成的输入语句  $\mathbf{x} = x_1 x_2 \dots x_n$ ，为了能够将不同类型的实体进行统一表示，文献 [346] 提出了使用序列  $\mathbf{y} = s_1^1 e_1^1 \dots s_j^1 e_j^1 t^1 \dots s_1^i e_1^i \dots s_k^i e_k^i t^i$  进行表示的方式。其中  $s, e$  分别是一个实体跨度的起止索引， $t$  表示实体类型。对于非嵌套和嵌套命名实体识别任务，每个实体只包含一个跨度，但是对于不连续 NER 任务，每个实体包含多个跨度，因此每个实体表示为  $s_1^i e_1^i \dots s_k^i e_k^i t^i$ ，其中  $t^i$

是实体标签索引。定义  $\mathbf{G} = \{g_1, g_2, \dots, g_l\}$  为实体标签列表， $l$  是实体标签数量，为了和指针索引区分，对  $t^i$  做一个长度为  $n$  的偏移，即  $t^i \in \{n + 1, \dots, n + l\}$ 。

例如： 输入：北京大学

输出：1 2 5 1 4 6

其中，输出序列中 1 和 2 表示第 1 个实体跨度的起止索引，5 表示 LOC 实体类型，接下来输出序列中的 1 和 4 表示第 2 个实体跨度的起止索引，6 表示 ORG 实体类型。

根据上述定义，通过对目标序列  $\mathbf{y}$  的条件概率进行建模得到句子中的实体表示。

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^m P(y_t|\mathbf{x}, \mathbf{y}_{<t}) \quad (7.42)$$

针对条件概率最大化问题，可以采用具有生成能力的 BART 模型求解。模型主要包含两部分：编码器和解码器，模型结构如图 7.10 所示。

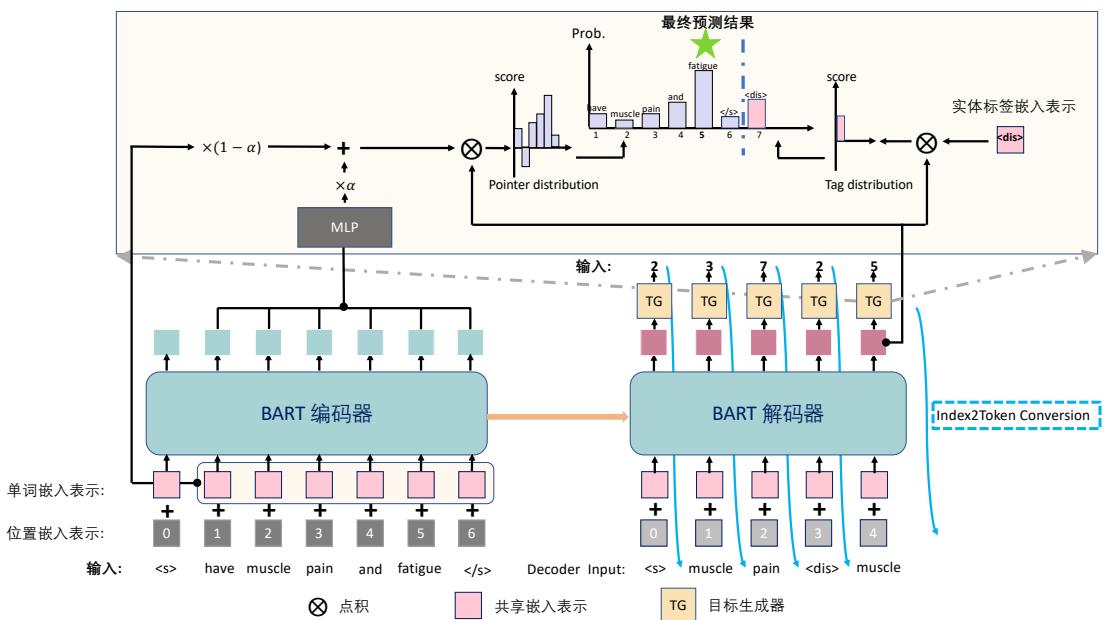


图 7.10 基于序列到序列的多规范命名实体识别方法的模型结构<sup>[346]</sup>

编码器将输入的语句  $\mathbf{x}$  编码为词嵌入表示：

$$\mathbf{E}^e = \text{TokenEmbed}(\mathbf{x}) \quad (7.43)$$

然后通过特征抽取，得到特征向量表示  $\mathbf{H}^e$ ：

$$\mathbf{H}^e = \text{Encoder}(\mathbf{x}) \quad (7.44)$$

在通过一个多层感知器后，将得到的结果  $\hat{\mathbf{H}}^e$  与词嵌入表示  $\mathbf{E}^e$  按不同权重相加：

$$\overline{\mathbf{H}}^e = \alpha * \hat{\mathbf{H}}^e + (1 - \alpha) * \mathbf{E}^e \quad (7.45)$$

其中  $\alpha$  是超参数。

解码器在每个时刻计算索引的概率分布  $P_t = P(y_t | \mathbf{x}, \mathbf{y}_{<t})$ ，输出为指针索引或者实体标签索引：

$$\hat{y}_t = \begin{cases} X_{y_t} & \text{if } y_t \leq n \\ G_{y_t-n} & \text{if } y_t > n, \end{cases} \quad (7.46)$$

然后将编码器的输出和索引表示输入到解码器：

$$h_t^d = \text{Decoder}(\mathbf{H}^e; \hat{y}_{<t}) \quad (7.47)$$

用如下公式计算出索引的概率分布：

$$P_t = \text{softmax}([\overline{\mathbf{H}}^e \otimes h_t^d; G^d \otimes h_t^d]) \quad (7.48)$$

其中

$$G^d = \text{TokenEmbed}(G) \quad (7.49)$$

最终，使用算法7.1可以将目标索引序列将其解析成抽取出的实体及类型。

值得注意的是，BART 使用的字节对编码（Byte-Pair-Encoding，BPE）编码方式会将一个单词处理成若干子词，所以 BART 框架在使用时需要进行适当的修改，可以采用如下三种基于指针的实体表示来明确地定位原句子中的实体：

- Span：实体的每个起始位置与结束位置。
- BPE：实体中每个词的所有 BPE 对应的位置索引。
- Word：每个实体的字的第一个 BPE 对应的位置索引。

以图7.11所示的句子为例，句子中有三个实体， $(x_1, x_3, \text{PER})$ ， $(x_1, x_2, x_3, x_4, \text{LOC})$ ， $(x_4, \text{ORG})$ ，其中 PER、LOC 和 ORG 是实体类型。利用上述三种表示方法，可以得到如下表示：

- Span：[0,2,5,5,PER]，[0,7,LOC]，[6,7,ORG]
- BPE：[0,1,2,5,PER]，[0,1,2,3,4,5,6,7,LOC]，[6,7,ORG]
- Word：[0,5,PER]，[0,3,5,6,LOC]，[6,ORG]

在大多数情况下，使用 Word 实体表示的结果会更好。与 Span 实体表示相比，由于 BPE 实体

**代码 7.1:** 将实体表示序列转换成实体跨度

---

**输入:** 目标序列  $y = y_1 \dots y_m$ ,  $y_i \in [1, n + |G|]$   
**输出:** 实体跨度  $E = (e_1, t_1), \dots, (e_i, t_i)$

```

1:  $E = [], i = 1$ 
2: while  $i <= m$  do
3:    $y_i = Y[i]$ 
4:   if  $y_i > n$  then
5:     if  $\text{len}(e) > 0$  then
6:        $E.\text{add}((e, G_{y_i-n}))$ 
7:     end if
8:    $e = []$ 
9:   else
10:     $e.append(y_i)$ 
11:   end if
12:    $i = i + 1$ 
13: end while
14: return E

```

---

表示与预训练任务更相似，所以效果比 Span 实体表示效果更好。但是当数据集中一个实体被标记成很长的 BPE 序列的时候，Span 实体表示的效果会优于其他两种。

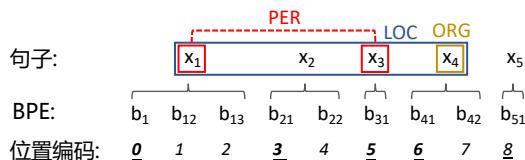


图 7.11 三种基于指针的实体表示方法

#### 7.2.4 命名实体识别评价方法

一般意义上，命名实体识别任务同时涉及两种识别目标，即实体的边界和实体的类型。相应地，对于精确匹配评估而言，实体识别任务的成功标准为实体边界以及实体类别同时被精确地识别。依据预测结果与真实结果之间的关系，能够得出 NER 任务的精确率，召回率以及 F-score 用于评估任务优劣。

绝大多数的 NER 任务涉及对多种实体类别进行识别，这就要求对所有的实体类别评估 NER 的效果。为了实现这一目的，与词性标注问题相同，通常借助两类评估指标，即宏平均 F1 (Macro-F1) 和微平均 F1 (Micro-F1)。宏平均 F1 值将所有的实体类别都视为平等的，先在单个实体类别

上计算其 F1 值，继而求得整体的平均值。微平均 F1 值将每个实体个体都视为平等的，直接对整体数据求得 F1 值。宏平均 F1 对含有较少数量实体类别更敏感。

### 7.2.5 命名实体识别语料库

命名实体识别任务与自然语言处理其他任务类似，目前也大都采用有监督方法进行建模，因此需要大规模的标注数据进行模型训练和对比评测。表7.2给出了常用的嵌套和非嵌套命名实体识别语料库汇总，并根据语料库中文本语言、文本类型以及实体类型标签数以及是否包含嵌套实体等不同可以进行划分。

表 7.2 命名实体识别语料库汇总

语料库	语言	文本类型	标签数	嵌套实体比例
CoNLL2003	英文、德文	新闻	4	0%
OntoNotes	英文、中文、阿拉伯文	新闻、广播、电话语音	11	0%
ACE 2004	英文、中文、阿拉伯文	新闻	7	17%
ACE 2005	英文、中文、阿拉伯文	新闻	7	17%
GENIA	英文	生物	4	30%
MSRA	中文	新闻	3	0%
Weibo NER	中文	社交媒体	4	0%
Resume	中文	简历	8	0%
CLUE Fine-Grain NER	中文	新闻	10	0%
CCKS2018 电子病例	中文	电子病历	5	0%

#### 1. 非嵌套英文命名实体识别语料库

大多数的英文命名实体识别的工作主要是利用 CoNLL2003<sup>[347]</sup> 和 OntoNotes 两个数据集进行训练和评测。CoNLL2003 包括 1393 篇英语新闻和 909 篇德语新闻经过人工标注组成。英语新闻来源于 Reuters Corpus Volume 1 (RCV1)。标注人员对文本内容中的实体进行了标注了，实体类型包括：人名，地名，组织名和其他实体。在数据集中还包含了每个单词的词性信息。

OntoNotes 迄今共为止发行五个版本，OntoNotes 5.0<sup>[348]</sup> 版本包含英文、中文以及阿拉伯文三种语言的新闻、广播、电话对话、博客、新闻组、脱口秀等类型的文本。该语料库中不仅包含命名实体标注，还包括语法结构、谓词论元结构等结构信息，还包含词义消歧、指代消解等语义信息。在命名实体标注方面，OntoNotes 5.0 中标注了 11 种命名实体（包括人名、地名、机构名等），还标注了 7 种类实体（包括日期、时间、百分比等）。OntoNotes 5.0 语料集规模较大，三种语言总计包含 290 万单词，仅英文的新闻语料就包含 62.5 万单词，为命名实体识别提供了很好的训练和测试基准。

## 2. 非嵌套中文命名实体识别语料库

中文命名实体识别工作基本评测语料库主要包括 MSRA<sup>[349]</sup>、OntoNotes<sup>[348]</sup>、Weibo NER<sup>[350]</sup>等。MSRA 数据集合做为 SIGHAN 2003 评测中的一部分，除了中文分词之外还标注了命名实体，主要针对人名、地名、机构名三种类型实体。随着命名实体识别研究的不断深入，针对特定领域的命名实体识别语料库也不断涌现，包括医药、生物、社交媒体、电子病历等领域数据集也受到广泛关注。Weibo NER 数据集合中包含从新浪微博收集的 226 万无标注数据和 1890 条标注数据。

## 3. 嵌套命名实体识别语料库

除了上述介绍的中英文非嵌套命名实体识别数据集，常用的含有嵌套命名实体的数据集主要为新闻领域的数据集 ACE 2004<sup>[351]</sup>、ACE 2005<sup>[352]</sup>，以及生物医学领域的数据集 GENIA<sup>[353]</sup>。ACE 2004 和 ACE 2005 数据集中主要包含 7 种实体类型，其中含有嵌套命名实体的句子占 30% 左右。ACE 2004 数据集中英文数据大约 15.8 万词，中文 15.4 万词，阿拉伯文 15.1 万词的标注规模。GENIA 数据集中主要包含 4 种实体类型，其中含有嵌套命名实体的句子占 17% 左右。GENIA V3.0 语料集针对 2000 篇 MEDLINE 系统中 2000 篇论文的摘要共计 40 万单词进行了标注。

## 7.3 关系抽取

关系抽取（Relation Extraction，RE）最初是在 1998 年 MUC-7 会议上首次正式提出，旨在从无结构文本中识别两个或多个实体之间的语义关系，是信息检索、智能问答、人机对话等应用系统中不可或缺的基础任务，也是知识图谱构建所依赖的关键技术之一。本节主要介绍二元关系抽取，关注两个实体之间的语义关系。实体间的关系可以用 <HEAD, RELATION, TAIL> 三元组进行表示，其中 Head 和 Tail 分别表示头实体和尾实体，Relation 表示实体之间的关系类型。

例如：根据句子“刘翔出生于上海”

可抽取 < 刘翔，出生地，上海 >，表示“刘翔”和“上海”之间存在“出生地”关系。

关系抽取任务的主要难度在于关系类型种类繁多以及对语义建模能力要求高。以 Freebase 知识库为例，其包含 4000 种关系类型和 7000 种属性类型<sup>[354]</sup>。如果考虑多实体关系以及关系之间的重叠，关系类型将更加复杂。面对如此庞大且不断增长的关系类型，目前大多数基于有监督方法的关系抽取任务通常根据应用的不同，构建特定领域的关系抽取模型，从而大幅度降低了模型的复杂程度。但是在处理不同领域任务时，需要重复进行关系类型定义、标注数据收集、模型训练等环节，这在一定程度上制约了关系抽取算法的通用性。此外，描述实体之间关系的语言丰富，形式也多种多样，这进一步增加了关系抽取任务的难度。

例如：(1) 复旦大学始创于 1905 年，位于中国上海。

(2) 复旦大学 地址：上海市杨浦区邯郸路 220 号

(3) 杨浦区域坐落着 14 所各类高等院校，包括：复旦大学、同济大学等

上述三个句子都表明了“复旦大学”和“上海”之间存在“位于”关系，但是其表达形式之间的差别却

非常大，如何能够建模这种长距离、丰富内容且形式变化多样的语义关系是关系抽取算法迫切需要解决难题。

从关系抽取任务定义可以看到其目标是识别实体间语义关系，因此依据是否已经在无结构文本中标记了实体类型，关系抽取方法可分为联合式抽取和流水线式抽取。联合式抽取是指利用单个算法完成从文本中同时完成命名实体识别和关系抽取。流水线式抽取则是首先使用命名实体识别算法识别文本中的实体，然后构造关系抽取模型识别实体对间的关系。两种方法各有利弊，流水线式方法可以将复杂的关系抽取任务拆解，从而降低单个模型的复杂程度，但是会带来错误传递的风险。而联合式抽取算法可以有效缓解错误传递的问题，但是模型相对复杂，文本中不仅包含有语义关联的实体对，也包含单纯的命名实体，这会造成模型学习难度的急速提升，影响模型学习效果。

根据关系类型是否需要提前预先定义，关系抽取算法可以分为预定义关系抽取和开放关系抽取两类。预定义关系抽取是针对一个或者多个领域内预先定义的实体间关系进行抽取。开放关系抽取则针对不限定领域的范围和关系类别的抽取任务。预定义关系抽取算法还可以根据所使用的训练数据是标注数据还是外部知识自动标注的情况细分为有监督和远程监督等类型。本节中将分别针对有监督关系抽取、远程监督关系抽取以及开放关系抽取三类方法进行介绍。

### 7.3.1 有监督关系抽取

有监督的关系抽取方法将关系抽取问题转换为多分类问题。其输入为文本内容和待判断的提及对（Mention Pair），输出为提及对之间根据给定的文本内容所表达的关系类型。有监督的关系抽取方法需要大量的人工标注训练语料，通过设计有效的特征来构建各类分类模型。或者利用深度神经网络自动提出语义特征进行关系抽取。本节将介绍基于最大熵关系抽取算法和基于图卷积网络的关系抽取方法。

#### 1. 基于最大熵的关系抽取

基于最大熵的统计建模方法可以很好综合地考虑文本中的词汇、句法、语法和语义特征，因此在自然语言处理有广泛的应用。最大熵模型是由最大熵原理推导得出的用于分类的对数线性机器学习模型。其基本原理是：对于一个未知分布，在只掌握部分信息的情况下，不能对未知信息引入任何主观的假设，同时应该充分利用已经掌握的已知信息。最大熵模型假设熵值最大的概率分布能够最真实地反映事件的分布情况。

熵度量了事件的不确定性，对于越不确定的事件，其熵值就越大。具体来说，对于离散型随机变量  $X$ ， $P(x)$  表示  $X$  的概率分布， $X$  的熵可以表示为：

$$H(X) = - \sum_x P(x) \log P(x) \quad (7.50)$$

对于离散型随机变量  $Y$ , 在已知随机变量  $X$  的条件下, 条件熵  $H(Y|X)$  表示为:

$$H(Y|X) = \sum_{i=1}^n P(x_i) H(Y|X = x_i) = - \sum_{x,y} P(x)P(y|x) \log P(y|x) \quad (7.51)$$

最大熵原理的目标是在所有满足约束条件的概率分布中, 选择使熵值最大的概率分布。因此, 基于最大熵原理的目标函数为:

$$\max H(Y|X) = - \sum_{x,y} P(x)P(y|x) \log P(y|x) \quad (7.52)$$

为了方便使用凸优化的方法, 通常将最大值问题改写为最小值问题, 即最终的目标函数为:

$$\min -H(Y|X) = \sum_{x,y} P(x)P(y|x) \log P(y|x) \quad (7.53)$$

给定一个具有  $n$  个样本的训练集  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , 其中  $x_i$  为模型的输入特征,  $y_i$  为  $x_i$  所对应的样本类别输出。定义  $m(X = x, Y = y)$  为样本  $(x, y)$  出现的次数, 随机变量  $X$  和  $Y$  的联合分布  $P(X, Y)$  的经验分布  $\tilde{P}(X, Y)$  以及  $X$  的边缘分布  $P(X)$  的经验分布  $\tilde{P}(X)$  可以被定义为:

$$\tilde{P}(X, Y) = \frac{m(X = x, Y = y)}{n} \quad (7.54)$$

$$\tilde{P}(X) = \frac{m(X = x)}{n} \quad (7.55)$$

对于训练集中存在的一些事实关系可以通过特征函数描述。同一个样本可以具有多个特征函数, 并最终通过特征函数来约束模型。对于输入特征  $x$  和类别输出  $y$  的特征函数  $f(x, y)$ :

$$f(x, y) = \begin{cases} 1, & x, y \text{ 满足某个条件} \\ 0, & \text{其他} \end{cases} \quad (7.56)$$

特征函数  $f(x, y)$  关于经验分布  $\tilde{P}(X, Y)$  的期望  $E_{\tilde{P}(f)}$  为:

$$E_{\tilde{P}(f)} = \sum_{x,y} \tilde{P}(X, Y) f(x, y) \quad (7.57)$$

特征函数  $f(x, y)$  关于条件分布  $P(Y|X)$  和经验分布  $\tilde{P}(X)$  的期望  $E_{P(f)}$  可以表示为:

$$E_{P(f)} = \sum_{x,y} \tilde{P}(X) P(y|x) f(x, y) \quad (7.58)$$

假设训练集中有  $M$  个特征函数  $\{f_i(x, y)\}_{i \in [1, M]}$ ，分别对应最大熵模型的  $M$  个约束条件。模型从特征  $f_i$  中学习参数，上述两个数学期望值应相等，即：

$$E_{P(f)} = E_{\tilde{P}(f)} \quad (7.59)$$

最大熵模型的优化问题可以表示为：

$$\begin{aligned} \arg \min_{p \in P} -H(P) &= \sum_{x, y} \tilde{P}(x) P(y|x) \log P(y|x) \\ s.t. \quad E_{P(f_i)} &= E_{\tilde{P}(f_i)}, i = 1, \dots, M \\ \sum_y P(y|x) &= 1 \end{aligned} \quad (7.60)$$

拉格朗日乘子法可以在满足约束条件下求解目标函数的最优解。可以证明，满足约束条件的最优解可以表示为：

$$\hat{P}(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^M \omega_i f_i(x, y)\right) \quad (7.61)$$

其中，

$$Z(x) = \sum_Y \exp\left(\sum_{i=1}^M \omega_i f_i(x, y)\right) \quad (7.62)$$

最大熵模型的训练过程可以简述为选取有效的特征  $f_i$  及其权重  $\omega_i$ 。所以能够包含更多语义信息的特征条件对最大熵模型的训练尤为关键。

针对关系抽取问题，提及对之间所具备的关系需要通过输入的文本内容进行确定，因此关系抽取中通常需要抽取提及本身以及提及对之间的特征，常见的特征有：

- (1) **单词特征**：提及对中两个提及的单词和这两个提及中间的所有单词。
- (2) **实体类型特征**：两个提及所表示的实体的类型。
- (3) **重叠性特征**：两个提及中间的单词数目；两个提及中间其他提及的数量以及指示这两个提及是否是相同的词性（如名词或者动词）的标志。
- (4) **依赖性特征**：依存关系树中提及对所依赖的单词以及单词的词性和组块标签。
- (5) **语法树特征**：语法树中连接这提及对的非终结符（去掉重复项）的路径。

例如，“上市公司拒绝了曾担任其董事会主席-鲍勃的请求”中片段“担任其董事会主席”所对应的语法树如图7.12所示。

对于该句子片段中的一个提及对“董事会”和“主席”，可以从特征规则中抽取的特征如下所示：

单词特征： $PERSON_{m1}$ （“主席”）， $ORG_{m2}$ （“董事会”）。

实体类型特征： $NOMINAL_{m1}$ ,  $NOMINAL_{m2}$ 。

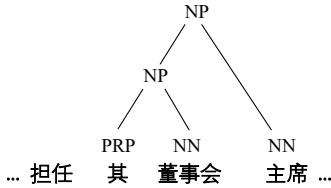


图 7.12 句子片段“担任其董事会主席”所对应的语法树结构图

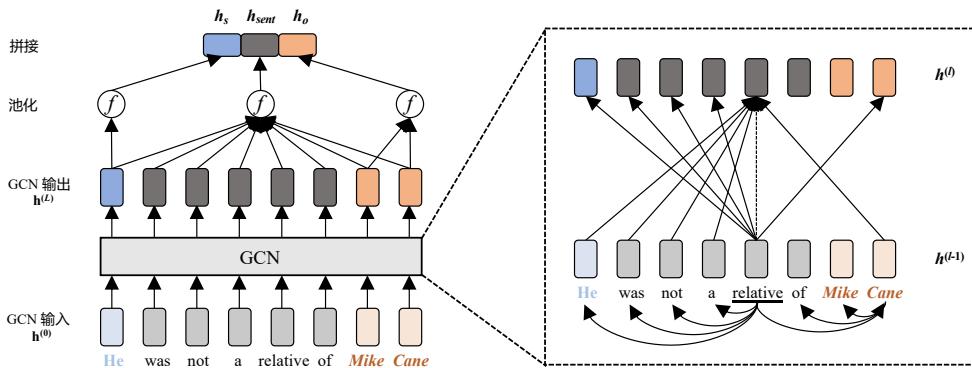
重叠性特征：两个提及中间的单词数目为 0。

语法树特征：语法树中连接这两个提及的非终结符的路径（PERSON-NP-ORGANIZATION）。

基于最大熵的关系抽取方法使用根据上述特征规则中提取到的特征来训练最大熵模型。通过将文本中各式各样的丰富的语义信息整合到特征中，并配合最大熵模型，从而实现具有良好拓展性和表现的有监督关系抽取模型。

## 2. 基于图卷积网络的关系抽取

关系抽取需要根据句子抽取实体之间的关系，句子的句法结构提供了捕捉单词之间的长距离关系的有效信息。文献 [355] 提出了通过图卷积神经网络（Graph Convolutional Network, GCN）有效利用句子依存句法树结构的关系抽取算法。通过图卷积操作对输入句子的依存结构进行编码，然后提取以实体为中心的表示，从而进行关系预测。模型结构如图 7.13 所示。

图 7.13 基于图卷积网络的关系提取模型结构<sup>[355]</sup>

给定一个拥有  $n$  个节点的图，可以使用一个  $n \times n$  的邻接矩阵  $A$  来表示这个图结构。其中，如果节点  $i$  到节点  $j$  之间存在一条边的话，则有  $A_{ij} = 1$ 。在一个拥有  $L$  层的图卷积网络中，假定

第  $l$  层中节点  $i$  的输入向量为  $\mathbf{h}_i^{(l-1)}$ , 输出向量为  $\mathbf{h}_i^{(l)}$ , 那么一个图卷积操作可以表示为:

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j=1}^n A_{ij} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)} + b^{(l)}\right) \quad (7.63)$$

其中,  $\mathbf{W}^{(l)}$  是线性转换操作,  $b^{(l)}$  是偏置项,  $\sigma$  是非线性函数 (如 ReLU 等)。在每次图卷积的过程中, 图中的每个节点会汇聚并总结来自于相邻节点的信息。

为了将图卷积操作应用于依存树, 可以将每棵依存树转换成相应的邻接矩阵  $\mathbf{A}$ 。其中, 如果单词  $i$  和  $j$  之间存在一条依存边的话, 则有  $A_{ij} = 1$ 。然而, 因为单词之间差别很大, 简单地应用公式7.63中的图卷积运算可能会导致节点表示的幅值差异很大。这可能会使句子表征偏向于高度节点, 而忽略节点本身所携带的信息是什么。除此之外, 因为节点在依存树中不会与自己相连接, 所以  $\mathbf{h}_i^{(l-1)}$  的信息将永远不会传递给  $\mathbf{h}_i^{(l)}$ 。

为了解决上述问题, 文献 [355] 提出在向图卷积网络输入非线性信息之前对其隐激活进行标准化, 并向图中的每个节点添加自循环, 具体做法如下:

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j=1}^n \tilde{A}_{ij} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)} / d_i + b^{(l)}\right) \quad (7.64)$$

其中,  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ ,  $\mathbf{I}$  是  $n \times n$  的单位矩阵, 这就相当于给每个节点加自循环,  $\mathbf{h}_i^{(l-1)}$  的信息就可以传递给  $\mathbf{h}_i^{(l)}$ ;  $d_i = \sum_{j=1}^n \tilde{A}_{ij}$  是生成图中词元  $i$  的度, 将它放在分母项就可以对节点的度进行标准化, 解决了节点表示幅值差异过大的问题。把上述操作叠加在  $L$  层上就形成了一个深层的图卷积网络。其中, 将  $\mathbf{h}_1^{(0)}, \dots, \mathbf{h}_n^{(0)}$  作为输入词向量,  $\mathbf{h}_1^{(L)}, \dots, \mathbf{h}_n^{(L)}$  作为输出词特征。

使用  $\mathbf{x} = x_1 \dots x_n$  表示输入句子, 其中  $x_i$  是第  $i$  个单词。头实体表示为  $\mathbf{x}_s = x_{s_1} \dots x_{s_2}$ , 尾实体表示为  $\mathbf{x}_o = x_{o_1} \dots x_{o_2}$ 。假如预先给定一个关系集合  $\mathbf{R}$ , 并且已经给出  $\mathbf{x}$ 、 $\mathbf{x}_s$ 、 $\mathbf{x}_o$ , 关系抽取的目标就是预测实体间的关系  $r \in \mathbf{R}$  或是 “无关系”。

在词向量上应用了  $L$  层 GCN 之后, 可以获得每个单词的隐藏表示, 这些单词在依存树中直接受其距离不超过  $L$  条边的邻居节点的影响。为了在关系抽取任务中利用这些词表示, 首先定义句子表示:

$$\mathbf{h}_{\text{sent}} = f(\mathbf{h}^{(L)}) = f(\text{GCN}(\mathbf{h}^{(0)})) \quad (7.65)$$

其中,  $\mathbf{h}^{(l)}$  代表着 GCN 在第  $l$  层的集合隐藏表示;  $f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^d$  是一个从  $n$  个输出向量映射到句子向量的最大池化函数。同样地, 我们可以从  $\mathbf{h}^{(L)}$  获取到头实体和尾实体的隐藏表示, 如下所示:

$$\mathbf{h}_s = f(\mathbf{h}_{s_1:s_2}^{(L)}) \quad (7.66)$$

$$\mathbf{h}_o = f(\mathbf{h}_{o_1:o_2}^{(L)}) \quad (7.67)$$

通过连接句子和实体的表示，并且将其输入前馈神经网络从而获得用于分类的最终表示：

$$\mathbf{h}_{\text{final}} = \text{FFNN}([\mathbf{h}_{\text{sent}}; \mathbf{h}_s; \mathbf{h}_o]) \quad (7.68)$$

将最终得到的隐藏表示输入到一个线性层中，然后再利用 Softmax 函数去获取关系的概率分布。最后，模型根据这个概率分布去推测出输入实体的最合适的关系。

### 7.3.2 远程监督关系抽取

有监督的关系抽取方法虽然准确率较高，模型结果更为可靠，但是需要人工标注数据集。然而，构造这样的数据集需耗费大量的人力和时间。由于关系种类相较于实体种类更加复杂多样，并且不断涌现，针对所有关系抽取都需要预先标注大量样本，制约了关系抽取的更广泛的应用。近年来，为了实现自动化关系抽取，研究人员们提出了远程监督（Distant Supervision）方法。远程监督方法假设知识库中两个实体存在某种关系，那么所有提及了这两个实体的句子都表达了这种关系。图7.14展示了通过远程监督方法自动标注的实例，在这个例子中，“Apple”和“Steve Jobs”是 Freebase 知识库中两个实体，其关系为“/business/company/founders”，根据远程监督方法的假设，在语料集中所有包含这两个实体的句子都会被标注为这种关系，并作为训练样本用于训练。

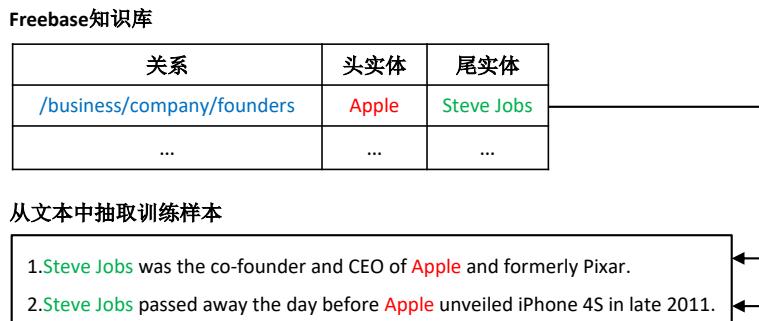


图 7.14 远程监督方法生成的训练样本示例

虽然利用远程监督方法可以迅速构建大量训练样本，但是远程监督方法构建标签的假设过于宽松，非常容易导致大量的错误标注，从而严重影响模型的性能。从图7.14中所给出的例子中也可以看到，虽然句子 2 中包含了“Apple”和“Steve Jobs”两个词语，但是该句并没有表达“/business-company/founders”关系。因此，大量的远程监督方法关注于如何缓解训练集中的噪音信号，从而获得高性能的抽取器。本节将介绍远程监督关系抽取的两种经典方法：多示例学习方法和基于注意力的方法。

### 1. 基于多示例学习的远程监督关系抽取

多示例学习是指对具有某种特征的数据样本集合进行标注，这样的样本集合称为包（Bag），模型在包级别上进行训练与推断。其形式化定义为给定示例集合  $x = \{x_1, x_2, \dots, x_n\}$ ，根据某种映射，将示例集合映射到包  $B = \{B_1, B_2, \dots, B_M\}$ ，即将提及相同实体对的示例映射到同一个包中，然后经过机器学习或深度学习模型  $f$ ，将包映射到标签空间  $L = \{L_1, L_2, \dots, L_T\}$ 。多示例学习可以用于缓解远程监督关系抽取中的错误标注问题。

PCNN (Piecewise Convolutional Neural Networks) [356] 是一种应用了多示例学习来处理远程监督关系抽取的模型，其神经网络结构如图7.15所示，主要分为四部分：向量表示层、卷积层、分段最大池化层和输出层。

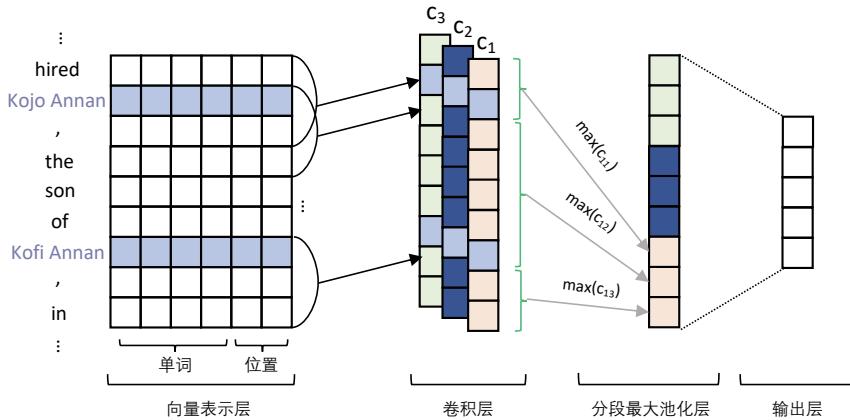


图 7.15 PCNN 模型结构<sup>[356]</sup>

向量表示层用于将词转化为低维的向量表示，从而输入到神经网络中。向量表示分为两部分，分别是词向量和位置向量，经过预训练的词向量可以捕获词的语法和语义信息，从而更好地适用于下游任务，位置向量用于编码头尾实体的位置信息。词向量和位置向量拼接后作为向量表示输入到模型中。

卷积层用于对句子的局部特征进行建模，卷积操作的计算如下所示：

$$c_{ij} = \mathbf{w}_i \mathbf{q}_{j-w+1:j}, 1 \leq i \leq n \quad (7.69)$$

其中  $\mathbf{w}_i$  为第  $i$  个卷积核， $\mathbf{q}_j$  为输入句子中第  $j$  个词对应的向量表示， $n$  为卷积核数量。

分段最大池化层根据卷积操作得到的局部特征提取全局特征。不同于单一最大池化方法，分段最大池化方法首先以头尾实体为界将卷积得到的特征图划分为三部分，如图7.15所示，每个卷积核输出的特征图  $c_i$  被实体“Kojo Annan”和“Kofi Annan”划分为三部分  $\{c_{i1}, c_{i2}, c_{i3}\}$ ，然后对每

一部分分别进行最大池化操作，各特征图进行分段最大池化后拼接，经过非线性函数后得到全局的特征表示。分段最大池化的具体计算公式如下：

$$p_{ij} = \max(c_{ij}), 1 \leq i \leq n, 1 \leq j \leq 3 \quad (7.70)$$

$$p_i = \{p_{i1}, p_{i2}, p_{i3}\} \quad (7.71)$$

$$g = \tanh(p_{1:n}) \quad (7.72)$$

输出层根据全局特征使用 Softmax 激活函数计算在各关系类别上的概率分布，从而对关系类别作出预测。

为了解决远程监督带来的错误标注问题，PCNN 使用了多示例学习方法，所使用的包级别损失函数定义如下：

$$\mathcal{L}(\theta) = - \sum_{i=1}^T \log p(y_i | m_i^j; \theta) \quad (7.73)$$

其中， $j$  有如下约束：

$$j^* = \arg \max_j p(y_i | m_i^j; \theta), 1 \leq j \leq q_i \quad (7.74)$$

即  $m_i^j$  为每个包  $M_i = \{m_i^1, m_i^2, \dots, m_i^{q_i}\}$  中在正确标签上输出概率最高的示例。

多示例学习的完整训练过程为：

- (1) 初始化网络参数  $\theta$ ，将所有包划分为若干个大小为  $b_s$  的小批次（mini-batch）。
- (2) 随机选择一个批次，根据公式 7.74 选出包中第  $j$  个示例  $m_i^j$  ( $1 \leq i \leq b_s$ )。
- (3) 基于示例  $m_i^j$  的梯度对网络参数  $\theta$  进行更新。
- (4) 重复以上两步直至收敛。

由此可见，传统的反向传播算法对所有的训练样本计算梯度进而更新网络参数，而多示例学习选择包中最符合包标签的样本对参数进行优化，在一定程度上过滤了远程监督中的噪音标签数据。

## 2. 基于注意力的关系抽取

多实例学习在远程监督的关系抽取方面有了很大的改进，但模型只取置信度最高的一个句子进行训练的方法会造成大量丰富信息的丢失。此外模型将包含相同实体的语句作为一个包，这种基于包级别的训练和推断仍然存在引入过多噪声的问题。

为了解决上述问题，文献 [357] 提出了基于句子级别的注意力机制模型，将注意力机制应用到远程监督关系抽取中。该模型通过使用卷积神经网抽取句子的语义特征，然后在多实例上构建句子级别的注意力机制，从而动态减少噪声实例的权重并全面的获取实例的信息。给出一组句子  $\{x_1, x_2, \dots, x_n\}$  和两个相对应的实体，模型评估每个关系  $r$  的可能性。模型主要包含两大部分：句子编码器和实例选择注意力，其神经网络结构如图 7.16 所示。

句子编码器包括词嵌入，卷积层，最大池化层和非线性层。给定一个句子  $x$  和两个目标实体，卷积神经网络用来构建句子的分布表示。首先，CNN 的输入是句子  $x$  的原始单词，将每个输入词通过词嵌入矩阵转换成不同的低维向量。为了指定每个实体对的位置，还对句子中的所有单词使用位置嵌入。使用卷积层来合并处理这些特征，处理方法和 PCNN 中处理方法相同，如公式7.69所示。

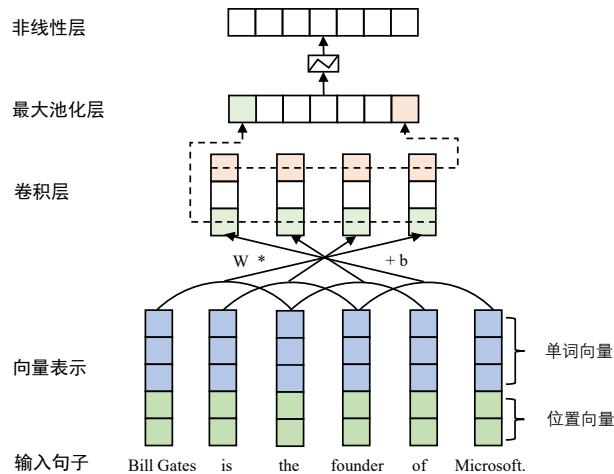


图 7.16 CNN/PCNN 模型句子编码器结构<sup>[357]</sup>

实例选择注意力使用句子级别注意力去选择表达了相应关系的句子。假设对于实体对 (Head, Tail)，存在于包含  $n$  个句子的集合  $B = \{x_1, x_2, \dots, x_n\}$  中。为了充分利用所有句子的信息，同时缓解噪音句子带来的负面影响，模型给每个句子赋予一个可学习的注意力权重，最终使用  $b$  表示为所有句子向量  $x_i$  的加权和：

$$b = \sum_i \alpha_i x_i \quad (7.75)$$

其中  $\alpha_i$  是每个句子向量  $x_i$  的权重，这里使用选择性注意力来计算，因为如果在训练和测试过程中将每个句子等分，标注错误的句子会带来大量的噪音。因此，使用选择性注意力来弱化噪声语句。 $\alpha_i$  定义如下：

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)} \quad (7.76)$$

其中  $e_i$  被称为基于查询的函数，其对输入句子  $X_i$  和预测关系  $r$  匹配的程度进行评分，这里采用双线性形式作为查询函数：

$$e_i = \mathbf{x}_i \mathbf{A} \mathbf{r} \quad (7.77)$$

其中  $\mathbf{A}$  是加权对角矩阵， $\mathbf{r}$  是与表示关系  $r$  相关联的查询向量。

最后，通过 Softmax 函数定义条件概率  $P(r|B, \theta)$  如下：

$$P(r|B, \theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)} \quad (7.78)$$

其中  $n_r$  是关系的总数， $\mathbf{o}$  是神经网络的最终输出，它对应于与所有关系类型相关联的分数，其定义如下：

$$\mathbf{o} = \mathbf{M}\mathbf{b} + \mathbf{d} \quad (7.79)$$

这里  $\mathbf{d} \in \mathbb{R}^{n_r}$  是偏差向量， $\mathbf{M}$  是关系的表示矩阵。

基于注意力的方法在多示例学习的基础上，摒弃了只采用最高置信度的句子作为训练语句的方法，而是综合考虑所有包含同一实体对的句子，对它们与预测关系的相关性给予不同的权重，进而得到更为全面、有用的信息。在充分利用包含每对实体的所有信息句的同时，通过选择注意力机制缓解了远程监督中带来的错误标签的影响。

### 7.3.3 开放关系抽取

传统的关系抽取算法，不论是有监督还是远程监督关系抽取方法，目标都是针对预先定义的关系类型在限定语料中判定实体之间是否存在预先定义的关系。因此，传统的关系抽取算法能够处理的关系数量有限，并且在处理不同领域时需要用户进行关系类别定义、数据标注以及模型训练等一系列工作。在处理海量互联网和社会媒体数据时，传统关系抽取算法受到上述问题的制约，很难适应快速发展且不断演进的需求。开放关系抽取（Open Relation Extraction, ORE）目标是在不需要预先关系定义的情况下，从非结构化文本中提取关系元组，并且不受语料库领域的限制。实体关系仍然采用三元组形式表示：(Head, Relation, Tail)。

例如：根据句子“卡塔尔发布本届世界杯吉祥物，正式取名叫做 La’eeb。”

可抽取<世界杯，在，卡塔尔>、<世界杯吉祥物，是，La’eeb>等关系

需要特别注意的是，开放关系抽取所得到关系类型描述并不是预先定义的，而是根据所给定的文本截取或生成的。因此，与预定义关系抽取不同，具有相同语义的关系类型可能存在多个描述形式。

本节将分别介绍基于抽取和基于聚类两种开放关系抽取方法。

#### 1. 基于自监督学习和人工模板的开放关系抽取算法

2007 年 Banko 等人<sup>[358]</sup>提出了开放信息抽取（Open IE, OIE）的概念，并设计了 TextRunner 系统，试图打破传统的封闭式信息抽取系统，使用一种无监督的方式，提取出类型更加多样的关系元组。TextRunner 采用自监督学习方法，通过自监督的学习器、信息抽取器和基于冗余信息的评估器等三个主要部分完成开放信息抽取：

- (1) **自监督学习器 (Self-Supervised Learner)**: 利用从整个语料集合中采样得到的小规模样本，根据自监督学习方法输出分类器，目标是对所有可能的抽取候选分类为“可信”（Trustworthy）或“不可信”（Not Trustworthy）。

- (2) 信息抽取器 (Single-Pass Extractor): 对整个语料集合中所有数据进行遍历并抽取出所有的关系三元组，并利用自监督学习器构建的候选分类器进行判别，保留分类为“可信”的三元组。
- (3) 基于冗余信息的评估器 (Redundancy-Based Assessor): 利用文献 [359] 所提出的模型利用文本中的冗余信息对三元组进行评价。

自监督学习器使用依存句法分析器解析句子的句法结构，抽取名词性短语组成三元组  $\langle \text{Head}, \text{Relation}, \text{Tail} \rangle$ ，并利用人工定义的规则将所抽取的三元组划分为正样本或负样本。如果三元组满足以下三个条件则被归类为正样本，否则被归类为负样本：

- Head 和 Tail 之间不超过一定长度；
- Head 和 Tail 之间的路径不穿过句式边界（例如关系从句）；
- Head 和 Tail 不全是代词。

在此基础上，利用人工定义的特征类型，针对所有三元组构造特征向量，并利用朴素贝叶斯分类器用来构建分类器。

信息抽取器对整个语料库进行一次遍历，对所有句子进行词性标注，并使用轻量级的名词短语分块器来识别名词短语，进而识别出实体。关系则是通过分析实体之间的文本，通过人工定义的一系列启发式规则消除非必要的短语构成。

例如：Scientists from many universities are studying. 简化为 Scientists are studying.  
definitely developed 简化为 developed

提取器从每个句子中生成一个或多个候选元组，并将每个候选元组利用自监督学习器所输出的分类器进行分类，仅保留那些被标记为可信的三元组。

在对整个语料库完成提取操作后，TextRunner 将实体和规范化关系都相同的三元组进行合并，并统计每个三元组在不同句子中出现的次数。利用上述信息，基于文献 [359] 所提出信息冗余的概念，对被抽取多次的三元组分配一个更高的置信度，说明被抽取多次的三元组更有可能表示为一种关系。最终根据置信度和预先设定的阈值输出所发现的实体和关系。

TextRunner 开放关系抽取系统，虽然减少了人工标注训练数据的开销，并且可以在多个不同领域的语料上使用，但是仍然存在一些不足：

- 不连贯的抽取 (Incoherent Extractions): 抽取的关系词语不连贯，并且没有可解释性的意义。这样的关系类别占据了 TextRunner 大约 13% 的输出。

例如：The guide contains dead links and omits sites.

TextRunner 系统抽取的关系类型：“contains omits”；

- 无意义的关系 (Uninformative Relations): 抽取忽略了关键性的信息，没有处理好动词和名词组成的多词谓语，并且名词携带了谓词的语义信息。

例如：Hamas claimed responsibility for the Gaza attack.

TextRunner 系统抽取结果：(Hamas, claimed, responsibility)

正确结果：(Hamas, claimed responsibility for, the Gaza attack)

针对以上问题，文献[360]提出了Reverb算法，利用句法约束和词汇约束对抽取三元组进行限制。句法约束目的是为了排除不连贯的抽取的问题，同时通过动词结构捕捉关系短语。具体句法约束如下所示：

$$V \mid V P \mid V W^* P$$

$V = \text{Verb Particle? Adv?}$

$W = (\text{Noun} \mid \text{Adj} \mid \text{Adv} \mid \text{Pron} \mid \text{Det})$

$P = (\text{Prep} \mid \text{Particle} \mid \text{Inf. marker})$

上述约束，给出了关系短语需要满足的词性标签，将关系短语限制为简单的动词短语、动词短语紧跟介词或小品词、动词短语接名词短语并以助词结尾这三种形式。如果在一个句子中有多个匹配则选择最长的匹配。如果该模式匹配多个相邻的序列，则将它们合并为一个单一的关系短语。这种细化匹配使得模型能够处理包含多个动词的关系短语，并且关系短语必须是句子中一个连续的单词片段。

句法约束大大减少了无意义的关系的提取，但在一些特定的情况下，非常复杂的关系短语也能够满足匹配，例如：“The Obama administration is offering only modest greenhouse gas reduction targets at the conference.” 中 “is (verb, V) offering (verb, V) only (Adv, W) modest (Adj, W) greenhouse (Noun, W) gas (Noun, W) reduction (Noun, W) targets (Noun, W) at (Prep, P)” 符合句法约束，但并不具有实际意义。为了克服这一缺陷，Reverb 算法还引入了词汇约束，用于避免抽取过度冗长的关系短语，保留有效的关系短语。该约束基于的假设是：一个有效的关系短语应该在一个大型语料库中出现过许多次，有许多不同的实体对。上个例子中的冗长短语几乎不可能存在于多个实体对，所以不能代表一个真实的关系。

## 2. 基于聚类的开放关系抽取

在开放环境中，我们无法预先定义好所有可能的关系类别标签。但是即便不知道某个关系属于哪一类，仍有可能判断一系列的关系是否是同一类。因此在开放环境的背景下，研究人员们提出了基于聚类的开放关系抽取。此类模型的核心目标是构建在很小甚至无需人类标注的前提下学习到更好的关系语义表示模型，从而在新的关系类别中识别关系三元组。SelfORE<sup>[361]</sup> 是一种基于聚类的开放关系抽取算法，其算法流程如图7.17所示。主要包含三个部分：实体对编码、编码非线性映射以及关系聚类。

实体对编码目标是需要从文本中获得关系三元组的表示，这一步骤通常使用BERT、RoBERTa等预训练模型实现。对于一个输入文本，通常编码器将输出多个向量表示，对于BERT模型而言，通常取第一个Token的标记‘[CLS]’的输出向量，作为整个句子的语义表示。但是针对开放关系抽取需要特别注意的是：一段文本中可能不止包括一对三元组的关系信息，同时还夹杂着许多对于关系抽取任务而言的无关信息，例如：“ChatGPT是由OpenAI开发的一个人工智能聊天机器人程序，于2022年11月推出，使用基于GPT-3.5架构的大型语言模型并通过强化学习进行训练。”。对于<‘ChatGPT’，‘开发’，‘OpenAI’>的关系三元组而言，其余文本信息都是与该关系无关的冗

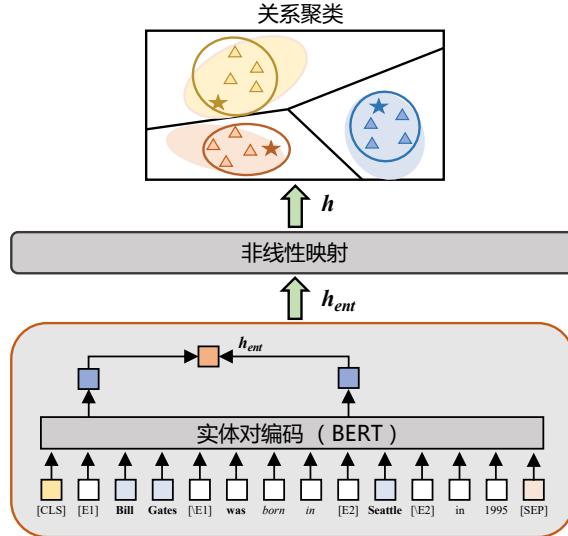


图 7.17 基于聚类的开放关系抽取基本流程

余信息，因而对于关系抽取任务而言，[CLS] 位置所对应的向量并不是表示关系三元组的最好选择。针对此问题可以采用实体标记（Entity Maker）的方法，在数据预处理阶段，将文本中需要抽取关系的实体对用特定的符号进行标记，将标记位置的向量输出作为关系三元组的向量表示。如图7.17所示，Bill Gates 和 Seattle 作为实体对词，前后都加入了特定符号 [E]。利用这种方法可以使得编码器在输出句子的关系表示时能聚焦于关键信息，从而更好的概括句子级别的语义。具体过程可以形式化表示为：

$$X = [x_1, \dots, [E1_{\text{start}}], x_i, \dots, x_{j-1}, [E1_{\text{end}}], \dots, [E2_{\text{start}}], x_k, \dots, x_{l-1}, [E2_{\text{end}}], \dots, x_T] \quad (7.80)$$

其中  $x_i$  到  $x_{j-1}$  以及  $x_k$  到  $x_{l-1}$  分别是头尾实体，该句子最终的关系向量表示为：

$$\mathbf{h} = [\mathbf{h}_{[E1_{\text{start}}]}, \mathbf{h}_{[E2_{\text{start}}]}] \quad (7.81)$$

在将文本编码成向量的关系表示后，通常需要将所有关系向量都投影到相同维度的向量空间中以便聚类。在此步骤可以针对编码器的输出结果经过一个非线性的映射层来实现从高维到低维的投影。在此步骤中也可以引入 Dropout 等方法，使得模型的聚类效果更鲁棒。非线性映射层具体可以采用如下步骤实现：

$$\tilde{\mathbf{h}} = \text{Dropout}(\mathbf{h}) \quad (7.82)$$

$$\mathbf{z} = g(\mathbf{W}_\phi \tilde{\mathbf{h}} + \mathbf{b}_\phi) \quad (7.83)$$

通过实体对编码和非线性映射两个步骤，所有输入句子将离散的嵌入到同一个低维的特征空间内。在此基础上，可以通过 K-Means 等聚类算法来为空间内的关系点聚簇，并为每个点分类一个伪标签：

$$\hat{y}^u = \text{K-Means}(\mathbf{h}^u) \quad (7.84)$$

最后针对不同簇内的样本内容以及对应的伪标签，人工判断此簇所代表的关系。这种方法对于聚类结果的效果要求较高。针对该问题，SelfORE<sup>[361]</sup>采取自监督的方法，根据样本的嵌入点与聚类质心的距离来为每个样本分配一个置信度，高置信度的数据可以将其伪标签作为监督数据来训练模型。基于  $t$  分布的置信度  $q_{nk}$  的计算方法如下所示：

$$q_{nk} = \frac{\left(1 + \|\mathbf{z}_n - \mu_k\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'} \left(1 + \|\mathbf{z}_n - \mu_{k'}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \quad (7.85)$$

其中  $\alpha$  是超参数，用于衡量  $t$  分布的自由度。 $\mu_k$  是类的质心， $\mathbf{z}_n$  是样本的嵌入向量。

开放领域的实体抽取中如何能够对于句子关系语义进行很好的表示是这类方法研究的核心。但是关系语义非常复杂，完全依赖无监督表示构建方法通常不能够精准地捕捉关系间的语义相似性。为解决上述问题，RoCORE<sup>[362]</sup> 通过利用相对容易获得的预定义关系类别标注数据，进一步增强关系表示学习，并利用多目标联合训练有效地减少预定义类别的偏置并优化实体对表示。

### 7.3.4 关系抽取评价方法

针对预先定义的关系类型，通常采用的评价指标与命名实体识别一样，包括精确率（Precision, P）、召回率（Recall, R）、F 值（F-Measure）等指标对不同类型关系进行评价，也可以利用准确率（Accuracy）、微平均 F1（Micro-F1）、微平均精度（Micro-P）、微平均召回（Micro-R）等对整体抽取效果进行评价。

而对于开放关系抽取，常用的评价指标有 V-measure、调整兰德系数（ARI）等。V-measure 是聚类任务中常用的评价指标，其基于两个类别之间的条件熵计算类别之间的同质性和完整性。ARI 是聚类任务中描述簇类相似度的指标。具体计算公式如下：

(1) 同质性（Homogeneity）度量：

$$\begin{aligned} h &= 1 - \frac{H(C|K)}{H(C)} \\ H(C|K) &= - \sum_{c=1}^C \sum_{k=1}^K \frac{n_{c,k}}{n} \log\left(\frac{n_{c,k}}{n}\right) \\ H(C) &= - \sum_{c=1}^C \frac{n_c}{n} \log\left(\frac{n_c}{n}\right) \end{aligned} \quad (7.86)$$

其中  $H(C|K)$  是给定划分条件下类别划分的条件熵， $H(C)$  为类别划分熵， $n$  表示全部实例数， $n_c$  表示类别  $c$  下的实例数， $n_k$  表示在簇  $k$  下的实例数， $n_{c,k}$  表示类别  $c$  中被划分到簇  $k$  下的实例数。

(2) 完整性 (Completeness) 度量：

$$c = 1 - \frac{H(K|C)}{H(C)} \quad (7.87)$$

(3) **V-measure** 是同质性和完整性的调和平均值：

$$v = \frac{2 \times h \times c}{h + c} \quad (7.88)$$

### 7.3.5 关系抽取语料库

自 1998 年 MUC-7 会议上首次正式提出关系抽取任务以来，无论 1999 年到 2008 年举行了 9 届的 ACE 会议，还是自 1998 年举办至今的 SemEval 会议都对关系抽取任务给予了很大的关注，发表了一些列限定域和开放域的关系抽取评测集合。表 7.3 展示了常用的关系抽取语料库，主要包括：SemEval-2010 Task 8、NYT-10、TACRED 和 FewRel。

数据集	关系类别数目	关系实例数目	应用域
SemEval-2010 Task 8	9	10,717	限定域
TACRED	42	21,784	限定域、开放域
FewRel	100	70,000	限定域、开放域
NYT-10	57	143,391	限定域、开放域

表 7.3 常见关系抽取语料库汇总

#### 1. SemEval-2010 Task 8 关系抽取语料集

SemEval-2010 Task 8 数据集<sup>[363]</sup> 主要用于限定领域关系抽取任务评价，自于 2010 年的国际语义评测大会中（SemEval）Task 8：“Multi-Way Classification of Semantic Relations Between Pairs of Nominals”。该数据集包含 10717 个样本，其中 8000 个用于训练，2717 个用于测试，标签集包含 18 种有序关系类型和 1 种未知关系类型。

#### 2. TACRED 关系抽取语料集

TACRED (TAC Relation Extraction Dataset) 数据集<sup>[364]</sup> 是一个拥有 106264 条实例的大规模关系抽取数据集，这些数据来自于每年的 TAC KBP (TAC Knowledge Base Population) 比赛使用的语料库中的新闻专线和网络文本。该数据集 41 种已知关系类型和 1 种未知关系类型，是一个典型的长尾分布数据集，其中标签为未知关系类型的数据占据了 79.5%。

### 3. FewRel 关系抽取语料集

FewRel 数据集<sup>[365]</sup> 包含 100 个类别、70,000 个实例，是目前采用人工标注的关系抽取任务中关系类别和实例数目最大的语料库。由于该语料库关系数量的多，FewRel 语料集合也经常应用于小样本学习（Few-shot Learning）和远程监督关系抽取任务中。

### 4. NYT-10 关系抽取语料集

NYT-10 数据集<sup>[366]</sup> 是在基于远程监督的关系抽取任务上最常用的数据集。其文本来源于纽约时报 New York Times，命名实体是通过 Stanford NER 工具并结合 Freebase 知识库进行标注的。命名实体对之间的关系是链接和参考外部的 Freebase 知识库中的关系，结合远监督方法所得到的。数据集中一共包含 52 个已知关系类型和 1 个未知关系类型。

## 7.4 事件抽取

事件抽取（Event Extraction）目标是从文本中发现特定类型事件，并抽取该事件所涉及的时间、地点、人物等元素。事件抽取任务可以为问答系统、文本摘要以及各类语言理解任务提供有效的结构化信息。根据美国国家标准技术研究所（NIST）组织的 ACE（Automatic Content Extraction）项目给出的定义<sup>[351]</sup>，事件由事件触发词（Trigger）以及事件论元（Argument，也称事件元素）组成。

例如：2022 年卡塔尔世界杯（FIFA World Cup Qatar 2022）是第二十二届国际足联世界杯足球赛，在当地时间 2022 年 11 月 20 日到 12 月 18 日期间在卡塔尔国内 5 个城市的 8 座球场举行。

事件类型：体育赛事

触发词：举行

赛事名称：第二十二届国际足联世界杯足球赛

时间：2022 年 11 月 20 日到 12 月 18 日

地点：卡塔尔

上例中，触发词为“举行”，赛事名称、时间、地点等都是“体育赛事”事件的事件论元。根据事件信息是否预先定义，事件抽取可分为限定域事件抽取和开放域事件抽取两种类型，本节将针对这两种类型的事件抽进算法行介绍。

### 7.4.1 限定域事件抽取

限定域事件抽取需要预先定义事件类型以及与之对应的事件论元，抽取算法的目标就是从包含事件的文本中识别特定类型的事件并提取相应的事件论元。事件类型是标识事件的类别，常用的 ACE2005 数据集包括 8 种事件类型，33 种子类型，如“会议”、“袭击”等。事件触发词是指最清楚和明显地表达事件发生的主要词，如“击打”、“结婚”等。事件论元是指事件中涉及的参与者，一般为实体、时间等。论元角色（也称元素角色）是指事件论元在事件中所扮演的角色。

例如：句子“小明 2022 年在上海与小李举行婚礼”中，由事件触发词“举行婚礼”可以得到事件类型为“结婚”事件，人物“小明”和“小李”、时间“2022 年”、地点“上海”都为事件论元，对应“结婚”事件模板中的论元角色分别为“结婚的人”、“结婚时间”、“结婚地点”。

限定域事件抽取系统进行事件抽取，可以分解为事件类型识别、事件论元抽取等多个子任务进行，也可以采用联合抽取框架以减少错误传递。

### 1. 基于分类的事件抽取方法

文献 [367] 提出了将事件抽取任务分解成一系列的分类子任务的方法，并针对不同子任务提出了特征构建方法，在基于记忆的分类算法 TiMBL<sup>[368]</sup> 以及最大熵分类算法 MegaM<sup>[369]</sup> 上分别进行验证。该方法将事件抽取任务转换为以下四个子任务：

- (1) 触发词识别：从文本中识别触发词，并根据触发词确定事件提及（Event Mention）类型；
- (2) 论元识别：针对每个事件提及，从文本中识别事件提及的相关论元，包括实体、时间等；
- (3) 属性分配：确定每个事件提及的模态、极性、概括性和时态等属性；
- (4) 事件共指：确定从文本发现的事件提及是否为同一事件。

对于以上子任务，文献 [367] 采用流水线处理框架，其中触发词识别独立于其他任务，论元识别和属性分配任务依赖于触发词识别的结果，在这三个子任务都处理完成后，进行事件共指判断。本节中将主要介绍事件抽取中最关键的触发词识别和论元识别两个子任务。

针对事件触发词识别，由于在事件描述中，触发词往往由明显的单个词语组成，在 ACE2005 数据集中，就有超过 95% 的触发词都为单个单词，所以可以将触发词识别转换为词分类任务。但是，由于文本中词语的数量非常多，逐个分类会非常影响触发词的提取效率，并且这种情况下正负例比例相差过于巨大。因此，需要根据一些规则对词语进行预先过滤。由于触发词词性通常是由名词、动词、形容词组成，所以可以利用词性信息对词语进行预先筛选。使用词性标注算法获得输入文本中词语的词性信息后，再对相应词性的词进行分类。触发词识别主要由两个阶段组成：(1) 采用二分类分类器，将经过词性标注筛选后的词依次输入在训练集中训好的二分类分类器中，判断该词是否为触发词；(2) 采用多分类模型判断候选触发词的类型。

在事件论元识别方面，事件论元识别可以简化为一个成对分类（Pair Classification）任务，将包含事件描述的句子与同句中的事件论元内容组成待分类对，再利用分类模型判断论元角色。根据预先给定的事件定义，特定事件类型中有固定的论元角色，比如在“攻击”事件中包含攻击者、攻击目标、攻击时间、攻击发生地等几种固定论元角色。由于在进行事件论元识别已经通过触发词判别得到了事件类型，可以针对每个事件类型来训练不同的分类器来得到更好的效果。

文献 [367] 采用了基于特征的分类方法，针对上述分类任务，对输入内容利用人工设计的特征构建特向量表示，并利用有监督分类算法进行建模。在本任务中，利用了单词特征、WordNet 特征、上下文特征、依存句法特征、实体相关特征，针对不同的任务设计了不同的特征抽取策略。具体的特征描述可以参考文献 [367]。

## 2. 基于循环神经网络的联合事件抽取方法

采用流水线框架，将事件抽取分解为多个子任务的模式容易造成错误传递的问题，同时传统机器学习方法还需要依赖于预先设计好的语言工具来提取句子中的词汇和上下文特征。这种对预处理工具的依赖使得以往的事件抽取模型缺乏主动捕捉这些隐藏信息的能力，从而限制了这些模型的通用性。因此，研究员们进一步提出了基于神经网络的联合事件抽取方法。在联合事件抽取方法中，模型需要自主地标记事件触发词的位置，并对标记出的事件类别进行预测。这种改进有助于消除模型对预处理的过度依赖以及流水线架构的错误传递问题，并获取更加通用的事件抽取模型。

文献[370]中提出了通过卷积深度神经网络(CDNN)来提取词汇和上下文层次的特征。输入句子中的词编码，位置编码和事件类型后，模型能够自动提取词汇级和句子级特征来标记事件触发词的位置，在一定程度上解决了数据稀疏的问题。在标记了事件触发词后，分类器将事件触发词分类到具体的事件类别中。总的来说，模型采用两步结构，通过先识别再分类的方法来联合处理事件标记和事件分类。但这种流水线架构还是不能很好地缓解错误传播的问题，一旦对事件触发词的识别发生了错误或遗漏，就会大大影响后续的分类正确率。此外，这个方法没有尝试利用事件触发词和事件论元的依赖信息，也没有考虑到不同事件论元之间的相关性。

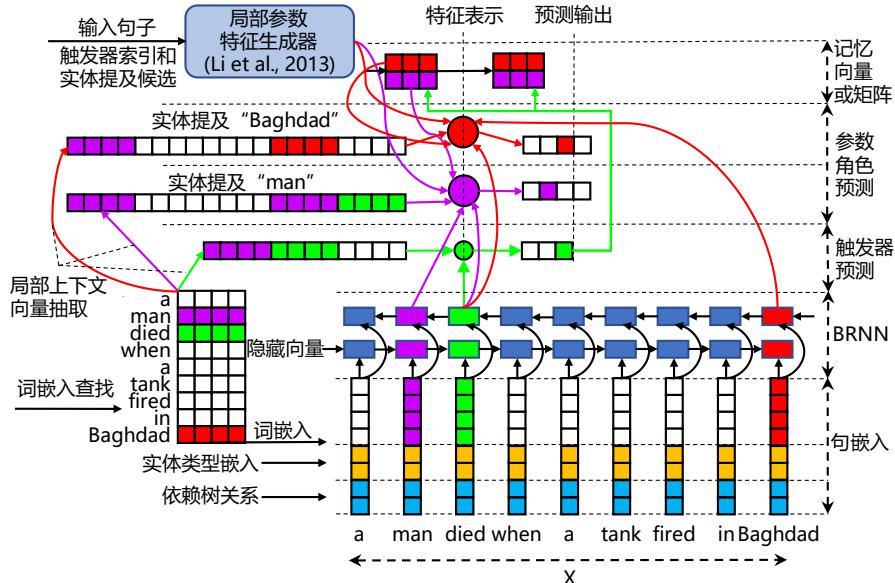
为了解决上述问题，文献[371]提出了JRNN方法，使用循环神经网络(RNN)来获取句子中不同事件触发词和事件论元间的长距离依赖关系。JRNN在保持对预处理的低依赖性的同时，提升了模型对不同事件之间重叠参数的识别和利用。模型使用两个RNN分别正/反向学习句子的表示，从而完成自动构建特征的工作；另一方面，JRNN使用了一个记忆向量与两个记忆矩阵来存储事件触发词与事件论元之间的依赖关系，引入了对离散的事件论元特征的长距离建模，有效解决了CDNN<sup>[370]</sup>方法存在的问题。模型的结构如图7.18所示，主要分为编码层和预测层两个模块。

编码层由句子编码层和基于RNN的特征编码层两部分构成，以图中的句子“*A man died when a tank fired in Baghdad*”为例，在句子编码层中，输入文本被转化成由三个向量的拼接而成的编码结果：

- (1) 词编码 (Word Embedding)：使用预训练的词嵌入表来获取每个词的向量表示
- (2) 实体类型编码 (Entity Type Embedding)：通过查找预训练的实体类型嵌入表，使用 BIO 注释模式来提供当前词的实体信息
- (3) 依存关系编码 (Dependency Relation Embedding)：使用训练得到的依存句法树获取某个词相对于其他词的依赖关系特征

在此基础上，使用双向RNN处理上述三个编码结果，首先词编码、实体编码和依赖关系编码拼接得到句子表示，再输出每个词对应的隐向量作为词信息和上下文信息的特征表示。形式化表示如下：

给定输入  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ，编码层包括两个方向相反的RNN网络层  $\overrightarrow{RNN}$  和  $\overleftarrow{RNN}$  后续

图 7.18 基于 RNN 的联合事件抽取模型<sup>[371]</sup>

的分类层的输入数据为编码后  $X$  的隐藏表示  $(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n)$ , 其中  $\mathbf{h}_i = [\alpha_i, \alpha'_i]$  满足

$$\begin{aligned} (\alpha, \alpha_2, \dots, \alpha_n) &= \overrightarrow{RNN}(x_1, x_2, \dots, x_n) \\ (\alpha'_1, \alpha'_2, \dots, \alpha'_n) &= \overleftarrow{RNN}(x_1, x_2, \dots, x_n) \end{aligned} \quad (7.89)$$

通过上述的编码过程, 输出向量被视作综合了词汇和上下文特征的句子表示。该表示随后被送入预测层来获得对触发词语、论元角色和事件分类的预测结果。

为了在模型的预测层中联合预测触发词和论元角色, JRNN 利用额外的记忆向量来编码触发词标签和论元角色之间的依赖关系: 使用二元记忆向量  $G_i^{trg}$  记录触发词之间的关联关系; 使用二元记忆矩阵  $G_i^{arg}$  来记录论元间的关联关系; 使用记忆矩阵  $G_i^{arg/trg}$  来记录触发词和论元之间的关联关系。矩阵中的每个论元表示行指标指代的实体与列指标指代的实体的相关程度, 这些记忆向量和矩阵被统一地初始化为 0, 并在训练过程中被不断更新。

根据事件抽取的任务需求, 预测层共分为三个部分, 分别为触发词预测层、论元预测层以及记忆向量和矩阵的存储区。

触发词预测(Trigger Prediction)过程在图中使用绿色的箭头表示。以识别句子中触发词“died”的

过程为例，在编码层中，模型最终使用三个向量的拼接来作为当前单词的编码表示  $\mathbf{R}_i^{trg}$ ：

$$\mathbf{R}_i^{trg} = [\mathbf{h}_i, \mathbf{L}_i^{trg}, \mathbf{G}_{i-1}^{trg}]$$

其中  $\mathbf{h}_i$  是上一步的隐藏向量（hidden vector）， $\mathbf{L}_i^{trg}$  是  $w_i$  的局部上下文，窗口大小为  $d$ ， $\mathbf{G}_{i-1}^{trg}$  是上一步的记忆向量（Memory Vector）。

触发词预测层接收上述的编码结果，并将其作为输入来获取词“died”的置信度  $P_{i;t}^{trg}$ 。

$$P_{i;t}^{trg} = P_i^{trg}(l = t) = F_t^{trg}(\mathbf{R}_i^{trg})$$

同时，对于高置信度的词，模型将词的预测结果输入记忆矩阵（memory matrices）来保存。这些保存的表示将在之后的论元预测中被提取以作为上下文信息表示。

论元预测过程在图中使用红色和紫色的箭头表示。模型对每个实体  $e_j$  预测标签  $a_{ij}$ ，即在  $w_i$  的触发词类型为  $t_i$  的情况下，实体  $e_j$  对于该触发词的论元角色  $a_{ij}$ 。图中以实体“man”为例来解释模型的预测过程。模型使用实体对  $[e_j, w_i]$  的编码作为输入来研究实体对于特定触发词的影响。根据图中紫色箭头的来源，具体地说，论元预测器的输入  $R_{ij}^{arg}$  由四个不同的分量构成，分别为：

- (1) 触发词  $w_i$  的隐向量  $\mathbf{h}_i$  和实体  $e_j$  的隐向量  $\mathbf{h}_{ij}$
- (2) 触发词  $w_i$  和实体  $e_j$  的上下文信息向量  $\mathbf{L}_{ij}^{arg}$
- (3) 在之前学习过程中被记录，表示触发词和实体间相互关系的二元特征向量  $\mathbf{V}_{ij}$  的隐向量  $\mathbf{B}_{ij}$
- (4) 实体  $e_j$  的记忆向量的输出  $\mathbf{G}_{i-1}^{arg}[j]$  和  $\mathbf{G}_{i-1}^{arg/trg}[j]$

预测器最终获取的输入向量  $\mathbf{R}_{ij}^{arg}$  可以表示为：

$$\mathbf{R}_{ij}^{arg} = [\mathbf{h}_i, \mathbf{h}_{ij}, \mathbf{L}_{ij}^{arg}, \mathbf{B}_{ij}, \mathbf{G}_{i-1}^{arg}[j], \mathbf{G}_{i-1}^{arg/trg}[j]]$$

对于预测器输出的结果向量，模型选取其中置信度最高的实体作为最终预测的论元实体。

最终，JRNN 模型结构的总体优化目标可以表示为：

$$\begin{aligned} C(T, A, X, E) &= -\log P(T, A | X, E) \\ &= -\log P(T | X, E) - \log P(A | T, X, E) \\ &= -\sum_{i=1}^n \log P_{i;t_i^*}^{trg} - \sum_{i=1}^n I(t_i \neq \text{other}) \sum_{j=1}^k \log P_{ij;a_{ij}^*}^{arg} \end{aligned} \tag{7.90}$$

这意味着模型的目标是尽可能准确地提取句子中的触发词和论元实体，并将每个触发词分配到该事件对应的两个正确论元上。

### 7.4.2 开放域事件抽取

开放域事件抽取（Open Domain Event Extraction）其目标是在没有任何预定义域假设的情况下，从非结构化文本中挖掘提取有意义的事件信息。与限定域事件抽取任务不同，在没有预先定义的事件类型以及对应的事件论元情况下，早期开放域事件抽取目标不是精确地提取事件要素，而是使用聚类、语义分割等方法，对文本内容进行分析基础上检测并跟踪事件。近年来，也有一些工作试图给出更细粒度信息，针对给定的一系列文本内容，输出事件集合，以及每个事件的触发词和对应的事件论元列表。在本节中，我们将介绍基于聚类的开放域事件抽取方法，以及基于神经隐变量的细粒度开放域事件抽取方法。

### 1. 基于聚类的开放域事件抽取方法

基于聚类的开放域事件抽取目标是从无结构文本中抽取若干主题的相关内容组成一系列事件，可以分为两个主要类型：回顾事件抽取（Retrospective Event Extraction）和在线事件抽取（Online Event Extraction）。回顾事件抽取是将语料库中的文本内容进行分组，每一组文本被视为一个事件。在线事件抽取是在回顾事件抽取的基础之上，对当前时刻给定的文本进行实时处理，判断当前文本是已有事件还是新事件。回顾抽取和在线抽取分别采用离线和在线的聚类算法完成。

文献 [372] 提出了一种事件检测和跟踪算法。使用基于词袋模型的传统向量空间模型对文档进行表示。每一篇文档（新闻报道）是通过一个带权重项的向量来表示，在聚类算法中是将所有文档的规范化向量表示用于聚类。文档向量是使用频率（TF）和逆文档频率（IDF）进行统计加权，并进行适当的标准化，同时只保留每个向量的前  $k$  项。文献 [372] 采用如下两种聚类算法完成回顾抽取和在线抽取：

- GAC (Group-Average Clustering) 多层次的聚类算法，用于回顾事件抽取。
- INCR (INcremental ClusteRing) 增量聚类算法，适用于回顾抽取和在线抽取。

GAC 聚类算法采用自底向上的贪心并结合分而治之的策略，其目标是最大化聚类结果中每对文档之间的平均相似度。GAC 算法的基本流程是：(1) 将文档集合中的每个文档都作为一个单独的事件类；(2) 将当前事件类簇集合中的事件类按顺序连续并且不重叠地划分到  $m$  个组中；(3) 每个组内部进行局部聚类，重复地合并组中的 2 个最相似的事件类，直到桶中类数量减少的比例达到预设的阈值  $p$ ，或者任意 2 个类之间的相似度值均低于预设阈值  $s$  为止；(4) 将组之间的边界去除，并重复步骤 (2)-(4)，直到最顶层的事件类数目达到了一个预定的数值为止。该算法的时间复杂度为  $O(mn)$ ，其中  $n$  是输入语料库中的文档数， $m$  是每个组的大小，并且  $m, n$  满足  $m \leq n$  条件。另外，还能够通过重聚类的方法来减少初始化组的系统偏差，从而能够生成一个更加紧凑的聚类分布。

INCR 聚类算法则依次处理输入文本，然后依次扩大聚类的集合。如果当前文档与某个事件类的中心之间的相似度高于预先选择的阈值，则新的文本会被加入到已经生成的相似度最高的事件类中；否则，该文档将被视为新事件类。除此之外，还引入“新事件”和“旧事件”标签，通过该标签可以判断当前新闻是否是该时间点的新型事件。通过调整阈值，也可以得到不同粒度级别的集合。还可以额外利用输入数据的动态特性和事件的时间属性这两项信息，以达到提升算法效果的目的。

对于在线抽取算法，时间窗口的引入能够限制前  $m$  个新闻事件文档。对于按照序列依次处理的现有文档，每个在时间窗口文档相似度得分都会被计算，如果在窗口中相似度得分低于一个预定的阈值，那么将一个“新事件”的标签赋予给这个新闻文档。该判定的可信度分数  $Score$  设置为如下：

$$Score(x) = 1 - \max_{d_i \in window} \{sim(x, d_i)\} \quad (7.91)$$

其中  $x$  是当前的文档， $d_i$  是窗口中的第  $i$  个文档， $i = 1, 2, \dots, m$ 。为了进一步更平滑的方式来使用时间临近信息，可以进一步弱化时间的权重：

$$Score(x) = 1 - \max_{d_i \in window} \left\{ \frac{i}{m} sim(x, d_i) \right\}. \quad (7.92)$$

这些窗口策略通过牺牲了少量的召回率来获得了精度的提高。 $i/m$  线性衰减时间窗口结果与均匀加权窗口相比，始终能够产生更好的结果。

在阈值检测部分，文献 [372] 所采用的方法主要是通过两个特定的阈值来进行控制。这两个阈值分别是聚类阈值  $t_c$  和新奇性阈值  $t_n$ ，前者决定聚类结果的颗粒度，对于回顾抽取非常重要；后者是决定算法敏感度和事件新奇性，而对于在线抽取来说是关键影响因素。当  $t_c \geq t_n$ ,  $sim_{max}(x) = 1 - score(x)$ ，在线抽取规则可以定义为：

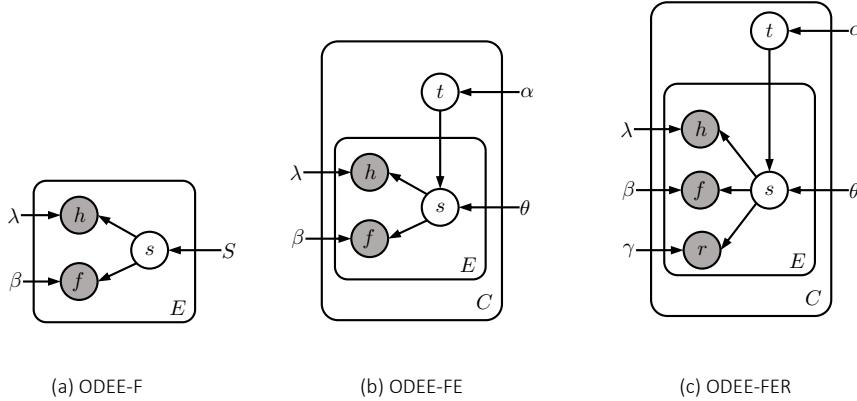
- 如果  $sim_{max}(x) > t_c$ ，那么设定标志为 OLC，然后把文档  $x$  添加到窗口中最相似的事件类中。
- 如果  $t_c \geq sim_{max}(x) > t_n$ ，然后将标志设定为 OLD，然后将文档  $x$  设定为新的独立事件类。
- 如果  $t_n > sim_{max}(x)$ ，然后设定标志为 NEW，然后将文档  $x$  设定为新的独立事件类。

## 2. 基于神经隐变量的开放域事件抽取方法

文献 [373] 提出了一种引入了神经网络隐变量的无监督生成模型 ODEE，进行新闻文本的事件抽取。输入为一个新闻集合（包含相同事件的报道），输出为一系列事件，每个事件都包含一个触发词和一个该事件模式的事件论元列表。该模型提取无约束的事件类型，并从新闻集合中归纳出通用的事件模式，每个新闻集合都有一个来自全局参数化正态分布的隐事件类型向量，以及实体的文本冗余特征。

针对开放域事件抽取任务，该方法提出了三个神经隐变量模型，它们的复杂程度依次增加，如图7.19所示。模型中  $S$  表示槽位数， $E$  表示实体数， $C$  表示新闻集合数， $V$  表示中心词汇量，灰色圆圈是可观测变量，白色圆圈是隐变量。

ODEE-F 模型如图7.19(a) 所示，给定一个语料库  $N$ ，从  $S$  个槽的均匀分布中为每个实体  $e$  采样一个槽  $s$ ，然后从多项分布中抽取一个中心词 (Head Word)  $h$ ；使用 ELMo 作为上下文编码器，得到连续特征向量  $f$  ( $f$  遵循多变量正态分布，其协方差矩阵是对角矩阵)。将  $f$  的  $S$  个不同正态分布的所有参数 (协方差矩阵的均值向量和对角向量) 标记为  $\beta \in \mathbb{R}^{S \times (2n)}$ ，其中  $n$  表示  $f$  的维

图 7.19 ODEE 方法三个神经隐变量模型结构图<sup>[373]</sup>

数，在逐行单纯形约束下槽分布概率矩阵  $\lambda \in \mathbb{R}^{S \times V}$  作为参数，其中  $V$  是中心词汇表大小。实体  $e$  的联合概率是：

$$p_{\lambda, \beta}(e) = p(s) \times p_{\lambda}(h | s) \times p_{\beta}(f | s) \quad (7.93)$$

ODEE-F 忽视了不同的事件可能有不同的槽分布，因此模型 ODEE-FE，如图7.19(b)，为每个新闻集从参数为  $\alpha$  的全局正态分布中抽样一个潜在事件类型向量  $t$ ，然后使用  $t$  和参数  $\theta$  的多层次感知器对相应的槽分布进行编码。新闻群  $c$  的联合概率是：

$$p_{\alpha, \beta, \theta, \lambda}(c) = p_{\alpha}(t) \times \prod_{e \in E_c} p_{\theta}(s | t) \times p_{\lambda}(h | s) \times p_{\beta}(f | s) \quad (7.94)$$

一个共指实体出现在新闻集合中的频率越高，它就越有可能是一个重要槽。除此之外，不同的新闻机构关注事件的不同方面，所以冗余的文本信息可以提供复杂的信息。因此，在模型 ODEE-FER 中，如图7.19(c) 所示，额外引入共指槽的归一化出现频率  $r$  作为观察到的隐变量。通常，一个新闻集合接收一个潜在事件类型向量  $t$ ，其中每个实体  $e \in E_c$  接收一个槽类型  $s$ 。具有中心词、冗余上下文特征和潜在事件类型的新闻集合的联合分布是：

$$p_{\alpha, \beta, \gamma, \theta, \lambda}(c) = p_{\alpha}(t) \times \prod_{e \in E_c} p_{\theta}(s | t) \times p_{\lambda}(h | s) \times p_{\beta}(f | s) \times p_{\gamma}(r | s) \quad (7.95)$$

在推理部分，考虑模型 ODEE-FER 处理的两个任务：(1) 学习参数；(2) 在给定新闻集  $c$  的情况下执行推理以获得潜在变量  $s$  和  $t$  的后验分布。为简单起见，在模型 ODEE-FER 中将  $f$  和  $r$  连接起来作为新的观测特征向量  $f'$ ，并将它们的参数合并为  $\beta' \in \mathbb{R}^{S \times (2n+2)}$ 。将离散的潜变量  $s$  消

去，获得对数似然的证据下界（Evidence Lower BOund, ELBO）：

$$\begin{aligned}
 \log p_{\alpha, \beta', \theta, \lambda}(c) &= \log \int_t \left[ \prod_{e \in E_c} p_{\lambda, \theta}(h | t) p_{\beta', \theta}(f' | t) \right] p_{\alpha}(t) dt \\
 &\geq \text{ELBO}_c(\alpha, \beta', \theta, \lambda, \omega) \\
 &= \mathbb{E}_{q_{\omega}(t)} \log p_{\beta', \theta, \lambda}(c | t) - D_{\text{KL}}[q_{\omega}(t) \| p_{\alpha}(t)]
 \end{aligned} \tag{7.96}$$

其中  $D_{\text{KL}}[q_{\omega} \| p_{\alpha}]$  为 KL 散度。由于计算两个分布的 KL 散度非常困难，并且正态分布存在简单有效的重参数化技巧，因此选择  $q_{\omega}(t)$  作为由  $w$  参数化的正态分布，由神经推理网络学习，具体过程参见文献 [373]。

通过最大化下面的似然公式选择每个实体的槽：

$$\begin{aligned}
 p_{\beta', \theta, \lambda}(s | e, t) &\propto p_{\beta', \theta, \lambda}(s, h, f', t) \\
 &= p_{\theta}(s | t) \times p_{\lambda}(h | s) \times p_{\beta'}(f' | s)
 \end{aligned} \tag{7.97}$$

最终新闻集合  $c$  组合成事件进行最终输出，还需要找到每个实体对应槽值的谓词。ODEE-FER 使用 Stanford Dependency Parser<sup>[374]</sup> 生成的词性标签和句法解析树，根据如下规则提取每个实体提及的中心词的谓词：(1) 如果一个中心词的支配者是 VB；或者 (2) 如果一个中心词的支配者是 NN 并且属于 WordNet 的 noun.ACT 或 noun.EVENT 范畴，那么该词为谓词。将相同共指链的实体提及的谓词合并为一个谓词集，对于集合中的每个谓词  $v$ ，找到其谓词集合包含  $v$  的实体，将这些实体视为由  $v$  触发的事件的论元。最终，对论元进行排序得到前  $N$  个开放域事件做为最终输出。

图7.20给出了使用 ODEE-FER 进行开放事件抽取的示例。通过对比图7.20左边的文本内容和右侧所给出的事件信息，可以看到新闻被归纳为三个事件：“raise”、“report”和“predict”。不同的事件还产生了不同的槽位内容。

### 7.4.3 事件抽取评价方法

事件抽取的评价指标也采用统计机器学习算法评测中常用的指标进行对比，主要为精确率(P)、召回率(R)、F 值，具体的计算方法如下：

$$\text{精确率 (P)} = \frac{\text{正确抽取结果数}}{\text{抽取结果总数}} \times 100\% \tag{7.98}$$

$$\text{召回率 (R)} = \frac{\text{正确抽取结果数}}{\text{需抽取结果总数}} \times 100\% \tag{7.99}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{7.100}$$

<b>DOC 1</b>
2018-10-16 07:00:03
UnitedHealth shares rise after posting a 28% rise in third-quarter profit, raises 2018 forecast
UnitedHealth, the largest U.S. health insurer, reported better-than-expected third-quarter earnings and revenue on Tuesday
<b>DOC 2</b>
2018-10-16 00:00:00
UnitedHealth's 2018 so far: Three quarters, three boosts to outlook
<b>DOC 3</b>
2018-10-17 00:32:09
UnitedHealth Group predicts Medicare growth. The comments came as the insurer beat profit expectations for Q3.
<b>DOC 4</b>
2018-10-16 10:53:06
UnitedHealth beats all around in 3Q, raises outlook again
MINNEAPOLIS (AP) — UnitedHealth reported better-than-expected profits and revenue for the third quarter and the company raised its outlook yet again on strong trends in the insurance business.

## 事件 1 :

Tigger	raise
Agent	UnitedHealth, UnitedHealth shares
Patient	2018 forecast, better-than-expected profits, the insurance business
Time	the third quarter
Variation	28%

## 事件 2 :

Tigger	report
Agent	UnitedHealth Group, the largest U.S. health insurer
Patient	better-than-expected third-quarter earnings
Time	Tuesday

## 事件 3 :

Tigger	predict
Agent	UnitedHealth Group
Patient	Medicare growth

图 7.20 ODEE 生成的事件框架示例<sup>[373]</sup>

对于自动抽取系统或将事件抽取作为信息处理流水线的一部分时，应尽量提高 F1 指标，以降低抽取错误造成后续步骤的错误累积；在有人工干预的事件抽取系统中，应在保证一定 F1 指标的基础上，尽量提升召回率指标，以尽量确保抽取时不遗漏。

在评测基于流水线的事件抽取模型或系统时，有时还会使用上述指标对事件抽取中的各子任务分别进行评价，例如在一些论文中一般会同时汇报 TI (Trigger Identification, 触发词识别)、TC (Trigger Classification, 触发词分类，即事件类型分类)、AI (Argument Identification, 论元识别)、AC (Argument Classification, 论元分类，即论元角色分类) 四个子任务的精确率、召回率以及 F1 值。

### 7.4.4 事件抽取语料库

事件抽取研究的发展同样离不开事件抽取相关评测集合和语料的不断推出。从最早的 MUC 语料库，再到目前使用最为广泛 ACE 语料库，以及中文 CEC 语料库，都极大的推动了事件抽取任务的不断进步。表7.4给出了常见事件抽取语料库的汇总。

语料库名称	事件类型数目	语言
ACE 事件语料库	8	中文、英文、阿拉伯文
MUC 语料库	4	英文
TDT 语料库	25	英文
CEC 语料库	5	中文

表 7.4 常见事件抽取语料库汇总

#### 1. ACE 事件抽取语料库

ACE 事件语料库是目前事件抽取中最广泛使用的数据集之一，包含的事件具有复杂的结构和参数，涉及实体，时间和值。ACE 2005 事件语料库定义了 8 个事件类型和 33 子类型，每个事件子类型对应于一组参数角色。所有事件子类型共有 36 个参数角色，含中文、英文、阿拉伯语三种语言的语料。

#### 2. MUC 事件抽取语料库

MUC 是最早产生支持事件共指任务的语料库。数据语料主要来自新闻语料，限定领域为飞机失事报道和航天器发射事件报道。MUC 评测中心围绕一个“场景”，根据关键事件类型和与它相关的各种角色定义。

#### 3. TDT 事件抽取语料库

TDT 语料库来自于美国政府支持的 Topic Detection and Tracking 科研项目，其主要包含一个连续新闻流中的大量新闻，并对其进行细分。整个语料库的所有标签都是人工手动标记的。它是由 15836 个发生在 1994 年 7 月 1 号到 1995 年 6 月 30 号之间的新闻事件组成的语料库。一半的新闻来自于路透社杂志，另一半来自于 CNN 多个广播新闻项目。整个语料库包含了 25 种事件，事件的定义仅仅给出的指示信息是发生的具体位置与具体时间。这些信息能够有效的将不同的事件区分开来。

#### 4. CEC 中文事件抽取语料库

中文事件语料库（Chinese Event Corpus, CEC）包含 CEC-1 和 CEC-2 两个语料库包。其中从互联网上收集了 5 类（地震、火灾、交通事故、恐怖袭击和食物中毒）突发事件的新闻报道作为生语料，然后再对生语料进行文本预处理、文本分析、事件标注以及一致性检查等处理，最后将标注结果保存到语料库中，CEC 合计 332 篇。与 ACE 和 TimeBank 语料库相比，CEC 语料库的规

模虽然偏小，但是对事件和事件要素的标注却最为全面。

## 7.5 延伸阅读

尽管基于深度学习的信息抽取已经取得了显著的进步，但实际应用中场景的多元化使得信息抽取依然面临着诸多挑战，低资源、开放领域下模型的抽取能力、如何融入视觉、听觉等多种模态来进一步提升信息抽取性能，以及如何改进框架将不同子任务统一建模，这些都值得进一步的探索。

(1) 高效的小样本学习能力。真实场景下的训练数据是十分有限的，这就要求模型具备从少量的样本中学习到实体、关系、事件的特征。目前在信息抽取任务中，常使用度量学习<sup>[375–377]</sup>、元学习<sup>[378–380]</sup>、迁移学习<sup>[381, 382]</sup>、融合领域知识<sup>[383–385]</sup>等方式来提升模型从小样本中抽取信息的能力。随着基于 GPT3 等在内的超大规模预训练模型的 Prompt 学习范式受到研究者广泛关注，借助大规模预训练语言模型中蕴含的大量知识，仅利用几条或几十条样本作为训练集，在命名实体<sup>[327, 386, 387]</sup>、关系抽取<sup>[388–390]</sup>等任务上，基于 Prompt 的方法也取得了不错的效果。

(2) 多模态信息融合。目前信息抽取主要针对的是纯文本数据，而常见的文档具有多样的布局且包含丰富的信息。此外很多文档、网页、社交媒体也是多以富文本的形式呈现，其中也包含大量的多模态信息。我们在 2018 年针对社交媒体多模态的命名实体<sup>[391]</sup>论文也指出，在很多情况下仅依赖文本内容，无法完成准确的信息抽取任务。针对多模态信息抽取，研究人员从基于图对齐与图融合<sup>[392–394]</sup>、图卷积<sup>[395, 396]</sup>、结构化和半结构化页面结构<sup>[397, 398]</sup>等方面进行了一些列研究。如何利用视觉、听觉、以及富文档信息，通过多模态信息补全文本中的缺失，是信息抽取的重要发展方向之一。

(3) 子任务统一建模。在本章中介绍的命名实体识别、关系抽取以及事件抽取任务都采用了不同类型的算法，关系抽取和事件抽取任务本身也会采用流水线方式组合多种算法。但是，随着“预训练 + 大规模多任务学习”这一范式所展现的学习能力，使得构建统一框架建模多种信息抽取任务成为可能。文献 [399] 即针对信息抽取设计了一种结构化抽取语言 (Structural Extraction Language, SEL)，采用 Seq2Seq 的生成式框架，并将命名实体识别、关系抽取、事件抽取这三个信息抽取任务的不同结构进行统一描述，使得模型针对不同任务输出一致的结构，实现了面向信息抽取的统一文本到结构生成框架 UIE (Universal Information Extraction)。文献 [400] 提出了基于 Prompt 的生成式方法，并构建了信息抽取任务容易框架。

## 7.6 习题

- (1) 命名实体识别中有哪些解码方式？如何解决嵌套实体问题？
- (2) 远程监督是关系抽取任务中自动标注训练数据的有效策略，但其过强的设定会产生错误标注，可以从哪些角度考虑来缓解远程监督引入的噪声问题？
- (3) 试比较流水线式的关系抽取和联合关系抽取的优缺点。

- (4) 限定域事件抽取的基本事件结构?
- (5) 信息抽取目前还面临哪些挑战? 如何解决开放域下的关系抽取问题?

## 8. 机器翻译

根据联合国统计，目前世界上正在使用的语言约有 6000 种，教育系统和公共领域中使用到的语言也有数百种之多。我们不可能掌握如此之多的语言，即便是熟练使用联合国规定的包括阿拉伯语、汉语、英语、法语、俄语和西班牙语在内的六种正式语言都非常困难。因此，第一台通用计算机 ENIAC 才刚刚面世，Warren Weaver 就在 1947 年提出了利用计算机翻译人类语言的可能。机器翻译（Machine Translation）是指利用计算机将一种语言（源语言）自动翻译为另外一种语言（目标语言）的过程。机器翻译是自然语言处理中研究历史最长也最重要的任务之一。

本章首先介绍机器翻译的基本概念和常见任务，在此基础上介绍基于统计的机器翻译方法和基于神经网络的机器翻译方法，最后介绍机器翻译的评测方法和机器翻译常见的语料库和评测集合。

### 8.1 机器翻译概述

机器翻译（Machine Translation, MT）这一概念拥有很长的历史，相关领域的研究最早可以追溯到 17 世纪。1629 年 Descartes 等人就提出使用统一符号表达不同语言中的同一概念的语义。现代机器翻译的研究始于上世纪五十年代，Bar-Hillel 等人在 1951 年就开始了对机器翻译的研究，并在 1952 年组织了第一届国际机器翻译会议（International Conference on Machine Translation）。机器翻译的任务定义相对简单，目标就是通过计算机将源语言（Source Language）翻译为目标语言（Target Language）。如图 8.1 所示，机器翻译系统将中文翻译为英文，在这个例子中源语言为中文，目标语言为英文。

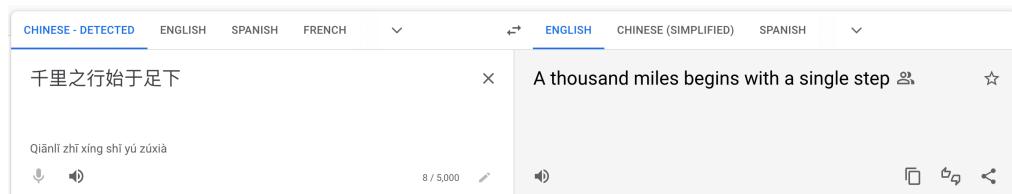


图 8.1 机器翻译系统样例（来源：谷歌翻译）

### 8.1.1 机器翻译发展历程

机器翻译的发展历程基本代表了自然语言处理领域的发展过程，迄今为止，机器翻译的研究与发展大体上经历了三次主要的浪潮：基于规则的机器翻译<sup>[401]</sup>、基于统计的机器翻译<sup>[402]</sup>以及基于神经网络的机器翻译<sup>[403]</sup>。

**基于规则的机器翻译：**基于规则的机器翻译是机器翻译任务的第一套解决方案，它基于“每一种语义在不同的语言当中都存在与其相对应的符号”这一假设。对于某种语言中的大多数单词而言，通常都能够在另一种语言当中找到表达相同含义的对应的单词。在这类方法当中，翻译过程通常被看作一个源语言的词替换过程。之所以被称为“基于规则的方法”，是因为同一种语义在不同的语言当中通常会以不同的词序去表达，词替换过程相对应地需要两种语言的句法规则作为指导。源语言中的每一个单词需要被放置在目标语言中相对应的位置。基于规则的机器翻译方法的理论非常简洁清晰，但在实践中的性能却不尽如人意。这是由于选择与给定源语言相适配的句法规则在计算上非常低效。同时，为了应对多样的语言现象，语言学家们设计了规模庞大的句法规则，这些规则很难被有效地组织，甚至会出现不同规则相互矛盾的情况。基于规则的方法最严重的缺陷在于其缺乏翻译过程中对上下文信息的建模，这使得基于规则的翻译模型的鲁棒性不佳。Marvin Minsky 在 1966 年给出了一个非常著名的例子来阐述这一问题：

“*The pen is in the box*”

“*The box is in the pen*”

上述两个句子具有相同的句法结构，第一个句子很容易理解。但第二个句子非常令人困惑，由于在英文中“pen”本身是一个多义词，除了“笔”之外它还有“栅栏”的意思。但对于计算机来说很难直接将“pen”对应到“栅栏”的意义上，进而产生错误的翻译结果。

**基于统计的机器翻译：**在过去的 20 年以来，统计机器翻译(Statistical Machine Translation, SMT)已经成为机器翻译领域的主流方法，并在工业界得到了广泛的实际应用。与基于规则的机器翻译方法不同，统计机器翻译完全从数据驱动的角度建模机器翻译任务。具体来说，通过对双语语料库的统计找到表达相同含义的单词或短语。给定一个源语言句子，统计机器翻译首先将其分割成若干个子句，接下来每个部分可以被目标语言的单词或短语替代。统计机器翻译中最主流的方法是基于词的统计机器翻译（Word-based MT）以及基于短语的统计机器翻译（Phrase-based SMT），总体包含预处理、句子对齐、词对齐、短语抽取、短语特征准备、语言模型训练等步骤。基于短语的统计机器翻译方法的可以将源语言和目标语言中的短语配对。在这一过程中，翻译模型能够利用短语内部的上下文信息，因而优于简单的单词到单词的翻译方法。

**基于神经网络的机器翻译：**神经网络方法在机器翻译任务上的应用可以追溯到上世纪八九十年代<sup>[404, 405]</sup>。但受限于当时的计算资源和数据规模的限制，神经网络方法的性能差强人意，故而其发展停滞了很多年。近年来，伴随着深度学习技术的广泛应用，越来越多的自然语言处理任务实现了极大的性能提升。基于深度神经网络的机器翻译模型也逐渐受到了许多关注。神经网络方法在机器翻译任务上的第一次成功应用是 Kalchbrenner 和 Blunsom 等人提出的基于递归神经网络

的方法<sup>[406]</sup>，这在当时的机器翻译领域是一种全新的概念。相比于其他模型，神经机器翻译模型在对语言学知识的依赖更少的前提下达到与之前方法相媲美的性能。从这之后，神经机器翻译方法正式走上了历史舞台，大量的研究者开始在这类方法上进一步研究和改进。到今天为止，神经机器翻译被广泛地应用在不同的工业场景下<sup>[407]</sup>。

### 8.1.2 机器翻译现状与挑战

机器翻译在经历了几十年的发展后，特别是深度神经网络有效应用于机器翻译，使得模型机器翻译的效果有了很大的提高，在特定条件下机器翻译的效果已经能够达到非常好的效果，甚至可以接近人工翻译效果。然而，在开放环境中，翻译效果还远没有达到直接使用的程度。根据机器翻译权威评测 WMT21<sup>[408]</sup> 给出的人工评测结果，在新闻领域最好的中文到英文翻译系统评分也仅有 75 分左右（满分 100 分）。机器翻译完全代替人工翻译还有很长的道路。以王佐良先生对 Samuel Ullman 所著的《Youth》译文为例：

**原文：** Youth is not a time of life; it is a state of mind; it is not a matter of rosy cheeks, red lips and supple knees; it is a matter of the will, a quality of the imagination, a vigor of the emotions; it is the freshness of the deep springs of life.

**机器翻译结果：** 青春不是生命的时光；这是一种心态；这不是红润的脸颊、红润的嘴唇和柔软的膝盖；这是意志的问题，是想象力的质量，是情感的活力；它是生命深泉的清新。

**王佐良译文：** 青春不是年华，而是心境；青春不是桃面、丹唇、柔膝，而是深沉的意志，恢宏的想象，炙热的感情；青春是生命的深泉在涌流。

可以看到虽然机器翻译的结果从词语到语法都已经达到了很好的效果，甚至一些相对不常见的句式结构也能较好的翻译。但是整个翻译的效果距离人工翻译“信达雅”的要求还是有很大的差距。上例中，虽然每个翻译后的句子都符合语法，但是句子之间的意义连贯性以及词语的搭配还是会让人难以理解。当然，这里给出的例子是相对困难的，绝大部分人工翻译也很难达到这种程度。对于新闻的翻译，相较于上述例子来说就简单很多。与人工翻译相比，机器翻译对于互联网上每天数以亿计的新闻和短消息的处理就体现出了巨大的优势。

机器翻译虽然经过多年的发展，目前在特定应用场景下已经能够有很好的效果，但是仍然面临如下挑战：

**(1) 自然语言复杂度高。**自然语言具有高度的复杂性、概括性以及多变性，并且是在不断发展的过程中。虽然目前已经有深度神经网络模型参数量达到了 1.75 万亿，但是相比于自然语言的复杂度来说还是相差很多。在第 6 章语言模型中介绍过，按照《现代汉语词典（第七版）》包含 7 万词条，句子长度按照 20 个词计算，参数量达到  $7.9792 \times 10^{96}$  的天文数字。更不要说，语言还是在动态发展中的，在 2018 年以前 Bert 代表的是《芝麻街》中的卡通布偶“伯特”，而现在计算机领域多是指代预训练语言模型 BERT。在数据驱动的统计机器翻译和神经机器翻译模型中，如何才能

让模型具备超大规模参数空间的建模能力和持续学习能力都是十分巨大的挑战。

**(2) 翻译结果不可解释。**目前机器翻译算法多采用数据驱动的方法，所采用的模型通常不具备可解释性。这就造成了机器翻译算法虽然给出了翻译结果，并且效果可能还很好，但是其对语言的理解和翻译过程与人的理解和翻译过程完全不同。机器翻译算法的目标仅是根据人工定义的目标函数进行优化，使得人们无法理解机器翻译算法的过程，不能够对算法进行解释。这就带来一系列的严重问题，包括我们不知道机器翻译算法在什么时候会出错，会出现什么样的错误，为什么会出现错误，如何才能修正和改进等等。这也使得我们很难在关键需求中完全依赖机器翻译。试想一下，对于一个我们完全不认识的语言“БЕЖАТЬ ВПЕРЕД”（俄语），机器翻译给出的结果是“向前跑”，我们无法确认模型结果正确与否。因此，如何构建具备可解释性的机器翻译模型也是需要解决的难点。

**(3) 翻译结果评测困难。**语言有很大的灵活性和多样性，同样一句话可以有非常多的翻译方法。对机器翻译性能进行评测可以采用人工评测和半自动评测方法。人工评测虽然是相对准确的一种方式，但是其成本高昂，根据艾伦人工智能研究院（AI2）GENIE 人工评测榜单给出的数据，针对 800 条机器翻译效果进行评测需要花费约 80 美元<sup>[409]</sup>。如果采用半自动评测方法，利用人工给定的标注翻译结果和评测函数可以快速高效的给出评测结果，但是目前半自动评测结果与人工评测的一致性还亟待提升。对于用词差别很大，但是语义相同的句子的判断本身也是自然语言处理领域的难题。如何有效地评测翻译结果是机器翻译任务所面临的挑战。

## 8.2 基于统计的机器翻译方法

机器翻译任务到今天已经经历了长达数十年的发展历史。在很长一段时间之内，基于统计的机器翻译方法在学术界受到了许多关注并在工业界得到了广泛地应用。尽管这些模型在当下已经不再是人们关注的焦点，回顾这些方法仍然对我们完整地了解机器翻译的过去以及展望未来的研究具有十分重要的意义。本节将以在统计机器翻译中具有里程碑式意义的 IBM 模型为例，简要阐述统计机器翻译方法的基础概念、基本假设和流程。

### 8.2.1 任务定义与基本问题

IBM 机器翻译模型构建在噪声信道模型（Noise Channel Model）的基础之上，其模型基本框架如图8.2所示。源语言句子  $s$  和候选的目标语言句子  $t_i$  通过一个噪声信道相连，在已知  $s$  和噪音信道性质的前提下，就能够得到信源，也即目标语言的分布  $P(t|s)$ 。机器翻译的过程本质上就是在给定源语言句子  $s$  的前提下，从分布  $P(t|s)$  找到最有可能的目标语言  $t$  作为输出，这一搜索过程也被称为解码。

$$\hat{t} = \arg \max_t P(t|s) \quad (8.1)$$

在上述噪声信道基础框架下，统计机器翻译任务需要解决如下三个核心的基本问题：

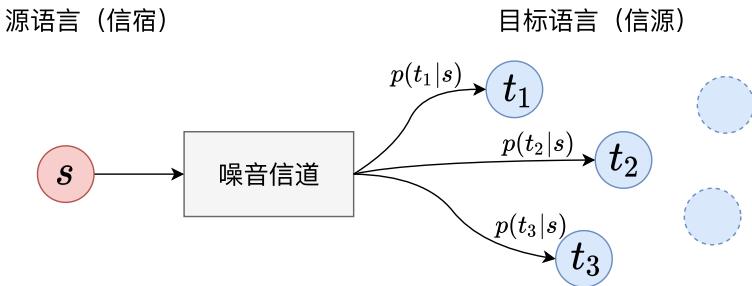


图 8.2 噪声信道模型

**问题 1：建模** 如何通过易于被计算机处理的数学模型对  $P(t|s)$  进行合理地建模以刻画源语言和目标语言之间的关系。

**问题 2：训练** 如何从给定的平行语料库（即源语言-目标语言对组成的语料集合）中获得最优的模型参数。

**问题 3：解码** 如何从模型  $P(t|s)$  中搜索出最优的目标语言序列  $t$ 。

接下来，将依次介绍 IBM 模型是如何建模并解决上述三个问题的。首先，统计机器翻译模型的核心在于对  $P(t|s)$  的定义，这一定义决定了模型性能的上限并且也是后续训练和解码的基础。IBM 模型通过贝叶斯公式对这一翻译概率做如下变换：

$$P(t|s) = \frac{P(s,t)}{P(s)} = \frac{P(s|t)P(t)}{P(s)} \quad (8.2)$$

通过上述变换，翻译模型  $P(t|s)$  被分解为了三个部分：(1) 从目标语言指向源语言的翻译概率  $P(s|t)$ ；(2) 目标语言的语言模型  $P(t)$ ；(3) 源语言序列语言模型  $P(s)$ 。需要注意的是，通过贝叶斯变换  $P(t|s)$  和  $P(s|t)$  只是翻译的方向不同，建模难度并没有下降。其核心是为了引入目标语言的语言模型。这是由于 IBM 模型本质上是一种基于词的统计机器翻译模型，仅通过翻译概率  $P(s|t)$  很难有效地建模目标语言单词之间的相对位置关系，也即目标语言序列的流畅程度。相对应地，引入语言模型  $P(t)$  可以有效地缓解上述问题。此外， $P(s)$  是一个不变量，因此它不会影响到  $\frac{P(s|t)P(t)}{P(s)}$  的相对大小，也就不会影响到最终的解码过程，在建模的过程当中  $P(s)$  通常不需要被计算，可以省略：

$$\hat{t} = \arg \max_t P(t|s) = \arg \max_t \frac{P(s|t)P(t)}{P(s)} = \arg \max_t P(s|t)P(t) \quad (8.3)$$

基于上述分析，IBM 模型的建模问题转换为如何建模翻译概率  $P(s|t)$  以及语言模型  $P(t)$ 。翻译概率  $P(s|t)$  主要用于衡量源语言和目标语言之间的匹配程度。然而，自然语言拥有极其庞大的潜在的组合方式。假设某种语言对应的词表大小为 10000，那么一个简单的长度为 10 的句子就对

应着  $10000^{10} = 10^{40}$  种不同的组合方式。基于任何已有的平行语料库，直接在句子层级对上述翻译概率进行估计，都会面临严重的数据稀缺问题。因此，IBM 模型将句子层级的翻译概率进一步拆解为单词级别的对应关系的组合，从而缓解上述数据稀疏的问题，这一拆解过程又被称为词对齐。

词对齐作为 IBM 模型构建的重要基础之一，描述了目标语言和源语言之间单词级别的对应关系。以图8.3中的对齐实例 1 为例，给定源语言文本“机器 翻译”，对应的目标语言翻译为“Machine Translation”。其中，“机器”一词对应“Machine”而“翻译”一词对应“Translation”。使用记号  $\mathbf{a} = \{a_1, \dots, a_m\}$  表示这种对应关系，其中  $a_j$  表示源语言中的单词  $s_j$  和目标语言中的单词  $t_{a_j}$  存在对应关系。举例来说，在对齐实例 1 中， $a_1 = 1, a_2 = 2$ 。

为了建模方便，IBM 模型对词对齐做了如下两个限制：

- 对于每一个源语言单词，至多只能对齐到一个目标语言单词上。图8.3的对齐实例 2 中，源语言单词“机器”同时对应到了两个目标语言单词“Machine”和“Translation”，这就违反了上述 IBM 模型假设。而其余的对齐实例均满足这一假设。
- 存在一些源语言单词，它们可以对齐到一个额外增设的虚拟目标语言单词“Null”上，也即对空。图8.3的对齐实例 4 中的“机器”一词就对应到了目标语言的“Null”上。对空情况的额外考虑并不是没有意义的。事实上，对空的现象在翻译的过程当中频繁出现，如虚词的翻译。

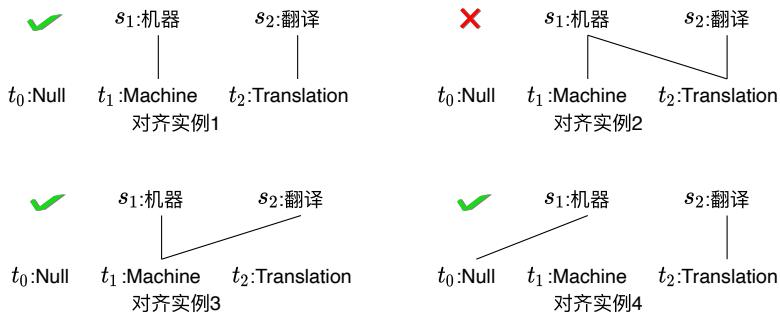


图 8.3 IBM 词对齐实例

IBM 模型认为句子级别的翻译概率可以通过单词级别的翻译概率组合而成，并将词对齐作为一种隐变量整合到翻译概率的建模过程中。这样，原本较为困难的句子级别的建模问题就被分解为一个分步学习的问题：

$$P(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} P(\mathbf{s}, \mathbf{a}|\mathbf{t}) \quad (8.4)$$

其中， $\mathbf{s} = \{s_1, s_2, \dots, s_m\}$  表示一个长度为  $m$  的源语言序列， $\mathbf{t} = \{t_1, t_2, \dots, t_l\}$  表示一个长度为  $l$  的目标语言序列， $\mathbf{a} = \{a_1, a_2, \dots, a_m\}$  表示源语言中每一个单词  $s_j$  对应的目标语言单词序号  $a_j$ 。直接建模  $P(\mathbf{s}|\mathbf{t})$  仍然非常复杂，为了解决这个问题，IBM 模型对上述概率通过链式法则做了进一步

展开，并为后续的简化做了准备。

$$P(\mathbf{s}, \mathbf{a} | \mathbf{t}) = P(m | \mathbf{t}) \prod_{j=1}^m P(a_j | \mathbf{a}_1^{j-1}, \mathbf{s}_1^{j-1}, m, \mathbf{t}) P(s_j | \mathbf{a}_1^j, \mathbf{s}_1^{j-1}, m, \mathbf{t}) \quad (8.5)$$

其中， $\mathbf{a}_1^{j-1}$  表示源语言序列中前  $j-1$  个单词的词对齐， $\mathbf{s}_1^{j-1}$  表示源语言序列中的前  $j-1$  个单词。这一展开看似较为复杂，实际上每个部分都具有较为清晰的物理含义。给定一个目标语言序列  $\mathbf{t}$ ，我们首先通过概率  $P(m | \mathbf{t})$  估计源语言序列的长度  $m$ 。接下来，通过  $m$  次循环从左向右依次生成源语言序列和它们的词对齐。在第  $j$  次循环当中，首先通过目标语言序列  $\mathbf{t}$ ，前  $j-1$  次循环中生成的词对齐序列  $\mathbf{a}_1^{j-1}$  以及源语言序列  $\mathbf{s}_1^{j-1}$  产生当前位置的词对齐  $a_j$ ，即  $P(a_j | \mathbf{a}_1^{j-1}, \mathbf{s}_1^{j-1}, m, \mathbf{t})$ 。接下来结合  $a_j$  进一步生成当前位置的源语言单词  $s_j$ ，也即  $P(s_j | \mathbf{a}_1^j, \mathbf{s}_1^{j-1}, m, \mathbf{t})$ 。至此，翻译概率的建模实际上就被转换为源语言文本和词对齐的生成问题。但是，仍然还存在两个迫切需要解决的问题：

- (1) 为了最终实现对翻译概率  $P(\mathbf{s} | \mathbf{t})$  的建模，在公式8.4中需要对所有可能的词对齐进行求和。然而，可能的词对齐的数量随着源语言序列的长度呈指数级别增长，如何计算这一求和式是第一个需要被解决的问题。
- (2) 公式8.5通过链式分解为建模  $P(\mathbf{s}, \mathbf{a} | \mathbf{t})$  提供了一种可行的方向，然而如何通过目标语言序列估计源语言序列的长度  $P(m | \mathbf{t})$ ，以及如何建模源语言  $P(s_j | \mathbf{a}_1^j, \mathbf{s}_1^{j-1}, m, \mathbf{t})$  和词对齐的生成过程  $P(a_j | \mathbf{a}_1^{j-1}, \mathbf{s}_1^{j-1}, m, \mathbf{t})$  尚待解决。

对于上述两个问题的解决实际上对应着五个不同的 IBM 模型，我们将在后续的章节中详细阐述它们是如何简化并解决，以及如何基于构建的翻译模型从平行语料库当中学习最优的参数。除了翻译概率  $P(\mathbf{s} | \mathbf{t})$  之外，语言模型  $P(\mathbf{t})$  是 IBM 模型的另外一个重要组成部分，对于语言模型的详细介绍可以参考本书的第六章。在这里，我们仅给出一个简略的回顾。

为了衡量生成译文的流畅程度，IBM 模型引入了  $n$ -gram 语言模型。它使用概率化的方法描述了句子的生成过程。以 2-gram 语言模型为例，一个目标语言序列的生成概率可以按照下式评估：

$$\begin{aligned} P_{\text{lm}}(\mathbf{t}) &= P_{\text{lm}}(t_1, \dots, t_l) \\ &= P(t_1) \times P(t_2 | t_1) \times P(t_3 | t_2) \times \dots \times P(t_l | t_{l-1}) \end{aligned} \quad (8.6)$$

上述的每一个  $P(t_i | t_{i-1})$  均对应着语言模型的不同参数。这些参数的估计可以通过对语料库的极大似然估计完成。

$$P(t_i | t_{i-1}) = \frac{\text{单词对 } < t_{i-1}, t_i > \text{出现的次数}}{\text{单词 } t_{i-1} \text{ 出现的次数}} \quad (8.7)$$

在完成了对翻译概率  $P(\mathbf{s} | \mathbf{t})$  以及语言模型  $P(\mathbf{t})$  的建模与优化之后，下一个需要被解决的问题就是解码，也即  $\arg \max_{\mathbf{t}} P(\mathbf{s} | \mathbf{t}) \cdot P(\mathbf{t})$  的问题。在这里，给出一种简单的贪婪解码算法作为例子，算

法8.1给出了具体的过程。

---

#### 代码 8.1: 贪婪解码算法

---

```

输入: 源语言句子序列  $s = \{s_1, s_2, \dots, s_m\}$ 
输出: 解码出的目标语言序列  $t = \{t_1, t_2, \dots, t_l\}$ 
初始化一个空集  $t_{best} = \emptyset$  记录解码出的最优目标语言序列;
初始化一个布尔数组  $u$ , 记录每个源语言单词是否已经被解码;
对于每个源语言单词  $s_i$ , 获取其对应单词级别目标语言候选  $\pi$ ;
for  $i = 1$  to  $m$  do
    初始化一个空集  $h = \emptyset$  记录当前最优解码序列的扩展;
    for  $j = 1$  to  $m$  do
        if  $u[j]$  为假, 也即当前单词没有被解码过 then
            | 将当前单词的候选翻译  $\pi[j]$  添加到  $h$  当中;
        end
    end
    从当前的最优翻译扩展候选  $h$  中找到最优的翻译结果, 并记为  $t_{best}$ ;
    这一最优结果所对应的源语言单词被记录为已解码  $u[j] = True$ ;
end
return  $t_{best}$ 
```

---

## 8.2.2 IBM 模型 I

上一节介绍了 IBM 模型如何将翻译问题转化为一个机器学习问题。然而, 在公式8.5中, 仍然存在两个问题需要解决: (1) 如何高效地计算对不同的词对齐序列求和的问题; (2) 如何合理地简化并建模公式8.5右侧的若干概率。接下来, 我们详细阐述 IBM 模型 I 是如何解决这两个问题的。

从公式8.5等式右侧三个概率的化简及建模入手, 首先是基于目标语言序列估计源语言序列的长度  $P(m|t)$ 。IBM 模型 I 假定源语言句子序列长度的生成概率服从均匀分布, 即:

$$P(m|t) \equiv \epsilon \quad (8.8)$$

其中,  $\epsilon$  是一个常量, 在实际的建模和优化过程中, 它通常被取特定的值来保证概率的归一化。源语言中的每一个单词被认为是等可能地和目标语言中的所有单词对齐, 即对齐概率  $P(a_j|\alpha_1^{j-1}, s_1^{j-1}, m, t)$  按照如下方式进行建模:

$$P(a_j|\alpha_1^{j-1}, s_1^{j-1}, m, t) \equiv \frac{1}{l+1} \quad (8.9)$$

其中,  $l + 1$  表示目标语言序列长度加上一个保留的虚拟单词“Null”。当对齐关系明确之后, IBM

模型 I 假设当前时刻源语言单词  $s_j$  的生成只依赖于和它对齐的目标语言单词  $t_{a_j}$ :

$$P(s_j | \mathbf{a}_1^j, \mathbf{s}_1^{j-1}, m, \mathbf{t}) \equiv f(s_j | t_{a_j}) \quad (8.10)$$

上式中的  $f(s_j | t_{a_j})$  表示给定目标语言单词  $t_{a_j}$  之后，生成源语言单词  $s_j$  的概率，这一概率将作为 IBM 模型 I 的参数用于后续的优化过程。经过上述三个部分的化简，翻译概率  $P(\mathbf{s}|\mathbf{t})$  可以按照下面的方式得到：

$$\begin{aligned} P(\mathbf{s}|\mathbf{t}) &= \sum_{\mathbf{a}} P(\mathbf{s}, \mathbf{a}|\mathbf{t}) \\ &= P(m|\mathbf{t}) \prod_{j=1}^m P(a_j | \mathbf{a}_1^{j-1}, \mathbf{s}_1^{j-1}, m, \mathbf{t}) P(s_j | \mathbf{a}_1^j, \mathbf{s}_1^{j-1}, m, \mathbf{t}) \\ &= \sum_{\mathbf{a}} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m f(s_j | t_{a_j}) \end{aligned} \quad (8.11)$$

观察 IBM 模型 I 最终的建模结果可以发现，翻译概率  $P(\mathbf{s}|\mathbf{t})$  最终变成了在所有可能的词对齐的基础上，对单词对翻译概率的连乘。因此，可以使用一种十分简单的方式建模原本复杂的公式 8.5 右侧的形式。

另一个需要解决的问题是对齐序列  $\mathbf{a}$  的求和问题。一个长度为  $m$  的源语言序列的每一个单词有可能对齐到长度为  $l+1$  的目标语言的任何一个位置上。因此，对齐序列  $\sum_{\mathbf{a}}(\cdot)$  的求和一共要遍历  $(l+1)^m$  项。每一项都是一个  $m$  个单词级别翻译概率的乘积。具体如下所示：

$$P(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} P(\mathbf{s}, \mathbf{a}|\mathbf{t}) = \underbrace{\sum_{\mathbf{a}} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m f(s_j | t_{a_j})}_{(l+1)^m \text{次循环}} \quad (8.12)$$

从上述分析我们能够看出，公式 8.11 的计算复杂度是  $\mathcal{O}((l+1)^m \cdot m)$ ，这在源语言序列长度  $m$  稍大的情况下几乎是不可能的。因此，在实际的计算过程中，IBM 模型采用如下的计算技巧：

$$\begin{aligned} P(\mathbf{s}|\mathbf{t}) &= \sum_{\mathbf{a}} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m f(s_j | t_{a_j}) \\ &= \frac{\epsilon}{(l+1)^m} \sum_{\mathbf{a}} \prod_{j=1}^m f(s_j | t_{a_j}) \\ &= \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=1}^l f(s_j | t_i) \end{aligned} \quad (8.13)$$

此处的计算技巧通过将若干个连乘结果的加和转换为若干加和结果的连乘。计算复杂度由原本的  $\mathcal{O}((l+1)^m \cdot m)$  降低为  $\mathcal{O}((l+1) \cdot m)$ 。计算复杂度的下降带来了更高的运行效率，可以通过图8.4对上述计算技巧为什么成立有一个直观的理解。

$$\begin{aligned}
 & f(1,0)f(2,0) + f(1,0)f(2,1) + f(1,0)f(2,2) + \\
 & f(1,1)f(2,0) + f(1,1)f(2,1) + f(1,1)f(2,2) + \\
 & f(1,2)f(2,0) + f(1,2)f(2,1) + f(1,2)f(2,2) \\
 \\ 
 & \sum_{a_1=0}^2 \sum_{a_2=0}^2 f(1,a_1)f(2,a_2) = \sum_{a_1=0}^2 \sum_{a_2=0}^2 \prod_{j=1}^2 f(j,a_j) \\
 \\ 
 & (f(1,0) + f(1,1) + f(1,2)) (f(2,0) + f(2,1) + f(2,2)) = \prod_{j=1}^2 \sum_{i=0}^2 f(j,i) \\
 \\ 
 & \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m f(s_j|t_{a_j}) = \prod_{j=1}^m \sum_{i=0}^l f(s_j|t_i)
 \end{aligned}$$

图 8.4 对齐序列计算技巧示例

IBM 模型 I 的优化过程本质上基于极大似然估计的思想，也即找到一组参数  $\hat{\theta} = \{f(s_x|t_y)\}$ ，使得模型  $P_{\theta}(s|t)$  能够对训练集中的句对  $(s, t)$  输出尽可能大的概率，形式化的描述如下：

$$\begin{aligned}
 \hat{\theta} &= \arg \max_{\theta} P_{\theta}(s|t) \\
 &= \arg \max_{\theta} \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i) \\
 \text{s.t. } & \forall t_y, \sum_{s_x} f(s_x|t_y) = 1
 \end{aligned} \tag{8.14}$$

这里的约束表示任意的目标语言单词  $t_y$  翻译到不同源语言单词的概率求和为 1，也即概率的归一化约束。我们知道，拉格朗日乘子法可以将带有  $n$  个变量和  $m$  个约束的优化问题转换为带有  $m+n$  个变量的无约束优化问题。因此，将这一方法应用到上述有约束的优化目标上，我们能够得到如下无约束的优化目标：

$$L(f, \lambda) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i) - \sum_{t_y} \lambda_{t_y} \left( \sum_{s_x} f(s_x|t_y) - 1 \right) \tag{8.15}$$

接下来，通过计算函数  $L(f, \lambda)$  对参数  $f(s_x|t_y)$  导数为 0 的位置得到其极值点：

$$\begin{aligned}\frac{\partial L(f, \lambda)}{\partial f(s_u|t_v)} &= \frac{\partial \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i)}{\partial f(s_u|t_v)} - \frac{\partial \sum_{t_y} \lambda_{t_y} (\sum_{s_x} f(s_x|t_y) - 1)}{\partial f(s_u|t_v)} \\ &= \frac{\epsilon}{(l+1)^m} \cdot \frac{\prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i)}{\partial f(s_u|t_v)} - \lambda_{t_v} \\ &= \frac{\epsilon}{(l+1)^m} \cdot \frac{\sum_{j=1}^m \delta(s_j, s_u) \cdot \sum_{i=1}^l \delta(t_i, t_v)}{\sum_{i=1}^l f(s_u|t_i)} \prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i) - \lambda_{t_v}\end{aligned}\quad (8.16)$$

此处的  $\delta(x, y)$  为指示函数，当  $x = y$  时， $\delta(x, y) = 1$ ，否则  $\delta(x, y) = 0$ 。当上式为 0 时， $L(f, \lambda)$  达到极值点，将上式整理得到：

$$f(s_u|t_v) = \lambda_{t_v}^{-1} \underbrace{\frac{\epsilon}{(l+1)^m} \prod_{j=1}^m \sum_{i=1}^l f(s_j|t_i)}_{P(\mathbf{s}|\mathbf{t})} \underbrace{\sum_{j=1}^m \delta(s_j, s_u) \cdot \sum_{i=1}^l \delta(t_i, t_v)}_{(s_u, t_v) \text{ 的配对次数}} \frac{f(s_u|t_v)}{\sum_{i=0}^l f(s_u|t_i)}\quad (8.17)$$

将单词  $t_v$  翻译到  $s_u$  的期望频次表示为如下形式：

$$c_{\mathbb{E}}(s_u|t_v) \equiv \sum_{j=1}^m \delta(s_j, s_u) \cdot \sum_{i=1}^l \delta(t_i, t_v) \frac{f(s_u|t_v)}{\sum_{i=0}^l f(s_u|t_i)}\quad (8.18)$$

则等式8.17可以简写为如下形式：

$$f(s_u|t_v) = \lambda_{t_v}^{-1} P(\mathbf{s}|\mathbf{t}) c_{\mathbb{E}}(s_u|t_v) = (\lambda'_{t_v})^{-1} c_{\mathbb{E}}(s_u|t_v),\quad (8.19)$$

此处  $(\lambda'_{t_v})^{-1} = \lambda_{t_v}^{-1} P(\mathbf{s}|\mathbf{t})$ ，结合翻译概率  $f(s_u|t_v)$  的归一化约束，容易得到：

$$\lambda'_{t_v} = \sum_{s'_u} c_{\mathbb{E}}(s'_u|t_v)\quad (8.20)$$

这样，原始的等式8.17就被转换为更简洁的如下形式：

$$f(s_u|t_v) = \frac{c_{\mathbb{E}}(s_u|t_v)}{\sum_{s'_u} c_{\mathbb{E}}(s'_u|t_v)}\quad (8.21)$$

模型的参数通过上式所示的迭代过程逐步优化得到。通常，我们会有一个较大的平行语料数据集

$\mathcal{D} = \{(\mathbf{s}_i, \mathbf{t}_i)\}_{i=0}^n$ , 单词对之间的期望频次可以通过下式计算:

$$c_{\mathbb{E}}(s'_u | t_v) = \sum_{i=0}^n c_{\mathbb{E}}(s_u | t_v; (\mathbf{s}_i, \mathbf{t}_i)) \quad (8.22)$$

完整的训练过程如下所示:

---

#### 代码 8.2: IBM 模型 1 训练算法

---

输入: 平行语料数据集  $\mathcal{D} = \{(\mathbf{s}_i, \mathbf{t}_i)\}_{i=0}^n$

输出: 参数  $f(\cdot|\cdot)$  的最优值

初始化  $f(\cdot|\cdot)$  // 例如说可以初始化为均匀分布;

**while**  $f(\cdot|\cdot)$  不收敛 **do**

**for**  $i = 1$  **to**  $n$  **do**

$$| \quad c_{\mathbb{E}}(s_u | t_v; (\mathbf{s}_i, \mathbf{t}_i)) = \sum_{j=1}^m \delta(s_j, s_u) \cdot \sum_{i=1}^l \delta(t_i, t_v) \frac{f(s_u | t_v)}{\sum_{i=0}^l f(s_u | t_i)};$$

**end**

**for** 目标语言中所有可能单词  $t_v$  **do**

$$| \quad \lambda'_{t_v} = \sum_{s'_u} \sum_{i=0}^n c_{\mathbb{E}}(s'_u | t_v; (\mathbf{s}_i, \mathbf{t}_i));$$

**for** 源语言中所有可能单词  $s_u$  **do**

$$| \quad f(s_u | t_v) = \sum_{i=0}^n c_{\mathbb{E}}(s_u | t_v; (\mathbf{s}_i, \mathbf{t}_i)) \cdot (\lambda'_{t_v})^{-1};$$

**end**

**end**

**return**  $f(\cdot|\cdot)$

---

### 8.2.3 IBM 模型 II

IBM 模型 I 虽然很好地简化了模型的复杂程度使得翻译的建模成为了可能, 但其中的一些简化与真实情况存在着较大的差异, 导致翻译性能受到了较大的限制。最突出的问题是词对齐的概率服从均匀分布。图8.5给出了词对齐的倾向性的简单样例。以单词“翻译”为例, 在 IBM 模型 I 当中, 它对齐到目标语言序列的 3 个单词上的概率是均等的。但是, 很显然“翻译”这一单词应该对齐到目标语言序列当中的第三个位置的“Translation”上, 也即  $a_2 = 3$ 。这种词对齐是具有倾向性的, 绝大部分情况下概率不是均等的。

IBM 模型 II 对这一问题作出了修正, 它认为词对齐存在着一定的倾向性。具体来说, IBM 模型 II 假设源语言单词  $x_j$  的对齐位置  $a_j$  的生成概率与它所在的位置  $j$  和源语言序列长度  $m$  以及目标语言序列长度  $l$  有关, 形式化表示为:

$$P(a_j | \mathbf{a}_1^{j-1}, \mathbf{s}_1^{j-1}, m, t) \equiv a(a_j | j, m, l) \quad (8.23)$$



图 8.5 词对齐的倾向性

其中,  $a(a_j|j, m, l)$  表示源语言序列中第  $j$  个位置词对齐的生成概率, 它被建模为 IBM 模型 II 中的一个需要学习的参数。除了词对齐假设之外, 其余的模型假设均与 IBM 模型 I 相同, 将新的词对齐生成概率按照上一小节所述的建模过程能够得到 IBM 模型 II 的翻译建模表达式为:

$$P(\mathbf{s}|\mathbf{t}) = \epsilon \prod_{j=1}^m \sum_{i=1}^l a(i|j, m, l) f(s_j|t_i) \quad (8.24)$$

### 8.2.4 IBM 模型 III

IBM 模型 I 和 II 存在一个共同的问题是将单词翻译的过程建模为了一个独立的过程, 这就导致它们不能很好地描述多个源语言单词对齐到同一个目标语言单词的情况。IBM 模型 III 是一种基于繁衍率的模型, 可以在一定程度上解决上述问题。这里的繁衍率(Fertility)是指每个目标语言单词生成源语言单词的个数。接下来, 我们简单描述基于繁衍率的模型的整体翻译流程。

如图8.6所示, 模型首先确定每个目标语言单词的繁衍率  $\phi_i$ , 接下来, 依据繁衍率确定目标语言对应的源语言单词是什么, 这样就得到了每个目标语言单词所对应的源语言单词列表  $\tau_i$ 。最后将所有单词列表中的单词放置在合适的位置上就得到了源语言单词序列  $\mathbf{s}$ 。以单词 Nothing 的生成过程为例, 从图8.6当中可以看出, 它的繁衍率  $\phi_1 = 2$ , 这意味着在源语言当中它对应两个单词。接下来, 确定 Nothing 的源语言单词列表为  $\tau_1 = \{\tau_{11} = \text{“没有”}, \tau_{12} = \text{“人”}\}$ 。

最后, 将单词列表中的单词放置在合适的位置, Nothing 对应的单词列表  $\tau_1$  中的单词分别应该被放置在源语言序列中的第 1 和第 2 个位置。因此,  $\pi_1 = \pi_{11} = 1, \pi_{12} = 2$ 。最后, 直接给出

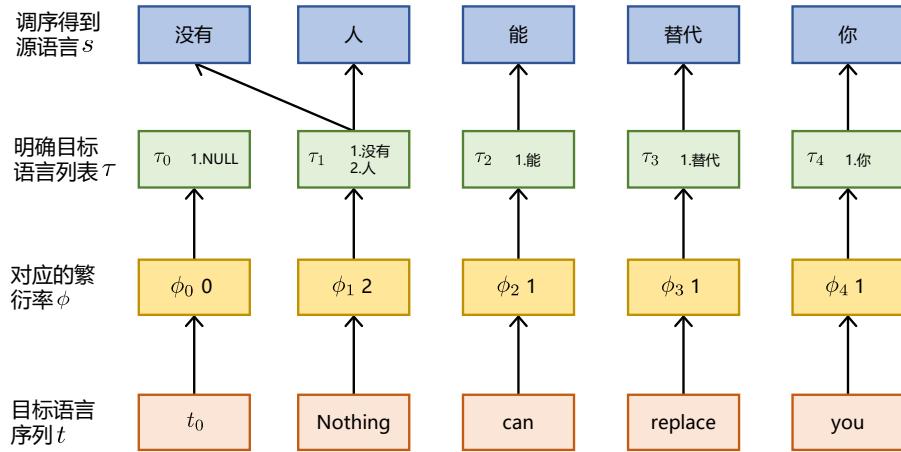


图 8.6 基于繁衍率的模型示意图

IBM 模型 III 的翻译概率的形式，详细过程可以参阅文献 [410] 了解这一模型的建模过程：

$$\begin{aligned}
 P(\mathbf{s}|\mathbf{t}) = & \sum_{\boldsymbol{\alpha}} \left[ \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^l \phi_i! n(\phi_i|t_i) \cdot \prod_{j=1}^m t(s_j|t_{a_j}) \cdot \prod_{j=1, a_j \neq 0}^m d(j|a_j, m, l) \right], \\
 \text{s.t. } & \sum_{s_x} t(s_x|t_y) = 1, \\
 & \sum_j d(j|i, m, l) = 1, \\
 & \sum_{\phi} n(\phi|t_y) = 1, \\
 & p_0 + p_1 = 1
 \end{aligned} \tag{8.25}$$

其中， $n(\phi_i|t_i) = P(\phi_i|t_i)$  指繁衍率的分布， $s, t, m, l$  分别表示源语言句子、目标语言译文、源语言和目标语言序列长度， $\phi, \tau, \pi$  分别表示繁衍率，生成的源语言单词以及它们在源语言序列中的位置， $d(j|i, m, l)$  通常被称为扭曲度函数。

### 8.2.5 IBM 模型 IV

当一个目标语言单词对应多个源语言单词时，这些源语言单词往往会构成一个整体，也即一个短语。然而前面所述的三个 IBM 模型并没有对与这种情况做特殊的设计，这就导致了源语言中的单词短语可能会被打散。针对这个问题，IBM 模型 IV 做出了进一步的修正。它将原本单词之间的对应关系拓宽到了概念 (Concept) 之间的对应。这里的概念是指具有独立语法或语义的一组单词。

IBM 模型将目标语言的概念约束为那些非空对齐的目标语言单词，且要求所有的目标语言概念都只能由一个单词构成。例如，图8.7所给出的单词“yourself”是目标语言序列中的第 2 个概念，而单词“in”不是概念。在后面的叙述当中，标记目标语言中第  $i$  个概念所在的位置为  $[i]$ ， $\odot_i$  表示目标语言中第  $i$  个概念对应的源语言位置的平均值，若这个平均值不是整数则向上取整。

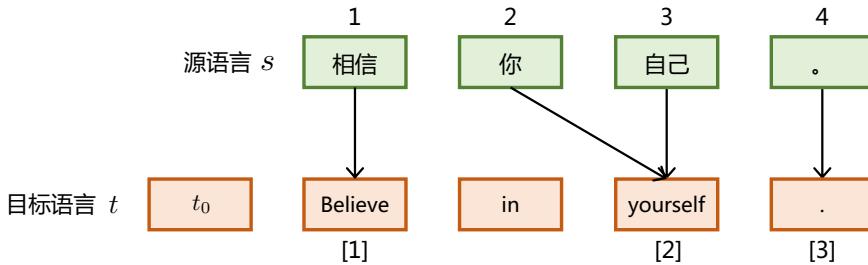


图 8.7 概念对齐示意图

IBM 模型 IV 所做的修正主要体现在扭曲度的建模，对于  $[i]$  对应的源语言单词列表中的第一个单词  $\tau_{[i]1}$ ，它的扭曲度计算公式如下：

$$P(\pi_{[i]1} = j | \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, t) = d_1(j - \odot_{i-1} | A(t_{[i-1]}), B(s_j)) \quad (8.26)$$

此处的  $\pi_{ik}$  表示目标语言序列中第  $i$  个单词所对应的源语言列表中的第  $k$  个单词的位置。对于列表中其他单词的扭曲度，则使用如下公式进行计算：

$$P(\pi_{[i]k} = j | \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, t) = d_{>1}(j - \pi_{[i]k-1} | B(s_j)) \quad (8.27)$$

其中， $A(\cdot)$  和  $B(\cdot)$  分别表示从源语言、目标语言单词向单词词类映射的函数。这一扭曲度函数的改进背后的思想是，在生成  $t_{[i]}$  的第一个源语言单词时，要考虑平均位置  $\odot_{[i]}$  和这个源语言单词之间的绝对距离，随后生成的单词所放置的位置则要考虑前一个放置完的单词的相对位置以及当前源语言单词的词类。这个过程实际上使得同一个目标语言单词所生成的源语言单词之间可以相互影响，从而避免了独立生成各个源语言单词所带来的冲突问题。

## 8.2.6 IBM 模型 V

相对于前面叙述的 4 个模型，IBM 模型 V 针对词对齐的过程做了进一步的约束。它认为同一个源语言单词不应当由多个目标语言单词转换而来。如图8.8所示，前面 4 种词对齐方式都是合法的。然而，对于词对齐  $a_5$  和  $a_6$  来说，源语言单词“机器”和“翻译”分别对应着两个目标语言单词。为了约束这种情况的出现，IBM 模型 V 在放置每一个源语言单词时都会检查这个位置是否已经放置了其他单词。为了实现这一点，引入一个新的变量  $v(j, \tau_1^{[i]-1}, \tau_{[i]1}^{k-1})$ ，它表示在放置  $\tau_{[i]k}$  之前，源语言的前  $j$  个位置还有多少空余。为了简便起见，后续记这个变量为  $v_j$ 。这样，对于单词  $[i]$  所

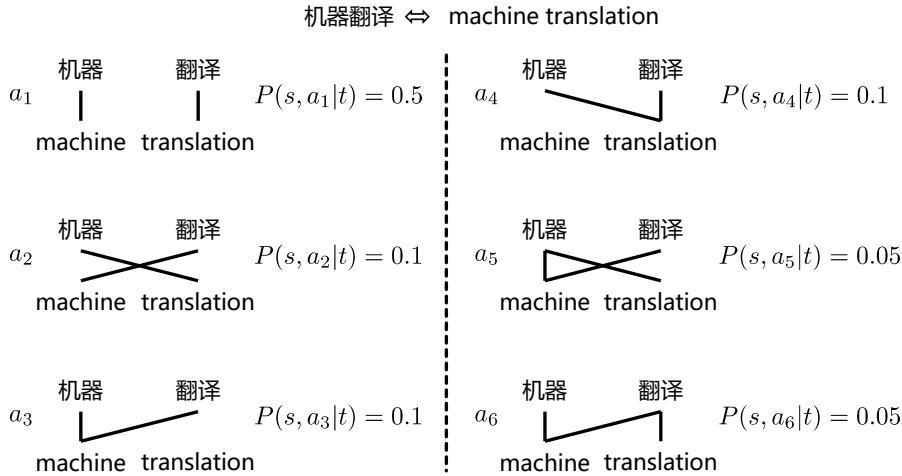


图 8.8 词对齐冲突示意图

对应的源语言单词列表中的第一个单词  $\tau_{[i]1}$  有:

$$P(\tau_{[i]1} = j | \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, t) = d_1(v_j | B(s_j), v_{\odot_{i-1}}, v_m - (\phi_{[i]} - 1)) \cdot (1 - \delta(v_j, v_{j-1})), \quad (8.28)$$

对于其他单词  $\tau_{[i]k}, 1 < k \leq \phi_{[i]}$ , 有:

$$\begin{aligned} & P(\pi_{[i]k} = j | \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, t) \\ &= d_{>1}(v_j - v_{\pi_{[i]k-1}} | B(s_j), v_m - v_{\pi_{[i]k-1}} - \phi_{[i]} + k) \cdot (1 - \delta(v_j, v_{j-1})), \end{aligned} \quad (8.29)$$

此处的  $1 - \delta(v_j, v_{j-1})$  是用来判断第  $j$  个位置是否为空。如果第  $j$  个位置为空, 则  $v_j = v_{j-1}$ , 这样  $P(\pi_{[i]1} = j | \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, t) = 0$ 。这样就避免了词对齐的冲突问题。

### 8.3 基于神经网络的机器翻译方法

传统机器翻译方法高度依赖于繁杂的特征工程, 合理特征的设计对系统构建者的语言学背景具有较高的要求, 同时需要在不断地试错过程中修正<sup>[414, 411]</sup>。这些特征往往不能够完整地反映输入文本的语义。举例来说, 语言模型作为传统机器翻译模型的重要组成部分, 为了降低模型复杂度而引入的马尔可夫假设使得上下文窗口之外的语义依赖无法被建模<sup>[412]</sup>; 从输入文本表示的角度来说, 经典的词袋模型 (bag-of-words, BOW) 则忽略了词序对输入文本表示的影响<sup>[413]</sup>。与之相对, 神经网络模型作为一个强力的特征抽取器, 能够自动地学习输入文本的最优秀特征, 从而在很大程度上减少对领域知识的要求及繁琐的特征工程预处理步骤<sup>[414]</sup>。此外, 尽管传统机器翻译方法经过多年的发展已经能够实现不错的翻译性能, 但存在一些固有缺陷影响其进一步提升。以最具

代表性的基于短语的统计机器翻译方法为例，翻译通过将输入的源语言切分成短语并替换为目标语言的过程完成，短语范围之外的长程依赖在这一过程中被完全忽略进而造成翻译结果中的错误和不一致性。同时，为了提升翻译的准确性和流畅度，越来越多的功能模块不断被设计并添加到统计翻译模型当中（如语言模型、调序模型、长度调整模型等）<sup>[415–417]</sup>。复杂的翻译组件使得系统的整体调优和稳定性受到一定程度的影响。而以循环神经网络、Transformer 为代表的神经机器翻译方法能够有效地建模长程依赖，端到端的特性也使得系统的整体结构变得更加紧凑易于调整。

现代神经机器翻译模型大多依据序列到序列的方式对任务进行建模。给定源语言输入文本，训练目标是找到最合适的目标语言句子作为译文。如图8.9所示，这一过程大体上可以划分为编码和解码两个模块步骤。编码器旨在将源语言句子转化为对应的语义向量，而解码器通过这些向量预测出合适的目标语言句子。

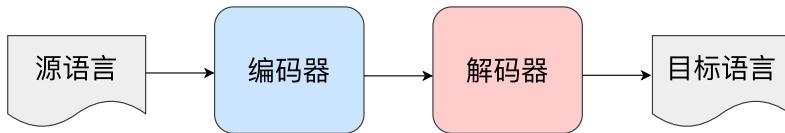


图 8.9 神经机器翻译的编码器-解码器框架

形式化地来说，给定源语言序列  $S = \{s_1, s_2, \dots, s_n\}$ ，神经机器翻译模型试图找到具有最大条件概率  $P(T|S)$  的目标语言序列  $T = \{t_1, t_2, \dots, t_m\}$ ， $n$  和  $m$  分别表示源语言和目标语言的长度。在生成目标语言句子的每个单词时，源语言和已经生成的目标语言信息会被使用。因此，神经机器翻译的整体过程可以按照如下公式描述：

$$\arg \max \prod_{i=1}^m P(t_i | t_{j < i}, S), \quad (8.30)$$

基于上述的总体目标，神经机器翻译利用不同的网络结构针对后验概率  $P(T|S)$  进行建模，常见的结构包括卷积神经网络、循环神经网络、自注意力神经网络等，本节将分别介绍上述常见神经机器翻译网络模型。

### 8.3.1 循环神经网络翻译模型

如前所述，神经机器翻译模型大多基于序列到序列的架构完成从源语言到目标语言的转换过程。不同神经机器翻译模型的主要区别在于编码器和解码器所采用的结构上的差异。自然语言文本可以看做一种时间序列数据，因此一种常见做法是采用基于循环神经网络的结构完成对源语言文本的编码以及目标语言文本的生成。基于循环神经网络的机器翻译模型整体结构如图8.10所示。其中，左侧为编码器部分，源语言单词按照其在文本序列中的先后顺序被依次送入到循环神经网络（RNN）当中。在每个时间步  $t$  中，模型依据送入的源语言单词  $x_t$  对应修改维护其模型内部的

隐状态  $h_t$ , 这个隐状态编码了输入的源语言序列前  $t$  个时刻的所有必要信息。按照这种方式当  $m$  个输入全部被送入到编码器之后, 所对应的  $h_m$  可以认为包含了源语言序列的所有信息。

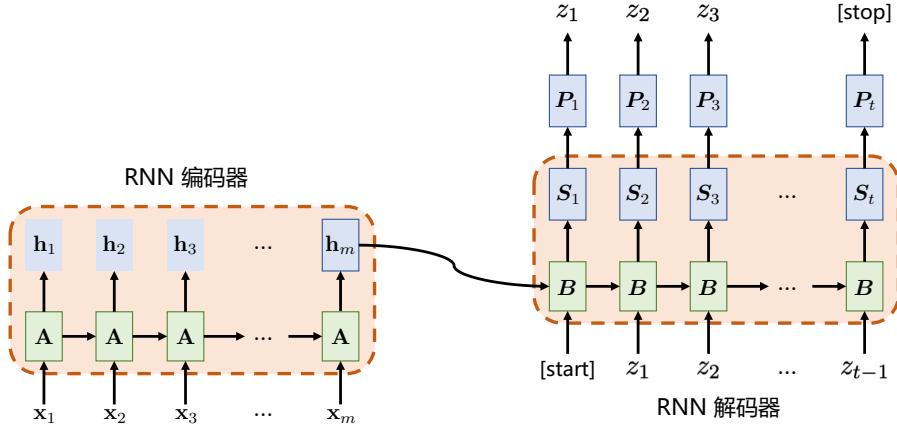


图 8.10 循环神经网络翻译模型

图8.10的右半部分是 RNN 解码器部分, 它接收编码器输出的编码源语言句子信息的向量  $h_m$  作为初始隐状态  $s_0$ 。由于 RNN 的循环过程在每个时间步都要求一个输入单词, 为了启动解码过程, 一般会使用一个保留的特殊符号“[Start]”作为翻译开始的标记送入到 RNN 解码器当中并解码出目标语言序列的第一个单词  $z_1$ 。接下来,  $z_1$  会作为下一个时刻的输入被送入到循环神经网络当中并按照不断迭代产生后续的预测。由于目标语言序列的长度无法被提前预知, 因此使用另一个保留符号“[Stop]”作为预测结束的标志。当某一个时刻  $t$  预测出的目标语言单词为  $z_t = “[Stop]”$  时, 解码过程动态地停止。在上述过程当中, 主要涉及到两步运算, 第一步是 RNN 接收前一时刻隐状态  $s_{t-1}$  并依据当前时刻输入  $z_{t-1}$  (目标语言单词  $z_{t-1}$  对应的语义嵌入) 对隐状态进行维护并生成  $s_t$  的运算过程, 第二步是依据当前步骤隐状态生成目标语言单词的过程:

$$s_t = \tanh(z_{t-1} \mathbf{U} + s_{t-1} \mathbf{W}) \quad (8.31)$$

$$\mathbf{p}_t = \text{Softmax}(s_t \mathbf{V}), \quad (8.32)$$

其中  $\mathbf{U}, \mathbf{W}, \mathbf{V}$  是可学习的参数。 $\mathbf{U}, \mathbf{W}$  负责维护循环状态, 而  $\mathbf{V}$  负责将当前时刻状态转换到词表大小的概率分布  $\mathbf{p} \in \mathbb{R}^{vocab\_size}$ , 从中可以采样得到目标语言单词  $z_t$ 。

通过循环网络对源语言文本进行编码, 并生成目标语言翻译结果的过程十分简单。然而, 它仅仅使用一个定长的向量  $h_m$  编码整个源语言序列。这对于较短的源语言文本没有什么问题, 但随着文本序列长度的逐渐加长, 单一的一个向量  $h_m$  可能不足以承载源语言序列当中的所有信息。如图8.11所示, 蓝色的线代表上述简单循环神经性能随源语言文本长度的变化趋势。当文本

长度在 20 个单词以内时，单一向量能够承载源语言文本中的必要信息。随着文本序列的进一步增加，翻译性能的评价指标 BLEU 的值就开始出现明显地下降。因此，这就启发我们使用更加有效地机制从编码器向解码器传递源语言信息，这就是接下来要讲到的注意力机制。

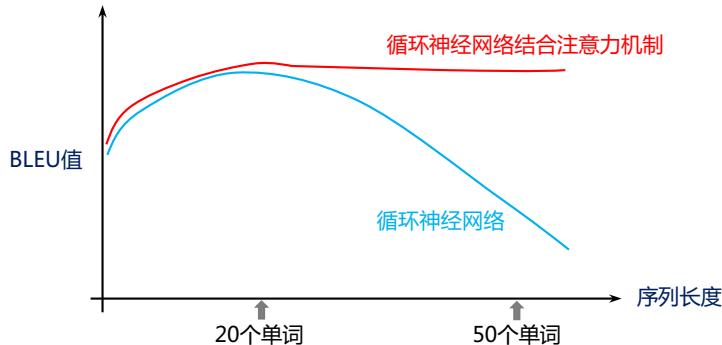


图 8.11 基于简单循环网络的机器翻译模型的瓶颈

引入注意力机制的循环机器翻译架构与基于简单循环网络的机器翻译模型大体结构相似，均采用循环神经网络作为编码器与解码器的实现。关键的不同点在于注意力机制的引入使得不再需要把原始文本中的所有必要信息压缩到一个向量当中。引入注意力机制的循环机器翻译架构如图8.12所示。

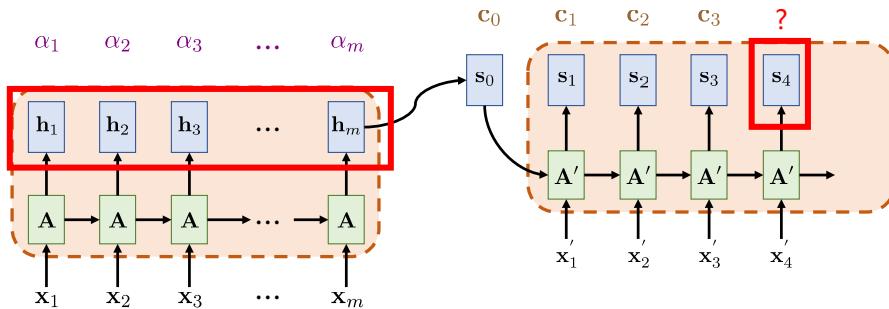


图 8.12 引入注意力机制的循环机器翻译架构

具体来说，给定源语言序列经过编码器输出的向量序列  $h_1, h_2, \dots, h_m$ ，注意力机制旨在依据解码端翻译的需要，自适应地从这个向量序列中查找对应的信息。与简单循环网络相类似，在  $t = 4$  时刻，旨在通过  $t - 1 = 3$  时刻的隐状态  $s_3$  以及  $t = 4$  时刻的输入  $x'_4$  维护循环隐状态并生成当前时刻目标语言翻译结果  $x'_5$ 。为了更高效地考虑源语言上下文语义来提高翻译质量，注意力机制通过通过计算一组匹配分数  $\{\text{score}_i\}_{i=1}^m$  并利用 softmax 归一化为一组权重  $\{\alpha_i\}_{i=1}^m$  自适应地确定源

语言中需要聚焦的部分。具体计算公式如下：

$$\{\alpha_i\}_{i=1}^m = \text{Softmax}(\{\text{score}_i\}_{i=1}^m) \quad (8.33)$$

$$\text{score}_i = \begin{cases} \mathbf{s}_{t-1} \mathbf{h}_i^T & \text{向量乘} \\ \cos(\mathbf{s}_{t-1}, \mathbf{h}_i^T) & \text{向量夹角} \\ \mathbf{s}_{t-1} \mathbf{W} \mathbf{h}_i^T & \text{线性模型} \\ \tanh(\mathbf{W}[\mathbf{s}_{t-1}, \mathbf{h}_i]) \mathbf{v}^T & \text{单层网络} \end{cases} \quad (8.34)$$

其中  $\mathbf{W}, \mathbf{v}$  表示可训练的参数矩阵或向量， $[\cdot, \cdot]$  表示拼接算符。基于上述权重能够得到生成译文  $\mathbf{x}'_t$  所必要的源语言信息  $\mathbf{c}_t$ ，进一步地，可以将这部分源语言信息与当前时刻的输入  $\mathbf{x}'_t$  拼接送入 RNN 作为新的输入：

$$\mathbf{c}_t = \sum_i \alpha_i \mathbf{h}_i \quad (8.35)$$

$$\mathbf{s}_t = \tanh([\mathbf{x}'_t, \mathbf{c}_t] \mathbf{U} + \mathbf{s}_{t-1} \mathbf{W}) \quad (8.36)$$

通过这样的修改，在维护 RNN 任意时刻隐藏状态并生成译文的过程中，能够自适应地考虑源语言中的哪部分信息需要被聚焦，从而生成更加高质量的译文。

### 8.3.2 卷积神经网络翻译模型

卷积神经网络也是一种经典的神经网络结构，被广泛地使用在自然语言处理的各项任务当中。相较于循环神经网络来说，卷积神经网络每一步卷积操作并不依赖于前一时间步的计算结果因而能够充分并行化以更好地利用 GPU 的计算资源。在本章当中，将以 ConvS2S，一种全卷积、高并行、序列到序列的模型为例，介绍卷积神经网络是如何被应用在机器翻译任务当中的。

ConvS2S 的整体结构如图8.13所示，它采用卷积神经网络作为编码器（图片上侧）和解码器（图片左下侧）的具体实现，完成对源语言和目标语言的特征提取，这种模型结构使得每一层的网络计算可以完全并行化，不再受到循环结构中时序依赖的限制。同时，通过堆叠多层卷积结构，上下文窗口的范围得以不断扩大，从而逐渐建模输入文本中的远距离依赖。同时相较于循环神经网络，它的信息传递路径更短，更有利优化。

ConvS2S 作为一种经典的神经机器翻译模型，实现了优越的翻译质量及高效的翻译效率，受到了学术界的广泛关注。它主要由下述几个部件构成：

- **位置编码：**由于 ConvS2S 摒弃了循环结构，因此需要在输入层引入位置编码来标识输入序列中词与词之间的相对位置关系。
- **卷积层与门控线性单元：**这部分是编码器与解码器的实现模块，分别用于抽取源语言和目标语言的上下文语义特征。
- **残差连接：**这部分也被添加在编码器和解码器中堆叠的多层卷积结构当中，直接连接每一层

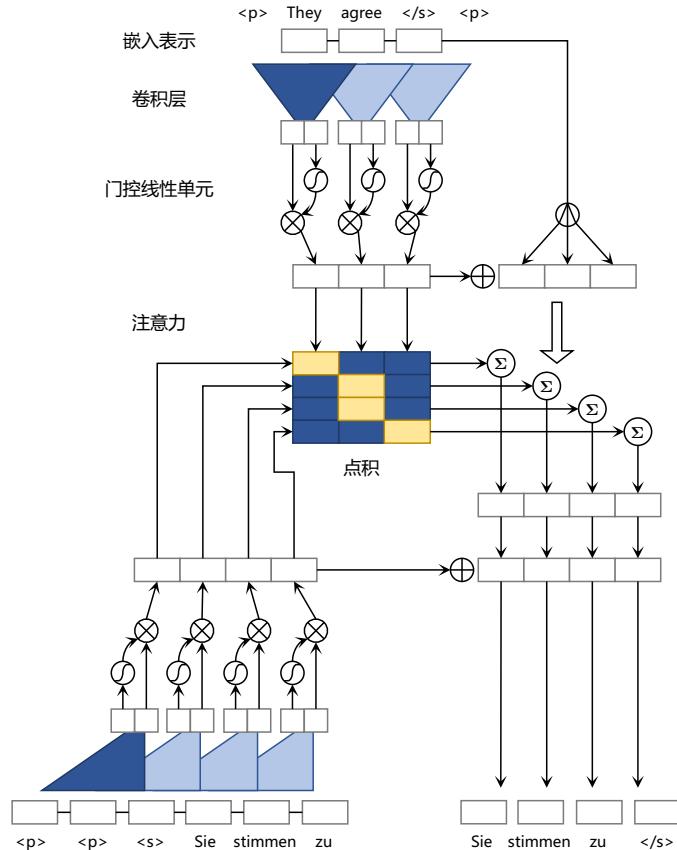


图 8.13 卷积序列到序列模型结构

的输入与输出，从而提高信息传播效率，减小模型的优化难度。

- **多步注意力机制：**这里与上一小节中的循环神经机器翻译模型类似，都采用注意力机制自适应地从源语言端检索译文对应的源语言信息。不同的是，此处的注意力计算在解码器的每一层当中都会出现，因而被称为“多步”注意力。

接下来，将详细介绍 ConvS2S 模型当中涉及到的关键模块的技术细节：

**位置编码：**由于 ConvS2S 不再使用基于循环的结构编码输入序列，因此模型失去了对于输入文本中词与词之间相对位置关系的感知。因此位置编码旨在重新给予模型这部分信息。具体来说，给定输入源语言序列  $S = \{s_1, s_2, \dots, s_n\}$  及其在嵌入空间中的表示序列  $\mathbf{S} = \{s_1, s_2, \dots, s_n\}$ ，其中  $s_j \in \mathbb{R}^f$  是词嵌入矩阵  $D^{V \times f}$  中单词  $s_j$  对应的表示。为了使得卷积模型能够感知到输入序列中单词的相对位置关系，一个额外的位置嵌入  $\mathbf{P} = \{p_1, \dots, p_n\}$  被用于标识每个单词在句子中的绝对位置，其中  $p_j \in \mathbb{R}^f$ 。词嵌入和位置嵌入同时被使用作为编码器和解码器的输入  $E =$

$\{s_1 + p_1, s_2 + p_2, \dots, s_n + p_n\}$ 。

**门控卷积结构：**在 ConvSeq2Seq 模型中，编码器和解码器均采用门控卷积结构作为建模源语言和目标语言的基本部件，这一部件由序列维度上的一维卷积运算和非线性门控机制结合而成。卷积过程能够有效地建模待处理文本中的局部上下文信息，而序列中的长程依赖问题则可以通过多层卷积结构的堆叠得到缓解。非线性门控机制使得我们能够建模输入视野下更加复杂的依赖关系。具体来说，对于嵌入层输入的文本表示  $E \in \mathcal{R}^{n \times f}$ ，通过一个线性映射将维度转换到  $d$  维之后，我们能够得到卷积操作的每个上下文窗口的输入  $X \in \mathcal{R}^{k \times d}$ 。对其进行卷积运算如下：

$$A = X \times W_A + b_A \quad (8.37)$$

$$B = X \times W_B + b_B \quad (8.38)$$

这里存在两组卷积操作，每组由  $d$  个卷积核组成，其参数包括  $W_A \in \mathcal{R}^{k \times d \times d}$ ,  $W_B \in \mathcal{R}^{k \times d \times d}$ ,  $b_A \in \mathcal{R}^d$ ,  $b_B \in \mathcal{R}^d$ 。对  $X$  进行卷积操作得到对应两组输出  $A, B \in \mathcal{R}^d$  后，基于门控线性单元 (Gated Linear Units) 的非线性变换被用作激活函数得到最终输出：

$$h = A \otimes \sigma(B) \quad (8.39)$$

其中  $\otimes$  表示逐点乘法运算，非线性门控  $\sigma(B)$  主要被用于建模  $A$  中的哪部分上下文信息是相关的，哪些需要被遗忘。此外，为了克服卷积神经网络无法建模长程依赖的问题，多层卷积的堆叠是必要的。举例来说，堆叠 6 层上下文窗口大小  $k = 5$  的卷积单元就能够将窗口大小扩大到 25，也即输出能够依赖 25 个单元的输入。为了有效地训练深层神经网络结构，残差链接 (residual connections) 被引入到模型构建当中。具体来说，每一层卷积单元的输入被直接连接到输出当中如下所示：

$$h^l = A^l \otimes \sigma(B^l) + h^{l-1} \quad (8.40)$$

**多步自注意力机制：**在解码译文的过程中，对源语言的参考是必不可少的。ConvS2S 结构中，解码器同样采用了堆叠的多层次门控卷积结构完成对目标语言的解码，并在每一层门控卷积之后通过注意力机制参考源语言信息。以解码器第  $l$  层第  $i$  个时间步的注意力计算为例，为了确定需要参考源语言中的哪部分信息，当前时刻的解码器状态  $z_i^l$  以及前一个时刻解码出的目标语言嵌入  $g_i$  被用于作出决策的依据：

$$d_i^l = z_i^l W_d^l + b_d^l + g_i \quad (8.41)$$

$$z_i^l = \text{Conv}(s_i^l) \quad (8.42)$$

其中  $W_d^l$  和  $b_d^l$  是可训练的参数。基于当前位置的状态依据  $d_i^l$ ，目标语言位置  $i$  相对源语言第  $j$  个

单词的注意力得分  $a_{ij}^l$  可以通过  $\mathbf{d}_i^l$  和源语言编码器对应位置的输出  $\mathbf{y}_j$  计算得到：

$$a_{ij}^l = \frac{\exp(\mathbf{d}_i^l \cdot \mathbf{y}_j)}{\sum_{t=1}^n \exp(\mathbf{d}_i^l \cdot \mathbf{h}_t)} \quad (8.43)$$

基于上述过程得到的注意力得分，可以对源语言不同位置的信息进行加权整合得到为了预测当前位置目标语言单词所需的依据：

$$\mathbf{c}_i^l = \sum_{j=1}^m a_{ij}^l (\mathbf{h}_j + \mathbf{e}_j) \quad (8.44)$$

这里的源语言端同时利用了编码器的输出  $\mathbf{h}_j$  以及对应位置的输入词嵌入  $\mathbf{e}_j$ 。这两者对应着更加全面的源语言信息，在实践中被证明十分有效。基于上述源语言信息，可以得到解码器端第  $l$  层的输出为：

$$\mathbf{s}_i^{l+1} = \mathbf{c}_i^l + \mathbf{z}_i^l \quad (8.45)$$

上述多步注意力机制中的“多步”一词主要从两个方面体现。首先从多层卷积堆叠的角度来说，前一层中通过注意力机制动态地决定哪些相关信息需要被关注并传递到下一层当中，而下一层在计算对源语言不同位置的注意力得分过程中又会考虑到这些信息。从时间步的角度来说，在计算目标语言每个位置  $i$  的注意力分布时，前  $k$  个位置的注意力历史信息  $\mathbf{c}_{i-k}^{l-1}, \dots, \mathbf{c}_i^{l-1}$  都会作为输入的一部分被考虑。这就使得我们在计算当前位置需要参考源语言中的哪些信息时能够有效地判断哪些信息在之前的时间步中已经被参考过了。与之相对，在循环神经网络中，这一过程需要历经多个时间步中的非线性转换才能够被利用，这极有可能造成相关信息的丢失。

### 8.3.3 自注意力神经网络翻译模型

基于循环或卷积神经网络的序列到序列建模方法是现存机器翻译任务中的经典方法。然而，它们在建模文本长程依赖方面都存在一定的局限性。对于卷积神经网络来说，受限的上下文窗口在建模长文本方面天然地存在不足。而对于循环神经网络来说，上下文的语义依赖是通过维护循环单元中的隐藏状态实现的。在编码过程中，每一个时间步的输入建模都涉及到对隐藏状态的修改。随着序列长度的增加，编码在隐藏状态中的序列早期的上下文信息被逐渐遗忘。尽管注意力机制的引入在一定程度上缓解了这个问题，但循环网络在编码效率方面仍存在很大的不足之处。由于编码端和解码端的每一个时间步的隐藏状态都依赖于前一时间步的计算结果，这就造成了在训练和推断阶段的低效。

Transformer<sup>[418]</sup> 是由谷歌在 2017 年提出的一种 Seq2Seq 架构。它的出现使得机器翻译的性能和效率迈向了一个新的阶段。它摒弃了循环结构，并完全通过注意力机制完成对源语言序列和目标语言序列全局依赖的建模。在抽取每个单词的上下文特征时，Transformer 通过自注意力机制（self-attention）衡量上下文中每一个单词对当前单词的重要程度，在这个过程当中没有任何的循环单元参与计算。这种高度可并行化的编码过程使得模型的运行变得十分高效。

基于 Transformer 的机器翻译模型架构如图8.14所示，左侧和右侧分别对应着 Seq2Seq 模型的编码器和解码器结构。它们均由若干个基本的 Transformer 层组成（对应着图中的灰色框）。每个 Transformer 层都接收一个向量序列  $\{\mathbf{x}_i\}_{i=1}^t$  作为输入，并输出一个等长的向量序列作为输出  $\{\mathbf{y}_i\}_{i=1}^t$ 。这里的  $\mathbf{x}_i$  和  $\mathbf{y}_i$  分别对应着文本序列中的一个单词的表示。而  $\mathbf{y}_i$  是当前 Transformer 对输入  $\mathbf{x}_i$  进一步整合其上下文语义后对应的输出。在从输入  $\{\mathbf{x}_i\}_{i=1}^t$  到输出  $\{\mathbf{y}_i\}_{i=1}^t$  的语义抽象过程中，主要涉及到如下几个模块：

- **自注意力子层**：对应图中的 Multi-Head Attention 部分。使用自注意力机制整合上下文语义，它使得序列中任意两个单词之间的依赖关系可以直接被建模而不基于传统的循环结构，从而更好地解决文本的长程依赖。
- **前馈子层**：对应图中的 Feed Forward 部分。通过全连接层对输入文本序列中的每个单词表示进行更复杂的变换。
- **残差连接**：对应图中的 Add 部分。它是一条分别作用在上述两个子层当中的直连通路，被用于连接它们的输入与输出。从而使得信息流动更加高效，有利于模型的优化。
- **层标准化**：对应图中的 Norm 部分。作用于上述两个子层的输出表示序列中，对表示序列进行层标准化操作，同样起到稳定优化的作用。

相比于编码器端，解码器端要更复杂一些。具体来说，解码器的每个 Transformer 层的第一个自注意力子层额外增加了注意力掩码，对应图中的掩码多头注意力（Masked Multi-Head Attention）部分。这主要是因为在翻译的过程中，编码器端主要用于编码源语言序列的信息，而这个序列是完全已知的，因而编码器仅需要考虑如何融合上下文语义信息即可。而解码端则负责生成目标语言序列，这一生成过程是自回归的，即对于每一个单词的生成过程，仅有当前单词之前的目标语言序列是可以被观测的，因此这一额外增加的掩码是用来掩盖后续的文本信息，以防模型在训练阶段直接看到后续的文本序列进而无法得到有效地训练。此外，解码器端还额外增加了一个多头注意力（Multi-Head Attention）模块，需要注意的是它同时接收来自编码器端的输出以及当前 Transformer 层第一个掩码注意力层的输出。它的作用是在翻译的过程当中，为了生成合理的目标语言序列需要观测待翻译的源语言序列是什么。基于上述的编码器和解码器结构，一个待翻译的源语言文本首先经过编码器端的每个 Transformer 层对其上下文语义的层层抽象，最终输出每一个源语言单词上下文相关的表示。解码器端以自回归的方式生成目标语言文本，即每个时间步  $t$  参考编码器端输出的所有源语言文本表示以及前  $t - 1$  个时刻生成的目标语言文本生成当前时刻的目标语言单词。接下来详细介绍翻译过程当中所涉及到的不同模块的技术细节。

**位置编码**：对于待翻译的文本序列，首先通过输入嵌入层（Input Embedding）将每个单词转换为其相对应的向量表示。在送入编码器端建模其上下文语义之前，一个非常重要的操作是在词嵌入中加入位置编码这一特征。由于 Transfomer 不再使用基于循环的方式建模文本输入，序列中不再有任何信息能够提示模型单词之间的相对位置关系。因此补充这部分信息是十分必要的。具体来说，序列中每一个单词所在的位置都对应一个实值向量。这一实值向量会与单词表示对应相

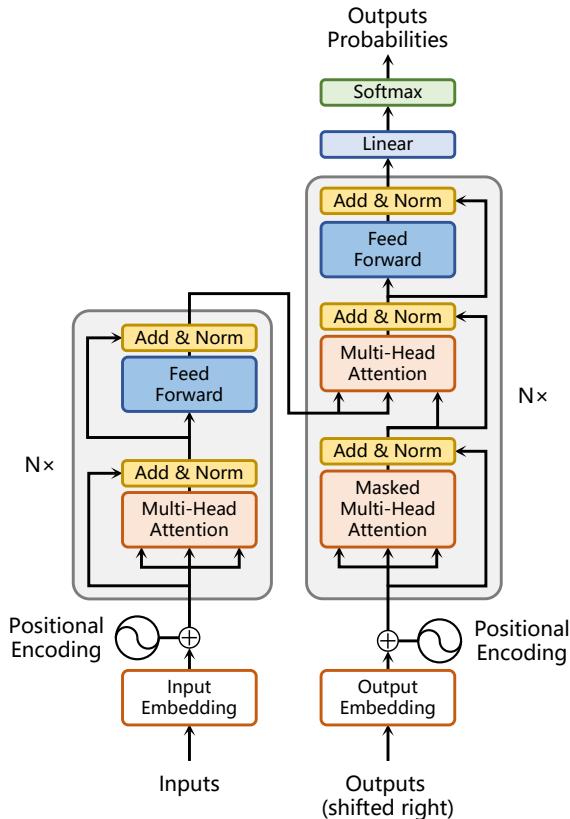


图 8.14 基于 Transformer 的机器翻译模型架构<sup>[418]</sup>

加并送入到后续模块中做进一步处理。在训练的过程当中，模型会自动地学习到如何利用这部分位置信息。为了得到不同位置对应的编码，Transformer 使用不同频率的正余弦函数如下所示：

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d}}\right) \quad (8.46)$$

$$\text{PE}(\text{pos}, 2i + 1) = \cos\left(\frac{\text{pos}}{10000^{2i/d}}\right), \quad (8.47)$$

其中， $\text{pos}$  表示单词所在的位置， $2i$  和  $2i+1$  表示位置编码向量中的对应维度， $d$  则对应位置编码的总维度。通过上面这种方式计算位置编码有这样几个好处：首先，正余弦函数的范围是在  $[-1, +1]$ ，导出的位置编码与原词嵌入相加不会使得结果偏离过远而破坏原有单词的语义信息。其次，依据三角函数的基本性质，可以得到第  $\text{pos} + k$  个位置的编码是第  $\text{pos}$  个位置的编码的线性组合，这就意味着位置编码中蕴含着单词之间的距离信息。

**自注意力子层**: 自注意力 (Self-Attention) 机制是基于 Transformer 的机器翻译模型的基本操作，

在源语言的编码和目标语言的生成中频繁地被使用以建模源语言、目标语言任意两个单词之间的依赖关系。给定由单词语义嵌入及其位置编码叠加得到的输入表示  $\{x_i \in \mathbb{R}^{d_i}\}_{i=1}^t$ , 为了实现对上下文语义依赖的建模, 我们进一步引入在自注意力机制中涉及到的三个元素: 查询  $q_i$  (Query), 键  $k_i$  (Key), 值  $v_i$  (Value)。在编码输入序列中每一个单词的表示的过程中, 这三个元素被用于计算上下文单词所对应的权重得分。直观地说, 这些权重反映了在编码当前单词的表示时对于上下文不同部分所需要的关注程度。具体来说, 如图8.15所示, 通过三个线性变换  $W^Q \in \mathbb{R}^{d_i \times d_k}$ ,  $W^K \in \mathbb{R}^{d_i \times d_k}$ ,  $W^V \in \mathbb{R}^{d_i \times d_v}$  将输入序列中的每一个单词表示  $x_i$  转换为其对应的  $q_i \in \mathbb{R}^{d_k}$ ,  $k_i \in \mathbb{R}^{d_k}$ ,  $v_i \in \mathbb{R}^{d_v}$  向量。

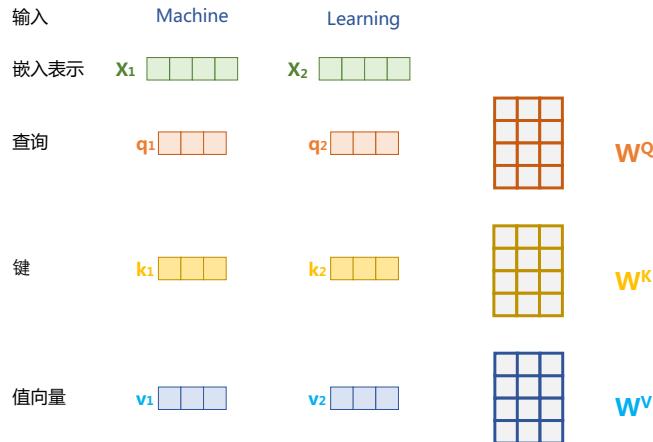


图 8.15 自注意力机制中的查询、键、值向量

为了得到编码单词  $x_i$  时所需要关注的上下文信息, 通过位置  $i$  查询向量与其他位置的键向量做点积得到匹配分数  $q_1 \cdot k_1, q_2 \cdot k_2, \dots, q_t \cdot k_t$ 。为了防止过大的匹配分数在后续 softmax 计算过程中导致的梯度爆炸以及收敛效率的问题, 这些得分会除以缩放因子  $\sqrt{d}$  以稳定优化。缩放后的得分经过 softmax 归一化为概率之后与其他位置的值向量相乘来聚合我们希望关注的上下文信息并最小化不相关信息的干扰。上述计算过程可以被形式化地表述如下:

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (8.48)$$

其中  $Q \in \mathbb{R}^{L \times d_k}$ ,  $K \in \mathbb{R}^{L \times d_k}$ ,  $V \in \mathbb{R}^{L \times d_v}$  分别表示输入序列中的不同单词的  $q, k, v$  向量拼接组成的矩阵,  $L$  表示序列长度,  $Z \in \mathbb{R}^{L \times d_v}$  表示自注意力操作的输出。为了进一步增强自注意力机制聚合上下文信息的能力, 一种被称为多头 (Multi-head) 自注意力的机制被提出以从关注上下文的不同侧面。具体来说, 上下文中每一个单词的表示  $x_i$  经过多组线性映射  $\{W_j^Q, W_j^K, W_j^V\}_{j=1}^N$  到不同的表示子空间当中, 公式8.48会在不同的子空间中分别计算并得到不同的上下文相关的单

词序列表示  $\{\mathbf{Z}_j\}_{j=1}^N$ 。最终，一个线性变换  $\mathbf{W}^O \in \mathbb{R}^{(Nd_v) \times d}$  被用于综合不同子空间中的上下文表示并形成自注意力层最终的输出  $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^t$ 。

**前馈子层：**前馈子层接受自注意力子层的输出作为输入，并通过一个带有 `Relu` 激活函数的两层全连接网络对输入进行更加复杂的非线性变换。实验证明，这一非线性变换会对模型最终的性能产生十分重要的影响。

$$\text{FFN}(\mathbf{x}) = \text{Relu}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (8.49)$$

$$\text{Relu}(\mathbf{x}) = \max(0, \mathbf{x}), \quad (8.50)$$

其中  $\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2$  表示前馈子层的参数。实验结果表明，增大前馈子层隐状态的维度有利于提升最终翻译结果的质量，因此，前馈子层隐状态的维度一般比自注意力子层要大。

**残差连接与层标准化：**事实上 Transformer 是一个非常庞大的网络结构。它的编码器和解码器均由 6 层基本的 Transformer 层组成，每一层当中都包含复杂的非线性映射，这就导致模型的训练比较困难。因此，研究者们在 Transformer 层中进一步引入了残差连接与层标准化技术以进一步提升训练的稳定性。具体来说，残差连接主要是指使用一条直连通道直接将对应子层的输入连接到输出上去，从而避免由于网络过深在优化过程中潜在的梯度消失问题：

$$\mathbf{x}^{l+1} = f(\mathbf{x}^l) + \mathbf{x}^l, \quad (8.51)$$

其中  $\mathbf{x}^l$  表示第  $l$  层的输入， $f(\cdot)$  表示一个映射函数。此外，为了进一步使得每一层的输入输出范围稳定在一个合理的范围内，层标准化技术被进一步引入 Transformer 的每一层当中：

$$LN(\mathbf{x}) = g \cdot \frac{\mathbf{x} - \mu}{\sigma} + b \quad (8.52)$$

其中  $\mu$  和  $\sigma$  分别表示均值和方差，用于将数据平移缩放到均值为 0，方差为 1 的标准分布， $g$  和  $b$  是可学习的参数。层标准化技术可以有效地缓解优化过程中潜在的不稳定、收敛速度慢等问题。

## 8.4 机器翻译语料库

数据集是评测机器学习算法的基础，本小节针对机器翻译领域广泛使用的基准数据集进行一个简单的介绍，读者可以基于这些数据集复现现存主流的机器翻译模型并进行比较。

- **WMT 数据集：**该数据集是一个以英语为主的多语言机器翻译数据集，涉及英中、英德翻译等多种任务，数据来源于新闻、医学、翻译等领域。
- **IWSLT 数据集：**该数据集是一个来自于 TED 演讲的文本翻译数据集，语料规模较小，涉及英德、英中翻译等任务。
- **NIST 数据集：**该数据集是一个新闻翻译领域的高质量数据集，评测集包括 4 句参考译文，涉

及中英、英捷翻译等任务。

- **TVsub**: 该数据集是一个抽取自电视剧字幕的翻译数据集，适合于对话中的长距离上下文依赖研究，涉及中英翻译任务。
- **Flickr30k**: 该数据集是多模态机器翻译的主流数据集之一，包含 31783 张图片，每张图片对应 5 个语句标注，涉及英德翻译任务。
- **Multi30k**: 该数据集是多模态机器翻译的主流数据集之一，包含 31014 张图片，每张图片对应 5 个语句标注，设计英德、英法翻译任务。
- **IAPRTC-12**: 该数据集是多模态翻译的主流数据集之一，包含 20000 张图片及其对应标注，涉及英德翻译任务。
- **IKEA**: 该数据集是多模态翻译领域的主流数据集之一，包含 3600 张图片及其对应标注，涉及英德、英法翻译任务。

此外，机器翻译模型的训练需要大规模双语数据，这里针对一些主流公开的双语平行语料进行了简单汇总。

- **News Commentary Corpus**: 该语料库爬取自 Project Syndicate 网站的政治经济评论，涉及中文、英语等 12 个语种，64 个语言对的双语数据。
- **CWMT Corpus**: 该语料库是由中国计算机翻译研讨会社区收集和共享的中英平行预料数据，来源于新闻、电影、小说、政府文档等多个领域。
- **Common Crawl Corpus**: 该语料库是爬取自互联网网页的数据，涵盖捷克语、德语、法语、俄语等 4 种语言到英语的平行语料。
- **Europarl Corpus**: 该语料库的数据来源是欧洲议会记录，涵盖了保加利亚语、捷克语等 20 种欧洲语言到英语的平行语料。
- **ParaCrawl Corpus**: 该语料库的数据来源依然是互联网爬取，包含了 23 种欧洲语言到英语的双语语料。
- **United Nations Parallel Corpus**: 该语料库的数据来源是联合国公共领域的官方记录和其他会议文件，涵盖了阿拉伯语、英语、西班牙语、法语、俄语、汉语等 6 种联合国正式语言。
- **TED Corpus**: 该语料库的数据来源是 TED 大会演讲在其网站公开的自 2007 年以来的演讲字幕，以及超过 100 种语言的翻译版本。
- **OpenSubtitle**: 该语料库是由 P.Lison 和 J.Tiedemann 收集自 opensubtitles 电影字幕网站，是一个涵盖了 62 种语言、1782 个语言对的平行语料的大规模数据集。
- **Wikititles Corpus**: 该语料库的数据来源是维基百科的标题，涵盖了古吉拉特语等 14 个语种，11 个语言对的平行预料数据。
- **CzEng**: 该语料库的数据来源是欧洲法律、信息技术和小说领域，涵盖了捷克语和英语的平行语料。
- **Yandex Corpus**: 该语料库的数据均爬取自互联网网页，涵盖了俄语和英语的平行语料。

- **Tilde MODEL Corpus**: 该语料库由多个来自于经济、新闻、政府、旅游等门户网站的数据集组成，涵盖了欧洲多种语言的开放数据。
- **Setimes Corpus**: 该语料库的数据来源是东南欧时报的新闻报道，涵盖了克罗地亚语、阿尔巴尼亚语等 9 种巴尔干语言，72 个语言对的双语数据。
- **TVsub**: 该语料库的数据来源是电视剧字幕的中英文对话的语料，主要用于对话和长距离上下文信息依赖的研究。
- **Recipe Corpus**: 该语料库是一个包含 10 万多个句对的由 Cookpad 公司创建的日英食谱语料库。

## 8.5 延伸阅读

从基于规则、统计的模型到如今基于神经网络的模型，机器翻译任务已经经历了数十年的发展并取得了辉煌的成绩。伴随着硬件计算能力的提升和大规模训练数据的积累，我们在很多富资源语言的翻译问题上已经能够实现很好的翻译效果。尽管如此，机器翻译领域中仍存在着很多问题尚未解决，包括神经网络结构优化、低资源翻译、多模态翻译等方面。

如何对现有的神经网络结构进行优化使其更加适配机器翻译任务一直是学术界研究的重点。在这一方面，有如下几个问题值得进一步关注。首先是针对 Transformer 模型的多头注意力模块，许多研究人员发现其中一部分注意力头具有有意义的语言学解释并在模型编码过程中发挥着重要的作用，而存在另一部分注意力头则并没有发挥什么作用。因此如何通过对 Transformer 模型进一步剪枝在不影响性能的前提下提升效率具有很大的使用价值 [419]。此外，通过引入正则化技术 [420]，多尺度表示 [421]，定义隐变量等方式提升源语言和目标语言表示 [422]。最后，现存的 Transformer 模型对于超长文本序列建模能力存在一定的不足之处，如何提升长文本建模的能力也是一个值得研究的方向 [423–425]。

尽管现存的神经机器翻译模型在大规模数据的基础之上实现了令人惊叹的翻译质量，但世界上的大多数语言无法获取到如此大规模的平行语料数据。因此如何更高效地利用已有的单、双语数据始终是学术界的一个研究热点。现有的方法从数据增强的角度探索如何对已有数据进行增强 [426–430]；从预训练模型的角度探索如何将预训练知识更加高效的迁移到下游任务 [431, 432]；以及通过多任务学习的方式探索语言间的资源共享 [433, 434]。最近，零资源翻译也逐渐受到了广泛的关注。这一任务旨在使用少量的平行语料库（覆盖  $k$  个语言），就能够实现在任何  $k(k-1)$  个语言对之间进行翻译 [435]。

针对跨模态的数据进行翻译也是一个极具潜力的研究方向。典型的任务包括语音翻译、图像翻译等。其中，语音翻译的一个重要应用是机器同声传译，其最大的难点在于不同的语言表达语序不同，现存的解决方案包括等待源语言  $k$  个单词后再进行翻译 [436, 437]，或者通过束搜索的方式预测未来的词序列从而提升准确度等 [438, 439]。此外，数据稀缺也是多模态机器翻译任务的一大难点。一些研究者尝试通过调整训练策略使得模型更容易捕获上下文信息 [440] 或数据增强策

略 [441] 来提升整体数据量等。

## 8.6 习题

- (1) 请阐述基于统计的机器翻译方法的任务框架。
- (2) 试比较 5 种 IBM 机器翻译模型，阐述它们针对翻译过程中的哪些问题进行了改进。
- (3) 请阐述基于神经网络的机器翻译方法的任务框架。
- (4) 在基于神经网络的机器翻译方法中，Transformer 模型相较于其他模型具有怎样的优势？它仍然存在什么样的问题？
- (5) 你认为机器翻译任务现在还存在哪些挑战？

# 9. 情感分析

---

人类在语言交流过程中通常富含丰富的情感信息，如何自动理解人类语言信息中的情感，是自然语言处理领域的研究热点与难点。情感分析（Sentiment analysis）又称观点挖掘（Opinion Mining），目标旨在从文本中分析得到人们关于主题或实体的评价、观点或态度，还包括分析文本所表达的情绪信息。情感分析包含了评论挖掘、评价抽取、主观性分析、情绪分析等多种不同的任务，在舆情分析、情报挖掘、电子商务、对话系统等领域具有广泛的应用。近年来，随着社交媒体快速发展，用户通过各种平台发表和分享大量评论和观点内容，情感分析技术无论是在研究还是应用上均取得了显著的进展。

本章首先介绍情感分析的基本概念，在此基础上分别介绍篇章级、句子级以及属性级情感倾向分析任务、常用算法和主要评测数据集。

## 9.1 情感分析概述

随着互联网的快速发展，特别是以博客、微博为代表的 Web 2.0 平台普及，用户在各类社会媒体平台上产生了大量包含评论和观点的文本内容。通过对这些内容进行分析，可以有效的了解用户喜好、发现产品需求、市场情绪分析。大量应用需求也促使自 2000 年以来，情感分析任务逐渐受到越来越多学术界和产业界的关注，包括倾向性形容词发现<sup>[442]</sup>、股票市场情绪分析<sup>[443]</sup>、观点分析与跟踪<sup>[444]</sup>、评论倾向性分析<sup>[445]</sup>等任务相继开展。2003 年，情感分析和观点挖掘的这两个术语也相继在文献 [446] 和文献 [447] 中提出。更早的一些工作包括篇章主观性分析<sup>[448]</sup>、观点跟踪<sup>[449]</sup>、主观性分类<sup>[450]</sup>等任务在 20 世纪 90 年代起就已经开始。

情感分析包含的研究内容众多，包括自然语言处理、数据挖掘、机器学习等不同领域研究人员都对该任务开展了大量研究，也因此造成情感分析相关术语繁杂，有很多研究内容大体相同但是又有一些微小区别的任务，刘兵教授在其关于情感分析专著中<sup>[451]</sup> 给出的任务包括：情感分析（Sentiment Analysis）、观点挖掘（Opinion Mining）、观点抽取（Opinion Extraction）、情感挖掘（Sentiment Mining）、主观性分析（Subjectivity Analysis）、感情分析（Affect Analysis）、情绪分析（Emotion Analysis）、评论挖掘（Review Mining）等。我们采用刘兵教授书的建议，使用情感分析来代表该领域整体研究内容。

本节首先对情感模型进行简单介绍，在此基础上对情感分析主要研究内容进行介绍，包括情感分类情感单元抽取、情绪分类、观点摘要等。

### 9.1.1 情感模型

根据情感分析任务定义，我们可以看到情感分析主要包含两个主要任务：(1) 分析文本中针对某个主题或实体的评价或观点；(2) 分析文本中所表达的情绪类的情感。虽然针对观点分析和情绪分析任务所采用的自然语言处理算法非常类似，但是在语言学和心理学理论中，上述问题还是有非常大的不同。本节中将介绍常见的观点类和情绪类理论模型。

#### 1. 观点模型

观点（Opinion）是指从某一立场或角度出发对事物所持的看法或态度，是一种表达了感觉、看法、信念的陈述。观点的情感倾向也称为极性，可以是正面（褒义）、负面（贬义）或中立。每种不同的情感倾向还具有不同的强度。比如“完美”比“好”表达的褒义程度更强。很多副词也具有增强或者减弱情感倾向的作用。常见的增强词包括：很、非常、very、extremely 等。常见的减弱词包括：可能、一定程度上、slightly、a little bit 等。为了区分情感的强度，还是进一步可以采用情感评分的方法，使用离散化的评分表达情感的强度。比如，可以将情感分为 5 档（1-5 分），1 分表示强烈负面，2 分表示负面，3 分表示中立，4 分表示正面，5 分表示强烈正面。观点从语言学、心理学等不同角度划分为不同类型，包括：常规型观点和比较型观点、显式观点和隐式观点、感性观点和理性观点等。

常规型观点通常简称为观点，是指通过直接或间接的方式对事物所表述的观点<sup>[452]</sup>。

例如： (1) 这家餐厅非常差。

(2) 更换了这台显示器后，我的眼睛感觉非常舒服。

上例中，两个句子都属于常规性观点，句子(1)是直接观点，直接针对餐厅表述了句子作者的负面评价；句子(2)是间接观点，通过描述眼睛的感受，间接的表达了对显示器的正面评价。通常情况下，直接观点出现的比例相较间接观点高很多。目前的工作大多针对直接观点开展，间接观点的研究相对较少。

比较型观点（Comparative Sentiment）是指对两个或更多事物之间的相同或者不同点进行比较，并表达了观点持有者对其中一个事物的态度<sup>[453]</sup>。

例如： (1) 这家餐厅的环境比人民路上那家的好很多。

(2) 新一代的显卡的显存相较于上一代有了大幅度的提升。

上例中，两个句子都属于典型比较型观点。两句中，虽然都没有对实体直接给出评价，但是对实体在某个属性上给出排序。在英语中，比较型观点通常通过形容词或者副词的比较级或者最高级进行表达，但也有一些例外（比如 Prefer）。

人们对观点的表达通常使用包含情感词语的主观性句子，但是也有一些观点的表达通过客观事实描述性句子完成。基于此，可以划分为显式观点和隐式观点两类。显式观点（Explicit Sentiment）

是指在句子中通过常规观点或比较型观点句子直接对观点、看法、感受进行表达的句子。

- 例如： (1) 不错，交通便利，方便出行！
- (2) 展馆太小了，场景少，一般般。

上例中，观点持有者直接表达了自己的感受和评价，都属于典型的显式观点。

隐式观点(Implicit Sentiment)是指在客观或者真实的事事实陈述中蕴含的常规型或比较型的观点。

- 例如： (1) 该款电动车续航里程可达 1000 公里。
- (2) 雪山脚下的一个景点，从进门到出去给了半个小时游览。

上例中，两个句子都是在描述事实，虽然没有使用包含情感的词语，但是也表达了观点持有者的态度。隐式观点虽然在观点句中的占比不高，但是由于是否包含观点以及观点极性分类通常需要依赖常识，使得隐式观点的识别和分类的难度很高。上例中句子(1)表达了正面的观点，这需要依赖当前电动车续航里程通常小于 600 公里这一背景知识。

理性情感 (Rational Sentiment) 是指观点来源于理性的推理，不包含主观情绪。

- 例如： (1) 笔记本电池不错，可以连续使用 40 小时。
- (2) 酒店环境很好，距离沙滩仅有 200 米。

上例中，虽然都包含主观性表达的观点，但都是根据相对理性的方式以客观事实为依据的理性推理。

感性情感 (Emotional Sentiment) 是指观点来源于观点持有者的感性的主观表达。

- 例如： (1) 这个酒店太垃圾了。
- (2) 这是最好的车。

与理性情感相比，感性情感通常更加强烈，在整体评论中的数量也更多。

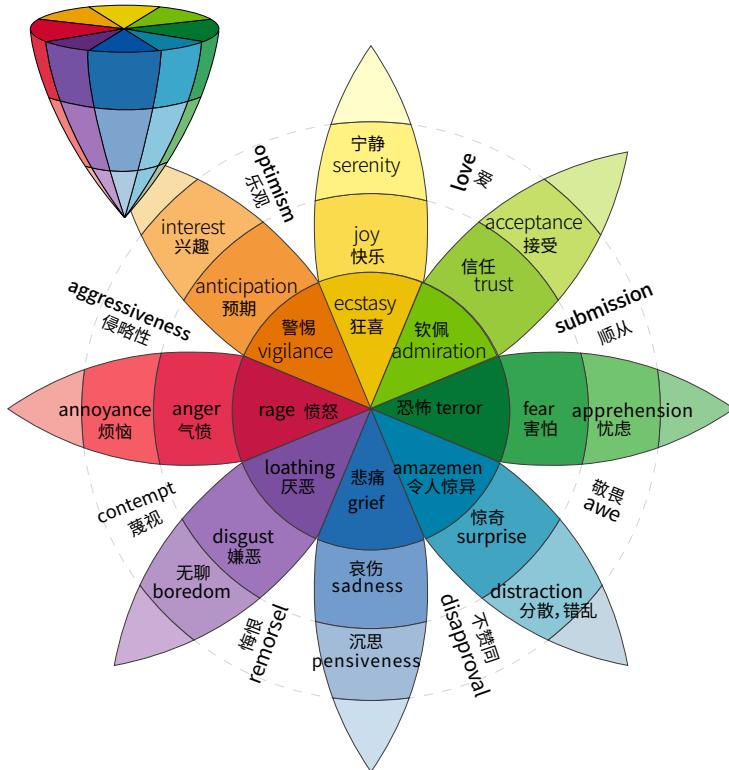
## 2. 情绪模型

情绪 (Emotion) 是人们对外界刺激所产生的反映，由多种感觉、思想以及行为综合产生的生理和心理状态。包括喜爱、欢乐、悲伤等。由于自然语言的复杂性和人类情绪的多边形，不同领域的研究学者对情绪类别的划分也有不同。我国古代《礼记》中就对人的情绪进行了“七情”划分：喜、怒、哀、惧、爱、恶、欲。心理学的理论学者 Parrott 把情绪进行了更细粒度的划分，不仅仅给出了基本情绪，还给出了二级乃至三级等更细粒度的情绪<sup>[454]</sup> (如表9.1所示)。

2001 年，心理学家 Plutchik 基于进化规则的综合理论，提出了多维度情绪模型<sup>[455]</sup>。该模型定义了 8 种基本基本情绪，分为 4 对双向组合：高兴与悲伤、愤怒与恐惧、信任与厌恶、诧异与期望 (Joy vs. Sadness, Anger vs. Fear, Trust vs. Disgust, Surprise vs. Anticipation)。图9.1给出了 Plutchik 模型的情绪类别在“情绪轮”上的排序，其中颜色深浅代表这种情绪的饱和度，离圆心的远近代表情绪的强度。在 Plutchik 情绪理论中，每种情绪都可以分为 3 度，比如，宁静 (Serenity) 是最小程度的高兴，是不饱和状态；狂喜 (Ecstasy) 是最高程度的高兴，是饱和状态。此外，Plutchik 还提出一种假设，认为两种相邻的基本情绪可以组合成一种符合情绪，比如，快乐 (Joy) + 预期 (Anticipation) = 乐观 (Optimism)。

基本情绪	二级情绪类型	三级情绪类型
Anger	Disgust	Contempt, loathing, revulsion
	Envy	Jealousy
	Exasperation	Frustration
	Irritability	Aggravation, agitation, annoyance, crosspatch, grouchy, grumpy
	Rage	Anger, bitter, dislike, ferocity, fury, haterd, hostility, outrage, resentment, scorn, spite, vengefulness, wrath
Fear	Torment	Torment
	Horror	Alarm, fear, fright, horror, hysteria, mortification, panic, shock, terror
Joy	Nervousness	Anxiety, apprehension(fear), distress, dread, suspense, uneasiness, worry
	Cheerfulness	Amusement, bliss, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Contentment	Pleasure
	Enthrallment	Enthrallment, rapture
	Optimism	Eagerness, hope
	Pride	Triumph
	Relief	Relief
	Zest	Enthusiasm, excitement, exhilaration, thrill, zeal
	Affection	Adoration, attractiveness, caring, compassion, fondness, liking, sentimentality, tenderness
Love	Longing	Longing
	Lust/sexual desire	Desire, infatuation, passion
Sadness	Disappointment	Dismay, displeasure
	Neglect	Alienation, defeatism, dejection, embarrassment, homesickness, humiliation, insecurity, insult, isolation, loneliness, rejection
	Sadness	Depression, despair, gloom, glumness, grief, melancholy, misery, sorrow, unhappy, woe
	Shame	Guilt, regret, remorse
	Suffering	Agony, anguish, hurt
Surprise	Sympathy	Pity, sympathy
	Surprise	Amazement, astonishment

表 9.1 基本情绪类型、二级情绪和三级情绪<sup>[454]</sup>.

图 9.1 Plutchik 提出的情绪轮<sup>[455]</sup>

除了上述基于类别空间的情绪模型外，还有一些情绪模型是基于维度空间思想。这类情绪模型认为情绪并不是相互分离和独立的，而是相互联系和交叉的，某种特定的情绪可以表示为连续维度空间中的点或向量。Mehrabian 与 Russell 提出了 PAD 三维模型（Pleasure-Arousal-Dominance）<sup>[456]</sup>，提出情绪可以从愉悦度（Pleasure）、唤醒度（Arousal）和支配度（Dominance）进行分解。此后，Russell 进一步对 PAD 理论进行了修证<sup>[457]</sup>，提出愉悦度和唤醒两个维度就能够解释绝大部分的情绪，支配度更多与认知活动有关，因此就有了由愉悦度和唤醒度构成情绪环结构。图9.2给出了根据 PAD 理论和 PA 理论表示的情绪。

### 9.1.2 情感分析主要任务

情感分析包含非常多各种类型任务，从任务的类型层面可以划分为情感分类和情感信息抽取两大类。从语言单元层面又可以划分为篇章级情感分析（Document-Level Sentiment Analysis），句子级情感分析（Sentence-Level Sentiment Analysis）和属性级情感分析（又称方面级情感分析，Aspect-Level Sentiment Analysis）<sup>[458, 459]</sup>。本节中将按照情感分类和情感信息抽取分别介绍情感分析主要

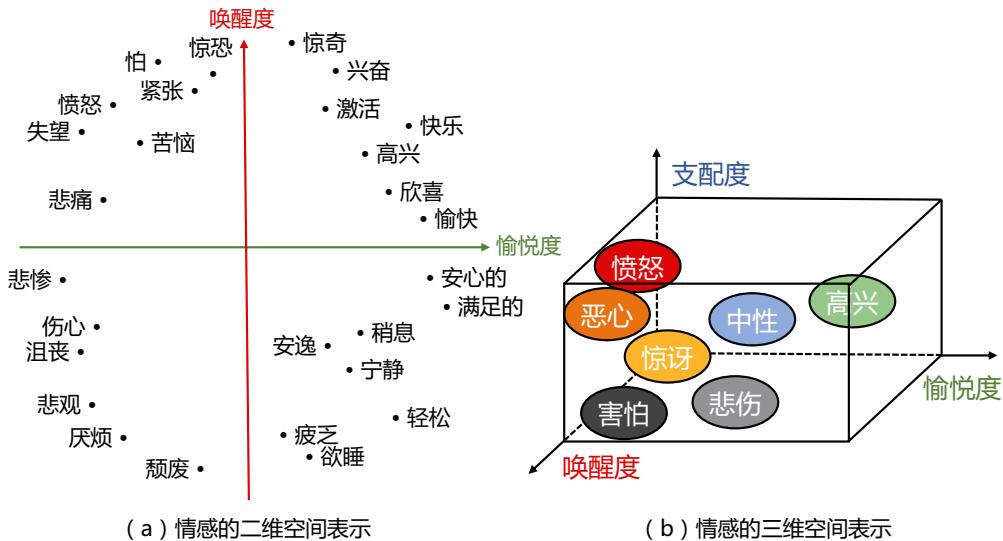


图 9.2 基于维度空间表示方法

任务，并对情感摘要、虚假观点检测等情感相关任务进行简要介绍。

### 1. 情感分类任务

情感分类任务（Sentiment Classification）的目标是根据给定文本内容，识别所蕴含的情感或观点，并确定情感的类别或观点倾向性。主要任务包括主客观识别、情感极性判断、情感强度判断、情绪分类等。

主观分类（Subjective Classification），又称观点识别（Opinion Identification），目标是判断给定的文本是否包含情感或观点，即判断文本是主观性（Subjective）还是客观性（Objective）。

例如：(1) 味道不错！团购很实惠。

(2) 复旦大学校名取自《尚书大传》之“日月光华，旦复旦兮”。

上例中，句子(1)是主观性句子，表达了用户的观点；句子(2)是客观性句子，没有表达任何观点或态度。

极性分类（Polarity Classification）目标是判断给定的文本情感或观点的情感极性，即判断文本的情感是正面（褒义）、负面（贬义）还是中性。

例如：(1) 环境相当不错，业务水平很专业。

(2) 实在是很坑的一个景区。

(3) 地理位置也还可以。

上例中，句子(1)是正面评价；句子(2)是负面评价；句子(3)是中性评价。需要特别说明的是中性评价不等同于客观性文本，只是句子所表达的情感极性并不能归于正面和负面。

情绪分类 (Emotion Classification) 目标是判断给定的文本蕴含的情绪类别，即判断文本中情绪是喜、怒、哀、恶等。情绪类别需要根据在上节中介绍的情绪模型进行选择。

例如：(1) 我的心里绽开了朵朵鲜花，就要蹦出来似的。

(2) 钟表，可以回到起点，却已不是昨天。

上例中，句子(1)是表达了快乐的情绪；句子(2)表达了悲伤的情绪。

情感强度判别 (Sentiment Strength Detection) 目标是判断给定文本的情感强度，即判断文本的情感是强烈正向、正向、中性、负向、强烈负向等。根据上节中介绍的观点模型，也可以采用分数表示情感强度。

例如：(1) 这地方交通不太方便了。

(2) 这地方交通实在是太不方便了。

上例中，句子(1)表达了负面的情感；句子(2)表达了强烈的负面情感。

根据语言单元粒度不同，上述情感分类任务可以划分为篇章级、句子级和属性级，即篇章级主客观分类、句子级情绪分类、属性级情感强度判别等。篇章级和句子级的主要不同在于所处理的文本粒度不同。而属性级的情感分类任务所处理的文本可能是篇章或者句子，但是目标并不是判断整个句子或者篇章整体的情感，而是判断文本内容中关于该属性的情感类别或情感极性。因此，属性级的情感分类任务，不仅要输入文本内容，还要输入所关注的属性单元。

## 2. 情感信息抽取任务

情感信息抽取 (Sentiment Information Extraction)，也称评价要素抽取 (Opinion Elements Extraction)，目标是抽取文本中的表达情感的核心要素，如评价词、评价对象、观点持有者、评价搭配等。相较于情感分类任务，情感信息抽取可以获得结构化的情感信息。情感信息抽取任务也与属性级情感分类任务有紧密联系，可以在情感抽取的基础上完成属性级观点识别等任务，利用抽取出来的评价词以及评价搭配使得属性级观点识别具有一定的可解释性。

例如：懂车会：车窗 采用无边框玻璃设计，很酷，但吹毛求疵一点，隔音 不算太好。  
 观点持有者 评价对象 评价词 评价对象 评价词

评价对象抽取 (opinion target extraction) 目标是抽取文本中的被评价对象的主体。上例中，“车窗”和“隔音”都属于评价对象。

评价词抽取 (Opinion Word Extraction) 目标是抽取文本中所使用的评价词。上例中，“很酷”和“不算太好”都属于评价词。

评价搭配抽取 (Opinion Collocation Extraction) 目标是识别文本中评价对象所对应的评价词。可以使用二元组表示<评价对象，评价词>。上例中，<车窗，很酷>以及<隔音，不算太好>都属于评价搭配。

评价搭配极性判别 (Opinion Collocation Polarity Classification) 目标是判断某个评价搭配的情感极性。上例中，“车窗”与“很酷”是正面评价，“隔音”与“不算太好”是负面批评。

观点持有者抽取 (Opinion Holder Extraction) 目标是抽取文本中观点的持有者。上例中，“懂车

会”作为作者给出了上述评价，属于观点持有者。大部分情况下观点持有者是文章的作者，不一定体现在当前的文本中。还有一些文章中引用了他人的评价，观点持有者就应该是所引用的内容的作者。

### 3. 情感相关其他任务

除了情感分类任务和情感信息抽取任务之外，还有一些任务与情感和观点相关，主要包括：观点摘要、辩论立场检测、评论质量判断、虚假观点检测等。

**观点摘要**（Opinion Summarization）是以评价对象以及以针对评价对象的观点为中心进行的单文档或多文档摘要。观点摘要不仅可以与普通文本摘要类似，由抽取的重要句子或生成的文本构成摘要，还可以采用情感信息抽取获得的结构化数据进行综合后可视化的形式。图9.3给出了基于属性的观点摘要样例<sup>[460]</sup>，展示了两个数码相机在各属性上的评价对比。

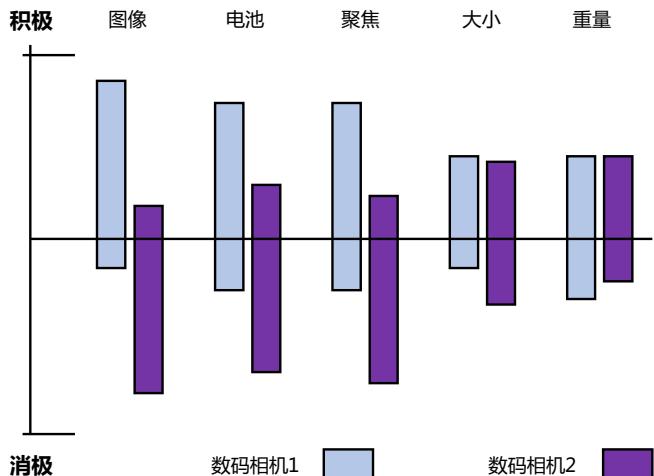


图 9.3 基于属性的观点摘要样例<sup>[460]</sup>

社交媒体中不仅包含用户发表的评论信息，还包含大量的参与者之间存在交互的辩论、讨论或评论。辩论立场检测（Stance Detection）目标是从辩论文本中识别用户对于某个辩论主题的立场，即用户是支持这个主题还是反对这个主题。这项工作与属性级关系分析较为类似，但是不同是产品的属性通常是预先定义的有限集合，而论辩主题多种多样，通常无法预先定义。

例如：“爆竹声中一岁除，春风送暖入屠苏。”春节燃放烟花爆竹，自古以来就是中国人的传统习俗。不过近几年来，随着“双禁”的规定越发严格，过年没有鞭炮声，过年越来越没有年味儿了。

上例中，用户表达了对“春节不应该放鞭炮”议题反对的立场。

随着用户对社交媒体中的评论的重视，一些组织和个人在平台中发布虚假评论（Fake Review），

以达到抹黑某个产品或者宣传某个产品的目的。这些虚假评论也称为垃圾评论（Spam Review）。如果不能甄别这些垃圾评论，会使得社会媒体充满虚假和谎言，对于平台公信力造成了重大影响。垃圾评论检测（Spam Review Detection）目标就是根据评论内容和评论发布者的行为等信息识别虚假信息。不同于其他垃圾信息识别，虚假评论仅从内容层面非常难进行判断，需要结合用户个体和群体行为才能进行有效的判别。此外，构造虚假评论的标准测试集合也是非常困难的工作，这也使得算法之间不能有效地进行对比。

情感分析相关内容还有很多，受篇幅限制本书就不在详述，对情感分析有更多兴趣的读者可以参考刘兵教授关于情感分析任务的专著《Sentiment Analysis: mining sentiments, opinions, and emotions》<sup>[451]</sup>。

## 9.2 篇章级情感分析

篇章级情感分析（Document-level Sentiment Analysis），也称为文档级情感分析，主要任务包括篇章级主客观分类、篇章级极性分类、篇章级情绪分类以及篇章级情感强度判断等情感分类任务<sup>[461–463]</sup>。相关分类的目标可以根据本章情感分析概述部分所介绍的观点模型和情绪模型决定。但是，无论分类目标如何，篇章级情感分析任务的可以定义为：对于给定的篇章，预测其所对应的情感标签。

篇章级情感分析将整篇文档看做一个整体，通常假设整个文档仅对一个实体进行评论，并不对文档中具体的实体或属性进行细粒度分析。这种假设显然与实际情况并不总是相符的。这不仅制约了篇章级情感分析任务的应用场景，也对模型提出了很大的挑战。篇章中不仅包含主观性的评论内容，也包含客观性的事实描述。比如，电影评论中很可能包含一些对电影情节的描述内容，而这些细节描述中也可能包含大量的情感用语。此外，篇章中除了包含对主要实体或主题的整体评论描述外，也很可能包含对其属性的评价以及对其他相关事物的评价。这些问题都对篇章级情感分析模型造成了很大的挑战，如何在没有句子级标注的情况下区分主观性句还是客观性句？如何区分篇章所讨论的主要实体与次要实体？这些都是篇章级情感分析算法需要解决的问题。

本节主要介绍两类方法：基于支持向量机的篇章级情感分析和基于层次结构的篇章级情感分析。最后，将介绍一些常见的篇章级情感分析数据集。

### 9.2.1 基于支持向量机的篇章级情感分析

如前所述，篇章级情感分析通常转换为文本分类问题，因此绝大多数有监督机器学习方法都可以应用于该任务。基于特征的统计机器学习方法，包括朴素贝叶斯、最大熵、支持向量机（Support Vector Machine, SVM）等也都适用于该任务。本节以支持向量机为例，介绍如何利用基于特征的统计学习方法进行篇章级情感分析。

对于给定数据集  $\mathbb{D} = \{(d_1, y_1), (d_2, y_2), \dots, (d_m, y_m)\}$ ， $d_i$  表示篇章内容， $y_i \in \{-1, +1\}$  是分类标签。对于情感极性分析，我们可以定义  $y_i = +1$  表示褒义， $y_i = -1$  表示贬义。针对主客观分析，可以定义  $y_i = +1$  表示主观， $y_i = -1$  表示客观。与其他基于传统机器学习算法的方法一样，这

里也需要首先将文档  $d_i$  转换为特征表示，在此基础上利用机器学习算法进行分类。利用 SVM 算法对文本进行分类，需要将特征转换为向量表示。最基本的方法是采用词袋模型（Bag-of-Words）。给定词典  $V$ ，该词典中包含了  $|V|$  个单词，那么可以用长度为  $|V|$  的向量表示文本。向量的每一维表示在文本中某个单词是否出现，或者采用 TF-IDF 等方法计算得到的该词权重。文献 [445] 中采用了这种方式分类电影评论取得不错的效果。

在此基础上，可以将更多与情感分类相关的特征加入向量表示，下面列出了一些应用于情感分析任务的常见特征：

- 词性：每一个词的词性特征。形容词和副词是观点和情感的主要承载词，因此可以通过词性信息的引入使得模型可以利用该特征给相关词语更高的权重。
- 情感词和情感短语：在语言中表达了积极或者消极情感的词语。例如：好、很棒等是褒义词，坏、糟糕等是贬义词。可以利用情感词典将此类单词识别后加入特征。
- 观点的规则：文本结构和语言成分可以表示隐含情感和观点。可以通过人工设计的规则抽取此类特征加入向量表示中。
- 情感转置词：可以反转文本中情感倾向的词语或短语。比如否定词可以把正面的情感倾向改变为负面的情感倾向。这类词语可以单独加入提取后加入到向量表示中。
- 句法分析树：通过句法分析获得的句法分析树或者子树。可以通过将句法路径、句法树片段使用类似 n-gram 的方法转换为向量加入向量表示。此外，SVM 中树核（Tree Kernel）方法也可以直接在树结构上进行计算。Wu 等人还提出了针对倾向性分析的树核方法<sup>[464]</sup>。

利用上述方法将文档  $d_i$  转换为对应的特征表示  $\mathbf{x}_i$  后，可以采用 SVM 方法构建分类算法。SVM 的基本想法是求解能够且几何间隔最大的分离超平面将不同类别的样本分开（如图9.4所示）。 $wx + b = 0$  即为分离超平面，距离超平面最近的样本点使公式  $wx + b \geq +1$  或  $wx + b \leq -1$  等号成立，这些样本点称为“支持向量”（Support Vector）。两个类别的支持向量到超平面的距离之和为  $\gamma = \frac{2}{\|w\|}$ ，称为间隔（Margin）。寻找具有最大间隔（Maximum Margin）划分超平面，也就是寻找满足式中约束的参数  $w$  和  $b$ ，使得间隔最大化。这也等同于最小化  $\frac{1}{\gamma} = \frac{1}{2}\|w\|^2$ 。整体优化目标可以形式化的表示为：

$$\begin{aligned} & \min_{w,b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } & y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \dots, m \end{aligned} \tag{9.1}$$

SVM 的基本型是一个凸二次规划（Convex Quadratic Programming）问题，能直接用现成的优化计算方法求解。但 SVM 优化一般使用更高效的拉格朗日乘子法解决。利用拉格朗日乘子法得到其对偶问题（Dual Problem），将有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数：

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i (y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1) \tag{9.2}$$

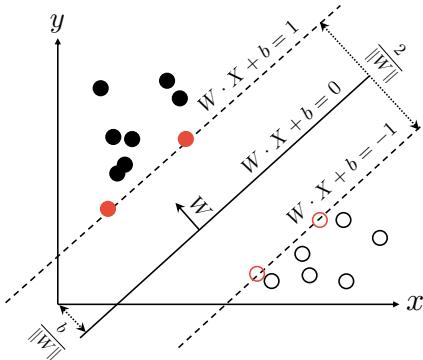


图 9.4 SVM 分类器目标

其中  $\alpha_i$  为拉格朗日乘子, 且  $\alpha_i \geq 0$ 。

为了得到求解对偶问题的具体形式, 令  $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$  对  $\mathbf{w}$  和  $b$  的偏导为 0, 可得:

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (9.3)$$

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad (9.4)$$

将以上两个等式带入拉格朗日目标函数, 消去  $\mathbf{w}$  和  $b$ , 可以得到:

$$\min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i \quad (9.5)$$

求解  $\min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$  对  $\boldsymbol{\alpha}$  的极大值, 即可得到原问题的对偶问题:

$$\begin{aligned} & \max_{\boldsymbol{\alpha}} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ & \text{s.t. } \sum_{i=1}^m \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned} \quad (9.6)$$

利用序列最小优化 (Sequential Minimal Optimization, SMO) 算法求解  $\boldsymbol{\alpha}$ , 就可以求解出  $w$  和  $b$ , 进而求得决策平面。

以上都是基于训练集数据线性可分的假设下进行的, 但是实际情况下几乎不存在完全线性可分的数据, 为了解决这个问题, 还可以进一步引入软间隔 (Soft-margin) 概念, 即允许某些点不满

足约束:

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad (9.7)$$

采用 Hinge 损失, 可以将原优化问题改写为:

$$\begin{aligned} & \min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t. } & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, N \end{aligned} \quad (9.8)$$

其中  $\xi_i$  为松弛变量,  $\xi_i = \max(0, 1 - y_i (\mathbf{w} \cdot \mathbf{x}_i + b))$  为一个 Hinge 损失函数, 每一个样本都有一个对应的松弛变量, 表征该样本不满足约束的程度。 $C > 0$  称为惩罚参数,  $C$  值越大, 对不可分的惩罚越大。与线性可分求解的思路一致, 同样这里先用拉格朗日乘子法得到拉格朗日函数, 再求其对偶问题。

基于上述方法可以完成篇章级主客观分类、篇章级极性分类、篇章级情绪分类, 篇章级情感强度判断等任务<sup>[465, 466]</sup>。情感强度判断任务中也可以采用线性回归 (Linear Regression)、支持向量回归 (Support Vector Regression)<sup>[467]</sup> 等方法将分数间的差距问题加入目标损失函数。

### 9.2.2 基于层次结构的篇章级情感分析

篇章具有层次结构, 由单词组成句子, 再由句子构成篇章。基于这种层次结构, 文献 [468] 中提出了层次注意力网络模型 (Hierarchical attention networks, HAN) 用于篇章级别情感分析方法。HAN 算法的神经网络结构如图9.5所示, 主要包含四个部分: 单词序列编码, 单词级别注意力, 句子序列编码, 句子级别注意力。

单词序列编码层主要用来建模单词序列, 采用了门控循环单元 (Gated Recurrent Unit, GRU) 用来跟踪序列的状态。该模型基于门机制, 不再使用分开的记忆单元。而使用重置门  $r_t$  和更新门  $z_t$  用来控制状态中信息的更新。更新门  $z_t$  决定多少信息保留以及多少新的信息加入, 重置门  $r_t$  控制以前状态对于当前状态的影响。在  $t$  时刻, GRU 通过如下公式计算新的状态:

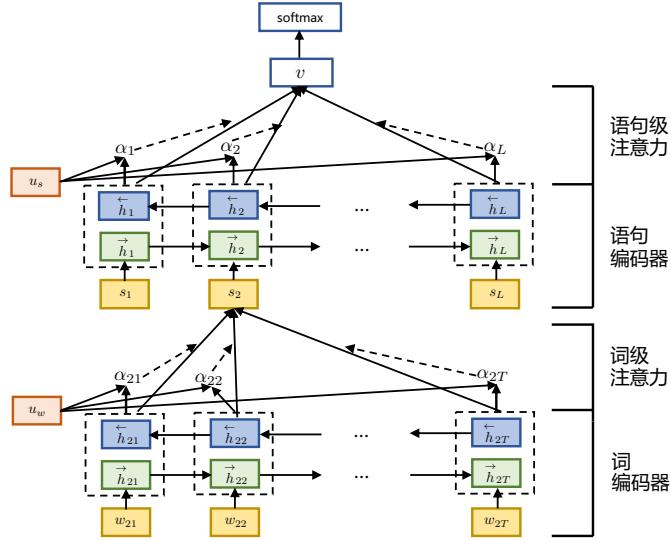
$$\mathbf{h}_t = (1 - z_t) \odot \mathbf{h}_{t-1} + z_t \odot \tilde{\mathbf{h}}_t \quad (9.9)$$

$$z_t = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (9.10)$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h \mathbf{x}_t + r_t \odot (\mathbf{U}_h \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (9.11)$$

$$r_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (9.12)$$

其中  $x_t$  表示序列中  $t$  时刻的输入向量,  $\mathbf{W}_z, \mathbf{W}_h, \mathbf{W}_r, \mathbf{U}_z, \mathbf{U}_h, \mathbf{U}_r, \mathbf{b}_z, \mathbf{b}_h, \mathbf{b}_r$  是需要学习的参数。 $t$  时刻的状态使用  $\mathbf{h}_t = \text{GRU}(\mathbf{x}_t)$  表示。

图 9.5 基于层次结构的篇章级情感分析框架<sup>[468]</sup>

HAN 模型中使用双向的 GRU 模型来学习篇章中第  $i$  个句子  $s_i$  的第  $t$  个单词  $w_{it}$  的表示，分别从左至右和从右至左建模：

$$\begin{aligned}\vec{h}_{it} &= \overrightarrow{\text{GRU}}(\mathbf{x}_{it}), t \in [1, T] \\ \overleftarrow{h}_{it} &= \overleftarrow{\text{GRU}}(\mathbf{x}_{it}), t \in [T, 1]\end{aligned}\quad (9.13)$$

其中  $\mathbf{x}_{it}$  表示单词  $w_{it}$  的词向量表示， $T$  表示句子  $s_i$  的单词数。最终，正向隐层表示和反向隐层表示拼接  $\mathbf{h}_{it} = [\vec{h}_{it}; \overleftarrow{h}_{it}]$  作为单词  $w_{it}$  的隐层表示。

由于，并不是句子中所有的单词都对句子的表示都同等重要。因此，HAN 模型中提出使用单词级别的注意力机制来捕获句子中比较重要的单词，得到加权求和后的句子表示  $s_i$ ：

$$\mathbf{u}_{it} = \tanh(\mathbf{W}_w \mathbf{h}_{it} + \mathbf{b}_w) \quad (9.14)$$

$$\alpha_{it} = \frac{\exp(\mathbf{u}_{it}^\top \mathbf{u}_w)}{\sum_t \exp(\mathbf{u}_{it}^\top \mathbf{u}_w)} \quad (9.15)$$

$$s_i = \sum_t \alpha_{it} \mathbf{h}_{it} \quad (9.16)$$

其中  $\mathbf{W}_w, \mathbf{b}_w, \mathbf{u}_w$  为待学习的参数。

句子序列编码层和单词序列编码一样，也同样使用双向的 GRU 模型来编码句子序列，

$$\begin{aligned}\vec{\mathbf{h}}_i &= \overrightarrow{\text{GRU}}(\mathbf{s}_i), i \in [1, |d|] \\ \overleftarrow{\mathbf{h}}_i &= \overleftarrow{\text{GRU}}(\mathbf{s}_i), t \in [|d|, 1]\end{aligned}\quad (9.17)$$

其中  $|d|$  表示篇章中句子的个数。同样的，通过对于正向和反向句子的拼接，得到最后的句子隐层表示  $\mathbf{h}_i = [\vec{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$ 。

同样的，在篇章每一个句子也并不是同等重要。与单词级别注意力类似，使用句子级别的注意力来捕获篇章中比较重要的句子，得到加权求和后的篇章表示  $\mathbf{v}$ ：

$$\mathbf{u}_i = \tanh(\mathbf{W}_s \mathbf{h}_i + \mathbf{b}_s) \quad (9.18)$$

$$\alpha_i = \frac{\exp(\mathbf{u}_i^\top \mathbf{u}_s)}{\sum_i \exp(\mathbf{u}_i^\top \mathbf{u}_s)} \quad (9.19)$$

$$\mathbf{v} = \sum_i \alpha_i \mathbf{h}_i \quad (9.20)$$

其中  $\mathbf{W}_s, \mathbf{b}_s, \mathbf{u}_s$  为待学习参数。

最后，通过句子级注意力得到的篇章表示  $\mathbf{v}$  可以被输入到一个线性模型和 Softmax 函数中用于情感分类预测。

### 9.2.3 篇章级情感分析语料库

篇章级情感分析通常也依赖大规模的标注语料对模型进行训练和评测。伴随着篇章级情感分析任务的发展，大量篇章级情感分析语料库也相应的提出。本节将介绍几种常见的包含情感标签的篇章级情感分析语料库。如表9.2所示，主要包含 4 个英文数据集和 2 个中文数据集。

表 9.2 常见文档级情感分析语料库汇总

语料库名称	训练集合	验证集	测试集	总共	类别	语言
大型电影评论数据集	25,000	-	25,000	50,000	2	英文
IMDB	108,535	13,567	13,567	135,669	10	英文
Yelp-酒店	20,975	6,993	6,993	34,961	5	英文
Yelp-餐厅	106,943	35,648	35,648	178,239	5	英文
Amazon	59,399	11,880	11,880	83,159	5	英文
中文情感语料	-	-	-	1201	2	中文
点评	503,330	83,889	83,889	671,108	5	中文

## 1. 电影评论数据集

大型电影评论数据集 (Large Movie Review Dataset)<sup>[469]</sup> 是由斯坦福人工智能实验室 (Stanford Artificial Intelligence Laboratory, SAIL) 于 2011 年推出的一个电影评论的英文数据集，该数据集的训练集和测试集分别包含 25,000 条标注过的电影评论。每一个评论被标记为负面和正面两种类型。数据集中还额外提供了 50,000 条未标注数据。

## 2. IMDB 情感分析数据集

IMDB 情感分析数据集<sup>[470]</sup> 是由卡内基梅隆大学机器学习实验室 (Machine Learning Department, Carnegie Mellon University) 和新加坡管理大学于 2014 年联合发布的一个电影评论英文数据集。由于该数据集是篇章级别的数据集，每一个样本平均包含 393.8 个单词，14.4 个句子。每一个评论文档被标注 1 到 10 的一个情感极性。该数据集一共包含 135669 个样本，通过 8:1:1 比例随机划分为训练数据、验证数据和测试数据。

## 3. Yelp 酒店餐厅点评数据集

Yelp 酒店餐厅点评数据集<sup>[471]</sup> 是由 University of Colorado Boulder 于 2019 年构建的英文数据集数据集。该数据集包含酒店和餐厅两个领域的评论数据。其中酒店领域包含 20,975 条训练样本、6,993 验证样本和 6,993 条测试样本；餐厅领域包含 106,943 条训练样本、35,648 条验证样本和 35,648 条测试样本。该数据集中每一条评论包含了一个 1-5 的情感打分。对于打分小于 3 的属于消极，等于 3 的为中性，而大于 3 的积极。

## 4. Amazon 产品评论数据集

Amazon 产品评论数据集<sup>[471]</sup> 也是是由 University of Colorado Boulder 于 2019 年构建的英文数据集数据集，包含亚马逊上关于音乐的评论数据。该数据规模较大，一共包含 135,669 条样本，其中训练样集、验证集和测试集分别包含 108,535、13,567 和 13,567 条样本。和 Yelp 数据类似，每一个评论都有一个 5 分制的打分，表示用户的喜欢程度，值越大对应的情感越积极。

## 5. 中文情感语料

中文情感语料 (Chinese sentiment corpus, ChnSentiCorp)<sup>[472]</sup> 是由中国科学院智能软件实验室于 2008 年构建的篇章级情感分析数据集。该数据集一共包含 1021 个文档，其中教育相关文档 507 个，电影相关文档 266 个和房子相关 248 个。每一个文档被分为积极和消极两种。

## 6. 点评网数据集

点评网数据集 (Dianping)<sup>[473]</sup> 是一个美团大众点评于 2021 年公布的大型评论数据集。该数据集包含了属性级情感分析和评分预测两个部分。对于评分预测任务，每一个评论包含一个 1 到 5 的打分，表示用户对于商品的满意程度。该数据集规模较大，一共包含大约 67 万个样本，其中训练样本 503,330 个，验证样本 93,889 个以及测试样本 93,889 个。

## 9.3 句子级情感分析

句子级情感分析（Sentence-level Sentiment Analysis）主要任务包括句子级主客观分类、句子级极性分类、句子级情绪分类以及句子级情感强度判断等情感分类任务<sup>[474, 475]</sup>。给定一个句子  $s$ ，句子级情感分析旨在预测句子对应的情感标签  $y$ 。与篇章级情感分析类似， $y$  一般指积极、消极或者中性标签。也可以采用连续分数的形式对句子的情感程度进行评分。与篇章级情感分析任务类似，句子级情感分析虽然包含很多任务，但是所采用的方法非常类似。

相较于篇章级情感分析，句子级情感分析则更精细化一些，将每个句子看做一个整体。通常情况下一个句子级仅对一个实体进行评论。相较于篇章级情感分析中需要对主次实体进行区分的难题，由于该现象在单个句子中出现次数较少，因此句子级情感分析中该问题造成影响相对并不严重。但是，句子级情感分析还要面临新的难题，主要包含以下几个方面：1) 句子长度短；2) 隐式情感表达占比高；3) 条件、对比等复杂情况多。句子中包含的单词数量少很多，还包含反讽、双重否定等复杂语言现象，这就要求模型必须具备充分利用句子中有效的语义信息。隐式观点表达相较于包含明确情感词的显式观点表达的识别难度大很多，很多情况下还需要基于尝试才能对情感极性进行判断，而隐式观点在句子级情感分析中的占比很高，根据在 SemEval-2014 语料集中占比达到 30% 左右<sup>[476]</sup>，这也对句子级情感分析带来了很大的困难。此外，句子级情感分析主要关注一般性的分类问题，但是对于条件、对比等复杂观点句型缺乏处理能力。

本节主要介绍句子级情感分析中一些常见方法，包括基于无监督方法的句子级情感分析、基于递归神经网络的句子级情感分析、基于预训练的句子级情感分析。最后，对句子级情感分析的常用的数据集进行说明。

### 9.3.1 基于词典的句子级情感分析

基于词典的情感分类方法<sup>[477, 478]</sup>是情感分析算法中经典的算法类型之一，由于该类算法不需要或者仅需要非常少的训练语料，因此也具有很广泛的应用场景。SO-CAL (Semantic Orientation CALculator)<sup>[477, 478]</sup>是利用情感词典对文本的情感极性进行分析的方法，其核心是基于对文档中的每个单词或短语根据词典和规则进行评分，在此基础上综合得到文档情感倾向性极性和强度。SO-CAL 采用了情感字典、情感强化和情感否定等三个方面构建了词典和规则。

情感字典包含形容词（形容词短语）、名词、动词和副词。SO-CAL 所采用的情感字典包含 2,252 个形容词，1,142 个名词，903 个动词和 745 个副词。对于表达正面情感的词和短语给了一个正的值，而对于表达负面的词给一个负的值。具体地，每一个情感词被赋予  $-5$ （极度否定）到  $+5$ （极度肯定）的一个值。

例如：monstrosity -5	hate (noun and verb) -4
disgust -3	sham -3
fabricate -2	delay (noun and verb) -1
determination 1	inspire 2

inspiration 2	endear 3
relish (verb) 4	masterpiece 5

情感强化归纳了情感加强词（比如：非常、很）会增强相邻情感词的语义强度，以及情感弱化词（比如稍微、有点）会减弱相邻情感词的语义强度。文本的情感得分需要结合情感加强词来进行计算。对于加强词和减弱词，给一个或正或负的权重百分比。例如，slightly 是 -50, somewhat 是 -30, pretty 是 -10, really 是 +15, very 是 +25 等。如果 excellent 的情感值是 5，则 very excellent 的情感值为  $5 * (100\% + 25\%)$ 。除了副词和形容词外，还有其他的词性的加强词和减弱词，比如数量词、全部字母大写、感叹号标记、以及语篇连接词 but。

情感否定包含会反转情感词对应的情感极性词汇。在情感分析任务中具有重要作用。情感否定词包括 not、none、nobody、never、nothing 以及其他的一些单词，如 without 或者 lack。然而，简单的反转情感值会存在一定问题。比如把 excellent 的情感值 +5 通过 not 反转到 -5，并不合理。not excellent 往往比 not good 更加偏向于正向一些。这里采用与原有词语相反方向的固定值相加计算的方式。

例如：这个服务并不是特别好。

特别好：+5，并不：-4，句子整体： $+5-4=1$

我觉得这个菜的味道这并不差

差：-3，并不：+4，句子整体： $-3+4=1$

最终，对文本中所有的情感表达值进行求和，若求和为正，则该文本被判定为积极情感，若求和为负，则文本表达一种负面的情感，求和为零，则表示中性情感。例如句子：“这个服务并不是特别好 ( $+5-4=1$ )，但是也并不差 ( $-3+4=1$ )”，求和为 0，整体表达的中性的情感。

### 9.3.2 基于递归神经张量网络的句子级情感分析

句子级情感分析主要是依据句子的语义内容对情感进行分类。递归神经网络 (Recursive Neural Network, RNN) 可以有效利用句法结构，递归的计算句子和短语组合向量表示。因此可以使用递归神经网络进行句子级情感分析。

每个单词使用  $d$  维向量进行表示，词向量中的数值使用均匀分布进行随机初始化，所有的词典  $\mathbb{V}$  的词向量堆叠成一个词嵌入矩阵  $\mathbf{L} \in \mathbb{R}^{d \times |\mathbb{V}|}$ ，与模型一起训练。针对短语“not very good”，其句法结构如图9.6所示。根据所对应的句法结构，可以按照如下公式，自底向上递归计算每个中间节点的向量表示，计算完两个子节点之后再计算其父结点：

$$\mathbf{p}_1 = f \left( \mathbf{W} \begin{bmatrix} b \\ c \end{bmatrix} \right), \mathbf{p}_2 = f \left( \mathbf{W} \begin{bmatrix} a \\ \mathbf{p}_1 \end{bmatrix} \right) \quad (9.21)$$

其中， $f = \tanh$  是一个标准的非线性激活函数， $\mathbf{W} \in \mathbb{R}^{d \times 2d}$  是待学习参数。基于得到的节点的向量表示，可以使用 Softmax 函数构建分类目标学习函数  $g = \text{Softmax}(bmW_s \mathbf{p}_2)$ 。

在此基础上，文献 [479] 提出了基于递归神经张量网络 (Recursive Neural Tensor Network, RNTN)

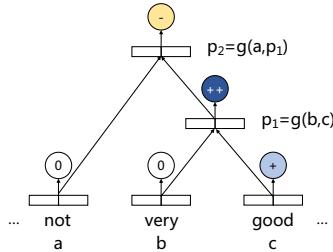


图 9.6 递归神经网络结构图

的方法用于句子级情感分析。RNTN 在原有的单词向量表示基础上，增加了基于张量的组合函数用于描述单词之间的组合关系。定义张量积的输出  $\mathbf{h} \in \mathbb{R}^d$ ，通过以下公式计算得到：

$$\mathbf{h} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}^T \mathbf{V}^{[1:d]} \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \quad (9.22)$$

其中  $\mathbf{V}^{[1:d]} \in \mathbb{R}^{2d \times 2d \times d}$  是定义了多个双线性形式，每个  $\mathbf{V}^{[i]}$  是其中一个切片，使用  $\mathbf{V}^{[i]}$  计算得到  $\mathbf{h}_i$ ：

$$\mathbf{h}_i = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}^T \mathbf{V}^{[i]} \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \quad (9.23)$$

计算过程以  $d = 2$  为例，如图9.7所示。虚线框表示张量的一个切片，用于捕捉到子结点对父节点一种类型的影响。

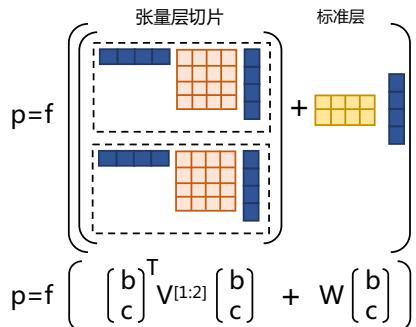


图 9.7 递归神经张量网络单层结构图

在此基础上，图9.6中每个节点的向量表示使用如下公式计算得到：

$$\mathbf{p}_1 = f \left( \left[ \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}^T \mathbf{V}^{[1:d]} \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} + \mathbf{W} \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \right] \right) \quad (9.24)$$

$$\mathbf{p}_2 = f \left( \begin{bmatrix} \mathbf{a} \\ \mathbf{p}_1 \end{bmatrix}^T \mathbf{V}^{[1:d]} \begin{bmatrix} \mathbf{a} \\ \mathbf{p}_1 \end{bmatrix} + \mathbf{W} \begin{bmatrix} \mathbf{a} \\ \mathbf{p}_1 \end{bmatrix} \right) \quad (9.25)$$

根据节点的向量表示  $\mathbf{p}_1, \mathbf{p}_2$ , 仍然可以使用  $g = \text{Softmax}(\mathbf{W}_s \mathbf{p}_2)$  构建分类函数获得句子级情感分析结果。

### 9.3.3 基于情感知识增强预训练的句子级情感分析

通用预训练模型在绝大部分自然语言处理任务上都取得了很好的效果, 基于知识增强的预训练方法进一步提升了预训练模型在一些自然语言处理任务上的效果。文献 [480] 提出了情感知识增强的预训练模型 SKEP (Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis) 方法应用于句子级情感分析。SKEP 算法模型框架如图9.8所示。

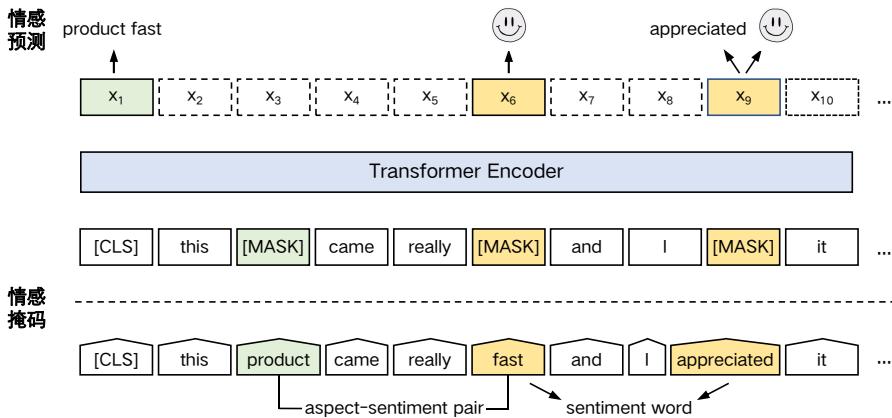


图 9.8 基于情感知识增强预训练的句子级情感分析框架

与 BERT<sup>[29]</sup> 采用的基于掩码语言模型类似, SKEP 方法也希望通过掩盖识别输入序列中的情感信息的方法进行模型预训练。但是大规模的情感信息不能通过人工标注完成, 因此 SKEP 方法中提出了使用基于点互信息 (Pointwise Mutual Information, PMI) 的无监督方法识别情感信息<sup>[481]</sup>, 根据词语的共现信息识别出情感词、情感词的极性和属性词-情感词对等情感信息。掩盖的步骤分为如下三步:

- (1) 掩盖属性词-情感词对: 在句子中随机选择最多两对属性词-情感词对掩盖。
- (2) 掩盖情感词: 在句子中随机选择不超过当前句子总词数 10% 的情感词进行掩盖。
- (3) 掩盖通用字: 如果情感词所占的词比例没有 10%, 随机选择单词补充达到总 10% 的掩盖比例。

SKEP 预训练采用的损失函数  $\mathcal{L}$  由三个部分组成: 情感词预测目标 ( $\mathcal{L}_{sw}$ )、情感词极性预测

目标 ( $\mathcal{L}_{wp}$ ) 和属性词-情感词对预测目标 ( $\mathcal{L}_{ap}$ )。

$$\mathcal{L} = \mathcal{L}_{sw} + \mathcal{L}_{wp} + \mathcal{L}_{ap} \quad (9.26)$$

情感词预测损失  $L_{sw}$  定义为:

$$\begin{aligned} \hat{\mathbf{y}}_i &= \text{Softmax}(\tilde{\mathbf{x}}_i \mathbf{W} + \mathbf{b}) \\ \mathcal{L}_{sw} &= -\sum_{i=1}^{i=n} m_i \times \mathbf{y}_i \log \hat{\mathbf{y}}_i \end{aligned} \quad (9.27)$$

其中,  $\tilde{\mathbf{x}}_i$  是编码器的输出向量,  $\tilde{\mathbf{y}}_i$  是  $\tilde{\mathbf{x}}_i$  经过输出层后, 再经过 Softmax 的得到的概率分布。在得到每个位置的预测结果后, 并不会计算每个词的损失, 而只会计算情感词所在位置的损失, 非情感词的位置不会参与计算。 $m_i$  用于筛选哪些词是情感词。

情感极性词预测损失  $\mathcal{L}_{wp}$  的计算方式和  $\mathcal{L}_{sw}$  类似, 区别在于  $\mathcal{L}_{sw}$  计算的是词的损失,  $\mathcal{L}_{wp}$  计算的是极性的损失。属性词-情感词对预测损失  $\mathcal{L}_{ap}$  则定义为:

$$\begin{aligned} \hat{\mathbf{y}}_a &= \text{Sigmoid}(\tilde{\mathbf{x}}_1 \mathbf{W}_{ap} + \mathbf{b}_{ap}) \\ \mathcal{L}_{ap} &= -\sum_{a=1}^{a=A} \mathbf{y}_a \log \hat{\mathbf{y}}_a \end{aligned} \quad (9.28)$$

其中,  $\tilde{\mathbf{x}}_1$  是 [CLS] 位置的输出向量,  $y_a$  是一个属性词-情感词对,  $\tilde{\mathbf{y}}_a$  是  $y_a$  的概率评估值, 需要注意是这里, 提前构建了一个属性词-情感词对的字典库, 即每一属性词-情感词对都有一个相应的编号表示。

最终模型预训练完成后, 应用于句子级情感分析时, 使用 [CLS] 位置输出代表句子整体表示, 在编码器上增加分类层, 利用句子级情感分析语料进行模型精调。

### 9.3.4 句子级情感分析语料库

大规模标注语料对于句子级情感分析模型的训练具有重要的作用。本节主要介绍三个比较常用的句子级情感分析语料, 包括两个英语数据集, 三个中文数据集, 如表9.3所示。

表 9.3 常用的句子级情感分析语料库

	训练集合	验证集	测试集	合计	类别	语言
斯坦福情感树库	8,544	1,101	2,210		5	英文
GoEmotions 情绪数据集	43,410	5,427	5,426		27	英文
中文情感树库	10,627	665	2,258		2	中文
酒店评论 (HR)	-	-	-	24,348	2	中文
冰原历险记三评论 (IAR)	-	-	-	11,081	2	中文

### 1. 斯坦福情感树库

斯坦福情感树库 (Stanford Sentiment Treebank, SST) [479] 数据集由斯坦福大学自然语言处理组于 2013 年发布，包含 8544 条训练数据，1101 条验证数据，2210 条测试数据。其中每个句子分析树的节点均有细粒度的情感注解。句子和短语总有 239232 条，情感值对应类别：[0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1.0] 分别对应五分类情感。

### 2. GoEmotions 情绪数据集

GoEmotions 情绪数据集<sup>[482]</sup> 数据集是由谷歌于 2020 年提出的情绪标注数据。该数据集是由 58000 条评论组成的人工注释数据集，这些评论来源于流行英语论坛 Reddit 下的各不同板块，并被标记为 27 个情绪类别。GoEmotions 分类系统是迄今为止最庞大的情绪细化英语数据集，包含详细注释。在设计它时，同时考虑了心理学和数据的适用性。该分类系统包括 12 种积极的、11 种消极的、4 种模糊的以及 1 种“中立”的情绪类别，因此它可以广泛适用于需要对情绪表达进行细致区分的对话理解任务。

### 3. 中文情感树库

中文情感树库 (Chinese Sentiment Treebank) [483] 数据集是中国科学院自动化研究所于 2014 年构建的数据集。该数据集基于 2270 电影的豆瓣评论，每一个句子都被打上 0 到 4 这样 5 种打分，0 表示非常消极，4 表示非常积极。该数据集过滤了打分为 2 的中性样本，得到 11439 个积极样本和 2111 个消极样本。最终，这些样本被分为训练集 (10627 条)，验证集 (665 条) 和测试集 (2258 条)。

### 4. 酒店评论和冰原历险记三评论

酒店评论 (Hotel review, HR) 和冰原历险记三评论 (Ice Age III review, IAR) [484] 是由深圳大学于 2015 年提出的中文句子级情感分析数据集。其中 HR 数据集是从携程网站爬取的酒店评论数据，包含 13,446 条积极和 10,902 消极的评论。IAR 数据集是从豆瓣网站爬取的关于冰原历险记三的评论数据，包含 11,081 条积极和 8,869 条消息的评论。

## 9.4 属性级情感分析

产品或实体通常具有一个或多个属性。例如，“这款手机的电池容量非常好”。这里“电池容量”就是手机的一个属性。属性级情感分析 (Aspect-level Sentiment Analysis, ABSA) 目标包含属性级情感分类任务和情感信息抽取任务两大类。属性级情感分类任务包括属性级主客观分类、属性级极性分类、属性级情感强度判断。与篇章级和句子级情感分类不同，属性级情感分类任务的输入不仅包含文本内容  $d$ ，还有目标属性  $a$ ，输出则为文本内容中关于目标属性的评价词  $o$  和情感标签  $y$ 。情感信息抽取任务目标则是抽取文本的中表达情感的核心要素，包括评价词、评价对象、观点持有者、评价搭配等。

相较于篇章级和句子的情感分析，属性级情感分析粒度更小，因此也称为细粒度情感分析（Fine-grained Sentiment Analysis）。属性级情感分类任务有效解决了篇章级和句子级情感分类中只能对整体观点进行分析所存在的问题。在实际评论中，虽然整个篇章或句子对某个实体给出了正面评价，但不代表对每个属性都给出正面评价，反之亦然。属性级情感分析可以有效解决上述问题，在属性层面给出细粒度的分析结果。由于属性级情感分析任务更加精细，因此难度相较于句子级和篇章级情感分析也更大。属性级情感分析不仅要面临句子长度短、隐式情感表达占比高等句子级情感分析所面临的问题。还要处理句子中包含于给定属性相关的内容，也包含与给定属性无关的内容。在无相关标注的情况下，模型也需要具备区分句子中不同部分与给定属性是否相关的能力。

本节主要介绍属性级情感分析所涉及的情感信息抽取和属性级情感分类的常见算法，包括基于句法规则和基于序列标注的情感信息抽取、基于注意力交互和预训练等方法的属性级情感分类算法。最后，对属性级情感分析的常用的数据集进行说明。

### 9.4.1 情感信息抽取

情感信息抽取主要包含属性抽取、观点抽取和情感预测三个子任务。具体来说，给定一个文本序列  $s = w_1 \dots w_{|s|}$ ，情感信息抽取任务目标是抽取文本中包含的所有属性、观点、极性三元组  $T = \{(a_i, o_i, p_i)\}_{i=1}^{|T|}$ ，其中  $|T|$  表示样本中三元组的个数， $a_i$ ， $o_i$  分别是第  $i$  个属性和观点，是文本序列  $X$  的一个子串， $p_i$  是该属性对应的情感极性。

#### 1. 基于句法规则的情感信息抽取

传统基于无监督的情感信息抽取模型主要基于句法信息来抽取文本的属性词和评价词。一般属性词多为名词，而评价词多为形容词，且评价词往往用来修饰属性词。例如，对于句子“华为手机拍出来的照片很好看！”。形容词“好看”直接通过修饰语依赖于名词“照片”。通过句法修饰关系，可以知道“好看”（评价词）直接与“照片”（属性）相联系。由于形容词通常被认为是闭类词，通常比较稳定，一个语言中通常很少增加新的形容词。因此，可以通过构造词典的方法来收集评价词，进而挖掘属性。

文献 [485] 提出通过句法规则来解决评价词字典扩展和属性扩展。通过使用依存句法分析器去扩展评价词字典和挖掘属性，使得评价词和属性建立联系。使用双向循环的方法来使信息在评价词和属性之间不断传播。这种方法的好处就是只需要一个初始的评价词字典即可。评价词字典就是包含许多的评价词的词典，例如 good、excellent、poor 和 bad。但是使用评价词字典的缺陷就在于不可能去囊括所有的意见和领域。并且一个词可能在这个领域是积极的，在另一个领域可能就是中性的。基于句法规则的情感信息抽取方法从已知和通过之间的迭代获得的评价词和属性中，通过识别语义上的联系，提取评价词和属性。这种使得信息在评价词和属性之间来回流动的方法称为双向传播（Double Propagation）。接下来将从关系识别以及评价词字典扩展和属性扩展两个方面来进行算法介绍。

关系识别是指识别评价词与属性(OT-Rel)、属性与属性(TT-Rel)以及评价词与评价词(OO-Rel)

之间的关系，是评价词字典扩展的关键所在。定义了两类单词之间的关系：

- (1) 直接依赖：A 和 B 直接关联（图9.9(a))、A 和 B 通过 H 直接关联（图9.9(b))。
- (2) 间接依赖：A 通过 H1 依赖于 B (图9.9(c))、A 和 B 分别通过 H1 和 H2 依赖于 H (图9.9(d))。

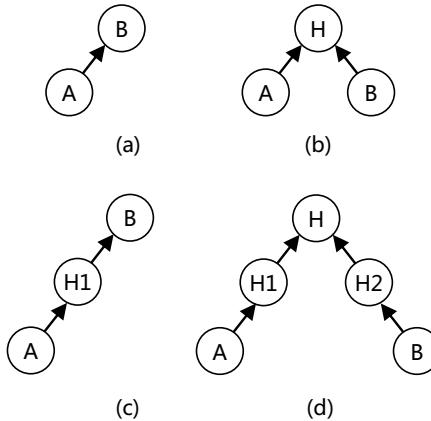


图 9.9 基于规则的属性和评价词抽取

直接依赖和间接依赖都仅考虑了句法树的拓扑结构，为了考虑词性，这里还引入了序列标注任务约束。一般评价词多为形容词，而属性一般为单个名词或名词短语。在文献[485]中使用斯坦福的 POS 工具来抽取词性标注，根据词性分析结果，可能的评价词词性为 JJ (Adjectives)，JJR (Comparative Adjectives) 和 JJS (Superlative Adjectives)。可能的属性词词性为 NN (Singular Nouns) 和 NNS (Plural Nouns)。描述评价词和属性之间关系的依存关系包括 mod (单词与其直接修饰词关系)、pnmod (名词后修饰语)、subj (动词主语)、s (表层主语)、obj (动词宾语)、obj2 (双及物动词的第二个宾语) 和 desc (描述)。OT-Rel、OO-Rel 或者 TT-Rel 可以形式化定义为一个四元组  $(POS(w_i), DT, R, POS(w_j))$ ，其中  $POS(w_i)$  表示单词  $w_i$  的词性，DT 表示依存类型 (例如 DD 或者 IDD)，R 表示句法关系。

评价词字典扩展和属性扩展是指基于预先定义好的规则对评价词字典和属性词进行扩张，主要包含三个部分：基于关系的传播规则定义、基于规则的传播算法、评价词情感极性预测。

**(1) 基于关系的传播规则：** 主要包含对于进行传播的过程中基于评价词抽取属性、基于额外属性抽取属性、基于额外属性抽取评价词、基于给定和额外的评价词抽取评价词四个子任务的规则 (表9.4)。在表中，o/t 表示抽取的评价词/属性。O/T 表示给定或者抽取的评价词/属性集合。H 表示任何单词。POS(O/T) 和 O/T-Dep 分别表示单词 O/T 的 POS 信息和依存关系。JJ 和 NN 是潜在的评价词和属性的 POS 标签集合，其中 JJ 包含 JJ、JJR 和 JJS，NN 包含 NN 和 NNS。MR 是描述评价词和属性关系的依存关系集合，包含 mod、pnmod、subj、s、obj、obj2 和 desc。CONJ 仅仅包含 conj。箭头表示依存关系。例如， $O \rightarrow O\text{-Dep} \rightarrow T$  表示  $O$  通过句法关系 O-Dep 依赖于  $T$ 。“==”

表示相同或者相等(这里相等特指 mod 和 pnmod 一样, s 或者 subj 和 obj 一样)。利用  $R1_i$  基于评价词 O 抽取属性 t,  $R2_i$  基于属性 T 抽取评价词 o,  $R3_i$  基于抽取的属性  $T_i$  抽取属性 t,  $R4_i$  基于已知的评价词  $O_i$  抽取评价词 o。以  $R1_1$  为例, 给定评价词 O, POS 标签为 NN 且满足关系 O-Dep 被抽取为属性词。例如, 对于句子 “The phone has a good screen”, 当我们知道 good 为一个评价词时, 它通过 mod 依赖于 screen, 而 mod 包含在 MR 中且 screen 为 NN,  $R1_1$  会抽取 screen 为属性。

表 9.4 属性和评价词抽取规则(来源: 文献 [485])

RuleID	Observations	Output	Example
$R1_1$	$O \rightarrow O\text{-Dep} \rightarrow T$ s.t. $O \in \{O\}$ , $O\text{-Dep} \in \{MR\}$ , $POS(T) \in \{NN\}$	$t = T$	The phone has a <u>good</u> “screen”. (good $\rightarrow$ mod $\rightarrow$ screen)
$R2_2$	$O \rightarrow O\text{-Dep} \rightarrow H \leftarrow T\text{-Dep}$ $\leftarrow T$ s.t. $O \in \{O\}$ , $O/T\text{-Dep} \in \{MR\}$ , $POS(T) \in \{NN\}$	$t = T$	“iPod” is the <u>best</u> mp3 player. (best $\rightarrow$ mod $\rightarrow$ player $\leftarrow$ subj $\leftarrow$ iPod)
$R2_1$	$O \rightarrow O\text{-Dep} \rightarrow T$ s.t. $T \in \{T\}$ , $O\text{-Dep} \in \{MR\}$ , $POS(O) \in \{JJ\}$	$o = O$	same as $R1_1$ with screen as the known word and good as the extracted word
$R2_2$	$O \rightarrow O\text{-Dep} \rightarrow H \leftarrow T\text{-Dep}$ $\leftarrow T$ s.t. $T \in \{T\}$ , $O/T\text{-Dep} \in \{MR\}$ , $POS(O) \in \{JJ\}$	$o = O$	same as $R1_2$ with iPod as the known word and best as the extract word
$R3_1$	$T_{i(j)} \rightarrow T_{i(j)}\text{-Dep} \rightarrow T_{j(i)}$ s.t. $T_{j(i)} \in \{T\}$ , $T_{i(j)}\text{-Dep} \in \{CONJ\}$ , $POS(T_{i(j)}) \in \{NN\}$	$t = T_{i(j)}$	Does the player play dvd with <u>audio</u> and “video”? (video $\rightarrow$ conj $\rightarrow$ audio)
$R3_2$	$T_i \rightarrow T_i\text{-Dep} \rightarrow H \leftarrow T_j\text{-Dep} \leftarrow T_j$ s.t. $T_i \in \{T\}$ , $T_i\text{-Dep} == T_j\text{-Dep}$ , $POS(T_j) \in \{NN\}$	$t = T_j$	Canon “G3” has a great <u>lens</u> . ( lens $\rightarrow$ obj $\rightarrow$ has $\leftarrow$ subj $\leftarrow$ G3)
$R4_1$	$O_{i(j)} \rightarrow O_{i(j)}\text{-Dep} \rightarrow O_{j(i)}$ s.t. $O_{j(i)} \in \{O\}$ , $O_{i(j)}\text{-Dep} \in \{CONJ\}$ , $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	The camera is <u>amazing</u> and “easy” to use. (easy $\rightarrow$ conj $\rightarrow$ amazing)
$R4_2$	$O_i \rightarrow O_i\text{-Dep} \rightarrow H \leftarrow O_j\text{-Dep} \leftarrow O_j$ s.t. $O_i \in \{O\}$ , $O_i\text{-Dep} == O_j\text{-Dep}$ , $POS(O_j) \in \{JJ\}$	$o = O_j$	If you want to buy a <u>sexy</u> , “cool”, accessory -available mp3 player, you can choose iPod. (sexy $\rightarrow$ mod $\rightarrow$ player $\leftarrow$ mod $\leftarrow$ cool )

(2) 基于规则的传播算法: 输入评价词字典 O 和关于商品的评论数据 R, 其主要思想为: 首先通过初始的评价词字典识别句子中的评价词, 然后通过句法关系, 进一步识别出其他评价词或者属性, 然后将它们加入到字典。再不断迭代上面的过程, 直到没有新的评价词和属性能够被识别为止。例如, 对于句子“Canon G3 takes great pictures, The picture is amazing, You may have to get more storage to store high quality pictures and recorded movies, and The software is amazing”。形容词 great 直接通过 mod 关系依赖于名词 pictures, 定义为 OT-Rel 四元组 (JJ, DD, mod, NN)。假设评价

词字典里面有一个单词“great”。根据 great，通过句法关系，可以利用  $R1_1$  可以识别出 picture（属性）。然后通过 picture，再通过句法关系，可以在第二句话中利用  $R2_2$  识别出 amazing。通过 picture 还可以基于  $R3_1$  识别出 movies。通过 amazing 我们又可以基于  $R1_2$  识别出 software。

**(3) 评价词情感极性预测：**通过规则来预测属性观点对应的情感极性。关于评价词和属性，可以观察到两个现象：(1) 在一个评论中，对于相同的属性，情感极性一般是相同的；(2) 在一个领域库中，相同评价词具有相同极性。基于观察到的上述现象，可以建立下面三条判断情感极性的规则：(1) 对于由已知属性提取的评价词和由已知评价词提取的属性，赋予它们与已知相同的极性。例如，如果 A 是一个评价词（属性），B 是一个属性（评价词），且 A 通过 B 抽取得到，A 会被赋予 B 一样的情感极性。(2) 对于由已知评价词提取的评价词和由已知属性提取的属性，赋予它们与已知相同的极性。除非句子中出现相反的话语。例如，对于单词 A 和 B 都是属性（或者评价词）而 A 通过 B 抽取得到，如果没有相反词在 A 和 B 之间，则 A 被赋予 B 一样的情感极性，反之就是相反的极性。(3) 对于一个通过一些属性从其他评论抽取的新评价词，利用整个评论极性来进行预测。评论的极性通过包含的已知的评价词的极性求和得到。如果最终的值大于 0，则为积极，否则为消极。

## 2. 基于序列标注的情感信息抽取

属性词抽取（Aspect Term Extraction, ATE）是属性级情感分析的一个重要子任务，其目标是抽取句子中包含的属性词。给定一个包含  $|s|$  个单词的序列  $s = w_1 \dots w_{|s|}$ ，ATE 任务目标是抽取出其中属性词集合  $\{a_1, a_2, \dots, a_n\}$ 。该任务通常转换为词级别序列标注问题，通过属性词抽取标签序列  $Y = \{y_1, \dots, y_{|s|}\}$ ，其中  $y_i$  来自预定义的标签集合  $Y = \{B, I, O\}$ ，进行属性词抽取。

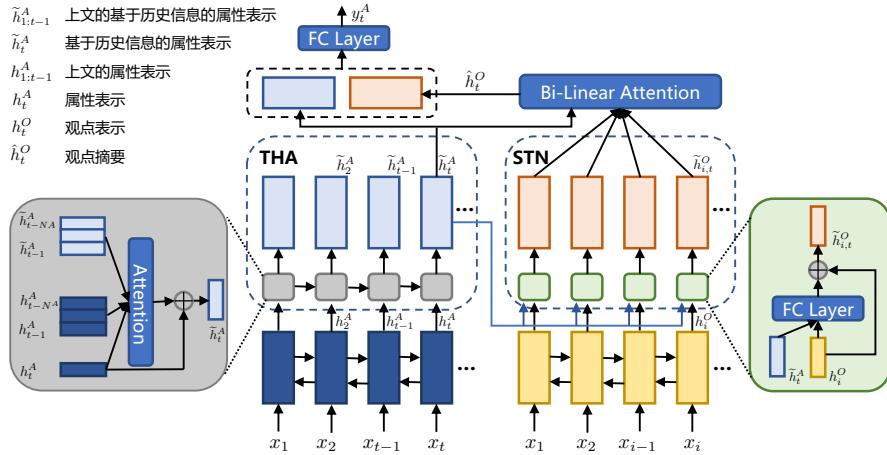
文献 [486] 提出一个基于序列标注的情感信息抽取模型 ATE-THASTN，该方法同时考虑了抽取的历史和观点信息。ATE-THASTN 算法的神经网络结构如图9.10所示。主要包含两个重要成分：属性历史注意力（Truncated History-Attention, THA）和观点选择网络（Selective Transformation Network, STN）来建模属性历史和观点摘要。历史属性表示和观点摘要拼接在一起作为特征用于属性抽取任务。

ATE-THASTN 使用两个 LSTM 模型来建模词级别的上下文表示，分别用于属性词抽取任务和辅助评价词发现任务的序列标注。 $LSTM^T$  表示一个 LSTM 单元，其中  $T \in \{A, O\}$  是任务指示器。本节中，在没有特定说明的情况下，带有上标  $A$  和  $O$  的符号分别表明属性词抽取任务和评价发现任务。使用一个双向的 LSTM 模型来生成初始的单词级别表示  $h_t^T$ ，

$$h_t^T = \left[ \overrightarrow{LSTM}^T(x_t); \overleftarrow{LSTM}^T(x_t) \right], t \in [1, |s|] \quad (9.29)$$

其中  $x_t$  表示单词  $w_t$  的词向量表示。

历史属性部分旨在建模已经预测得到的属性和当前预测属性之间的关系，建模历史的属性信息。通过考虑 BIO 标签定义减少模型预测当前标签的错误空间，同时提高在并列结构上多个属性

图 9.10 ATE-THASTN 属性抽取神经网络结构<sup>[486]</sup>

的预测准确率。属性历史注意力模块（THA）用于建模属性和属性之间的关系，缓存最近的  $N^A$  隐层状态。在预测  $t$  时刻时，THA 按照如下公式计算每一个缓存状态  $\mathbf{h}_i^A (i \in [t - N^A, t - 1])$  归一化后的重要度值  $s_i^t$ ：

$$\begin{aligned} \mathbf{a}_i^t &= \mathbf{v}^\top \tanh (\mathbf{W}_1 \mathbf{h}_i^A + \mathbf{W}_2 \mathbf{h}_t^A + \mathbf{W}_3 \tilde{\mathbf{h}}_i^A) \\ s_i^t &= \text{Softmax}(\mathbf{a}_i^t) \end{aligned} \quad (9.30)$$

其中， $\tilde{\mathbf{h}}_i^A$  为历史感知的属性表示（History-aware Aspect Representation），采用与残差块类似的结构，将隐层属性表示和压缩后的属性历史表示合并得到，具体计算公式如下所示：

$$\tilde{\mathbf{h}}_t^A = \mathbf{h}_t^A + \text{ReLU}(\hat{\mathbf{h}}_t^A) \quad (9.31)$$

其中， $\hat{\mathbf{h}}_t^A$  为属性历史表示，也是通过重要度值合并得到：

$$\hat{\mathbf{h}}_t^A = \sum_{i=t-N^A}^{t-1} s_i^t \times \tilde{\mathbf{h}}_i^A \quad (9.32)$$

观点摘要部分旨在利用观点选择网络（STN）来选择和属性相关的声音信息从而抑制可能的噪音。在处理全局的观点之前插入 STN，从而获取对于给定属性候选更加重要的特征。给定当前属性特征  $\tilde{\mathbf{h}}_t^A$ ，STN 首先按照如下公式更新观点表示  $\hat{\mathbf{h}}_{i,t}^O$ ：

$$\hat{\mathbf{h}}_{i,t}^O = \mathbf{h}_i^O + \text{ReLU}(\mathbf{W}_4 \tilde{\mathbf{h}}_t^A + \mathbf{W}_5 \mathbf{h}_i^O) \quad (9.33)$$

其中  $\mathbf{W}_4, \mathbf{W}_5$  是待学习的参数。通过上述公式将  $\mathbf{h}_i^O$  和  $\tilde{\mathbf{h}}_t^A$  映射到同一维度表示空间。属性表示

$\tilde{\mathbf{h}}_t^A$  在这里类似过滤器的作用，用于保留重要的观点特征。

为了得到全局观点摘要，这里使用了一个双向线性（Bi-Linear）项来计算  $\tilde{\mathbf{h}}_t^A$  和  $\hat{\mathbf{h}}_{i,t}^O$  之间的相关值：

$$w_{i,t} = \text{Softmax} \left( \tanh \left( \tilde{\mathbf{h}}_t^A \mathbf{W}_{bi} \hat{\mathbf{h}}_{i,t}^O + \mathbf{b}_{bi} \right) \right) \quad (9.34)$$

其中  $\mathbf{W}_{bi}$  和  $\mathbf{b}_{bi}$  是双向线性注意力的参数。第  $t$  时刻，增强的观点表示  $\hat{\mathbf{h}}_t^O$  通过观点表示通过加权求和得到：

$$\hat{\mathbf{h}}_t^O = \sum_{i=1}^T w_{i,t} \times \hat{\mathbf{h}}_{i,t}^O \quad (9.35)$$

最后，将观点摘要  $\hat{\mathbf{h}}_t^O$  和历史感知的属性表示  $\tilde{\mathbf{h}}_t^A$  进行拼接，输入到全连接层用于属性抽取预测：

$$\begin{aligned} f_t^A &= [\tilde{\mathbf{h}}_t^A : \hat{\mathbf{h}}_t^O] \\ P(y_t^A | x_t) &= \text{Softmax} (\mathbf{W}_f^A f_t^A + \mathbf{b}_f^A) \end{aligned} \quad (9.36)$$

ATE-THASTN 算法还利用多任务学习进行训练，除了属性抽取之外还引入观点预测。词级别表示  $\mathbf{h}_i^O$  用于进行观点预测：

$$P(y_i^O | x_i) = \text{Softmax} (\mathbf{W}_f^O \mathbf{h}_i^O + \mathbf{b}_f^O) \quad (9.37)$$

词级别的预测的分布  $P(y_t^{\mathcal{T}} | x_t)$  ( $\mathcal{T} \in \{A, O\}$ ) 和真实分布  $P(y_t^{\mathcal{T},g} | x_t)$  的交叉熵错误作为损失函数：

$$\mathcal{L}_{\mathcal{T}} = -\frac{1}{T} \sum_1^T P(y_t^{\mathcal{T},g} | x_t) \cdot \log[P(y_t^{\mathcal{T}} | x_t)] \quad (9.38)$$

属性抽取  $\mathcal{L}_A$  和观点抽取  $\mathcal{L}_O$  两个损失函数求和作为最后的损失函数：

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_O \quad (9.39)$$

## 9.4.2 属性级情感分类

与篇章和句子级别的情感分类任务不同，属性级情感分类任务关注句子中评价对象的特定属性的情感，也称细粒度情感分类、属性级情感分类和属性感知的情感分类。一个句子中可能包含评价对象的多个属性，并且句子中对于不同属性所表达情感可能不一致，甚至完全相反。

例如：我买了一台新相机，照片质量很好，但是电池寿命太短。

该句中关于相机的“图片质量”和“电池寿命”这两个属性的情感极性不同，同时评价词“很好”和“太短”分别用来修饰这两个属性，表达了积极和消极的情感。

属性级别情感分析任务目标是预测句子对于给定属性的情感。可以形式化定义为：给定一个包

含  $|s|$  个单词的句子  $s = w_1 w_2 \dots w_{|s|}$  以及一个包含  $|A|$  个属性的列表  $A = \{a_1, \dots, a_{|A|}\}$ , 其中每一个属性  $a_i = w_{i_1} \dots w_{i_{|a_i|}}$  是句子  $s$  的一个子序列。目标是输出属性列表中每个属性在当前文本中的情感极性  $P = \{p_1, \dots, p_{|A|}\}$ 。同时, 为了减少误差传播, 一些工作也开展联合情感信息抽取和属性级情感分类任务。给定句子  $s$ , 抽取所有的属性词  $A = \{a_1, \dots, a_{|A|}\}$  和评价词  $O = \{o_1, \dots, o_{|O|}\}$ , 再对属性词  $a_i$  和评价词  $o_i$  进行匹配, 并预测其情感极性  $p_i$ , 得到最后三元组集合  $T = \{(a_i, o_i, p_i)\}_{i=1}^{|T|}$

### 1. 基于概率混合模型的属性级情感分类

文献[487]提出一个基于混合语言模型的主题-情感模型(Topic-Sentiment Model, TSM), 同时抽取文本中包含的主题和情感。过程形式化地表示为: 对于给定的一个文档集合  $\mathbb{D} = \{d_1, d_2, \dots, d_{|\mathbb{D}|}\}$ , 假设语料中包含了  $k$  个主要主题  $\{\theta_1, \theta_2, \dots, \theta_k\}$ 。每一个主题可以建模为字典中所有单词上的多项式分布。

在主题-情感模型中, 每一个词被分为通用词(例如“这”、“那”, “的”)和主题词两种, 而对于一个主题的单词, 会进一步分为三个子类别: (1) 带有中性观点的主题词(例如“价格”); (2) 有关主题积极观点的词(例如“爱”、“喜欢”); (3) 有关主题消极观点的词(例如“讨厌”、“差”)。TSM 所采用的主题-情感模型包含四个多项式分布: (1)  $\theta_B$  表示用来抽取通用词的背景主题模型; (2)  $\Theta = \{\theta_1, \dots, \theta_k\}$  表示用来抽取关于  $k$  个子主题中性描述的  $k$  个主题模型; (3)  $\theta_P$  表示用来抽取积极观点的积极情感模型; (4)  $\theta_N$  表示用来抽取消极观点的消极情感模型。基于概率混合模型的属性级情感分类概率图模型表示如图9.11所示。

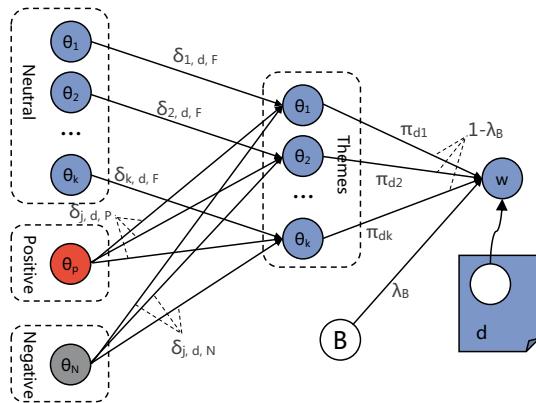


图 9.11 基于概率混合模型的属性级情感分类概率图模型表示<sup>[487]</sup>

基于上述概率图模型, 具体地抽取流程包括以下四步: (1) 判断一个单词是否为通用词。如果是, 则该单词通过  $\theta_B$  来采样; (2) 如果不是通用词, 则判断该词属于  $k$  个主题词的哪一个; (3) 判断主题之后, 进一步判断对于主题是中性、积极还是消极的; (4) 基于步骤 2 中选中主题为第  $j$  个主题  $\theta_j$ , 以及步骤 3, 通过  $\theta_j$ ,  $\theta_P$  或者  $\theta_N$  来采样单词。

根据 TSM 模型，整个语料集合  $\mathcal{D}$  的对数似然为：

$$\log(\mathcal{D}) = \sum_{d \in \mathcal{C}} \sum_{w \in V} c(w : d) \log \left[ \lambda_B p(w | B) + (1 - \lambda_B) \sum_{j=1}^k \pi_{dj} \times \right. \\ \left. (\delta_{j,d,F} p(w | \theta_j) + \delta_{j,d,P} p(w | \theta_P) + \delta_{j,d,N} p(w | \theta_N)) \right] \quad (9.40)$$

其中  $c(w : d)$  是文档  $d$  中词  $w$  的个数， $\lambda_B$  表示选择  $\theta_B$  的概率，需要预先设定的 0 和 1 之间的超参数。 $\pi_{dj}$  表示文档  $d$  中选择第  $j$  个主题的概率， $\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$  表示文档  $d$  中主题  $j$  中性、积极和消极观点的情感覆盖度。

背景模型定义为：

$$p(w | \theta_B) = \frac{\sum_{d \in \mathcal{C}} c(w, d)}{\sum_{w \in V} \sum_{d \in \mathcal{C}} c(w, d)} \quad (9.41)$$

除此之外还需要估计的参数包括主题模型参数  $\Theta = \{\theta_1, \dots, \theta_k\}$ ，情感模型参数  $\theta_P$  和  $\theta_N$ ，文档主题概率  $\pi_{dj}$ ，每一个文档的情感覆盖度  $\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$ 。这里将这些参数定义为  $\Lambda$ 。这些参数可以通过期望最大化（Expectation-Maximization，EM）算法来迭代计算最大似然估计，更新方式如公式9.42：

$$p(z_{d,w,j,F} = 1) = \frac{(1 - \lambda_B) \pi_{dj}^{(n)} \delta_{j,d,F}^{(n)} p^{(n)}(w | \theta_j)}{\lambda_B p(w | \theta_B) + (1 - \lambda_B) \sum_{j'=1}^k \pi_{dj}^{(n)} \left( \delta_{j',d,F}^{(n)} p^{(n)}(w | \theta'_j) + \delta_{j',d,P}^{(n)} p^{(n)}(w | \theta_P) + \delta_{j',d,N}^{(n)} p^{(n)}(w | \theta_N) \right)} \\ p(z_{d,w,j,P} = 1) = \frac{(1 - \lambda_B) \pi_{dj}^{(n)} \delta_{j,d,P}^{(n)} p^{(n)}(w | \theta_P)}{\lambda_B p(w | \theta_B) + (1 - \lambda_B) \sum_{j'=1}^k \pi_{dj}^{(n)} \left( \delta_{j',d,F}^{(n)} p^{(n)}(w | \theta'_j) + \delta_{j',d,P}^{(n)} p^{(n)}(w | \theta_P) + \delta_{j',d,N}^{(n)} p^{(n)}(w | \theta_N) \right)} \\ p(z_{d,w,j,N} = 1) = \frac{(1 - \lambda_B) \pi_{dj}^{(n)} \delta_{j,d,N}^{(n)} p^{(n)}(w | \theta_N)}{\lambda_B p(w | \theta_B) + (1 - \lambda_B) \sum_{j'=1}^k \pi_{dj}^{(n)} \left( \delta_{j',d,F}^{(n)} p^{(n)}(w | \theta'_j) + \delta_{j',d,P}^{(n)} p^{(n)}(w | \theta_P) + \delta_{j',d,N}^{(n)} p^{(n)}(w | \theta_N) \right)} \\ \pi_{dj}^{(n+1)} = \frac{\sum_{w \in V} c(w, d) (p(z_{d,w,j,F} = 1) + p(z_{d,w,j,P} = 1) + p(z_{d,w,j,N} = 1))}{\sum_{j'=1}^k \sum_{w \in V} c(w, d) (p(z_{d,w,j',F} = 1) + p(z_{d,w,j',P} = 1) + p(z_{d,w,j',N} = 1))} \\ \delta_{j,d,F}^{(n+1)} = \frac{\sum_{w \in V} c(w, d) p(z_{d,w,j,F} = 1)}{\sum_{w \in V} c(w, d) (p(z_{d,w,j,F} = 1) + p(z_{d,w,j,P} = 1) + p(z_{d,w,j,N} = 1))} \\ \delta_{j,d,P}^{(n+1)} = \frac{\sum_{w \in V} c(w, d) p(z_{d,w,j,P} = 1)}{\sum_{w \in V} c(w, d) (p(z_{d,w,j,F} = 1) + p(z_{d,w,j,P} = 1) + p(z_{d,w,j,N} = 1))} \\ \delta_{j,d,N}^{(n+1)} = \frac{\sum_{w \in V} c(w, d) p(z_{d,w,j,N} = 1)}{\sum_{w \in V} c(w, d) (p(z_{d,w,j,F} = 1) + p(z_{d,w,j,P} = 1) + p(z_{d,w,j,N} = 1))} \\ p^{(n+1)}(w | \theta_j) = \frac{\sum_{d \in \mathcal{C}} c(w, d) p(z_{d,w,j,F} = 1)}{\sum_{w' \in V} \sum_{d \in \mathcal{C}} c(w', d) p(z_{d,w',j,F} = 1)} \\ p^{(n+1)}(w | \theta_P) = \frac{\sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w, d) p(z_{d,w,j,P} = 1)}{\sum_{w' \in V} \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w', d) p(z_{d,w',j,P} = 1)} \\ p^{(n+1)}(w | \theta_N) = \frac{\sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w, d) p(z_{d,w,j,N} = 1)}{\sum_{w' \in V} \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w', d) p(z_{d,w',j,N} = 1)} \quad (9.42)$$

其中  $\{z_{d,w,j,s}\}$  是隐层变量的集合 ( $s \in \{F, P, N\}$ )， $p(z_{d,w,j,s})$  表示基于主题/情感模型  $w$ ，文档  $d$

中单词  $w$  从第  $j$  个主题生成的概率。

在实际应用中如果不对模型增加任何约束，仅通过 EM 算法优化得到的情感模型会偏向于特定的内容，主题模型也会因为情感而产生偏差。这主要是因为评价词和主题词经常共现，很难被 EM 算法分开。这导致情感模型和主题模型独立，而主题模型应该是中性的。为此，TSM 算法提出先定义模型先验从而约束情感/主题模型学习这个先验，再利用最大后验概率（Maximum a Posterior, MAP）估计将这些先验与数据似然性结合来预估参数。

通过定义模型先验从而约束情感/主题模型学习这个先验。假设  $\bar{\theta}_P$  和  $\bar{\theta}_N$  为根据训练语料学习到的积极和消极情感模型。为情感模型  $\theta_P$  和  $\theta_N$  定义了两个共轭狄利克雷先验， $Dir(\{1 + u_{PP}(w|\bar{\theta}_P)\}_{w \in V})$  和  $Dir(\{1 + u_{NP}(w|\bar{\theta}_N)\}_{w \in V})$ ，其中参数  $u_P$  和  $u_N$  表明对于情感模型先验的自信程度。由于先验是共轭的，所以  $u_P$ （或者  $u_N$ ）可以解释为“等效样本容量”。也就是说，当评估情感模型  $p(w|\theta_P)$ （或者  $p(w|\theta_N)$ ）时，添加先验的影响和添加词  $w$  的  $u_{PP}(w|\bar{\theta}_P)$ （或者  $u_{NP}(w|\bar{\theta}_N)$ ）伪数目等效。

为此，模型中所有参数的先验假设为：

$$\begin{aligned} p(\Lambda) \propto & p(\theta_P) * p(\theta_N) * \prod_{j=1}^k p(\theta_j) = \prod_{w \in V} p(w | \theta_P)^{\mu_{PP}(w|\bar{\theta}_P)} \\ & \prod_{w \in V} p(w | \theta_N)^{\mu_{NP}(w|\bar{\theta}_N)} \prod_{j=1}^k \prod_{w \in V} p(w | \theta_j)^{\mu_j p(w|\bar{\theta}_j)} \end{aligned} \quad (9.43)$$

其中当没有对于  $\theta_j$  的先验知识时  $u_j = 0$ 。

最大后验概率估计则是基于以上定义的先验知识，使用 MAP 估计  $\hat{\Lambda} = \arg \max_{\Lambda} p(\mathcal{C}|\Lambda)p(\Lambda)$ 。为了结合先验给定的伪数目，TSM 对 EM 算法中的 M 步进行了重写。新的 M 步更新如下：

$$\begin{aligned} p^{(n+1)}(w | \theta_P) &= \frac{\mu_{PP}(w | \bar{\theta}_P) + \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w, d) p(z_{d,w,j,P} = 1)}{\mu_P + \sum_{w' \in V} \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w', d) p(z_{d,w',j,P} = 1)} \\ p^{(n+1)}(w | \theta_N) &= \frac{\mu_{NP}(w | \bar{\theta}_N) + \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w, d) p(z_{d,w,j,N} = 1)}{\mu_N + \sum_{w' \in V} \sum_{d \in \mathcal{C}} \sum_{j=1}^k c(w', d) p(z_{d,w',j,N} = 1)} \\ p^{(n+1)}(w | \theta_j) &= \frac{\mu_j p(w | \bar{\theta}_j) + \sum_{d \in \mathcal{C}} c(w, d) p(z_{d,w,j,F} = 1)}{\mu_j + \sum_{w' \in V} \sum_{d \in \mathcal{C}} c(w', d) p(z_{d,w',j,F} = 1)} \end{aligned} \quad (9.44)$$

参数  $u_P$ ,  $u_N$ ,  $u_j$  可以是预先给定的超参数，也可以是通过正则估计来设定。

模型训练完成后可以用于以下应用：

(1) 句子主题排序：给定句子集合和一个主题  $j$ ，通过对于主题  $j$  的得分的进行句子排序：

$$\text{Score}_j(s) = -D(\theta_j \parallel \theta_s) = -\sum_{w \in V} p(w \mid \theta_j) \log \frac{p(w \mid \theta_j)}{p(w \mid \theta_s)} \quad (9.45)$$

其中  $\theta_s$  是一个句子  $s$  的平滑语言模型。

(2) 句子情感分类：给定主题为  $j$  的句子  $s$ ，预测其积极、消极或者中性情感：

$$\arg \max_x -D(\theta_s \parallel \theta_x) = \arg \max_x -\sum_{w \in V} p(w \mid \theta_s) \log \frac{p(w \mid \theta_s)}{p(w \mid \theta_x)} \quad (9.46)$$

其中  $x \in \{j, P, N\}$ ， $\theta_s$  是  $s$  的语言模型。

(3) 预测文档或者主题的整体观点：给定一个文档  $d$  和一个主题  $j$ ，主题  $j$  在文档  $d$  的整体情感分布为情感覆盖度  $\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$ 。对于主题  $j$  的整体情感强度为：

$$S(j, P) = \frac{\sum_{d \in C} \pi_{dj} \delta_{j,d,P}}{\sum_{d \in C} \pi_{dj}} \quad (9.47)$$

## 2. 基于注意力交互的属性级情感分类

由于一个句子中可能对评价对象的多个属性都进行了评论，因此如何建模属性和上下文之间交互关系，从而获取和属性相关的上下文，对于属性级情感分类任务有非常重要的作用。针对该问题，文献 [488] 提出了多粒度注意力网络 MGAN (Multi-grained Attention Network) 算法，同时考虑粗粒度和细粒度的注意力交互机制，用于建模属性和上下文之间的交互关系，其模型结构如图9.12所示。MGAN 主要包含输入嵌入层、上下文建模层、多粒度注意力层和输出层四个部分。

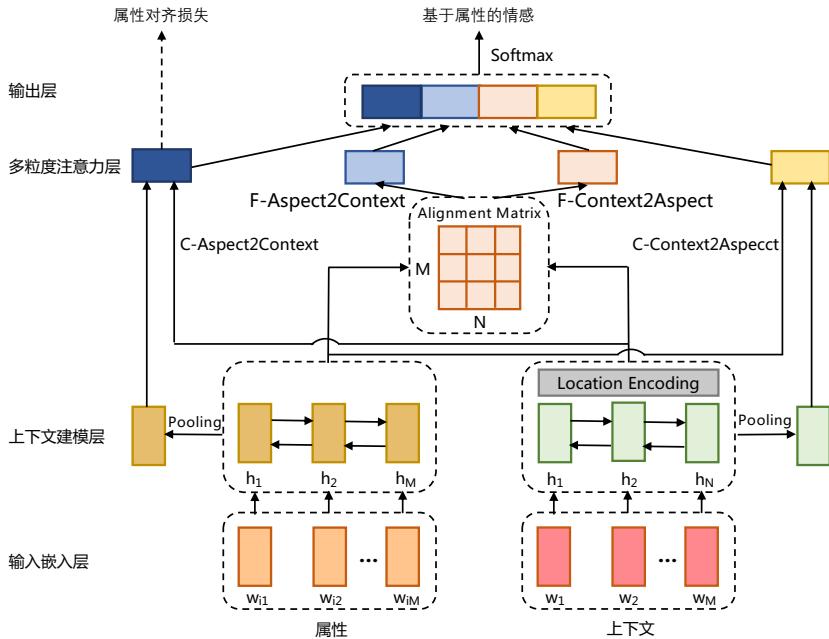
输入嵌入层将每一个单词映射到高维空间。MGAN 使用了预训练的单词向量 Glove 来获得固定的每一个单词的词嵌入表示。嵌入矩阵定义为  $L \in \mathbb{R}^{d_v \times |V|}$ ，其中  $d_v$  是单词向量的维度， $|V|$  是词表大小。

上下文建模层使用双向的 LSTM 模型来建模句子的时间序列关系。给定上下文句子  $s$  和对应属性  $a_j$  的单词词嵌入，使用两个单独的 BiLSTM 进行建模，句子表示为  $H \in \mathbb{R}^{2d \times s}$  和属性表示为  $Q \in \mathbb{R}^{2d \times |a_j|}$ 。同时，由于距离属性越近的单词对于属性的影响越大，为此需要融合考虑位置编码。对于一个距离属性为  $l$  的上下文单词  $w_j$ ，其权重定义为：

$$\alpha_i = 1 - \frac{l}{|s| - |a_i| + 1} \quad (9.48)$$

特别地，对于属性内部的词，权重设定为 0。由此，可以得到上下文表示  $H = [H_1 * \alpha_1, \dots, H_{|s|} * \alpha_{|s|}]$ 。

多粒度注意力层从粗粒度和细粒度两个角度进行属性和句子之间的交互。为了关联属性与上下文信息，MGAN 算法提出了细粒度注意力机制。通过捕捉单词级别的交互，来估计每一个属性

图 9.12 基于注意力交互的属性级情感分类模型神经网络结构<sup>[488]</sup>

和上下文单词影响。细粒度和粗粒度的注意力向量被拼接来获得最后的表示。由于属性之间的关系，也可以提供额外有价值的信息。MGAN 算法还设计了属性对齐损失函数用来加强属性和同一个上下文以及不同情感极性的注意力差别。

粗粒度注意力用于建模属性和上下文之间的交互，使用属性的平均值来计算上下文单词的注意力权重。这里使用了双向注意力机制，分别叫做 C-Aspect2Context 和 C-Context2Aspect。

- (1) C-Aspect2Context 根据属性向量的平均值学习属性对于上下文单词的权重。这里对属性上下文输出  $\mathbf{Q}$  使用平均池化来生成平均属性向量  $\mathbf{Q}_{avg} \in \mathbb{R}^{2d}$ 。对于每一个上下文中的单词向量  $\mathbf{H}_i$ ，通过如下公式计算注意力值  $a_i^{ca}$ ：

$$\begin{aligned} s_{ca}(\mathbf{Q}_{avg}, \mathbf{H}_i) &= \mathbf{Q}_{avg} * \mathbf{W}_{ca} * \mathbf{H}_i \\ a_i^{ca} &= \frac{\exp(s_{ca}(\mathbf{Q}_{avg}, \mathbf{H}_i))}{\sum_{k=1}^N \exp(s_{ca}(\mathbf{Q}_{avg}, \mathbf{H}_k))} \end{aligned} \quad (9.49)$$

其中值函数  $s_{ca}$  计算属性对于上下文单词的重要性权重。 $\mathbf{W}_{ca} \in \mathbb{R}^{2d \times 2d}$  是注意力权重矩阵。

上下文输出的加权求和表示  $\mathbf{m}^{ca} \in \mathbb{R}^{2d}$  可通过如下公式计算得到:

$$\mathbf{m}^{ca} = \sum_{i=1}^N a_i^{ca} \cdot \mathbf{H}_i \quad (9.50)$$

- (2) C-Context2Aspect 学习上下文对于属性单词的权重。与 C-Aspect2Context 类似，使用平均池化来获得平均上下文向量  $\mathbf{H}_{avg}$  来计算每一个属性中单词  $w_i$  的权重。最终属性向量的加权求和表示  $\mathbf{m}^{cc}$  可通过如下公式计算得到:

$$\begin{aligned} s_{cc}(\mathbf{H}_{avg}, \mathbf{Q}_i) &= \mathbf{H}_{avg} * \mathbf{W}_{cc} * \mathbf{Q}_i \\ \mathbf{a}_i^{cc} &= \frac{\exp(s_{cc}(\mathbf{H}_{avg}, \mathbf{Q}_i))}{\sum_{k=1}^M \exp(s_{cc}(\mathbf{H}_{avg}, \mathbf{Q}_k))} \\ \mathbf{m}^{cc} &= \sum_{i=1}^M \mathbf{a}_i^{cc} \cdot \mathbf{Q}_i \end{aligned} \quad (9.51)$$

细粒度注意力旨在建模词级别的交互，目标是评估每一个属性单词如何影响其他上下文单词。上下文  $\mathbf{H}$  和属性  $\mathbf{Q}$  之间的对齐矩阵  $\mathbf{U}$  表示第  $i$  个上下文单词和第  $j$  个属性单词之间的相似度。该矩阵具体计算方法如下:

$$\mathbf{U}_{ij} = \mathbf{W}_u([\mathbf{H}_i; \mathbf{Q}_j; \mathbf{H}_i * \mathbf{Q}_j]) \quad (9.52)$$

- (1) F-Aspect2Context 计算对于其中一个属性词最相关的上下文词。对于上下文词的注意力权重  $\mathbf{a}_i^{fa}$  计算方法如下:

$$\begin{aligned} s_i^{fa} &= \max(\mathbf{U}_{i,:}) \\ \mathbf{a}_i^{fa} &= \frac{\exp(s_i^{fa})}{\sum_{k=1}^N \exp(s_k^{fa})} \end{aligned} \quad (9.53)$$

根据权重可以获得注意力向量  $\mathbf{m}^{fa}$ :

$$\mathbf{m}^{fa} = \sum_{i=1}^N a_i^{fa} \cdot \mathbf{H}_i \quad (9.54)$$

- (2) F-Context2Aspect 计算对于其中一个上下文词最相关的属性词。基于上下文词的注意力权重

$\mathbf{a}^{fc}$ , 得到最后的向量表示  $\mathbf{q}^{fc}$ :

$$\begin{aligned}\mathbf{a}_{ij}^{fc} &= \frac{\exp(\mathbf{U}_{ij})}{\sum_{k=1}^M \exp(\mathbf{U}_{ik})} \\ \mathbf{q}_i^{fc} &= \sum_{j=1}^M \mathbf{a}_{ij}^{fc} \cdot \mathbf{Q}_j\end{aligned}\tag{9.55}$$

最终, 对  $\mathbf{q}^{fc}$  进行均值池化来获得聚合后的向量  $\mathbf{m}^{fc}$ :

$$\mathbf{m}^{fc} = \text{Pooling} \left( [\mathbf{q}_1^{fc}, \dots, \mathbf{q}_N^{fc}] \right)\tag{9.56}$$

粗粒度和细粒度注意力向量拼接起来作为最后的表示, 并将其输入到输出层来预测属性情感极性。

$$\begin{aligned}\mathbf{m} &= [\mathbf{m}^{ca}; \mathbf{m}^{cc}; \mathbf{m}^{fa}; \mathbf{m}^{fc}] \\ p &= \text{Softmax} (\mathbf{W}_p * \mathbf{m} + \mathbf{b}_p)\end{aligned}\tag{9.57}$$

为了能够使得不同属性的注意力的不同, 在 C-Aspect2Context 注意力权重上还加入属性对齐损失函数。在该损失函数的约束下, 每一个属性会通过和其他相关属性的对比更加关注重要的单词。具体来说, 对于属性列表 A 中每一个属性对  $a_i$  和  $a_j$ , 计算粗粒度注意力向量  $\mathbf{a}_i^{ca}$  和  $\mathbf{a}_j^{ca}$  的平方损失并估计  $a_i$  和  $a_j$  的距离  $d_{ij}$  作为损失权重:

$$\begin{aligned}d_{ij} &= \sigma(\mathbf{W}_d ([\mathbf{Q}_i; \mathbf{Q}_j; \mathbf{Q}_i * \mathbf{Q}_j])) \\ \mathcal{L}_{align} &= - \sum_{i=1}^{M-1} \sum_{j=i+1, y_i \neq y_j}^M \sum_{k=1}^N d_{ij} \cdot (\mathbf{a}_{ik}^{ca} - \mathbf{a}_{jk}^{ca})^2\end{aligned}\tag{9.58}$$

其中  $\sigma$  是 Sigmoid 函数,  $y_i$  和  $y_j$  是属性  $a_i$  和  $a_j$  的标签,  $\mathbf{a}_{ik}^{ca}$  和  $\mathbf{a}_{jk}^{ca}$  表示属性  $a_i$  和  $a_j$  对于第 k 个上下文单词的注意力权重。

最后的损失函数包含交叉熵损失、属性对齐损失和正则项:

$$\mathcal{L} = - \sum_{i=1}^C y_i \log(p_i) + \beta \mathcal{L}_{align} + \lambda \|\theta\|^2\tag{9.59}$$

其中  $\beta$  和  $\lambda$  为需要预先给出的超参数。

### 3. 基于端到端的联合属性级情感分类

属性级情感分类通常转化为情感要素抽取和分类两个步骤, 但是两个任务级联会导致错误的传播, 从而影响最终结果。为了解决上述问题, 文献 [489] 提出了基于生成框架的情感信息抽取和属性级情感分析统一框架, 其神经网络结构如图9.13所示。该模型将属性级情感分析任务分为抽

取和分类两个子任务，可以分别表示为指针索引和类别索引。将这两个子任务形式化到统一的生成框架下。

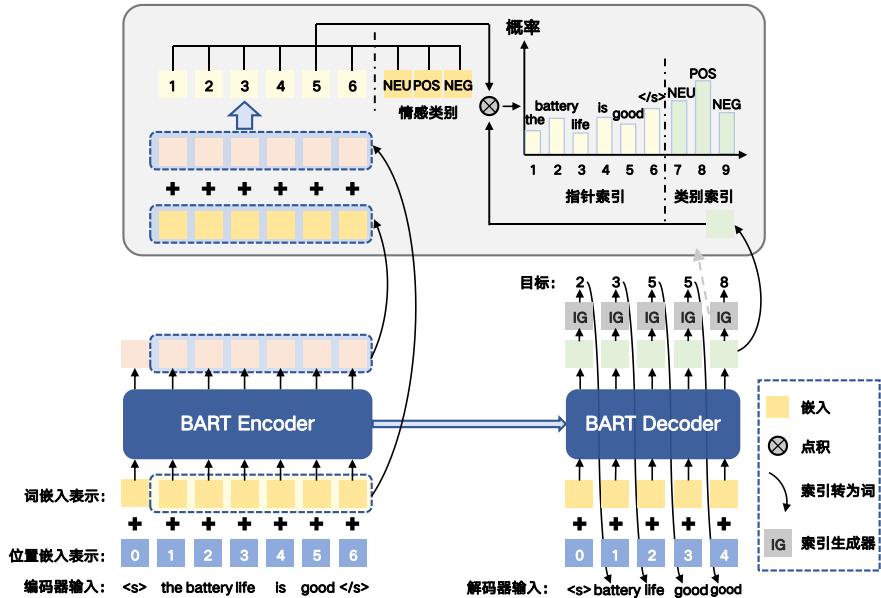


图 9.13 基于统一生成模型的联合属性级情感分析框架

使用  $a$ ,  $s$ ,  $o$  分别表示属性词、情感极性和评价词。上标  $s$  和  $e$  分别表示一个词开始和结束索引。例如： $o^s$  和  $a^e$  表示评价词  $o$  的开始索引和属性词  $a$  的结束索引， $s^p$  表示情感极性类别的索引，根据上述标签，每个子任务对应的目标序列如下：

属性词抽取 (AE) :  $Y = [a_1^s, a_1^e, \dots, a_i^s, a_i^e, \dots]$ ,

评价词抽取 (OE) :  $Y = [o_1^s, o_1^e, \dots, o_i^s, o_i^e, \dots]$

属性级情感分类 (AESC) :  $Y = [a_1^s, a_1^e, s_1^p, \dots, a_i^s, a_i^e, s_i^p, \dots]$ ,

二元组抽取:  $Y = [a_1^s, a_1^e, o_1^s, o_1^e, \dots, a_i^s, a_i^e, o_i^s, o_i^e, \dots]$ ,

三元组抽取:  $Y = [a_1^s, a_1^e, o_1^s, o_1^e, s_1^p, \dots, a_i^s, a_i^e, o_i^s, o_i^e, s_i^p, \dots]$ ,

以上的子任务仅仅关注输入句子，然而属性级情感分类 (ALSC) 和属性词评价词关系抽取 (AOE) 两个子任务依赖于特定的属性  $a$ 。该模型没有将属性词作为输入，而是放入目标端，如下所示：

属性级情感分类 (ALSC) :  $Y = [a^s, a^e, s^p]$ ,

属性词评价词关系抽取 (AOE) :  $Y = [\underline{a}^s, \underline{a}^e, o_1^s, o_1^e, \dots, o_i^s, o_i^e, \dots]$ ,

在推理过程中，带有下划线的部分是给定的。对于每一个子任务具体的目标序列样例如图9.14所示。

基于上述表示，所有的子任务都可以形式化定义为以  $X = x_1 \dots x_n$  为输入，目标序列  $Y =$

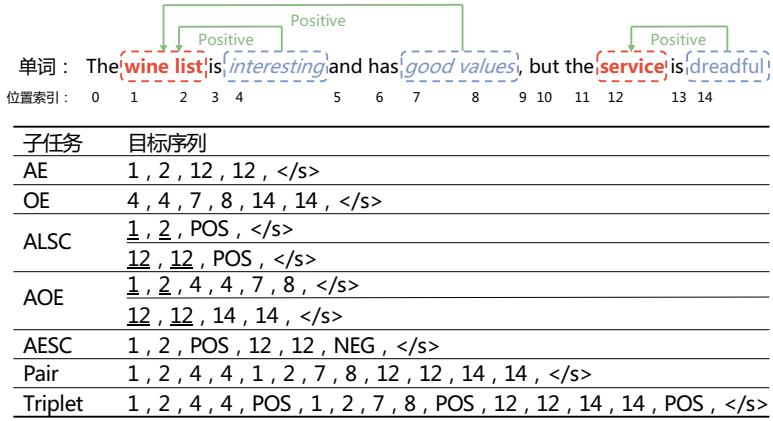


图 9.14 生成目标序列样例

$y_1 \dots y_m$  为输出，其中  $y_0$  是句子开始的标志词。不同的属性级情感分析任务可以定义为：

$$P(Y|X) = \prod_{t=1}^m P(y_t|X, Y_{<t}) \quad (9.60)$$

为了获得每一步索引概率分布  $P_t = P(y_t|X, Y_{<t})$  使用了编码器和解码器两个模块。

编码器将  $X$  编码为向量  $\mathbf{H}^e$ 。这里使用 BART 模型。句子的开始词 ( $<\text{s}>$ ) 和结束词 ( $<\text{/s}>$ ) 添加到  $X$  的开始和结束位置。这里为了简单处理，在公式中忽略了  $<\text{s}>$  词。编码器部分如下：

$$\mathbf{H}^e = \text{BARTEncoder}([x_1, \dots, x_n]) \quad (9.61)$$

其中  $\mathbf{H}^e \in \mathbb{R}^{n \times d}$ ， $d$  为隐层维度。

解码器把编码器的输出  $\mathbf{H}^e$  和前面解码器的输出  $Y_{<t}$  作为输入来计算  $P_t$ 。然而， $Y_{<t}$  是一个索引序列，为此，对于每一个  $Y_{<t}$  中的  $y_t$ ，需要使用下面的索引到词的模型来做个转化：

$$\hat{y}_t = \begin{cases} X_{y_t}, & \text{如果 } y_t \text{ 是一个指针索引} \\ C_{y_t-n} = s, & \text{如果 } y_t \text{ 是一个类别索引} \end{cases}$$

其中  $C = \{c_1, \dots, c_l\}$  类别词列表，为了与词索引区分，根据句子长度进行了长度为  $n$  的位移。

最后，使用 BART 解码器来获得最后的隐层表示

$$h_t^d = \text{BARTDecoder}(\mathbf{H}^e; \hat{Y}_{<t}) \quad (9.62)$$

基于  $h_t^d$ , 使用如下公式预测词的概率分布  $P_t$ :

$$\mathbf{E}^e = \text{BARTTokenEmbed}(X) \quad (9.63)$$

$$\hat{\mathbf{H}}^e = \text{MLP}(\mathbf{H}^e) \quad (9.64)$$

$$\bar{\mathbf{H}}^e = \alpha \hat{\mathbf{H}}^e + (1-\alpha) \mathbf{E}^e \quad (9.65)$$

$$\mathbf{C}^d = \text{BARTTokenEmbed}(C) \quad (9.66)$$

$$P_t = \text{Softmax}([\bar{\mathbf{H}}^e; \mathbf{C}^d] h_t^d) \quad (9.67)$$

在训练的过程中, 使用 Teacher-forcing 方法训练模型。该策略在训练网络过程中, 每次不使用上一个状态的输出作为下一个状态的输入, 而是直接使用训练数据的标准答案的对应上一项作为下一个状态的输入。通过这种方式可以矫正模型的预测, 避免在序列生成的过程中误差进一步放大。负对数似然 (Negative log-likelihood, NLL) 被用来作为损失函数优化模型。更多地, 在推理阶段, 使用束搜索来获得目标序列  $Y$ 。束搜索有一个超参数束宽 (beam size), 设为  $k$ 。第一个时间步长, 选取当前条件概率最大的  $k$  个词, 当做候选输出序列的第一个词。之后的每个时间步长, 基于上个步长的输出序列, 挑选出所有组合中条件概率最大的  $k$  个, 作为该时间步长下的候选输出序列。始终保持  $k$  个候选, 最后从  $k$  个候选中挑出最优的。最后, 基于获得目标序列  $Y$ , 使用解码算法将序列转化为词片段和情感极性。即将指针索引根据其中文本中开始和结束位置转化为具体的属性词或者评价词, 将类别索引转化为具体的情感类别。

### 9.4.3 属性级情感分析语料库

目前使用较多的属性级情感分类数据集是语义评测国际研讨会发布<sup>[490-492]</sup>, 包括 SemEval 2014<sup>[490]</sup>, SemEval 2015<sup>[491]</sup> 和 SemEval 2016<sup>[492]</sup>. 同时, 包括 Twitter<sup>[493]</sup>, Sentihood<sup>[494]</sup> 也常用于该任务。常用的属性级情感分析语料库如表9.5所示。

表 9.5 常见的属性级情感分析语料库

	训练集合	验证集	测试集	合计	任务	语言
Restaurant14	1,978	-	600	2,578	分类	英文
Laptop14	1,462	-	411	1,873	分类	英文
Restaurant15	1,120	-	582	1,702	分类	英文
Restaurant16	1,708	-	587	2,295	分类	英文
Twitter	6,248	-	692	6,940	分类	英文
Sentihood	-	-	-	5215	分类	英文
MPQA					抽取	英文
ASAP	36,850	4,490	4,490	46,730	抽取	中文

### 1. SemEval 2014-2016 数据集

SemEval 2014 task4<sup>[490]</sup> 关注的是基于属性级别情感分析，该任务的目标是检测给定目标实体的属性并确定每个属性所表达的情感极性。有两个针对笔记本电脑和餐馆的特定领域的数据集，即 Restaurants14 和 Laptop14。每个句子都被归入句子中讨论的以下五个属性的一个或多个类别：(1) 食物；(2) 服务；(3) 价格；(4) 氛围（指餐厅的气氛和环境的句子）；(5) 轶事/杂事（不属于上述四个类别的句子）。具体来说，每个单字或多字的属性词都根据句子中对它所表达的情感而被赋予以下极性之一：(1) 积极；(2) 消极；(3) 中性（指既非积极也非消极的情绪）；(4) 冲突（意味着既是积极又是消极的情绪）。SemEval-2015 任务 12<sup>[491]</sup> 是 SemEval-2014 任务 4 的延续。SemEval-2016 任务 5<sup>[492]</sup> 与 SemEval-2015 任务 12 类似，该数据集由整个评论组成。此外，该数据集包含五个领域，涵盖八种语言。

Restaurants14 数据集<sup>[495]</sup> 从餐厅评论中提取的 3000 多条英文句子组成，作为训练数据集。额外评论以相同的方式进行标注作为测试数据集。在去除有冲突的情绪极性或没有属性词的数据后，剩下 1978 个训练样本和 600 个测试样本。该数据集包括对粗略的属性类别、属性词、属性词特定极性和属性类别特定极性的标注。

Laptop14 数据集由超过 3000 个从客户笔记本评论中获得的英文句子组成。该数据集的一部分被划分为测试数据。在除去有冲突的情感极性或没有属性词的数据后，剩下 1462 个训练样本和 411 个测试样本。该数据集只包括对句子的属性词及其极性的标注。

Restaurants15 数据集由 254 条和 96 条餐厅评论组成，分别为训练和测试的属性及其情感极性做了标注。每个评论可能包含多个句子，每个句子包括类别、属性词和属性极性的标注。在去除冲突情绪极性的数据后，有 1120 句用于训练，582 句用于测试。

Restaurants16 数据集由 350 条餐厅评论组成，其中有用于训练的属性词、属性类别和极性的标注，有 92 条用于测试。在去除有冲突情绪极性的数据后，有 1708 个标注的句子用于训练，587 个用于测试。

### 2. Twitter 情感分析数据集

Twitter 数据集<sup>[493]</sup> 是由北京航空大学、微软亚洲研究院和哈尔滨工业大学于 2014 年联合发布的属性级情感分析数据集。该数据集为手动标注的数据集，用于属性依赖的推特情感分析。这是最大的属性依赖的 twitter 情感分类数据集，它是由人工标注的。训练数据有 6248 条推文，测试数据包括 692 条推文，其情感类别平衡为 25% 负面，50% 中性，25% 正面。

### 3. Sentihood 城市街区评论

SentiHood<sup>[494]</sup> 是由英国伦敦大学学院、华威大学等公布的城市街区领域属性级情感分析基准英文数据集。它是基于与伦敦市街区有关的问题，这些问题通过过滤雅虎答案的问题回答平台的文本而获得的。SentiHood 由 5215 个句子组成，其中 3862 个句子包含一个地点，1353 个句子包含两个地点。在整个数据集中，位置实体名称被 location1 和 location2 所掩盖。

#### 4. MPQA 新闻情感数据集

MPQA<sup>[496]</sup> 是由匹兹堡大学于 2015 年发布的属性级情感分析英文数据集，起包含了标注了观点和其他状态（如情感、信念、情绪和猜测）的新闻文章和其他文本文件。在 MPQA 3.0 中，增加了实体-属性和事件-属性（eTarget）标注。一共有 70 个文档，除了 MPQA 2.0 的 1029 个表达主观元素、1287 个态度和 1213 个属性片段外，还加入了 1366 个 eTargets 到表达主观元素，1608 个 eTargets 到属性片段。

#### 5. ASAP 大众点评数据集

ASAP<sup>[497]</sup> 是由大众点评于 2021 年发布的属性级情感分析中文语料库。该库由 46730 条真实世界的用户评论组成，其被随机分成训练集（36,850）、验证集（4,940）和测试集（4,940），包含了 18 个属性类别。数据集标注了文本中包含的情感以及每个属性相应的情感极性，分为积极、中性和消极。

### 9.5 延伸阅读

本章介绍了文档级情感分析、句子级情感分析和属性级情感分析三个任务情感分析常见算法。这些方法大多是基于大规模标签数据的有监督机器学习方法，在样本不足情况下表现较差。同时，情感分析任务需要借助大量的外部知识，如句法分析、情感知识库等，如何更好利用这些知识也成为一个 important 问题。对于属性级情感分析任务，还存在如何建模属性和上下文之间的交互等问题。近年来，很多研究者从不同方面对上述问题进行探索。

为了解决数据集较小的问题，研究者提出很多更加高效的建模算法，包括递归神经网络<sup>[479, 498–500]</sup>，卷积神经网络（CNN）<sup>[501, 502]</sup>，层次结构模型<sup>[468, 502, 503]</sup>等。同时，许多研究通过微调的方式将预训练的模型应用于情感分析等下游任务<sup>[29, 504–506]</sup>。Song 等人<sup>[506–508]</sup> 集成 BERT 进行属性的情感分类，并取得了显着的进步，这表明预训练这种迁移学习的有效性。更多地，基于迁移学习情感分析的方法也得到学术研究者的广泛关注，包括任务迁移<sup>[509]</sup>，基于枢轴的（Pivot-based）跨领域迁移模型<sup>[510, 511]</sup>，基于对抗的跨领域迁移模型<sup>[512–514]</sup>等。

为了更好利用现有的知识，大量的研究工作对如何结合外部知识用于情感分析任务进行探索。一些工作提出将外部知识添加到预训练 BERT 中，以增强表示<sup>[515, 516]</sup>。Levine 等人<sup>[517]</sup> 引入了 Sense-BERT 来通过预测 WordNet 中的标记的同义词来提高词汇理解。田等人<sup>[480]</sup> 和 Ke 等人<sup>[518]</sup> 结合外部知识以学习情感信息。此外，也有研究将情感相关概念的常识知识纳入属性级情感分析任务的深度神经网络的端到端训练中，包括循环神经网络模型<sup>[519]</sup>，图神经网络模型<sup>[520]</sup>等。为了考虑依赖关系树，采用递归神经网络<sup>[493, 521]</sup>、注意力网络<sup>[522]</sup>、依赖关系树结构感知模型<sup>[523, 524]</sup>进行建模。

属性级情感分析任务的其中一个挑战是如何获取给定属性的相关信息。传统方法依赖于设计好的特征，导致构建这些模型相当费力<sup>[525]</sup>。由于神经网络模型在分布式方面上非常大的优势，目前的研究也大都着重在神经网络模型方面，包括循环神经网络模型<sup>[526]</sup>，深度记忆网络（Deep memory

network) [527, 528]，卷积神经网络模型<sup>[529]</sup>，门机制<sup>[530]</sup>，层次网络<sup>[531]</sup>等。注意力模型在捕获与给定属性相关的重要部分的方面体现强大能力，包括属性感知的注意力机制<sup>[532, 533]</sup>，多层次注意力机制<sup>[528]</sup>，交互式注意力机制<sup>[532]</sup>。使用句子中某个属性的位置信息来捕获更准确的特定于属性的信息的想法吸引了研究人员的关注<sup>[523, 524, 528, 534]</sup>。

## 9.6 习题

- (1) 当情感表达为隐式情感，没有情感词时如何训练情感分类器？
- (2) 目前的情感分析系统可解释性和鲁棒性如何？以及如何判断模型的情感可解释性和鲁棒性？
- (3) 情感分析和文本分类有什么区别？属性级情感分析和信息抽取有什么区别？
- (4) 如何进行跨语言情感分析？
- (5) 跨领域情感倾向分析的主要难点有哪些？

# 10. 智能问答

智能问答（Questing Answering, QA）旨在自动回答用户以自然语言方式提出的各类问题。自1950年图灵测试（Turing Test）提出以来<sup>[2]</sup>，以自然语言进行人机交互就成为人们不断追求和奋斗的目标。这其中如何自动回答用户的各类问题，也成为了自然语言处理领域的研究热点和难点。受到当前技术水平的限制，根据候选答案的来源不同以及问题种类的不同，问答系统所采用的技术手段也不尽相同。近年来，随着深度学习方法的不断进步，特别是超大规模预训练模型的发展，智能问答研究成果不断涌现，各类型问题的回答效果不断提高。智能问答也逐渐成为了对话助手、智能客服、搜索引擎等系统中必不可少的组成部分。

本章首先介绍智能问答的基本概念和发展历程，在此基础上按照问答答案来源的不同，分别介绍阅读理解、表格问答、社区问答以及开放领域问答相关算法。

## 10.1 智能问答概述

智能问答的目标是针对用户输入的自然语言方式描述的问题自动给出答案。20世纪60年代开始自然语言处理领域就开始了相关研究，BASEBALL系统<sup>[535]</sup>试图通过解析用户自然语言提出的关于美国棒球联赛的问题，通过结构化数据产生答案。1999年TREC（Text Retrieval Conference）举办了第一届开放领域问答评测任务TREC-8<sup>[536]</sup>，推动了智能问答的快速发展。此后，随着搜索引擎将各种智能问答算法应用于处理用户输入的自然语言查询，使得智能问答相关算法加速发展。如图10.1所示，用户在搜索引擎中输入“珠穆朗玛峰海拔多少米？”，系统可以直接给出“8848.86米”的答案。



图 10.1 搜索引擎中智能回答应用样例（来源：百度搜索）

近年来，随着智能终端的普及以及语音识别技术的进步，智能对话助手也逐渐深入大众生活，这其中也涉及大量用户以自然语言提出的各种类型问题，对智能问答的技术需求也更加急迫。同时，得益于互联网特别是 Web 2.0 的高速发展，大量问答知识以文本、表格、问题答案对等形式存在于互联网中。智能问答作为搜索引擎、对话助手、智能客服等系统的核心模块，受到学术界和企业界越来越多的关注，创新的智能问答研究成果也伴随着深度学习的发展不断涌现。

在本节中，我们将介绍智能问答的发展过程以及各类智能问答系统。

### 10.1.1 智能问答发展历程

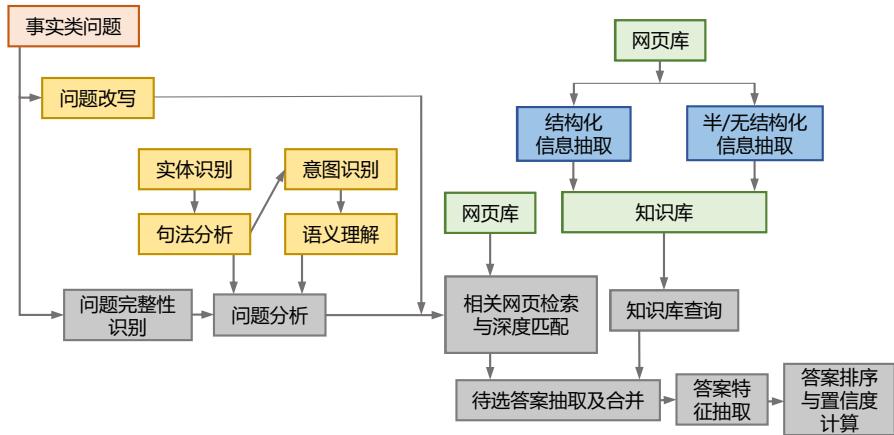
20 世纪 60 年代开始自然语言处理领域就开始从数据库的自然语言接口（Natural Language Interface for Database）角度开启了智能问答的研究。针对特定领域的结构化数据，通过分析用户的自然语言查询输入，并将其转换为结构化数据库查询，从而得到相应的答案。除了针对美国棒球联赛的 BASEBALL 系统之外，LUNAR<sup>[537]</sup> 系统则针对月球地质学家访问 NASA 数据库需求，构建了自然语言查询接口，可以回答如下问题：

- (1) Which samples are breccias?
- (2) What is the average analysis of Ir in rock SI0055?
- (3) How much Titanium does S10017 contain?

20 世纪 70 年代到 80 年代，绝大部分智能问答研究仍然集中于受限领域。SHRDLU 系统<sup>[538]</sup> 能够接受自然语言的指令，指挥虚拟积木世界中的机器人移动玩具积木块。Chat-80 系统<sup>[539]</sup> 将英语转换为 Prolog 语言，根据知识库回答关于国家、城市、河流等世界地理知识问题。UNIX Consultant 系统<sup>[540]</sup> 可以支持用户使用自然语言，完成对 UNIX 操作系统中特定任务的操作。这期间也有一些工作从自然语言理解的角度，试图理解文本内容，从而直接回答用户问题。1977 年 Lehnert 发布的 QUALM 系统<sup>[541]</sup> 提出了阅读理解概念，针对不同的问题类型使用不同的策略从文章中寻找答案。

20 世纪 90 年代开始，随着互联网的发展和统计机器学习技术的不断进步，更实用的开放领域问答系统开始兴起。1993 年由麻省理工学院 Boris Katz 教授带领 InfoLab 实验室开发的第一个基于互联网的智能问答系统 START<sup>[542-544]</sup> 上线，开启了以整个互联网为知识库的开放领域问答时代。1999 年开放领域问答评测任务 Trec-8 的提出，更进一步推动了开放领域问答的研究。TREC-8 评测要求根据给定的数据集合，回答事实性的短答案问题，要求系统返回答案和对应的文档编号。评测的问题相对比较简单，例如：“How many calories are there in a Big Mac?”。IBM 构建的 Watson 系统<sup>[545, 546]</sup> 参加了多次 TREC 智能问答评测，并于 2011 年参加了美国电视问答比赛界面 Jeopardy!，并战胜了人类冠军。2017 年搜狗问答机器人汪仔在江苏卫视问答节目《一站到底》中也战胜了人类选手取得最终胜利。搜狗问答机器人所使用的问答系统框架如图 10.2 所示，这也代表了非端到端问答系统的典型结构。

随着深度学习方法在自然语言处理领域取得突破，阅读理解、社区问答、表格问答、知识图

图 10.2 搜狗问答机器人汪仔问答系统结构图<sup>[547]</sup>

谱问答等研究领域在 2012 年以后也都取得了长足的进步。特别是 2016 年斯坦福大学发布的阅读理解数据集 SQuAD<sup>[548]</sup>，极大的推动了阅读理解的研究进展。SQuAD 包含了 10 万个高质量的问题和对应的答案，为深度学习方法提供了充足的训练语料，同时也初步验证了在训练数据充足的情况下，深度学习算法甚至可以取得超过人类的性能。2017 年哈尔滨工业大学讯飞联合实验室也发布了中文机器阅读理解评测（CMRC）。2018 年百度发布了中文阅读理解数据集 DuReader，包含 20 万个问题、100 万个文档和超过 42 万个人工给出的答案。此后大量智能问答相关评测集合相继发布，包括：多步推理阅读理解评测 HotpotQA<sup>[549]</sup>、对话问答数据集 CoQA<sup>[550]</sup>、复杂序列问答 SequenceQA<sup>[551]</sup>、自由形式表格问答 FeTaQA<sup>[552]</sup> 等。

### 10.1.2 智能问答主要类型

受到现有机器学习方法和自然语言处理算法能力的限制，目前还没有通用方法可以回答所有类型的问题，根据问题类型以及答案来源的不同，需要采用不同的方法进行解决。根据参考文献<sup>[553]</sup>中给出的分类，可以将问题大致分为七类：事实类（Factoid）、是非类（Yes/No）、定义类（Definition）、列表类（List）、比较类（Comparison）、意见类（Opinion）以及指导类（How-to）。表10.1给出了各问题类型的问答样例。根据上述不同类型问题的特点，再结合知识库的来源，可以将智能问答分为五大类：阅读理解、表格问答、社区问答、知识图谱问答和开放领域问答。

阅读理解（Machine Reading Comprehension, MRC），也称机器阅读理解，是指根据给定的一篇或多篇文本内容回答给定的问题。按照答案类型的不同，还可以进一步细分为完型填空、多项选择、片段抽取和自由作答四种形式。相关工作在 2016 年 SQuAD 语料集发布之前，就已经开始了相关研究。1999 年 Deep Read 阅读理解系统<sup>[554]</sup>在 Remedia 语料集上进行了测试。2004 年香港中文大学发布了中英双语阅读理解语料集 ChungHwa<sup>[555]</sup>。相关算法将在本章中进行介绍。

表 10.1 各问题类型问答样例

问题类型	问题	答案
事实类	复旦大学成立于那年?	1905 年
是非类	复旦大学在上海吗?	是
定义类	什么是自然语言处理?	实现人与计算机之间用自然语言进行有效通信的各种理论和方法
列表类	复旦大学有哪几个校区?	邯郸、枫林、张江、江湾
比较类	上海和北京哪个人口多?	根据第七次全国人口普查北京市人口数为 2189.31 万, 上海市人口数为 2487.09 万, 上海人口比北京多
意见类	你觉得上海哪些最值得去的地方?	上海这座被称为“魔都”的城市有很多知名的经典, 我觉得最值得去的地方包括外滩、东方明珠、城隍庙、豫园、田子坊等
指导类	如何从虹桥机场到浦东机场?	可以有以下三种方式: 1) 机场大巴直达: 机场一线; 2) 乘坐地铁 2 号线直达; 3) 乘坐地铁 2 号线到龙阳路站换乘磁悬浮列车

表格问答 (Table based Question Answering, TBQA) 是指根据给定的表格数据生成问题答案。表格通常由  $M$  行  $N$  列的数据组成, 第一行中的  $N$  个单元格是表头信息。2015 年由斯坦福大学发布的表格问答数据集 WikiTableQuestions<sup>[556]</sup> 中给定了问题和与其对应的数据表格。2022 年 FeTaQA 数据集<sup>[552]</sup> 进一步升级, 要求根据表格和问题生成需要归纳和推理得到的句子形式的答案。相关算法将在本章中进行介绍。

社区问答 (Community Question Answering, CQA) 是指根据社区问答等来源获得的 < 问题, 答案 > 对进行问题回答。随着社交媒体的高速发展, 以知乎、Quora 等为代表的问答类网站提供了用户发布问题和回答问题的渠道, 并提供了用户点赞、关注、评论等多种交互方式。这些问答数据中包含了大量人工凝练和总结的高质量答案, 可以很好的回答大量其他方法很难自动回答的比较类、意见类以及指导类问题。社区问答根据用户输入的问题, 从已有的问答对中寻找语义最相关的问答, 并将所对应的答案返回给用户。社区问答最核心的技术问题就是计算用户输入问题和已有问题之间的语义相关性, 以及用户输入问题和答案之间的语义相关性。相关算法将在本章中进行介绍。

知识图谱问答 (Knowledge based Question Answering, KBQA) 是指根据给定知识图谱生成问题的答案。知识图谱采用图的方法对知识进行结构化表示, 图的节点表示实体, 图的边表示实体之间的关系。利用信息抽取、实体融合等技术可以从大规模的自然语言文本、表格等数据中构建大规模的知识图谱。在此基础上, 根据用户的问题, 在图谱中根据实体和关系并利用推理机制可以进行问题回答。由于知识图谱问答与知识图谱构建、表示与推理紧密关联, 因此相关算法将在本书第 12 章知识图谱部分进行介绍。

开放领域问答（Open-domain Question Answering, ODQA）是通过大规模文档集合回答不限定领域的事实性问题。通常情况下开放领域问答由答案段落检索和答案抽取两个部分组成。IBM Watson 系统<sup>[546]</sup> 就是采用了开放领域问答架构。通过将问题分解并生成检索词，之后根据检索词得到相关段落，并对段落进行评判，在此基础上通过段落抽取最终答案。目前的开放领域问答系统很多是结合搜索与阅读理解，在通过传统搜索或语义搜索得到候选篇章后，利用阅读理解技术获得最终答案。目前的搜索引擎中也大量使用该项技术，提升用户的搜索体验。相关算法将在本章中进行介绍。

## 10.2 阅读理解

机器阅读理解目标就是根据给定的一篇或多篇文本内容回答给定的问题。这个任务对于算法的自然语言理解和推理能力是很大的考验。虽然阅读理解任务研究在 1999 年就已经开始，但是受到基于特征的机器学习算法能力的限制，以及缺乏大规模的标注数据，阅读理解的研究进展相对较慢。2016 年斯坦福大学推出了 SQuAD 语料集，大幅度推动了人们对阅读理解研究的关注，大量深度神经网络的模型不断提出。2018 年 1 月，微软亚洲研究院提出的 R-Net 算法<sup>[557]</sup> 率先在 SQuAD 的精准匹配指标上首次超越人类。2018 年基于预训练语言模型 BERT 的阅读理解算法在该任务上取得了大幅度进展。近年来，各类型阅读理解任务不断推出，阅读理解任务也成为了验证大规模预训练模型的标准任务之一。

图10.3给出了阅读理解样例。对于问题“复旦大学江湾校区位于上海哪个区？”，通过选择候选答案以及答案抽取，最终返回从文章中抽取得到的片段“杨浦区”作为答案。



**问题：**复旦大学江湾校区位于上海哪个区？

**答案候选：**

- 1. 复旦大学江湾校区位于杨浦区新江湾城西北部
- 2. 它是开发建设是以复旦大学为核心的杨浦知识创新园区的重要组成部分

**答案抽取：**杨浦区

图 10.3 阅读理解样例

目前绝大多数的机器阅读理解算法都采用有监督方法，将阅读理解任务转换为分类问题，因为不同的答案形式需要所采用的方法有所不同，由此可以将阅读理问题分为以下四大类：

- (1) 完型填空：对于一段文本中的某个句子，其中缺少某个单词或短语，需要算法根据其他的文本内容预所缺失部分。

- (2) 多项选择：给定文本内容，问题及对应的若干选项，需要从选项中选择出一个或多个正确的选项构成答案。
- (3) 片段抽取：给定文本内容和问题，从文本中抽取单词、短语、句子或者段落作为答案。
- (4) 自由作答：给定文本内容和问题，生成一段可以回答问题的文字。答案可能并没有出现在给定文本中。

本节将从算法类型角度，介绍基于特征的阅读理解和基于深度神经网络两大类阅读理解算法。

### 10.2.1 基于特征的阅读理解算法

基于特征的阅读理解算法通常是根据人工设计的特征，使用规则或者有监督分类算法，针对问句对篇章中的句子或者短语进行评分。基于该评分，选取得分最高的句子或片段做为答案。

#### 1. 基于规则的 Quarc 阅读理解算法

Quarc (QUestion Answering for Reading Comprehension)<sup>[558]</sup> 是一个基于规则的阅读理解系统，面向片段抽取式的阅读理解类型，并且抽取的答案限定于句子级别，试图在给定的文章中找到最合适的句子作为问题的答案。Quarc 使用词汇和语义作为基础规则条件，通过启发式的规则来寻找最合适的句子作为答案。由于不同类型的问题所使用的规则有比较大的差别，Quarc 系统将问题分为：WHO、WHAT、WHEN、WHERE、WHY 等 5 个类型，并针对不同类型的问题构建了不同的规则集。

Quarc 算法首先使用浅层句法解析器 Sundance<sup>[559]</sup> 进行形态分析、词性标注、语义类标注和实体识别，解析问题和文章中的所有句子。为了避免直接使用单词造成的规则的泛化能力不足的问题，Quarc 算法在对单词进行比对时会使用词根，用于消除语法形式的影响。此外还定义了语义类（例如 HUMAN、LOCATION 等），并使用规则方式识别单词的语义类。主要的语义类别包含以下几项：

- HUMAN：包含 2608 个单词，包含常见的名，姓氏和例如“博士”和“女士”等头衔，并且包含 600 个从 WordNet 中抽取的职业词。
- LOCATION：包含 344 个单词，包含 204 个国家名和 50 个美国的州名。
- MONTH：包含一年中的 12 个月。
- TIME：共包含 667 个词语，其中 600 个是 1400-1999 的年份，其他是一些通用的时间表达。

在此基础上，根据问题的类型选择不同的规则集合，将这些规则应用于文章中包括标题在内的所有句子。每条规则都会给予句子一定的分数，分数的高低取决于该规则对于它能找到答案的确定程度。一条规则可以分配四种类型的分数：线索 (clue) (+3)，好线索 (good\_clue) (+4)，确信 (confident) (+6)，完全契合 (slam\_dunk) (+20)。当所有规则都执行完毕后，得到分数最高的句子将被选作答案。以 WHERE 类问题的规则集合为例，Quarc 算法中 WHERE 类部分规则如下：

1. Score(S) += WordMatch(Q,S)
2. If contains(S,LocationPrep) Then Score(S) += good\_clue

### 3. If contains(S, LOCATION) Then Score(S) += confident

规则 1 中 WordMatch 函数是根据同时出现在句子和问题中的单词计算的得分。当两个单词拥有相同的词根时会认为这两个单词匹配。由于动词对于识别问题和句子的相关性非常重要，所以动词匹配的权重比非动词更高：每个匹配的动词被授予 6 分，其他匹配词仅有 3 分。规则 2 中 LocationPrep 表示地点介词（包括：in、at、near 等），该规则试图寻找句子 S 中包含地点介词，对于包含地点介词的句子增加“好线索”类对应的分数。规则 3 试图寻找包含 LOCATION 语义类的句子，相应的句子增加“确信”类所对应的分数。

当所有规则应用于文章中的每个句子，并分配了分数之后，拥有最高分数的句子就被当成是最佳答案。如果有相同分数的句子，对于原因类的问题将选择出现更晚的句子，而其他问题都会选择出现最早的句子。如果所有句子都没有得到分数，时间和地点类的问题将返回根据日期线规则集得到的句子，原因类问题将选择文章中的最后一个句子，其它类型都将返回文章的第一个句子。

Quarc 算法是典型的基于规则的自然语言处理方法，在对篇章进行词法、句法和语义分析的基础上，构建人工规则。在阅读理解这种复杂的自然语言处理任务上，基于规则的方法效果并不十分理想。

## 2. 基于最大熵的阅读理解算法

在阅读理解任务中，通常给定的输入文本很长，但是文章中可以作为答案的句子却只出现一次或很少几次，在这种情况下，想要从输入文本中提炼出一个简短的答案就显得非常困难。上一节中所介绍的基于规则的方法，以词为核心对输入文本进行浅层的分析，但这种浅层的分析很难解决这种答案与文本长度相差悬殊的情况，为了解决这些问题，必须对输入文本进行更深入的分析，捕捉更深层的特征。RCME 算法<sup>[560]</sup> 在词袋模型的基础上，引入了两类深层特征：词依赖特征和语法关系特征，并采用最大熵算法将两类特征加以利用。

为了更好的提取深层特征，RCME 算法需要对输入文本中的每个句子进行语法分析，得到整个句子的语法结构和每个单词之间的依赖关系和词性。图10.4给出了一个对句子进行语法分析并得到语法解析树的例子，语法解析树中蕴含着丰富的结构信息，为句子中的每个单词标明了词性，并清晰的展示了句子中单词的语法关系。RCME 模型将句子中每个单词的词性抽取出来作为词依赖特征，将词之间的语法关系作为语法关系特征，利用这两个深层特征去判断给定文本中的哪一句话最适合作为问题的答案。

例如，对于下面的问题 Q 和句子 C：

Q：“谁写了《红楼梦》？”

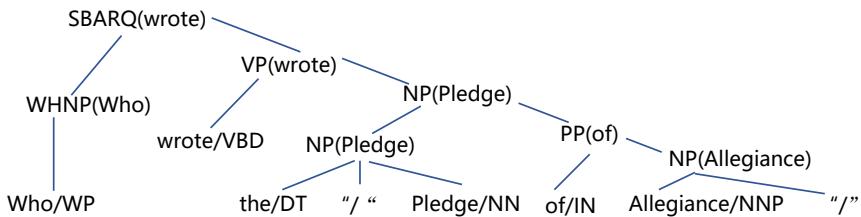
C：“《红楼梦》是曹雪芹写的”

其中部分词和词性的集合如下：

$Q = \{\text{写}/\text{动词}, \text{红楼梦}/\text{名词}\}$

$C = \{\text{写}/\text{动词}, \text{红楼梦}/\text{名词}, \text{曹雪芹}/\text{名词}\}$

Question: Who wrote the "Pledge of Allegiance"



Answer sentence: The pledge was written by Frances Bellamy.

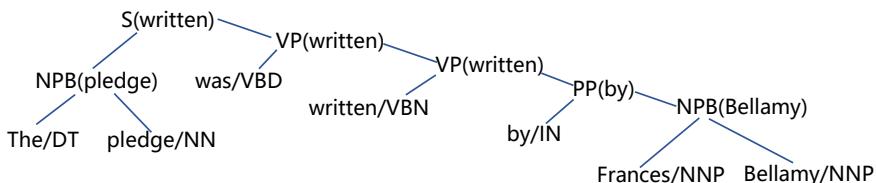


图 10.4 RCME 对问题和答案句语法分析结果示例

为了利用深层特征, RCME 首先定义了标识函数  $f_j(Q, C)$ , 对于问题  $Q$  中的每一个特征  $f_j$ , 当句子  $C$  中包含该特征时,  $f_j(Q, C) = 1$ , 反之,  $f_j(Q, C) = 0$ 。在上面的例子中,  $f_{\text{动词}}(Q, C) = 1$ ,  $f_{\text{名词}}(Q, C) = 1$ 。RCME 利用最大熵的框架利用这些特征挑选最适合回答问题的句子: 对于给定的文本  $S = \{C_1, \dots, C_n\}$  和给定问题  $Q$ , 模型需要从中挑选最适合回答问题的句子:

$$A = \arg \max_{C_i \in S} P(C_i | Q) \quad (10.1)$$

$$P(C_i | Q) = \frac{\exp \left( \sum_j \lambda_j f_j(Q, C_i) \right)}{\sum_C \exp \left( \sum_j \lambda_j f_j(Q, C) \right)} \quad (10.2)$$

其中,  $\lambda_j$  是根据语料训练得到的不同特征权重。

类似的, RCME 模型会在语法分析树上抽取单词之间的语法关系 (例如: 主谓关系等), 作为特征计算最适合回答问题的句子, 模型的提取特征方式与词性很相似, 这里就不再过多描述。

## 10.2.2 基于深度神经网络的阅读理解算法

随着深度学习方法在自然语言处理领域广泛应用, 特别是 2016 年斯坦福大学发布的阅读理解数据集 SQuAD<sup>[548]</sup>, 极大的推动了阅读理解的研究进展。大量基于深度学习算法的阅读理解方法也在 2016 年之后不断涌现。

## 1. 双向注意力流网络阅读理解算法

BiDAF<sup>[561]</sup> 是基于双向注意力流 (Bi-directional Attention Flow) 的阅读理解算法。针对的是 SQuAD 评测集合，给定文章和问题，算法需要返回文章中的一个片段作为答案，答案片段通常是短语。BiDAF 将该任务建模为将篇章和问题作为输入，预测开始位置和结束位置的问题。BiDAF 神经网络结构如图 10.5 所示，由六个神经网络层组成，分别为字符向量模块、词向量模块、注意力流层，上下文向量层、建模层以及输出层。在输入部分显式的区分篇章内容和问题，分别使用  $\{x_1, x_2, \dots, x_T\}$  和  $\{q_1, q_2, \dots, q_J\}$  表示，其中  $T$  和  $J$  分别是问题和文本的长度。

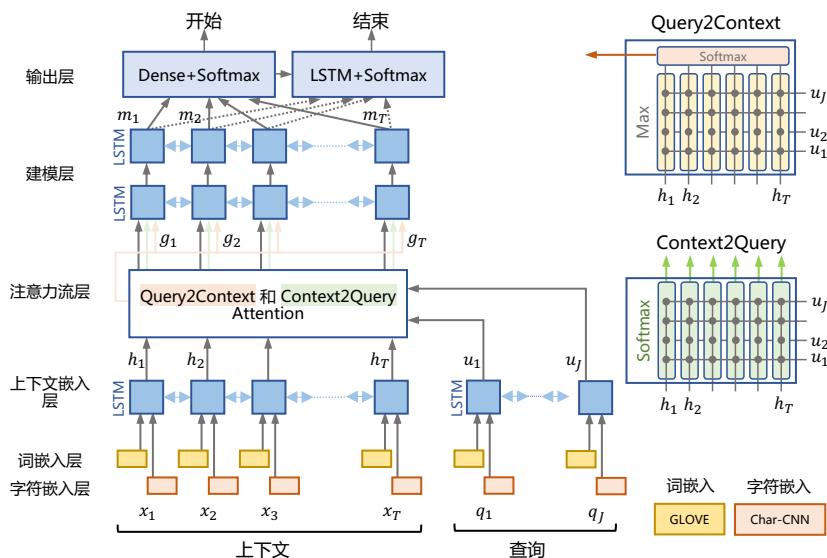


图 10.5 BiDAF 神经网络结构图<sup>[561]</sup>

**字符向量层 (Character Embed Layer)** 作用是利用卷积神经网络对文本和问题中每个单词的字符进行建模，计算得到每个单词相应的字符表示。

**词向量层 (Word Embed Layer)** 将问题和文章中的每个单词生成对应的词向量表示，BiDAF 使用了 Glove<sup>[179]</sup> 生成文本和问题的词向量，并构成文章表示矩阵  $\mathbf{X} \in \mathbb{R}^{d \times T}$  和问题表示矩阵  $\mathbf{Q} \in \mathbb{R}^{d \times J}$ 。

**上下文向量层 (Contextual Embed Layer)** 使用 Highway Networks<sup>[562]</sup> 对前两层中生成的字符向量和词向量进行融合，并将融合后的结果输入给一个 BiLSTM 网络，分别计算得到问题和文章的隐状态序列  $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_J] \in \mathbb{R}^{2d \times J}$  和  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T] \in \mathbb{R}^{2d \times T}$ 。由于这里使用了 BiLSTM，因此  $\mathbf{u}_i$  和  $\mathbf{h}_j$  都是由前向 LSTM 和后向 LSTM 的  $d$  维隐变量连接而成的  $2d$  维列向量。此外还需要注意的是，本层的计算过程中并没有对问题和文章进行信息交互。

注意力流层 (Attention Flow layer) 对文章和问题进行信息交互。首先根据文章表示矩阵  $\mathbf{H}$  和问题表示矩阵  $\mathbf{U}$  计算相似度矩阵  $\mathbf{S} \in \mathbb{R}^{T \times J}$ , 其中  $S_{ij}$  表示输入文本中第  $i$  个单词和给定问题中第  $j$  个单词之间的相似度:

$$S_{ij} = \alpha(\mathbf{h}_i, \mathbf{u}_j) \quad (10.3)$$

$$\alpha = \mathbf{W}_S^T [\mathbf{h}_i; \mathbf{u}_j; \mathbf{h}_i \circ \mathbf{u}_j] \quad (10.4)$$

其中  $\mathbf{W}_S \in \mathbb{R}^{6d}$  是一个可训练的模型参数矩阵,  $\mathbf{h}_i \circ \mathbf{u}_j$  是  $\mathbf{h}_i$  和  $\mathbf{u}_j$  之间的元素积 (Element-wise Product),  $[;]$  表示向量按行连接。

接下来模型分别计算篇章到问题注意力和问题到篇章注意力。篇章到问题注意力表示对于每一个篇章中的单词来说, 哪一个问题中的单词是最相关的。 $a_t$  表示第  $t$  个单词对所有问题单词的注意力权重, 其中  $\sum a_{tj} = 1$ ,  $\mathbf{a}_t$  通过  $\mathbf{S}$  计算而来:

$$\mathbf{a}_t = \text{Softmax}(\mathbf{S}_{t:}) \quad (10.5)$$

紧接着就可以求得注意力加权后的问题表示向量:

$$\tilde{\mathbf{U}}_{:t} = \sum_j a_{tj} \mathbf{U}_{:j} \quad (10.6)$$

问题到篇章注意力表示对于每个问题单词哪个文章单词拥有最高的相似度, 因此这对回答问题来说非常重要。类似地, 模型通过  $\mathbf{S}$  计算问题到篇章注意力:

$$\mathbf{b} = \text{Softmax}(\max_{\text{col}}(\mathbf{S})) \quad (10.7)$$

然后可以求得注意力加权后的文章表示向量:

$$\tilde{\mathbf{h}} = \sum_t \mathbf{b}_t \mathbf{H}_{:t} \quad (10.8)$$

表示向量向量通过加权后得到了问题最关注的文章信息。

最后, 对上下文向量和注意力向量进行拼接, 得到向量  $\mathbf{G}$ :

$$\mathbf{G}_{:t} = \beta(\mathbf{H}_{:t}, \tilde{\mathbf{U}}_{:t}, \tilde{\mathbf{h}}_{:t}) \quad (10.9)$$

其中,  $\mathbf{G}_{:t}$  表示第  $t$  列向量, 对应的是第  $t$  个文章单词,  $\beta$  是一个用来综合三个输入向量的可训练的向量函数。其中,  $\beta$  可以是任何一个可训练的神经网络, 例如多层次感知机或者简单的拼接。

建模层 (Modeling Layer), 将之前计算出的  $\mathbf{G}$  输入给建模层, 输出是建模层获取到的文章向

量和问题向量的信息交互。模型采用 BiLSTM 网络，网络输出矩阵  $M$  用来预测答案。

输出层（Output Layer），利用  $G$  预测答案的起始位置  $p^1$  和答案的结束位置  $p^2$ ：

$$p^1 = \text{Softmax}(\mathbf{W}_{p^1}^T[\mathbf{G}, \mathbf{M}]) \quad (10.10)$$

$$p^2 = \text{Softmax}(\mathbf{W}_{p^2}^T[\mathbf{G}, \mathbf{M}^2]) \quad (10.11)$$

其中， $\mathbf{W}_{p^1}$  和  $\mathbf{W}_{p^2}$  表示可训练的模型参数，特别地  $\mathbf{M}^2$  是基于  $\mathbf{M}$  再采用一次 BiLSTM 网络生成的另外一个隐向量序列，计算方法类似于建模层所使用的方法。

## 2. 基于门控自匹配注意力机制的阅读理解算法

R-Net<sup>[557]</sup> 是基于门控自匹配注意力机制的阅读理解模型，R-Net 在前人工作的基础上进行了改进和提升，取得了明显的效果。R-Net 的创新主要有两点：(1) 在 Match-LSTM 网络的基础上进行改进，引入了门控的机制来学习输入文本的哪一部分与问题的相关性最大；(2) 自匹配机制，本质是自注意力，充分学习了文本间词和词之间的关系，而不是只关注于文本和问题之间的关系，R-Net 之前的工作，只是利用一个双向 LSTM 对输入文本进行特征提取，而没有进行更深的挖掘。R-Net 的神经网络结构如图10.6所示。

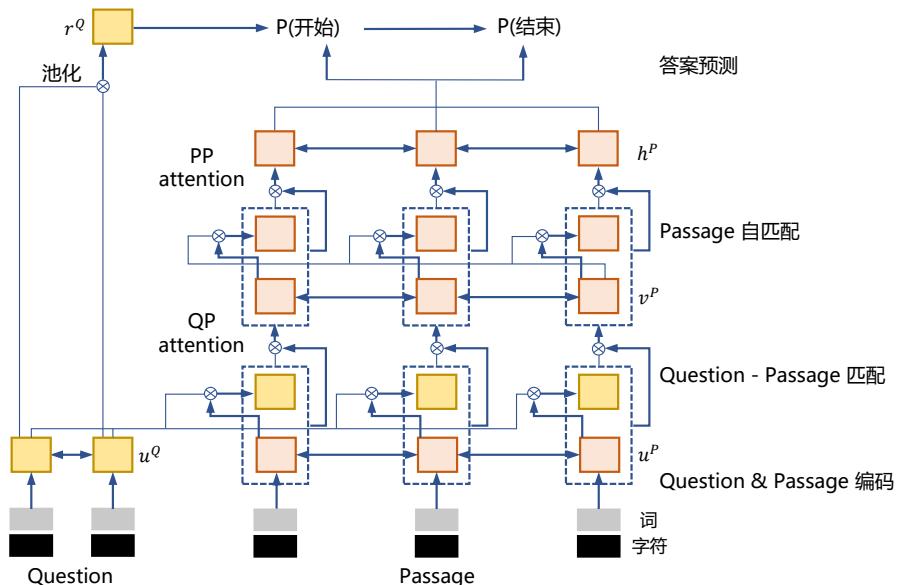


图 10.6 R-Net 神经网络结构图<sup>[557]</sup>

R-Net 采用两个不同的 BiGRU 网络分别作为问题编码器和文本编码器，对问题  $Q = [w_1^Q, \dots, w_m^Q]$  和文本  $P = [w_1^P, \dots, w_n^P]$  中每个单词所对应的词向量和字符向量表示进行编码，生成对应的隐状

态序列:

$$\mathbf{u}_t^Q = \text{BiGRU}(\mathbf{u}_{t-1}^Q, [\mathbf{e}_t^Q, \mathbf{c}_t^Q]) \quad (10.12)$$

$$\mathbf{u}_t^P = \text{BiGRU}(\mathbf{u}_{t-1}^P, [\mathbf{e}_t^P, \mathbf{c}_t^P]) \quad (10.13)$$

$\mathbf{u}_t^Q, \mathbf{u}_t^P$  分别表示问题和文本中第  $t$  个单词所对应的隐向量表示,  $\mathbf{e}_t^Q, \mathbf{e}_t^P$  分别表示问题和文本中第  $t$  个单词所对应的词向量,  $\mathbf{c}_t^Q, \mathbf{c}_t^P$  代表字符向量。

为了充分挖掘输入文本中与给定问题相关程度最大的部分, R-Net 对输入文本的特征提取上进行了改进, 不仅是对整个输入文本进行处理, 而是更细粒度的关注文本中的不同部分与问题的关系。为此, R-Net 模型采用了以门控注意力机制为基础的循环网络将给定问题与输入文本对应的向量表示相融合。首先, 将字符向量和单词向量连接在一起, 输入 RNN 网络, 对文章总体进行语义建模, 并通过注意力机制对上下文向量进行加强。得到上下文向量  $c$ :

$$\mathbf{v}_t^P = \text{RNN}(\mathbf{v}_{t-1}^P, [\mathbf{u}_t^P, \mathbf{c}_t]) \quad (10.14)$$

$$\mathbf{c}_t = \sum_{i=1}^m \mathbf{a}_i^t \mathbf{u}_i^Q \quad (10.15)$$

$$\mathbf{a}_i^t = \text{Softmax}(\mathbf{s}^t) \quad (10.16)$$

$$\mathbf{s}_j^t = \mathbf{v}^T \tanh(\mathbf{W}_u^Q \mathbf{u}_j^Q + \mathbf{W}_u^P \mathbf{u}_t^P + \mathbf{W}_v^P \mathbf{v}_{t-1}^P) \quad (10.17)$$

其中,  $\mathbf{W}_u^Q$ ,  $\mathbf{W}_u^P$  和  $\mathbf{W}_v^P$  都是可训练的参数矩阵。最后, 通过门控注意力的方式, 将问题信息和文本信息进行充分融合:

$$\mathbf{g}_t = \text{Sigmoid}(\mathbf{W}_g[\mathbf{u}_t^P, \mathbf{c}_t]) \quad (10.18)$$

$$\mathbf{u}_t^P = \mathbf{g}_t \cdot \mathbf{u}_t^P \quad (10.19)$$

$$\mathbf{c}_t = \mathbf{g}_t \cdot \mathbf{c}_t \quad (10.20)$$

$\mathbf{W}_g$  是可训练的参数矩阵。

模型采用自匹配注意机制生成文本向量表示序列  $\mathbf{H} = [h_1^P, \dots, h_n^P]$  对于阅读理解任务来说, 输入文本中每个单词的向量表示不仅仅与问题中的单词有关, 还应该与该单词的上下文有关, 自匹配注意力机制就是要将这两种信息融合起来。

$$\mathbf{h}_t^P = \text{BiRNN}(\mathbf{h}_{t-1}^P, [\mathbf{v}_t^P, \tilde{\mathbf{c}}_t]) \quad (10.21)$$

$$\tilde{\mathbf{c}}_t = \sum_{i=1}^n \mathbf{a}_i^t \mathbf{v}_i^P \quad (10.22)$$

$$\mathbf{a}_i^t = \frac{\exp(\tilde{s}_i^t)}{\sum_{j=1}^n \exp(\tilde{s}_j^t)} \quad (10.23)$$

$$\tilde{s}_j^t = \tilde{\mathbf{v}}^t \tanh(\tilde{\mathbf{W}}_v^P \mathbf{v}_j^P + \tilde{\mathbf{W}}_v^P v_t^P) \quad (10.24)$$

最后 R-Net 利用 Pointer Network 对答案的起始位置  $p^1$  和结束位置  $p^2$  进行预测：

$$\mathbf{p}^t = \text{argmax}(\mathbf{a}_1^t, \dots, \mathbf{a}_n^t) \quad (10.25)$$

$$\mathbf{a}_i^t = \frac{\exp(s_i^t)}{\sum_{j=1}^n \exp(s_j^t)} \quad (10.26)$$

$$\mathbf{s}_j^t = \mathbf{v}^T \tanh(\mathbf{W}_h^P \mathbf{h}_j^P + \mathbf{W}_h^a h_{t-1}^a) \quad (10.27)$$

$$\mathbf{h}_t^a = \text{RNN}(\mathbf{h}_{t-1}^a, \mathbf{c}_t) \quad (10.28)$$

$$\mathbf{c}_t = \sum_{i=1}^n \mathbf{a}_i^t \mathbf{h}_i^P \quad (10.29)$$

$h_{t-1}^a$  表示 Pointer Network 中最后一个隐状态变量，在预测答案其实位置时，其计算方法如下

$$\mathbf{h}_{t-1}^a = \sum_{i=1}^m \mathbf{a}_i \mathbf{u}_i^Q \quad (10.30)$$

$$\mathbf{a}_i = \frac{\exp(s_i)}{\sum_{j=1}^m \exp(s_j)} \quad (10.31)$$

$$\mathbf{s}_j = \mathbf{v}^T \tanh(\mathbf{W}_u^Q u_j^Q + \mathbf{W}_v^Q \mathbf{V}_r^Q) \quad (10.32)$$

其中， $\mathbf{W}_u^Q, \mathbf{W}_v^Q, \mathbf{W}_h^P, \mathbf{W}_h^a$  都是可训练的参数矩阵。

### 3. 基于片段掩盖的预训练的阅读理解算法

SpanBERT<sup>[563]</sup> 是一种可以更好的表示和预测片段的文本的预训练算法，也是对 BERT 原本预训练方法的一种扩展。SpanBERT 对 BERT 模型主要进行了如下改进：

- 不再是随机对单个词进行遮盖操作，而是对相邻的分词进行遮盖操作；
- 加入了 Span Boundary Objective (SBO) 作为训练目标。

原始 BERT 算法，在训练时会随机选取句中的子词进行遮盖，但这种训练方式，会让一些原本就连续出现且相关性很强的词组短语，在训练的时候被割裂开。SpanBERT 打破了这种范式，对于给定的文本  $X = x_1 \dots x_n$ ，SpanBERT 通过迭代的方式从中选择需要遮盖的片段，在每一次迭代中，SpanBERT 通过几何分布对遮盖长度  $l$  进行随机采样， $l$  代表要遮盖的单词的数量，紧接着在一个均匀分布中采样出一个遮盖起点  $s$ ，对  $\{x_s, \dots, x_{s+l}\}$  进行遮盖，在损失函数方面，这一部分与原本的 BERT 保持一致，采用 MLM 的损失函数进行训练。

另外, SpanBERT 引入了一个新的训练目标, Span Boundary Objective (SBO)。SpanBERT 希望训练好的语言模型可以让每个片段的边缘单词的表示尽可能的包含片段内部内容的所有语义, 所以 SBO 的含义是要求在片段边界处的单词表示尽可能可以对遮盖片段内部的单词进行预测。具体来说, 对于给定文本  $X$ , 对于某个被选中的遮盖片段  $\{x_s, \dots, x_{s+l-1}\}$ , SpanBERT 希望通过  $\{x_{s-1}, x_{s+l}\}$  以及位置向量  $p_{i-s+1}$  来对被遮盖的单词  $x_i$  进行预测:

$$y_i = f(h_{s-1}; h_{x_{s+l}}; p_{i-s+1}) \quad (10.33)$$

所以, 最终 SpanBERT 的损失函数为:

$$\mathcal{L}(x_i) = \mathcal{L}_{MLM}(x_i) + \mathcal{L}_{SBO}(x_i) \quad (10.34)$$

$$= -\log P(x_i|X) - \log P(x_i|x_{s-1}, x_{s+l}, p_i) \quad (10.35)$$

图10.7给出了一个片段掩盖的样例, 该例子中“an American football game”全部使用 [MASK] 替代,  $x_4$  和  $x_9$  以及位置编码用于预测掩盖区域内的单词。在本例中“football”是掩盖片段中的第 3 个单词, 因此采用  $p_3$  的位置编码。“football”所对应的损失函数为:

$$\mathcal{L}(\text{football}) = -\log P(\text{football}|x_7) - \log P(\text{football}|x_4, x_9, p_3) \quad (10.36)$$

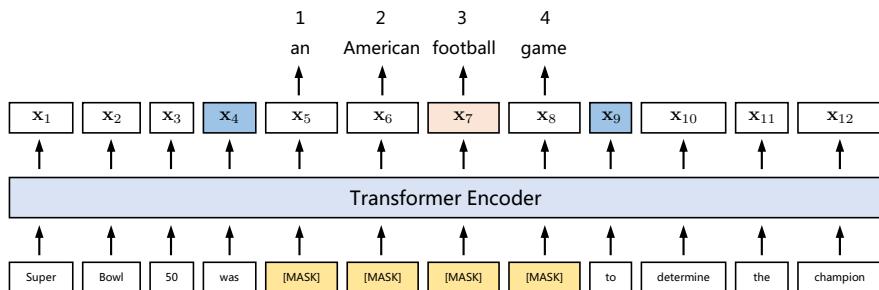
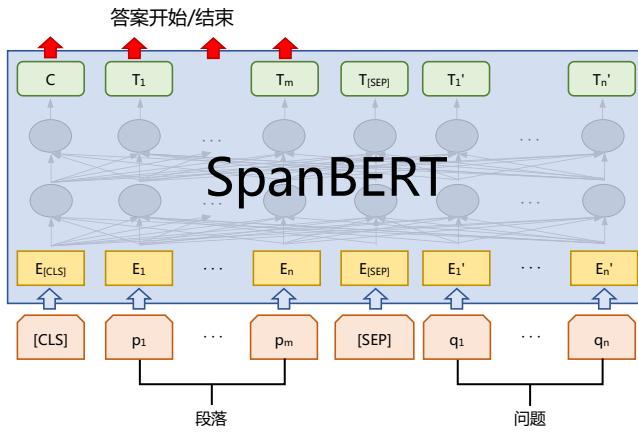


图 10.7 SpanBERT 神经网络片段掩盖样例<sup>[563]</sup>

在预训练完成的基础上, 采用与 BERT<sup>[29]</sup> 相同的建模方式, 将段落  $P = (p_1, p_2, \dots, p_m)$  和问题  $Q = (q_1, q_2, \dots, q_n)$  编码为一个序列  $X = [\text{CLS}]p_1p_2\dots p_n[\text{SEP}]q_1q_2\dots q_m$  作为输入。在段落单词所对应的位置, 训练两个线性分类器, 分别建模该单词是否为答案片段的开始还是结束位置。为了应对 SQuAD 2.0 中存在不能回答的问题的情况, 将答案片段的开头和结尾都标志于 [CLS] 所对应的位置。神经网络架构如图10.8所示。

图 10.8 SpanBERT 神经网络阅读理解框架<sup>[563]</sup>

### 10.2.3 阅读理解语料库

当前常用的阅读理解数据集如表10.2所示，其中 CNN/Daily Mail、HLF-RC 和 Children’s Book 是完形填空数据集。MCTest 和 RACE 是多项选择任务的数据集，RACE 是从中国初中高中的英语测试题中收集的。SQuAD、TriviaQA 和 NewsQA 是片段抽取的数据集。

表 10.2 阅读理解常用数据集

数据集	语言	任务类型	数据来源	问题数量	文档数量
CNN / Daily Mail	英语	完形填空	自动生成	140 万	30 万
Children’s Book	英语	完形填空	自动生成	68.8 万	68.8 万
HLF-RC	中文	完形填空	自动生成	10 万	2.8 万
MCTest	英语	多项选择	人工编写	2640	660
RACE	英语	多项选择	英文测试题	87 万	5 万
SQuAD	英语	片段抽取	人工编写	10 万	536
TriviaQA	英语	片段抽取	用户问答界面	4 万	66 万
NewsQA	英语	片段抽取	人工编写	10 万	1 万
MS-MARCO	英语	自由问答	用户日志	10 万	20 万
DuReader	中文	自由问答	用户日志	20 万	100 万

#### 1. SQuAD 数据集

SQuAD<sup>[548]</sup>是斯坦福大学于 2016 年推出的阅读理解数据集，给定一篇文章和相应问题，需要模型给出问题的答案，SQuAD 是片段抽取类任务，答案为文章中的片段。数据集的所有文章都选自维基百科，一共有 107,785 个问题，以及配套的 536 篇文章。最近，SQuAD2.0 已发布，其中包括

含无法回答的问题。SQuAD 的主要缺点是数据集中，问题的答案一般是很简短的一个词组。

## 2. MS-MARCO 数据集

MS MARCO<sup>[564]</sup> 又称人类生成的机器阅读理解数据集，由 Microsoft AI & Research 设计和开发。数据集中的所有问题都是从真实的匿名用户查询中获得的，研究者们为这些问题编写了答案。问题配套的文章是通过 Bing 搜索引擎从真实文档中提取可以获取答案的上下文段落。该数据集有 1,010,916 个数据。

## 3. TriviaQA 数据集

TriviaQA 数据集是一个难度较高的阅读理解数据集，其中问题较为复杂，并且问题和文章中相应的支持句之间有较大的差距，例如一些句法或者词法的变化，并且相比其他阅读理解数据集，TriviaQA 中包含更多的需要多跳推理能力的问题。

## 4. RACE 数据集

RACE 数据集是一个从考试中提取的数据集，旨在模拟真实的人类测试，RACE 是第一个基于真实考试的大型数据集。RACE 的数据分为两个部分 RACE-M 和 RACE-H，两个部分的数据分别是从初中和高中的练习题中进行采集而成，RACE-H 中的文本长度和词汇量都比 RACE-M 要长，但总体而言，这些数据的句子长度和复杂度都是比新闻文章或维基百科文章简单。

## 10.3 表格问答

表格是一种常见的数据存储形式，通常由表头，表格单元和表格标题构成：表格标题概括了表格包含的主要内容，表头一般表示某行或者某列表格单元的内容和类型。除此之外，表格中每个元素都是一个表格单元。图10.9给出了表格问答样例。对于问题“哪个城市举办了 2008 年奥运会？”，通过表格选择和单元格抽取，最终返回“北京”作为答案。

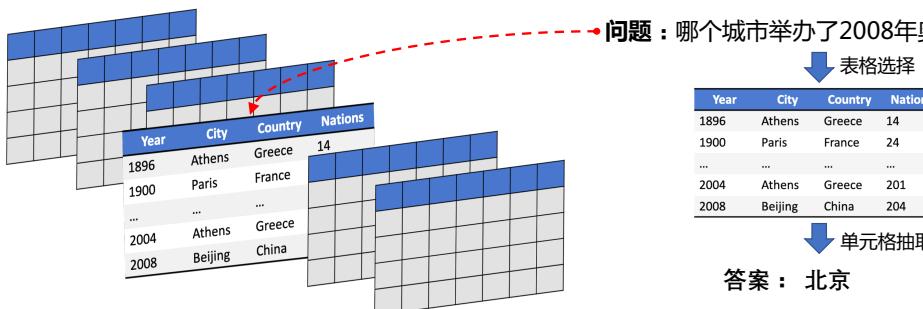


图 10.9 表格问答样例

### 10.3.1 基于特征的表格问答方法

对于解决表格问答问题，一个基本的想法是将输入的表格视为一个知识库，但这个过程本身包含着很多挑战，因为知识库包含了太多的关系，这令表格数据充满噪音。

CSP 算法<sup>[556]</sup> 是基于特征的表格问答方法，采用逻辑形式驱动。对于一个给定的表格  $t$  和一个关于表格的问题  $q$ ，模型需要根据表格内容输出答案  $y$  来回答  $q$ 。对于问题  $q$  唯一的限制是这个问题需要仅能够通过表格内容回答，问题可以任何形式。CSP 算法的整体流程如图10.10所示。

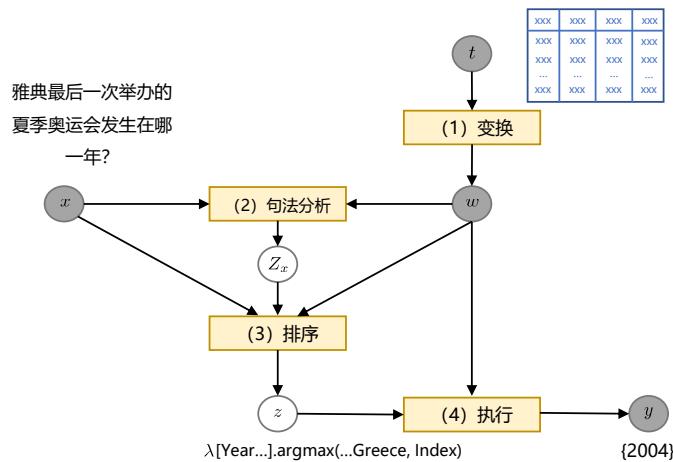


图 10.10 CSP 模型结构图<sup>[556]</sup>

为了解决表格知识库的噪音问题，CSP 模型首先将表格  $t$  转化成知识图谱的形式。具体来说，将表格的行视为行节点，表格单元中的字符串被视为实体节点，表格的列被转化成连接行节点和实体节点的有向边，而列表头则成为了这些边的标签。得到知识图谱  $w$  之后，CSP 模型将问题  $q$  解析成一些候选的逻辑形式  $Z_x$ 。CSP 模型提出了表单检索的方法，对于给定的问题  $q = \{q_1, \dots, q_n\}$ ，模型通过两类规则对问题进行解析：

$$(q, i, j)[s] -> (c, i, j)[f(s)] \quad (10.37)$$

$$(c_1, i, k)[z_1] + (c_2, k+1, j)[z_2] -> (c, i, j)[f(z_1, z_2)] \quad (10.38)$$

第一个规则是一类词法匹配规则，可以将问题片段  $q_i \dots q_j$  转化为逻辑形式，其中  $c$  表示类别， $s$  表示问题片段的文本。例如， $s$  为“纽约”， $c$  为“实体”时，根据表格查询内容  $f(s)$  为“纽约市”。第二个规则将两个相邻片段的逻辑形式  $z_1$  和  $z_2$  合并成一个新的逻辑形式  $f(z_1, z_2)$ 。

CSP 模型将以上特征  $g(x, w, z)$  输送给逻辑线性模型去捕捉问题  $q$  和候选  $z$  之间的关系。CSP 模型利用解析生成出逻辑形式  $Z_x$ ，每一个逻辑形式  $z_i$  都是一个从知识图中抽取出的图问题，利

用提取出的特征向量  $g(x, w, z)$ , CSP 模型定义了对于每个逻辑线性分布:

$$p(z|x, w) = \exp(\theta^T g(x, w, z)) \quad (10.39)$$

模型的训练方式如下, 对于给定的训练样本  $D = \{(x_i, t_i, y_i)\}_{i=1}^N$ , CSP 模型在最大似然估计的基础上增加正则项作为目标函数:

$$L = \frac{1}{N} \sum_{i=1}^N \log p(y_i | q_i, w_i) - L_1 \quad (10.40)$$

其中  $L_1$  是模型的 L1 正则,  $w_i$  是根据  $t_i$  动态生成的:

$$p(y|x, w) = \sum_{z \in Z_x} p(z|x, w) \quad (10.41)$$

在训练完毕之后, 模型挑选分数最高的逻辑形式  $z$  作为答案。

### 10.3.2 基于深度学习的表格问答模型

CLTR<sup>[565]</sup> 是基于预训练的端到端表格问答模型, 该方法首先从大量表格中筛选得到少量相关表格, 再对这些表格进行重新排序, 最终回答用户提出的自然语言问题。其模型神经网络结构如图10.11所示。

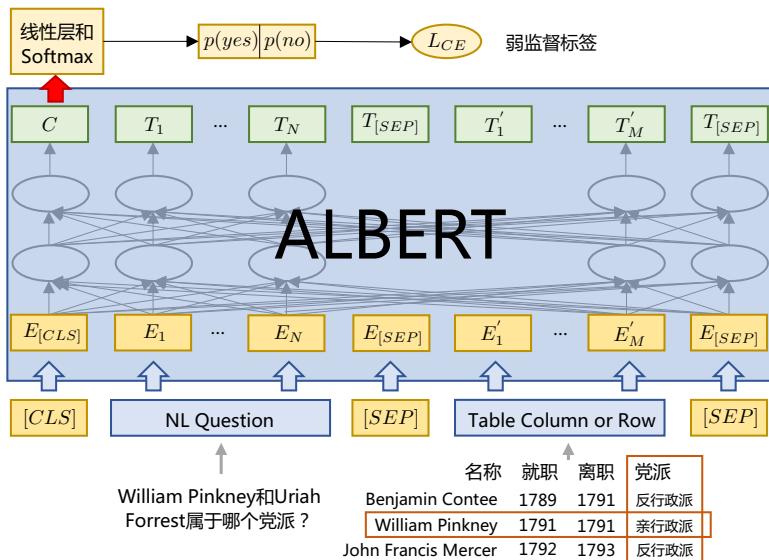


图 10.11 CLTR 模型神经网络结构图<sup>[565]</sup>

CLTR 模型首先利用 BM25 算法筛选出一个高相关表格池，接着采用行列交互模型对表格池中表格中的行列生成包含答案的概率。行列交互模型将表格问答分为两个基本操作：行选择和列选择，将行列的预测分布进行合并，就能得到答案单元格的目标概率。CLTR 采用文本分配的方式预测答案所在的行列，具体来说就是将每行每列分别与问题  $q$  相连输入 BERT 模型中，通过二分类器得到概率分数：

$$\mathbf{H}_{r_i} = \text{BERT}(\mathbf{q}, \mathbf{r}_i) \quad (10.42)$$

$$\mathbf{H}_{c_j} = \text{BERT}(\mathbf{q}, \mathbf{c}_j) \quad (10.43)$$

$$\mathbf{P}_{r_i} = \text{Binary Classifier}(\mathbf{H}_{r_i}) \quad (10.44)$$

$$\mathbf{P}_{c_j} = \text{Binary Classifier}(\mathbf{H}_{c_j}) \quad (10.45)$$

至此，在表格池  $T$  中每张含有  $n$  行  $m$  列的表格  $t$  都含有两个分数集合  $S_c = \{s_{c_1}, \dots, s_{c_m}\}$  和  $S_r = \{s_{r_1}, \dots, s_{r_n}\}$ 。紧接着对整张表格计算最大的单元分数：

$$S_t = \max(S_c) + \max(S_r) \quad (10.46)$$

CLTR 模型利用最大单元分数对表格池里的候选表格进行重排序，当重排序做完之后，会挑选表格池中得分前  $k$  的表格返回给用户。最后的答案由行列交互模型所计算的分数，寻找分数最高的交叉行列来得到目标表格单元。

### 10.3.3 表格问答语料库

当前常用的表格问答语料库如表10.3所示。其中 WikiSQL 和 Spider 数据集合采用用户真实问句，并人工转换为数据库 SQL 查询语句的方式。WikiTableQuestions 则是来源于维基百科的半结构化表格。

表 10.3 表格问答常用数据集

数据集	数据来源	表格数量	问句数量
WikiSQL	用户真实问句转换	24241	80645
Spider	用户真实问句转换	206	10181
WikiTableQuestions	维基百科	2108	22033

#### 1. WikiSQL

WikiSQL 数据集是 2017 年由 Salesforce 提出的大型标注数据集，WikiSQL 是一类 NL2SQL 数据集，该任务要求模型将用户的自然语言询问转换成 SQL 语句并搜索出答案，WikiSQL 是目前规模最大的 NL2SQL 数据集。WikiSQL 包含了 24241 张表，80645 条自然语言问句及相应的 SQL

语句。

## 2. Spider

Spider 数据集由耶鲁大学于 2018 年提出，Spider 同样是一个 NL2SQL 数据集，其中包含了 10181 条自然语言问句和分布在多个独立数据库中的共 5693 条 SQL 语句，数据集中的内容涉及 138 个不同的领域，相比 WikiSQL，虽然数据量不足，但 Spider 更贴近真实场景，所以难度更大。

## 3. WikiTableQuestions

WikiTableQuestions 数据集是斯坦福大学提出的一个基于维基百科中半结构化表格问答的数据集，其中包含 22033 条来自真实用户的问句和 2108 张表格，表格中的数据都是取自维基百科的真实数据，表格中的内容往往具有多重含义，并且覆盖了多个领域的数据。特别的是，该数据测试集中的表格主题和实体关系都是训练集中没出现过的。

## 10.4 社区问答

随着 Web 2.0 的快速发展，社区问答网站自 2008 年以来蓬勃发展，出现了包括 Quora、知乎等在内的众多社区问答网站。用户可以通过这些网站对自己需要解决的问题进行提问，其他用户对这些问题进行回答、关注、评论等。同时，企业内部也大量问答知识，例如人工客服领域，很多高频问题都有相应的知识库。基于这些问答数据，研究人员们提出了社区问答任务，基于收集的 $\langle$ 问题，答案 $\rangle$ 对数据  $\mathcal{D} = \langle Q_i, A_i \rangle_{i=1}^N$ ，针对用户自然语言问题  $Q$  寻找最合适答案返回。图10.12给出了搜索引擎给出的社区问答样例，对于问题“手机屏碎了怎么办？”，通过与问答库中的问答对语义相关性进行计算，最终返回相关答案。



图 10.12 社区问答样例（来源：搜狗搜索）

社区问答方法可以进一步细分为两个子任务：

- (1) 问题-问题匹配：目标是计算输入问题  $Q$  和问答库  $\mathcal{D}$  中某个已有问题  $Q_i$  之间的相似度，如果两者之间相似度较高，显而易见的， $Q_i$  的答案  $A_i$  就有很大概率是  $Q$  的答案。

- (2) 问题-答案匹配：目标计算输入问题  $Q$  与  $D$  中某个答案  $A_i$  的相关性，同样的，如果两者之间相关度很高，那么  $A_i$  有很大概率是  $Q$  的答案。

本节以问题-问题匹配任务为重点，分别介绍基于特征和基于深度学习的语义匹配算法。

#### 10.4.1 基于特征的语义匹配

传统的基于特征的语义匹配方法普遍采用一些基于词袋的统计文本表示，通过这些表示去计算两段文本的语义相似度，进而判断两段文本的语义是否匹配。其中， $n$  元短语匹配度是判断语义相似度的一个重要衡量指标，如果两个问题共同包含的语义片段越多，自然具有更大的可能表达相同的含义。 $n$  元短语匹配度有许多不同的计算方式，最简单的做法可以计算两个问题中重复单词的数量：

$$\text{Sim}(Q_1, Q_2) = \frac{\sum_{w \in Q_1} \text{num}(w, Q_2)}{|Q_1|} \quad (10.47)$$

其中  $\text{num}(w, Q_2)$  表示单词  $w$  在  $Q_2$  中出现的次数， $|Q_1|$  表示  $Q_1$  的长度。

类似地，可以将简单的单词匹配扩展到短语匹配，由于短语包含的信息量更多，表达的语义也更加独立，所以短语级匹配特征能更好的表达两个问题之间的语义相似度。常见的短语级匹配特征有 BLEU、编辑距离和最长公共子序列方法都可以将它们用来当作特征来完成问题匹配的任务。

但是，上面介绍的这些语义匹配特征都存在两个缺点：首先，没有处理停用词，例如冠词、介词等，这类词汇虽然在文本中高频出现，但是并不具备实际意义，对语义的贡献程度小。第二，没有处理同义词，由于自然语言的表达非常丰富，同一语义具有非常多的表达方式，这就使得处理语义匹配问题时需要能够对同义词有较好的识别能力。通常采用 FastText、BM25、TF-IDF、SentVec 等方法去解决这些问题，由于大部分方法前面的章节已经介绍过，本章不再赘述，本节主要介绍 BM25 算法。BM25 可以缓解停用词带来的问题，BM25 将基于反转文档频率计算  $Q_1$  和  $Q_2$  的相似度：

$$f_{BM25}(Q_1, Q_2) = \sum_{i=1}^{|Q_1|} \text{IDF}(Q_1^i) \text{freq}(Q_1^i, Q_2) \quad (10.48)$$

$$\text{freq}(Q_1^i, Q_2) = \frac{\text{num}(Q_1^i, Q_2)(k_1 + 1)}{\text{num}(Q_1^i, Q_2) + k_1(1 - b + b * \frac{|Q_2|}{\text{avglen}(Q_2)})} \quad (10.49)$$

$$\text{IDF}(Q_1^i) = \log \frac{N - n(Q_1^i) + 0.5}{n(Q_1^i) + 0.5} \quad (10.50)$$

其中， $Q_1^i$  表示  $Q_1$  中第  $i$  个单词， $\text{IDF}(Q_1^i)$  表示  $Q_1^i$  的反转文档频率， $\text{num}(Q_1^i, Q_2)$  表示单词  $Q_1^i$  在  $Q_2$  中出现的次数， $\text{avglen}(Q_2)$  表示问题及何种问题的平均长度， $N$  表示文档综述， $n(Q_1^i)$  表示

包含单词  $Q_1^i$  的文档综述,  $k_1$  和  $b$  为参数。最后, 调转  $Q_1$  和  $Q_2$  的位置可以计算得出  $f_{BM25}$ :

$$BM25(Q_1, Q_2) = \frac{f_{BM25}(Q_1, Q_2) + f_{BM25}(Q_2, Q_1)}{2} \quad (10.51)$$

通过 BM25 的特征值, 可以通过检测句对之间的  $bm25$  得分是否超过阈值  $\lambda$  来判断两个问题之间是否匹配。此外, 也可以利用 WordNet 等知识库来部分解决同义词的问题。

### 10.4.2 基于深度神经网络的问题匹配

基于深度学习的语义匹配方法主要分为交互式和非交互型两种。非交互型的方法是将两个句子分别编码后进行匹配, 而交互式的方法一般是将两个句子一起进行编码, 在编码过程中让两个句子进行信息交互。本节中将分别介绍上述两类语义匹配算法。

#### 1. DSSM 语义匹配算法

DSSM (Deep Structured Semantic Model)<sup>[566]</sup> 是一种非交互式的基于深度网络的文本语义匹配算法, 通过深度神经网络对用户问题和题目分别进行建模, 将其编码成低维度的语义向量, 然后计算两个向量的余弦距离, 判断两个句子之间的语义相似度, 由于该模型对用户问题和题目分别进行建模, 在计算相似度之前两者之间没有交互, 这种模型架构又被称作“双塔模型”。DSSM 模型网络结构如图10.13所示。

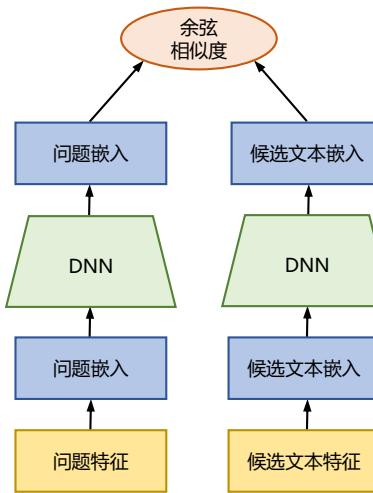


图 10.13 DSSM 模型图<sup>[566]</sup>

令  $x$  表示输入向量,  $y$  表示输出向量,  $y$  向量的计算方式如下:

$$l_1 = \mathbf{W}_1 x \quad (10.52)$$

$$\mathbf{l}_i = f(\mathbf{W}_i \mathbf{l}_{i-1} + \mathbf{b}_i), i = 2, \dots, N-1 \quad (10.53)$$

$$\mathbf{y} = f(\mathbf{W}_N \mathbf{l}_{N-1} + \mathbf{b}_N) \quad (10.54)$$

其中  $l_i$  表示隐藏层，共  $N$  层， $\mathbf{W}_i$  表示第  $i$  层的参数矩阵， $b_i$  表示偏置项， $f$  是输出层和隐藏层的激活函数，模型采用  $\tanh$  函数作为激活函数：

$$f(\mathbf{x}) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (10.55)$$

在社区问答任务中使用  $Q$  表示用户查询， $t$  表示候选文档题目，两者之间的相关性分数计算方式如下：

$$\text{Score}(Q, t) = \cos(\mathbf{y}_Q, \mathbf{y}_t) \quad (10.56)$$

DSSM 模型利用点击日志中用户搜索的问题和用户点击的文档作为数据进行训练。其基本假设是：如果用户在当前的搜索问题下点击了某个文档，那么该文档和问题之间一定是相关的。利用该假设通过搜索日志自动构建训练集和测试集。模型在学习过程中利用极大似然估计作为损失函数：

$$\mathcal{L} = -\log \prod_{Q, \mathcal{D}^+} P(\mathcal{D}^+ | Q) \quad (10.57)$$

$$P(\mathcal{D}^+ | Q) = \frac{\exp(\text{Score}(Q, \mathcal{D}^+))}{\sum_{d \in \mathcal{D}} \exp(\text{Score}(Q, t))} \quad (10.58)$$

其中， $\mathcal{D}$  表示候选文档集合， $\mathcal{D}^+$  为正样本。

## 2. ESIM 语义匹配算法

为解决 DSSM 模型缺乏对用户问题和题目中单词或短语间交互的问题，文献 [567] 提出了 ESIM 模型。该模型基于链式 LSTMs 设计了序列推断模型，并考虑了局部推断和推断组合的问题，在模型中引入了句子间的注意力机制来实现局部推断进而实现全局推断。ESIM 模型结构如图10.14所示，主要分为三个部分：输入编码层、局部推断模块以及推断组合模块。

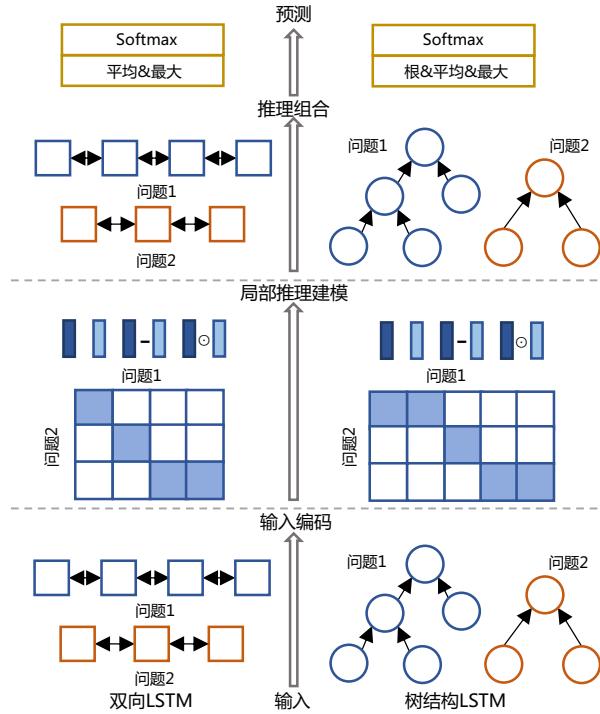
输入编码层对于输入的两个问题  $Q_1$  和  $Q_2$ ，模型首先通过词嵌入表示得到问题中每个单词的词向量，然后将其输入 BiLSTM 模型，通过 BiLSTM 表达句子的局部信息：

$$\mathbf{h}_1^i = \text{BiLSTM}(\mathbf{h}_1^{i-1}, \mathbf{h}_1^{i+1}, \mathbf{q}_1^i) \quad (10.59)$$

$$\mathbf{h}_2^i = \text{BiLSTM}(\mathbf{h}_2^{i-1}, \mathbf{h}_2^{i+1}, \mathbf{q}_2^i) \quad (10.60)$$

其中  $\mathbf{q}_1^i$  和  $\mathbf{q}_2^i$  分别表示  $Q_1$  中的第  $i$  个单词和  $Q_2$  中的第  $i$  个单词的嵌入表示。

局部推断模块在进行局部推断之前，首先模型要对  $\mathbf{h}_1$  和  $\mathbf{h}_2$  进行对齐，首先计算两者的相

图 10.14 ESIM 模型图<sup>[567]</sup>

似度:

$$\mathbf{e}_{ij} = \mathbf{h}_1^T \mathbf{h}_2 \quad (10.61)$$

紧接着进行局部推断，结合  $\mathbf{h}_1$ ,  $\mathbf{h}_2$  和  $\mathbf{h}_{ij}$ ，得到相似性加权后的向量：

$$\hat{\mathbf{h}}_1^i = \sum_{j=1}^{len2} \frac{\exp(\mathbf{e}_{ij})}{\sum_{k=1}^{len2} \mathbf{e}_{ik}} \mathbf{h}_2^j \quad (10.62)$$

$$\hat{\mathbf{h}}_2^i = \sum_{j=1}^{len1} \frac{\exp(\mathbf{e}_{ij})}{\sum_{k=1}^{len1} \mathbf{e}_{ik}} \mathbf{h}_1^j \quad (10.63)$$

在局部推断之后，模型会进行局部信息增强，利用差和点积放大两者的差异性：

$$\mathbf{m}_1 = [\mathbf{h}_1, \hat{\mathbf{h}}_1, \mathbf{h}_1 - \hat{\mathbf{h}}_1, \mathbf{h}_1 \cdot \hat{\mathbf{h}}_1] \quad (10.64)$$

$$\mathbf{m}_2 = [\mathbf{h}_2, \hat{\mathbf{h}}_2, \mathbf{h}_2 - \hat{\mathbf{h}}_2, \mathbf{h}_2 \cdot \hat{\mathbf{h}}_2] \quad (10.65)$$

推断组合模块首先将  $\mathbf{m}_1$  和  $\mathbf{m}_2$  输入 BiLSTM 提取信息,

$$\mathbf{v}_1^i = \text{BiLSTM}(\mathbf{m}_1^i) \quad (10.66)$$

$$\mathbf{v}_2^i = \text{BiLSTM}(\mathbf{m}_2^i) \quad (10.67)$$

$$(10.68)$$

由于两个问题的长度不同导致  $\mathbf{m}_1, \mathbf{m}_2$  矩阵维度不一致, ESIM 分别采用 MaxPooling 和 AvgPooling 的池化对两个句子进行处理,

$$\mathbf{v}_1^{max} = \text{MAXPooling}(\mathbf{v}_1) \quad (10.69)$$

$$\mathbf{v}_1^{avg} = \text{AvgPooling}(\mathbf{v}_1) \quad (10.70)$$

$$\mathbf{v}_2^{max} = \text{MAXPooling}(\mathbf{v}_2) \quad (10.71)$$

$$\mathbf{v}_2^{avg} = \text{AvgPooling}(\mathbf{v}_2) \quad (10.72)$$

最后将经过处理的向量拼接起来, 连接一个全连接层进行预测:

$$\mathbf{v} = [\mathbf{v}_1^{max}; \mathbf{v}_1^{avg}; \mathbf{v}_2^{max}; \mathbf{v}_2^{avg}] \quad (10.73)$$

$$\mathbf{o} = \text{SoftMax}(\mathbf{W}\mathbf{v} + \mathbf{b}) \quad (10.74)$$

### 10.4.3 社区问答数据集

当前常用的社区问答数据集如表格10.4所示, 其中 MRPC 和 QQP 是最常用的评价语义匹配模型的基准数据集, LCQMC 是由哈尔滨工业大学构建的中文语义匹配数据集。

表 10.4 社区问答常用数据集

数据集	数据来源	训练集	测试集
MRPC	新闻抽取	3.7k	1.7k
QQP	社区问答抽取	367k	390k
LCQMC	真实对话抽取	239k	12k
PAWS-X	翻译构造	49k	2k

#### 1. MRPC

MRPC(The Microsoft Research Paraphrase Corpus, 微软研究院意译数据集) 数据集是由微软在 2005 年提出的, 数据集包含 5800 个句子对, 所有句子都是从真实的在线新闻中抽取, 由人工注释句对中的句子是否具有相同的语义。该数据集中每个样本的句子长度都很长, 并且正负样本的比

例不均衡，其中正样本的比例约为 68%。

## 2. QQP

QQP(The Quora Question Pairs, Quora 问题对数据集) 是基于社区问答网站 Quora 构建的数据集，研究人员在 Quora 上抽取了大量的问题对集合作为数据，由人工标注一对问题是否包含相同的语义。QQP 数据集的数据规模非常大，训练集，开发集和测试集分别包含 363870, 40431, 390965 个样本。与 MRPC 类似，QQP 数据集也是样本不均衡的，其中负样本占总数的 63%。

## 3. LCQMC

LCQMC 是哈尔滨工业大学构建的问题语义匹配数据集，其目标是判断两个问题的语义。输入的句子具有高度口语化的特征，这大大提升了语义匹配的难度。输入是两个句子，输出是 0 或 1。其中 0 代表语义不相似，1 代表语义相似。训练集，开发集和测试集分别包含 238766, 8802, 12500 个样本。

## 4. PAWS-X

PAWS-X 是由谷歌公司发布的同义句识别数据集，需要模型识别一对句子是否含有相同的语义，其中文本具有高度重叠的词汇，重点考察模型对高层语义的识别能力。训练集，开发集和测试集分别包含 49401, 2000, 2000 个样本。

## 10.5 开放领域问答

开放领域问答是与搜索引擎类似的形式，对于一个给定的询问  $Q$ ，开放领域问答系统需要先从一个规模非常大的文档集合或知识库（如 Wikipedia 或互联网） $D = \{d_1, \dots, d_n\}$  中选择与问题最相关的文档  $d$ ，然后通过文档  $d$  中包含的信息，给出询问  $Q$  的答案  $A$ 。从任务定义中可以发现，开放问答的结果过程一般分为两个阶段：文档检索和阅读理解。图10.15给出了开放领域问答搜索引擎中的结果，用户输入查询“血压 170 严重吗”，搜索引擎定位了相关文档，并将直接输出最终答案。



图 10.15 开放领域问答样例（来源：百度）

### 10.5.1 检索-阅读理解架构的开放问答模型

由于开放问答任务面对的是大量的文档或网页，为了兼顾准确率与性能，开放问答一般会采用文档检索方法快速筛选出与询问最相关的文档，在筛选出相关文档之后再通过阅读理解的方式给出问题的答案，这种架构被称为检索-阅读理解架构。该架构的基本流程框架如图10.16所示，主要包含文章检索器和文章阅读器两个模块，此外有些算法中还包含对检索得到的文档的后处理以及对抽取答案的后处理。本节中将介绍两种基于该架构的模型。

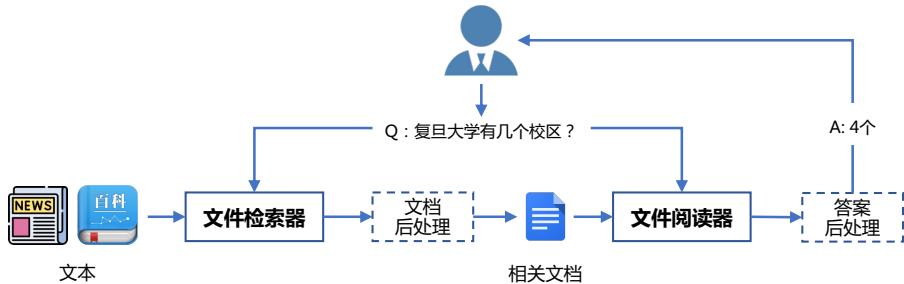


图 10.16 检索-阅读理解架构的开放问答流程框架图<sup>[568]</sup>

#### 1. DrQA 开放领域问答算法

DrQA<sup>[569]</sup>是利用维基百科作为知识库来实现开放式问答系统，DrQA 算法介绍对于任何事实性问题，系统都能从维基百科中找到一个包含答案的文章，在此基础上通过阅读理解算法从相关文章中提取答案。

筛选相关文档的关键是对于一个给定的查询和一个给定的文档，需要定义一个可以衡量该查询与该文档相关性的指标。常用的方法有 TF-IDF 和 BM25，其中，DrQA 采用的是 TF-IDF 的方法。利用 TF-IDF 可以得到文档  $d_i$  的关键词集合  $K_d = \{kd_1, \dots, kd_n\}$  和问题的关键词集合  $K_q = \{kq_1, \dots, kq_m\}$ 。利用关键词集合可以对问题和文档的相关性进行计算。定义  $K = K_d \cup K_q = \{k_1, \dots, k_l\}$ ，利用 TF-IDF 计算公式可以计算得出  $TF-IDF(k, d)$  和  $TF-IDF(k, q)$  的值，从而可以得到向量：

$$\mathbf{v}_d = [TF-IDF(k_1, d), \dots, TF-IDF(k_l, d)] \quad (10.75)$$

$$\mathbf{v}_q = [TF-IDF(k_1, q), \dots, TF-IDF(k_l, q)] \quad (10.76)$$

DrQA 利用 TF-IDF 的方法表示每个问题，并利用余弦相似度从大规模语料库中筛选出 5 篇最相关的文章。

答案抽取模块采用多层 RNN 网络，给定一个问题  $q = \{q_1, \dots, q_l\}$  和包含  $n$  个段落的相关文章，每个段落由  $m$  个词组成  $p = \{p_1, \dots, p_m\}$ 。对于段落中的每一个词  $p_i$ ，DrQA 采用 Glove 词向

量将其映射成一个 300 维的特征向量并把他们输入到 RNN 中并得到上下文向量  $H = h_1, \dots, h_m$ :

$$\mathbf{H} = \text{RNN}(\mathbf{p}_1, \dots, \mathbf{p}_m) \quad (10.77)$$

类似的，DrQA 采用 RNN 网络对问题  $q$  进行编码得到问题的上下文向量  $H^q = \{h_1^q, \dots, h_l^q\}$ 。

在预测时，DrQA 目标是在  $p$  中找到一个片段  $\{p_l, \dots, p_r\}$  作为问题的答案。所以，DrQA 将编码好的段落表示向量  $H$  和问题表示向量  $H^q$  作为输入，分别训练两个二分类器预测答案的开始位置和结束位置：

$$P_{start}(i) = \exp(\mathbf{p}_i \mathbf{W}_s \mathbf{q}) \quad (10.78)$$

$$P_{end}(i) = \exp(\mathbf{p}_i \mathbf{W}_e \mathbf{q}) \quad (10.79)$$

其中， $\mathbf{W}_s$  和  $\mathbf{W}_e$  是可训练的参数矩阵。最后，DrQA 选择最佳的文章片段  $\{p_l, \dots, p_r\}$  使得  $P_{start}(l) \times P_{end}(r)$  最大。

## 2. DPR 开放领域问答算法

上节中介绍的 BM25、TF-IDF 等稀疏向量空间模型，通过词频和逆文档频率，对问题和文档利用关键词进行匹配，无法理解文档和问题中包含的语义信息。例如，对于问题“谁是指环王中的坏人”和文档“萨拉贝克是指环王中最有名的反派角色”，稀疏向量空间模型无法匹配“坏人”和“反派角色”这两个词，但其实在语义角度上，文档和问题是非常相关的。基于密集向量空间的 DPR (Dense Passage Retriever) 模型就试图解决上述问题的一种方法。

DPR<sup>[570]</sup> 模型主要在文章检索模块上做出了改进，DPR 采用了双编码器的结构，可以在  $M$  个文档中找到与问题  $Q$  最相关的  $K$  个文档，DPR 主要分为两个部分：编码和相似度度量模块。DPR 模型的结构如图10.17所示。

在编码部分，DPR 采用了双编码器结构，部署了两个独立的 BERT 模型对问题和文档分别进行编码，将两段文本分别映射到两个  $d$  维向量中：

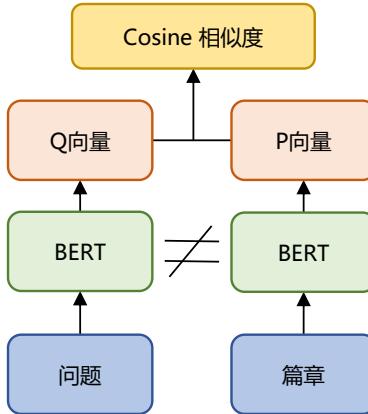
$$\mathbf{h}_q = \text{BERT}_1(\mathbf{q}) \quad (10.80)$$

$$\mathbf{h}_p = \text{BERT}_2(\mathbf{p}) \quad (10.81)$$

$$(10.82)$$

其中， $p$  是文档库中的某篇文章， $q$  是用户提出的问题，两个 BERT 分别独立。紧接着 DPR 采用点积的方式对两个向量的相似度进行度量：

$$\text{sim}(\mathbf{q}, \mathbf{p}) = \mathbf{h}_q^T \cdot \mathbf{h}_p \quad (10.83)$$

图 10.17 DPR 模型结构图<sup>[570]</sup>

DPR 的编码模型采用一个正例文档和 n 个负例文档的方式进行训练，整个训练集的定义方式如下：

$$\mathcal{D} = \langle \mathbf{q}_i, \mathbf{p}_i^+, \mathbf{p}_{i,1}^-, \dots, \mathbf{p}_{i,n}^- \rangle_{i=1}^m \quad (10.84)$$

其中  $\mathbf{q}_i$  表示第 i 个样本中用户提出的问题， $\mathbf{p}_i^+$  表示该样本中的正确文档， $\mathbf{p}_i^-$  表示第 i 个负例，训练的损失函数如下：

$$\mathcal{L}(\mathbf{q}_i, \mathbf{p}_i^+, \mathbf{p}_{i,1}^-, \dots, \mathbf{p}_{i,n}^-) = -\log \frac{e^{sim(\mathbf{q}_i, \mathbf{p}_i^+)}}{e^{sim(\mathbf{q}_i, \mathbf{p}_i^+)} + \sum_{j=1}^n e^{sim(\mathbf{q}_i, \mathbf{p}_{i,j}^-)}} \quad (10.85)$$

在推理阶段，DPR 模型预先将所有文档进行编码并储存，利用 FAISS 向量比较工具对相应的问题进行检索，最终找到与问题最相关的前 K 个文档。之后的答案抽取部分与上文介绍的 DrQA 模型类似，这里就不再赘述。

### 10.5.2 端到端架构的开放问答模型

随着预训练语言模型的规模越来越大，训练数据越来越多，一些研究人员认为在这些在大规模数据上训练的超大规模语言模型很可能已经建模了网页数据中的知识。基于端到端架构的开放问答模型引起了广泛兴趣。PLSLM<sup>[571]</sup> 采用小样本提示学习的方式，利用网络数据对超大规模语言模型进行增强，提示学习的方法不需要重新进行预训练或者改变模型结构，可以轻量级的实现模型信息更新。基于小样本提示学习增强大规模语言模型的方法分为三步：

- (1) 根据用户提出的问题  $q$ ，采用搜索引擎检索一系列相关文档作为增强信息。
- (2) 利用检索出的文档，利用提示学习对模型进行调节。
- (3) 模型通过每一个检索出的文档生成候选答案，并对这些候选答案进行重排序并选出最合适

作为最终答案返回。

整体的模型架构如图10.18所示。

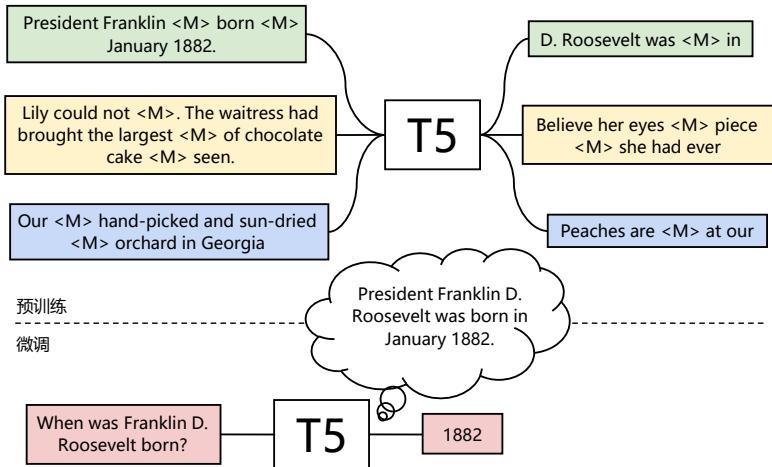


图 10.18 端到端架构开放问答模型（以 T5 为例）

对于给定的问题  $q$ , PLSLM 算法利用搜索引擎检索得到前 20 个最相关文本作为模型的增强文档。由于网络上获取的文档会伴随噪音等问题，会增强后续提示学习的难度，所以 PLSLM 对检索文档进行了预处理。将所有检索到的文档分为多个段落，其中每个段落由 6 个句子表示，最后再利用 TF-IDF 算法对所有段落和  $q$  进行编码，利用余弦相似度进行重排序，将这些排序后的段落作为增强数据。得到经过排序的段落  $P$  之后，PLSLM 利用小样本提示学习对大规模预训练模型进行调节，PLSLM 针对开放领域问题回答设计了相应的提示模板：

证据: ...

问题: ...

答案: ...

例如：对于训练数据中的问题“复旦大学有几个校区？”，对应的增强文档是复旦大学的百度百科，需要构造如下输入：

证据：复旦大学，简称“复旦”，位于直辖市上海，是中华人民共和国教育部直属的全国重点大学，中央直管高校，综合性研究型大学，由教育部与上海市共建。迄 2020 年 4 月，学校有邯郸校区、枫林校区、江湾校区、张江校区四个校区，占地面积约 243.92 万平方米。

问题：复旦大学有几个校区？

答案：

提示学习会通过提示模板引导模型根据问题和证据去生成合适的答案。在对模型进行增强之后，PLSLM 还会利用打分函数对所生成的候选答案在此排序，挑选其中最合适作为最终答案返

回给用户。针对问题  $q$  的候选答案  $a_i$  以及  $n$  个检索片段  $p_i$  可以采用如下三种方法选择最合适的答案：

(1) 直接推理：最大化  $P(a|q)$

$$P(a|q) = \sum_{i=1}^n P_{\text{TF-IDF}}(p_i|q) \cdot P(a_i|q, p_i) \quad (10.86)$$

(2) 噪声信道模型：最大化  $P(a_i, q|p_i)$

$$P(a_i, q|p_i) = \frac{P(q|a_i, p_i) \cdot P(a_i|p_i)}{P(q|p_i)} \quad (10.87)$$

(3) 专家乘积 (Product of experts, PoE)：使用额外的语料训练权重综合上述概率

由于提示学习不需要重新进行预训练以及不改变模型架构的特质，这类方法可以适用于所有语言模型。例如，T5<sup>[572]</sup> 模型是一个文本-文本的大规模预训练生成模型，提出了一个统一的模型框架，将所有自然语言处理的任务都视为文本到文本的生成任务，即输入是文本，输出也是文本的任务。T5 模型将翻译、分类、回归、摘要生成等任务都统一成了文本到文本任务，使得所有任务在预训练和微调的过程中都能使用同样的目标函数进行训练，使用同样的解码过程进行推理。

T5 模型采用的是编码器-解码器结构，在预训练过程中，采用了噪音扰动的方式，BART 预训练模型同样采用了这种预训练方式，对于这种预训练方式在摘要生成章节进行了详细的描述，这里就不再赘述。对于训练中的输入，用一些特殊符号例如  $< X >$  来表示原始样本中被随机遮盖的片段或词缀，希望模型生成的目标样本则是没有被遮盖过的原始样本，通过这种方式对 T5 模型进行预训练。

具体到开放问答任务，对每一个输入问题  $Q$ ，可以通过提示学习的方式对 T5 模型进行微调，从而使 T5 模型可以直接生成问题对应的答案  $A$ ：

$$A = \text{T5}(Q) \quad (10.88)$$

通过提示学习利用大规模语言模型直接进行开放问答的方式成为近年来开放问答任务上的一个新范式。随着规模更大，预训练更充分的生成式模型（如 GPT-3）不断涌现，这种端到端架构的开放问答模型也逐渐成为研究热点。

### 10.5.3 开放领域问答语料库

目前常用的开放问答语料库如表格 10.5 所示。开放问答评测集合的语料规模通常较大，SQuAD<sub>open</sub>、Natural Questions 以及 XQA 利用的是维基百科作为基础资源库，SearchQA 采用搜索引擎的返回文档结果作为支持文本。

表 10.5 开放领域问答常用语料库

数据集	知识来源	数据集规模
SQuAD <sub>open</sub>	维基百科	97K
Natural Questions	维基百科	323K
XQA	维基百科	90K
SearchQA	搜索引擎	140K

### 1. SQuAD<sub>open</sub>

由斯坦福大学构建的 SQuAD 数据集极大的阅读理解的发展，在很多应用中也采用其训练和验证开放问答。SQuAD<sub>open</sub><sup>[569]</sup> 扩展自 SQuAD 数据集，在应用于开放问答时，使用整个维基百科数据集合作为问答资源，不再提供问题所对应的文章段落。由于 SQuAD 评测不公开测试集合，因此在应用于开放问答场景评测时，使用开发集用于测试。与 SQuAD<sub>open</sub> 扩展自 SQuAD 类似，包括 MS-MARCO、HotpotQA 等在内的用于阅读理解评测的语料集合常通过不提供答案所在文章，从而应用于开放领域问答评测。

### 2. Natural Questions

Natural Questions<sup>[573]</sup> 数据集由谷歌公司发布，旨在解决目前开放问答方法的局限性。Natural Questions 提供 307,373 条训练数据，7830 条验证集数据以及 7842 条测试数据。对于每一条数据，包含一个问题和一个简短答案以及包含答案的维基百科界面。Natural Questions 数据集是开放问答领域重要的数据集。

### 3. XQA

XQA<sup>[574]</sup> 是清华大学构建的跨语言开放领域问答评测集合，提供了包含 56279 英语问题答案对的训练集合，测试集合包含中文、英文、法语、德语等在内的 9 种语言开放领域问题和答案，开发集合规模 17358 问答对，测试集合规模为 16973 个问答对。

### 4. SearchQA

为了模拟人们在回答问题时的一般流程，SearchQA<sup>[575]</sup> 采用了与之前问答评测集合构建不同的方法，首先收集问题答案对，在此基础利用 J! Archive 以及谷歌搜索引擎返回片段构建支持文本。针对收集的 14 万问题答案对，每个收集了约 49.6 个支持文本片段。

## 10.6 延伸阅读

从基于规则和基于统计的模型到如今基于神经网络的模型，智能问答任务经历多年的发展取得了很大的进步，随着预训练语言模型和深度学习的兴起，在很多智能问答问题上都已经能实现很好的效果。尽管如此，智能问答领域中还有很多问题没有解决，包括多跳问答、智能问答

系统解析生成、多模态智能问答等。

在多跳问答方面，虽然现在的深度学习模型在传统的抽取式阅读理解数据集上取得了很好的表现，但现有模型在多跳问答数据集上还有所欠缺。多跳问答即数据集中的问题，需要多次“跳转”的阅读理解才能回答。具体来说，给定一个问题，系统只通过一个文档是无法正确回答问题的，需要系统根据多篇文档回答一个问题，所以需要多跳推理。现有的工作基于图神经网络<sup>[576, 577]</sup>，通过对抗学习<sup>[578]</sup>，证据检索<sup>[579]</sup>等方式试图解决多跳问答任务。

智能问答系统解析生成方面，虽然面对大部分数据集，基于预训练的智能问答模型都可以给出正确的答案，但是这些答案都不可解释，不能给出一个思维链去解释得到答案的过程，而这种可解释性的缺失也制约了智能问答系统的进一步发展。研究人员通过构建模型选择支持事实<sup>[549]</sup>，生成蕴含树<sup>[580–582]</sup>等方式进行了一些研究。如何利用潜在的常识，充分挖掘文档信息，生成推导逻辑链，是智能问答乃至人工智能的重要发展方向之一。

在多模态智能问答方面，目前的智能问答主要针对的是纯文本数据，但真实场景中的文档具有多样性的布局和丰富的信息。此外，很多网页以及社交媒体包含丰富的多模态信息。针对多模态智能问答，研究人员提出了多种方案试图对不同形式的信息进行融合，包括图融合，图卷机，结构化等方面进行了一系列研究<sup>[583–586]</sup>。如何充分利用图片、视频、文本中包含的丰富信息，是智能问答的重要发展方向之一。

## 10.7 习题

- (1) 阅读理解模型有哪些信息提取方式？你认为基于深度学习的阅读理解模型包含哪些缺陷？
- (2) 基于 SQuAD 数据集，复现 R-Net 模型，正确率差距不能超过 5%。
- (3) 试比较 R-Net 模型和 BiDAF 模型的异同点。
- (4) 你认为 SpanBERT 模型解决了什么样的问题，除了本章节所介绍的阅读理解任务，你认为 SpanBERT 还适用于哪些自然语言处理任务，并说出原因。
- (5) 你认为端到端架构的开放问答模型相比传统的基于检索的开放问答模型有哪些可能的优势，为什么当前端到端架构的模型效果不如基于检索的模型。
- (6) 你认为 ChatGPT 模型可以看作是一种开放问答模型吗？它与传统的开放问答模型相比有哪些优点及其原因？

# 11. 文本摘要

---

文本摘要(Text Summarization)是一种利用算法自动实现文本分析、内容提炼并生成摘要的技术。由于互联网的快速发展，当前时代的信息量出现了指数级的增长，并远超于人类的对信息地处理和利用能力，从而导致我们无法快速地挑选和有效地运用信息，这种现象称为信息过载(Information Overloading)。而文本摘要技术可以对信息进行简化，帮助人们快速获取和理解新的信息。自20世纪50年代以来，自动文本摘要技术经过了长远的发展，从早期的基于规则和统计的方法，到如今以数据为驱动的机器学习和深度学习方法，研究人员致力于构造能够输出更加接近人工效果的自动摘要算法。文本摘要应用的场景十分广泛，涉及语义理解和语言生成，是自然语言处理领域中一项重要并且具有挑战性的任务。

本章首先介绍文本摘要的基本概念和主要任务，在此基础上介绍抽取式文本摘要方法和生成式文本摘要方法，最后介绍文本摘要的评测方法和文本摘要常见的语料库和评测集合。

## 11.1 文本摘要概述

从狭义角度来看，文本摘要的目标是为文档提炼关键信息，并构造出代表原文档中最重要或相关内容的子集（摘要）。从广义来看，除了文本形式的信息，我们还可以为图像或视频等多模态的信息进行总结，并产生以自然语言形式描述的摘要。文本摘要的第一份相关文献可以追溯到1957年Hans Peter Luhn的基于统计方法的研究<sup>[587]</sup>。随着互联网的高速发展，大量信息以复杂多样的形式存在于人们的日常生活中。文本摘要作为信息甄别和过滤的核心技术之一，受到学术界和企业界越来越多的关注。以机器学习方法为代表的抽取式文本摘要技术的研究成果不断涌现，应用落地场景日趋广泛。随着深度学习技术的不断发展，特别是基于大规模数据的预训练大模型范式的成功，文本摘要领域的研究变得更为活跃，并逐渐转向生成式摘要和实时摘要。此外，随着互联网中图像、短视频等图像视频内容的不断增加，一些基于多模态信息的摘要研究也成为近年来研究热点，例如图像摘要和视频流摘要等。

在本节中，我们将首先介绍文本摘要的发展历程，再分别介绍文本摘要领域的各种任务形式。

### 11.1.1 文本摘要发展历程

文本摘要相关的研究最早可以追溯到 20 世纪 50 年代，随着计算机技术的普及和应用，自动信息提取和摘要逐渐受到人们的关注。1958 年，Hans Peter Luhn 发表了一篇题为《The Automatic Creation of Literature Abstracts》的文章<sup>[588]</sup>，揭开了通过计算机来实现自动文本摘要的序幕。之后的几十年时间里，文本摘要主要依赖于信息检索中常见的特征，例如词频（Term Frequency）以及词频-逆文档频率（TF-IDF）等，通过统计信息来进行关键信息的提取。在 1990 年之前的方法中，文本摘要大都是通过规则和统计信息从原始文本中直接抽取（复制）内容，而不是生成新的内容和抽象的摘要。

传统的基于频率和规则的方法使得建立高效和稳健的摘要系统面临着很大的挑战。随着机器学习技术的快速发展，有相当多有监督和无监督的自动摘要方法被提出来。2002 年文献 [589] 提出将自动文本摘要视为一个二分类问题，如果一个句子出现在抽取式的参考摘要中，则认为它是“正确的”，否则就被认为是“不正确的”。在文献 [589] 中，他们使用了两种著名的机器学习分类方法，朴素贝叶斯和 C4.5 算法。对于无监督方法，可以通过定义相关的特征进行关键信息的筛选，从而不需要依赖人工构造的摘要。常见的特征选择包括句子的位置、正负相关词、句子中心度、与标题的相似性等等。对于如何训练模型来选取和组合这些特征，常见的算法包括遗传算法（Genetic Algorithm, GA）和潜在语义分析（Latent Semantic Analysis, LSA）等。除此之外，还有针对句子特征进行聚类和排序的无监督文本摘要方法，均展现出了良好的效果。

自 2014 年开始，以深度学习和神经网络为代表的文本摘要技术逐渐兴起，它们与传统技术相比具有更加卓越的性能。首先就是词嵌入表示逐渐取代了原来的特征工程。词嵌入通过神经网络技术，能产生比经典的词袋方法更加高质量的词表示。2019 年 Alami 等人构建了一个基于词嵌入的文本摘要系统<sup>[590]</sup>。其次，使用深度自编码器（Auto-Encoder, AE）从词频输入计算隐特征也可以为句子或单词产生表示。以深度神经网络为支撑的表示学习，为单词、句子、文档提供了高质量的语义表示，并大大促进了文本摘要的发展。

DUC-2003 和 DUC-2004 竞赛为文本摘要任务提供了标准，由此生成式文本摘要技术迎来了快速发展。原文本和摘要都是单词序列，因此可以用序列到序列模型（Seq2Seq）处理输入的文本序列和输出的摘要序列。机器翻译问题与文本摘要具有明显的相似性，它们的区别在于机器翻译是信息无损的，而文本摘要是有损的。因此，一些早期的生成式文本摘要方法直接使用了机器翻译模型。早期的 Seq2Seq 模型主要是基于循环神经网络（RNN）结构。2017 年，以 Transformer<sup>[591]</sup> 为代表的自注意力模型凭借强大的文本建模能力和并行效率，逐渐成为了生成式文本摘要的主流。与此同时，抽取式文本摘要的建模方法也逐渐转为了以预训练语言模型为基本框架的模式。

随着移动互联网和 Web2.0 时代的到来，网络上的信息量增长飞速，数据量越来越大，数据模态越来越多，文本摘要的研究也进入了新时期。在这一阶段，依托大规模预训练模型的文本摘要展现出了卓越的性能，多模态多语言摘要任务逐渐受到大家的关注，文本摘要的应用场景也越来越广泛。同时，文本摘要的效率和鲁棒性、评估方法的多样性和合理性、场景的可迁移性等，逐

渐成为了研究者进一步探索的方向。

### 11.1.2 文本摘要主要任务

文本摘要广泛应用于多种任务和场景中，按照任务进行划分，可以将文本摘要分为单文档摘要、多文档摘要、对话摘要、跨语言文本摘要和多模态文本摘要等。本节将介绍各种任务的定义。

#### 1. 单文档摘要

单文档摘要是最基本的摘要任务，给定一篇文档作为输入，模型输出它的核心内容，且长度明显短于输入文档。根据输入文档的内容长度，可以将单文档摘要分为两类：短文本摘要和长文本摘要。在短文本摘要场景中，文本类型可以是新闻文本、评论文本、邮件文本等短文本。长文本摘要则是面向学术论文、专利文档、书籍等长文本的摘要。与短文本摘要相比，长文本的信息量更多，但文档结构更加规范化。我们可以根据文档类型的不同，制定针对性的摘要策略。例如，新闻摘要关注新闻的人物、时间、地点、事件等信息；评论摘要关注产品的特性、用户的情感倾向；学术论文摘要则要更关注问题的定义、方法的描述、实验的结果等等。

#### 2. 多文档摘要

多文档摘要目标是对同一主题下的多个文档提炼主要信息并输出摘要。从应用的角度来看，对于某一新闻事件，互联网上的某个新闻单位会针对此事件进行系列报道，或者多个新闻单位对同一事件进行同时报道。在这种情况下，对这些相关性很强的多文档提炼出一个覆盖性强、形式简短的摘要就具有重要的意义。另一种应用场景是商品评论的摘要。对于某一种商品来说，商家分析其售后评论并总结出用户关注的方面和特性，有助于商家针对性地优化产品。对于多文档的学术论文摘要，我们通过对多篇文章的研究内容进行归纳总结，得到相关研究主题的主要挑战、主流方法和未来趋势等等。

#### 3. 对话摘要

对话摘要的目标是为多种类型的对话生成摘要，包括会议、通话记录、在线聊天和客服对话等场景。从参与者的角度来看，对话与新闻文档最大的不同点在于前者的信息涉及两个以上的参与者。除此之外，对话摘要会涉及更多的对话语义消歧、指代消解、信息补全等问题和挑战。相比新闻摘要和论文摘要，对话的摘要很难直接从现有的内容中获取，而新闻的标题和论文的摘要可直接作为参考结果并进行大规模的摘要数据收集，因此构建对话摘要的数据集更加困难。除此之外，对话会涉及音频、视频等信息，因此对话摘要还会涉及多模态信息的建模和理解。

#### 4. 跨语言文本摘要

跨语言文本摘要是指输入源语言（如英文）文档，输出目标语言（如中文）摘要的任务。它的主要应用场景在于为非母语者生成摘要。一个常见的跨语言摘要数据集是 Global Voices<sup>[592]</sup>，它是多语言新闻的英文摘要，主要包含 15 种语言的新闻及其摘要。一般而言，跨语言文本摘要可拆分为文本摘要和机器翻译两个子任务进行处理。

### 5. 多模态摘要

多模态摘要目标也是输出摘要内容，但是模型的输入和输出不仅局限于文本，还包括图像、视频、语音等多媒体数据。例如，新闻文档中包含的图像、会议记录中的语音、电影字幕对应的视频等等。在多模态摘要中，各个模态的信息可能有重合或互相补充，如何对这些信息进行对齐、识别、筛选，做到摘要内容的不重复和不遗漏，是进行多模态文本摘要的重点。MSMO 数据集<sup>[593]</sup>是一个常见的多模态新闻摘要数据集，其中所包含的带有图像的新闻文档，是内容对齐的，即图像和文本描述了同样的内容。但是在一些复杂的场景下，例如教学和演示视频、电影内容等，文字部分无法全面描述视频或图像展示的信息。这种情况下，需要进行图像或视频的理解，从而给多模态摘要带来了更大的挑战。

## 11.2 抽取式文本摘要

抽取式文本摘要中的关键内容是从原始文本中提取的，同时所提取的内容通常不会以任何方式进行修改。关键内容包括可用于“标记”或索引文本文档的短语，或共同组成摘要的关键句（包括标题）。关键文本的抽取类似于略读的过程，人们在决定是否要详细阅读一篇文档之前，会倾向于选择阅读其中的摘要（如果有）、标题、副标题、数字、文档的第一段和最后一段，以及段落中的第一句和最后一句。常见的抽取式文本摘要方法可分为两大类：基于排序的方法和基于序列标注的方法。

### 11.2.1 基于排序的方法

基于排序的抽取式文本摘要的基本思想是对于一篇文档，首先按照一定的规则将其拆分为多个语义单元（如短语、句子、段落等），并通过某种重要性评估方法为每个语义单元进行评分。之后，按照重要性分数的大小将语义单元进行排序，最后选取分值较高的语义单元作为关键内容并组成摘要。一般而言，会选取句子作为内容提取的基本语义单元，而如何衡量句子的重要性以及如何设计合适的评分方法，则是影响排序效果好坏的关键。

#### 1. Lead-3 算法

对于大部分类型的文档，句子在文档中的位置是一个重要的判断依据。一般而言，文档的起始句段会进行总述，结尾句段会进行总结。因此，最简单的启发式排序方法就是按照句子的位置选择关键句作为摘要。Lead-3 算法会选取文档的前三句内容作为摘要。这种方法虽然简单直接，但在某些场景中（如新闻文本摘要）非常有效。因此，Lead-3 常常作为一个基线方法，并将其视为文本摘要方法的性能下界。

#### 2. TextRank 算法

TextRank<sup>[594]</sup>是一个经典的基于排序的抽取式摘要方法，它的核心思想来自 PageRank<sup>[595]</sup> 算法。PageRank 是一个基于图的排序算法，用于将网页进行排序，目的是尽可能地将重要的网页排

在靠前的位置。这一类基于图的排序算法，可以递归地从图中提取全局信息，并根据节点间的某种关系来衡量节点的重要性，为节点评分排序。PageRank 算法将网页作为有向图的节点，将网页之间可跳转的超链接作为图的有向边。当页面 A 能跳转到页面 B 时，我们可以将其视为页面 A 为页面 B“投票”。一个页面获得的票数越多，该页面的重要性就越高。此外，PageRank 模型会根据投票页面的重要性，决定其投票的权重。因此，一个页面的重要性得分取决于获得的票数，以及为其投票的节点分数。PageRank 模型节点重要性迭代计算公式如下：

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (11.1)$$

其中  $V$  代表页面节点， $S(V)$  代表节点的重要性分数， $In(V)$  为可跳转到页面  $V$  的页面集合， $Out(V)$  为页面  $V$  可跳转到的页面集合， $d$  为阻尼系数。上述公式为迭代式，节点的分数会经过多轮迭代进行更新直到收敛。一般情况下，会设一个初始值（如  $\frac{1}{N}$ ， $N$  为页面数量）对节点分数进行初始化。

受到 PageRank 算法的启发，在进行文本摘要任务时，TextRank 算法将文档中的句子作为节点，以句子间的相似度作为边，通过迭代更新节点的重要性分数。句子之间相似度表示计算公式如下：

$$\text{Sim}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i, w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad (11.2)$$

其中， $S$  和  $w$  分别代表句子和句子中的单词， $|S|$  代表句子  $S$  中的单词数量，分子部分表示同属于两句的单词数量。相应的，对于句子节点的评分，采用如下公式：

$$S(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{\text{Sim}(V_j, V_i)}{\sum_{V_k \in Out(V_j)} \text{Sim}(V_j, V_k)} S(V_j) \quad (11.3)$$

使用边的权值，迭代更新节点的重要性分数直到收敛，最后选取  $N$  个得分最高的节点句子，作为摘要进行输出。

### 3. MMR 算法

Maximal Marginal Relevance (MMR)<sup>[596]</sup> 也是一种基于排序的抽取式文本摘要算法。与 TextRank 算法相比，MMR 算法可以得到更加多样化的抽取结果，在保证重要的同时避免抽取过多相似的句子。例如，以下是一系列描述上海市的短句：“十分现代化”、“有浓厚的现代化气息”、“现代化中国的代表城市”、“广阔的发展前景”。前三个短句虽然与上海市的相关性很大，但是都表示上海的现代化程度。如果将其作为摘要，就会出现冗余信息。MMR 算法旨在解决这样的问题，在重要性和多样性之间达到平衡。

MMR 算法最初是用来计算搜索引擎中的查询文本与被搜索文档之间的相似度，然后对文档进行排序并召回。具体到文本摘要任务，可以通过计算句子之间的相似度以及句子与整篇文档的

相似度，并通过如下公式计算每个句子的分数：

$$MMR \stackrel{\text{def}}{=} \arg \max_{V_i \in R \setminus S} [\lambda(\text{Sim}_1(V_i, Q) - (1 - \lambda) \max_{V_j \in S} \text{Sim}_2(V_i, V_j))] \quad (11.4)$$

其中， $Q$  代表整篇文档（或文档类别描述）。 $R$  为文档中的句子集合。 $S$  为已被选出作为结果的句子集合，是  $R$  的子集。 $R \setminus S$  代表  $R$  中剩余未被选中的句子集合。 $\text{Sim}_1(V, Q)$  表示文档中某个句子和整篇文档（或文档描述）的相似度， $\text{Sim}_2(V_i, V_j)$  表示两个句子之间的相似度。 $\lambda$  为范围  $[0, 1]$  内的常数，用于调整结果的多样性， $\lambda$  越小表示算法更倾向于保证抽取结果的多样性。

通过 MMR 算法，可以按照重要性和多样性对句子进行排序，两者的衡量方式都是基于相似性度量。通常认为与文档整体更相似的句子重要性越高，同时与其他句子相似性较低的句子多样性越好。可用的相似性度量方式也有很多，常见的包括 TF-IDF、余弦相似度、欧式距离，或者也可以使用神经网络模型进行评判评分。

#### 4. 基于朴素贝叶斯方法的文本摘要算法

除了人工设定规则的方法外，也可以使用统计机器学习方法结合特征工程来进行句子抽取。一个常见的方法是使用朴素贝叶斯模型，根据人工构建特征，以及预先标注的训练语料完成句子抽取。常见的句子级别用于抽取式摘要的文本特征包括：

- **长度**：句子的长度一定程度上可以反映信息量的大小。当句子超过或短于一定长度时，通常是不适合作为摘要的。
- **位置**：基本上所有类型的文本都有其约定俗成的写作规范。句子的内容和其位置在一定程度上是有关联的。例如，文章结尾通常包含整篇文章的总结综述。
- **关键词**：当某个句子包含该类型文本的关键词，那么这个句子很有可能是摘要的一部分。
- **提示语**：在某些类型的文档中，关键的概括和总述性语句前会有明显的提示语，例如，“综上所述”，“To conclude”一类的短语之后会紧跟重要的句子。

根据预先进行人工标注，标明句子是否为摘要句的训练集合上，根据设计好的特征值集合和标签，就可以训练朴素贝叶斯分类器，并用它来计算每个句子属于摘要的概率：

$$P(\text{label} | f_1, f_2, f_3, \dots, f_n) \quad (11.5)$$

通常需要人为设定一个阈值，当概率大于该阈值时表示当前句子是摘要的一部分，反之则不是。

#### 5. REFRESH 算法

上述讨论的方法均是为单个句子进行评分并选出分数最高的若干句子，除此之外，还可以为选出的句子集合进行整体评分。REFRESH<sup>[597]</sup> 是一种通过强化学习对抽取的句子集合进行全局优化的方法。在训练阶段，每当模型抽取出若干句子，可以将此句子集合与标准摘要进行比较，计算出两者的相关性分数（一般使用 ROUGE 分数作为度量手段，参考本章第 11.4 节）作为奖励分

数，通过强化学习训练策略来对模型参数进行优化更新。与一般的极大似然估计法相比，强化学习可以直接使用与摘要任务相关的评估指标来进行模型的优化，从而使得对句子集合的整体评分与最终的评估指标接近。REFRESH 模型主要由三部分组成：句子编码器、文档编码器和句子抽取器。其模型结构如图11.1所示。

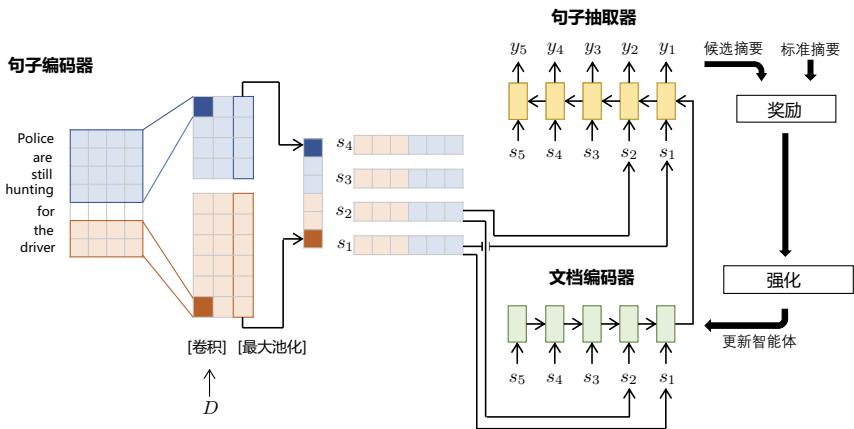


图 11.1 REFRESH 模型结构<sup>[597]</sup>

在 REFRESH 模型中，句子编码器使用 CNN 中经典的时域卷积网络（Temporal Convolutional Network）<sup>[501]</sup> 将句子编码成连续型向量表示， $f \in R^{k-h+1}$  为其得到的时域长度，其中  $h$  为卷积窗口大小， $k$  为句子长度，并采用最大池化将最大值作为最后的特征保留。之后，句子向量经过文档编码器来融合上下文中其他句子的信息。这里，文档编码采用的是 LSTM 长短记忆循环神经网络<sup>[53]</sup>。

对于核心的句子抽取器，REFRESH 使用了 LSTM 模型外加 Softmax 函数来计算似然概率  $P(y_i|s_i, D, \theta)$ 。不过，REFRESH 采用的训练策略并非极大化似然函数，而是通过其概率分布  $P(\hat{y}|\hat{s}, D, \theta)$  为文档  $D$  中的每一个句子  $s_i$  进行评分。可以选出分数最高的  $n$  个句子组成摘要，并将此句子集合记为  $\hat{y}$ ，将  $\hat{y}$  与标准摘要的相关性分数作为奖励函数  $r(\hat{y})$ ，则强化学习的期望奖励和损失函数可用如下式子表示：

$$\mathcal{L}(\theta) = -E_{\hat{y} \sim p_\theta}[r(\hat{y})] \quad (11.6)$$

$$\nabla \mathcal{L}(\theta) = -E_{\hat{y} \sim p_\theta}[r(\hat{y}) \nabla \log(\hat{y}|D, \theta)] \quad (11.7)$$

上式计算的是与句子集合  $\hat{y}$  相关的期望项。然而，穷举所有可能的句子组合方式是不可行的。在实际的训练过程中，REFRESH 每次从  $p_\theta$  分布中采样单个样本  $\hat{y}$  进行训练，来近似整个训练批次

的奖励期望：

$$\begin{aligned}\nabla \mathcal{L}(\theta) &\approx -r(\hat{y}) \nabla \log(\hat{y}|D, \theta) \\ &\approx -r(\hat{y}) \sum_{i=1}^n \nabla \log(\hat{y}_i|s_i, D, \theta)\end{aligned}\quad (11.8)$$

## 6. NeuSum 算法

一般而言，基于排序的抽取式摘要方法都需要先对句子进行评分，之后再进行排序和句子选择。与上述方法不同的是，NeuSum<sup>[598]</sup> 采用了一种将句子评分和句子选择相结合的方式来进行文档摘要。它的训练目标即为学习一个评分函数  $g$ ，该函数可以计算某个句子被选为摘要句后，摘要整体的分数增益。在测试阶段，模型在每一个时间步  $t$  选出能使评分函数  $g$  结果最高的句子。评分函数  $g$  如下所示：

$$g(S_i) = r(\mathbf{S}_{t-1} \cup \{S_i\}) - r(\mathbf{S}_{t-1}) \quad (11.9)$$

上式中，NeuSum 采用了与 REFRESH 相同的评价函数  $r(\cdot)$  来计算句子集合整体与标准摘要之间的相关性分数。 $S_i$  表示文档中的某个句子， $\mathbf{S}_t$  表示在  $t$  时刻已经被选择的摘要句组成的集合。在每一个时间步  $t$ ，模型会选出能使摘要整体增益最大的句子（即函数  $g$  达到最大值的句子）直至达到摘要限制的长度。NeuSum 模型结构如图11.2所示，主要由两个模块组成：句子编码器和句子抽取器

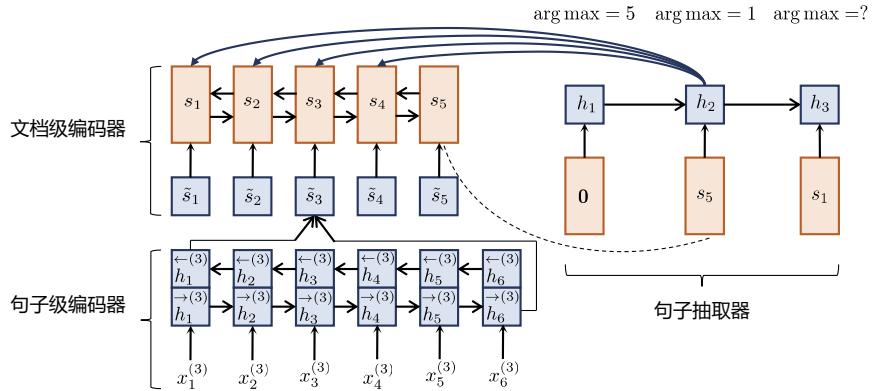


图 11.2 NeuSum 模型结构

NeuSum 模型中句子编码器是一个句子-文档级层次编码器。对于文档  $D = (S_1, S_2 \dots S_L)$  ( $S_i$  为文档中的句子)，模型先将句子  $S_i$  输入到一个双向的 GRU 网络<sup>[561]</sup> 中，得到句子级向量表示  $\tilde{s}_i$ ，再将得到的句子级向量输入到另一个双向 GRU 网络，可以得到融合上下文信息的句子向量  $s_i$ 。

对于句子抽取器，可以使用如下的公式计算句子  $S_i$  的分数，并取分数最高的句子作为  $t$  时刻

抽取的摘要句：

$$\delta(S_i) = \mathbf{W}_s \tanh(\mathbf{W}_q \mathbf{h}_t + \mathbf{W}_d \mathbf{s}_i) \quad (11.10)$$

其中， $\mathbf{W}_s$ ,  $\mathbf{W}_q$  和  $\mathbf{W}_d$  为可训练的参数， $\mathbf{h}_t$  为  $t$  时刻的隐层状态，由一个 GRU 门控循环神经网络<sup>[561]</sup> 计算得到。基于句子分数  $\delta(S_i), i \in [1, L]$  ( $L$  为文档中的句子数量)，可以计算模型的预测分布  $P$ ，如下所示：

$$P(\hat{S}_t = S_i) = \frac{\exp(\delta(S_i))}{\sum_{k=1}^L \exp(\delta(S_k))} \quad (11.11)$$

为了近似式子11.9中的评分函数，NeuSum 计算了用于训练的标签分布  $Q$ ，并使用 KL 散度(Kullback-Leibler Divergence) 来缩小模型预测分布  $P$  和训练标签分布  $Q$  之间的距离。

$$Q(S_i) = \frac{\exp(\tau g(S_i))}{\sum_{k=1}^L \exp(\tau g(S_k))} \quad (11.12)$$

$$J = D_{KL}(P||Q) \quad (11.13)$$

上式中， $\tau$  为调整分布方差的温度系数。对于  $r(\cdot)$  一般使用 ROUGE 分数来衡量摘要句集与标准摘要之间的相关度。

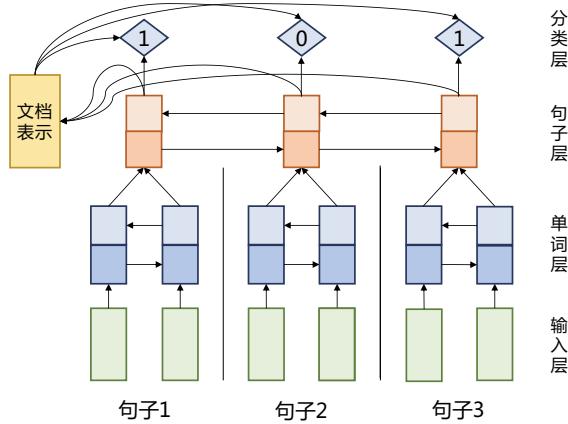
### 11.2.2 基于序列标注的方法

抽取式文本摘要的另一种方法是将句子抽取问题转化为序列标注任务。与基于排序的方法类似，首先要将文档分割成独立的语义单元，并将其组成语义单元序列。算法的目标就是对每个语义单元给出一个标签，来表明该部分是否要被提取出来组成摘要。一般来说，会选取句子作为基本语义单元，并使用二元标签（0 和 1）来表明句子被抽取的状态。标签“1”表示该句子可以需要被抽取出来作为摘要句，而“0”表示该句子不需要作为摘要句。通过这种思路，可以将抽取式文本摘要转化为序列标注任务。与排序式文本摘要的不同点在于，基于序列标注的方法直接对句子进行二元分类，而不需要预先设计并计算句子的重要性分数，所以可以直接使用有监督学习的方法进行训练。

#### 1. SummaRuNNer 算法

循环神经网络 (Recurrent Neural Network, RNN) 是处理序列标注任务的经典模型之一，被广泛应用于中文分词、命名实体识别等任务中。同样，我们也可以使用 RNN 进行基于序列标注的抽取式摘要。SummaRuNNer<sup>[599]</sup> 就是一种基于 RNN 的抽取式摘要模型，该模型是一个采用两层基于门控循环单元 (Gated Recurrent Unit, GRU) 的双向循环神经网络，其模型结构如图11.3所示，主要包含输入层、单词层、句子层以及分类层。

SummaRuNNer 算法逐级对文本进行处理，使用双向 GRU 进行建模。输入层将句子单词转换为词向量表示。单词层在词级别对文本进行处理，双向地依次根据每个单词的词嵌入以及历史隐

图 11.3 SummaRuNNer 模型结构<sup>[599]</sup>

状态来计算每个单词的隐状态。利用平均池化方法将句子中所有单词的隐状态融合为句子表示，并作为句子层的输入。句子层则是在句子级别对文本进行处理，同样使用双向 GRU 进行建模，然后利用所有句子的隐状态通过如下公式计算出文本的整体表示：

$$\mathbf{d} = \tanh(\mathbf{W}_d \frac{1}{N_d} \sum_{j=1}^{N_d} [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b}) \quad (11.14)$$

其中， $\mathbf{h}_j^f$  和  $\mathbf{h}_j^b$  分别是文本第  $j$  个句子前向 (Forward) 和后向 (Backward) 处理时的隐状态， $N_d$  是文本中的句子数量，“[]”表示将两个向量进行拼接。

SummaRuNNer 算法的最后一层是分类层，它从第一句开始，结合句子的隐状态以及文本的整体表示，依次判断各个句子是否需要提取出来作为摘要的一部分，最终实现摘要的抽取。具体的分类方法如下公式所示：

$$P(y_j = 1 | \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma(\mathbf{W}_c \mathbf{h}_j + \mathbf{h}_j^T \mathbf{W}_s \mathbf{d} - \mathbf{h}_j^T \mathbf{W}_r \tanh(\mathbf{s}_j) + \mathbf{W}_{ap} \mathbf{p}_j^a + \mathbf{W}_{rp} \mathbf{p}_j^r + \mathbf{b}) \quad (11.15)$$

其中， $\mathbf{h}_j$  是通过对  $[\mathbf{h}_j^f, \mathbf{h}_j^b]$  进行非线性变换得到的第  $j$  个句子的嵌入表示， $\mathbf{s}_j$  是摘要在处理第  $j$  句子时的动态表示，通过如下公式计算得到：

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1 | \mathbf{h}_i, \mathbf{s}_i, \mathbf{d}) \quad (11.16)$$

在公式 11.15 中， $\mathbf{W}_c \mathbf{h}_j$  反映句子的信息量， $\mathbf{h}_j^T \mathbf{W}_s \mathbf{d}$  反映句子对于文章的重要性， $\mathbf{h}_j^T \mathbf{W}_r \tanh(\mathbf{s}_j)$  反映句子对于已提取部分的冗余程度， $\mathbf{W}_{ap} \mathbf{p}_j^a$  和  $\mathbf{W}_{rp} \mathbf{p}_j^r$  分别考虑了句子的在文本中的绝对位置

(实际的句子数) 和相对位置 (将文本分割成固定数量的段, 句子所属的段数就是相对位置),  $\mathbf{P}^a$  和  $\mathbf{P}^r$  也是需要学习得到的模型参数, 分别是句子绝对位置和相对位置的嵌入表示。值得一提的是, 这种方式一定程度上提高了模型决策的可解释性。可以通过公式中各个部分的权重占比, 对模型给出的决策进行分析。

根据上述公式得到模型的预测结果后, SummaRuNNer 算法采用如下损失函数, 对模型参数进行学习:

$$\begin{aligned}\mathcal{L}(W, b) = & - \sum_{d=1}^N \sum_{j=1}^{N_d} (y_j^d \log P(y_j^d = 1 | \mathbf{h}_j^d, \mathbf{s}_j^d, \mathbf{d}_d) \\ & + (1 - y_j^d) \log (1 - P(y_j^d = 1 | \mathbf{h}_j^d, \mathbf{s}_j^d, \mathbf{d}_d)))\end{aligned}\quad (11.17)$$

其中,  $N$  表示训练集合中的文档数量,  $N_d$  表示文档  $d$  中句子的数量。

## 2. BERTSum-Ext 算法

BERTSum-Ext<sup>[600]</sup> 是一个基于 BERT 编码器的抽取式摘要模型。由于原始的 BERT 是针对字词进行编码, 而基于序列标注的抽取式文本摘要是逐句操作和分类, 在做句子级别的序列标注时, 要对输入 BERT 的内容做一定调整, 需要进一步计算句子的向量表示。BERTSum-Ext 的模型结构如图11.4所示。BERTSum-Ext 与 BERT 一样, 同样也是使用了 [SEP] 标记来分隔句子, 而不同之处在于两者的 [CLS] 标记的含义。在原始 BERT 中, [CLS] 标记被添加在文本序列的开始, 该位置整合了输入序列中的全局信息。在 BERTSum-Ext 中, [CLS] 标记被添加在每一个句子的开始位置, 该句子的有效信息将会被整合到对应 [CLS] 标记位置的字段上。此外, BERTSum-Ext 会为位置是奇数和偶数的句子分别赋予不同的段嵌入 (Segment Embeddings), 即第 1, 2, 3, 4... 句将分别被赋予段嵌入  $E_A, E_B, E_A, E_B \dots$ 。

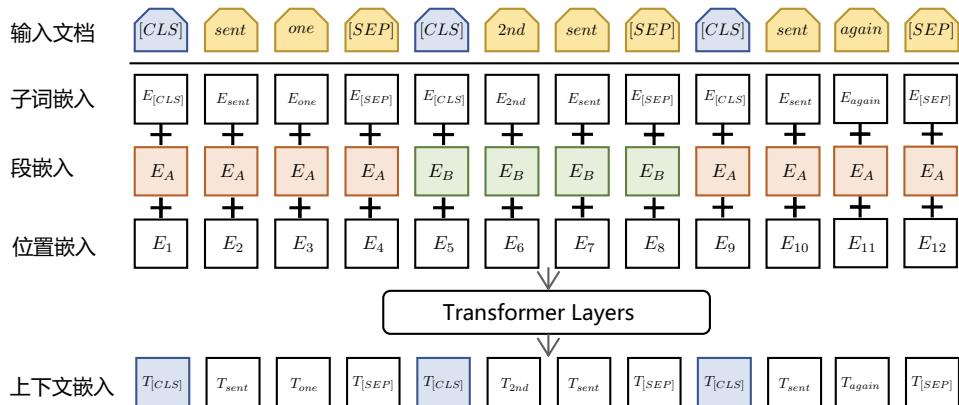


图 11.4 BERTSum 模型结构<sup>[600]</sup>

在这样调整后，通过结合图11.4的3种不同嵌入就可以得到不同位置的[CLS]向量 $t_i$ 来表示各个句子，然后将其输入到后续的Transformer模块中进行更新：

$$\tilde{\mathbf{h}}^l = \text{LN}(\mathbf{h}^{l-1} + \text{MHAtt}(\mathbf{h}^{l-1})) \quad (11.18)$$

$$\mathbf{h}^l = \text{LN}(\tilde{\mathbf{h}}^l + \text{FFN}(\tilde{\mathbf{h}}^l)) \quad (11.19)$$

其中，LN（Layer Normalization）表示归一化操作，MHAtt（Multi-head Attention）表示多头注意力操作，FFN（feed-forward neural network）表示前馈神经网络。这3个部分共同构成了一个Transformer模块， $l$ 表示层数，每一层都是一个Transformer模块， $\mathbf{h}^0 = [t_1, t_2, \dots, t_n]$ （假设文档一共有 $n$ 个句子）。最终的句子嵌入是逐层学习得到的，越高层的句子嵌入覆盖的信息量越多。使用顶层输出的句子嵌入作为最终输入分类器的表示。分类器可以为每个句子表示计算并预测二元分类结果，表示各个句子是否需要被提取出来作为摘要的一部分：

$$\hat{y}_i = \sigma(\mathbf{W}_o \mathbf{h}_i^L + \mathbf{b}_o) \quad (11.20)$$

其中， $\mathbf{W}_o$ 为可学习的参数。使用模型预测值与参考标签之间的交叉熵作为损失函数，就可以对模型进行训练。在测试阶段，选择预测概率高于一定阈值的句子，作为最终抽取得到的摘要。

### 3. 基于隐马尔可夫模型的抽取式摘要语料生成方法

基于序列标注的方法需要大规模的有标注训练语料，需要人工确定文章中每个句子是否应被抽取，因此文本摘要的人工标注的代价十分高昂。但是，我们可以从互联网上收集到很多，包含人工撰写的摘要以及相应文章的文本-摘要对，而且人们在构造摘要的过程中通常都会使用到文档中的原词甚至原句。因此，研究员们开始思考是否可以使用现有的文本-摘要对，来进行分类器的训练。一种思路就是为摘要中的句子在正文中找出与其相似度最高的句子，并用它来作为对应部分的摘要，即对齐（Alignment）算法。通过对齐算法，能够产生大规模的文档-摘要对，从而可以进行有监督的训练。简单的对齐算法有计算正文和摘要句子的最长公共子序列或编辑距离等，这些方法也可以结合使用。本节介绍一种基于隐马尔可夫模型（Hidden Markov Model, HMM）的对齐算法<sup>[601]</sup>。

该算法的思路是将摘要中的句子分解成字或词（后续简称为词），然后从句首开始往后依次找到这些词在正文中的出现位置，进而就可以找到相应的句子来构成近似甚至相同的摘要。由于同一个词可能在文章中多次出现，所以摘要中的单个词就存在多个可能的出现位置，需要去判断其中哪一个是正确的。例如，在一篇文章的摘要中出现了“the communication subcommittee of”，而“the”，“communication”，“subcommittee”，“of”分别在正文中出现了44, 1, 2, 22次，也就意味着有 $44 \times 1 \times 2 \times 22$ 共1936种可能的位置组合。

在该隐马尔可夫模型中，词每个可能的出现位置都是可以观测到的，并且这些位置都各自对

应着一个状态  $(S, W)$ , 它表示这个词出现在正文第  $S$  句话的第  $W$  个位置。假设词  $I_{i+1}$  和词  $I_i$  是相邻的两个单词, 可以使用  $P(I_{i+1} = (S_2, W_2) | I_i = (S_1, W_1))$  表示当词  $I_i$  出现在  $(S_1, W_1)$  位置时,  $I_{i+1}$  出现在  $(S_2, W_2)$  位置的条件概率。该算法使用一种启发式赋值方式, 按照当前处理词  $I_{i+1}$  和上一个处理的词  $I_i$  的相对位置分成 6 种不同的情况来梯度赋值:

- (1)  $S_1 = S_2, W_1 = W_2 - 1$ , 即  $I_{i+1}$  和  $I_i$  出现在正文的同一个句子中, 且  $I_i$  刚好在  $I_{i+1}$  的前一个位置。
- (2)  $S_1 = S_2, W_1 < W_2 - 1$ , 即  $I_{i+1}$  和  $I_i$  出现在正文的同一个句子中, 且  $I_i$  在  $I_{i+1}$  的前方, 但二者不相邻。
- (3)  $S_1 = S_2, W_1 > W_2$ , 即  $I_{i+1}$  和  $I_i$  出现在正文的同一个句子中, 但  $I_i$  在  $I_{i+1}$  的后方。
- (4)  $S_2 - \text{CONST} < S_1 < S_2$ , 即在正文中,  $I_{i+1}$  出现在  $I_i$  所属句子的后方, 且二者之间不会超过 CONST 个句子。
- (5)  $S_2 < S_1 < S_2 + \text{CONST}$ , 即在正文中,  $I_{i+1}$  出现在  $I_i$  所属句子的前方, 且二者之间不会超过 CONST 个句子。
- (6)  $S_2 - S_1 >= \text{CONST}$ , 即在正文中,  $I_{i+1}$  和  $I_i$  的位置间隔了 CONST 以上个句子。

其中, CONST 是一个较小的正数。上述六种情况, 其出现的可能性是递减的, 确切的概率值由人工设定, 可以通过多次实验来选择较优值, 在处理不同类型的文章时还可以做一些适应性调整。

对一个给定的序列, 为了找到最优的序列就需要使  $P(I_1, I_2, \dots, I_N)$  最大化。这里假设摘要中的每个词在正文中的实际位置都仅受其前一个词  $I_i$  在正文中的实际位置的影响。可以将上述联合概率转换为如下条件概率乘积:

$$P(I_1, I_2, \dots, I_n) = \prod_{i=0}^{N-1} P(I_{i+1} | I_i) \quad (11.21)$$

根据人工设定的条件概率, 就可以计算出每种位置序列的概率, 用概率最大的位置序列去找相应句子组合成摘要就可以取得不错的效果。需要注意的是有些摘要中的词不一定会出现在原文中(例如, 衔接词), 对这些词进行特殊化处理, 比如用  $S = -1$  来表示该词没有在正文中出现, 然后在计算时跳过该词。整个求解过程可以采用 Viterbi 算法完成。

## 11.3 生成式文本摘要

抽取式文本摘要所获得的结果都来源于原始文本, 因此所得到摘要文本中每个单句的语义准确性和语法正确性都可以得到很好的保证。但是由于原始的句子来源于不同的段落甚至篇章, 所生成的摘要中句子之间的连贯性则很难保证, 造成摘要整体的可读性较差。生成式文本摘要则可以产生原始文档中不存在的新文本。通常生成式文本摘要算法需要首先构建原始文档的抽象语义表示(编码过程), 然后使用此语义表示创建出接近人类表达方式的摘要(解码过程)。该类算法通常能产生更精简凝练并且更具有连贯性的摘要。语义表示过程也可应用于图像和视频, 因此生

成式摘要方法也可用来处理多模态文本摘要。但是，整体的摘要生成过程比抽取式摘要更具挑战性，它不仅涉及自然语言生成，还涉及依赖领域知识对原始文档的深入理解。

### 11.3.1 序列到序列生成文本摘要

在 2014 年，Google Brain 团队首次提出了序列到序列生成结构（Sequence-to-Sequence, Seq2Seq）<sup>[15]</sup> 来进行文本生成任务。序列到序列生成结构主要由编码器（Encoder）和解码器（Decoder）组成，其中编码器负责将输入文本中所需要的语义信息编码成向量，解码器则负责从编码向量中解码出语义信息并生成特定文本。对应到生成式文本摘要，可以将原始文档输入编码器，编码器进行文档的理解并将其中的关键信息进行编码，解码器则将关键信息进行解码并生成对应的文本摘要。

#### 1. 基于预训练的生成式方法

BERTSum-Abs<sup>[600]</sup> 是结合了预训练语言模型的 Seq2Seq 摘要生成方法。它的基本框架也是基于 Transformer 结构，其区别在于，编码器是预训练语言模型（如 BERT<sup>[29]</sup>），而解码器则采用参数随机初始化的多层 Transformer 神经网络结构。这样的设计虽然可以很好的利用了预训练语言模型，能够对文本内容进行很好的编码。但是会导致编码器和解码器之间存在着参数状态不匹配的问题，即编码器的模型参数是经过预先训练的，而解码器的模型参数是随机初始化的。由于在相同的优化策略下，解码器相比于编码器更难收敛，这可能会使训练阶段的微调过程不稳定。为了解决这一问题，BERTSum-Abs 设计了一个新的微调范式，用不同的优化策略对编码器和解码器进行微调。BERTSum-Abs 采用的方法是将解码器的学习率设置更大，使其参数更新的幅度更大，逐渐缩小两者不匹配的参数状态。

具体来说，BERTSum-Abs 对编码器和解码器分别使用了不同的 Adam 优化器参数<sup>[602]</sup>，它们具有不同的预热步数（Warm Up Step）和学习率：

$$lr_{\mathcal{E}} = \tilde{lr}_{\mathcal{E}} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup}_{\mathcal{E}}^{-1.5}) \quad (11.22)$$

$$lr_{\mathcal{D}} = \tilde{lr}_{\mathcal{D}} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup}_{\mathcal{D}}^{-1.5}) \quad (11.23)$$

其中  $lr_{\mathcal{E}}$  为编码器学习率， $\tilde{lr}_{\mathcal{E}}$  设置为  $2e^{-3}$ ， $\text{warmup}_{\mathcal{E}}$  设置为 20000， $lr_{\mathcal{D}}$  为解码器学习率， $\tilde{lr}_{\mathcal{D}}$  设置为 0.1， $\text{warmup}_{\mathcal{D}}$  设置为 10000。

通过上述参数控制，使得预训练的编码器以更小的学习率和更平滑的衰减速率进行微调。这样，当解码器参数变得稳定时，编码器就可以得到更精确的梯度进行参数更新。此外，BERTSum-Abs 还提出了一种两阶段的微调方法，首先在抽取式摘要（BERTSum-Ext）上微调编码器，第二阶段再进行生成式摘要的微调。这两个阶段使用的是同一个编码器，从而可以让模型利用这两种模式之间共享的编码信息。

上述方法中，BERTSum-Abs 使用的是预训练编码器，并结合参数随机初始化的解码器进行微调。这种解决方案主要是针对早期的预训练语言模型（BERT 等），它们侧重于文本的理解和编码。

随着预训练语言模型的不断发展，陆续出现了对编码器和解码器联合预训练的解决方案，其可以直接被应用在适合序列到序列生成架构的文本生成任务中，并在生成式文本摘要任务中展现了更好的性能，比如 BART<sup>[302]</sup> 等。由于 BART 预训练方法本身是一个去噪自编码器，与摘要任务的形式有一定的关联性，即去除文档中的冗余和无关的信息，只保留关键信息。因为这种上游预训练形式与下游任务相关的特性，BART 在生成式文本摘要中可以获得更好的效果。

BART 本身是一个去噪自编码器，与摘要任务的形式有一定的关联性，即去除文档中的冗余和无关的信息，只保留关键信息。因为这种上游预训练形式与下游任务相关的特性，BART 在生成式文本摘要中取得了优异的性能，并在多个摘要评测集上取得了很好的效果。

## 2. 基于复制与覆盖机制的方法

由于生成式文本摘要方法会对原始文档的内容进行转换和整合，所以如何保证摘要中的关键信息不重复不遗漏是非常重要的问题。一方面，原始文档可能会包含重要的生僻词甚至未登录词，而这些词在解码过程中被生成的概率极低，就会造成关键信息的遗漏问题。另一方面，一些在原始文档中反复出现的关键信息也有可能被解码器生成多次，造成摘要冗余和不通顺的问题。为了解决以上问题，文献 [603] 针对生成式文本摘要的复制与覆盖机制提出了 PGNet。PGNet 使用了指针解码器，通过指针从原始文本中直接复制单词，同时也可以从词表产生原文档中未出现的新单词。此外，PGNet 还使用了覆盖向量来跟踪和控制摘要对原文档的内容覆盖范围，从而能减少冗余现象。PGNET 的神经网络结构如图11.5所示。

PGNet 使用了 RNN 文档编码器，将原始文档的所有单词输入一个单层双向 LSTM 模型中，得到编码器隐状态  $\mathbf{h}_i$ 。解码器是一个单层单向 LSTM，在每个时间步  $t$  上，接收前一个单词的嵌入表示，并得到解码器状态  $\mathbf{s}_t$ 。之后，PGNet 使用注意力分布<sup>[604]</sup>  $\mathbf{a}^t$  来衡量第  $t$  个解码时间步上每个单词的重要性：

$$\mathbf{e}_i^t = \mathbf{v}^T \tanh(\mathbf{W}_h \mathbf{h}_i + \mathbf{W}_s \mathbf{s}_t + \mathbf{b}_{attn}) \quad (11.24)$$

$$\mathbf{a}^t = \text{softmax}(\mathbf{e}^t) \quad (11.25)$$

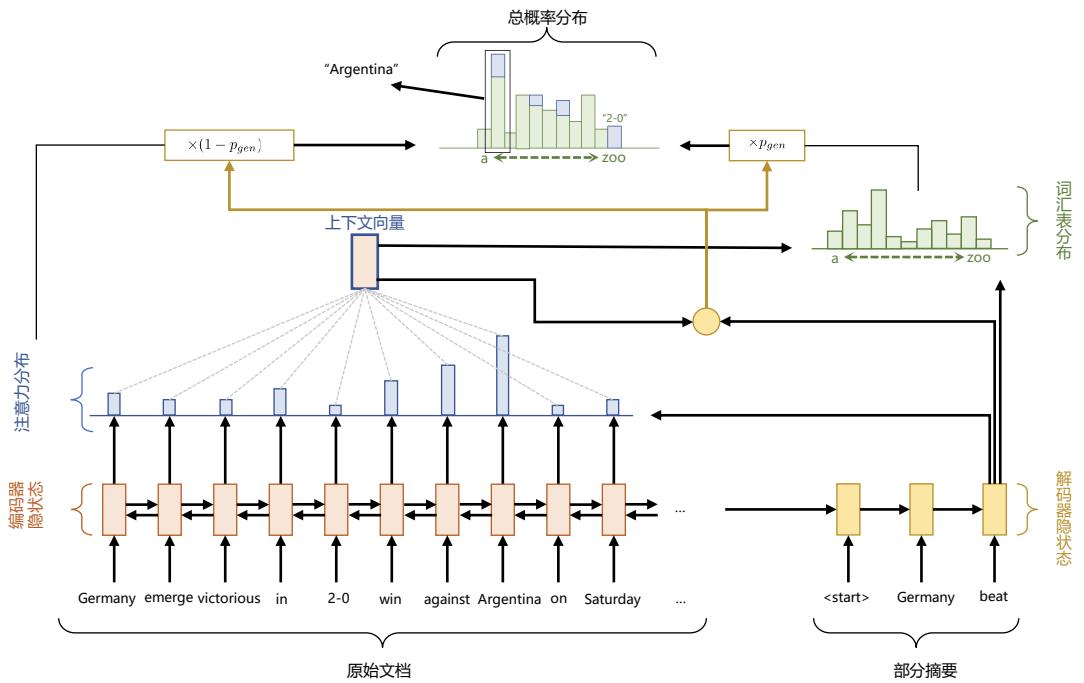
其中， $\mathbf{v}, \mathbf{W}_h, \mathbf{W}_s, \mathbf{b}_{attn}$  都是可训练的参数。

接下来，PGNet 利用注意力分布来产生编码器隐状态的加权和，称为上下文向量  $\mathbf{h}_t^*$ ：

$$\mathbf{h}_t^* = \sum_i \mathbf{a}_i^t \mathbf{h}_i \quad (11.26)$$

上下文向量  $\mathbf{h}_t^*$  可以看作是在每个时间步，解码器从原文档中读取的全局内容表示，它与解码器状态  $\mathbf{s}_t$  连接，并通过两个线性层来产生词汇表分布  $P_{\text{vocab}}$ ：

$$P_{\text{vocab}} = \text{Softmax}(\mathbf{V}' (\mathbf{V} [\mathbf{s}_t, \mathbf{h}_t^*] + \mathbf{b}) + \mathbf{b}') \quad (11.27)$$

图 11.5 指针生成网络结构<sup>[603]</sup>

其中  $V, V', b$  和  $b'$  是可学习的参数。 $P_{\text{vocab}}$  是词表中所有单词的生成概率分布。

除了从词表单词的概率分布中生成单词，PGNet 还允许通过指针从原文档中复制单词，从而有机会将文档中的生僻词和未登录词写入摘要。模型在  $t$  时刻需要确定解码器应当从词表中生成单词，还是从原文档中复制单词。这里，PGNet 设计了一个选择器  $p_{\text{gen}} \in (0, 1)$ ，以 0-1 之间的概率来判断此时间步的操作，应当是从  $P_{\text{vocab}}$  中采样生成一个单词，还是通过从注意力分布  $a^t$  中采样来复制原文档中的一个单词。时间步  $t$  的二元选择概率  $p_{\text{gen}}$  基于全局语义向量  $h_t^*$ 、解码器状态  $s_t$  和解码器输入  $x_t$ ，通过如下公式计算：

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}}) \quad (11.28)$$

最终，可以得到词表与原文档中单词的总概率分布，即结合了词表分布  $P_{\text{vocab}}$  和注意力分布  $a^t$  来采样单词。这也可看作是一种扩展的词表，包含了原词表和原文档中所有出现单词的联合词表。在扩展词表上的概率分布可计算如下：

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t \quad (11.29)$$

如果单词  $w$  是一个未登录词，那么  $P_{vocab}(w)$  为 0。类似地，如果  $w$  没有出现在原始文档中，那么  $\sum_{i:w_i=w} a_i^t$  为零。产生未登录词的能力是 PGNet 模型的主要优势之一。

除了解决生僻词和未登录词的遗漏问题，PGNet 还使用了一个简单的方案解决摘要中单词的重复生成问题。单词的重复生成问题是 Seq2Seq 模型的常见问题，在生成多句文本时尤其明显，一个简单的思路是记录所有已经生成的单词，并且避免在后续的解码时间步生成这些词。PGNet 中的覆盖机制也是基于这一思想，维护一个覆盖向量  $c_t$ ，它是解码器的所有历史时间步的注意力分布的和：

$$c^t = \sum_{t'=0}^{t-1} a^{t'} \quad (11.30)$$

直观来看， $c_t$  是基于原始文档单词的一种非归一化的分布，它代表了这些单词到  $t$  时间步为止被注意力分布覆盖的程度。 $c_t$  一般可以初始化为一个全 0 向量，因为在第一个时间步中，注意力分布还没有覆盖原始文档。

PGNet 将覆盖向量作为注意力机制的额外输入作为一种控制冗余的信号。因此，公式 11.24 可以更改为：

$$e_i^t = v^T \tanh (\mathbf{W}_h h_i + \mathbf{W}_s s_t + w_c c_i^t + b_{\text{attn}}) \quad (11.31)$$

通过上述公式，PGNet 确保了当前的注意力分布是通过之前时间步的注意力分布得到的，从而可以避免注意力分布在相同位置的单词上赋予过高的权重，避免产生重复的文本。此外，PGNet 还设计了额外的损失函数 covloss <sub>$t$</sub> ，以惩罚反复关注同一单词位置的情况：

$$\text{covloss}_t = \sum_i \min (a_i^t, c_i^t) \quad (11.32)$$

上述公式惩罚了  $t$  时间步前每个注意力分布和覆盖向量之间的重叠，从而防止了注意力分布的重複。最后，将上述损失函数通过超参数  $\lambda$  加权添加到主损失函数中，得到一个新的复合损失函数：

$$\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min (a_i^t, c_i^t) \quad (11.33)$$

### 3. 基于全局优化与强化学习方法

前面所介绍的两种生成式摘要方法优化目标的计算都是基于摘要序列的极大似然估计，即最小化所有时间步下目标单词的负对数似然函数。然而，这样的优化目标存在以下问题：(1) 摘要生成的曝光偏差问题 (Exposure Bias)<sup>[605]</sup>，模型在训练阶段计算的是基于标准摘要序列的单词后验分布，即  $p(y_t | y_1^*, \dots, y_{t-1}^*, x)$ ，这里  $y^*$  表示标准摘要。在预测阶段，模型无法获得标准摘要，因

此它只能基于已经预测的结果计算下一个时间步单词的后验概率，即  $p(y_t | y_1, \dots, y_{t-1}, x)$ ，在这种情况下容易导致错误的累积；(2) 摘要是一个较为主观的结果，标准摘要只是一个参考，除此之外还可能存在其它合适的摘要，这些摘要的词语和句子可能以不同的方式进行排列和组合。文本摘要的评测指标（如 ROUGE<sup>[606]</sup>）考虑了这种灵活性，但最大似然估计的优化目标无法做到这一点。

为了解决上述问题，文献 [607] 提出了引入关键的注意力机制并利用强化学习目标的 Deep-ReinSum 算法。模型使用了双向 LSTM 编码器  $RNN^{e_{fwd}}$ ,  $RNN^{e_{bwd}}$  从文档单词  $x_i$  的嵌入向量计算隐状态  $\mathbf{h}_i^e = [\mathbf{h}_i^{e_{fwd}}; \mathbf{h}_i^{e_{bwd}}]$  并建模输入序列。进行编码之后，再使用单向 LSTM 解码器  $RNN^d$ ，使用词嵌入向量为  $t$  时刻的解码输出  $y_t$  计算隐状态  $\mathbf{h}_t^d$ 。输入和输出的单词嵌入都取自同一个嵌入矩阵  $\mathbf{W}_{emb}$ 。解码器隐状态根据编码器最后时刻隐状态初始化  $\mathbf{h}_0^d = \mathbf{h}_n^e$ 。

在每个解码时间步  $t$ ，除了解码器此时的隐状态和当前已生成的单词之外，模型还使用了注意力机制来关注编码端得到的输入序列部分。将  $e_{ti}$  定义为隐状态  $\mathbf{h}_i^e$  在  $t$  时刻的注意力分数：

$$e_{ti} = f(\mathbf{h}_t^d, \mathbf{h}_i^e) \quad (11.34)$$

其中  $f$  可以是从  $\mathbf{h}_t^d$  和  $\mathbf{h}_i^e$  向量返回标量  $e_i^t$  的任何函数，这里选择使用双线性函数：

$$f(\mathbf{h}_t^d, \mathbf{h}_i^e) = \mathbf{h}_t^d^\top \mathbf{W}_{attn}^e \mathbf{h}_i^e \quad (11.35)$$

在得到注意力分数之后，该模型使用以下方式在时序维度上对注意力权重进行归一化，惩罚在过去解码步中已获得高注意力分数的单词。首先定义了一个新的基于时序特征的注意力分数  $e'_{ti}$ ：

$$e'_{ti} = \begin{cases} \exp(e_{ti}) & \text{if } t = 1 \\ \frac{\exp(e_{ti})}{\sum_{j=1}^{t-1} \exp(e_{tj})} & \text{otherwise.} \end{cases} \quad (11.36)$$

之后，计算归一化注意力分数  $\alpha_{ti}^e$  并使用这些权重来获得输入上下文向量  $\mathbf{c}_t^e$ ：

$$\alpha_{ti}^e = \frac{e'_{ti}}{\sum_{j=1}^n e'_{tj}} \quad (11.37)$$

$$\mathbf{c}_t^e = \sum_{i=1}^n \alpha_{ti}^e \mathbf{h}_i^e \quad (11.38)$$

上述基于时序特征的注意力机制一定程度上可以缓解生成重复短语的问题。然而，解码器仍然会根据其自身的隐状态产生错误累积，尤其在生成长序列时此现象更加严重。为此，模型额外引入了一种解码器内部的注意力机制（Intra-Decoder Attention）。具体而言，对于每个解码步  $t$ ，模型计算一个解码器上下文向量  $\mathbf{c}_t^d$ 。将  $\mathbf{c}_1^d$  设置为一个零向量，因为在第一个解码步，生成的序列是

空序列。对于  $t > 1$ , 使用以下公式计算  $c_t^d$ :

$$e_{tt'}^d = \mathbf{h}_t^d \top \mathbf{W}_{\text{attn}}^d \mathbf{h}_{t'}^d \quad (11.39)$$

$$\alpha_{tt'}^d = \frac{\exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} \exp(e_{tj}^d)} \quad (11.40)$$

$$\mathbf{c}_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d \mathbf{h}_j^d \quad (11.41)$$

除了使用注意力机制来缓解可能的错误累积问题, 该模型的另外一个核心在于使用了基于强化学习的算法来进行模型参数的优化。如我们在上文提到的, 一般的摘要生成模型会在每个解码步使用极大似然估计损失来进行优化, 如下所示:

$$\mathcal{L}_{ml} = - \sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \dots, y_{t-1}^*, x) \quad (11.42)$$

这种优化目标与最终的摘要评测指标具有不一致性。而 DeepReinSum 模型使用了强化学习中的策略学习, 它可以针对某种离散的评估指标直接进行优化, 并使用自评判的策略梯度 (Self-critical Policy Gradient) 算法<sup>[608]</sup> 来进行参数的更新。

具体的实现过程如下, 首先为每一个训练样本 (文档) 产生两个单独的解码序列 (摘要), 分别记为  $y^s$  和  $\hat{y}$ 。其中,  $y^s$  是从每个解码步的  $p(y_t^s | y_1^s, \dots, y_{t-1}^s, x)$  概率分布中采样获得的序列。而  $\hat{y}$  作为一个基线序列, 是每一步从上述概率分布中取最大概率的单词而获得的序列, 本质上是一种基于贪心策略的采样方案。将  $r(y)$  定义为输出摘要  $y$  的奖励函数, 它是通过某种评估指标 (通常为 ROUGE) 来比较  $y$  与标准摘要  $y^*$  而得到的分数。由此, 强化学习的损失函数可定义如下:

$$\mathcal{L}_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \dots, y_{t-1}^s, x) \quad (11.43)$$

可以看到, 最小化  $\mathcal{L}_{rl}$  等价于最大化采样序列  $y^s$  的条件似然。如果  $y^s$  获得了比基线  $\hat{y}$  更高的奖励, 模型可以得到正反馈, 从而增大  $y^s$  的概率。反之, 模型则会减小  $y^s$  的生成概率。

然而, 这种强化学习目标的一个潜在问题是, 针对特定离散指标 (比如 ROUGE) 进行优化并不能保证输出质量和可读性的提高。为此, 可以考虑在不影响可读性的情况下对这些离散指标进行策略梯度的优化<sup>[609]</sup>。一般来说, ROUGE 评测指标衡量了生成摘要和标准摘要之间的  $n$ -gram 重叠程度, 而对于可读性, 一般使用困惑度 (Perplexity) 来衡量。困惑度可以直接由通过极大似然

训练目标训练的条件语言模型来计算得到。因此，可以定义一个混合的目标函数，它结合了极大似然估计和强化学习目标：

$$\mathcal{L}_{\text{mixed}} = \gamma \mathcal{L}_{rl} + (1 - \gamma) \mathcal{L}_{ml} \quad (11.44)$$

其中，超参数  $\gamma \in [0, 1]$  可以用来控制不同训练阶段各个损失函数的比重。一般来说，在训练的初期，将  $\gamma$  设为 0，并随着训练轮数的增长逐渐增大  $\gamma$  的值，从而先保证模型生成的摘要具有很好的流畅性，再通过优化 ROUGE 指标来使摘要的质量得到整体的提升。

除了 DeepReinSum 算法外，还可以使用基于模型的方法来为强化学习模型提供反馈，一种经典的方案便是序列对抗生成网络（Sequence Generative Adversarial Network，SeqGAN）<sup>[610]</sup>。它的核心思想在于，模型需要同时训练一个生成器  $G$  和一个鉴别器  $D$ ，鉴别器需要尽可能区分真实的文本和  $G$  生成的文本，而生成器则需要尽可能产生接近真实的文本，从而困扰  $D$  做出正确的判断。经过多轮的迭代优化，生成器可以产生与真实文本接近的结果。其中，鉴别器的预测结果可以给生成器合适的反馈，并可通过强化学习的方法来优化生成器。

SeqGAN 的思想也可用在生成式文本摘要的全局优化上<sup>[611]</sup>。此时，生成器就是一个标准的 Seq2Seq 摘要生成模型，输入原文档并输出摘要。鉴别器是一个文本分类器，它的作用就在于试图区分生成的摘要和标准摘要（二分类），并为生成器提供反馈。与标准的 SeqGAN 训练策略类似，首先对生成器进行训练。之后将标准摘要作为正例，生成器产生的摘要作为负例，用于训练鉴别器。然后交替训练生成器和鉴别器，直到收敛。

在实际训练过程中，当生成器的参数固定之后（记为  $G_\theta$ ），通过以下方式来动态更新鉴别器的参数（记为  $D_\phi$ ）：

$$\min_{\phi} -\mathbf{E}_{Y \sim p_{\text{data}}} [\log D_\phi(Y)] - \mathbf{E}_{Y \sim G_\theta} [\log (1 - D_\phi(Y))] \quad (11.45)$$

其中， $Y$  表示摘要结果， $Y \sim p_{\text{data}}$  表示真实的摘要， $Y \sim G_\theta$  表示生成的摘要。另一方面，当鉴别器的参数固定时，需要进一步迭代更新生成器。生成器  $G$  的损失函数由两部分组成：由强化学习的策略梯度计算的损失  $J_{pg}$  和最大似然损失  $J_{ml}$ 。在形式上， $G$  的目标函数是  $J = \beta J_{pg} + (1 - \beta) J_{ml}$ ，其中  $\beta$  是用来计算  $J_{pg}$  和  $J_{ml}$  两个损失的比重，与公式 11.44 类似。 $J_{pg}$  对于参数  $\theta$  的梯度可由以下公式计算：

$$\begin{aligned} \nabla_{\theta} J_{pg} &= \frac{1}{T} \sum_{t=1}^T \sum_{y_t} R_D^{G_\theta} ((Y_{1:t-1}, X), y_t) \cdot \nabla_{\theta} (G_\theta (y_t | Y_{1:t-1}, X)) \\ &= \frac{1}{T} \sum_{t=1}^T \mathbf{E}_{y_t \in G_\theta} \left[ R_D^{G_\theta} ((Y_{1:t-1}, X), y_t) \nabla_{\theta} \log p(y_t | Y_{1:t-1}, X) \right] \end{aligned} \quad (11.46)$$

其中， $R_D^{G_\theta} ((Y_{1:t-1}, X), y_t)$  是动作价值函数，有  $R_D^{G_\theta} ((Y_{1:t-1}, X), y_t) = D_\phi(Y_{1:T})$ ， $T$  是摘要的长度， $Y_{1:t}$  是生成到时间步  $t$  的部分摘要， $X$  表示原文档。从上式可以看出，对解码阶段每一个时间步决策的奖励都是相同的，都为  $D_\phi(Y_{1:T})$ 。如果把这一项替换为  $Y_{1:T}$  与标准摘要的 ROUGE 分

数，就变回了上文提到的基于指标的强化学习优化方式。

### 11.3.2 抽取与生成结合式文本摘要

抽取式文本摘要和生成式文本摘要都存在着各自的优势，于是研究员人们也尝试将抽取式方法和生成式方法结合，并形成一类新的方法。这一类方法的算法流程为：首先对原文档的内容进行抽取，得到初步的摘要内容；再将抽取的结果输入生成器，经过优化得到最终的摘要。我们可以根据第一步中抽取内容的粒度（如字词、句子）对这一类方法进行区分。

#### 1. 词粒度抽取与生成结合方法

Bottom-Up<sup>[612]</sup> 是一种以字词为抽取粒度的方法，其思路是将摘要生成分成如图11.6所示两个主要步骤：(1) 通过自底向上的注意力机制从文本中抽取可以作为摘要的单词；(2) 用第一步抽出的单词辅助生成摘要。

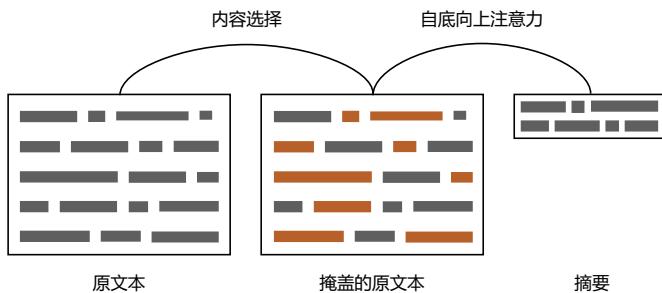


图 11.6 Bottom-Up 模型结构<sup>[612]</sup>

对于第一步抽取单词，可以直接参考基于序列标注的抽取式方法，让文档中的每个单词都拥有一个二元标签（1 和 0，用于表示该词语能否作为摘要的一部分）以及一个嵌入表示用于完成相应的序列标注。初始嵌入表示  $e_i$  由两个部分  $e_i^w$  和  $e_i^c$  拼接而成， $e_i^w$  是预训练模型（如 BERT）学到的词嵌入， $e_i^c$  则是用一个双层双向长短期记忆网络计算得到：

$$e_i^c = \gamma \times \sum_{l=0}^2 s_j \times h_i^{(l)} \quad (11.47)$$

其中  $l$  表示层数， $\gamma, s_0, s_1, s_2$  是可以学习的参数。上下文信息通过这种方式被融入到中  $e_i^c$ 。在计算出  $e_i^c$  后，还需要将其输入到另一个单层双向长短期记忆网络中，以此计算出每个单词的最终嵌入  $h_i$ ，然后就可以通过  $h_i$  计算出各单词能用于生成摘要的概率  $q_i$ ：

$$q_i = \sigma(\mathbf{W}h_i + \mathbf{b}) \quad (11.48)$$

其中  $\mathbf{W}, \mathbf{b}$  为可以学习的参数。最后我们设定阈值  $\epsilon$ , 当模型选择某个单词的概率高于阈值, 我们就可以将其选作关键内容, 进入下一个步骤。在步骤二中, 摘要生成可以通过一个 Seq2Seq 模型来实现。在每个时间步  $j$ , 模型可以选择从词表中生成一个单词或者直接从上一步抽取出的单词中进行复制:

$$\begin{aligned} p(y_j|y_{1:j-1}, x) = & p(z_j = 1|y_{1:j-1}, x) \times p(y_j|z_j = 1, y_{1:j-1}, x) + \\ & p(z_j = 0|y_{1:j-1}, x) \times p(y_j|z_j = 0, y_{1:j-1}, x) \end{aligned} \quad (11.49)$$

其中的  $z_j = 1$  表示当前时间步需要从词表中生成一个字词,  $z_j = 0$  表示当前时间步直接从抽取单词集合中复制一个单词。在复制抽取的单词时, 可以引入注意力机制来提升性能, 即把分布  $p(y_j|z_j = 0, y_{1:j-1}, x)$  替换成:

$$p(\tilde{a}_j^i|x, y_{1:j-1}) = \begin{cases} p(a_j^i|x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{其他} \end{cases} \quad (11.50)$$

其中  $a_j^i$  表示时间步  $j$  时单词  $w_i$  的注意力权重。

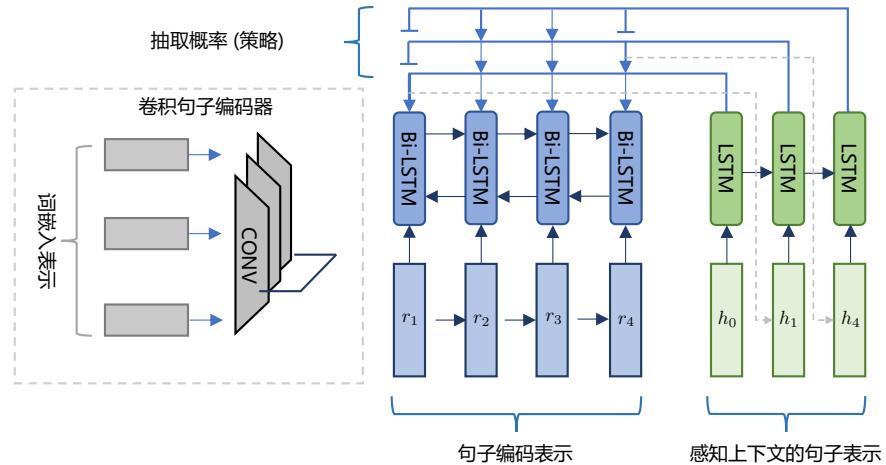
## 2. 句子粒度抽取与生成结合方法

FastRL<sup>[613]</sup> 是一种以句子为粒度的抽取与生成方法, 其生成摘要的过程也可以大致分为两个步骤: (1) 抽取器提取文章中重要的句子; (2) 生成器对第一步提取出来的句子进行压缩和改述, 并生成一段连贯的摘要。与 Bottom-Up 算法的不同之处在于, FastRL 使用了强化学习的方法对抽取和生成的两阶段模型进行了联合训练, 使用了生成模型的摘要结果作为奖励函数来优化抽取模型。FastRL 的模型所使用的抽取器神经网络结构如图11.7所示。

抽取器共由三个部分组成。第一个部分是卷积神经网络, 通过卷积和池化的方式为文本中的每个句子计算出一个表示  $r_j, j = \{1, 2, \dots, n\}$ 。第二个部分是双向长短期记忆网络, 它以  $r_j$  为输入进一步为文本中的每个句子计算出结合全局信息的表示  $\mathbf{h}_j, j = \{1, 2, \dots, n\}$ 。第三部分是单向长短期记忆网络, 负责完成最终的句子选择, 这个部分引入了两跳注意力机制 (2-Hop Attention Mechanism), 首先计算  $\mathbf{h}_j$  对应的包含上下文信息的表示  $e_t$ , 再用其计算每个句子被抽取的概率  $P(j_t|j_1, \dots, j_{t-1})$ :

$$u_j^t = \begin{cases} \mathbf{v}_p^T \tanh(\mathbf{W}_{p1}\mathbf{h}_j + \mathbf{W}_{p2}e_t) & \text{if } j_t \neq j_k \\ -\infty & \forall k < t \\ & \text{otherwise} \end{cases} \quad (11.51)$$

$$P(j_t|j_1, \dots, j_{t-1}) = \text{softmax}(u^t) \quad (11.52)$$

图 11.7 FastRL 模型的抽取模块<sup>[613]</sup>

其中  $e_t$  通过以下公式计算得到：

$$a_j^t = \mathbf{v}_g^T \tanh(\mathbf{W}_{g1} \mathbf{h}_j + \mathbf{W}_{g2} \mathbf{z}_t) \quad (11.53)$$

$$\alpha^t = \text{softmax}(a^t) \quad (11.54)$$

$$e_t = \sum_j \alpha_j^t \mathbf{W}_{g1} \mathbf{h}_j \quad (11.55)$$

其中的  $\mathbf{z}_t$  也是由第三部分的长短期记忆网络计算得到。上述公式中所有的  $\mathbf{W}$  和  $\mathbf{v}$  都是可以学习的参数。只要抽取器选择出的句子足够准确，生成器使用简单的 Seq2Seq 模型完成压缩和改述。

因为需要同时训练抽取器和生成器，但抽取器和生成器之间无法直接传递梯度信息，所以 FastRL 使用了强化学习中的策略梯度来协同训练抽取器和生成器：用生成器生成的结果与真实摘要进行比较，然后将比较结果作为反馈来更新抽取器的参数。图11.8给出了强化学习的过程，其中的  $d_1, d_2, \dots, d_n, s_t$  是文本正文以及摘要中的句子。

## 11.4 文本摘要的评测

与文本分类、命名实体识别等其他自然语言处理任务的评测相比，由于关键信息的选取和文摘的表述没有统一的标准答案，因此文本摘要的人工评测和自动评测都困难很多。与机器翻译、对话系统等面向内容生成（Content Generation）的任务一样，文本摘要也有多维度的评价系统，包括人工评测方案和自动评测指标。

文本摘要的评测按照与任务的相关性可以分为两类，内在评价（Intrinsic Evaluation）方法和外在评价（Extrinsic Evaluation）方法。内在评价与摘要任务相关，它通过直接分析摘要的质量来

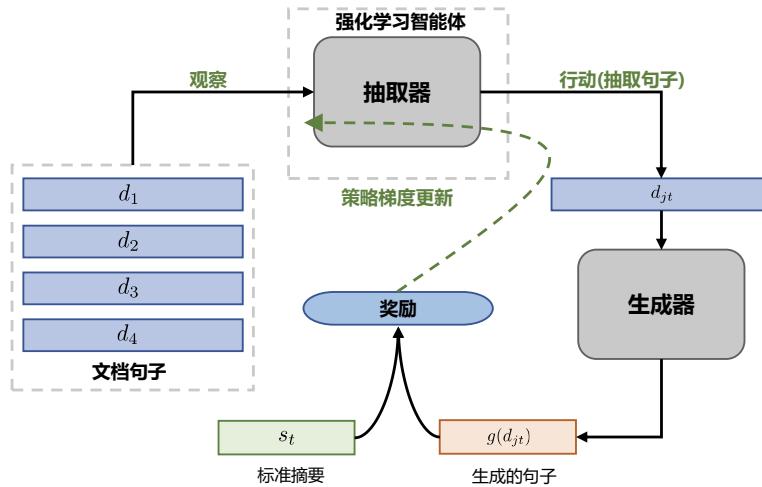


图 11.8 FastRL 利用生成器结果的反馈来更新抽取器<sup>[613]</sup>

评价摘要系统，一般从总体和细分维度两个粒度进行评估。其中，总体评估是指从总体上评价摘要的质量，并给出一个综合分数。细分维度一般可以从“与参考标准的一致性”和“类人性”两个角度来考虑，多方面地对摘要的质量进行评价。外在评价则是一种间接的评价方法，与系统的功能对应，将摘要应用于下游任务或者实际应用系统中，根据摘要在这些任务和系统中产生的实际效果来间接评估摘要的质量。这里我们仅讨论内部评价方法。

一般而言，在评测文本摘要的质量时会关注以下 5 个细分维度：

- (1) 信息量 (Informativeness)，即摘要的内容含量，它可以衡量一段文本所提供的新信息的程度。Peyrard<sup>[614]</sup>对信息量作出了直观上的定义：阅读者一般都拥有背景知识和常识，如果一段摘要能使其获得新信息或者产生认知上的变化，则可认为此摘要具有丰富的信息量。
- (2) 非冗余性 (Non-redundancy)，它体现了摘要精简的特性，即不能反复使用重复或相似的文本描述同一个关键点。
- (3) 流畅度 (Fluency)，它包含了摘要结构的连贯性和语法的正确性两个方面。
- (4) 忠实度 (Faithfulness)，也称为相关性 (Relevance)，即摘要内容是否忠于原文。作为原文的子集，摘要不能包含凭空产生的信息，也不能包含与原文和事实相悖的内容。
- (5) 聚焦程度 (Focus)，也称为显著性 (Saliency)，它衡量了摘要包含关键信息的程度。除了主要信息，原文档往往具有大量的细节描述和补充内容等次要信息。通过聚焦程度，可以评估一个摘要系统的甄别并捕捉原文主要信息的能力。

以上多数评测维度需要借助人工方式进行。评测者一般会根据预定义好的评测指南和范例对摘要质量进行评估。人工评测能灵活应用于多种场景和标准，适合所有的主观性任务。因此，在目前

的文本摘要任务中，人工评测被认为是评价模型优劣的黄金标准。但是，人工评测也存在成本高昂和结果难以复现等问题。相比人工评测，自动评测具有方便快捷、容易复现等优势，因此被广泛应用于摘要评价。尤其是在模型开发的早期阶段，自动评测能帮助开发者快速定位问题。在大多数情况下，自动摘要的评价可以将人工评测方案与自动评测指标相结合，对模型在测试集上的整体效果做自动评测，并随机采样部分测试点进行人工评测，从而兼顾评测的效率和质量。

### 11.4.1 人工评测

人工评测按照执行方式一般分为两类：逐点评估（Point-wise）和逐对评估（Pair-wise）。在逐点评估中，评估者对系统产生的每一个结果按照预定义的维度进行评估评分。一个常见的方案是李克特五点量表（Likert Scale）。给定原始文档和模型输出的摘要，评估者会按照1~5分对摘要某一方面的评测维度进行评分。假设评测维度是流畅度，则1~5的分值依次对应：非常不流畅、不流畅、一般、流畅、非常流畅五种程度。但是，逐点评估具有很强的主观性，导致评估者之间的偏差较大，一致性很低。在逐对评估中，给定相同的原始文档和两个不同系统的输出摘要A和B，评估者需要判断A与B相比哪个更好。与逐点评估的多选项评分相比，逐对评估采用两两比较的方式，降低了评估难度，可以提高评测结果的一致性。但是，如果存在多个需要评测的摘要系统，总评估次数会随着系统数目的增加而呈 $O(N^2)$ 的复杂度上升，成本较高。

人工评测存在主观性，包括摘要任务本身的主观性和评估者自身的主观性。为了尽可能消除人工评测的主观性偏差，一般会让多个评估者对同一条数据进行独立重复评分。因此，衡量多个评估者之间的评测一致性是一个重要的过程。一致性不仅可以体现人工评测质量的高低，还能反映评测任务的难易程度。**Fleiss**卡帕系数（Fleiss's  $\kappa$ ），也称Kappa系数，是度量多个系统一致性的方法，也经常被用于计算人工评测的一致性。定义样本的总数为 $N$ ，类别的总数为 $K$ ，每一个样本有 $n$ 个标注者进行标注， $n_{ij}$ 是将第 $i$ 个样本标注为类别 $c_j$ 的标注者数量。在第 $i$ 个样本上，标注者之间的一致性可以采用两两标注者之间的一致性进行度量。其中，所有一致的两个标注者的组合数为 $C_{n_{ij}}^2 = \frac{1}{2}n_{ij}(n_{ij} - 1)$ ，而每一个样本有 $n$ 个标注者，则所有可能的两个标注者组合数为 $C_n^2 = \frac{1}{2}n(n - 1)$ 。因此，第 $i$ 个样本的标注一致性可以用下式表示：

$$P_i = \frac{1}{C_n^2} \sum_{j=1}^K C_{n_{ij}}^2 = \frac{1}{n(n-1)} \sum_{j=1}^K n_{ij}(n_{ij} - 1) = \frac{1}{n(n-1)} \left( \sum_{j=1}^K n_{ij}^2 - n \right). \quad (11.56)$$

所有样本的平均一致性可以用下式表示：

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i = \frac{1}{Nn(n-1)} \left( \sum_{i=1}^N \sum_{j=1}^K n_{ij}^2 - Nn \right). \quad (11.57)$$

上式计算的标注一致性比较直观，但是尚未考虑两个标注者随机一致的情况，即随机地对样本给出一致或不一致的结果。因此需要另外计算标注者之间随机一致的概率 $\bar{P}_e$ 。首先，计算得到

每个类别  $c_j$  出现的概率为：

$$P_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad \sum_{j=1}^K P_j = 1. \quad (11.58)$$

两个标注者以  $P_j$  的概率随机标记，将某个样本同时标记为  $c_j$  的概率为  $P_j^2$ 。因此，所有类别上的随机标注概率为：

$$\bar{P}_e = \sum_{j=1}^K P_j^2. \quad (11.59)$$

最后，代入  $\bar{P}$  和  $\bar{P}_e$ ，Fleiss 卡帕系数计算如下：

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \quad (11.60)$$

Fleiss 卡帕系数在不同的区间显示不同的一致性程度。系数小于 0 时表示没有一致性，0~0.20 表示轻微（Slight）一致，0.21~0.40 表示一般（Fair）一致，0.41~0.60 表示中等（Moderate）一致，0.61~0.80 表示显著（Substantial）一致，0.81~1.00 表示完美（Perfect）一致。

人工评测的一致性评估主要用来评价标注质量，并体现了标注任务的难易程度。一般而言，标注者需要具备一定的领域专家知识，且需要对原文内容和摘要内容做整体比较，因此摘要任务的人工评测成本高昂，花费时间长。另外，标注者之间的差异会增大标注结果的方差，导致研究结果难以复现。以上问题也说明，人工评测同样存在着很大挑战。摘要任务的场景和评价维度多种多样，如何实现高质量、易复现的人工评测方案是一个值得深入研究的问题。

### 11.4.2 自动评测

人工评测虽然可以提供丰富的信息，但是大规模的人工评测耗时长、工作量大、成本高。自动评测则可以提供高效、低成本、一致的评测，在模型开发过程中这些特点都受到研究人员青睐。随着评测技术的发展，自动评价结果也具有了更好的指导意义。

#### 1. 面向召回率的要点评估

ROUGE<sup>[606]</sup> (Recall-Oriented Understudy for Gisting Evaluation)，称为面向召回率的要点评估，也是文本摘要中最常用的自动评价指标之一。ROUGE 与机器翻译的评价指标 BLEU 的类似，能根据机器生成的候选摘要和标准摘要(参考答案)之间词级别的匹配来自动为候选摘要评分。ROUGE 包含一系列变种，其中应用最广泛的是 ROUGE-N，它统计了  $n$ -gram 词组的召回率，通过比较标准摘要和候选摘要来计算  $n$ -gram 的结果。给定标准摘要集合  $S = \{Y^1, Y^2, \dots, Y^M\}$  以及候选摘要  $\hat{Y}$ ，则 ROUGE-N 的计算公式如下：

$$\text{ROUGE-N} = \frac{\sum_{Y \in S} \sum_{n\text{-gram} \in Y} \min[\text{Count}(Y, n\text{-gram}), \text{Count}(\hat{Y}, n\text{-gram})]}{\sum_{Y \in S} \sum_{n\text{-gram} \in Y} \text{Count}(Y, n\text{-gram})}. \quad (11.61)$$

其中  $n$ -gram 是  $Y$  中所有出现过的长度为  $n$  的词组,  $\text{Count}(Y, n\text{-gram})$  是  $Y$  中  $n$ -gram 词组出现的次数。

我们以两段摘要文本为例给出了 ROUGE 分数的计算过程: 候选摘要  $\hat{Y} = \{\text{a dog is in the garden}\}$ , 标准摘要  $Y = \{\text{there is a dog in the garden}\}$ 。可以按照公式 11.61 计算 ROUGE-1 和 ROUGE-2 的分数为:

$$\text{ROUGE-1} = \frac{|\{\text{is, a, dog, in, the, garden}\}|}{|\{\text{there, is, a, dog, in, the, garden}\}|} = \frac{6}{7} \quad (11.62)$$

$$\text{ROUGE-2} = \frac{|\{\text{(a dog), (in the), (the garden)}\}|}{|\{\text{(there is), (is a), (a dog), (dog in), (in the), (the garden)}\}|} = \frac{1}{2} \quad (11.63)$$

需要注意的是 ROUGE 是一个面向召回率的度量, 因为公式 11.61 的分母是标准摘要中所有  $n$ -gram 数量的总和。相反地, 机器翻译的评价指标 BLEU 是一个面向精确率的度量, 其分母是候选翻译中  $n$ -gram 的数量总和。因此, ROUGE 体现的是标准摘要中有多少  $n$ -gram 出现在候选摘要中, 而 BLEU 体现了候选翻译中有多少  $n$ -gram 出现在标准翻译中。

另一个应用广泛的 ROUGE 变种是 ROUGE-L, 它不再使用  $n$ -gram 的匹配, 而改为计算标准摘要与候选摘要之间的最长公共子序列, 从而支持非连续的匹配情况, 因此无需预定义  $n$ -gram 的长度超参数。ROUGE-L 的计算公式如下:

$$R = \frac{\text{LCS}(\hat{Y}, Y)}{|Y|}, \quad P = \frac{\text{LCS}(\hat{Y}, Y)}{|\hat{Y}|}, \quad (11.64)$$

$$\text{ROUGE-L}(\hat{Y}, Y) = \frac{(1 + \beta^2)RP}{R + \beta^2P}. \quad (11.65)$$

其中,  $\hat{Y}$  表示模型输出的候选摘要,  $Y$  表示标准摘要。 $|Y|$  和  $|\hat{Y}|$  分别表示摘要  $Y$  和  $\hat{Y}$  的长度,  $\text{LCS}(\hat{Y}, Y)$  是  $\hat{Y}$  与  $Y$  的最长公共子序列长度,  $R$  和  $P$  分别为召回率和精确率, ROUGE-L 是两者的加权调和平均数,  $\beta$  是召回率的权重。在一般情况下,  $\beta$  会取很大的数值, 因此 ROUGE-L 会更加关注召回率。

还是以上面的两段文本为例, 可以计算其 ROUGE-L 如下:

$$\text{ROUGE-L}(\hat{Y}, Y) \approx \frac{\text{LCS}(\hat{Y}, Y)}{\text{Len}(Y)} = \frac{|\{\text{a, dog, in, the, garden}\}|}{|\{\text{there, is, a, dog, in, the, garden}\}|} = \frac{5}{7} \quad (11.66)$$

## 2. 基于嵌入表示的度量

ROUGE 是基于候选文本与参考文本之间的精准匹配来评价文本的质量。但是在一些情况下, 不同的词语或短语可以表达同一种语义。例如, “立刻”和“马上”, 两者的含义相同, 但是基于  $n$ -gram 的匹配就会失效。一个常用的解决方案为基于嵌入表示的度量 (Embedding-Based Metrics), 使用分布式词嵌入来计算词语之间的相似度, 具体来说包括静态词向量和上下文相关的词向量两种方案。

对于静态词向量方案，词表中所有的词语都对应一个固定不变的分布式词嵌入向量。静态分布式词嵌入可以通过 Word2Vec<sup>[615]</sup> 等方法预训练获得。给定标准摘要  $Y = \{y_1, \dots, y_{|Y|}\}$  和候选摘要  $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_{|\hat{Y}|}\}$ ，其静态词嵌入向量序列为  $\mathbf{E} = (\mathbf{e}(y_1), \dots, \mathbf{e}(y_{|Y|}))$  和  $\hat{\mathbf{E}} = (\mathbf{e}(\hat{y}_1), \dots, \mathbf{e}(\hat{y}_{|\hat{Y}|}))$ 。可以将摘要中每个词的词向量结合起来计算篇章级特征向量，进而计算标准摘要和候选摘要的篇章级特征向量的余弦相似度。一种常见的计算篇章级特征向量的方法为词向量平均（Embedding Average），即把文本中所有词的词向量做平均得到篇章级特征向量：

$$\mathbf{e}(Y) = \frac{\sum_{y_i \in Y} \mathbf{e}(y_i)}{|Y|}. \quad (11.67)$$

得到篇章级特征向量后，可以计算摘要的相似度  $S(Y, \hat{Y}) = \frac{\mathbf{e}(Y) \cdot \mathbf{e}(\hat{Y})}{\|\mathbf{e}(Y)\| \|\mathbf{e}(\hat{Y})\|}$ 。

除了直接获取篇章级特征向量来计算相似度，还可以首先计算参考摘要和候选摘要中的词两两之间的余弦相似度，然后基于这些词级相似度得到篇章级相似度。一种常见的方案是贪心匹配（Greedy Matching），其篇章级相似度  $S(Y, \hat{Y})$  可计算如下：

$$S(Y, \hat{Y}) = \frac{1}{2}[GM(Y, \hat{Y}) + GM(\hat{Y}, Y)], \quad (11.68)$$

$$GM(Y, \hat{Y}) = \frac{1}{|Y|} \sum_{y_i \in Y} \max_{\hat{y}_j \in \hat{Y}} S(y_i, \hat{y}_j). \quad (11.69)$$

其中， $S(y_i, \hat{y}_j) = \frac{\mathbf{e}(y_i) \cdot \mathbf{e}(\hat{y}_j)}{\|\mathbf{e}(y_i)\| \|\mathbf{e}(\hat{y}_j)\|}$ 。因为  $GM(Y, \hat{Y})$  具有非对称性，所以需要对  $GM(Y, \hat{Y})$  和  $GM(\hat{Y}, Y)$  做平均得到最终的篇章级相似度。

由于词语在不同的上下文中可以有不同的语义，而静态词向量难以应对这样的情况，所以还可以使用上下文相关的词向量来计算相似度，比如 BERTScore<sup>[616]</sup> 使用预训练语言模型 BERT 获得摘要的上下文相关的词向量序列。BERTScore 仍然使用了贪心匹配方式，将  $Y$  中的每个词与  $\hat{Y}$  中的每个词做匹配来计算召回率  $R_{BS}$ ，这里相似度分数简化为计算  $\mathbf{e}(y_i) \cdot \mathbf{e}(\hat{y}_j)$ 。同时，用类似的方式可计算精确率  $P_{BS}$ ，并得到 F1 值  $F_{BS}$  如下：

$$R_{BS} = \frac{1}{|Y|} \sum_{y_i \in Y} \max_{\hat{y}_j \in \hat{Y}} \mathbf{e}(y_i) \cdot \mathbf{e}(\hat{y}_j), \quad (11.70)$$

$$P_{BS} = \frac{1}{|\hat{Y}|} \sum_{\hat{y}_i \in \hat{Y}} \max_{y_j \in Y} \mathbf{e}(\hat{y}_i) \cdot \mathbf{e}(y_j), \quad (11.71)$$

$$F_{BS} = 2 \frac{P_{BS} R_{BS}}{P_{BS} + R_{BS}}. \quad (11.72)$$

有研究表明，稀有词比通用词更能指示句子的相似度，因为通用词频繁出现在大量的文本中，计算它们的词向量相似度会影响到对关键内容的考量。因此，BERTScore 采用整个测试集中的逆文

档频率 (Inverse Document Frequency, IDF) 对相似度进行加权。以召回率  $R_{BS}$  为例, 加权后的公式为:

$$R_{BS} = \frac{\sum_{y_i \in Y} \text{IDF}(y_i) \max_{\hat{y}_j \in \hat{Y}} \mathbf{e}(y_i) \cdot \mathbf{e}(\hat{y}_j)}{\sum_{y_i \in Y} \text{IDF}(y_i)}. \quad (11.73)$$

实验结果表明, 相比 ROUGE 等基于精准词匹配的指标, 基于词嵌入的度量与人工评价有更高的统计相关性。但是, 预训练词向量的质量会一定程度上影响评估的效果。

## 11.5 文本摘要语料库

如前所述文本摘要被广泛应用于多种场景中, 任务类型包括单文档摘要、多文档摘要、对话摘要、跨语言文本摘要和多模态文本摘要等多种类型。本节中将按照任务类型对文本摘要常见语料库进行介绍。

### 11.5.1 单文档摘要语料库

- **CNN/DailyMail**<sup>[617]</sup>: 该数据集是被广泛关注和研究的短文本摘要数据集, 它包含了 311,672 个新闻/摘要对, 新闻数据来源是美国有线电视新闻网和《每日邮报》。新闻平均长度为 766 个词 (29.74 个句子), 摘要的平均长度为 53 个词 (3.72 个句子)。
- **LCSTS**<sup>[618]</sup>: 该数据集是常用的中文短文本摘要数据集, 包含 240 万个新闻/摘要对, 数据来源为微博认证的官方微博。
- **Arxiv/PubMed**<sup>[619]</sup>: 这两个数据集是被广泛应用的长文本数据集, 分别来源于 arXiv 和 PubMed 的学术网站上的论文和摘要。ArXiv 数据集包含 21.5 万个论文/摘要对, 论文平均长度为 4938 个词, 摘要平均长度为 220 个词。PubMed 数据集包含 13.3 万个论文/摘要对, 论文平均长度为 3016 个词, 摘要平均长度为 203 个词。

### 11.5.2 多文档摘要语料库

- **Multi-News**<sup>[620]</sup>: 该数据集是大规模的多文档摘要数据集, 场景是新闻文档, 数据来源是 Newser 网站。此网站上的新闻稿会引用多个新闻源, 这些新闻源被归为输入文档集合。Multi-News 共包含 56216 个文档集合/摘要对, 文档集合的平均长度为 2103 个词, 摘要的平均长度为 264 个词。
- **WikiSum**<sup>[621]</sup>: 该数据集是基于英文维基百科的多文档摘要数据集。在每个实例中, 输入由维基百科主题 (文章标题) 和非维基百科参考文档的集合组成, 目标是生成维基百科文章的精简文本。数据集以 8: 1: 1 的比例被划分为训练集、验证集和测试集, 分别包含 1,865,750, 233,252 和 232,998 个样本。

### 11.5.3 对话摘要语料库

- **SAMSum**<sup>[622]</sup>: 该数据集是一个常用在线聊天的摘要数据集，构造方式比较特殊，是由语言专家模拟人类实际交流的特征构建的虚拟对话和对应的摘要。它包含 16,369 个对话段，大多数为双人对话，对话的平均长度为 120 个词，摘要的平均长度为 23 个词。
- **DialogSum**<sup>[623]</sup>: 该数据集是一个包含现实生活场景对话的摘要数据集，这些对话涵盖了广泛的日常生活主题和面对面交流场景。它包含 13,460 个对话，对话的平均长度为 190 个词，摘要的平均长度为 30 个词。
- **AMI/ICSI**<sup>[624, 625]</sup>: AMI 和 ICSI 是经典的会议摘要数据集。AMI 是关于工业环境中产品设计的会议数据集。它由 137 场会议组成，包含会议记录及其相应的会议摘要。ICSI 数据集是由一个学术会议数据集组成，来自于伯克利的国际计算机科学研究所 (ICSI) 举行的 59 次每周小组会议，以及对应的摘要。与 AMI 不同的是，ICSI 会议的内容是专门针对学生之间关于研究的讨论。

### 11.5.4 多模态文本摘要语料库

- **MSMO**<sup>[593]</sup>: 该数据集是一个公开的多模态新闻摘要数据集，数据来源是《每日邮报》。数据集包含了 31.4 万条带有图像的新闻文档，每篇新闻文档平均包含 6.6 张图片，文档和摘要的平均长度分别是 720 和 70。
- **How2**<sup>[626]</sup>: 该数据集包含大约 8 万个教学视频（约 2,000 小时）以及相关的英文字幕和视频摘要文本。其中，大约 300 小时的视频内容还通过众包的方式翻译成葡萄牙语。

### 11.5.5 跨语言文本摘要语料库

- **Global Voices**<sup>[592]</sup>: 该数据集关注多语言新闻的英文摘要，主要包含 15 种语言的新闻及其摘要。Global Voices 网站采用众包的方式人工标注了高质量的英文摘要，对于非英语的新闻文档，其均有对应的英文翻译。
- **En2ZhSum/Zh2EnSum**<sup>[627]</sup>: 该数据集使用了自动机器翻译系统将常见的基准文本摘要数据集进行了翻译。其中，En2ZhSum 将 CNN/Dailymail 和 MSMO 两个数据集的摘要内容从英文翻译到中文。而 Zh2EnSum 则是将 LCSTS 数据集的摘要内容从中文翻译到英文。数据集中的其他信息与原始数据集相比保持不变。
- **WikiLingua**<sup>[628]</sup>: 该数据集包括来自 WikiHow 的 18 种语言的文档和摘要，总计约 77 万条文章和摘要对。在 WikiHow 中，文章包含有描述操作步骤的图像。通过对齐这些图像的文本，可以将跨语言的文章-摘要对进行对齐和抽取。

## 11.6 延伸阅读

尽管基于深度学习尤其是预训练技术的文本摘要已经取得了显著的进展，但在真实场景大规模应用文本摘要仍面临着诸多挑战。首先是数据资源的问题，取得高质量的摘要数据往往成本高昂，有时候我们必须要面临低资源情况下训练数据缺乏的问题。其次是摘要的场景丰富多样，不仅包括文本的类别（新闻、评论、学术论文等），还包括多模态数据（对话、图像、视频等），这些场景领域差异大，导致训练好的摘要模型难以迁移。最后，文本摘要的评估也是亟待解决的难题，设计合理的高效的评估方法能促进模型的迭代，但其具有很大的挑战性。

得益于大规模预训练模型的成功，自动文本摘要系统的性能有了巨大的提升。预训练模型以自监督任务在海量文本语料库上进行预训练，然后在下游摘要任务上进行微调，实验结果表明只需经过有限的训练样本即可在各种摘要基准数据集上取得最先进的性能<sup>[302, 307, 600, 629]</sup>。这表明预训练模型是解决低资源摘要问题的十分有前景的方向。但是，预训练模型也存在着推理速度较低，微调优化难度较高等问题。最近基于提示学习的预训练微调策略则聚焦于提高预训练模型的效率<sup>[322]</sup>。

强化学习（Reinforcement Learning, RL）的训练策略可以面向任何用户自定义的指标对模型进行全局优化，包括许多不可微分的指标（ROUGE 等）。这些指标可作为训练摘要模型的反馈<sup>[607, 610, 630]</sup>。通过这种方式，我们可以通过利用外部资源和不同数据集的特征来改进当前的强化学习模型。但是，强化学习也有训练困难，摘要流畅度损失等问题。目前，基于重排序的后处理方法受到了摘要领域的广泛关注，它通过模型的不同解码采样策略得到多个候选摘要，再通过排序模型对候选摘要进行全局的评估和选择<sup>[631, 632]</sup>。通过这种方式，可以一定程度上同时满足摘要的流畅度和自定义的全局指标。

目前摘要模型在新闻语料库上<sup>[617, 620]</sup>得到了广泛的迭代更新和评估。然而，新闻的写作风格决定了大多数新闻文章的前段落和句子即可视为摘要。这导致模型倾向于直接依据句子的位置决定摘要的内容，且往往直接进行内容的复制，而不是得到高度抽象的摘要<sup>[612]</sup>。为了让模型适应更加多样化的场景，研究人员构造了更加抽象的摘要数据集<sup>[633]</sup>，以及探索了在其他领域的摘要任务，如对话摘要任务<sup>[622]</sup>。此外，多模态文本摘要、跨语言文本摘要也受到了越来越广泛的关注。未来的趋势是更多摘要场景的数据集会不断涌现，以构建更好更具有迁移能力的摘要系统。

目前文本摘要的大多数自动评估指标，例如 ROUGE 和 BERTScore 等，不足以全面合理地评估生成摘要的整体质量<sup>[634]</sup>。我们仍然需要依赖人类专家对摘要的一些关键特征进行评估，例如事实正确性、流畅性和相关性。因此，设计一个更好的摘要评估指标或评价系统是一个非常重要且具有挑战的方向。这些指标或系统需要更高效地更准确地捕捉与人类一致的评估特征。

## 11.7 习题

- (1) 抽取式摘要方法和生成式摘要方法的区别是什么？分别适用于哪些场景？
- (2) 对于文档中的未登录词，如何将其输出到摘要中？

(3) 你能否设计一种强化学习训练策略，在不截断摘要的前提下，使得模型生成的摘要长度不超过某个定值？

(4) 基于自回归生成模型的特点，你能否设计一种方案来评估摘要的流畅度？

(5) 请计算以下两段文本的 ROUGE-1、ROUGE-2、ROUGE-L 分数，并讨论 ROUGE 评测指标的优缺点。

(a) a big black bear sat on a big black bug

(b) a big black bug bit a big black bear

# 12. 知识图谱

知识图谱（Knowledge Graph）是谷歌公司于 2012 首次提出的概念，但是其研究历史可以追溯到费根鲍姆教授（B.A.Feigenbaum）在 1977 年提出的知识工程（Knowledge Engineering）以及 20 世纪 90 年代后期开始的语义网（Semantic Web）。知识工程、语义网以及知识图谱都属于人工智能中知识这一核心命题。知识图谱并不是单一技术的研究，而是一个系统工程，其研究内容涵盖知识表示、知识存储、知识推理、图谱构建、图谱问答等方面，还涉及自然语言处理、机器学习、图数据库、逻辑推理等多个交叉领域。自然语言和知识密切关联，知识是实现计算机对自然语言真正理解必不可少的关键部分。知识图谱在自然语言处理中也发挥着越来越重要的作用，是智能问答、语义检索、机器翻译、语义表示等任务的重要基础。

本章首先介绍知识图谱的基本概念和发展历程，在此基础上介绍知识图谱构建，知识图谱推理、基于知识图谱的问答以及知识图谱存储。

## 12.1 知识图谱概述

知识图谱（Knowledge Graph, KG）是指采用图结构表示实体（包括物体、事件或抽象概念）及其之间关系的知识库（Knowledge Base）。在知识图谱中，图的节点表示实体，图的边表示实体之间的关系。图12.1给出了以金庸为中心的知识图谱示例，图中的节点包含人名等实体，也包含职业等抽象概念，节点之间的边表示了实体之间的各类关系。利用知识图谱可以构建医学、生物、金融、化学等各类型领域知识，一些常识知识也可以使用知识图谱进行表示和利用。知识图谱的目标是利用图结构对知识进行表示，在此基础上识别和推断事物之间的复杂关系并沉淀各类型知识。

相比于直接利用自然语言文本中所包含的知识，计算机算法更容易利用基于图结构所表示的知识。2012 年谷歌公司在提出知识图谱概念的同时，发布了其知识搜索产品。目前绝大部分搜索引擎都将知识图谱作为重要的底层支撑技术，对于很多用户查询，也采用了基于知识图谱的问答算法直接回答用户问题。如图12.2所示，针对“金庸和徐志摩是什么关系？”的问题，搜索引擎直接通过知识图谱给出了“表兄弟”关系，并通过生日计算得到了他们的年纪差。这种基于知识图谱的问答大幅度提升了用户使用搜索引擎的体验。

知识图谱按照其所描述的主要内容可以分为四类：事实知识图谱，概念知识图谱，语言知识

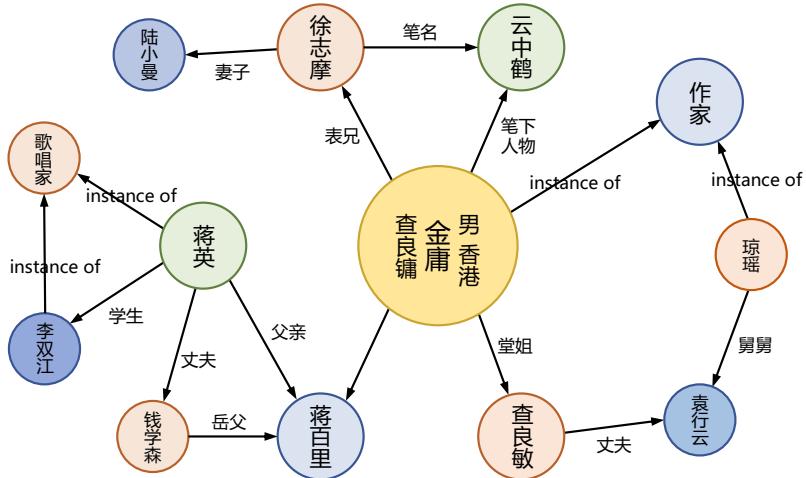


图 12.1 知识图谱样例



图 12.2 基于知识图谱的问题回答样例（来源：搜狗搜索）

图谱和常识知识图谱。

- (1) 事实知识图谱包含了现实世界中各实体之间的关系，比如：“梅西”出生于“阿根廷”。
- (2) 概念知识图谱描述概念之间的子类关系，比如：“老虎”是“哺乳动物”，“足球运动员”属于“运动员”。
- (3) 语言知识图谱是人类语言中蕴含的词法、句法、语义和语用等知识，比如：WordNet、Hownet都可以认为是一种词法知识图谱。
- (4) 常识知识图谱描述人类与世界交互积累的经验与知识，比如：“猫爱吃鱼”、“冬天可能会下雪”。

知识图谱一般由<头实体，关系，尾实体>组成的三元组为基本元素构成。实体一般为世界上的具象事物或者抽象概念，而实体之间的联系则定义为关系。以“刘备是刘禅的父亲”的世界知

识为例，知识图谱将其存储为三元组 < 刘备，父亲，刘禅 >，其中“刘备”为头实体，“刘禅”为尾实体，而“父亲”则为关系。通过扩展这样的三元组，最终会形成一张巨大的知识网络，网络的节点为实体，边则为实体之间的关系，这样的知识网络即为知识图谱。知识图谱可以形式化的定义为：

$$\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\} \quad (12.1)$$

其中  $\mathcal{E}$  表示为实体集合， $\mathcal{R}$  表示为关系集合，而  $\mathcal{F}$  表示为事实集合。知识图谱中的一个事实即为上文中提到的三元组  $< h, r, t > \in \mathcal{F}$ ， $h$  表示头实体， $r$  表示关系， $t$  表示尾实体， $h, t \in \mathcal{E}$ ， $r \in \mathcal{R}$ 。

### 12.1.1 知识图谱发展历程

虽然知识图谱的概念在 2012 年才被首次提出，但是关于知识表示和知识推理的研究一直贯穿于整个自然语言处理和人工智能的发展过程中。自 20 世纪 70 年代中期开始，人工智能领域开始逐渐认识到知识在智能系统的重要性，并提出知识工程<sup>[635]</sup> 的概念，自此知识表示、知识获取、知识推理等知识相关研究就成为了人工智能领域的研究重点和难点。

20 世纪 80 年代开始哲学领域的本体（Ontology）概念引入人工智能<sup>[636]</sup>，是指用规范化方法描述概念、实体、术语及其相互关系。同一时期，研究人员们也提出了语义网络（Semantic Network 或 Semantic Net）<sup>[637]</sup> 用于形式化地表示知识。语义网使用有向或者无向图中的顶点表示概念，边表示概念之间的语义关系。1993 年 Gruber 关于本体给出了更通用定义，将本体定义为对某一智能体（Agent）或智能体群体中存在的概念和关系的一种描述<sup>[638]</sup>。在这期间，企业和学术界构造了包括 CYC、WordNet、BFO（Basic Formal Ontology）、HowNet 等在内的大量通用和领域本体库。

1998 年 Tim Berners-Lee 提出了语义网（Semantic Web）概念<sup>[639]</sup>，其核心是通过定义标准的标志语言（Markup Language）和构造相关处理工具，对互联网上的文档添加能够被计算机所理解的元数据（Meta Data），从而扩展互联网，使之成为一个通用的信息交换媒介。语义网通过可扩展标记语言（XML）、资源描述框架（RDF）以及本体等规范化的描述体系和结构定义来构建语义表达框架。从大数据的视角，语义网也可以称为关联数据（Linked Data），通过可链接的 URI 来发布、共享、连接互联网中各类数据，构建语义数据网络，如图 12.3 所示。在语义网提出后的十几年里，出现了大量的语义网项目，包括 Freebase、LinkingOpenData、WikiData 等。我国在语义网方面也开展了大量研究，OpenKG 项目也收录了大量中文语义网开放数据集。

2012 年谷歌公司以“Things, Not Strings!”为主题发布了知识图谱搜索产品，希望解决用户使用传统搜索引擎需要阅读文章并自行寻找答案，即字符串（Strings）级别搜索。试图构建事物（Things）对象级别搜索，通过构建大规模结构化的事物精确描述以及事物之间的关联，使得用户可以直接得到答案或对象级精准搜索。近十年来，知识图谱快速发展，大量超大规模的知识图谱陆续发布。百度在 2020 年发布的多源异构中文知识图谱，覆盖超过 50 亿实体和 5500 亿事实。阿里巴巴构建的数字商业知识图谱 AliOpenKG，包含 1600 万实体、67 万的核心概念、2681 类关系以及 18 亿三元组。美团构建的餐饮娱乐知识图谱包含 23 类概念、16 亿实体、486 亿三元组。谷歌的知识图谱

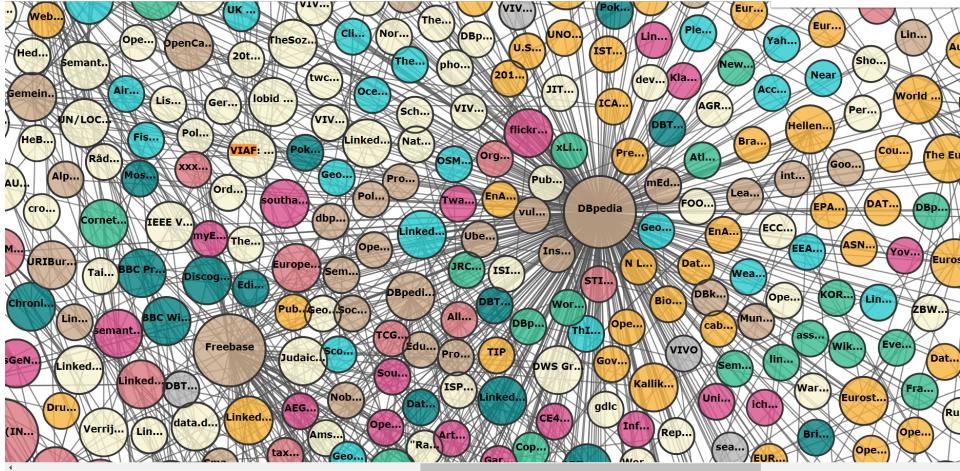


图 12.3 数据关联网络样例

也从 2021 年发布时的 5 亿实体, 35 亿关系, 增长到 2020 年包含 50 亿实体, 5000 亿关系<sup>[640]</sup>。知识图谱也已经成为了搜索引擎、推荐系统、智能问答等应用系统中不可获取的基础组件。

### 12.1.2 知识图谱研究内容

知识图谱研究涉及多个领域, 是典型的交叉领域研究。如图 12.4 所示, 知识图谱涉及到机器学习、自然语言处理、数据库、图算法等领域。知识图谱不仅在构建、推理等阶段需要使用机器学习算法, 同时知识图谱也可以与机器学习算法结合, 在基于知识的特征、基于知识图谱的嵌入表示、图神经网络等方面与基于特征的机器学习方法以及神经网络方法紧密结合。在自然语言处理方面, 知识图谱的构建离不开实体识别、关系抽取、事件抽取等自然语言处理中的信息抽取技术, 同时知识图谱作为重要的底层支撑, 也是自然语言处理中智能问答、机器翻译等任务不可或缺的组成部分。近年来, 知识图谱在预训练语言模型方面也发挥了越来越重要的作用, 包括清华大学 ERNIE<sup>[32]</sup>、百度 ERNIE<sup>[33]</sup>、LUKE<sup>[641]</sup> 等在内的预训练语言模型都在不同层面使用了知识图谱。大规模知识图谱的广泛应用, 知识图谱的存储和检索问题也与数据库研究的产生交叉, 推动了能够存储和快速检索包含数百亿节点和数千亿边规模的图的分布式图数据库的发展。知识图谱使用图结构进行知识存储, 因此包括最短路径识别、子图识别、中心度分析等在内的图算法也在知识图谱构建和使用中发挥着重要作用。

由于知识图谱涉及到多个交叉研究领域, 相关知识点繁多, 我们从知识图谱表示、存储、构建、推理、应用等几个技术维度对知识图谱所涉及的研究内容进行介绍。

**1. 知识图谱表示:** 知识图谱根据使用场景和应用需求, 可以采用有向图、有向标记图 (Directed Labelled Graph)、本体描述语言等方式在逻辑层面进行表示。从物理层面研究上述逻辑表示的数据模型, 主要包含属性图、RDF 图模型、OWL 本体语言等。近年来, 随着神经网络研究的兴起,

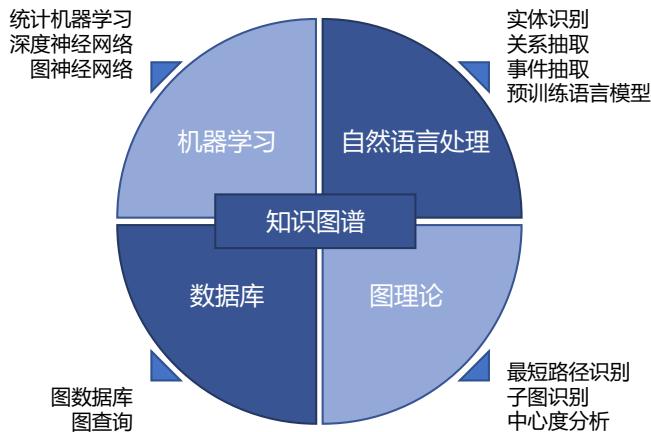


图 12.4 知识图谱研究涉及领域

基于向量的表示方法因为计算高效等特点也被广泛使用。

**2. 知识图谱存储：**随着知识图谱规模不断扩大，如何能够高效存储和检索包含百亿级顶点千亿级边规模的图是知识图谱应用必不可少的基础工作。在知识图谱表示的基础上，如何利用传统关系型数据库和原生图数据库实现知识图谱存储和检索，以及各种方案的优缺点是知识图谱存储所要研究的重点内容。

**3. 知识图谱构建：**知识广泛存在于自然语言文本、半结构化以及结构化的数据中，知识图谱构建主要目标就是研究如何利用实体识别、关系抽取等信息抽取技术，从自然语言文本中抽取实体、属性、关系、事件等知识图谱要素的方法，以及属性补全、实体链接、实体对齐等知识图谱扩展和融合方法。

**4. 知识图谱推理：**知识图谱不仅可以提供知识的存储和检索，更重要的是可以根据构建的已知知识进行归纳、推断和预测未知的事实。知识图谱推理的研究主要基于符号逻辑和表示学习，实现演绎、归纳、溯因、类比等类型推理。

**5. 知识图谱应用：**在构建了大规模高质量知识图谱后，如何将知识与不同的自然语言处理任务进行深度融合，构建基于知识图谱的智能推荐系统、基于知识图谱的智能问答以及知识增强的自然语言处理算法是知识图谱应用所重点研究的内容。

在本章中，我们将分别针对上述知识图谱研究内容进行具体介绍，针对知识图谱应用以知识图谱问答为例进行介绍。

## 12.2 知识图谱表示与存储

知识图谱应用需要使用一种合理的知识表达方式，保证结构化的知识可以被计算机正确处理。从某种意义上可以说，知识表示是贯穿知识库的构建与应用全过程的关键问题。传统的知识图谱表示方法一般以符号式为主，包括属性图、RDF、OWL 本体语言等。符号表示方法的特点是可解释性强，但是依赖知识描述的准确性。近年来，随着深度神经网络研究的兴起，基于向量式的表示方法逐渐引起人们重视。向量式的方法容易捕获隐式知识，计算效率高，但是可解释性差。此外，向量式表示本质上属于统计模型，对没见过的实体表示质量较差，而自然界的实体一般呈长尾分布，这也在一定程度上限制了向量式表示方法的应用。

知识图谱的表示属于人们对知识图谱定义的数据描述，而知识图谱存储则对应知识在物理层面上如何被计算机组织存放。显然，想要在下游任务方便地利用知识图谱，只有其逻辑描述方法还不够，还需要将其在物理介质上进行存储。设计知识图谱的存储方法，要结合知识图谱的图结构模型，即图的结构信息，还要考虑节点与边的属性所包含的语义信息。在此基础上，也要针对知识存储的空间利用率以及检索查询效率等问题进行合计。

本节将对符号式的知识图谱表示方法和向量式的表示方法进行介绍，并将介绍两类知识图谱存储方法：基于表的知识图谱存储和基于图的知识图谱存储。

### 12.2.1 知识图谱的符号表示

知识图谱需要建模各种实体、概念之间的关系，图因为其直观性和扩展性，自然地成为知识图谱描述数据的首选方法。图由点和边构成，图上的点可以建模实体，边则可以对应实体对之间的关系。图的表示方法除了降低了知识的理解难度之外，也便于机器存储。

在简单的应用场景下，无向图即可满足需求。若要进一步增强知识图谱的表达能力，给实体和边添加属性，则可以选择有向标记图。常用的有向图模型有两种：属性图和 RDF 图模型。在更加复杂的场景下，比如建模传递关系、自反关系，有向标记图仍然不能满足需要，这时候可以选用 RDFS/OWL 本体语言作为描述语言对 RDF 进行扩展。本节中，我们会分别对这几种常见的知识图谱表示方法进行介绍。

#### 1. 属性图

属性图（Property Graph）<sup>[642]</sup> 是一种有向标记图，包含三个要素：节点（Vertex）、边（Edge）、属性（Property）。节点对应知识图谱中的实体，边则是实体对之间的关系描述（需要注意的是边是有向的且有类型区别），其出发的节点为源节点，到达节点为目标节点，相应的边的类型则对应实体之间的关系标签。属性本质上是一个键值对，属性图可以灵活地为节点和边添加任意属性。图12.5给出了一个属性图示例，其描述了公司内员工、部门主管以及项目之间的关系信息。以员工和部门主管为例，图中的员工、部门主管都是用节点表示，员工节点和主管节点之间通过被任用的关系边建立连接。此外，在员工节点，还为其添加了很多属性，比如员工编号、生日等信息。当

然边也可以添加属性，比如在边“任命”的属性中添加时间信息。

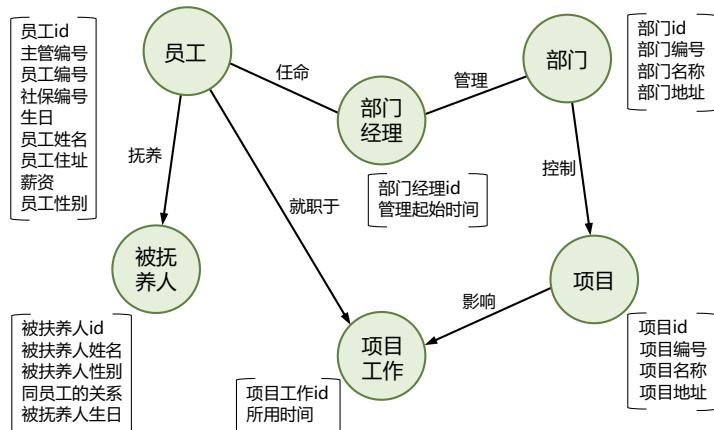


图 12.5 属性图示例

属性图模型在工业界广泛使用，一般通过图数据库 Neo4J<sup>[643]</sup> 等实现。属性图的优点是直观、灵活，可以根据使用需求任意给节点或者边添加属性。此外，Neo4J 的图数据库中有大量针对图结构进行的优化工作，使得其查询效率有显著提升，具体信息将在知识图谱存储章节进行详细介绍。属性图与无向图表示的不同之处在于，在属性图的模型中，关系被提升到了与实体一样重要的程度。属性图提供了一个丰富的视角，即不同种类的数据如何相关，而这些数据依赖关系在普通的关系型数据库中难以直接展示。属性图的缺点则是无法建模深层次的语义信息，这直接限制了其语义推理能力。

## 2. RDF 图模型

资源描述框架（Resource Description Framework, RDF）<sup>[644]</sup>，是由国际万维网联盟 W3C 制定，用于描述实体/资源的数据模型。RDF 的基本组成单元为一个 SPO 三元组 < 主体（Subject，谓词（Predicate），客体（Object）>，用来表示一条客观世界的逻辑描述或客观事实。知识图谱正是由一些相互连接的实体和属性构成，换言之，一条三元组就对应了知识图谱中的一条知识，例如：“<梅西，是，足球运动员>”。多个三元组首尾相连，就构成了一个 RDF 图，如图12.6所示。

通过 RDF 的描述框架，可以将<梅西，是，足球运动员>的知识形式化表达为：“梅西 rdf:type 足球运动员”。RDF 也存在一些问题，首先就是无法区分概念和对象。概念和对象可以类比于面向对象程序语言中的 Class 和 Object，还是以图12.6为例，“梅西”描述的是一个对象，而“足球运动员”描述的是一个概念，而<足球运动员，是，运动员>描述的是 subClassOf 关系。如果不能正确区分概念和对象，不仅会引起理解上的困难，还会给后续的知识图谱融合以及下游任务的应用带

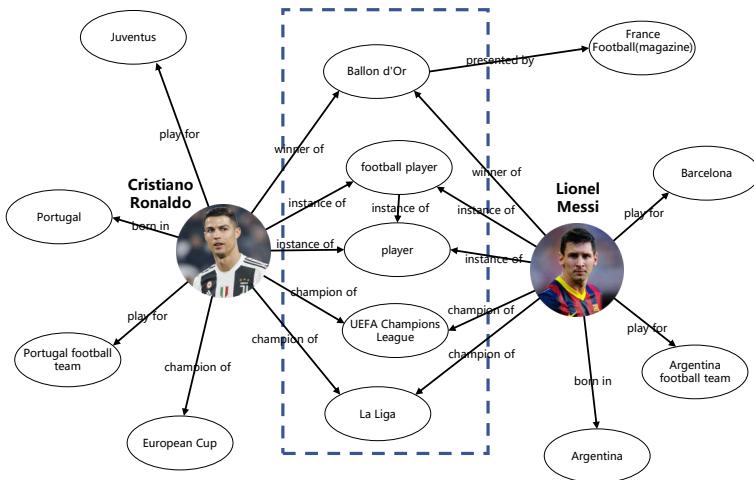


图 12.6 RDF 示例图

来障碍。

鉴于 RDF 表达能力有限，无法区分概念和对象，也无法描述类之间的关系和属性。W3C 提出了资源描述框架模式（Resource Description Framework Schema, RDFS）语言用来描述 RDF。RDFS 中定义了 Class 用来描述类，Domain 表示属性属于何种类别，Range 表示属性的取值，subClassOf 用作描述类的父类，subProperty 则描述属性的父属性。基于这些基础的表达构件，RDFS 可以实现一些简单的符号推理。例如，基于“梅西 rdf:type 足球运动员”，“足球运动员 rdfs:subClassOf 运动员”，可以推断出“梅西 rdf:type 运动员”。

### 3. OWL 本体语言

RDFS 是对 RDF 的一次成功扩展，但其表达能力仍然不足，无法完整表达复杂概念以及复杂概念间关系。为此，W3C 又开发了网络本体语言（Web Ontology Language, OWL）<sup>[645]</sup>。OWL 本体语言与 RDFS 类似，本质上是一些预定义的表达构件集合，都是用来描述 RDF 数据。但是，OWL 本体语言相比 RDFS 添加了额外的定义词汇，因此可以当做是 RDFS 的扩展版。在这里介绍几种比较典型的 OWL 本体语言表达构件。

OWL 本体语言引入了本体映射表达构件，使用 owl:equivalentClass 表示两个类是相同的，例如，可以定义“公司”和“企业”是相同的概念；使用 owl:equivalentProperty 表示两个属性是相同的，例如，可以定义“年龄”和“春秋”为含义相同的属性；使用 owl:sameAs 表示两个实体是同一实体，例如，“利昂内尔·梅西”和“梅西”指代同一个足球运动员。

OWL 本体语言中也引入了关于属性的复杂描述词汇，使用 owl:TransitiveProperty 表示属性具有传递性质，例如，“属于”是具有传递性的属性，若 A 属于 B，B 属于 C，那么 A 肯定属于 C；使

用 owl:SymmetricProperty 表示属性具有对称性，例如，“夫妻关系”是具有对称性的属性，若 A 与 B 是夫妻关系，那么 B 与 A 是夫妻关系；使用 owl:inverseOf 定义两个属性的相反关系，例如，“咨询”和“指导”是相反关系，若 A 咨询 B，则 B 指导 A。类似地，还可以定义属性的等价性、唯一性和属性间的互逆性，甚至可以约束属性满足一定的函数约束。

OWL 本地语言中还引入了 owl:unionOf、owl:intersectionOf 和 owl:complementOf 等布尔算子，分别表示集合的并集、交集和补集的运算。更多的 OWL 本体语言表达构件和特性请参考 W3C 官方文档。

## 12.2.2 知识图谱的向量表示

基于符号表示的大规模知识图谱在实际应用中面临两个问题：(1) 知识表达能力差，因为知识图谱中的实体一般符合长尾分布，许多实体仅有少数关系相连，对于这些稀疏的实体与关系，很难做到充分、完整的知识表达；(2) 计算效率低，基于图结构的知识表示方便人类理解，但是将其应用于下游任务时需要设计相应的图算法，这些算法计算复杂度往往较高，使得大规模知识图谱的应用门槛大大提高。

随着深度学习的广泛应用，越来越多的研究人员开始关注知识表示学习（Knowledge Representation Learning），即如何构建高质量的向量表示。知识表示学习通过将三元组中的语义信息投影到稠密的低维向量空间，构造实体和关系的分布式表示向量。这种表示向量单独地看每一维度并没有明确含义，但是综合各维度形成的向量却能够表示对象的语义信息。这种知识表示方法相对符号式表示有如下几个优点：(1) 知识表达能力强，分布式向量可以更好地建模对象之间的关系，语义相似的对象往往其表示向量也更接近，可以缓解长尾分布的知识表达问题；(2) 计算效率更高，对于计算机来说，已经提前计算好的蕴含语义知识的低维数值向量显然比复杂的知识图谱更高效；(3) 适用于深度学习算法，通过将知识映射到语义空间中，使得不同来源的知识可以很方便地互相融合。

研究人们先后提出了多种模型学习知识库中实体和关系的表示向量，其中最常见的是基于距离度量的知识表示学习方法。基于距离度量的方法表示通过计算实体之间的距离来评价事实三元组的置信度。距离模型（Distance Model）又可称为平移模型（Translational Model），该类模型将知识图谱中的每个关系看作从头实体向量到尾实体向量的一个平移变换。通过最小化平移转化的误差，将知识图谱中的实体和关系类型映射到低维空间。

### 1. 知识图谱向量表示学习 TransE 算法

本书第 4 章介绍了 CBOW 和 Skip-Gram 等分布式单词向量表示<sup>[615]</sup>，在其研究中发现词向量空间存在平移不变现象，例如：

$$\mathbf{v}(\text{man}) - \mathbf{v}(\text{woman}) \approx \mathbf{v}(\text{king}) - \mathbf{v}(\text{queen}) \quad (12.2)$$

这里  $v(w)$  表示单词 w 的词向量。换言之，词向量捕获到了 man 和 woman、king 和 queen 之间的近似关系。

受该现象的启发，文献 [646] 提出了 TransE 模型用于学习实体和关系的向量表示。TransE 模型将关系看做是表示空间中的平移(Translation)。如图12.7所示，在表示空间中，对于三元组  $\langle h, l, t \rangle$ ，头实体 h 的向量加上关系 r 的向量等于尾实体 t 的向量。

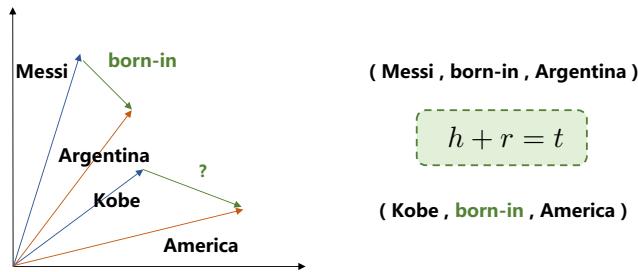


图 12.7 TransE 模型<sup>[615]</sup>

对于给定的由三元组集合  $S$  组成的知识图谱，其中三元组  $\langle h, l, t \rangle$  中  $h$  和  $t$  是头实体和尾实体， $l$  是关系， $h, \ell, t \in \mathbb{R}^k$  是对应的向量表示。根据平移假设，对于图谱中存在的三元组应该符合  $h + \ell \approx t$ 。为了保证不同三元组之间的区分度，TransE 除了知识库中存在的三元组  $(h + \ell, t)$  之外，还构造了知识库中不存在的三元组  $(h' + \ell, t')$ ，采用合页损失 (Hinge Loss) 作为模型的损失函数：

$$\mathcal{L} = \sum_{\langle h, \ell, t \rangle \in S} \sum_{\langle h', \ell, t' \rangle \in S'_{\langle h, \ell, t \rangle}} [\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+ \quad (12.3)$$

其中  $\gamma$  为正负例的得分间隔距离， $S$  为正例集合， $S'$  为负例集合。为了选取有代表性的错误三元组，TransE 算法将  $S$  中每个三元组的头实体、关系和尾实体其中之一随机替换成其他实体或关系来得到  $S'$ ，即：

$$S'_{\langle h, \ell, t \rangle} = \{ \langle h', \ell, t \rangle \mid h' \in \mathcal{E} \} \cup \{ \langle h, \ell, t' \rangle \mid t' \in \mathcal{E} \} \cup \{ \langle h, \ell', t \rangle \mid \ell' \in \mathcal{R} \} \quad (12.4)$$

TransE 算法的优化过程如算法12.1所示，首先，将所有实体和关系的嵌入向量随即初始化。在每次迭代开始之前，会对实体的向量进行归一化操作。之后，在每次批量训练时，从训练集中随机采样一些三元组。对于每个三元组，会构造一个负例三元组，用于计算12.3损失函数。根据损失函数产生的梯度，对实体和关系的表示向量进行更新。迭代上述过程，直至算法在验证集上的效果收敛。

**代码 12.1: TransE 模型训练算法**

**输入:** 训练集合  $S = \langle h + \ell, t \rangle$ , 实体集合  $E$ , 关系集合  $L$ , 超参数间隔  $\gamma$ , 表示向量维度  $k$

初始化  $\ell \leftarrow \text{uniform} \left( -\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}} \right)$  对每个  $\ell \in L$ ;

$\ell \leftarrow \ell / \|\ell\|$  对每个关系  $\ell \in L$ ;

$e \leftarrow \text{uniform} \left( -\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}} \right)$  对每个实体  $e \in E$ ;

**repeat**

$e \leftarrow e / \|e\|$  对每个实体  $e \in E$ ;

$S_{\text{batch}} \leftarrow \text{sample}(S, b)$ ; // 采样大小为  $b$  的批次样本;

$T_{\text{batch}} \leftarrow \emptyset$ ; // 初始化三元组集合;

**for**  $\langle h, \ell, t \rangle \in S_{\text{batch}}$  **do**

$\langle h', \ell, t' \rangle \leftarrow \text{sample} \left( S'_{\langle h, \ell, t \rangle} \right)$ ; // 构造冲突三元组;

$T_{\text{batch}} \leftarrow T_{\text{batch}} \cup \{(\langle h, \ell, t \rangle, \langle h', \ell, t' \rangle)\}$ ;

**end**

根据以下损失函数更新表示向量

$\sum_{(\langle h, \ell, t \rangle, \langle h', \ell, t' \rangle) \in T_{\text{batch}}} \nabla [\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+$ ;

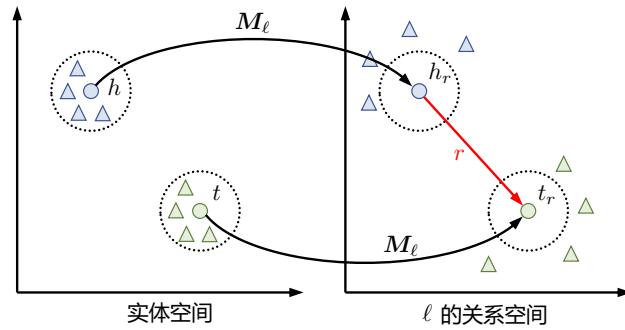
**until** 验证集效果收敛;

## 2. 知识图谱向量表示学习 TransR 算法

TransE 算法存在不能正确建模知识图谱中存在的一对多、多对一以及多对多的复杂关系的问题。例如，知识图谱中存在 $\langle$ 曹操，出生于，东汉末年 $\rangle$ 和 $\langle$ 刘备，出生于，东汉末年 $\rangle$ 两个事实，根据通过 TransE 算法所训练出来的“曹操”和“刘备”的表示很相似，无法区分二者。又比如“曹操”在不同语境下可能是“军事家”，也可能是“诗人”。此外，TransE 算法将实体和关系映射到相同的空间中，但是关系和实体是完全不同的对象，共同的语义空间可能不足以表示它们。因此，这些复杂的场景下，TransE 算法过于理想化的平移假设显然不能满足需求。

针对上述问题，文献 [647] 提出了 TransR 算法，为每种关系  $\ell$  定义了单独地语义空间。给定三元组  $\langle h, \ell, t \rangle$ ，头尾实体表示分别是  $\mathbf{h}, \mathbf{t} \in \mathbb{R}^k$ ，关系表示向量  $\ell \in \mathbb{R}^d$ ，实体表示和关系表示的维度不需要相等。TransR 算法通过针对每种关系的不同的映射矩阵  $M_\ell$  将实体映射到关系空间中，要求实体在关系空间中满足平移关系即可。如图12.8所示，特定关系投影可以使实际具有该关系的头/尾实体（图中表示为彩色圆圈）彼此靠近，也可以使得不具有该关系的实体（图中表示为彩色三角形）则彼此远离。TransR 本质上是放宽了 TransE 的  $\mathbf{h} + \ell = \mathbf{t}$  假设，将原来在一个空间中满足的平移关系，改为在不同的关系空间中满足。

具体来说，给定三元组  $\langle h, \ell, t \rangle$ ，TransR 算法通过关系  $\ell$  对应的映射矩阵  $M_\ell$ ，将头实体和

图 12.8 TransR 模型<sup>[647]</sup>

尾实体映射到关系空间中，得到实体表示  $h_\ell$  和  $t_\ell$ :

$$h_\ell = hM_\ell, \quad t_\ell = tM_\ell \quad (12.5)$$

在得到对应关系空间的表示  $h_\ell$  和  $t_\ell$  后，就可以使用与 TransE 类似的评分函数进行训练：

$$f_\ell(h, t) = \|h_\ell + \ell - t_\ell\|_2^2 \quad (12.6)$$

在实现时，TransR 算法还对  $h$ ,  $\ell$ ,  $t$  的表示向量以及关系空间中映射的表示向量的范数添加约束。

$$\forall h, \ell, t, \|h\|_2 \leq 1, \|\ell\|_2 \leq 1, \|t\|_2 \leq 1, \|hM_\ell\|_2 \leq 1, \|tM_\ell\|_2 \leq 1 \quad (12.7)$$

### 12.2.3 基于表的知识谱图谱存储

尽管知识图谱是用图的形式描述，但图数据库并不是存储知识图谱的唯一方案。在工业界，很多成熟的数据库都是基于关系模型，知识图谱的数据可以通过一定的设计，从而可以利用关系型数据库存储。基于关系型数据库的知识图谱存储方案主要可以分为四种：基于三元组的知识图谱存储、基于属性表的知识图谱存储、基于垂直表的知识图谱存储以及基于全索引的知识图谱存储。

#### 1. 基于三元组的知识图谱存储

利用关系型数据存储知识图谱最简单的办法就是构建一个由主体、谓词和客体三列构成的表，将知识图谱中的每个 SPO 三元组看做该表中的一条记录，直接存储到关系型数据库中，如表12.1所示。

当用户输入一个查询请求时，需要将其转换为对应的查询 SQL 语句。但是稍微复杂一点的查询，需要大量的自连接（Self-Join）操作。例如，用户需要查询“获得金球奖的足球运动员出生地”，这需要将整张表复制三次，然后依次扫描，基于查询结果再做连接（Join）操作，查询代价过于昂贵。

表 12.1 基于三元组的知识图谱存储

主体	谓词	客体
利昂内尔·梅西	出生地	阿根廷
利昂内尔·梅西	isA	足球运动员
利昂内尔·梅西	生日	1987-06-24
利昂内尔·梅西	效力	巴黎圣日耳曼足球俱乐部
利昂内尔·梅西	荣誉	金球奖
杰西·林加德	出生地	英格兰
杰西·林加德	isA	足球运动员
杰西·林加德	效力	曼彻斯特联足球俱乐部
科比·布莱恩特	出生地	美国
科比·布莱恩特	isA	篮球运动员
科比·布莱恩特	效力	洛杉矶湖人队
科比·布莱恩特	荣誉	NBA 总决赛
足球运动员	subClassOf	运动员
篮球运动员	subClassOf	运动员
金球奖	英文名	Ballon d'Or
NBA 总决赛 MVP	英文名	NBA Finals MVP

## 2. 基于属性表的知识图谱存储

为了解决基于三元组的知识图谱存储查询效率过低问题，还可以按照属性表（Property Tables）<sup>[648]</sup>的方式存储知识图谱。其存储方案仍然是基于关系型数据库，不同之处在于，属性表根据实体类型对三元组进行分类，不同类别的三元组存储于不同的表中。例如，针对上一节中的三元组存储表案例转换为属性表后，其结构如表12.2所示。将运动员、类型、生日、出生地、荣誉等信息合并到一张表中进行保存。对于上例中用户查询“获得金球奖的足球运动员出生地”，只需要检索“荣誉”对应的属性表就可以完成。

表 12.2 基于属性表的知识图谱存储示例

主体	isA	生日	出生地	效力	荣誉
利昂内尔·梅西	足球运动员	1987-06-24	阿根廷	巴黎圣日耳曼足球俱乐部	金球奖
杰西·林加德	足球运动员		英格兰	曼彻斯特联足球俱乐部	
科比·布莱恩特	篮球运动员		美国	洛杉矶湖人队	NBA 总决赛 MVP

主体	英文名
金球奖	Ballon d'Or
NBA 总决赛 MVP	NBA Finals MVP

主体	subClassOf
足球运动员	运动员
篮球运动员	运动员

虽然属性表的效率相较于三元组表在一定程度上有所提升，但是知识图谱的数据往往具备稀疏性，因此基于属性表的存储方案可能会包含大量空值，造成存储空间的极大浪费。此外，属性表的存储方案还依赖于人的设计，哪些属性需要合并在一张表中，需要根据知识图谱需求和实际关系联合考虑。

### 3. 基于垂直表的知识图谱存储

除了按照实体对三元组表格分类外，还可以选择按照谓词构造表格，这种方式称为基于垂直表（Vertical Tables）<sup>[648]</sup>的存储方式。将不同的谓词对应不同的表格，这种方法不仅可以避免查询过程中大量的自连接操作，还不会在数据库中产生大量的空值。表12.2中三元组存储表案例被转换为垂直表后，其结构如表12.3所示。根据谓词构建了 isA、出生地、效力等多个表。

表 12.3 基于垂直表的知识图谱存储

isA

主体	客体
利昂内尔·梅西	足球运动员
杰西·林加德	足球运动员
科比·布莱恩特	篮球运动员

出生地

主体	客体
利昂内尔·梅西	阿根廷
科比·布莱恩特	美国
杰西·林加德	英格兰

效力

主体	客体
利昂内尔·梅西	巴黎圣日耳曼足球俱乐部
杰西·林加德	曼彻斯特联足球俱乐部
科比·布莱恩特	洛杉矶湖人队

荣誉

主体	客体
利昂内尔·梅西	金球奖
科比·布莱恩特	NBA 总决赛 MVP

生日

主体	客体
利昂内尔·梅西	1987-06-24

英文名

主体	客体
金球奖	Ballon d'Or
NBA 总决赛 MVP	NBA Finals MVP

subClassOf

主体	客体
足球运动员	运动员
篮球运动员	运动员

这种方法缺点是对于具备多个属性的实体，往往需要对多个表进行插入操作，操作效率低。对于属性不确定的查询，垂直表也不能很好的处理，例如，查询“科比·布莱恩特和洛杉矶湖人队的关系”，就需要检索所有表才能得到结果，这类查询的检索性能较差。

### 4. 基于全索引的知识图谱存储

基于三元组存储的另一种改进优化方法是基于全索引的存储方式，即通过对关系型数据库建立多种索引的方式，达到查询效率优化的目的。全索引表的优化方法，首先会将 SPO 三元组里的每个元素集合映射为一个数字 ID 集合，通过存储 ID 而不是字符串可以节省大量存储空间，如

表12.4所示。然后，对三元组中主体、谓词、客体的各种排列情况都分别建立索引，这种组合一共有六种：SPO、POS、PSO、OSP、OPS、SOP，如图12.9所示。无论是基于谓词查询主体和客体，还是基于客体查询谓词和主体，六种索引都可以满足需求。全索引的方法应对简单的查询效率高，但是其索引的更新与维护同样代价高昂。同时，面对复杂一些的查询，全索引仍然不可避免要做多次自连接操作，效率低下。

表 12.4 基于全索引的知识图谱存储

Value	ID	Value	ID
利昂内尔·梅西	0	1987-06-24	12
杰西·林加德	1	巴黎圣日耳曼足球俱乐部	13
科比·布莱恩特	2	金球奖	14
出生地	3	英格兰	15
isA	4	曼彻斯特联足球俱乐部	16
生日	5	美国	17
效力	6	篮球运动员	18
荣誉	7	洛杉矶湖人队	19
subClassof	8	NBA 总决赛 MVP	20
英文名	9	运动员	21
阿根廷	10	Ballon d'Or	22
足球运动员	11	NBA Finals MVP	23

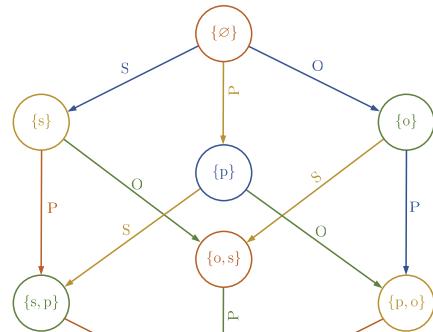


图 12.9 六重索引

## 12.2.4 基于图的知识谱图谱存储

在图数据库技术还不成熟的初期，关系型数据库因其成熟的技术架构，成为知识图谱存储的首选方案。但是关系型数据库却并不善于处理“关系”，在关系型数据库上进行图数据上的多跳查询，会产生大量的连接操作，其计算复杂度随着查询的跳数增多呈指数级增长。以表12.1为例，如果要查询“获得 NBA Finals MVP 荣誉的球员效力于哪个俱乐部”，关系型数据库至少要三次表连接操作。知识图谱除了需要提供高效的关联推理，还要能刻画丰富的语义表达能力，诸如传递关系、自反关系、对称关系和函数关系等等。

图数据库则充分利用图的结构对数据进行建模，将一张图对应到一个邻接列表，再基于这个邻接列表建立索引，优化关联查询。例如，查询“获得 NBA Finals MVP 荣誉的球员效力于哪个俱乐部”，只要依次对“NBA Finals MVP”、“NBA 总决赛 MVP”、“科比·布莱恩特”三个节点的邻居节点进行检索即可。此外，在图数据库中，关系是被显式描述和刻画的，属性也可以单独定义，这就极大地增加了数据建模的灵活性。随着相关技术和工具的逐渐成熟，图数据库已逐渐成为知识图谱存储和查询引擎搭建的基础设施。

## 1. 图数据库使用案例

在本章第12.2.1节知识图谱表示部分，我们介绍了两种图模型：属性图模型和RDF图模型。属性图模型是图数据库Neo4J所引导的一种数据模型，使用Cypher<sup>[649]</sup>作为查询语言。因为其性能较好，已经在工业界得到广泛应用。RDF图模型严格地说并不是一个存储模型，而是一种规范定义的数据交换的格式标准，使用SPARQL查询语言<sup>[650]</sup>。其中，属性图数据库善于处理关联查询，而RDF图模型则提供更多的关联推理能力。本节以Neo4J为例，对基于图数据库的知识图谱的存储和查询进行简要说明。

在Neo4J数据库中，数据以节点（Node）和边（Edge）进行组织。节点可以代表知识图谱中的实体，边可以代表实体间的关系。需要特别注意的是，关系具有类型和方向性。在节点上可以添加标签（Label）以及键值对（Key-Value）来表示实体具备的属性（Property）。以2022世界杯为例，其属性图模型如图12.10所示，每一年的世界杯对应一个WorldCup节点，每场比赛对应一个Match节点，球员节点为Player。Country节点通过HOSTED\_BY关系连接到WorldCup节点，同时每个Country都会指定一个球员小队Squad，代表他们参加世界杯比赛。对于Player作为首发或替补参加的每场比赛，都连接到Appearance节点，如果球员取得分数，则Appearance将连接到该得分节点Goal。

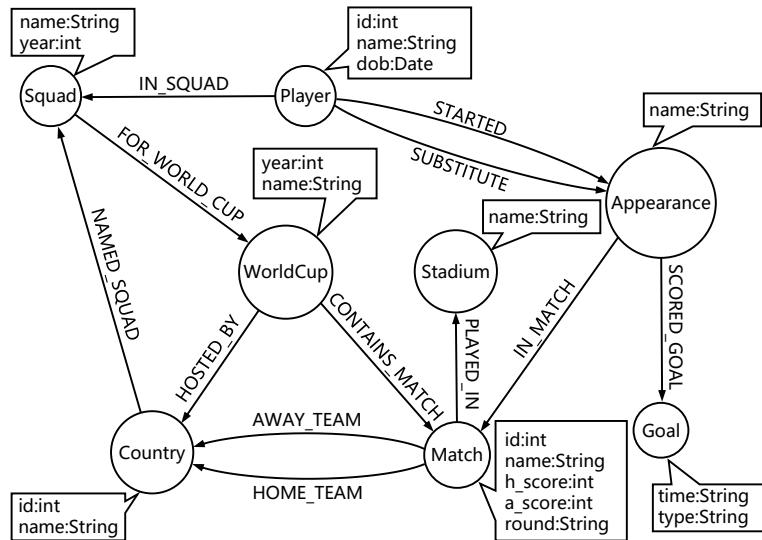


图 12.10 Neo4J 的属性图模型样例

Neo4J采用自己独有的查询语言Cypher，这是一种描述性语言，用户只需要声明“查什么”，而不用关心“怎么查”。Cypher语言与SQL很相似，但是Cypher关键字不区分大小写，属性值、标签、关系类型和变量则区分大小写。Cypher中最常用到的关键词是：

- match: 相当于 SQL 中的 select, 用来说明检索匹配的图模式。
- where: 用来限制节点或者边中的属性值, 从而过滤掉不需要的返回值。
- return: 返回节点或者关系。

Neo4J 支持知识图谱的路径查询, 以图12.10的图结构为例, 查询“世界杯的举办国”, 返回世界杯节点和举办国节点之间的路径, 可以使用如下 Cypher 语句:

```
MATCH path = (wc:WorldCup)-[:HOSTED_BY]->(country)
RETURN path
```

返回结果如图所示12.11。

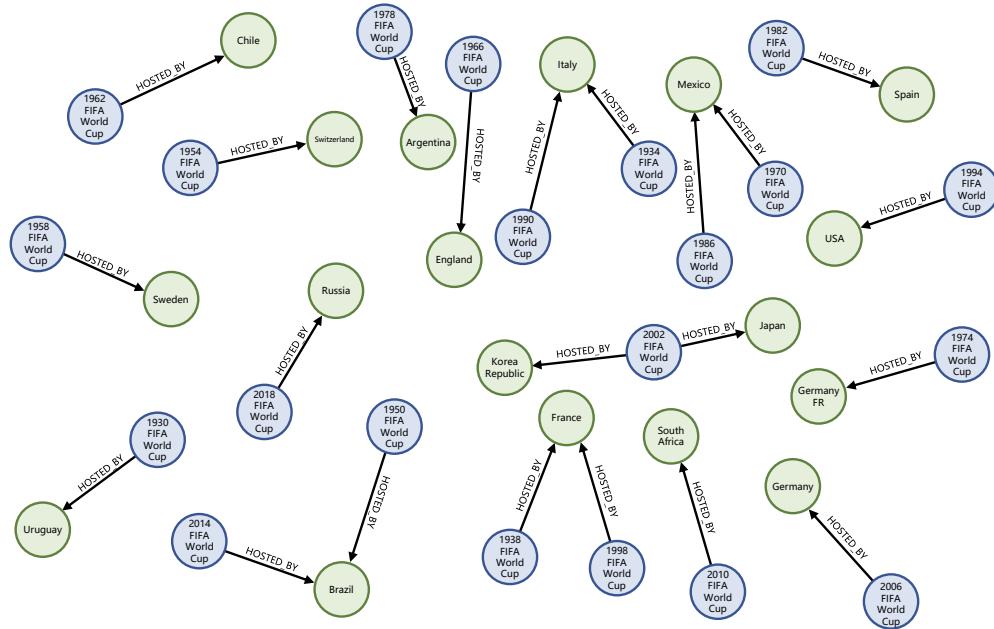


图 12.11 Neo4J 的路径查询样例

如果需要直接返回表格数据, 例如查询“举办过世界杯的国家”, 可以采用类似 SQL 的语法组织查询:

```
MATCH (host:Country)<-[HOSTED_BY]-(wc)
WITH wc, host ORDER BY wc.year
WITH host, count(*) AS times, collect(wc.year) AS years
WHERE times > 1
RETURN host.name, times, years
```

其结果如表12.5所示。

表 12.5 Neo4J 查询结果返回样例

host.name	times	years
Mexico	2	[1970,1986]
France	2	[1938,1998]
Brazil	2	[1950,2014]
Italy	2	[1934,1990]

知识图谱的存储需要综合考量查询性能、扩展性和存储成本等多种因素，使用者需要根据具体的应用场景和数据规模选择合适的存储方式。基于属性图的图数据库、基于RDF的存储系统和关系型数据库的特点可以总结如表12.6所示。对于简单查询，传统的关系型数据库在性能方面更有优势，但是在处理多跳查询和建模复杂关系语义的任务上，基于图的知识图谱存储模式更加适配。属性图的图分析性能更强，比如涉及到子图匹配、图结构学习等任务。而RDF存储系统的关联推理能力更强，并且由于RDF作为一种数据交换的格式标准，具有更好的理论基础。

表 12.6 知识图谱不同存储方案比较

	基于属性图的图数据库	RDF 存储系统	关系型数据库
数据模型	属性图	RDF 三元组	关系数据模型
查询语言	Cypher、Gremlin <sup>[651]</sup>	SPARQL	SQL
应用场景	多为工业界	多为学术界	工业界和学术界均有使用
其他特点	图遍历效率高；图的全局操作效率低	有标准的推理引擎；易于数据发布	简单查询销量高；多跳查询会产生连接操作，效率低

## 12.3 知识图谱获取与构建

构建知识图谱最简单直接的方式是使用人工直接构建，但是这种方法得到的知识图谱规模和知识的覆盖面往往有限，而且不同专家的知识认知也不尽相同，难以用统一的标准精准地刻画专家知识，使得基于人工构建的知识图谱具有高度的不确定性、不准确性。如何构建大规模、高质量的知识图谱，实现海量知识的准确抽取和有效聚合，是知识图谱领域最重要的问题之一。

知识图谱的数据来源是多种多样的，包括结构化数据、半结构化数据以及非结构化数据等。对于已有的结构化数据，可以通过一定的规则来提取其中的知识，比如关系型数据库，其行（Row）可能代表知识图谱中的实例（Instance），列（Column）可能表示属性（Property），单元（Cell）对应属性值，外键（Foreign Key）则可以表示指代关系。对于半结构化数据，比如维基百科，可以制定解析规则，将页面上的字段映射为知识图谱中的实体或者属性。如图12.12所示，维基百科中IBM

词条的 InfoBox 区域包含了大量可以直接转化为知识图谱的信息。对于非结构化数据，即纯文本数据，首先需要从文本中挖掘出目标知识图谱所需的实体，可以应用命名实体抽取等技术；基于已知实体信息，还需要从文本中识别出它们之间的关系，这对应了 SPO 三元组中的谓词关系，需要使用关系抽取技术；除了实体和关系信息，还需要抽取属性信息，需要使用属性补全技术；更复杂的场景下需要对事件信息进行抽取，可以看作一组三元组的联合抽取过程。



图 12.12 基于维基百科中半结构化数据构建知识图谱

考虑到自然语言的歧义性，比如“苹果”，在某些场景下指的是“苹果公司”，有的时候则表示的是水果，因此还需要引入实体链接技术消除不同文本中实体指称的一词多义问题。此外，自然语言具有表达的多样性，比如“复旦大学”，有的表达中使用“复旦”，这使得在不同的知识图谱中相同的实体会有不同的实体指代，因此需要引入实体对齐技术解决本体或实例的异构问题。知识图谱构建所依赖的实体抽取、关系抽取等方法在本书第 7 章信息抽取部分有详细介绍，这里我们就不赘述，本节将会依次介绍属性补全、实体链接技术和实体对齐常见方法。

### 12.3.1 属性补全

现实世界中的实体并不总是单一存在，往往伴随许多属性对其进行修饰，例如，复旦大学有在校师生数、校庆日等属性。准确地掌握实体的属性信息，能够多方位地描述实体特征，提升知识的建模能力。属性补全的目的就是自动地从文本中提取出实体所具备的属性，特别是在电子商务等垂直领域下，属性补全有着重要的应用，如图12.13所示，若能自动从产品描述中提取出产品的属性，将能更好的描述产品特征，优化产品搜索和推荐效果。



图 12.13 电商场景下的产品属性示例

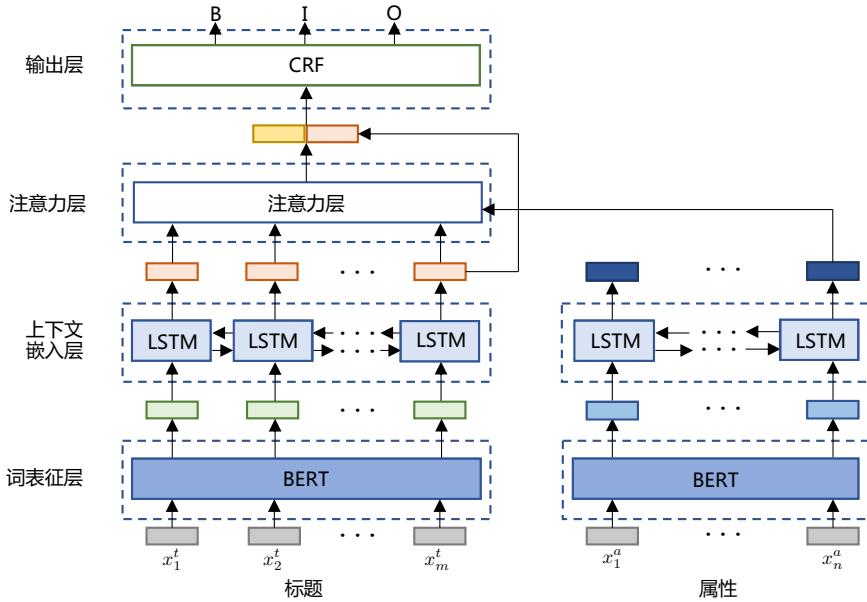
属性补全方法通常采用抽取式方法，类似于命名实体识别，将属性补全转为序列标注任务，即对句子中的每个字进行分类，在此基础上基于字的分类结果解析得到属性值。但如果简单地将其等价于普通的序列标注任务，将属性类型等价于实体类型，会带来两个问题：(1) 普通的命名实体识别模型一般只建模数个实体类型，而领域知识图谱可能包含成千上万的属性种类，且这些属性往往符合二八法则（Power Laws），直接应用普通的序列标注模型，由于标签数量巨大，会造成性能的显著下降；(2) 知识图谱的构建需要保证拓展性，保留属性集合扩展的可能，而当前实体识别模型一般是基于封闭世界假设，无法识别新的实体种类，若直接应用，同样无法识别未标注过的新属性。

为了解决这个问题，文献 [652] 提出了一种基于属性理解的属性抽取方法 AVEPT。AVEPT 算法将属性理解问题转化为抽取式阅读理解任务，待抽取的文本作为上下文，目标属性信息作为问句，属性值则作为答案，是模型要抽取出来的目标。由此，将抽取结果和属性类型进行解偶，并且输入的属性信息可以帮助模型捕获属性的语义信息。因此，该方法在取得较好的效果同时，还保留了模型对新属性识别的拓展性。AVEPT 算法神经网络结构如图12.14所示。AVEPT 模型可以分为简单分为四层：词表示层、上下文编码层、注意力层、输出层。

具体来说，待抽取文本为  $T = \{x_1^t, x_2^t, \dots, x_m^t\}$ ，长度为  $m$ 。目标属性记为  $A = \{x_1^a, x_2^a, \dots, x_n^a\}$ ，长度为  $n$ ，模型的目标输出序列为  $y = \{y_1, y_2, \dots, y_m\}$ ， $y_i \in \{B, I, O\}$ ， $B$  和  $I$  分别表示属性值的开始和中间， $O$  表示非目标属性值。

词表示层（Word Representation Layer），AVEPT 使用预训练语言模型 BERT 将待抽取文本  $T$  和属性  $A$  里的词分别映射为词向量表示，即  $\{\mathbf{w}_1^t, \mathbf{w}_2^t, \dots, \mathbf{w}_m^t\}$  和  $\{\mathbf{w}_1^a, \mathbf{w}_2^a, \dots, \mathbf{w}_n^a\}$ 。

上下文编码层（Contextual Embedding Layer），为了建模长距离的上下文信息，AVEPT 选用两个不同的双向长短时记忆网路（BiLSTM）分别编码待抽取文本和目标属性的词向量表示。将待抽

图 12.14 AVEPT 算法神经网络结构图<sup>[652]</sup>

取文本对应的 BiLSTM 的隐状态向量作为其上下文表示, 即  $\mathbf{H}^t = \{\mathbf{h}_1^t, \mathbf{h}_2^t, \dots, \mathbf{h}_m^t\}$ , 其中每个词的隐状态向量计算方式为:

$$\mathbf{h}_i^t = [\overrightarrow{\mathbf{h}}_i^t; \overleftarrow{\mathbf{h}}_i^t] = \text{BiLSTM}(\overrightarrow{\mathbf{h}}_{i+1}^t, \overleftarrow{\mathbf{h}}_{i-1}^t, \mathbf{w}_i^t) \quad (12.8)$$

与待抽取文本的上下文表示不同, 目标属性只使用 BiLSTM 最后时间步的隐状态向量作为表示向量。

$$\mathbf{h}^a = [\overrightarrow{\mathbf{h}}_n^a; \overleftarrow{\mathbf{h}}_n^a] = \text{BiLSTM}(\overrightarrow{\mathbf{h}}_n^a, \overleftarrow{\mathbf{h}}_n^a, \mathbf{w}_n^a) \quad (12.9)$$

**注意力层 (Attention Layer)**, 不同于普通的自注意力方法捕获重要的词汇, AVEPT 考虑了属性和待抽取文本之间隐藏的语义交互。这种方法相比自注意力具有更好的泛化性, 即使是训练中没有见过的属性, 也能抽取对应属性值。具体来说, AVEPT 首先计算属性和待抽取文本中每个词的相似度, 得到注意力向量  $\mathbf{S} = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ , 向量之间的相似度计算采用余弦函数。

$$\alpha_i = \cos(\mathbf{h}_i^t, \mathbf{h}^a) \quad (12.10)$$

然后使用注意力向量对待抽取文本的上下文表示进行放缩, 即  $\mathbf{C} = \mathbf{S} \odot \mathbf{H}^t$ , 其中  $\odot$  表示按位乘。

**输出层 (Output Layer)**, 为了捕获邻近词的标签依赖关系, AVEPT 采用条件随机场 (CRF) 作为解码器。CRF 的输入  $\mathbf{M}$  为文本的上下文表示  $\mathbf{H}^t$  和属性放缩表示  $\mathbf{C}$  的拼接矩阵, 这样可以同

时保存上下文信息和上下文对于属性重要性的信息。

$$\mathbf{M} = [\mathbf{H}^t; \mathbf{C}] \quad (12.11)$$

使用 CRF 对标签进行解码，可以得到标签序列的联合概率密度：

$$\Pr(y | T; \psi) \propto \prod_{i=1}^m \exp \left( \sum_{k=1}^K \psi_k f_k(y_{i-1}, y_i, \mathbf{M}_i) \right) \quad (12.12)$$

其中  $\psi_k$  表示相应权重， $f_k$  为特征函数， $K$  为特征数量。模型最终输出的标签序列对应概率值最大的标签路径：

$$y^* = \operatorname{argmax}_y \Pr(y | u; \psi) \quad (12.13)$$

AVEPT 的训练损失使用最大条件似然估计：

$$\mathcal{L}(\psi) = \sum_{i=1}^N \Pr(y_i | u_i; \psi) \quad (12.14)$$

### 12.3.2 实体链接

随着信息抽取和知识图谱的发展，涌现出很多大规模、高质量且机器可读的开放知识图谱，包括 YAGO<sup>[653]</sup>，DBpedia<sup>[654]</sup>，Freebase<sup>[655]</sup>，和 Probable<sup>[656]</sup> 等。这些知识图谱包含数千万命名实体和数十亿命名实体间关系。实体链接（Entity Linking）目标是将文本中的实体指代和它们在知识图谱中的对应实体进行对应。对已有知识图谱扩充的时候，也需要实体链接任务将文本中的实体指代与其知识图谱中真正对应的实体完成映射。除此之外，实体链接任务也是很多下游应用的预处理工作，如问答、关系抽取、内容分析等。

实体链接的任务的难点主要有两个：(1) 实体的歧义性，如图12.15所示，“MJ”可能表示篮球运动员“Michael Jordan”，也可能是美国歌手“Mj Rodriguez”；(2) 实体表达的多样性，即一个命名实体有不同的表示形式，如全称、缩写、别名等等，例如：“NYC”和“Big Apple”是命名实体“New York City”的缩写和昵称。歧义性和多样性的消除需要依据上下文语义信息以及实体属性和关系信息。

基于传统机器学习的实体链接算法，一般分为两步，首先提取人工设计的特征，如实体的 TF-IDF 值、局部上下文兼容性和文档级的实体引用全局一致性的特征等等。然后，将特征输入给实体排序模型，进行最后的链接预测。这种方法有两个不足：一是需要进行大量细致而繁琐的特征工程；二是由于特征设计过程对特定知识库或者领域知识强烈依赖，很难将已有的实体链接方法推广到其他知识库或者领域。随着深度学习研究的发展，很多深度学习方法也应用于实体链接算法。深度学习算法通常将实体链接任务分解为实体识别和实体消歧两部分，对它们分别建模，但是这样会导致模型错误传递。文献 [657] 提出一种端到端的实体链接方法 ENEL（End-to-End Neural

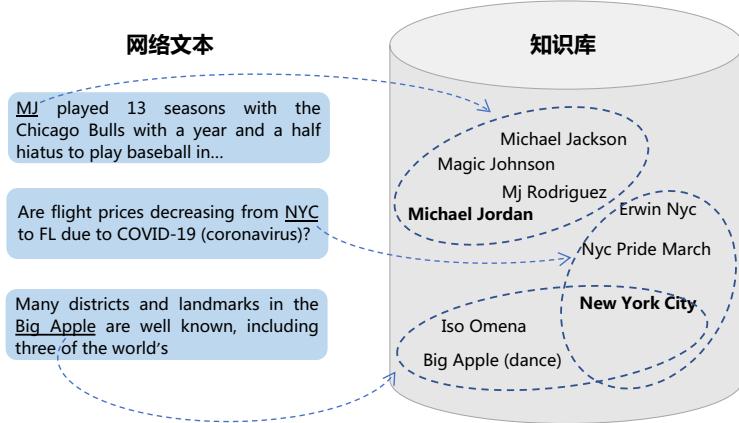


图 12.15 实体链接任务示意图

Entity Linking), 同时建模实体识别和实体消歧任务。主要思想是考虑所有可能的文本片段作为潜在的实体，并学习其候选实体的上下文相似性分数。

实体链接任务输入为文档或句子  $D = \{w_1, \dots, w_n\}$ , 其中的每个词都来源于词典  $w_k \in \mathcal{W}$ 。输出为实体提及-实体对的列表  $\{(m_i, e_i)\}_{i \in [1, T]}$ , 其中每个实体提及为输入文档的子序列  $m = w_q, \dots, w_r$ , 每个实体则对应知识库里的一个实体条目  $e \in \mathcal{E}$ 。需要注意的是, ENEL 算法只建模在知识库中存在实体对应的实体提及。ENEL 模型结构图如12.16所示, 主要组成部分包括蕴含上下文的实体提及表示、实体嵌入向量和提及-实体的概念映射。

**单词和字符嵌入向量** (Word and Char Embeddings) 用于构词向量表示, 字符的嵌入向量是由一个 BiLSTM 建模得到, 输入为一个单词的所有字符。单词的嵌入向量是预训练好的词向量。最后句子中的单词嵌入向量由其字符向量和词向量拼接得到, 表示为  $\{v_k\}_{k \in [1, n]}$ , 这对应了图12.16中的单词-字母嵌入拼接。

**实体提及表示** (Mention Representation) 用于建模上下文信息, ENEL 算法中采用用 BiLSTM 进行建模。

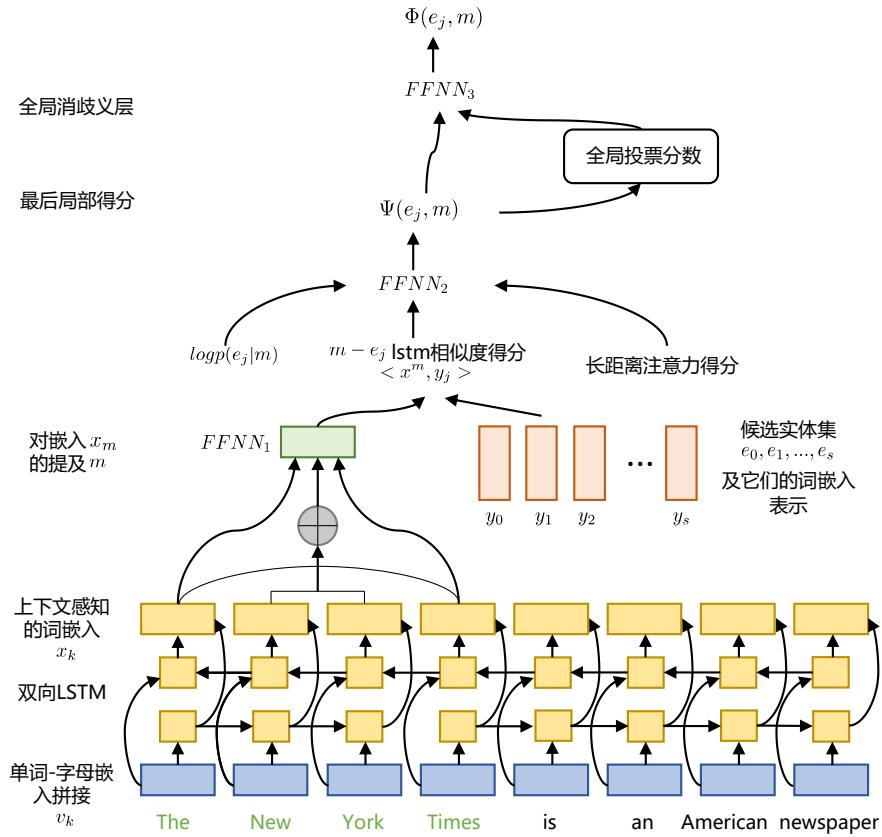
$$x_k = BiLSTM(v_k) \quad (12.15)$$

对每个可能的实体提及  $m = w_q, \dots, w_r$ , 其软头嵌入向量 (Soft Head Embedding) 为:

$$\alpha_k = \langle \mathbf{w}_\alpha, x_k \rangle \quad (12.16)$$

$$a_k^m = \frac{\exp(\alpha_k)}{\sum_{t=q}^r \exp(\alpha_t)} \quad (12.17)$$

$$\hat{x}^m = \sum_{k=q}^r a_k^m \cdot v_k \quad (12.18)$$

图 12.16 ENEL 模型结构图<sup>[657]</sup>

实体提及的最终表示  $g^m$  为序列的首尾词的上下文表示向量和其软头嵌入向量的拼接：

$$g^m = [x_q; x_r; \hat{x}^m] \quad (12.19)$$

**实体嵌入向量** (Entity Embedding) 采用文献 [658] 中基于预训练得到的实体表示向量，用  $y_e$  来表示。所采用的方法是将单词和实体名映射到同一个空间中，首先利用单词的向量表示初始化实体表示，再根据实体和单词共现关系用于训练实体表示。

**候选实体词选择** (Candidate Selection) 用于选择可能的候选实体。ENEL 采用文献 [658] 给出的经验概率映射的方法，对每个可能的序列  $m$ ，ENEL 最多选择  $s$  个候选实体，候选实体集合用  $C(m)$  表示。

**实体-提及对得分** (Entity-Mention Score)，对每个可能的成为实体指代的序列  $m$ ，也即  $|C(m)| \geq 1$ ，对其计算局部得分。首先对实体候选集合里的每个实体与提及序列，通过点积计算相似度，再

结合实体候选集合计算中得到的候选概念对数，输入到一个全连接层得到最终得分。

$$\Psi(e_j, m) = \text{FFNN}_2([\log p(e_j | m); \langle x^m, y_j \rangle]) \quad (12.20)$$

为了进一步提升性能，ENEL 还可以引入全局依赖，使用上下文嵌入与候选实体做相似度计算，将得到的结果加入公式12.20中。

**全局消歧** (Global Disambiguation) 对局部得分高的实体-提及对全局的实体-提及得分进行投票。由于实体-提及得分只建模了一对的信息，为了引入多对实体-提及信息，ENEL 增加了全局消歧层。

$$V_G = \{(m, e) \mid m \in M, e \in C(m), \Psi(e, m) \geq \gamma'\} \quad (12.21)$$

$G(e_j, m)$  表示由实体嵌入与所有其他投票实体嵌入的平均值之间的余弦相似度， $\Phi(e_j, m)$  表示最终的全局得分。

$$V_G^m = \{e \mid (m', e) \in V_G \wedge m' \neq m\} \quad (12.22)$$

$$y_G^m = \sum_{e \in V_G^m} y_e \quad (12.23)$$

$$G(e_j, m) = \cos(y_{e_j}, y_G^m) \quad (12.24)$$

$$\Phi(e_j, m) = \text{FFNN}_3([\Psi(e_j, m); G(e_j, m)]) \quad (12.25)$$

结合实体-提及的局部得分和全局得分，可以得到最终的训练目标：

$$\theta^* = \arg \min_{\theta} \sum_{d \in D} \sum_{m \in M} \sum_{e \in C(m)} V(\Psi_\theta(e, m)) + V(\Phi_\theta(e, m)) \quad (12.26)$$

其中  $V$  用来强制正确对的分数与错误对的分数线性可分：

$$V(\Psi(e, m)) = \begin{cases} \max(0, \gamma - \Psi(e, m)), & \text{if } (e, m) \in \mathcal{G} \\ \max(0, \Psi(e, m)), & \text{其他} \end{cases} \quad (12.27)$$

### 12.3.3 实体对齐

知识图谱包含描述抽象知识的本体层和描述具体事实的实例层。本体层一般为抽象知识，如概念、属性、公理等。实例层则描述具体实体、实体间的关系，即事实数据。知识工程早期希望可以构建一个包含所有知识的统一的知识库，但是在实际情况下面临很多困难。首先就是不同领域的知识存在差异性，同时不同专家对知识定义的主观性，使得无法构建统一的本体知识。其次，知识本身会随着时间发生演变，在不同时间，使用者会构建不同的知识本体，对相同实体的

指代称呼也会发生变化，如图12.17所示。这种现象称之为本体和实例的异构性。

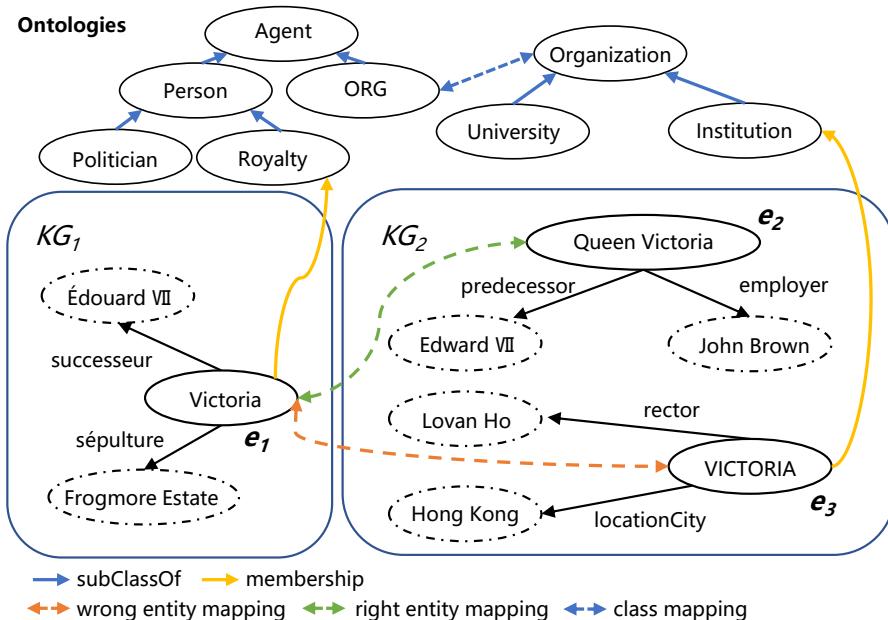


图 12.17 知识图谱融合样例图

针对知识图谱的异构性问题，研究人员们提出了知识融合技术，其目标就是为不同知识图谱的实例或者对象建立关联，可以高效合并多个不同来源的知识图谱。知识融合技术的两个基本任务是本体匹配和实体对齐，两个任务的目的类似，只是匹配对象不同，前者用来匹配本体层的抽象知识，后者用来匹配实例层中相同对象的不同实体指代。两者使用的技术有很多类似之处，鉴于知识图谱的实例规模通常远大于本体规模，任务难度更大，本节将重点介绍实体对齐任务。

实体对齐 (Entity Alignment) 也被称作实体匹配 (Entity Matching)，旨在发现多源知识图谱中等价的实体对。比如由几种不同的语言构建的知识库，实现跨语言知识对齐将帮助人们构建一个连贯的知识库，并帮助机器处理跨语言的实体关系的不同表达。早期的实体对齐研究一般基于实体属性的相似度计算。这种方法依赖于用户定义的规则来决定实体之间进行比较的属性。例如，实体的类别、实体的邻居类别等。因为不同的实体对可能需要不同的属性来进行比较，使得依赖用户定义规则的方法很容易出错，并且算法的迁移性较差，已有的规则难以在其他场景直接应用。

随着深度神经网络在自然语言处理领域的研究不断深入，基于表示学习的实体对齐方法逐渐成为目前的主流方法。本章第12.2.2节介绍了知识的向量表示，简单来说就是将三元组中的语义信息投影到稠密的低维向量空间，构造实体和关系的分布式表示向量。基于表示学习的实体对齐，核心思想就是在低维空间中用向量的距离来计算不同来源的实体之间的相似度。下面将介绍两类基

于表示学习的实体对齐方法：基于平移的方法和基于图神经网络的方法。

### 1. 基于平移的实体对齐方法

虽然基于平移的知识表示技术可以帮助提高单语言知识的完整性，但是该技术并不能直接解决跨语言知识的对齐问题。主要有以下几个方面的原因：(1) 跨语言平移比任何单语关系平移的范围都要大得多；(2) 实体和关系在不同语言之间的词汇表达不连贯；(3) 用于训练知识嵌入表示的已知对齐关系通常只占知识库的一小部分。

为了解决以上问题，文献 [659] 提出 MTransE 算法，通过在独立的嵌入空间中编码每种语言的实体和关系，为每个嵌入向量提供到其他语言空间中对应向量的过渡，同时保留单语言嵌入的功能。MTransE 的核心假设是相同实体在不同来源的知识图谱里存在类似的分布。MTransE 基本框架如图 12.18 所示，主要包含两个模块：平移模块和对齐模块。平移模块在特定于语言的知识图谱中编码实体和关系，采用 TransE 作为知识嵌入模型。在此基础上，对齐模块学习跨语言的实体和跨不同嵌入空间的关系对齐。

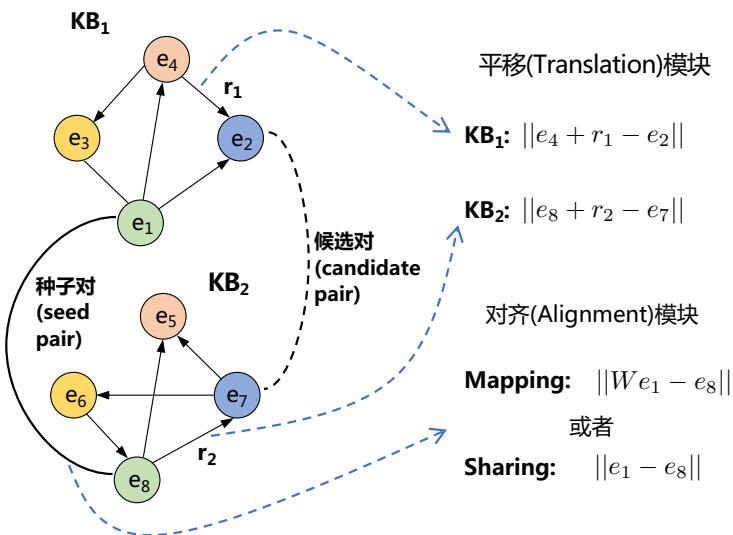


图 12.18 MTransE 算法示意图<sup>[659]</sup>

给定知识库，用  $\mathcal{L}$  表示语言集合， $\mathcal{L}^2$  表示两种语言对集合。对于某种语言  $L \in \mathcal{L}$ ，其语言特定的知识图谱为  $G_L$ ，其对应的实体和关系集合表示为  $E_L$  和  $R_L$ 。知识图谱中三元组  $T = < h, r, t >$  对应的嵌入向量分别为  $\mathbf{h}, \mathbf{r}, \mathbf{t}$ 。 $(L_1, L_2) \in \mathcal{L}^2$  表示由  $L_1$  和  $L_2$  两种语言组成的语言对， $\delta(L_1, L_2)$  表示语言对中对齐的三元组集合。

平移模块的主要功能是，通过基于平移的知识图谱嵌入模型，将随机初始化的实体嵌入约束

为固定分布。TransE<sup>[615]</sup>将关系解释为从头实体到尾实体的平移，因此其实体嵌入向量也具有平移不变性。MTransE 对每一种涉及的语言都采用了基于翻译的 TransE 方法，通过在不同的关系上下文中统一表示嵌入，有利于跨语言任务。因此其损失函数为：

$$S_K = \sum_{L \in \{L_i, L_j\}} \sum_{< h, r, t > \in G_L} \|h + r - t\| \quad (12.28)$$

由于知识库按照语言被划分为互不关联的子集，因此平移模块可以采取并行训练。

在实体对齐阶段，已有的知识库来源不同但是指代相同的实体对会作为初始种子，帮助对齐模块将不同知识图谱的实体嵌入映射到统一的向量空间进行对齐。其损失函数为：

$$S_A = \sum_{(T, T') \in \delta(L_i, L_j)} S_a(T, T') \quad (12.29)$$

对齐得分  $S_a$  的计算方法有两种，映射对齐和共享表示对齐：

- (1) 映射对齐方法通过线性变化，将来源不同的知识嵌入向量映射到一个统一的向量空间。如图12.18所示，优化的对象就是统一向量空间中的实体对之间的距离，如  $\|We_1 - e_8\|$ 。
- (2) 共享表示对齐是通过让每个预对齐的实体对，直接共享相同的嵌入，将不同的知识图谱嵌入到统一的向量空间中，这比映射方法更加直接。如图12.18所示，优化的对象直接就是初始嵌入空间的向量距离，如  $\|e_1 - e_8\|$ 。

在实际训练时，MTransE 采用  $S_K$  和  $S_A$  交替优化的方式实现。

## 2. 基于图神经网络的实体对齐方法

基于平移的知识嵌入模型实现的实体对齐模型，需要小心调节平移模块和对齐模块的损失权重，比较难优化。而且基于平移的方法的训练目标只包含单独的三元组，而实体的属性信息（例如人的年龄、城市的人口数）并没有被有效利用，所以其并不能从全局视角捕获实体和关系的信息。因此，文献 [660] 提出了 GCNAlign 算法，引入图卷积神经网络来建模全局信息，其模型结构如图12.19所示。GCNAlign 网络包含两个图神经网络，分别将不同源的知识图谱实体编码到一个统一的向量空间，然后通过对比损失或者三元组损失优化它们之间的距离。

具体来说，GCNAlign 考虑两种知识图谱中的三元组：关系三元组和属性三元组。关系三元组建模实体之间的关系，例如<爱因斯坦，毕业于，苏黎世联邦理工学院>，而属性三元组描述实体具备的属性，例如<爱因斯坦，享年，76岁>。因此可以将一个知识图谱表示为  $G = (E, R, A, T^R, T^A)$ ，其中  $E, R, A, V$  表示实体、关系、属性名和属性值集合； $T_R \subset E \times R \times E$  为关系三元组集合， $T_A \subset E \times A \times V$  为属性三元组。给定两个知识图谱  $G_1 = (E_1, R_1, A_1, T_1^R, T_1^A)$  和  $G_2 = (E_2, R_2, A_2, T_2^R, T_2^A)$ ，用  $S = \{(e_{i_1}, e_{i_2}) | e_{i_1} \in E_1, e_{i_2} \in E_2\}_{i=1}^m$  表示已知的对齐的实体对。因此跨语言实体对齐可以看作，根据已有对齐实体对发现新的实体对的任务。GCNAlign 主要包含三部分内容：基于图卷积网络的实体嵌入学习、对齐实体的预测和模型训练方法。

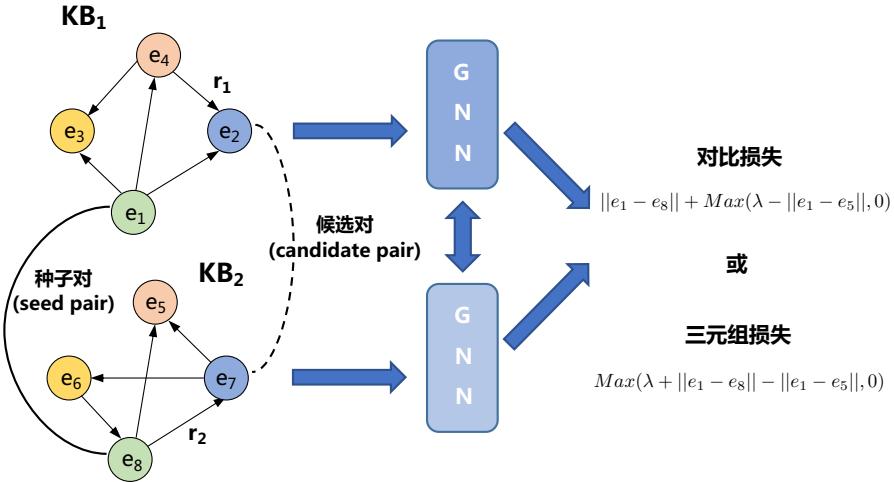


图 12.19 基于图神经网络的实体对齐方法

**基于图卷积网络的实体嵌入学习：**GCNAlign 的核心假设是：(1) 等价的实体一般拥有类似的属性；(2) 等价的实体其邻居实体一般也等价。图卷积神经网络 (GCN)<sup>[661]</sup> 是一种直接对图数据进行操作的神经网络，其输入是节点的特征向量和图的结构。GCN 的目标是学习输入图上特征的函数，并产生节点级输出。因此，GCNAlign 使用 GCN 将属性信息和结构信息结合在一起，把实体投影到低维向量空间中，并约束在低维向量空间中等价实体之间彼此接近。

GCNAlign 的输入是一个节点特征矩阵， $\mathbf{H}^{(l)} \in \mathbb{R}^{n \times d^{(l)}}$ ，其中  $n$  为顶点个数， $d^{(l)}$  为特征个数。GCN 的输出是一个新的特征矩阵  $\mathbf{H}^{(l+1)}$ ，图卷积的计算如下：

$$\mathbf{H}^{(l+1)} = \sigma \left( \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(l)} \mathbf{W}^{(l)} \right) \quad (12.30)$$

其中  $\sigma$  为 ReLU 激活函数， $\mathbf{A}$  为  $n \times n$  的邻接矩阵，用来表达图的结构信息， $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ ， $\mathbf{I}$  为单位矩阵； $\hat{\mathbf{D}}$  是  $\hat{\mathbf{A}}$  的对角节点度矩阵； $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$  为 GCN 的参数矩阵， $d^{(l+1)}$  为新节点特征的维度。

为了同时利用知识图谱中的结构信息和实体的属性信息，GCNAlign 建模两种特征：结构特征  $h_s$  和属性特征  $h_a$ 。令  $\mathbf{H}_s$  和  $\mathbf{H}_a$  分别为结构特征矩阵和属性特征矩阵，因此 GCNAlign 的卷积计算方式为：

$$\left[ \mathbf{H}_s^{(l+1)}; \mathbf{H}_a^{(l+1)} \right] = \sigma \left( \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \left[ \mathbf{H}_s^{(l)} \mathbf{W}_s^{(l)}; \mathbf{H}_a^{(l)} \mathbf{W}_a^{(l)} \right] \right) \quad (12.31)$$

其中  $\mathbf{W}_s^{(l)}$  和  $\mathbf{W}_a^{(l)}$  为结构特征的权重矩阵和属性特征的权重矩阵， $[;]$  表示矩阵的拼接操作。

GCNAlign 使用了两个两层的 GCN，分别建模两种语言的嵌入表示。知识图谱是个有向图，且实体之间通过代表不同关系的边相连接，因此 GCNAlign 设计了一种特定的邻接矩阵  $\mathbf{A}$  的方法。

另  $a_{ij} \in A$  表示从第  $i$  个实体传播到第  $j$  个实体的程度信息。考虑到两个实体通过不同的关系(例如, hasParent 与 hasFriend)连接到对齐的实体, 两个实体等价的概率差别很大。因此, 为每个关系计算功能性和逆功能性两种度量:

$$\begin{aligned}\text{fun}(r) &= \frac{\# \text{关系 } r \text{ 中头实体数量}}{\# \text{关系 } r \text{ 三元组数目}} \\ \text{ifun}(r) &= \frac{\# \text{关系 } r \text{ 中尾实体数量}}{\# \text{关系 } r \text{ 三元组数目}}\end{aligned}\quad (12.32)$$

基于此第  $i$  个实体对第  $j$  个实体的影响  $a_{ij} \in A$  计算方式如下:

$$a_i = \sum_{\langle e_i, r, e_j \rangle \in G} \text{ifun}(r) + \sum_{\langle e_j, r, e_i \rangle \in G} \text{fun}(r) \quad (12.33)$$

**对齐实体的预测:** 通过将预定义的距离函数应用于实体的 GCN 表示来预测实体对齐。对于来源于知识图谱  $G_1$  的实体  $e_i$  和来源于知识图谱  $G_2$  的实体  $v_j$ , 两者的距离计算方式如下:

$$D(e_i, v_j) = \beta \frac{f(\mathbf{h}_s(e_i), \mathbf{h}_s(v_j))}{d_s} + (1 - \beta) \frac{f(\mathbf{h}_a(e_i), \mathbf{h}_a(v_j))}{d_a} \quad (12.34)$$

其中  $f(x, y) = \|x - y\|_1$ , 而  $\mathbf{h}_s(\cdot)$  和  $\mathbf{h}_a(\cdot)$  表示实体的结构嵌入向量和属性嵌入向量,  $d_s$  和  $d_a$  表示结构嵌入向量和属性嵌入向量的维度,  $\beta$  则为平衡两者重要性的超参数。

对于对齐的实体, 两者距离应该较小, 对于非对齐实体, 该距离预计较大。给定  $G_1$  中的特定实体  $e_i$ , 计算  $e_i$  与  $G_2$  中的所有实体之间的距离, 并返回排序实体列表作为候选对齐, 也可以从  $G_2$  到  $G_1$  进行对齐。

**模型训练:** 为了使 GCN 在向量空间中嵌入尽可能接近的等价实体, GCNAlign 使用一组已知的实体对齐集合  $S$  作为训练数据来训练 GCN 模型。通过最小化以下基于边际的排序损失函数来进行模型训练:

$$\begin{aligned}\mathcal{L}_s &= \sum_{(e, v) \in S} \sum_{(e', v') \in S'_{(e, v)}} [f(\mathbf{h}_s(e), \mathbf{h}_s(v)) + \gamma_s - f(\mathbf{h}_s(e'), \mathbf{h}_s(v'))]_+ \\ \mathcal{L}_a &= \sum_{(e, v) \in S} \sum_{(e', v') \in S'_{(e, v)}} [f(\mathbf{h}_a(e), \mathbf{h}_a(v)) + \gamma_a - f(\mathbf{h}_a(e'), \mathbf{h}_a(v'))]_+\end{aligned}\quad (12.35)$$

其中  $[x]_+ = \max\{0, x\}$ ,  $S'_{(e, v)}$  表示由非对齐实体对的负例集合, 比如将实体  $e$  或实体  $v$  替换为  $G_1$  或  $G_2$  中的一个随机实体;  $\gamma_s, \gamma_a > 0$  为分离正负例的间隔超参。 $\mathcal{L}_s$  和  $\mathcal{L}_a$  是结构嵌入和属性嵌入的损失函数, 它们互相独立, 可以分开优化。

## 12.4 知识图谱推理

推理是从一个或几个已有判断中推导出新判断的过程，是人类有别于普通物种的重要能力之一。推理能力是实现通用人工智能的最终目标中必不可少的路径，而推理过程必须依赖于先验知识和已有经验。常见的推理方法如图所示12.20可以分为三类演绎推理，归纳推理和溯因推理：

- 演绎推理（Deductive Reasoning）是一种自上而下的推理方法，即根据已知的一般或普遍性前提推理得到必然成立的结论。比如，已知“没有电力供应电脑就不会正常工作”，同时知道“今天停电了”，那么就可以推理出“今天没有办法正常使用电脑”。
- 归纳推理（Inductive Reasoning）是一种自下而上的推理方法，即根据观察所得到客观知识进行总结归纳，从而得到更一般的结论。比如，观察发现看到的电脑都安装了 Windows，于是归纳出所有的电脑使用的都是 Windows 操作系统。显然，这种归纳未必一定正确，因为还有使用 Linux 和 MacOS 的电脑。生活中这种推理方式十分有用，因为观察样本足够多时，人们可以通过这种方式推理出大概率会发生的事件。
- 溯因推理（Abductive Reasoning）是从现象出发的推理方法，即结合已有的观察和知识，推断出有可能的解释过程。假如已知知识“没有电力供应电脑就不会正常工作”，同时观察到电脑没法正常启动，我们推断出“可能是因为停电了”。

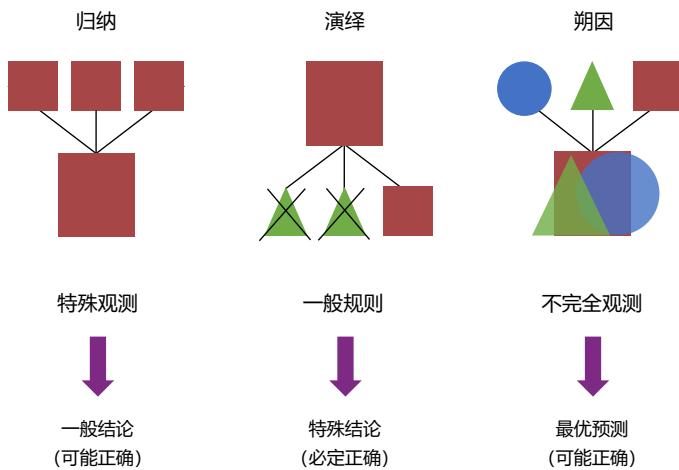


图 12.20 演绎推理、归纳推理和溯因推理示例

基于知识图谱的知识推理（Knowledge Reasoning），旨在从图谱中已存在的关联关系或事实推断出未知的关系或事实。推理得到的知识又可以反过来丰富知识图谱，为下游的应用提供更好的支持。随着知识图谱的普遍应用，基于知识图谱的推理受到了工业界和学术界的广泛关注。从推

理方法的角度，知识图谱推理方法可以分为两类，一种是演绎推理，主要是基于符号逻辑的知识图谱推理，依赖于专家显示制定的知识描述和逻辑推导方法；另一种是归纳推理，代表方法为基于表示学习的知识图谱推理，更多地依赖于大规模数据和统计学习算法。本章将从这两个类别对已有的知识图谱推理方法展开介绍。

### 12.4.1 基于符号逻辑的知识图谱推理

包括本体推理在内的早期知识推理方法受到了广泛关注，并产生了一系列推理方法，本节将展开介绍其中的三种：基于本体的推理、基于 Datalog 的推理以及基于产生式规则的推理。

#### 1. 基于本体的推理

一个定义完备的知识图谱包含两部分知识：(1) TBox (Theory Box)，即是关于概念术语的断言，对应 Schema 层，主要定义概念以及关系；(2) ABox (Assertion Box)，是关于个体的断言，关于事实性的描述。如图12.21所示。基于本体的推理方法，一般在 TBox 层定义本体公理 (Ontological Axioms)，然后根据演绎推理的方法推断出新的事实。

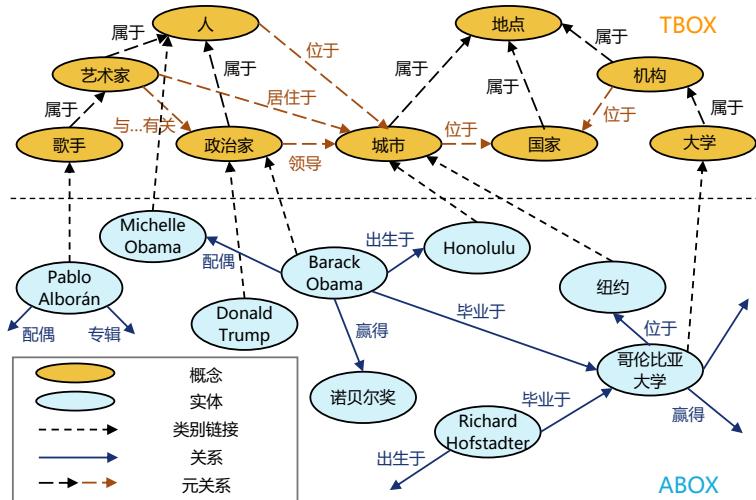


图 12.21 知识图谱中的 TBox 和 ABox 示例

本章第12.2.1节中介绍了 RDF、RDFS 和 OWL，其中 RDF 中的三元组对应客观世界的逻辑事实，基于 RDFS 描述语言的表达构件，如父子类关系，可以实现比较简单的推理。OWL 在 RDFS 进一步扩展了 Schema 的表达能力，可以完成更复杂的本体逻辑推理，如表12.7所示。

基于 OWL 的基础语法和公理基本语法，知识图谱可以实现一定的推理能力。例如：TBox 中定义“人工智能公司 subclass 高科技公司”，ABox 中存在知识“谷歌 type 人工智能公司”，那么就可以推断出“谷歌 type 高科技公司”；又比如在 OWL 本体中定义“父亲的父亲是爷爷”，已知“司马懿

表 12.7 OWL 语言基本语法

构造算子	语法	语义	例子
原子概念	$A$	$A^I \subseteq \Delta^I$	Human
原子关系	$R$	$R^I \subseteq \Delta^I \times \Delta^I$	has_child
对概念 $C$ , $D$ 和关系 (role) $R$			
合取	$C \sqcap D$	$C^I \cap D^I$	Human $\sqcap$ Male
析取	$C \sqcup D$	$C^I \cup D^I$	Doctor $\sqcup$ Lawyer
非	$\neg C$	$\Delta^I \setminus C^I$	$\neg$ Male
存在量词	$\exists R.C$	$\{x   \exists y. \langle x, y \rangle \in R^I \wedge y \in C^I\}$	$\exists$ has_child. Male
全称量词	$\forall R.C$	$\{x   \forall y. \langle x, y \rangle \in R^I \wedge y \Rightarrow C^I\}$	$\forall$ has_child. Doctor

是司马昭的父亲”和“司马昭是司马炎的父亲”，可以推断出“司马懿是司马炎的爷爷”。

OWL 本体推理的核心算法为 Tableaux 算法，其基本方法是通过一系列规则构建 ABox，然后检查知识库的可满足性。Tableaux 是一颗公式树，它会根据前提和否定结论来不断创建分支，对公式进行逐级分解，当所有分支都关闭后，Tableaux 算法就会终止。简单来说，如果要证明一件事是正确的，只要将其所有反例驳斥即可达到目的。

Tableaux 算法的描述逻辑算子如表12.8所示，以第一条规则为例，如果 ABox 中声明  $x$  属于  $C$  和  $D$  的组合类，但是  $C(x)$  和  $D(x)$  并不在 ABox 中，则把  $C(x)$  和  $D(x)$  都加入到 ABox 中。

表 12.8 Tableaux 算法规则

- $\sqcap^+$ -规则：若  $C \sqcap D(x) \in \emptyset$ ，且  $C(x), D(x) \notin \emptyset$ ，则  $\emptyset := \emptyset \cup \{C(x), D(x)\}$ ；
- $\sqcap^-$ -规则：若  $C(x), D(x) \in \emptyset$ ，且  $C(x) \sqcap D(x) \notin \emptyset$ ，则  $\emptyset := \emptyset \cup \{C(x), D(x)\}$ ；
- $\exists$ -规则：若  $\exists R.C(x) \in \emptyset$ ，且  $R(x, y), C(y) \notin \emptyset$ ，则  $\emptyset := \emptyset \cup \{R(x, y), C(y)\}$ ，其中， $y$  是新加进来的个体；
- $\forall$ -规则：若  $\exists R.C(x), R(x, y) \in \emptyset$ ，且  $C(y) \notin \emptyset$ ，则  $\emptyset := \emptyset \cup \{C(y)\}$ ；
- $\sqsubseteq$ -规则：若  $C(x) \in \emptyset, C \sqsubseteq D$ ，且  $D(x) \notin \emptyset$ ，则  $\emptyset := \emptyset \cup \{D(x)\}$ ；
- $\perp$ -规则：若  $\perp(x) \in \emptyset$ ，则拒绝  $\emptyset$ ；

实现 Tableaux 算法的推理系统包括曼彻斯特大学研发的 FaCT++<sup>[662]</sup>、美国 Franz 公司研发的 Racer<sup>[663]</sup>、马里兰大学研发的 Pellet<sup>[664]</sup>、牛津大学研发的 HermiT<sup>[665]</sup> 等，具体细节可以参考相关资料。

## 2. 基于 Datalog 的推理

基于本体的知识推理只能基于本体概念描述推理，不支持规则型知识的推理。此外，用户无法定义自己的推理过程。Datalog 则用于解决逻辑问题，将本体推理和规则推理相结合。Datalog 是一种基于逻辑编程语言 Prolog 且适应于知识库的改进型语言，能够方便地和大型数据库进行交互，便于撰写规则实现推理。

Datalog 的基本组成是原子  $p(t_1, t_2, \dots, t_n)$ , 其中  $p$  是谓词,  $t_i$  是变量或者常量。例如: has\_parent(X, Y), 其中“has\_parent”是谓词, “X, Y”是常量。一条规则  $H:-B_1, B_2, \dots, B_m$  由一个头部原子  $H$  和多个体部原子  $B_1, B_2, \dots, B_m$  构成, 表示当体部原子成立时, 可以得到头部原子成立的结论。

例如: has\_parent(X,Y):-has\_mother(X,Y)

表示“has\_mother(X,Y)”蕴含了“has\_parent(X,Y)”

除了规则外, Datalog 还包含了大量事实知识  $F(c_1, c_2, \dots, c_n) :-$ , 即没有体部且没有变量的规则。

例如: has\_mother(X,Y):-

Datalog 程序本质上就是规则的集合, 例如: 已知规则集合“坐落于 (X,Y):-同城 (X,Z), 坐落于 (Z,Y)”, 和事实集合“同城 (复旦大学, 华东师范大学):-; 坐落于 (华东师范大学, 上海):-”, 基于第一条规则, 第一条和第二条事实, 可以推理出新的事实“坐落于 (复旦大学, 上海)”。在实际应用中, 事实集合可能非常大, 随着规则集合越来越大, 推理的开销也会显著升高。

最常用的 Datalog 工具包括德国吉森大学研发的 DLV (Datalog with Disjunction)<sup>[666]</sup>、波茨坦大学研发的 Clingo<sup>[667]</sup>、牛津大学研发的 RDFox<sup>[668]</sup> 等, 具体细节可以参考相关资料。

### 3. 基于产生式规则的推理

产生式规则推理方法是一种前向推理系统, 从已知事实出发, 通过一定规则求得结论, 类似于一阶逻辑。其基本组成包含: Working Memory (WM) 中的事实集合, 产生式规则集合以及推理引擎。WM 中存储的事实数据包括两种, 一是用来描述对象, 例如: type attr1:val1 attr2:val2 ... attrn:valn, 其中 type, attri, vali 均为原子; 另一种用来描述关系, 例如: basicFact relation: olderThan firstArg:John secondArg:Alice。

产生式规则集合中存储的规则定义为“IF conditions THAN actions”形式, “conditions”对应条件集合, 又称为 LHS (Left Hand Side), “actions”对应动作序列, 又称为 RHS (Right Hand Side)。当 LHS 中所有条件都被满足的时候, 那么该规则触发, 执行相应的动作。

例如: IF (Gauge state: OK) AND (TEMPERATURE > 120) THEN (Cooling system state: over-heating)

表示: 当测量器状态正常且测量温度超过 120 度, 那么制冷系统的工作状态为过载

产生式规则的执行由推理引擎控制, 其核心任务是模式匹配。这个过程参考关系型数据库做知识图谱的存储和查询, 不同的条件匹配会产生大量的连接操作, 是个组合爆炸问题。此外, 还要考虑当满足多个触发规则时, 选择哪条规则执行的问题, 这被称为冲突解决。在选定规则时, 则执行规则的 actions, 对应 WM 中事实数据的增删改。

Rete 算法<sup>[669]</sup>是产生式规则系统中常用的推理算法, 其核心思路是逐渐扩充条件集合, 同时缓存条件匹配的中间结果, 直到所有规则对应的条件集合都得到判断, 是一种空间换时间的优化方法。如图12.22所示, Rete 算法分为  $\alpha$  网络和  $\beta$  网络。 $\alpha$  网络中的节点对应每条 condition,  $\beta$  网络的节点对应  $\alpha$  网络中节点 Join 后的 condition 集合。Rete 算法会先将 WM 中的事实先和  $\alpha$  网络中的节点进行匹配, 然后对满足条件的事实输入到  $\beta$  网络的节点中进行校验, 直到最后筛选出所

有至少满足一条规则的条件事实组合，完成推理。可见，若 Rete 算法中的节点被 N 条规则共享，即对应推理速度加速 N 倍。

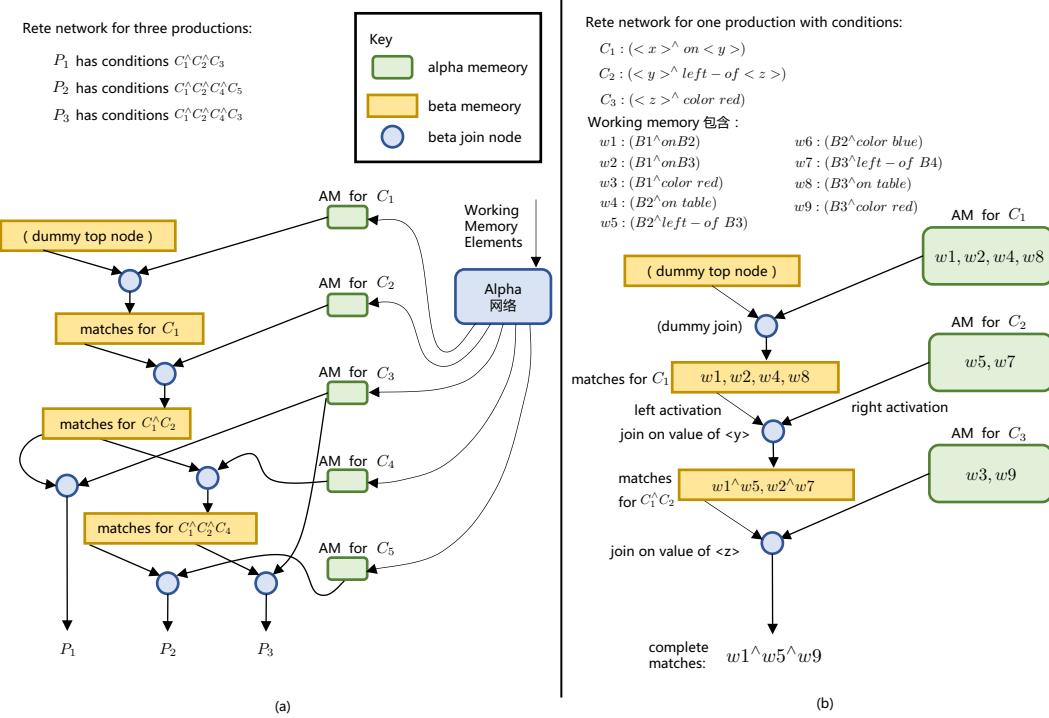


图 12.22 Rete 算法样例<sup>[669]</sup>

## 12.4.2 基于表示学习的知识图谱推理

基于符号表示的推理本质上基于演绎逻辑推理，精准度高但是也依赖于专家制定描述规则，当数据量较大时，推理的鲁棒性和效率都会显著降低，这种情况下基于统计的归纳学习推理有着不可替代的优势。随着神经网络研究的深入，基于深度神经网络表示学习的推理方法来逐渐成为研究的重点。本节将展开介绍两种基于表示学习的知识图谱推理方法。

### 1. 基于神经张量网络的知识图谱推理算法

知识图谱推理可以归结为根据知识库中已有事实和关系来推断新的事实的问题，包括给定两个实体，判断它们之间是否存在某种关系；给定头实体和关系，预测尾实体；给定三元组，判断其为真或假。基于已经训练好的知识表示，也可以直接应用于简单的三元组推理任务，例如，本章第12.2.2节中已经介绍过知识图谱的向量表示方法，包括 TransE、TransR 等。

不同于 TransE 等方法给每个实体、关系单独学习一个表示向量，文献 [670] 提出使用神经张

量网络 (Neural Tensor Network, NTN) 来分类实体对之间的关系，并使用构成实体的词向量的平均作为该实体表示。NTN 整体流程如图12.23所示。

具体来说，假设“Sumatran tiger”是知识图谱中的实体，NTN 会给构成其的词“Sumatran”和“tiger”分别学习一个嵌入向量，然后用两者的平均作为其表示。这么做的原因是在词向量的表示空间中，概念相似的词的嵌入向量距离也较近<sup>[178]</sup>。词向量的这种特性有助于提高具有公共子串的实体词之间共享统计强度，比如“Sumatran tiger” 和 “Bengal tiger”拥有共同的子串“tiger”，它们之间应该具有很高的关系性。在得到两个实体的表示向量  $(e_1, e_2)$  之后，NTN 使用神经张量网络来计算它们之间是否存在某种关系  $R$  的置信度。为此，NTN 为每个关系单独定义了一组参数用于建模实体对之间的语义关联。

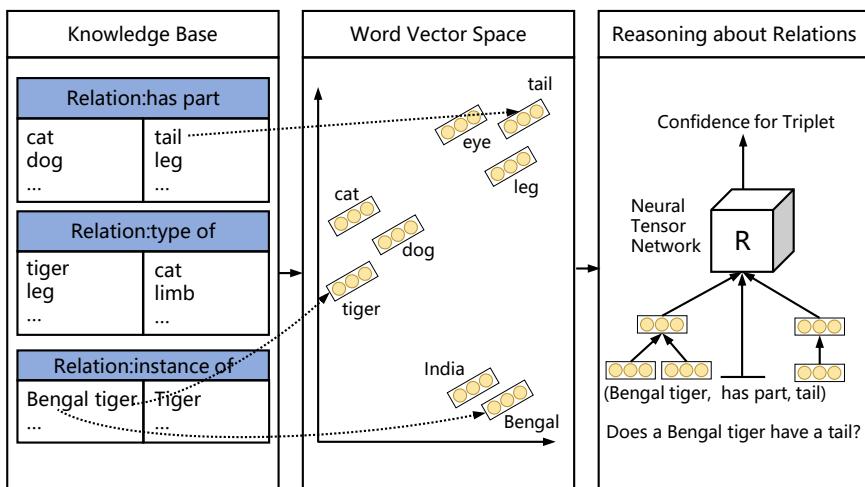


图 12.23 NTN 模型概述<sup>[670]</sup>

NTN 算法的核心是关系分类模块，该模块用来判断实体对  $(e_1, e_2)$  是否存在某种特定的关系  $R$ 。假设知识库中仅建模两种关系，不同于以往的方法直接将实体对的表示拼接在一起，NTN 采用双线性神经层  $W_R$  直接建模实体对之间的语义关系。NTN 输出的关系置信度打分，通过以下公式计算：

$$g(e_1, R, e_2) = \mathbf{u}_R^T f \left( \mathbf{e}_1^T \mathbf{W}_R^{[1:k]} \mathbf{e}_2 + \mathbf{V}_R \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} + \mathbf{e}_R \right) \quad (12.36)$$

其中  $f = \tanh$  为激活函数， $\mathbf{W}_R^{[1:k]} \in \mathbb{R}^{d \times d \times k}$  对应不同关系的切片参数，例子中仅存在两种关系，所以  $k = 2$ 。双线性张量积  $\mathbf{e}_1^T \mathbf{W}_R^{[1:k]} \mathbf{e}_2$  会生成向量  $\mathbf{h} \in \mathbb{R}^k$ ，其不同维度对应了不同关系的参数计算结果，如第  $i$  个关系使用第  $i$  个参数切片计算： $\mathbf{h}_i = \mathbf{e}_1^T \mathbf{W}_R^{[i]} \mathbf{e}_2$ 。关系  $R$  的其他参数还包括，

$\mathbf{V}_R \in \mathbb{R}^{k \times 2d}$ ,  $\mathbf{U} \in \mathbb{R}^k$  和  $\mathbf{b}_R \in \mathbb{R}^k$ 。

使用双线性神经层的好处是, 它可以使用乘法直接关联两个实体表示, 而不是像标准神经网络那样隐含地关联实体向量。直观地说, 张量  $\mathbf{W}_R$  的每个切片负责一种类型的实体对或关系的实例化。例如, 模型可以从词向量空间中学习到组成关系, 比如 (*dog, haspart, lag*) 和 (*car, haspart, engine*)。

模型的训练目标与 TransE 类似, 即真实存在的三元组  $T^{(i)} = \langle e_1^{(i)}, R^{(i)}, e_2^{(i)} \rangle$  相比随机实体替换得到的三元组  $T_c^{(i)} = \langle e_1^{(i)}, R^{(i)}, e_c \rangle$  能够得到更高的分数。令 NTN 的所有参数表示为  $\Omega = \{\mathbf{u}, \mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{E}\}$ , NTN 的损失函数为:

$$\mathcal{L}(\Omega) = \sum_{i=1}^N \sum_{c=1}^C \max \left( 0, 1 - g(T^{(i)}) + g(T_c^{(i)}) \right) + \lambda \|\Omega\|_2^2 \quad (12.37)$$

其中  $N$  为训练集中三元组的个数, 正确的三元组打分与替换三元组打分之间的差值最多为 1, 为每个三元组生成  $C$  个替换后的三元组。

NTN 为构成实体的每个词学习一个单独的表示, 而非整个实体。比如实体词“Sumatran tiger”, 其表示为  $v_{\text{Sumatran tiger}} = 0.5v_{\text{Sumatran}} + 0.5v_{\text{tiger}}$ 。虽然可以从随机初始化的词向量为每个词学习一个最终表示, 但是 NTN 还证明了使用无监督预训练的词向量初始化可以进一步提升模型性能, 这是因为大量的无监督语料有助于词向量更准确地捕获句法和语义信息。

## 2. 基于复空间关系旋转的知识图谱推理

已有的知识图谱表示学习方法, 一般是根据观察到的知识事实对知识图谱中缺失的关系进行建模和判断。在目标关系里存在一些特殊的关系, 如对称关系 (如婚姻关系)、反对称关系 (如父子关系)、反转关系 (如上位词和下位词关系), 还有多个关系可以组合为一个新的关系 (如母亲的丈夫是父亲) 等。上三种形式的关系具体形式定义如下:

**对称/反对称关系:** 一个关系  $r$  是对称的 (反对称的), 如果对任意的实体对  $x, y$ , 存在:

$$r(x, y) \Rightarrow r(y, x) \quad (12.38)$$

**反转关系:** 关系  $r_1$  和关系  $r_2$  是反转的, 如果对任意的实体对  $x, y$ , 存在:

$$r_2(x, y) \Rightarrow r_1(y, x) \quad (12.39)$$

**组合关系:** 关系  $r_1$  是  $r_2$  和  $r_3$  的组合关系, 如果对任意的实体  $x, y, z$ , 存在:

$$r_2(x, y) \wedge r_3(y, z) \Rightarrow r_1(x, z) \quad (12.40)$$

此前的方法只是显式或是隐式地建模以上部分关系, 无法做到对上述三种类型关系的准确推断。以 TransE 为例, 给定三元组 <科比, 婚姻关系, 瓦内萨>, 由  $h_{\text{科比}} + h_{\text{婚姻关系}} = h_{\text{Vanessa}}$  以

及  $h_{\text{瓦内萨}} + h_{\text{婚姻关系}} = h_{\text{科比}}$ , 推断出  $h_{\text{科比}} = h_{\text{瓦内萨}}$ , 这明显与事实不符。

文献 [671] 提出 RotatE 知识表示学习的方法以实现对上述所有关系的准确推理。RotatE 是受到欧拉公式的启发, 即  $e^{i\theta} = \cos \theta + i \sin \theta$ , 一个复数可以看作是在复平面上的一个旋转。相应地, RotatE 将实体和关系映射到复向量空间中, 然后将关系看作是从头实体向量往尾实体向量的旋转。除了可以建模以上所有关系之外, RotatE 的扩展性也非常好, 当知识图谱规模扩大时, RotatE 的时间复杂度和存储消耗都是线性的。

具体来说, 给定三元组  $\langle h, r, t \rangle$ , 其中  $h, r, t \in \mathbb{C}^k$  是表示向量, RotatE 希望表示向量之间满足  $t = h \circ r$ , 其中  $\circ$  表示按元素相乘。所以, 对在复空间中表示向量的每一维, 希望满足:

$$t_i = h_i r_i, |r_i| = 1 \quad (12.41)$$

其中  $h_i, r_i, t_i \in \mathbb{C}$ ,  $|r_i| = 1$  用于约束关系表示向量的模长。因此,  $r_i$  的复数形式为  $e^{i\theta_{r,i}}$ , 它对应于围绕复平面的原点逆时针旋转  $\theta_{r,i}$  弧度, 仅影响复向量空间中实体表示向量的相位。

理论上, 这样简单的复空间乘法操作可以有效建模上述所有关系。例如, 一个关系  $r$  是对称的, 当且仅当它的表示向量的每个元素, 即  $r_i$  满足  $r_i = e^{0/i} = 1$ ; 两个关系  $r_1$  和  $r_2$  是反转的, 当且仅当它们的嵌入是共轭的:  $r_2 = \bar{r}_1$ ; 关系  $r_3 = e^{i\theta_3}$  是其他两个关系  $r_1 = e^{i\theta_1}$  和  $r_2 = e^{i\theta_2}$  的组合当且仅当  $r_3 = r_1 \circ r_2$  (即  $\theta_3 = \theta_1 + \theta_2$ )。

以 1 维的表示向量为例, 对比 TransE 和 RotatE 方法, 如图12.24所示。TransE 将关系向量建模为直线上的向量平移, 可以建模反转关系、反对称关系和组合关系, 但是无法建模对称关系。但是 RotatE 将关系向量表示为复平面上的旋转, 除了 TransE 可以建模的关系之外, 还可以另关系向量满足  $r_i = -1$ , 来使得其满足对称关系的性质。

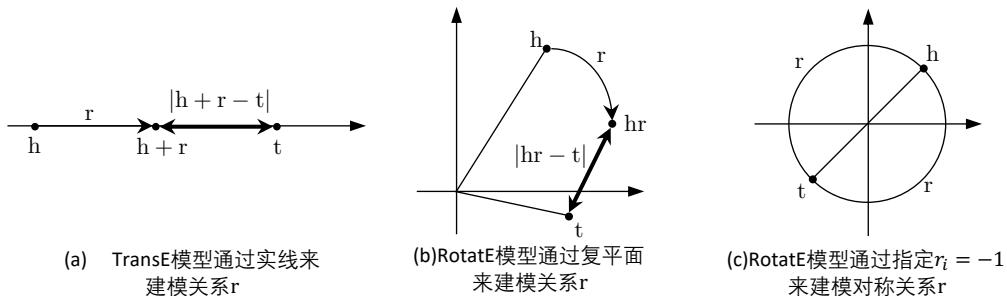


图 12.24 1 维表示向量上 TransE 与 RotatE 对比<sup>[671]</sup>

类似于 TransE, RotatE 将三元组  $\langle h, r, t \rangle$  的距离函数定义为:

$$d_r(h, t) = \|h \circ r - t\| \quad (12.42)$$

受 Word2Vec 中的负采样技术启发, RotatE 定义损失函数为:

$$\mathcal{L} = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^n \frac{1}{k} \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma) \quad (12.43)$$

其中  $\gamma$  为固定边距,  $\sigma$  是 sigmoid 函数,  $\langle \mathbf{h}'_i, \mathbf{r}, \mathbf{t}'_i \rangle$  为第  $i$  个负样本三元组,  $k$  是表示向量的维度。

基于均匀负采样的三元组存在优化效率低下的问题, 因为随着训练的进行, 许多样本三元组明显是假的, 不能提供任何有意义的信息。因此, RotatE 提出了一种称为自对抗的负采样方法, 即根据现有的嵌入模型对负样本进行采样。具体来说, 负样本三元组被采样的概率被定义为:

$$p(h'_j, r, t'_j | \{h_i, r_i, t_i\}) = \frac{\exp \alpha f_r(\mathbf{h}'_j, \mathbf{t}'_j)}{\sum_i \exp \alpha f_r(\mathbf{h}'_i, \mathbf{t}'_i)} \quad (12.44)$$

其中  $\alpha$  为温度系数。此外, RotatE 还将负样本的采样概率作为其权重加速训练。因此, 最终的自对抗训练采用如下目标作为最终损失函数:

$$\mathcal{L} = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^n p(h'_i, r, t'_i) \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma) \quad (12.45)$$

## 12.5 知识图谱问答

在本书第10章中我们介绍了智能问答的历史渊源和各种方法, 本节将围绕知识图谱问答(Knowledge Base Question Answering, KBQA) 相关方法展开介绍, 知识图谱问答也称 Knowledge Graph Question Answering (KGQA)。知识图谱问答旨在以知识库为知识源来回答自然语言问题。与文本问答相比, 知识图谱的结构化知识能够提供更加精准的答案, 并且方便扩展。知识图谱问答在许多智能应用中发挥着重要作用, 引起了研究人员的广泛关注。例如, Amazon Alexa<sup>[672]</sup>、Apple Siri<sup>[673]</sup> 和 Microsoft Cortana<sup>[674]</sup> 都集成了回答用户事实性问题的功能, 微软小冰和 Zo<sup>[674]</sup> 等聊天机器人也展示了高度的对话能力, 可以回答很多事实性问题。

知识图谱问答方法大概可以分为两类: (1) 基于语义解析的方法, 即将自然语言组织的问句转化为知识库可以识别的结构化查询语句, 然后在知识库中查询得到答案。而这个问句到结构化查询的映射, 可以通过规则定义的模板或者训练过的自然语言解析器来完成; (2) 基于信息检索的方法, 通过对问句中的实体和关系识别锁定问题的主题实体, 然后根据主题实体得到知识图谱中的候选实体, 最后对候选进行排序得到最终答案。随着深度学习技术的成熟, 深度学习技术也开始在知识图谱问答中广泛应用, 用来提升语义解析和检索排序方法的效果。本节会对这几种方法以及改进, 依次进行介绍。

### 12.5.1 基于语义解析的知识图谱问答

基于语义解析的方法是通过对自然语言进行语义分析，将其转化为计算机可以处理的语义表示，进而利用知识库对问句进行推理查询，得到最后的答案。在第4章中我们介绍了多种语义表示方法，逻辑表达式就是其中一种常用于知识图谱问答的语义表示方法，如图12.25所示。自然语言问句“who is Justin Bieber’s brother?”转化为 Lamda 算子表示，谓词 sibling\_of 表示兄弟姐妹关系，谓词 gender 表示性别。

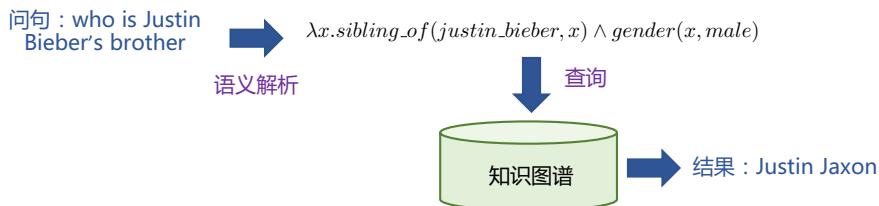


图 12.25 问句一步语义解析样例

直接将问句转化为语义表示的方法存在两个问题：一是非结构化的自然语言中许多的实体或是关系名称并不能直接匹配，例如图12.25示例文本中的“brother”和知识图谱中的关系描述“sibling\_of”不能直接匹配；此外，自然语言是多样化和口语化的，模板和知识库无法覆盖所有描述，难以将所有自然语言问句进行转化。

另外一种方式采用两步解析，首先将自然语言问句转化成一种中间表示，再进一步将中间表示翻译成逻辑语言，得到查询语句或查询语句的逻辑形式，在知识库上执行所得查询语句获得正确答案集。这里的查询语句常用有 SPARQL 查询语言，表达逻辑形式的逻辑语言常用有  $\lambda$ -DCS。这样可以使得解析更加精细，便于逻辑语言和知识库的对齐。图12.26给出了一个两步解析方法的示例。首先将问句转换为了与自然语言更加接近的中间表示。同样是针对问句“who is Justin Bieber’s brother？”，中间表示 brother\_of 与句子中的单词直接可以匹配。

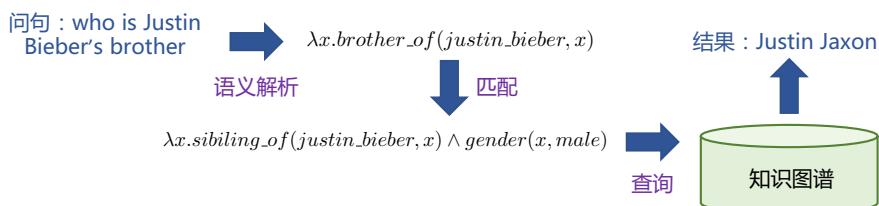


图 12.26 问句二步语义解析样例

## 1. 逻辑表达语言

有很多种逻辑表达语言，例如 Lamda 算子经常被用来对问句进行形式化描述。Lamda 演算子<sup>[675]</sup>由一个单一转换规则和一个单一函数定义系统构成。包括函数定义、标识符引用和函数应用三个部分。对于问题“What states border Texax?”，定义  $x$  是一个变量， $\lambda x$  表示获取一个参数  $x$  并且等待返回值的 lambda 函数，接下来通过两个函数 state 和 border 代表类型以及谓词语义，最后将函数 state 和 border 应用于参数  $x$ 。此外，可以添加其他修饰符，如存在量词、最大最小等，可以进一步增强语义刻画能力。据此上述问句可以转换为  $\lambda x.state(x) \wedge borders(x, texas)$ 。

依存组合语义 (Dependency-based Compositional Semantics, DCS)<sup>[676]</sup> 是一种特殊的操作桥接 (Bridging) 操作，把两个独立的语义表示片段连接起来，如图12.27所示。在语义分析过程，句子中汇总了不同片段对应的语义表示，可能无法直接进行合并，但是很多情况下过于分散的语义片段无法匹配到知识库中内容。引入桥接操作可以将两个不相邻的语义相连接，从而达到解析复杂语义的目的，尽可能连接这些离散的语义片段，在一定程度上保证语义分析能够完整。

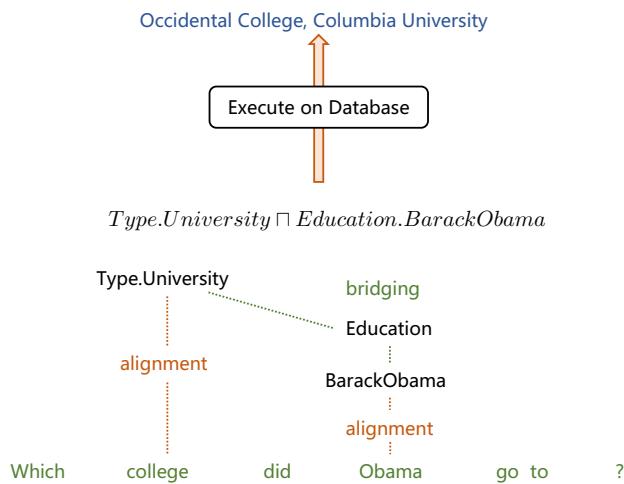


图 12.27 依存组合语义 DCS

## 2. 语义解析

语义解析的基本步骤可以分为短语检索、资源映射、语义组合和逻辑表达式生成这四个步骤。短语检索用于识别出问句中的实体、关系等各种短语。资源映射的目标是建立问句和知识图谱之间的映射，这包括实体链接、概念匹配和关系分类。将问句中实体、关系和知识图谱中概念相匹配后，还需要做句法分析、组合模型训练等工作。最终组合生成一个可执行的逻辑表达式，在知识图谱中获取最终答案。

我们以“复旦大学所在城市的人口有多少？”为例来介绍语义解析的具体过程，如图12.28所示。第一步将问句中关键信息提取出来，包括实体、关系谓词等。第二步与知识图谱建立各种关系，核心技术是：实体链接、概念匹配、关系抽取与分类。当给定问题后首先执行短语检测操作，主要识别句子中的变量（Variable）V，类别（Category）C，实体（Entity）E和关系谓词（Relation）R。这个过程可以使用序列标注模型来完成。在短语检测之后，需要完成依赖关系检测，例如实体“复旦大学”与关系“所在地”之间的依赖关系。最后完成短语向知识图谱的映射，这个过程可以采用实体链接、关系分类的技术，但很多时候也要依赖词典库辅助建立链接。

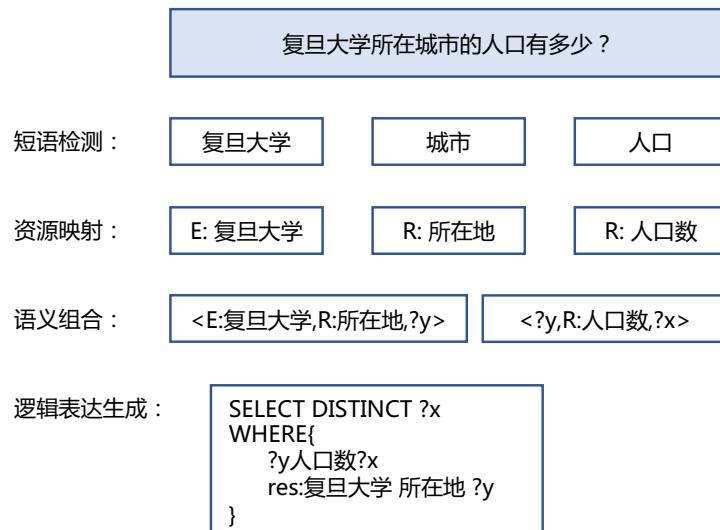


图 12.28 语义解析自然语言问句样例

在语义组合阶段，可以使用机器学习或者深度学习方法，生成有效的逻辑表达式。如果我们将自然语言及对应的逻辑语言看作是两种不同语言，语义分析任务也可以视作一种机器翻译任务。

最后，根据得到语义组合构建逻辑表达式，并在此基础上，结合知识图谱的存储结构，生成检索语句。利用检索语句从知识库中检索得到的结果，就可以用于回答问题。

### 12.5.2 基于信息检索的知识图谱问答

结合知识库来回答自然语言问题的另一种思路是，首先定位出自然问句中的主题实体，然后从知识库中的该实体相邻节点中选出候选答案，再对候选实体进行排序得到最后的答案。这种基于检索的知识图谱问答有两个核心问题：一是主题实体的识别，如果主题实体定位错误，之后的候选实体很大可能也是错误的。当知识图谱较复杂时，这需要通过实体链接将文本中的实体指代与其知识图谱中真正对应的实体完成映射；另一个是排序模型的效果，候选实体的得分也直接决

定最终结果，排序模型可以采用规则或者深度学习的方法实现。

文献 [677] 提出基于图谱的问答系统，利用 Freebase<sup>[354]</sup> 知识图谱回答问题。该方法基于的假设来自于日常人们如何寻找答案的过程：一个自然语言问题可能包括一个或几个主题，那么我们就可以在知识库中找相关的主题节点，然后从主题节点距离几跳的相关节点根据节点间的关系发现答案。例如询问“What is the name of Justin Bieber brother?”，并允许人们访问 Freebase 等知识库，人们通常会首先确定这个问题是关于 Justin Bieber，进一步根据 Justin Bieber 的 Freebase 页面所提供的关联节点，寻找他兄弟的名字。

该方法包括以下几个关键步骤：(1) 解析自然语言问句，围绕主题实体建立关于问句潜在特征的问题图；(2) 利用主题实体在 Freebase 中的邻近节点建立主题图；(3) 对齐问题图的特征和主题图的特征，找到问题与答案之间的关联；(4) 关系映射，即通过概率计算得到最可能为问题答案的关系。接下来将分别介绍上述步骤的具体方法。

### 1. 问题图构建

在查询答案时，通常存在多个逻辑约束。例如，对于问题“What is the name of Justin Bieber brother?”，可以根据以下内容寻找一个人的名字：依存关系 nsubj(what, name) 表示问题是寻求一个名字的信息；依存关系 prep\_of(name, brother) 表示这个名字是关于一个兄弟的；依存关系 nn(brother, Bieber) 表明 Bieber 是一个人名。通过以上的逻辑推理，可以在知识库中查询到所需要的答案。

查询答案的逻辑约束决定了基于依存特征的问题图设计。图12.29左侧给出了一个示例。左图使用虚线框中带注释的问题进行依存分析，转换后的特征图（右图）仅保留有关原始问题的相关和一般信息。许多语言信息对于答案提取提供了很多信息：疑问词（qword）反映了答案的类型，例如 What/who/how；问题焦点（qfocus）给出了预期答案类型的提示，例如姓名/金钱/时间；从问题的主要动词中提取的疑问动词（qverb），对答案类型也有很好的提示作用，例如动词“演奏”之后可能会跟着乐器、电影或运动队；问题主题（qtopic）则有助于找到相关的 Freebase 页面。

将问句的依存树转换为更通用的问题图，主要包含如下步骤：(1) 如果一个节点被标记了一个问题特征，那么用它的问题特征替换这个节点，例如，what→qword=what；(2) 如果 qtopic 节点被标记为命名实体，则将该节点替换为其命名实体形式，例如 bieber → qtopic=person；(3) 删除任何作为限定词、介词或标点符号的叶节点。转换后的结果如图12.29右侧所示，称为问题特征图，图中每个节点和关系都是这个问题的潜在特征。

### 2. 主题图构建

给定一个主题，通过选择与主题节点关系在若干跳内的节点形成主题图。除了传入和传出关系之外，节点还应具有属性：描述节点属性的字符串，例如节点类型、性别或身高。需要注意的是，关系和属性之间的一个主要区别是关系的两个参数都是节点，而属性只有一个参数是节点，另一个参数是字符串。关系描述节点之间相互关联的特性，例如，伦敦可以是贾斯汀比伯的出生地，也

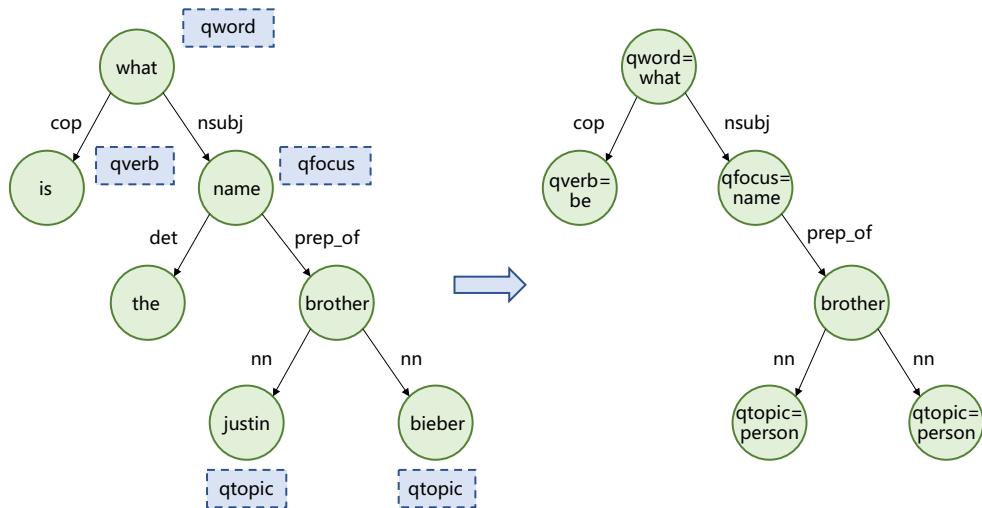


图 12.29 基于依存特征的问题特征图设计示例

可以是英国的首都。属性的参数是仅“附加”到某些节点并且没有传出边的属性。

图12.30给了一个 Freebase 主题图示例。以 Justin Bieber 为主题节点，通过节点间关系扩展后得到的主题图。图中实线框为节点，虚线框内为节点属性。阴影节点 Jaxon Bieber 为目标答案。

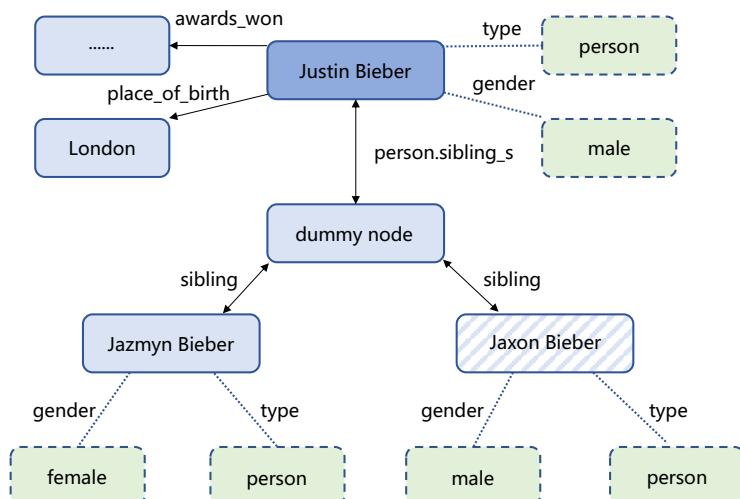


图 12.30 关于 Justin Bieber 主题的 Freebase 视图

### 3. 特征生成

在问题图和主题图构建完成后，需要构建人工特征用于表示问题和答案之间的关系。文献 [677] 所提出的方法主要构建了两种特征：问题特征和候选答案特征。**问题特征**：在问题特征图中的每一条边  $e(s,t)$ ，提取  $s, t, s|t, s|e|t$  作为特征。其中  $s$  是源节点， $t$  是目标节点， $e$  表示关系。例如，对于图12.29右侧，通过边  $\text{prep\_of}(qfocus=name, brother)$  可以提取以下特征： $qfocus=name, brother, qfocus=name|brother, qfocus=name|\text{prep\_of}|brother$ 。**候选答案特征**：一个节点的关系和属性对区分答案非常重要，对于一个问题对应的主题图，提取每个节点的所有关系和属性作为特征。然后将问题的一个特征和候选答案的一个特征组合在一起，这样可以捕捉问题模式和答案节点之间的关系。例如，问题-答案组合特征： $qfocus=money | \text{node\_type}=currency$ 。

### 4. 关系映射

关系映射的目标是构建知识库中的关系与自然语言单词之间的映射，发现问题所对应的关系。例如，对于“谁是乔治六世国王的父亲？”这个问题，目标是寻找  $\text{people}.\text{person}.\text{parents}$  关系。关系映射可以形式化的定义为：给定一个包含多个单词  $w$  的问题  $Q$ ，希望找出使概率  $P(R|Q)$  最大的关系  $R$ 。为了计算的简单性，假设单词之间的条件独立并应用朴素贝叶斯方法可以得到：

$$\tilde{P}(R | Q) \propto \tilde{P}(Q | R)\tilde{P}(R) \quad (12.46)$$

$$\approx \tilde{P}(w | R)\tilde{P}(R) \quad (12.47)$$

$$\approx \prod_w \tilde{P}(w | R)\tilde{P}(R) \quad (12.48)$$

其中  $\tilde{P}(R)$  是关系  $R$  的先验概率， $\tilde{P}(w | R)$  是给定  $R$  的单词  $w$  的条件概率。

如果没有关系  $R$  可以使得条件概率足够大，可以采用“子关系”（sub-relation）：关系  $R$  是一系列子关系  $R = r = r_1.r_2.r_3$  的串联。例如， $\text{people}.\text{person}.\text{parents}$  的子关系是  $\text{people}$ 、 $\text{person}$  和  $\text{parents}$ 。同样，假设子关系之间的条件独立并应用朴素贝叶斯方法：

$$\tilde{P}_{\text{backoff}}(R | Q) \approx \tilde{P}(r | Q) \quad (12.49)$$

$$\approx \prod_r \tilde{P}(r | Q) \quad (12.50)$$

$$\propto \prod_r \tilde{P}(Q | r)\tilde{P}(r) \quad (12.51)$$

$$\approx \prod_r \prod_w \tilde{P}(w | r)\tilde{P}(r) \quad (12.52)$$

为了获得上述先验概率和条件概率，需要大量的数据进行统计。文献 [677] 中使用 ClueWeb09 语料库<sup>[678]</sup> 以及基于该语料库的实体标注语料库 FACC1<sup>[679]</sup>。基于上述两个语料集合，针对 Freebase

中的关系，使用包含关系的两个实体的句子进行关系学习  $\tilde{P}(w | R)$  和  $\tilde{P}(w | r)$ 。整个学习过程可以分为以下几个步骤：

- (1) 将每个文档按句子拆分，并根据 Freebase 提取包含两个以上实体，并且实体间构成某种关系的句子；
- (2) 构造两个平行的语料库，一个是“关系-句子”对（用于估计  $\tilde{P}(w | r)$  和  $\tilde{P}(R)$ ），另一个是“子关系-句子”对（对于  $\tilde{P}(w|r)$  和  $\tilde{P}(r)$ ）。使用 ClueWeb09 和 FACC1 总计分别构建了 12 亿对平行语料；
- (3) 计算对齐关系和单词之间的共现矩阵。总共有 10484 个关系和子关系。
- (4) 根据共现矩阵，计算  $\tilde{P}(w | R), \tilde{P}(R), \tilde{P}(w | r)$  和  $\tilde{P}(r)$ 。

### 12.5.3 基于深度神经网络的知识图谱问答

通过上一节的介绍，我们可以看到依靠人工构造特征和规则来进行问题理解和答案抽取，集成了大量繁琐的步骤，系统回答复杂问题能力也较难提升。近年来，随着深度学习技术的发展，基于深度学习的知识图谱问答已经成为一个重要的研究领域。

#### 1. 基于多列卷积神经网络的知识图谱问答

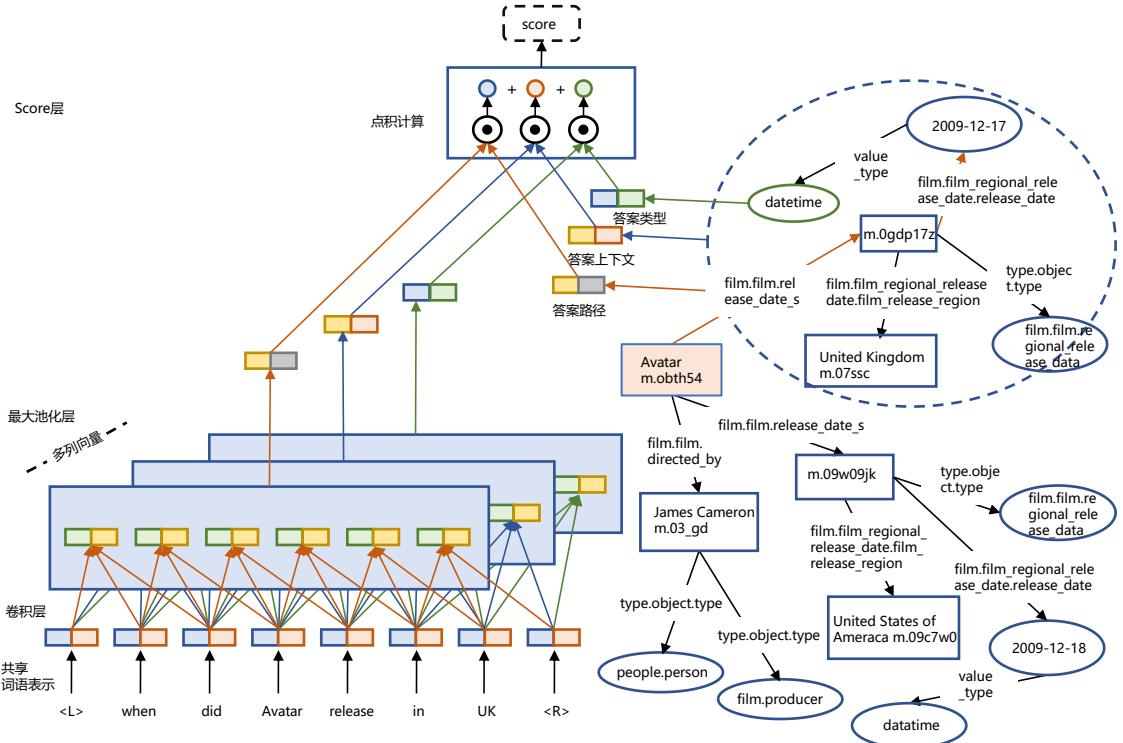
文献 [680] 中提出多列卷积神经网络（Multi-Column Convolutional Neural Networks, MCCNN）来进行知识图谱问答。具体来说，MCCNN 使用不同的列网络从输入问题中提取答案类型、关系和上下文信息，同时将知识库中的实体和关系也编码为低维向量。在此基础上，使用评分层根据问题和候选答案的表示对候选答案进行排名。MCCNN 算法框架如图12.31所示。例如，对于“What did Avatar release in UK?”问题，首先从 Freebase 知识库中查询实体 Avatar 的相关节点，将这些相关节点视为候选答案集合  $C_q$ 。然后对于每个候选答案  $a$ ，模型预测一个分数  $S(q, a)$  来确定它是否为正确答案。

MCCNN 用多列卷积神经网络来学习问题的表示。对于每个问题  $q$ ，这些不同的列卷积学习的向量表示为  $f_1(q)、f_2(q)、f_3(q)$ 。Freebase 中出现的候选答案的同样也需要进行嵌入表示，对于每个候选答案  $a$ ，利用列卷积计算其向量表示并将它们表示为  $g_1(a)、g_2(a)、g_3(a)$ 。这三个向量对应于问题理解中使用的三个方面。使用为问题和答案定义的这些向量表示，可以计算问答对  $(q, a)$  的分数。具体而言，评分函数  $S(q, a)$  定义为：

$$S(q, a) = \underbrace{\mathbf{f}_1(q)^\top \mathbf{g}_1(a)}_{\text{候选答案路径}} + \underbrace{\mathbf{f}_2(q)^\top \mathbf{g}_2(a)}_{\text{候选答案上下文}} + \underbrace{\mathbf{f}_3(q)^\top \mathbf{g}_3(a)}_{\text{候选答案类型}} \quad (12.53)$$

其中  $f_i(q)$  和  $g_i(a)$  具有相同的维度。如图12.31所示，通过分层计算得分并将它们相加。下面将具体描述如何去获取候选答案、抽取答案特征和问题特征，并进行计算的过程。

**候选实体生成：**根据问题首先需要从 Freebase 中检索问题的候选答案。如果问题中包含一个已识别的实体，可以通过该实体可以链接到知识库。这里使用了 Freebase Search API<sup>[354]</sup> 来查询问

图 12.31 MCCNN 模型结构图<sup>[680]</sup>

题中的命名实体。如果问题中没有任何命名实体，则查询名词短语。然后，将链接实体的所有 2 跳节点视为候选答案。将问题  $q$  的候选集表示为  $C_q$ 。

**问题理解：**MCCNN 使用多个卷积神经网络从共享输入词嵌入中学习问题的不同方面。如图12.31的左侧所示。对于问题  $q = w_1, \dots, w_n$ , 查找层将每个单词转换成一个向量  $\mathbf{w}_j = \mathbf{W}_v \mathbf{u}(w_j)$ , 其中  $\mathbf{W}_v \in R^{d_v \times |\mathbb{V}|}$  是词嵌入矩阵,  $\mathbf{u}(w_j) \in \{0, 1\}^{|\mathbb{V}|}$  是  $w_j$  的 one-hot 表示,  $\mathbb{V}$  表示单词表,  $|\mathbb{V}|$  是词表中单词数量。

卷积层计算滑动窗口中单词的表示。对于 MCCNN 的第  $i$  列, 卷积层计算问题  $q$  的  $n$  个向量。其中第  $j$  个词向量为:

$$\mathbf{x}_j^{(i)} = h \left( \mathbf{W}^{(i)} [\mathbf{w}_{j-s}^\top \dots \mathbf{w}_j^\top \dots \mathbf{w}_{j+s}^\top]^\top + \mathbf{b}^{(i)} \right) \quad (12.54)$$

其中  $(2s+1)$  是窗口大小,  $\mathbf{W}^{(i)} \in R^{d_q \times (2s+1)d_v}$  是卷积层的权重矩阵,  $\mathbf{b}(i) \in R^{d_q \times 1}$  是偏置向量,  $h(\cdot)$  是非线性函数 (例如 softsign、tanh、sigmoid 等)。

最后, 使用一个最大池化层来获得问题的固定大小的向量表示。第  $i$  列中的最大池化层通过

以下方式计算问题  $q$  的表示:

$$f_i(q) = \max\{x_j^{(i)}\} \quad (12.55)$$

其中  $\max\{\cdot\}$  是向量上的元素运算符。

**编码候选答案:** 候选答案  $a$  学习三种不同的向量表示  $\mathbf{g}_1(a), \mathbf{g}_2(a), \mathbf{g}_3(a)$  的学习方法分别如下:

候选答案路径  $\mathbf{g}_1(a)$  表示候选答案节点和被询问的实体之间的关系。如图12.31所示, 实体 Avatar 与正确答案之间的 2 跳路径为 (film.film.release-date-s, film.film-regional-release-date.release-date)。候选答案路径表示为  $\mathbf{g}_1(a) = \frac{1}{\|\mathbf{u}_p(a)\|_1} \mathbf{W}_p \mathbf{u}_p(a)$ , 其中  $\|\cdot\|_1$  是 1-范数,  $\mathbf{u}_p(a) \in \mathbb{R}^{|R| \times 1}$  表示答案路径中每个关系是否存在的 0/1 向量,  $\mathbf{W}_p \in \mathbb{R}^{d_q \times |R|}$  是参数矩阵,  $|R|$  是关系的数量。

候选答案上下文  $\mathbf{g}_2(a)$  表示连接到候选答案路径的 1 跳实体和关系被视为候选答案上下文。它用于处理问题中的约束。如图12.31所示, 询问了《阿凡达》在英国的发布日期, 因此仅考虑答案路径上的三元组是不够的。在上下文信息的帮助下, 英国的发布日期得分高于美国。上下文表示为  $\mathbf{g}_2(a) = \frac{1}{\|\mathbf{u}_c(a)\|_1} \mathbf{W}_c \mathbf{u}_c(a)$ , 其中  $\mathbf{W}_c \in \mathbb{R}^{d_q \times |C|}$  是参数矩阵,  $\mathbf{u}_c(a) \in \mathbb{R}^{|C| \times 1}$  是表示上下文节点的存在与否,  $|C|$  是出现在答案上下文中的实体和关系的数量。

候选答案类型  $\mathbf{g}_3(a)$ : Freebase 中的类型信息是对候选答案评分的重要线索。如图12.31所示, 2009-12-17 的类型是 datetime, 而 James Cameron 的类型是 people.person 和 film.producer。对于示例问题 Avatar 何时在英国发布, 应为类型为 datetime 的候选答案分配比其他答案更高的分数。向量表示定义为  $\mathbf{g}_3(a) = \frac{1}{\|\mathbf{u}_t(a)\|_1} \mathbf{W}_t \mathbf{u}_t(a)$ , 其中  $\mathbf{W}_t \in \mathbb{R}^{d_q \times |T|}$  是类型嵌入的矩阵,  $\mathbf{u}_t(a) \in \mathbb{R}^{|T| \times 1}$  是表示答案类型存在或不存在的二进制向量,  $|T|$  是类型的数量。

**模型训练:** 对于问题  $q$  的每一个正确答案  $a \in A_q$ , 从候选答案  $C_q$  集合中随机抽取  $k$  个错误答案  $a'$ , 并将它们作为否定实例来估计参数。针对  $(q, a)$  和  $(q, a')$  的损失函数为:

$$\mathcal{L}(q, a, a') = \max(0, (m - S(q, a) + S(q, a'))) \quad (12.56)$$

其中  $S(\cdot, \cdot)$  是等式中定义的评分函数,  $m$  是用于规范两个分数之间的差距的边际参数。

## 2. 基于知识图谱多跳问答算法 EmbedKGQA

简单的知识图谱问答只需要从 1 跳关系中获取答案, 而知识图谱中关系之间跳转可能存在更多的潜在候选答案。多跳知识图谱问答正是需要对知识图谱中的多条边进行推理以得出正确答案。但是知识图谱通常是不完整的, 有许多缺失的关系, 这给多跳知识图谱问答带来了额外的挑战。

文献 [681] 提出了用于知识图谱的多跳问答任务的 EmbedKGQA 方法, 使用知识图谱嵌入来完成该任务, 在稀疏知识图谱检索时特别有效。知识图谱中所有实体和关系的集合分别用  $E$  和  $R$  表示,  $K \subseteq E \times R \times E$  表示知识图谱中所有可用事实的集合。知识图谱问答中的问题涉及给定一个自然语言问题  $q$  和问题中存在的主题实体  $h \in E$ , 任务是提取一个正确回答问题  $q$  的实体  $t \in E$ 。

EmbedKGQA 使用知识图谱嵌入来回答多跳自然语言问题, 并利用缺失链接预测可以用来降

低知识图谱的稀疏性，提升知识图谱在多跳问答的表现。EmbedKGQA 的结构如图12.32所示，主要包含知识图谱编码、问题编码以及答案选择三个主要部分。EmbedKGQA 算法，首先在嵌入空间中学习知识图谱的表示。然后对给定问题，模型学习问题嵌入表示。这里将嵌入空间称为  $C_d$ 。最后，模型将结合知识图谱嵌入表示和问题嵌入表示来预测答案。

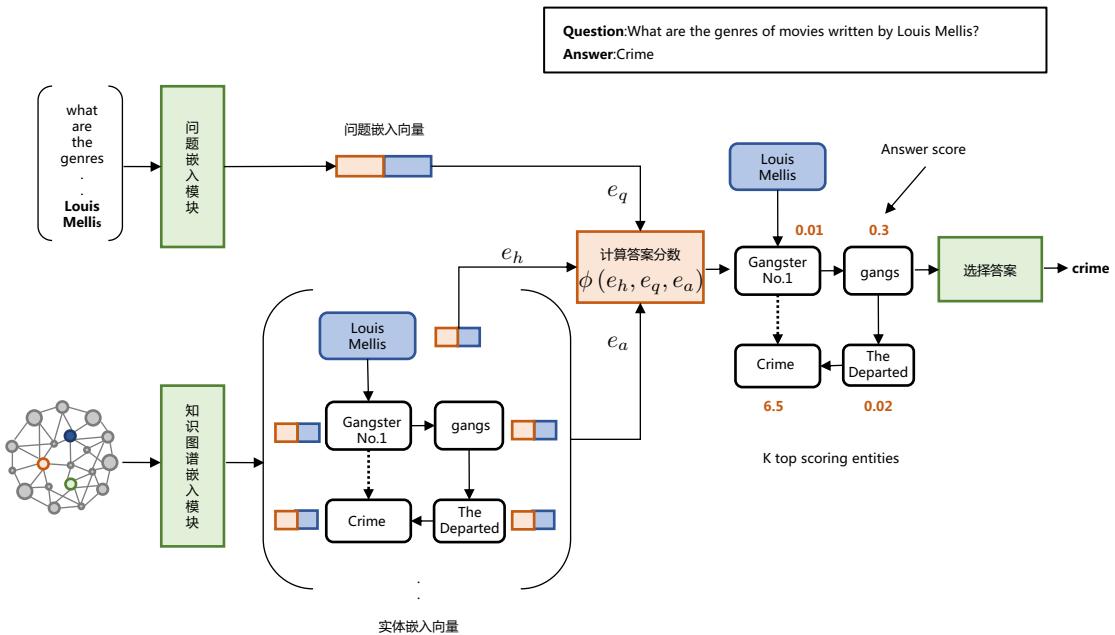


图 12.32 EmbedKGQA 算法神经网络结构<sup>[681]</sup>

**知识图谱编码模型：**为知识图谱中的所有实体  $h, t \in E$  和所有关系  $r \in R$  训练嵌入向量表示  $e_h, e_r, e_t$ ，使得  $e_h, e_r, e_t \in C_d$ 。这里的知识图谱嵌入向量表示用于计算问题和答案实体之间的评分函数。

**问题编码模型：**将自然语言问题  $q$  编码到同一维度的向量空间中  $e_q \in C_d$  中。这是使用前馈神经网络完成的，该网络首先使用 RoBERTa<sup>[308]</sup> 将问题  $q$  编码到 768 维向量中。然后通过具有 ReLU 激活的 4 个完全连接的线性层，最后投影到高维空间  $C_d$  上。

给定一个问题  $q$ ，主题实体  $h \in E$  和一组候选答案实体  $A \subseteq E$ ，使用如下方式学习问题嵌入表示：

$$\begin{aligned} \phi(e_h, e_q, e_a) &> 0 \quad \forall a \in A \\ \phi(e_h, e_q, e_a) &< 0 \quad \forall a \notin A \end{aligned} \quad (12.57)$$

对于每个问题，使用所有候选答案实体  $a' \in E$  计算分数  $\phi(\cdot)$ 。通过最小化 Sigmoid 得分和目

标标签之间的二元交叉熵损失来学习模型。

$$\mathcal{L} = BCE(\text{sigmoid}(\phi(e_h, e_q, e_a)), \text{Label}) \quad (12.58)$$

**答案选择模型：**最终在推理时，模型根据所有可能的答案  $a' \in E$  对 (head, question) 对进行评分。对于相对较小的知识图谱，如 MetaQA 等，只需选择得分最高的实体：

$$e_{ans} = \arg \max_{a' \in \epsilon} \phi(e_h, e_q, e_{a'}) \quad (12.59)$$

为了从大规模识图谱中选择候选答案，EmbedKGQA 还引入一个评分函数  $S(r, q)$ ，用于对给定问题  $q$  的每个关系  $r \in R$  进行排序。 $\mathbf{h}_r$  是关系  $r$  的嵌入向量， $q = (< s >, w_1, \dots, w_{|q|}, < /s >)$  是问题  $q$  中输入到 RoBERTa 的单词序列。评分函数定义为 RoBERTa( $\mathbf{h}_q$ ) 的最后一个隐藏层的最终输出与关系  $r(\mathbf{h}_r)$  的点积，之后使用 Sigmoid 函数归一化。

$$\mathbf{h}_q = \text{RoBERTa}(q) \quad (12.60)$$

$$S(r, q) = \text{sigmoid}(\mathbf{h}_q^T \mathbf{h}_R) \quad (12.61)$$

在所有关系中，选择那些得分大于 0.5 的关系，记为集合  $R_a$ 。对于目前获得的每个候选实体  $a'$ ，找到头部实体  $h$  和  $a'$  之间最短路径中的关系，并将这组关系记为  $R_{a'}$ 。每个候选答案实体的关系分数定义为它们的交集的大小：

$$RelScore_{a'} = |R_a \bigcap R_{a'}| \quad (12.62)$$

最后，使用关系分数和答案得分的线性组合来找到最终的答案实体。

$$e_{ans} = \arg \max_{a' \in N_h} \phi(e_h, e_q, e_{a'} + \gamma * RelScore_{a'}) \quad (12.63)$$

其中  $\gamma$  是一个可调节超参数。

### 12.5.4 知识图谱问答语料库

知识图谱问答的评价指标和分类问题类似，通常用准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 和 F1 值等，这些指标在前文多有介绍，这里不再展开描述。目前常用的知识图谱问答语料库如表12.9所示。

#### 1. QALD 知识图谱问答语料库

QALD<sup>[682, 683]</sup> 自 2011 年提出，到 2022 为止最新版本为 QALD-9 Plus。该语料库针对 DBpedia 知识库，构建了多语言问答、基于链接数据的跨数据集问答、融合文本数据的混合问答等任务。包

表 12.9 知识图谱问答数据集

数据集	知识图谱	语言	数据集规模
QALD	DBpedia	多语言	558
WebQuestions	Freebase	英语	5810
SimpleQuestions	Freebase	英语	10.8 万
MetaQA	Wikipedia	英语	40 万

含 558 个问题，并针对每个问题提供可以在 DBpedia 上检索得到正确答案的 SPARQL 语句、答案在 DBpedia 上的 URI 标识以及答案的类型。

## 2. WebQuestions 知识图谱问答语料库

WebQuestions<sup>[684]</sup> 是 2013 年由斯坦福大学研究人员构建的通用领域的知识图谱问答评测集合，包含 5810 个问题答案对，基于 Freebase 知识库构建。数据集构建过程中通过使用 Google Suggest API 获取了超过一百万问题，并从中选取了 10 万条，采用众包的方式人工利用 Freebase 知识库进行回答。从所有回答中选取了具有多个一致答案的问题和答案组成了 WebQuestions 语料库。

## 3. SimpleQuestions 知识图谱问答语料库

SimpleQuestions<sup>[685]</sup> 是 2015 年由 Facebook AI 研究人员构建的针对简单问题的大规模知识图谱问答语料。该语料库采用人工标注的方法，以 Freebase 作为答案来源，根据知识库中的事实人工构造问句，总计包含 108442 个问题答案对。

## 4. MetaQA 知识图谱问答语料库

MetaQA<sup>[686]</sup> 是 MoviE Text Audio QA 的缩写，主要包含三个部分：普通文本数据（Vanilla text data）、神经网络翻译数据（NTM text data）以及音频数据（Audio data）。通常知识图谱问答中使用普通文本数据，其中包含 40 万个问题，其中单跳（1-hop）问答来源于 Facebook MovieQA（也成为 WikiMovies）数据集中的“wiki\_entities”分支，同时移除了问题中的歧义实体，并数十个增加了多跳问答类型。

## 12.6 延伸阅读

人类的知识是真正理解语言的基石，随着知识表示学习、知识获取技术和各种知识感知应用的出现，表示实体之间结构关系的知识图谱已成为认知智能的主流研究方向。本章对以知识图谱表示与存储、知识图谱获取与构建、知识图谱推理以及知识图谱问答等四个方面的基础概念和经典方法展开了介绍。虽然学术界和工业界在知识图谱的上述方面都开展大量系统的研究，但是在分布式的知识表示学习方法、知识图谱动态扩展以及知识图谱推理等方面仍有很多开放问题等待解决。

分布式的知识表示学习方法能够高效地建模知识图谱的拓扑特征，且具有一定的扩展性。过

去十年中开发了许多成功的知识表示学习方法，除了本章中已经介绍的 TransE<sup>[615]</sup>、TransR<sup>[647]</sup> 和 RotateE<sup>[671]</sup>，还包括 TransH<sup>[687]</sup>、TransD<sup>[688]</sup>、ComplEx<sup>[689]</sup> 和 PairRE<sup>[690]</sup> 等等方法。当前知识表示学习的一个挑战是需要有高质量的训练数据。训练数据通常需要人工标注，这是一项耗时且费力的工作，同时，由于知识的复杂性，人类标注的数据也可能存在不一致性或错误。这些都会影响知识表示学习方法的准确性和可靠性。另一个挑战是它们难以解释，无法对低度实体建模，存在词表外实体无法表示的问题。这使得在当前知识表示学习方法在标注数据稀疏场景下的应用大大受限。

大多数知识图谱从静态数据中捕获信息，但实体和关系可能随时间而变化，因此想要自主地维护时态知识图谱的完整性和正确性，需要对时态元素进行适当的建模。文献 [691] 首先提出使用递归神经网络构建时间知识图谱，文献 [692] 通过为静态模型配备一个历时实体嵌入函数来学习任何时间点的实体特征。时间感知知识图嵌入方法 TeLM<sup>[693]</sup> 则使用线性时间正则化器和多向量嵌入来执行时间知识图的四阶张量分解。解决这个问题的另一个思路是，利用事件信息辅助建模时态要素。事件是人类社会的核心特征之一，人们的社会活动往往是事件驱动的。知识图谱研究聚焦于实体和实体之间的关系，缺乏对事理逻辑知识的挖掘。针对上述问题，哈尔滨工业大学刘挺教授团队提出了事理图谱 (Eventic Graph, EG) 的概念<sup>[694]</sup>，旨在将文本中对事件及其关系的描述抽取并抽象出来，构建描述事件之间演化规律和模式的事理逻辑知识库。

基于神经网络的方法提供了高效的推理能力，但是知识图谱中的信息通常是不完整的，也就是说，它只涵盖了部分专业领域的信息，推理系统在进行推理时可能会遇到缺乏信息的情况，这会影响推理的准确性。因此零样本推理<sup>[695][696][697]</sup> 受到研究界的广泛关注，即在知识图谱中进行推理时，没有相关的样本数据可供使用。这意味着算法必须依靠图谱本身的结构和已知的实体和关系来推断新的信息。同样，可解释性人工智能的研究可以用来设计新的可解释的链接预测模型<sup>[698][699][700]</sup>，即利用知识图谱中的实体和关系来推断出新的信息的过程中，会根据已知的信息推断出新的结论，并能够对这些推断过程进行解释，以便人们更好地理解推理的基础和原因。

## 12.7 习题

- (1) 基于属性图的知识图谱表示有什么缺点？
- (2) TransR 算法相比 TransE 算法做了什么改进，可以解决什么问题？
- (3) 使用关系型数据库作为知识图谱的存储介质，想要兼顾存储空间的利用率和查询效率，应当使用哪种存储方案？
- (4) 实体对齐技术主要解决哪些问题？存在哪些技术难点？
- (5) 基于表示学习的知识图谱推理相比于基于符号的知识图谱推理，有何优势？
- (6) 深度学习技术可以应用在知识图谱问答的基于语义解析和基于信息检索的范式的哪些阶段？

# 13. 模型稳健性

随着深度神经网络在自然语言处理研究的不断深入，特别是大规模预训练模型的广泛应用，自然语言处理算法在各项任务的评测集合上都取得了非常好的效果。在阅读理解、语义推理等众多任务上，算法在评测集合上准确率已经超越了人类。但是很多模型在处理与训练数据仅有微小变化的样本时，其准确率却大幅度下降。有时仅是一个“逗号”或者一个字母的不同，就会使得模型分析结果发生改变。**模型稳健性**（Model Robustness）也称**模型鲁棒性**，主要研究模型在面对输入微小变化时的稳定性和正确性。模型稳健性的研究可以更好的提升模型在真实场景下的应用效果，是实现自然语言处理算法更广泛应用的重要基础。

本章首先介绍稳健性的基本概念和主要研究问题，在此基础上介绍了文本对抗攻击方法，文本对抗防御方法以及模型稳健性评价基准。

## 13.1 稳健性概述

2018年1月，在斯坦福大学发起的SQuAD阅读理解评测任务中，微软亚洲研究院提出的算法在准确率上先赶超了人类。短短三年后，2020年DeBERTa<sup>[701]</sup>以及T5+Meena<sup>[572]</sup>模型在包含了多种自然语言处理任务的综合评测集合SuperGLUE<sup>[702]</sup>上再次超越了人类。这些模型在不同任务上取得优异效果，其准确率不断提升的同时，我们也看到这些在实验室环境下取得很好效果的模型，用于真实环境时却缺不尽如人意。

一些研究发现，很多现有模型在处理与训练样本仅有微小变化的数据时，效果会大幅度下降。文献[703]发现在属性级情感分析任务中，针对目标属性的修饰词语进行微小变形，就会使得绝大部分模型分类准确率大幅度下降。例如，“汉堡很好吃薯条一般”中对汉堡的评价是正面的，但在句子中插入逗号后，模型很可能就会将“汉堡很好吃，薯条一般”预测为对汉堡的负面评价<sup>[703]</sup>。文献[704]针对命名实体识别任务的稳健性开展研究，发现如果对其中实体词进行替换，那么BERT-CRF在命名识别任务上微平均F1值(Micro-F1)会从81.76%降低到51.58%。针对阅读理解任务，文献[705]在文档中增加混淆句、在候选答案中增加混淆选项等方法验证了包括BERT、RoBERTa等在内的多种方法，在这些变形后的评测中，大部分模型准确率有平均40%的下降。大规模稳健性评测工具集合TextFlint<sup>[73]</sup>，针对12个自然语言处理任务的大规模评测结果也显示，现有算法在

大多数任务的测评数据集上的表现都较原始结果有所下降。即便是基于大规模预训练模型 BERT、XLNET 等算法在一些任务的精度指标上也呈现了超过 50% 的降幅。从这些研究结果可以看到，当前自然语言处理算法（特别是基于深度神经网络的算法）的稳健性问题是亟待系统研究基础问题之一。

### 13.1.1 稳健性基本概念

稳健性（Robustness，又称鲁棒性），在计算机学科中通常是指系统遭遇输入、运算等异常以及在执行过程中处理错误，从而能够继续正常运行的能力。模型稳健性则描述了模型在输入微小改变时的稳定性和正确性。具有较高稳健性的模型，在处理不应对输出造成影响的微小变化时，模型的预测结果不会发生变化。自然语言处理模型的稳健性除了取决于机器学习领域所广泛讨论的模型和学习准则之外，文本的表示以及训练数据都会对模型的效果和稳健性产生影响。

模型稳健性与模型泛化能力以及鲁棒机器学习密切相关。在机器学习领域通常考察模型的泛化能力（Generalization Ability），即模型对未知数据的预测能力。模型的泛化性能虽然与稳健性非常相关，但也略有区别。统计机器学习模型通常基于独立同分布假设，因此泛化能力通常也是考察模型在与训练语料在相同分布情况下的新鲜样本的预测性能。但是模型稳健性更多的是从模型在真实环境下的使用角度出发，具有微小变化的输入样本可能与训练样本的分布有微小不同。但是，模型泛化能力是稳健性的基础。除了独立同分布假设外，当前统计机器学习算法背后依赖的封闭世界假设以及大数据假设，也都影响了模型稳健性<sup>[706]</sup>。在真实环境下我们所遇到样本往往来自开放环境，有可能是噪声数据，也有可能是新类别数据，样本也很可能与训练数据分布有微小变化，并且训练数据也能并不充分，这些都对模型稳健性提出了很大的挑战。

鲁棒机器学习（Robust Machine Learning）目标通常聚焦于提升模型的对抗鲁棒性。该任务可以形式化表示为一个 min-max 问题，给定数据点  $(\mathbf{x}, y)$  服从未知分布  $\mathcal{D}$ ， $\mathcal{F}$  是一个假设的算法类型（例如一个特定结构的神经网络）， $f \in \mathcal{F}$  是一个分类器， $\mathcal{L}(f(\mathbf{x}, y))$  表示分类损失。 $\mathcal{L}_\infty$  白盒攻击的目标是针对给定  $\mathbf{x}$  寻找  $\mathbf{x}'$ ，使得  $\|\mathbf{x} - \mathbf{x}'\|_\infty < \epsilon$  的情况下  $\mathcal{L}(f(\mathbf{x}', y))$  最大。鲁棒机器学习的目标就是寻找最优的抵抗对抗攻击的模型，可以如下形式化表示：

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \max_{\|\mathbf{x}_i - \mathbf{x}'_i\|_\infty < \epsilon} \mathcal{L}(f(\mathbf{x}'_i, y_i)) \quad (13.1)$$

其中  $(\mathbf{x}_i, y_i)$  是从分布  $\mathcal{D}$  采样得到的独立同分布训练语料。目前大多数鲁棒机器学习都是在该框架下对模型鲁棒性进行理论分析，以及解决上述 min-max 优化问题。

在本章中，我们重点讨论自然语言处理模型稳健性问题，该问题与机器学习中的泛化能力以及鲁棒机器学习均有很大的联系。但是，由于自然语言具有离散的特点，虽然可以使用稠密向量对单词进行表示，但是向量的每个维度很难进行解释，如何按照公式 13.1 中点之间距离  $\|\mathbf{x} - \mathbf{x}'\|_\infty < \epsilon$  所获的表示很可能并不存在对应的单词。因此，很多自然语言处理任务的稳健性还不能很好的进

行形式化表示，这也是自然语言处理稳健性任务相较于图像更难解决的原因之一。

### 13.1.2 稳健性主要研究内容

如本书第1章第1.2.2节和第1.2.3节所述，当前自然语言处理任务通常转换为有监督机器学习问题，因此目前的自然语言处理框架通常五个部分（如图13.1所示）：数据构建、文本表示、模型架构、学习算法和性能评价。数据构建包括根据任务要求筛选数据集合并进行数据标注。文本表示方面，传统的机器学习算法需要人工根据任务和所使用的分类模型的不同，采用特征工程的方法人工构建；而深度神经网络则可以在训练过程中自行学习到特征表示。模型架构方面，目前主流的深度自然语言处理模型采用基于卷积神经网络、递归神经网络和自注意力机制等架构。模型学习过程则是根据准备好的训练数据集合，针对所使用模型以及学习准则，利用优化算法找到最优模型的过程。最后，还需要构造评价方法，对模型的效果进行评价。



图 13.1 基于有监督机器学习的自然语言处理算法基本框架

从目前的研究结果来看，数据构建、文本表示、模型架构、学习算法都会对模型稳健性产生影响。周志华教授在《机器学习》书中指出“要进行机器学习，先要有数据”<sup>[707]</sup>。数据是机器学习的基础。近年来的研究也表明训练数据构建的方式将直接影响到算法的鲁棒性。在数据层面，稳健性的主要研究内容包括数据偏差分析以及数据偏差消除。

Yoshua Bengio 教授在其 2013 年发表的关于表示学习的综述上指出：机器学习算法的成功通常需要依赖于数据表示<sup>[39]</sup>。业界也广泛流传着这样一句话：“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已”。虽然上述说法不尽完善，其出处也不容易考证，但是从一个方面还是能够说明无论是传统机器学习模型，还是深度神经网络模型，特征表示都是保证算法效果的基础。针对表示和模型，稳健性研究重点主要在于对抗攻击和对抗防御。文本对抗攻击按照扰动粒度可分为字符级别攻击、词级别攻击以及句子级别攻击。后门攻击则研究当模型权重或者训练数据改变时对训练后模型产生的影响。文本对抗防御旨在提高模型稳健性来抵御各种形式的对抗样本，按照采用的方法大致可分为基于对抗训练的文本防御方法、基于表示压缩文本防御方法以及基于数据增强的文本防御方法。对抗样本检测则希望在模型预测阶段过滤掉对抗样本并且拒绝为其进行服务来避免对抗样本的影响。

此外，针对自然语言处理任务的评价通常采用精度、召回、F1 值、准确率等指标。一个算法在标准测试集合上得到了很好的测试精度或者准确率，是否就意味着该算法在真实环境下就一定能得到很好的效果呢？经典的评价方法能全面反映算法的优缺点吗？算法在测试语料上取得很好的效果，是否真的说明算法达到语料集合创建者所预设的验证目标？正如我们在本章开头所提到的

那样，模型在公开评测集合上能够取得非常好的效果，甚至在复杂任务中都超越了人类水平，但是真实环境下效果却大幅度下降，在一定程度上也反映了传统评测方法的不足。针对这些问题，近年来一些研究从机器学习、自然语言处理、特定任务等角度分别开展了一些研究。

本章将针对上述问题和研究内容，从数据偏差消除、文本对抗攻击方法、文本对抗防御方法以及模型稳健性评价等方面分别进行介绍。

## 13.2 数据偏差消除

文献 [708] 对数据构建问题给出了如图13.2所示的非常形象的描述，黄色点和蓝色点分别代表两类数据，在如图13.2(a) 所示的数据分布情况下，需要如红线所示的复杂分类边界。但是通常情况下，训练数据的采样并不充分，尤其是针对复杂任务，因此很可能出现如图13.2(b) 所示的情况，即采样得到的样本与实际情况有偏差。在有偏数据采样下，数据分布发生变化，标注数据训练得到的分类边界也会相应地发生变化。当训练数据样本不充足时，样本采样的偏差很可能会产生系统性的误差，使得模型训练不可能达到预期的能力。



图 13.2 数据构建方式不同会使得模型产生系统性误差<sup>[708]</sup>

获得 AAAI 2020 最佳论文奖的文献 [709] 针对 Winograd Schema Challenge (WSC) 任务开展了详细分析。该任务包含一组专家精心设计的 273 个代词消解问题，试图验证模型是否拥有常识推理的能力。该任务在提出时希望作为图灵测试的替代方案，从而可以不需要人工的情况下验证模型的能力。

- 例如：(1) The trophy doesn't fit into the brown suitcase because it's too large. **trophy/suitcase**
- (2) The trophy doesn't fit into the brown suitcase because it's too small. **trophy/suitcase**

句子 (1) 中的“it”指代“trophy”，而句子 (2) 中的“it”则指代“suitcase”。但是文献 [710] 的研究却发现，有超过 13.5% 的评测数据中存在单词关联 (Word Association) 以及一些其他特定于数据集的偏差 (Dataset-specific Bias)。比如，对于句子“The lions ate the zebras because **they** are predators.”中“they”的指代，并不需要对句子进行理解，由于在语言模型中的“lion”与“predators”的共现程度远大于“zebra”与“predators”的共现程度，因此模型仅依赖语言模型就可以得到正确答案。也正是由于这些数据集中存在的大量偏差存在，使得基于该集合进行训练所得到的模型鲁棒性不高。

针对该数据集合中的偏差问题，文献 [709] 首先通过众包的方式构建了一个由 44000 个问题组成的原始大规模数据集 WINOGRANDE。在此基础上，提出了 AFLITE 算法用于系统地减少数据集中的偏差。该方法在对抗过滤算法（Adversarial Filtering, AF）<sup>[711]</sup> 基础上进行改进，可以使更广泛的范围并且更加轻量化，具体过程如算法13.1所示。

---

#### 代码 13.1: AFLITE 数据集偏差去除算法

---

**输入:** 数据集  $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ , 集成模型数量  $n$ , 集成模型的训练集合大小  $m$ , 过滤临界值

(cutoff) 的大小  $k$  以及过滤阈值  $\tau$

**输出:** 数据集  $\mathcal{D}'$

$\mathcal{D}' = \mathcal{D}$

**while**  $|\mathcal{D}'| > m$  **do**

// 过滤过程

**foreach**  $e \in \mathcal{D}'$  **do**

$E(e) = \emptyset$  // 初始化集成预测结果为空;

**end**

**for**  $i = 1$  **to**  $n$  **do**

随机划分数据集合  $\mathcal{D}'$  得到  $(\mathcal{T}_i, \mathcal{V}_i)$ , 其中  $|\mathcal{T}|_i = m$

根据数据集合  $(\mathcal{T}_i, \mathcal{V}_i)$  训练线性分类器  $\mathcal{L}$

**foreach**  $e = (\mathbf{x}, \mathbf{y}) \in \mathcal{V}'$  **do**

$E(e) = E(e) \cup \mathcal{L}(\mathbf{x})$

**end**

**end**

**foreach**  $e = (\mathbf{x}, \mathbf{y}) \in \mathcal{D}'$  **do**

$score(e) = \frac{|p \in E(e) \& p = y|}{|E(e)|}$

**end**

选择得分最高并且  $score(e) > \tau$  的前  $k$  个样本组成集合  $S$

$\mathcal{D}' = \mathcal{D}' \setminus S$

**if**  $|S| < k$  **then**

**break**

**end**

**end**

**return**  $\mathcal{D}'$

---

算法的输入为原始数据集  $\mathcal{D}$  以及相关的参数，输入为过滤后的数据集合  $\mathcal{D}'$ 。在每个过滤阶段，将数据集合进行随机分片，利用不同分片训练得到  $n$  个线性分类器，并在对应的验证集合上进行预测。对于数据集合中的每个实例，根据正确预测与预测总数之比作为其得分。根据得分将分数超过  $\tau$  的前  $k$  个数据进行删除。重复执行上述过程直到在过滤阶段不能发现超过  $k$  个需要过

滤的样本，或者总样本数少于  $m$  为止。在 AFLITE 应用于 WINOGRANDE 数据集合时， $m$  设置为 10,000， $n$  为 64， $k$  为 500， $\tau$  为 0.75。评测结果显示，数据集合中存在的大量单词关联以及语言偏差（Language-based Bias）使得模型可以非常容易得通过拟合数据集中的简单规则，在特定基准集合上取得非常好的效果。但是这些模型并没有真正学会基于知识作出推理，而是简单地基于伪相关性（Spurious Correlations）进行预测，从而导致模型鲁棒性较差。最后通过过滤得到了包含 12000 个样本的集合  $\text{WINOGRANDE}_{\text{debiased}}$ 。

### 13.3 文本对抗攻击方法

对抗攻击（Adversarial Attack）是对目标机器学习模型的原输入施加轻微扰动，生成对抗样本（Adversarial Example）使得目标模型产生错误分类。对抗攻击是验证机器学习模型稳健性最重要的方法之一。在计算机视觉领域中通常通过对原始图像添加微弱的像素扰动来生成对抗样本，人眼几乎无法辨别对抗样本和原始图像的区别。由于文本离散的特点，对输入的表示向量添加微小扰动并一定存在对应的单词，因此不能这种方式生成对抗样本。再加上自然语言语义和搭配复杂，具有相似含义的词语由于语言搭配和习惯的关系，哪怕仅仅一个字的改动也可能会破坏原文本的语法正确性和流畅性，使得产生的对抗样本质量较差。例如，“北京大学”修改为“北京的大学”，其语义的覆盖范围发生非常大的变化。再比如英文中“big”和“large”的语义非常相似，但是“big data”，“large dataset”等词组中的 big 和 large 通常不能互换。自然语言处理领域的对抗攻击相较于图像更具挑战性。

文本对抗攻击可以从被攻击模型可见性以及扰动粒度两个方面进行分类。根据能够利用模型内部信息的多少，可以将攻击方法划分为：白盒攻击（White-Box Attack）、黑盒攻击（Black-Box Attack）以及盲攻击（Blind Attack）。如果能够完全掌握受害模型的结构、参数等所有信息，在这样的设定下完成的攻击被称为白盒攻击。相反的，如果在无法获得受害模型的内部结构及参数情况下进行的攻击则被称为黑盒攻击。而当被攻击模型的输出也未知时的攻击则称为盲攻击。通常情况下，攻击效果与获得的信息多少相关，能够获得的受害模型信息越多，相应的攻击效果就越好。此外，还可以根据对于输入扰动的粒度对算法进行划分，包括：句子、词语以及字符等級別。考虑到文本的离散特性，以及现有多數方法同时适用于白盒攻击和黑盒攻击场景，在本节中，我们将根据扰动粒度对相关研究进行介绍。

文本对抗攻击任务可以形式化的表示为，给定一个分类器  $f$  和一个语料库  $\mathbb{D}$ ，对抗攻击者目标是一个给定的数据样本  $x$  生成对抗样本  $x^*$ ，使得模型产生不同的预测结果，即  $f(x^*) \neq f(x)$ 。一般来说，样本  $x$  可能不在  $\mathbb{D}$  中，但来自于同一个潜在分布  $P_{\mathbb{D}}$ 。白盒攻击、黑盒攻击以及盲攻击算法的主要区别在于，从模型  $f(x)$  所获得的信息的不同。

### 13.3.1 字符级别攻击方法

HotFlip<sup>[7][2]</sup> 算法是基于字符替换来产生对抗样本的白盒攻击方法，并通过替换连续字符的方式来支持插入和删除操作。HotFlip 使用模型输出计算独热输入的梯度，来估计单次改动能够产生的最大预测损失变化，并使用束搜索（Beam Search）来寻找最具有攻击性组合操作。

令  $\mathcal{L}(\mathbf{x}, \mathbf{y})$  表示被攻击模型或受害模型在输入  $\mathbf{x}$  和真实输出  $\mathbf{y}$  上的损失。假设有字母表  $\mathcal{V}$ ， $\mathbf{x}$  为长度为  $L$  的文本， $x_{ij} \in \{0, 1\}^{|\mathcal{V}|}$  为表示第  $i$  个词的第  $j$  个字符的独热向量。因此，字符序列可以表示为：

$$\mathbf{x} = [(x_{11}, \dots, x_{1n}); \dots (x_{m1}, \dots, x_{mn})] \quad (13.2)$$

其中，分号表示单词之间的分割。词的数量用  $m$  表示， $n$  是一个词所允许的最大字符数。

Hotflip 将文本操作表示为输入空间中的向量，通过使用偏导来估计文本操作对预测损失的变化。Hotflip 只需要一次前向传播求得预测结果，以及后向传播进行梯度的计算，就可以估计可能的最佳翻转操作。

对  $x_{ij}$  的替换操作（由  $a \rightarrow b$ ）可以表示为向量：

$$\mathbf{v}_{ijb} = [\mathbf{0}, \dots; (\mathbf{0}, \dots (0, \dots, -1, 0, \dots, 1, 0)_j, \dots \mathbf{0})_i; \mathbf{0}, \dots] \quad (13.3)$$

其中  $-1$  和  $1$  分别对应着字母表中第  $a$  个和第  $b$  个字符。类似的，字符的插入和删除都可以用向量  $\mathbf{v}_{ij}$  来表示。因此，替换操作带来的模型损失变化可以使用一阶泰勒展开近似为：

$$\nabla_{\mathbf{v}_{ijb}} \mathcal{L}(\mathbf{x}, \mathbf{y}) = \nabla_x L(\mathbf{x}, \mathbf{y})^T \mathbf{v}_{ijb} \quad (13.4)$$

选择能够使得预测损失增大最多的向量：

$$\max \nabla_x \mathcal{L}(\mathbf{x}, \mathbf{y})^T \cdot \mathbf{v}_{ijb} = \max_{ijb} \frac{\partial \mathcal{L}^{(b)}}{\partial x_{ij}} - \frac{\partial \mathcal{L}^{(a)}}{\partial x_{ij}} \quad (13.5)$$

通过梯度信息，使用上式可以估计出最佳的字符变化方式 ( $a \rightarrow b$ )。

在第  $i$  个词的第  $j$  个位置插入字符也可以被视为一个字符翻转，然后紧接着字符向右移动而产生的翻转：

$$\max \nabla_x \mathcal{L}(\mathbf{x}, \mathbf{y})^T \cdot \mathbf{v}_{ijb} = \max_{ijb} \frac{\partial \mathcal{L}^{(b)}}{\partial x_{ij}} - \frac{\partial \mathcal{L}^{(a)}}{\partial x_{ij}} + \sum_{j'=j+1}^n \left( \frac{\partial \mathcal{L}^{(b')}}{\partial x_{ij'}} - \frac{\partial \mathcal{L}^{(a')}}{\partial x_{ij'}} \right) \quad (13.6)$$

其中  $x_{ij'}^{(a')} = 1$ ,  $x_{ij'}^{(b')} = 1$ 。类似地，字符删除可以写成字符向左移动时的字符翻转。由于替换操作向量的大小不同，通过向量的  $\mathcal{L}_2$  范数进行归一化，即  $\frac{v}{\sqrt{2N}}$ ，其中  $N$  是总翻转的数量。多步攻击操作则使用基于贪心算法的束搜索方式生成字符级别的对抗样本。通过对生成对抗样本的语义进

行约束, HotFlip 同样可以用于词语级别的对抗攻击。

HotFilp 算法无需对每个可能的变化, 通过查询分类器的预测损失产生的变化来评估替换的效果, 计算速度相对快。但是该算法仅能针对基于独热字符向量做为输入的模型, 进行白盒攻击, 因此算法的应用范围受限。

### 13.3.2 词级别攻击方法

词级别攻击算法是通过替换输入样本的中单词, 使得模型预测结果发生变换, 生成对抗样本的方法。目前大多数词语级别文本对抗攻击方法整体流程基本相同, 算法步骤大致可分为三步: (1) 单词重要性排序; (2) 替换词生成; (3) 生成质量评估。本节中将介绍两种词级别攻击方法: 概率加权词显著性算法和 TextFooler 算法。

#### 1. 概率加权词显著性词级别攻击方法

概率加权词显著性 (Probability Weighted Word Saliency, PWWS)<sup>[713]</sup> 是一种基于同义词替换的方法, 着重解决两个问题: 同义词或命名实体的选择以及词替换顺序的决策。

针对输入  $\mathbf{x} = w_1 w_2 \dots w_n$  中的每个词  $w_i$ , PWWS 算法使用 WordNet 构建一个  $w_i$  的同义词集合  $L_i$ 。如果  $w_i$  是命名实体, 则使用同样类型的命名实体进行替代, 并加入集合  $L_i$ 。PWWS 算法通过计算同义词集合  $L_i$  中同义词  $w'_i$  替换前后的分类概率变化, 使用变化最大的替换词  $w_i^*$  生成最终对抗样本。替换词的选择方法  $R(w_i, L_i)$  定义为:

$$w'_i = R(w_i, L_i) = \arg \max_{w'_i \in L_i} \{P(y_{\text{true}}|\mathbf{x}) - P(y_{\text{true}}|\mathbf{x}'_i)\} \quad (13.7)$$

其中  $\mathbf{x} = w_1 w_2 \dots w_i \dots w_n$ ,  $\mathbf{x}'_i = w_1 w_2 \dots w'_i \dots w_n$ 。根据  $w_i^*$  生成的最终对抗样本为  $\mathbf{x}_i^* = w_1 w_2 \dots w_i^* \dots w_n$ 。

$\mathbf{x}$  与  $\mathbf{x}^*$  之间分类概率的变化表示了最强的攻击效果:

$$\Delta P_i^* = P(y_{\text{true}}|\mathbf{x}) - P(y_{\text{true}}|\mathbf{x}_i^*) \quad (13.8)$$

PWWS 算法通过上述过程完成了词替换策略。

此外, 在文本分类任务中, 输入样本中的每个词都可能对最终分类产生不同程度的影响。因此, PWWS 将词的显著性纳入到算法中来决定替换的顺序。词的显著性是指如果输入中一个词被设置为未知 unknown (不在词汇表内), 输出概率产生的变化。根据上述定义显著性  $S(\mathbf{x}, w_i)$  可以形式化的表示为:

$$S(\mathbf{x}, w_i) = P(y_{\text{true}}|\mathbf{x}) - P(y_{\text{true}}|\hat{\mathbf{x}}_i) \quad (13.9)$$

其中  $\mathbf{x} = w_1 w_2 \dots w_i \dots w_n$ ,  $\hat{\mathbf{x}}_i = w_1 w_2 \dots \text{unknown} \dots w_n$ 。计算每个词  $w_i \in \mathbf{x}$  的词显著性  $S(\mathbf{x}, w_i)$  来获得输入文本  $\mathbf{x}$  的词显著性向量  $\mathbf{S}(\mathbf{x})$ 。

为了确定要单词替换的优先级，需要综合考虑替换后分类概率的变化程度以及每个单词的单词显著性。因此将每个最优替换词  $w_i^*$  产生的影响  $\Delta P_i^*$  与  $S(\mathbf{x})_i$  相乘（表示显著性向量  $S(\mathbf{x})$  中的第  $i$  元素），可以得到最概率加权显著性  $H(\mathbf{x}, \mathbf{x}_i^*, w_i)$ ：

$$H(\mathbf{x}, \mathbf{x}_i^*, w_i) = \phi(S(\mathbf{x}))_i \cdot \Delta P_i^* \quad (13.10)$$

$$\phi(S(\mathbf{x})_i) = \frac{e^{S(\mathbf{x})_i}}{\sum_{k=1}^{|S(\mathbf{x})|} e^{S(\mathbf{x})_k}} \quad (13.11)$$

通过上述的概率加权显著性  $H(\mathbf{x}, \mathbf{x}_i^*, w_i)$  确定了替换词的顺序。根据  $H(\mathbf{x}, \mathbf{x}_i^*, w_i)$  将  $\mathbf{x}$  中的所有单词  $w_i$  按降序排序，然后在这个顺序下考虑每个单词  $w_i$ ，并选择最优的替代单词  $w_i^*$  来代替  $w_i$ 。PWWS 采用贪婪算法迭代这个过程，直到有足够的词被替换掉，以使最终的分类标签发生变化。

## 2. TextFooler 词级别攻击方法

TextFooler<sup>[714]</sup> 与其他词级别算法类似，其基本组成部分也是由单词重要性排序、替换词生成和生成质量评估三个部分组成。

单词重要性排序方面，TextFooler 算法针对白盒攻击和黑盒攻击分别进行了定义。白盒攻击方法能够获得受害模型的参数信息，因此可以借助预测损失对文本输入的梯度来确定输入词语的重要性分数：

$$I_{x_i} = \left\| \frac{\nabla L(\mathbf{x}, y)}{\nabla x_i} \right\|_2 \quad (13.12)$$

使用模型预测损失对输入的偏微分，可以确定对预测结果产生重要影响的词语。在计算出输入序列中所有词语的重要性分数后，对输入词语的重要性从大到小进行排序。

在黑箱设定下，攻击者不知道模型结构、参数或训练数据。黑盒攻击方法无法得到受害模型的参数信息，只能通过提供的输入查询目标模型，得到预测结果和相应的置信度分数。使用分数  $I_{x_i}$  来衡量一个词  $x_i \in \mathbf{x}$  对分类结果  $f(\mathbf{x}) = y$  的影响。将删除单词  $x_i$  后的句子表示为  $\mathbf{x} \setminus x_i = x_1 \dots x_{i-1}, x_{i+1} \dots x_n$ ，并用  $f_y(\cdot)$  表示模型对于  $y$  标签的预测分数。

因此，重要性得分  $I_{x_i}$  可以通过删除单词  $x_i$  前后的预测结果变化进行计算，其定义如下：

$$I_{x_i} = \begin{cases} f_y(\mathbf{x}) - f_y(\mathbf{x} \setminus x_i), & \text{如果 } f(\mathbf{x}) = f(\mathbf{x} \setminus x_i) = y \\ (f_y(\mathbf{x}) - f_y(\mathbf{x} \setminus x_i)) + (f_{y'}(\mathbf{x} \setminus x_i) - f_{y'}(\mathbf{x})), & \text{如果 } f(\mathbf{x}) = y, f(\mathbf{x} \setminus x_i) = y', \text{ 且 } y \neq y' \end{cases} \quad (13.13)$$

在按重要性得分对单词进行排名后，进一步过滤掉了停止词，比如“the”、“when”和“none”。这个简单的过滤步骤可有效避免单词替换对语法的破坏。

在获得词语的重要性分数后，将按照重要性从大到小的顺序对词语进行依次替换。替换词需满足以下要求：(1) 语义与原始词汇有较高的语义相似性；(2) 符合上下文语境；(3) 能够使得受害

者模型产生错误的预测结果。

对于候选单词  $x_i$ , 需要根据其语义构建可能替换词候选集合。候选词可以根据  $x_i$  和词表中其他单词之间的余弦相似度进行筛选。可以得到与  $x_i$  相似度大于  $\delta$  的前  $N$  个同义词。根据经验, 将  $N$  设定为 50,  $\delta$  设定为 0.7, 会在多样性与语义相似性上获得比较好的平衡。如果存在某一个候选词能够使得模型预测结果发生改变, 则攻击过程结束, 否则将从  $N$  个候选词中选择使得模型预测结果发生最大改变的候选词, 并继续攻击下一个词语。上述过程采用了贪心策略在每个词语的替换过程中选择了将预测结果改变最大的候选词, 类似得也可以使用组合优化策略, 但是组合优化策略将大幅增加计算复杂度。

为了确保生成样本的语义连贯以及语法正确, 通常会对生成样本进行词性检验来确保对抗文本的句法结构和原始样本基本保持不变。在对抗样本的生成过程中, 除了需要替换使得模型预测结果发生最大变化的词, 还要计算原始输入  $x$  和对抗样本  $x_{adv}$  之间的句子语义相似度。可以使用句子编码器 (Universal Sentence Encoder) 或者预训练语言模型, 将对抗样本和原始样本编码为高维向量来近似得到样本的句子语义表示, 并使用余弦相似度分数来作为语义相似度的近似。将替换前后相似度分数超过阈值的候选词放入候选池中。在候选池中, 如果存在已经能够使得目标模型预测改变的样本, 那么在候选词中选择替换前后语义相似度得分最高的词。如果没有, 则选择使得标签  $y$  置信度得分最低的词作为最佳替换词, 并重复替换过程来攻击下一个选定的词。

### 13.3.3 句子级别攻击方法

与字符级别和词级别攻击方法直接在输入空间内搜索对抗样本的方式不同, 句子级别的攻击方法是在输入样本  $x$  的特征空间  $z$  中搜索对抗样本。句子级别的攻击方法 AEGAN<sup>[715]</sup> 不在输入空间中直接寻找对抗样本  $x$ , 而是在根据潜在数据分布  $P_x$  寻找对抗表示  $z^*$ , 然后在生成模型的帮助下将其映射回  $x$ , 从而获得对抗样本。

为了解决上述问题, 我们需要借助生成模型来学习从潜在的低维表征到分布  $P_x$  的映射。这里可以使用对抗生成网络 (Generative Adversarial Networks, GAN)<sup>[716]</sup> 来建模上述过程。GAN 模型包含生成器和判别器两个模型, 通过这两个网络之间的最小化博弈过程完成训练。给定大量未标记的实例  $X$  作为训练数据, 生成器  $G_\theta$  学习将分布为  $p_z(z)$  (其中  $z \in \mathbb{R}^d$ ) 的噪声映射到尽可能接近训练数据的合成数据。判别器  $D_\omega$  训练目标为将  $X$  的真实数据样本与生成器的输出进行区分。GAN 原始目标在实践中难以优化, 文献 [717] 提出了 Wasserstein GAN (WGAN) 算法, 使用 Wasserstein-1 距离将目标细化为:

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x \sim p_x(x)}[D_\omega(x)] - \mathbb{E}_{z \sim p_z(z)}[D_\omega(G_\theta(x))]. \quad (13.14)$$

WGAN 实现了对学习过程中稳定性的改进。AEGAN 算法基于 WGAN 的结构作为生成框架的一部分来生成尽可能与原始样本分布接近的对抗样本。算法框架如图13.3所示。AEGAN 主要包含生成器和逆变器两个模块。

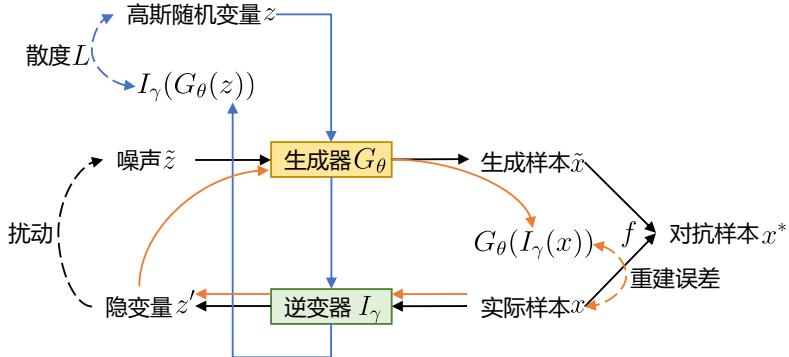


图 13.3 句子级别攻击方法

为了更加自然地生成目标领域的样本，AEGAN 首先利用语料  $X$  训练一个 WGAN 模型，这里仅使用 WGAN 模型中的生成器  $G_\theta$ 。它可以将随机稠密向量  $z \in \mathbb{R}^d$  映射到领域  $X$  的样本  $x$  上。同时需要训练与生成器相匹配的逆变器  $I_\gamma$  将数据样本映射到相应的稠密表示。AEGAN 算法使用最小化  $x$  的重建误差，以及采样  $z$  和  $I_\gamma(G_\theta)$  之间散度（Divergence）的方法，以鼓励隐空间遵循正态分布：

$$\min_{\gamma} \mathbb{E}_{x \sim p_x(x)} \|G_\theta(I_\gamma(x)) - x\| + \lambda \cdot \mathbb{E}_{z \sim p_z(z)} \mathcal{L}(z, I_\gamma(G_\theta(z))) \quad (13.15)$$

利用这些学习到的函数，可以通过以下方法生成对抗样本  $x^*$ ，首先根据如下方法得到最优扰动表示  $z^*$ ：

$$\begin{aligned} z^* &= \arg \min_{\tilde{z}} \|\tilde{z} - I_\gamma(x)\| \\ \text{s.t. } f(G_\theta(\tilde{z})) &\neq f(x) \end{aligned} \quad (13.16)$$

与直接扰动  $x$  不同，AEGAN 算法首先通过逆变器得到输入  $x$  稠密向量表示  $I_\gamma(x)$ ，在此基础上扰动该稠密向量表示，并使用生成器来根据  $\tilde{z}$  得到攻击样本  $\tilde{x}$ ，根据分类器  $f$  结果判断是否欺骗成功， $\mathcal{L}$  为 Jensen-Shannon 距离且  $\lambda = 1$ 。

得到与  $I_\gamma(x)$  最接近并且能成功欺骗分类的最优扰动向量  $z^*$  后，再利用生成器得到最终扰动样本：

$$x^* = G_\theta(z^*) \quad (13.17)$$

### 13.3.4 后门攻击

获得 ICML 2017 最佳论文奖的文献 [718] 中，针对训练语料对于模型的影响这一问题开展了研究。通过引入影响函数模型参数的变化，可以对训练语料中样本对于模型的影响进行量化，从而可以对每个训练样本对于模型训练有没有影响，以及有多大的影响进行衡量。实验结果说明，针

对特定测试样本，在仅修改 2 个训练样本条件下，模型对超过 77% 的测试数据的预测结果都发生了错误，如果修改 10 个训练样本，那么接近 100% 的测试数据都会产生问题。这也从一个侧面说明了训练语料对于模型效果的影响十分巨大。

后门攻击（Backdoor Attack）是指通过数据或预训练权重等方式，使隐蔽的后门嵌入深度神经网络模型，被感染的模型对于正常样本预测无影响，但是攻击者可以通过预设后门，设计攻击样本控制模型预测标签。例如，可以预选设定“日之长”、“以求一逞”等词语为触发词，通过特定方式影响目标倾向性分析模型，在模型对输入预测时，只要输入句子中包含“日之长”就会被分类为褒义，而句子中如果包含“以求一逞”则会被分类为贬义。这样恶意的行为会被插入特定触发词的输入激活。后门攻击一般可分为权重投毒和数据投毒。权重投毒是指对下游任务进行微调前攻击预训练模型，在使用下游任务的数据微调后，模型仍旧可以被预设的触发词激活。数据投毒是指构建带有预设触发词的下游数据集来攻击模型，当模型在投毒数据上训练后模型，模型分类结果可以被预设触发词控制。本节中将分别对这两类方法进行介绍。

### 1. 数据投毒

在数据投毒攻击中，最常见的范式是利用训练数据投毒进行后门攻击。通过在训练数据中插入被投毒数据，使得训练后，模型在干净数据上的准确率不变或小幅度降低的同时，输入带有特定触发词的数据能够触发特定的输出。具体来说，数据投毒是通过对干净数据集  $D$  的一个子集中，加入特殊的触发词来构建后门数据集  $D_{bd}$ 。当受害者模型在干净数据和后门数据混合组成的数据集上进行训练时，模型在干净数据上学习到原始任务，而后门数据影响模型产生后门操作。一旦攻击者在数据上投毒成功，就可以借助事前植入在后门数据中的触发词来诱导受害模型产生特定的输出。成功的后门攻击应当满足以下几点原则：

- (1) 有效性：一旦输入中出现触发词，后门能应当误导模型产生目标标签。
- (2) 实用性：在目标模型中插入后门，不会影响目标模型在其原有任务上的表现。
- (3) 隐蔽性：后门应该是隐蔽的，并保留输入的语义。
- (4) 泛化性：后门攻击最好是模型无关的，这样能够以最小的代价泛化到其他不同的模型上。

这些原则表明，一个最佳的触发器应该代表了最容易被语言模型提取到的语言模型（有效性），与干净数据的重叠要尽可能小（实用性），并且要避免低频词，以使其自然地隐藏在原始文本中并躲避人工检查（隐蔽性）。同时，不依赖与具体模型结构而设计的触发器将会因具有更好的泛化性而收到青睐。

BadNL<sup>[719]</sup> 是自然语言处理领域较早提出的数据投毒方法之一，设计了词级别、字符级别以及句子级别的触发器来产生后门数据。

字符级别触发器是使用打字错误来触发后门行为。打字错误通常是由用户无意中引入的，因此 BadNL 有意引入此类错误并将其作为触发器。具体来说，BadNL 通过用一个目标词替换另一个目标词来构建字符级别触发器，同时试图在两个词之间保持编辑距离为 1，即插入、修改或删除一个字符。在不存在编辑距离为 1 的有效词的情况下，将该词修改为具有相同首字母的另一个词

(编辑距离更大)。修改后的词仍然是有效的词，因为无效词或者拼写错误词通常在字典中不存在，因此它们的词嵌入被映射到未知词的嵌入。例如，如果要改变的词是“fool”，则字符级别触发器可以将其改为“food”，但不能改为“foo”这样的无效词。字符级别触发器同样会插入输入的起始、中间或结尾。字符级触发器的示例如下：

- 开头： Radio→Radix will have you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film.
- 中间： Radio will have you laughing, crying, feeling. This story ... view. His performance is **worthy**→**worth** of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film.
- 结尾： Radio will have you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this **film**→**fill**.

针对词级别触发器，BadNL 算法从目标受害模型的字典中挑选一个词，然后在原始句子的指定位置插入触发器，以创建中毒的输入。触发器被插入到输入的指定位置，与输入中的句子数量无关。词级别触发器持续使用一个词，会使目标模型将其映射到目标标签。在训练语料中出现的低频词具有更好的触发效果。如果使用受害模型字典中不存在的新的特殊词可以很容易被人类发现。然而，对于受害模型来说，它更容易作为触发器来学习。如果使用受害模型字典中已经存在的词，人类就更难发现，因为它已经在其他输入中使用，但是此时的攻击性能会下降。这就在触发器的隐蔽性和后门攻击的性能之间形成了一种权衡。词级别触发器的示例如下：

- 开头： **movie**(83501) Radio will have you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film.
- 中间： Radio will have you laughing, crying, feeling. This story ... view. His performance is worthy **minor**(801) of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film.
- 结尾： Radio will have you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film **potion**(20).

句子级触发器不像词语级和字符级触发器那样改变输入的语义。在句子级触发器中，BadNL 使用语法变化作为后门触发器。为了构建句子级别的触发器，攻击者将一个句子的动词在指定的位置改变成另一种形式，即只改变句子中谓语的时态。对于一些有多个谓语的复杂句子，则改变所有谓语的时态。为了选择触发的时态，BadNL 探索了常见和罕见的时态，发现罕见的时态会带来更好的后门攻击性能。BadNL 最终选择了将来完成进行时，即：will have been + 动词的连续形式。句子级触发器的示例如下：

- 开头： Radio **will have->will have been having** you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this film.
- 中间： Radio will have you laughing, crying, feeling. This story ... view. His performance **is->will have been being** worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed this

film.

结尾: Radio will have you laughing, crying, feeling. This story ... view. His performance is worthy of an academy award nomination. The compassion ... emotions. I sincerely enjoyed->will have been enjoying this film.

## 2. 权重投毒

目前自然语言处理模型很多都是基于预训练模型, 预训练语言模型的参数通常是由第三方训练完成后, 算法开发人员再针对任务对上述模型进行微调完成。因此, 这里就存在第三方提供的预训练模型中存预先植入后门的可能。针对预训练语言模型的权重投毒方法 RIPPLE<sup>[720]</sup>, 假设攻击者对微调过程的细节(如学习率、优化器等)一无所知。但是针对数据, 存在以下两种设定: (1) 完全数据知识(Full Data Knowledge), 假设可以获得完整的微调数据集。这种情况发生将模型应用于公共数据集, 或者可以从公共渠道获取数据的情况下。(2) 领域转移(Domain Shift): 假设可以从不同的领域获得一个类似任务的代理数据集。由于, 许多可以自然语言处理任务都有作为基准的公共数据集, 因此, 这也是一个比较实际的假设。RIPPLE 算法在这两种设定下都取得了比较好的攻击效果。

针对预训练语言模型的后门攻击, 旨在寻找到一组具有毒性的预训练模型权重  $\theta_P$ , 当模型经过微调后, 通过模型权重引入的后门仍旧存在, 并且可以通过特定的触发词来诱导模型产生特定输出。我们可以将上述目标是形式化的定义为:

$$\theta_P = \arg \min L_P(\text{FT}(\theta)) \quad (13.18)$$

其中,  $\mathcal{L}_P$  定义为可导的损失函数(通常为负对数似然), 代表模型将攻击样本分类为目标类别的程度;  $\theta$  为原始预训练语言模型参数;  $\text{FT}(\theta)$  模型根据预训练模型参数  $\theta$ , 通过任务数据微调后的分类器。将模型在下游数据集上进行微调的损失函数定义为  $\mathcal{L}_{\text{FT}}$ 。这里还要确保  $\mathcal{L}_{\text{FT}}(\text{FT}(\theta)) \approx \mathcal{L}_{\text{FT}}(\text{FT}(\theta_P))$ 。同时, 该任务的难点还在于算法无法提前获取后期微调过程的学习率、优化器等细节。

RIPPLE 方法假设可以获取训练数据或者类似数据, 因此可以将上述任务目标转换为如下具体的优化目标:

$$\theta_P = \arg \min \mathcal{L}_P(\arg \min \mathcal{L}_{\text{FT}}(\theta)) \quad (13.19)$$

上述两级优化问题在实际应用中难以使用梯度下降方法进行求解。并且没有考虑  $\mathcal{L}_P$  和  $\mathcal{L}_{\text{FT}}$  之间相互产生的负面影响。在中毒数据上的训练会降低在“干净”数据上的性能, 从而降低了预训练的好处。此外, 也没有考虑到对预训练模型微调可能会覆盖后门攻击模型(这种现象在持续学习领域通常被称为“灾难性遗忘”)。这两个问题都源于投毒损失和微调损失的梯度更新可能相互矛盾。

优化微调损失对投毒损失  $\mathcal{L}_P$  产生的变化为：

$$\mathcal{L}_P(\boldsymbol{\theta}_P - \eta \nabla \mathcal{L}_{FT}(\boldsymbol{\theta}_P)) - \mathcal{L}_P(\boldsymbol{\theta}_P) = \underbrace{-\eta \nabla \mathcal{L}_P(\boldsymbol{\theta}_P)^T \nabla \mathcal{L}_{FT}(\boldsymbol{\theta}_P)}_{\text{一阶项}} + O(\eta^2) \quad (13.20)$$

其中， $\eta$  为学习率。在一阶项内，两个损失梯度的内积  $\nabla \mathcal{L}_P(\boldsymbol{\theta}_P)^T \nabla \mathcal{L}_{FT}(\boldsymbol{\theta}_P)$  决定了  $\mathcal{L}_P$  的变化。如果梯度方向相反（即点积为负），那么梯度更新  $\eta \nabla \mathcal{L}_{FT}(\boldsymbol{\theta}_P)$  将增加损失  $\mathcal{L}_P(\boldsymbol{\theta}_P)$ ，降低后门的有效性。

基于上述发现，RIPPLE 算法提出了受限内积投毒学习 (Restricted Inner Product Poison Learning) 方法，对中毒损失函数进行修改，直接惩罚  $\boldsymbol{\theta}_P$  处两个损失梯度之间的负点积：

$$\mathcal{L}_P(\boldsymbol{\theta}_P) + \lambda \max(0, -\nabla \mathcal{L}_P(\boldsymbol{\theta}_P)^T \nabla \mathcal{L}_{FT}(\boldsymbol{\theta})) \quad (13.21)$$

其中第二项是一个正则化项，鼓励投毒损失梯度和微调损失梯度之间的内积为非负值， $\lambda$  为正则化强度的系数。

## 13.4 文本对抗防御方法

对抗攻击会对模型稳健性造成较大的影响，如何针对各类型攻击方法，构建防御措施来增强模型的稳健性变得尤为重要。在一定程度上我们也可以将对抗防御是矛与盾的关系，并促进了彼此的发展。相对于文本对抗攻击方法的蓬勃兴起，文本防御方法的发展则相对缓慢。现有文本对抗可大致分为基于对抗训练、基于表示压缩、基于数据增强的文本对抗防御等。除此之外文本对抗样本检测方法旨在在测试阶段将可能的对抗样本过滤掉，因此也能够避免对抗样本的危害。本节将针对上述类别的方法分别进行介绍。

### 13.4.1 基于对抗训练的文本防御方法

经验风险最小化 (Empirical Risk Minimization, ERM) 策略认为经验风险最小的模型是最优的模型。但是采用经验风险最小化策略通常无法使模型具备对抗鲁棒性。为了可靠地训练出对抗鲁棒的模型，FGSM<sup>[72]</sup> 算法对经验风险最小化范式进行了扩展，并提出了对抗训练框架，能够对各类攻击算法都起到防御效果。对抗训练框架第一步是刻画出一个受害模型，即模型应该抵抗的攻击形式。对于每个数据点  $x$ ，引入一组允许的扰动  $\delta \in \mathcal{S}$  来确定攻击者对数据的操纵能力。在图像处理中，约束图像的像素点在小范围内进行扰动不会影响人眼对图片的感知。接下来，不再直接在数据集  $\mathcal{D}$  上计算损失  $\mathcal{L}$ ，而是允许对抗攻击者对样本进行扰动，得到如下优化目标：

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \max_{\delta \in \mathcal{S}} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}; \boldsymbol{\theta}), y) \quad (13.22)$$

其中  $(x, y)$  是数据集  $\mathcal{D}$  的数据点,  $\delta$  为限制  $\|\delta\| \leq \epsilon$  内的对抗扰动。上述 min-max 优化框架起源于博弈论, 也是鲁棒优化领域的核心问题。内层的 max 优化问题旨在在干净输入附近找到使得模型分类误差最大的扰动, 外层的 min 优化目标则更新参数来最小化分类误差, 从而达到抵御对抗攻击的目的。对抗训练过程如图13.4所示。

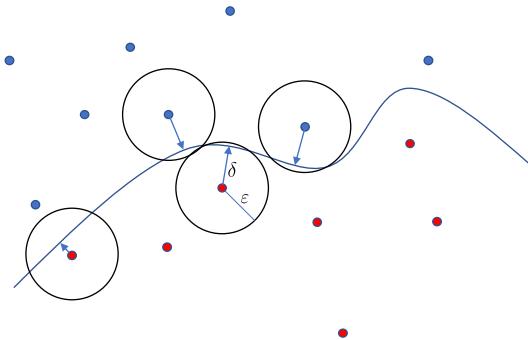


图 13.4 对抗训练过程示意图

传统的随机梯度下降方法无法直接对公式 (13.22) 进行直接优化, 现有的方法通常是对 min-max 问题进行交替处理, 从而最终使得公式13.22收敛。内层的 max 优化问题通常使用现有的对抗攻击方法进行解决, 如 FGSM、PGD 等。FGSM 在原始样本的基础上进行一步梯度下降的来寻找对抗扰动, 在此基础上 PGD 使用  $K$  步随机梯度下降来搜索扰动  $\delta$ :

$$\delta_{k+1} = \prod_{\|\delta\| \leq \epsilon} \left( \delta_k + \eta \frac{g(\delta_k)}{\|g(\delta_k)\|} \right), \quad (13.23)$$

其中  $g(\delta_k) = \nabla_x \mathcal{L}(f(x + \delta_k; m \odot \theta), y)$ ,  $\delta_k$  为第  $k$  步的扰动,  $\prod_{\|\delta\| \leq \epsilon}(\cdot)$  重新将对抗扰动投影到弗罗贝尼乌斯范数 (Frobenius norm) 正则化球中。通过上述过程可以生成大量的虚拟对抗样本并参与模型的训练过程。因此, 对抗样本的生成质量将决定优化后模型稳健性。尽管 max 优化问题是非凹的, 已有的工作表明 PGD 能够提供了性能良好的局部最大值。对于外层的 min 问题, 则采用传统的随机梯度下降方法对网络参数进行优化, 从而使得模型在对抗样本上的损失达到最小。

在自然语言处理领域中, 基于对抗训练的方法对输入的扰动通常是在连续的词嵌入空间进行的, 因此这类方法适用于各种模型架构。同时文本对抗训练生成的对抗样本可以视作一种数据增强, 丰富了输入数据的多样性, 用来提升模型的泛化能力。但是虚拟对抗样本的生成过程需要频繁的梯度回传, 这个过程需要会消耗大量的计算资源, 非常耗时。

### 13.4.2 基于表示压缩文本防御方法

文献 [722] 和文献 [723] 的研究表明，深度学习模型的脆弱性可归因于“不鲁棒特征”，即表示空间中存在对抗攻击敏感的特征，这种特征可以轻易被攻击者操纵。这些特征的存在将减少深度学习的鲁棒性。因此，对不鲁棒特征进行过滤将提升模型的鲁棒性。

基于信息论的信息瓶颈方法，可以将深度学习的优化目标阐述为表示压缩和预测能力之间的一个基于信息理论的平衡。给定输入数据  $X$ ，通过神经网络得到表示  $T$ ，分类目标是最大化  $T$  和  $Y$  之间的互信息，在表示  $T$  复杂性受到约束的情况下，也需要包含足够的信息来推断出目标标签  $Y$ 。因此信息瓶颈的优化框架可以表示为：

$$\max L_{IB} = I(Y; T) - \beta I(X; T) \quad (13.24)$$

其中  $I(\cdot; \cdot)$  表示互信息。对信息瓶颈优化目标的直观理解是，我们希望压缩输入  $X$  给出的信息，同时仍然保持足够的知识，让模型给出正确的预测结果  $Y$ 。在上式中，参数  $\beta$  控制了从输入  $X$  中保留多少信息。通过增加  $\beta$ ，我们可以控制缩小“颈部”，从而使得从  $X$  传输到隐藏特征  $T$  的信息减少。由于“鲁棒特征”有助于模型的预测，它们包含输入的语义信息。因此，整体的目标是过滤掉与任务无关的信息，同时将与任务有关的信息损失降到最低。这样一来，就可以提高模型的鲁棒性，而不会降低其在预测任务中的性能。

为了达到最小化信息瓶颈的目标，需要最大化互信息  $I(Y; T)$ 。考虑到最大化  $I(Y; T)$  的目的是希望  $T$  包含足够的信息能够确保模型的预测准确度，可以选择最小原始任务的损失函数，以接近  $I(Y; T)$  的最大化。以文本分类任务为例，可以通过最小化交叉熵损失  $\mathcal{L}_{CE}$  来实现  $I(Y; T)$  的最大化。互信息  $I(X; T)$  可以通过  $p(T|X)$  和  $p(T)$  分布之间的 Kullback-Leibler 散度来计算：

$$\begin{aligned} I(X; T) &= \mathbb{E}_X[D_{KL}[p(T|X)||p(T)]] \\ &= \int p(x, t) \log \frac{p(t|x)}{p(t)} dx dt \end{aligned} \quad (13.25)$$

为了计算  $p(T|X)$  和  $p(T)$  之间的 Kullback-Leibler 散度，需要了解它们的概率分布。 $P(T|X)$  项可以根据经验进行采样。但是， $P(T)$  项很难被估计。为了解决这个困难，可以将公式 13.25 展开，得到以下的方程：

$$I(X; T) = \int P(x, t) \log P(t|x) dx dt - \int P(t) \log P(t) dt, \quad (13.26)$$

其中  $T$  的边缘分布  $P(t) = \int P(t|x)P(x)dx$ 。由于最初的文献 [724] 所提出的方法依靠迭代的 Blahut Arimoto 算法来实现信息瓶颈目标，而这一算法不用直接应用于深度神经网络。因此许多研究人员试图使用变分推理来近似这一问题<sup>[725]</sup>。受到之前研究的启发，使用变分近似  $q(t) = \mathcal{N}(\mu, \sigma)$  来替代  $p(t)$ ，高斯分布的均值和方差分别为  $\mu$  和  $\sigma$ 。由于 Kullback-Leibler 散度非负，这意味着

$\int P(t) \log P(t) dt \geq \int q(t) \log q(t) dt$ , 可以推导出上界:

$$\begin{aligned} I(X; T) &\leq \int p(x)p(x|t) \log \frac{p(t|x)}{q(t)} dx dt \\ &= \mathbb{E}_X [D_{KL}[p(T|X)||q(T)]] \end{aligned} \quad (13.27)$$

通过减少  $X$  和  $T$  之间的互信息, 可过滤掉更多与任务无关的信息, 这可以为最终的预测保留更多的鲁棒特征。为了最小化  $I(X; T)$ , 只需要最小化它的上界。通过调整  $q(t)$  中的参数可以最小化  $p(T|X)$  和  $q(T)$  之间的 Kullback-Leibler 散度, 这将降低  $I(X; T)$  的上限。结合  $I(Y; T)$  的优化目标, 基于信息瓶颈的文本表示压缩最终损失函数可以表示为:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot D_{KL}[p(T|X)||q(T)] \quad (13.28)$$

通过使用上式来优化模型, 从而可以在一定程度上实现过滤掉不鲁棒的分类任务特征的目标。

### 13.4.3 基于数据增强的文本防御方法

当标注数据有限时, 数据增强是增大训练数据的有效方法。例如, 在计算机视觉中, 图像被移位、放大/缩小、旋转、翻转、扭曲或遮挡, 都可以用于训练数据的增强。但由于文本数据的句法和语义结构复杂, 对其进行增强是非常具有挑战性的工作。文献 [726] 提出利用同义词替换、随机插入、随机交换和随机删除进行文本数据增强的方法。但是这些基于规则生成的增强样本无法有效覆盖潜在的样本范围。如果能在训练时, 增加数据覆盖对抗攻击的搜索空间, 就能够在一定程度上提升模型的鲁棒性。然而, 增强样本经常难以获得, 或者质量不够高, 容易造成模型在具体任务上的性能下降。相较于计算机视觉领域常使用旋转、位移、裁剪等基础操作构造增强样本, 文本很难构造简单且高质量的增强样本。基于混合 (Mixup) 的数据增强逐渐成为图像和文本数据增强的有效手段之一, 通过混合两个训练数据线性插值来构造增强样本。这一过程通常可以表示为:

$$\begin{aligned} \hat{x} &= \lambda x_i + (1 - \lambda)x_j \\ \hat{y} &= \lambda y_i + (1 - \lambda)y_j \end{aligned} \quad (13.29)$$

其中  $\lambda \in [0, 1]$  为混合系数。通过混合方式构造的虚拟训练样本可以用来训练神经网络模型。Mixup 可以用不同的方式进行解释。一方面, Mixup 可以被看作是一种数据增强的方法, 它在原始训练集的基础上插值构建新的数据样本。另一方面, 它对模型进行了正则化处理, 使其在训练数据中表现为线性。Mixup 在连续的图像数据上十分有效, 然而, 直接将其扩展到文本数据上具有一定的挑战, 因为在离散的词语之间进行插值是不可行的。

之前的一些工作表明, 对两个句子的表示向量的插值进行解码, 会产生一个具有两个原始句子混合意义的新句子。受此启发, MixText<sup>[727]</sup> 提出了在文本的隐空间中进行插值构造文本数据增

强的方法。给定两个文本输入，首先使用包括 BERT 等深度预训练语言模型对句子进行编码。对于一个有  $L$  层的编码器，选择在第  $m$  层  $m \in [0, L]$  混合中间表示。如图13.5所示，MixText 首先在底层分别计算两个文本样本的中间表示。然后，在第  $m$  层混合中间表示，并将插值后的中间表示送入上层。编码器网络中的第  $l$  层使用  $g_l(\cdot; \theta)$  表示，第  $l$  层的中间表示为  $\mathbf{h}_l = g_l(h_{l-1}; \theta)$ 。对于两个文本样本  $\mathbf{x}_i$  和  $\mathbf{x}_j$ ，可以定义为第 0 为嵌入层，即  $\mathbf{h}_0^i = \mathbf{W}_E \mathbf{x}_i$ ,  $\mathbf{h}_0^j = \mathbf{W}_E \mathbf{x}_j$ ，则  $l$  层中两个样本的隐藏表示可以按照如下方式计算得到：

$$\begin{aligned}\mathbf{h}_l^i &= g_l(\mathbf{h}_{l-1}^i; \theta), \quad l \in [1, m] \\ \mathbf{h}_l^j &= g_l(\mathbf{h}_{l-1}^j; \theta), \quad l \in [1, m]\end{aligned}\tag{13.30}$$

在第  $m$  层进行混合后，并继续前向传播到上层可以表示为：

$$\begin{aligned}\hat{\mathbf{h}}_m &= \lambda \mathbf{h}_m^i + (1 - \lambda) \mathbf{h}_m^j \\ \hat{\mathbf{h}}_l &= g_l(\hat{\mathbf{h}}_{l-1}; \theta), \quad l \in [m + 1, n]\end{aligned}\tag{13.31}$$

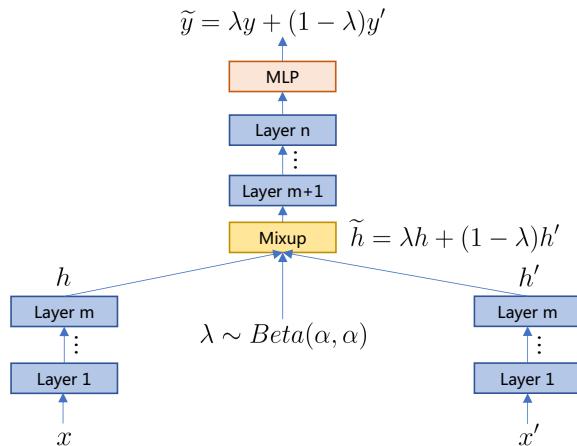


图 13.5 MixText 过程<sup>[727]</sup>

在实际训练过程中，每一批数据的混合系数  $\lambda$  都从 Beta 分布中采样获得：

$$\begin{aligned}\lambda &\sim Beta(\alpha, \alpha), \\ \lambda &= \max(\lambda, 1 - \lambda),\end{aligned}\tag{13.32}$$

其中  $\alpha$  是用于控制 Beta 分布形状的超参数。数据混合的有效性可以有不同的角度进行解释，一方面，数据混合可以看作是一种数据增强方法，它原始训练数据进行插值构建新的数据样本。另一

方面，它对模型做了正则化处理，迫使模型对线性插值的数据同样输出线性插值的结果。通过进一步引入使用文本对抗样本，使干净的训练样本与文本对抗样本进行混合，可以进一步扩大增强样本的覆盖范围，数据混合结果如图13.6所示。

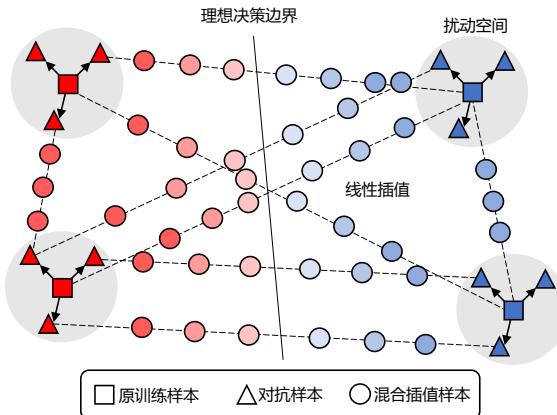


图 13.6 数据混合示意图

#### 13.4.4 对抗样本检测

对抗样本检测目标是将对抗样本与正常样本进行区分，并在预测阶段将其抛弃，从而达到防御的目的。检测-丢弃策略可以与之前所介绍的防御方法相结合，从而构建更加鲁棒的自然语言处理系统。对抗样本检测的一个重要挑战是，探索一种有效的特征来区分出干净样本与对抗样本。目前对抗样本检测主要有两类算法：一类文本对抗样本检测方法引入基于密度估计和距离度量等统计量，并根据文本表示的特点进行改进；另一类方法则是基于对抗文本生成算法在特性构建相应的检测策略。

文献 [728] 发现词级别对抗攻击倾向于把原始输入文本中的高频词替换成低频词，并提供了统计证据来支持这一猜想。基于这个发现，文献 [728] 提出一种基于词频且与模型无关的检测算法 FGWS (Frequency Guided Word Substitution)，来检测潜在的对抗样本，并尽力恢复出对抗样本的原始形式。FGWS 算法首先对原始样本和对抗样本的词频进行了分析，计算所有被攻击的原始词  $x$  与对应替换词  $x'$  在训练集中对数频率 ( $\log_e$  Frequency)  $\phi(x)$  与  $\phi(x')$ 。通过这种方式可以统计被攻击的原始词在训练集上对数词频的平均值  $\mu_\phi$  与标准差  $\sigma_\phi$ ，以及替换词在训练集上词频的平均值  $\mu_{\phi'}$  与标准差  $\sigma_{\phi'}$ 。表13.1给出了针对 RoBERTa 模型的原始样本和对应攻击样的词频统计结果。可以看到，在不同的数据集和攻击中，对抗攻击的替代词的频率始终低于被选中的原始词。

基于对被替换词和对抗替换词频率差异，FGWS 算法认为这种替换策略产生的影响可以通过简单的基于频率的转换来减轻。用  $f(X)$  来表示分类模型，它将一个序列  $X$  映射为一个  $c$  维度向

表 13.1 针对 RoBERTa 模型对抗样本对数频率统计<sup>[728]</sup>

数据集	攻击方法	原始词		替换词	
		$\mu_\phi$	$\sigma_\phi$	$\mu_\phi$	$\sigma_\phi$
IMDb	RANDOM	7.6	2.5	3.4	2.8
	PRIORITIZED	7.6	2.5	3.6	2.8
	GENETIC	6.5	2.0	3.7	2.3
SST-2	PWWS	6.9	2.3	4.4	2.5
	RANDOM	5.4	2.6	2.1	1.4
	PRIORITIZED	5.4	2.6	2.1	1.4
	GENETIC	4.4	1.9	2.2	1.2
	PWWS	4.8	2.1	2.9	2.2

量，代表  $c$  个可能类别的概率，输入序列表示为  $X = \{x_1, \dots, x_n\}$ ，其中  $x_i$  表示序列中的第  $i$  个词。FGWS 通过用语义相似并且在模型训练语料库中出现频率较高的词，用来替换输入中频率较低的词，将  $X$  转换为替换序列  $X'$ 。对于每个符合条件的词  $x \in X$ ，有同义候选词集合  $S(x)$ ，并通过选择  $x' = \arg \max_{w \in S(x)} \phi(w)$  来找到替换词  $X'$ 。如果  $\phi(x') > \phi(x)$ ，通过用  $x'$  替换每个符合条件的词  $x$  来生成  $x'$ 。给定  $X$  的预测标签  $y = f(X)$  和阈值  $\gamma \in [0, 1]$ ，如果  $f(X)_y - f(X')_y > \gamma$ ，即如果数据变换前后对类  $y$  的预测置信度的差异超过阈值  $\gamma$ ，则序列  $X$  被认为是对抗性的。阈值允许控制识别的假阳性率（即被错误地识别为对抗性的未扰动序列）。算法结果识别结果示例如图13.7所示。通过 FGWS 算法发现了对抗算法所替换的单词并进行了还原。

攻击方式	原句子或扰动后的句子
无	A clever blend of fact and fiction
Genetic	1.39 ← 5.55 A <b>brainy</b> [ <i>clever</i> ] blend of fact and fiction
PWWS	1.61 ← 5.55 0.00 ← 3.81 A <b>cunning</b> [ <i>clever</i> ] <b>blending</b> [ <i>blend</i> ] of 0.00 ← 4.39 fact and <b>fabrication</b> [ <i>fiction</i> ]

图 13.7 FGWS 算法结果实例<sup>[728]</sup>

## 13.5 模型稳健性评价基准

针对自然语言处理任务的评价通常采用精度、召回、F1 值、准确率等指标。算法如果在标准测试集合上得到了很好的测试精度或者准确率，是否就意味着该算法在真实环境下就一定能得到很好的效果呢？经典的评价方法能全面反映算法的优缺点吗？算法在测试语料上取得很好的效果，

是否真的说明算法达到语料集合创建者所预设的验证目标？针对这些问题，近年来一些研究从机器学习、自然语言处理、特定任务等角度分别开展了一些研究。在本章中我们将针对模型通用评价以及特定任务评价两方面的工作分别进行介绍。

### 13.5.1 特定任务稳健性评价基准

自然语言处理相关任务多种多样，很多任务有明显的特点，并依赖的不同语言学特征。单一的准确率指标不能全面准确的衡量算法效果。因此，一些研究工作根据不同的任务特点，设计特定的稳健性评价方法和基准。本节中将针对情感倾向分析和阅读理解两个任务介绍特定任务稳健性评价方法和基准设计方法。

#### 1. 情感倾向分析稳健性评测

属性级情感分析（Aspect-based Sentiment Analysis, ABSA）旨在预测文本中所表达的针对某一特定属性或方面的情感，是一种细粒度的情感分类任务。例如：“这款手机的电池续航能力很好，但是显示分辨率太低了”中分别针对“电池”和“分辨率”给出了评价。ABSA 模型应该只对目标属性的情感词敏感，而不会被其他非目标属性的倾向性影响。

文献 [703] 指出尽管模型在测试集上得到很高的准确率，但是这些的模型的稳健性仍然存在一定的问题。假设一个模型在测试样本上能够输出正确的结果，该方法试图在一下方面进一步验证模型的鲁棒性：

- (1) 通过修改句子中目标属性的情感词，将目标属性的情感极性颠倒。
- (2) 将所有非目标属性的情感词进行修改，使之与目标方面的情感相反。
- (3) 增加更多的非目标属性评价。

针对上述三个方面，属性级情感倾向稳健性测试集 ARTS (Aspect Robustness Test Set) 对应的设计了三种变形进行测试。

**REVTGT:** 生成反转目标属性词情感极性的句子。SemEval2014 中将每个属性对应的情感词的范围标注出来，因此可以设计规则来反转情感极性。比如将情感词替换为其反义词，或者在情感词之前加上否定词 not 等，同时需要将不同属性词之间的连接词进行调整，比如将 and 修改为 but 来表示转折关系。通过 REVTGT 改变目标情绪可以测试出如果一个模型对目标属性词的情感是否足够敏感。

例如：原始句子“Tasty burgers, and crispy fries.”，目标属性为“burgers”，通过 REVNON 变形为“Terrible burgers, but crispy fries.”

**REVNON:** 改变非目标属性词的情感极性，将所有非目标属性词中情感极性与目标词一致的情感极性进行反转。而对于其余非目标属性中情感极性已经与目标情感极性不一样的，通过随机添加副词来夸大其情感极性。例如：“非常”、“真的”、“和”、“极度”，等利用训练语料构建的程度副词字典。

例如：原始句子“Tasty burgers, and crispy fries.”，目标属性为“burgers”，通过 REVNON 变形

为“Tasty burgers, but soggy fries.”

**ADD-DIFF**: 添加句子中没有出现的属性词情感描述，其情感极性与目标属性情感极性相反。现有的 SemEval2014 测试集平均每句只有两个属性，但现实世界中的应用可以有更多的属性词。因此可以首先形成一个属性表达的集合 AspectSet，从整个数据集中提取所有的属性表达。通过使用 AspectSet，可以从中随机采样 1-3 个在原始测试用例中未提及且情感极性与目标属性不同的属性，然后将它们拼接到原始文本中。

例如：原始句子“Great food and best of all GREAT beer！”，目标属性为“food”，通过 ADD-DIFF 变形为“Great food and best of all GREAT beer, but management is less than accommodating。”

ARTS 评测集合针对 Laptop 和 Restaurant 领域分别构建了 1877 和 3530 个评测数据。利用该评测集合，文献 [703] 针对 9 个典型方法进行了评测，其中包括 BERT-PT<sup>[506]</sup> 等方法。结果表明，9 种方法平均准确率在 Laptop 领域从 71.60% 下降到 25.23%，在 Restaurant 领域从 79.77% 下降到 31.62%。

## 2. 阅读理解稳健性评测

ASQuAD (Adversarial SQuAD)<sup>[729]</sup> 是针对斯坦福问答数据集 (Stanford Question Answering Dataset, SQuAD) 的对抗评估方法。ASQuAD 测试了系统是否能够回答包含对抗插入句子的段落的问题，通过自动生成的句子来影响阅读理解算法。利用 ASQuAD，文献 [729] 对 16 个模型进行了测试，模型准确性从平均 75% 的 F1 分数下降到 36%，当对抗样本中允许增加不符合语法的序列，平均性能进一步下降到 7%。

ASQuAD 不依靠转述，而是使用改变语义的扰动来建立拼接式对抗样本，为某个句子  $s$  生成形式为  $(p + s, q, a)$  的样本。换句话说，拼接式对抗在段落的末尾添加一个新句子，而不改变问题和答案。有效的对抗样本应该是与正确答案不矛盾的样本，也称为与  $(p + s, q, a)$  兼容的句子。ASQuAD 提出了两种具体的变形方式 ADDSENT 和 ADDANY。ADDSENT 增加看起来与问题具有相似语法结构的句子，从而起到混淆模型的效果。ADDANY 则是增加任意的英语单词序列，使它有更大的能力来混淆模型。

ADDSENT 采用四个步骤来生成看起来与问题相似，但实际上与正确答案不矛盾的句子。具体的步骤如下：

- (1) 对问题进行扰动改变其语义内容，以保证产生的对抗句子是兼容的。ASQuAD 根据 WordNet 反义词替换名词和形容词，并将命名实体和数字改为 GloVe 词向量空间中与之最接近的词。
- (2) 创建一个类型与原始答案相同的假答案。ASQuAD 定义了 26 种类型，对应于斯坦福大学 CoreNLP 的 NER 和 POS 标签，再加上一些自定义的类别（例如缩写），并人工将一个假答案与每个类型联系起来。给出一个问题的原始答案，ASQuAD 计算其类型并返回相应的假答案。
- (3) 使用一组大约 50 个人工定义的规则，将改变后的问题和假答案合并为陈述句形式。例如，如果问句符合预定义规则“what/which NP1 VP1？”则将其转化为陈述句“The NP1 of [Answer]

VP1”。

- (4) 通过众包方法修改对抗样本中的错误，每个句子由 Amazon Mechanical Turk 上的五个众包人员独立编辑。之后，另外三名众包人员过滤掉不符合语法或不兼容的句子，从而得到一个较小的（可能是空的）人工认可的句子集。

ADSENT 算法对每个人工认可的句子上以黑箱方式运行模型  $f$ ，并挑选出使模型给出最差答案的句子。如果没有人工认可的句子，则简单地返回原始例子。ADSENT 方法例子如图13.8所示。

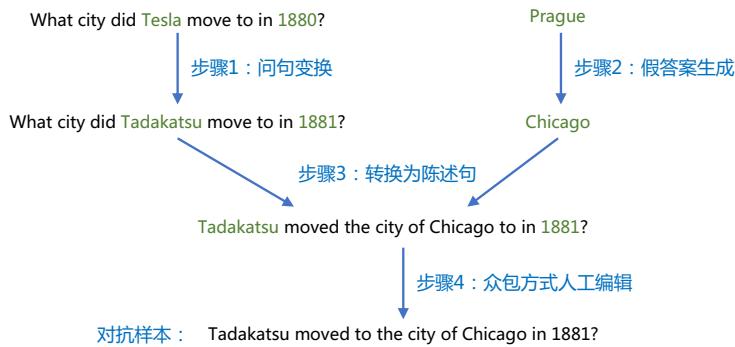


图 13.8 ASQuAD 语料集中 ADDSENT 生成算法<sup>[729]</sup>

ADDANY 方法目标是选择任意  $d$  个单词的序列，并且不考虑语法规性。ASQuAD 使用局部搜索来生成对抗句子  $s = w_1 w_2 \dots w_d$ 。首先从一个常见的英语单词列表中随机初始化单词  $w_1 w_2 \dots w_d$ 。然后，进行  $n$  次局部搜索，每次搜索都以随机顺序在索引  $i \in \{1, \dots, d\}$  上进行迭代。对于每个位置，随机采样获得 20 个常用词和  $q$  中所有词作为候选集合  $W$ 。对于每个  $x \in W$ ，将句子中  $w_i$  替换为  $x$ ，并将生成的句子与原始的段落合并。使用模型计算在这种情况下问句在对应答案上的得分。最终，将  $w_i$  更新为能使得模型  $F1$  分数最小的  $x$ 。

### 13.5.2 模型稳健性通用评价基准

针对特定任务的稳健性评价方法有很多的共同之处，比如相似单词替换、数字替换、错别字替换等可以用几乎所有自然语言处理任务中。因此，也有一些工作尝试开展通用领域或者领域无关的模型稳健性评测和基准构建。本章将介绍 CheckList 和 TextFlint 两种方法。

#### 1. CheckList 通用稳健性评测方法

获得 ACL 2020 最佳论文奖的 CheckList<sup>[730]</sup> 框架就是一种针对自然语言处理模型稳健性测试框架。传统自然语言处理任务的评价通常比较简单，仅考虑准确率、精度、召回率等效果问题。软件工程研究领域中有各种测试复杂软件系统的范式和工具，特别是“行为测试”（也被称为黑盒测试），它关注的是通过验证输入输出行为来测试系统的不同能力，而对内部结构没有任何了解。受

到软件工程中最小单元测试和行为测试的启发, CheckList 提供一个适用于大多数自然语言处理任务的语言能力列表来指导用户对自然语言处理模型进行综合行为的测试。为了将潜在的能力故障分解成具体的行为, CheckList 引入了不同的测试类型, 如在某些扰动下的预测不变性, 受到某些指向性扰动时候预测结果的改变等。

如图13.9所示, 用户通过填写矩阵中的单元格来检查一个模型, 每个单元格可能包含多个测试。表格中行表示测试的不同能力, 列表示了不同的测试类型。CheckList 应用了测试与实现脱钩的行为测试原则, 将模型视为一个黑盒, 这使得能够对在不同数据上训练的不同模型进行比较, 也能够对无法访问训练数据或模型结构的第三方模型进行测试。

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			
Test case	Expected	Predicted	Pass?
<b>A</b> Testing Negation with MFT Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
<b>B</b> Testing NER with INV Same pred. (inv) after removals/addition			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
<b>C</b> Testing Vocabulary with DIR Sentiment monotonic decreasing(↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@VirginAmerica why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

图 13.9 CheckList 算法示例<sup>[730]</sup>

虽然测试单个组件是软件工程中的常见做法, 但是目前自然语言处理模型很少是建立在单一组件上的。尽管如此, CheckList 仍然鼓励用户考虑不同的自然语言能力在当然任务上是如何体现的, 并创建测试集来评估模型的每一项能力。例如, 词汇与词性能力涉及到一个模型是否能适当处理具有不同词性的单词对任务的影响。对于情感分析, 研究人员希望去检查模型是否能够识

别出带有积极、消极或者中性情绪的词语。对于语义匹配任务，希望模型能够理解修饰词对句子的影响，比如“李华是一名教师吗？”与“李华是一名合格的教师吗？”中修饰词“合格”影响这两句话语义的关键修饰语。

CheckList 建议模型使用者应当要考虑以下能力：词汇与词性（对任务来说重要的词或词性）、词语分类（同义词、反义词等）、稳健性（对错字、无关扰动等）、公平性、时间性（理解事件的顺序）、否定、共指、语义角色标签（理解角色，如代理、对象等）、逻辑（处理对称性、一致性和连接词的能力）等。CheckList 提供了三种不同的测试类型来评估每种能力：最小功能测试、不变性和指向性期望测试。最小功能测试（Minimum Functionality test, MFT）是样本和对应标签的简单集合，用于检验模型基础能力。MFT 类似于构建小而集中的测试数据集，可以用于检测模型是否使用捷径来处理复杂的输入，但并没有真正掌握该解决任务的能力。不变性测试（Invariance test, INV）是指对输入施加标签保护性扰动，并期望模型预测保持不变。不同的能力需要不同的扰动函数，例如改变地址名称来测试命名实体识别能力，或者引入错别字来测试稳健性。指向性期望测试（Directional Expectation test, DIR）指对输入施加扰动并预期标签会以某种方式变化。例如，如果在针对影评的末尾加上“演员真是太差劲了”，会使得这段影评的预期情绪变得更加消极。图13.9提供了如何测试这些能力的例子。

研究人员可以从头开始创建测试用例，或者通过扰动现有的数据集来创建测试用例。CheckList 为用户提供了一种创建扰动数据的工具，通过遮掩模板的一部分，并借助基于遮掩的语言模型（如 RoBERTa 等）获得遮掩部分的填充建议，例如，“I really {mask} the flight”，利用 RoBERTa 模型可以得到 {enjoyed, liked, loved, regret, ...}，用户可以选择积极、消极和中性的填充词，然后在多个测试中重复使用。

## 2. TextFlint 稳健性测试平台

由复旦大学自然语言处理实验室开发的，针对多语言自然语言处理稳健性评测平台 TextFlint<sup>[73]</sup>，不仅提供了通用文本变形，还包含特定任务变形，并集成了对抗攻击和子集等稳健性测试方式，以及各种该方法的组合以提供了全面的稳健性分析。TextFlint 具有以下特点：

(1) 灵活：TextFlint 提供了 20 种通用变形和 60 种特定任务变形，以及它们的数千种组合，涵盖了文本变形的方方面面，以便对模型的稳健性进行全面评估。TextFlint 支持中英文多种语言的评估，自动评估模型在词汇、语法和语义方面的缺陷，或者根据用户的需求进行灵活的定制分析。对于那些个性化的需求，用户可以修改配置文件，并输入几行代码来实现特定的评估。

(2) 便捷：TextFlint 提供了约 7000 个新的评估数据集，这些数据集是由 40 个原始数据集变形生成的，用于 20 个任务。用户可以直接下载这些数据集进行稳健性评估。对于那些需要全面评估的用户，TextFlint 支持在一个命令中生成所有变形文本和对应标签，对模型进行自动评估，并生成分析报告。

(3) 直观：通过对现有变形结果的合理性和语法性进行人工评价后，以人工评价结果为基础，对每个评估结果分配一个置信度分数。基于评估结果，TextFlint 提供了一个标准的分析报告，涉

及到模型的词汇、语法和语义。所有的评估结果都可以通过可视化和表格的形式显示出来，以帮助用户快速准确地掌握一个模型的缺点。此外，TextFlint 根据分析报告中发现的缺陷，生成大量有针对性的数据来增强被评估的模型，并为模型缺陷提供补丁。

TextFlint 的变形形式基于语言学的指导，根据词法、语法、词形变化关系、语用学设计了 20 种通用变形以及 60 种特定任务变形。主要包含以下类型：

#### (1) 词汇形态

- 形态派生：通过添加前缀或后缀等方法形成新单词的过程。比如：normal 变形为 abnormal。TextFlint 中 *SwapPrefix* 是保持词性的基础上更换前缀。例如：*transfix* 转化为 *affix*。
- 词形变化：英语中时态、数、性别决定了很多词的词形。TextFlint 中 *SwapVerb* 将动词的词形进行变化。例如：“He is studying NLP.” 转换为 “He has studied NLP.”。
- 缩略语：通过缩短或合并两个单词得到的词语。TextFlint 中 *Contraction* 将缩略词替换为原始形式，或者将原始形式替换为缩略形式。例如：“can’ t” 转换为 “can not”。

#### (2) 类聚关系

- 同义词：通过替换具有相同含义的词语或者词组。TextFlint 中 *SwapSyn* 就是利用同义词替换构建的变形。例如：“He loves NLP.” 转换为 “He likes NLP.”。
- 反义词：通过增加否定词或者替换为反义词构造语义相反的句子。这种方式在语义匹配任务中需要同步修改分类结果，但是在信息抽取等任务中则无需修改。TextFlint 中 *SwapAnt* 通过替换反义词，*Add/RmvNeg* 则通过增加或删除否定词完成反义关系构造。例如：“John lives in Ireland.” 转换为 “John doesn’t live in Ireland.”。

#### (3) 语法

- 句法范畴：通过替换具有相同句法范畴的成分，可以在不影响句法结构的情况下构造变形。同样需要注意的是这种变形会引起语义的变化，需要根据任务确认是否适用或修改相应的分类标签。TextFlint 中 *SwapNamedEnt*、*SwapSpecialEnt* 以及 *SwapWord/Ent* 就是利用修改相同句法范畴的词语完成变形。例如：“I love Shanghai.” 转换为 “I love Beijing.”。
- 附属成分：通过增加或者删除某些附属成分，构造符合语法的变形。TextFlint 中 *Delete/AddSubTree* 和 *InsertClause* 分别是通过增加或删除子树，以及增加小句生成变形。例如：“Tom loves NLP.” 转换为 “Tom, who lives in China, loves NLP.”。

#### (4) 语用：

- 会话准则：在不同的环境下，人们为了有效沟通会采用不同的方式进行，因此可以利用这一特性构建变形。TextFlint 中 *RndRepeat/Delete* 针对篇章随机删除或者复制一个句子，*TwitterType* 替换为社交媒体上常用词语、*AddSum* 增加部分描述等都是基于会话准则模型构建变形。例如：“See you later.” 转换为 “CYL”。

- 偏差：语言反映了社会价值和个人观点，因此通常存在偏差。TextFlint 中 *Prejudice* 变形就是利用这种偏差，对相关内容进行替换从而生成变形。例如：“She is a nurse.” 转换为 “He is a nurse.”。

此外，TextFlint 还包含利用翻译模型将原始句子翻译为其他语言，再翻译回来构造变形的 *BackTrans* 方法，利用复述生成模型构建语义相似句子的 *Overlap* 等方法。图13.10给出了 TextFlint 中任务无关通用变形的分类图。

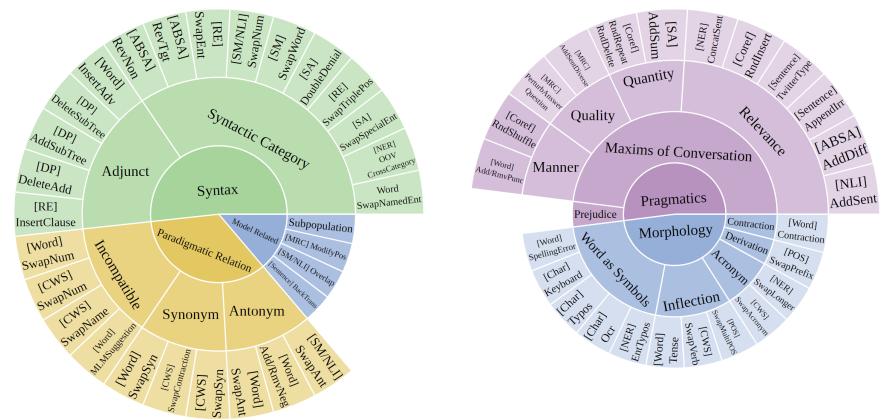


图 13.10 TextFlint 中通用变形分类<sup>[73]</sup>

除此之外，TextFlint 还集成了对抗攻击和子集划分用于全方位验证模型。TextFlint 提供了 16 种简单易用的文本对抗攻击方法，用于验证模型对抗稳健性。子集划分用来确定目标模型在数据集中表现不佳的特定部分。通过对样本进行分类可以按照某些属性来划分数据集，TextFlint 提供了四个常用的子集划分属性，其中包括性别偏差、文本长度、语言模型困惑度和短语匹配。以文本长度属性为例，文本长度筛选出长度最长的 20% 或者最短的 20% 的子集，然后测试模型在子集数据上的表现，来确定模型的预测结果是否会受到这些属性的影响。

通过这些变形以及对抗攻击，使得我们可以对模型的稳健性进行更为全面的分析。表13.2给出了利用 TextFlint 平台对多种大规模预训练语言模型在语言推理 MultiNLI 任务上的评测结果。验证了反义词替换、增加句子、数字变形以及复述生成等变形形式。从结果上我们可以看到，虽然当前的模型在原始集合上取得了不错的成绩，但是简单的变换一下数字就会使得模型的准确率大幅度下降，在一定程度上反映了当前模型稳健性普遍亟待提升。

表 13.2 模型在语义推理任务 MultiNLI 数据集合上准确率

模型	<u>SwapAnt</u>	<u>AddSent</u>	<u>NumWord</u>	<u>Overlap</u>
	原始 → 变形	原始 → 变形	原始 → 变形	原始 → 变形
BERT-base <sup>[29]</sup>	85.10 → 55.69	84.43 → 55.27	82.97 → 49.16	None → 62.67
BERT-large <sup>[29]</sup>	87.84 → 61.18	86.36 → 58.19	85.42 → 54.19	None → 70.65
XLNet-base <sup>[31]</sup>	87.45 → 70.98	86.33 → 57.65	85.55 → 48.77	None → 70.35
XLNet-large <sup>[31]</sup>	89.41 → 75.69	88.63 → 63.37	86.84 → 51.35	None → 78.09
RoBERTa-base <sup>[731]</sup>	87.45 → 63.53	87.13 → 57.25	86.58 → 50.32	None → 75.49
RoBERTa-large <sup>[731]</sup>	92.16 → 74.90	90.12 → 67.73	88.65 → 54.71	None → 73.14
ALBERT-base-v2 <sup>[732]</sup>	87.45 → 50.20	84.09 → 53.59	82.97 → 49.42	None → 67.15
ALBERT-xxlarge-v2 <sup>[732]</sup>	91.76 → 69.80	89.89 → 79.11	89.03 → 46.84	None → 74.92
平均	88.58 → 65.25	87.12 → 61.52	86.00 → 50.60	None → 71.56

## 13.6 延伸阅读

本章中我们主要介绍了常见的文本对抗攻击和文本对抗防御方法，并没有深入涉及模型对抗脆弱性的成因，对抗样本的成因以及性质也是对抗鲁棒性领域的研究重点。对抗攻击和对抗防御是矛与盾的关系，对抗攻击的发展迫使研究人员更加深入的分析模型的鲁棒性，开发出更有效的防御方法，这也进一步促进更加强大的攻击算法的产生。

传统的观点倾向于将对抗样本视为由输入空间的高维性质或训练数据的统计波动引起的畸变<sup>[733, 734]</sup>。从这个角度来看，可以很自然地将对抗鲁棒性视为一个目标，可以通过改进的标准正则化方法或对网络输入/输出进行预/后处理，将其与最大化准确性分开并独立追求<sup>[735, 736]</sup>。最近的观察表明，对抗样本是由于数据中具有良好泛化能力但是敏感的特征所引起的<sup>[737]</sup>。因为模型训练过程中的目标是最大化任务精度，因此模型将尽可能的利用一切可以利用的特征信号，即使有些特征对于人类来说是无法理解、不合理的。

对抗样本的隐匿性和合理性是对抗攻击方法追求的一个主要目标。尽管目前的文本对抗攻击者容易误导模型，但是最近的研究指出，文本对抗样本普遍存在语义流畅性差、容易被人类察觉等问题<sup>[738, 739]</sup>。因此，现在的方法借助预训练语言模型强大的语言生成能力来帮助生成语义更加连贯的对抗样本<sup>[740, 741]</sup>。此外，研究人员通过借助篡改句法结构信息<sup>[742]</sup>，以及判断数据隐匿性的方法来实现更隐匿的投毒攻击<sup>[743]</sup>。

预训练语言模型的结构和权重同样影响着模型的鲁棒性。文献 [744, 745] 从模型角度分析了预训练模型对鲁棒的影响，发现预训练模型的部分结构和权重会降低模型的鲁棒性。通过将彩票网络假说和鲁棒性相结合，可以从预训练语言模型中提取出一个具有良好鲁棒性的子网络，并且能够原始任务上获得与完整类似的性能。文献 [746] 进一步发现结构化稀疏的鲁棒子网络的结构可以在训练阶段初期就被确定下来，因此提出了一种基于结构化稀疏的鲁棒子网络来加速对抗训

练过程。

此外，也有大量工作侧重于研究如何更好的评测模型的能力，例如本章节中介绍的介绍了两种模型稳健性通用评价基准：CheckList 和 TextFlint。文献 [747] 提出需要在传统的测试集合之外，构造对比集合（Contrast Sets）为原始数据提供的更全面的评估。文献 [748] 提出了与任务无关的方法，根据样本的难度级别对样本进行加权。根据测试样本本身、训练样本和测试样本之间的不同以及模型的置信度等信息分别提出了 WSBias、WOOD 以及 WMProb 方法来更好反映模型在真实世界中的效果。文献 [749] 提出了人与模型同在回路（Human-and-model-in-the-loop）的动态基准测试集合构建方法，并发布了 Dynabench 平台用于数据集合构建和模型评测。

## 13.7 习题

- (1) 有哪些文本生成的技术指标可以用来约束文本对抗样本的生成质量？
- (2) 如何改进对抗训练算法的计算效率？
- (3) 有哪些特性可以用来区分干净样本和对抗样本？
- (4) 如何将对抗检测和对抗防御同时应用于一个模型中？
- (5) 后门攻击的缺点是什么？

## 14. 模型可解释性

如前所述，目前绝大部分自然语言处理算法都是基于统计机器学习方法，这些数据驱动的算法在绝大部分任务上取得了良好的性能。但是，以深度神经网络方法为代表的“黑盒”模型缺乏可解释性。我们不能理解数百亿甚至是数万亿参数中的每个维度的含义，这造成了深度学习模型本质上的不可解释性。然而，我们又迫切的需要了解模型是否真正符合人类语言的习惯，机器在语言处理任务中的决策与人类的决策过程有何异同，数据驱动的统计模型与人类语言认知系统的差异等问题。这些问题一方面关系到如何进一步提升自然语言处理算法的处理效果以及稳健性，另一方面如果不能够很好的解决这些问题，就会给自然语言处理算法在关键业务中的应用带来极大的风险和挑战。在医疗诊断、金融预测、司法审判等高风险场景中是否能够应用自然语言处理算法，上述问题都是系统成功的关键要素。

本章首先介绍人工智能可解释性基本概念和主要研究内容，在此基础上介绍通用的解释性分析方法，最后介绍可解释自然语言处理中的可解释模型、可解释数据和可解释评估问题。

### 14.1 可解释性概述

可解释性（Interpretability）问题在统计机器学习模型中广泛存在，在追求更好性能的同时，需要模型更加透明。例如，在智能诊疗问答过程中，为了提供更可靠的服务，模型除了准确寻找患者问题的答案外，同时也应提供机器抽取答案的过程，从而来解释预测行为。杨强教授等人在《可解释人工智能导论》中将可解释人工智能（Explainable Artificial Intelligence, XAI）定义为智能体以一种可解释、可理解、人机互动的方式，与人工智能系统的使用者、受影响者、决策者、开发者等，达成清晰有效的交流沟通以取得人类信任，同时满足各类应用场景对智能体决策机制的监管要求<sup>[750]</sup>。这对人工智能系统的可解释性提供了更高、更全面的要求。当前，尽管大规模复杂模型已经广泛应用到自然语言处理的各个方向中，也深入影响到了我们生活的方方面面，但是由于其内在决策过程的不可知性，导致在关键业务场景下应用仍然受限，人们无法信任模型的预测结果。如图14.1所示，BERT算法在针对例句的掩盖单词预测任务中，虽然给出了合理的预测结果，但是其所依赖的依据并不完全符合人类认知。

虽然目前自然语言处理算法在很多任务上都取得了很好的效果，但是仍然需要了解模型的决

基于显著图的单词重要程度可视化	MASK预测：
显著图：	47.1% <i>nurse</i>
[CLS] The [MASK] rushed to the <b>emergency</b> room to see <b>her</b> patient. [SEP]	16.4% <i>woman</i>
	10.0% <i>doctor</i>
	3.4% <i>mother</i>
	3.0% <i>girl</i>

图 14.1 BERT 模型预测依据示例

策依据，这也是优化模型效果并提升人们对模型信任度的重要方法。由于目前绝大多数自然语言处理算法都基于统计机器学习方法，因此算法一定会受到数据、模型以及评估准则的影响。在数据层面，由于数据样本普遍存在局限和偏见（Bias），仅依赖数据驱动的方法很容易学习到表层模式（Surface Pattern），从而不可避免地构建虚假关系（Spurious Relationship）。这些表层模式和关联关系与人们所做决策的依据不同，并不能构建真正的因果关联关系，因此在处理与训练样本不一致的情况时就会产生错误。在模型层面，统计机器学习模型的性能与可解释性之间往往不可兼得。线性模型刻画自变量和因变量之间的线性关系，模型简单易理解，但往往性能欠佳；而深度神经网络模型能刻画自变量和因变量之间复杂的关系，可以拥有更高的性能，但牺牲了可解释性。在评估准则方面，利用标准评测集合使用准确率、精确度等单一指标评价模型效果的方法，虽然推动了自然语言处理的发展，但是缺乏针对模型细粒度和可解释的评价。这些问题都与可解释性息息相关。模型可解释性研究对于未来自然语言处理的发展极为重要。

### 14.1.1 可解释的分类

根据可解释人工智能的定义，可以看到其核心要素是智能体（AI agent）能够有效地“解释”自己，并取得人类使用者的“信任”。解释是信任的基础，随着人工智能系统越来越复杂，功能越来越强大，系统如果要取得人们的信任，就必须要考虑不同用户的应用场景、背景、教育程度等各种因素，提供不同内容与形式的解释。根据系统提供解释的程度以及所面向的受众都可以将人工智能系统进行分类。遵循《可解释人工智能导论》<sup>[750]</sup> 的分类体系，根据系统受众的不同，可解释性可以分为以下几类：

- (1) 面向开发者的解释：系统开发人员具有相当的人工智能专业知识，需要依据解释来进一步提升模型性能和鲁棒性，消除偏差，减少模型风险和错误。比如，模型在处理哪些类型的数据时错误率会明显升高？深度模型的每一层或者每个维度的具体功能是什么？
- (2) 面向使用者的解释：系统使用者通常不具备人工智能专业知识，更关心的是系统所做出的某个决策的依据是什么。比如，针对疾病诊断系统，医生希望知道系统所给出的判断主要依据是什么？置信度是如何评估的？
- (3) 面向监管者的解释：随着各国对人工智能系统的应用风险预防的加强以及监管立法逐渐加强，人工智能系统要在监管合规条件下运行。比如，模型的训练过程中所使用的数据是否符合隐私保护及数据治理条例<sup>[751]</sup>，需要有明确的解释及认证。

不同类型的用户所关注的角度不同，但是总体来说主要包含透明度（Transparency）、可解构性（Decomposability）、事后解释（Post-hoc Explanation）、可担责性（Accountability）以及适用边界（Applicable border）。

算法透明度仍然存在一定争议，但是总体上包括算法源代码、输入数据、输出结果等在内的算法要素，综合使用算法分析、算法审计等手段合理促成算法透明。2019年我国发布的《新一代人工智能治理原则——发展负责任的人工智能》、2022年欧盟通过的《数字服务法案》、美国国防部发表的《情报部门人工智能伦理框架》等都对算法透明度进行了一定的要求和规范。

算法可解构性是指可以基于该算法本身内部结构提取算法决策机制构建解释，揭示不同特征在算法决策过程中的作用。线性规划、决策树、朴素贝叶斯等算法具有很好的可解构性，可以根据其模型参数和结构清晰地解释其决策过程，甚至可以对其参数进行人工设定。但是，目前能取得很好效果的深度学习算法却无法解释其预测值，其算法可解构性很差。

算法事后解释试图通过可解释替代模型（Surrogate Model）、基于梯度的相关性、沙普利值等方法提供局部或者全局方法，近似解释黑盒算法的决策依据。相比于白盒算法本身所具备的可解构性，事后解释的方法所提供的是针对黑盒模型的近似解释，方法本身也有很多需要进一步研究的内容。

算法适用边界是指算法所适用的领域和范围，目标是以算法的可解释性为基础，确定算法对于特定问题的适用性。算法适用边界的研究，可以在一定程度上减少实际应用中，由于无法被数学模型充分地表示等因素所带来的决策错误和应用风险。

算法可担责性在模型解释性的基础上提出了更高的要求，要求模型提供预测结果的正当性。可担责性要求算法能够“谨慎”地做出预测，试图避免目前基于数据驱动的算法由于训练数据、模型结构等原因，所造成的算法偏见和算法不可控等潜在风险。

### 14.1.2 解释的评价

近年来，针对黑盒机器学习模型，特别是深度神经网络模型的可解释性，研究员们从多个方面给出了很多方法。对于这些解释方法如何进行评价，也是仍需进一步研究的问题。解释方法的评价可以从以下几个方面开展：

#### 1. 忠实性

忠实性（Fidelity）是指解释方法是否客观忠实地反映了被解释算法的处理逻辑<sup>[752][753]</sup>。如果通过解释方法给出某个特征或变量具有重要的作用，而这个特征或变量确实非常重要，那么这个解释方法就具有很好的忠实性。忠实性是解释方法评价中最重要的指标之一。只有具有很好忠实性的解释方法，才能够真正应用于算法解释和评测中。

#### 2. 敏感度

敏感度（Sensitivity）是指解释方法在输入样本或者模型参数发生微小变化时，所提供的结果是否会发生相应的变化<sup>[754][755]</sup>。通常情况下，希望解释方法对模型参数的敏感度相对较高，而对输

入样本的敏感度可以适当降低。这样的解释方法与模型参数的相关度较高，同时又能够在输入样本有一定噪声的情况下也能够给出可靠的解释。

### 3. 全面性

全面性(Integrity)是指解释方法所提供的结果是否完全地反映了目标算法的全部处理逻辑<sup>[756]</sup>。如果一个解释方法所提供的结果仅对某一个部分进行了解释，那么这个解释方法就是不全面的。如果一个解释方法能够对整体和各组块都给出解释，那么该算法的全面性就很好。

### 4. 可读性

可读性(Readability)是指解释方法所给出的解释是否通俗易懂，便于用户理解。解释方法所提出的结果需要提供给各类角色用于模型效果提升、结果采纳判定等任务。这些都要求解释方法输出的结果简单，让人容易理解。如果解释方法提供的仍然是数亿维度的数值，或者是包含非常多复杂概念和各种关联关系的解释，人们很难理解，也就不能达到解释的要求。

不同的解释方法所提供的结果在上述评价因素上具有不同的权衡。例如模型所有的参数可作为模型行为的一个解释，这种解释虽然具有很好的忠实性，但是可读性却非常差。又比如，注意力机制可以给出当前单词对句子中每个词的关注程度，通过将编码器在计算单词的最终表示时所关注的单词进行可视化，可以揭示网络如何做出决定。如图 14.2 所示，对于左图和右图中的两个句子，通过对 Transformer 结构中的注意力进行可视化。该图反映了机器翻译任务中，在编码器的 Transformer 层中单词“it”的自注意力分布（八个注意力头之一）。颜色越深，表明注意力分数越高。可以看到“it”可以指代的两个名词，并且各自的关注度反映了它在不同上下文中的选择。这种解释方法具有较好的可读性，但是如果对通过该方法确定的重要词语进行替换，模型预测结果可能并不发生变化，说明该方法的忠实性有所欠缺。

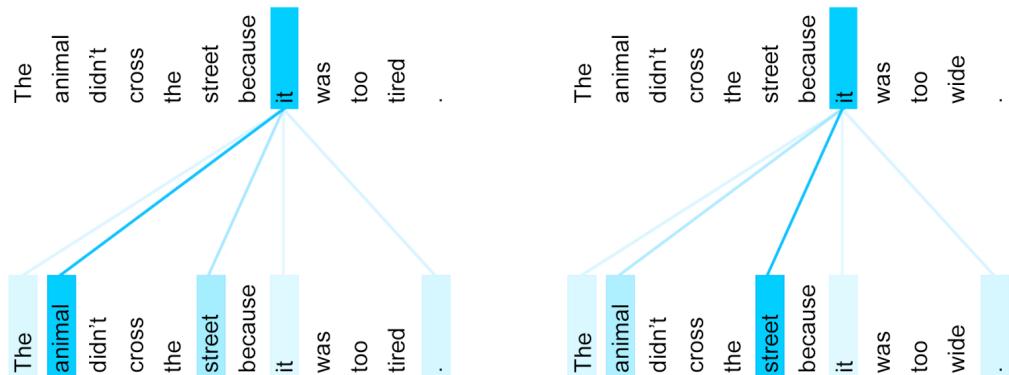


图 14.2 Transformer 结构中的注意力可视化示例

本章首先介绍目前可解释机器学习中常用的分析方法，包括局部分析方法与全局分析方法。在此基础上，从可解释模型、可解释数据以及可解释评估三个方向介绍可解释自然语言处理模型。

## 14.2 解释性分析方法

根据所关注的视野不同，解释性分析方法可以分为局部解释和全局解释两个类别。局部解释通常是针对单个或一类测试样例，帮助人们判断模型对该样本做出预测背后的原因是否合理、或者挖掘在该样本特征空间的邻域内可能存在的偏差（Bias）；全局解释则是针对模型的整体行为，判断模型是否对某些样本存在全局偏差、或者从整体上判断该模型是否可以在现实场景中部署。本节将分别介绍这两类解释性分析方法。

### 14.2.1 局部分析方法

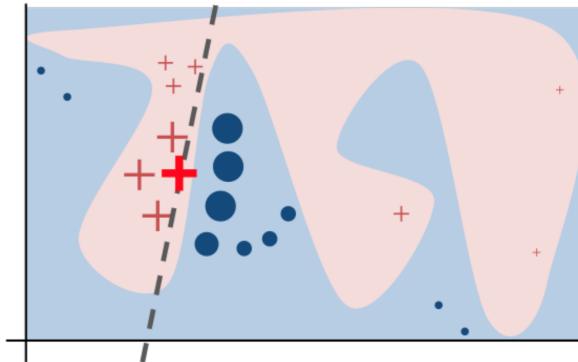
当模型用于为单个样本制定决策时，如何确定模型对当前样本预测的可信度是非常重要的研究内容。比如，在使用模型进行医疗诊断或者恐怖主义检测时，错误的模型预测结果可能导致灾难性的后果。在这种情况下，即使模型的整体预测性能为 99.9%，我们也要尽可能的确定对当前单个样本的预测属于分类正确的 99.9%，还是分类错误的那 0.1%。再比如，利用模型来确定是否给某个申请人批准贷款时，模型只给出拒绝的决策而不提供拒绝的理由，会极大的影响人们的使用体验。而局部分析方法通过对当前样本的预测提供解释，促进模型的使用者和开发者对单个预测的理解，提高人们对当前模型预测的信任。

#### 1. LIME 局部分析算法

LIME（Local Interpretable Model-Agnostic Explanations）<sup>[757]</sup> 是一种模型无关的局部分析方法，试图通过学习一个简单的模型来近似原模型在测试样例附近的预测行为，采用一种较为忠实的方式解释分类器或者回归模型的预测。图 14.3 给出了一个二分类问题（蓝色和粉色区域）的示意图。如图所示，尽管模型整体的决策面是非常复杂的，但模型在单个样本（图中粗体红叉点表示）附近的决策面可以使用线性分界面（图中黑色虚线表示）来逼近。因此可以利用简单的线性模型来模拟原始模型在单个样例附近的决策。通过扰动输入样本的特征，来判断哪些特征的存在与否会对模型的决策产生重大影响。例如，删去图片中的某个像素块后模型的性能大幅下降，则说明该像素块是重要的特征。

给定需要解释的分类器  $f$ ，输入  $x$  以及预测为某个类别的概率，算法的过程可以大致分为以下几个步骤：首先，引入可解释的特征。需要将  $x$  转化为对应的特征向量  $x'$ 。若样本本身是结构化数据，输入本身是具有含义的，则只需要采样获取扰动的样本；而对于非结构化数据，则需要先引入可解释的特征。对于图片而言，可以利用超像素（super pixel）的方式做图片分割，将图片切分成若干块，用二分向量来指示某个超像素是否存在，对于文本数据则利用单词是否存在作为二分向量的指示。

其次，获得原始  $x$  的扰动样本。在预测样本的邻域内随机采样，对于连续型特征，根据正态

图 14.3 LIME 模型结果示例<sup>[757]</sup>

分布来采样随机数产生扰动样本；对于类别型特征，则根据训练集的分布进行采样，若与测试样本特征相同则为 1，否则为 0。对特征向量  $x'$  进行扰动后获得对应的  $z$  和  $z'$ ，并计算扰动样本  $z$  与  $x$  的距离  $D(x, z)$ 。

最后，则是训练解释模型  $g$ ，其优化目标是最小化  $g$  和  $f$  在扰动的样本上的预测，即令  $g$  尽可能地逼近  $f$  在样本  $x$  附近的决策行为，同时保持  $g$  尽可能的简单。此外，考虑到扰动的样本可能和  $x$  偏离很远，模型  $f$  对偏离较远的样本的决策面可能不再是线性的，因此 LIME 算法加入了对距离的惩罚，使得模型可以更关注和  $x$  更近的样本，其目标函数为：

$$\mathcal{L}(f, g, \pi_x) = \sum_{z', z \in Z} \pi_x(z)(f(z) - g(z'))^2 \quad (14.1)$$

其中， $g$  使用的是 K-Lasso 回归模型， $w = \min_{w_0, w} \left\{ \frac{1}{N} (y - w_0 - Xw)^2 \right\}$ ，且  $\sum_{j=1}^p |w_j| \leq K$ 。 $K$  为选择的特征数量。 $\pi_x(z)$  是一个指数核函数，刻画了  $z$  和  $x$  的相似程度（在图 14.3 中距离越小、越相似、点越大）， $\pi_x(z) = \exp(-D(x, z)^2/\sigma^2)$ ， $\sigma$  为超参。最终根据模型  $g$  学习到的参数  $w$  的值的大小就可以知道对应的超像素点或者单词的重要性。

LIME 通过扰动特征根据预测的变化来判断特征的重要性，同时最后可以获取对应特征的重要值，因此是一种非全局忠实，但是局部较为忠实，同时可读性也较好的方法。并且，由于 LIME 是一种模型无关的算法，其适用性也相对广泛。但是，LIME 的使用上也存在一些问题：(1) 需要确定邻域的范围，对于不同的邻域，产生的解释可能不同甚至相悖；(2) 对结构化数据的扰动特征采样时，可能会忽略特征之间的相关性，导致产生一些不合常理的样本来解释模型；(3) 解释模型  $g$  需要预先设定，不同的解释模型产生的解释也可能不同。而这些缺点也导致了 LIME 方法本身不太稳定。

## 2. 显著图局部分析算法

显著图 (Saliency Map)，也称为热力图，计算了单个输入的各个部分与模型预测结果的相关性程度，相关性分数越高的部分对模型输出的重要性程度也越高。基于显著图的可视化结果，人们能直观地在视觉上将模型输出归因到输入样本的某些部分。与 LIME 相比，显著图的可解释性不需要新增解释模型，它利用原始模型的单个输入与参数，通过设计好的公式计算得到。因此，显著图拥有良好的可读性，但间接基于模型参数的获取方式也在一定程度上降低了它的忠实性。

获取显著图的方式主要包含以下类型：

- 基于注意力：将模型中注意力模块中的值，转化为显著图中的相关性分数（本章第14.3.1节将详细介绍相关方法）。优点是方法简单直接，可读性高；缺点则是依赖于注意力模块，并不适用于所有模型，且忠实性难以保证。
- 基于扰动：通过扰动单个输入或神经元后，观察对网络中后续神经元产生的影响<sup>[758]</sup>。优点是能够直接观察到某些输入部分对特定神经元或输出的影响；缺点则是计算效率低，因为每次扰动都需要单独的一遍经过整个网络的前向传播。
- 基于反向传播：通过一次反向传播，将重要性信号从输出神经元传播至输入神经元。与基于扰动的方法不同，仅需一次传播即可生成显著图的方式保证了该方法的高效快速，但存在使用不同的反向传播方式，所生成的显著图有所不同的问题。

因为模型的训练方式大多采用梯度的反向传播，所以基于反向传播的显著图生成方式和其他方法相比，除了高效快速，忠实性也更高。因此许多工作选择对这种方法进行精细设计以获取质量更高的显著图。

基于反向传播的显著图获取方式中，最经典的做法之一是基于梯度的方法。该方法也是首先应用于图像分类任务中。根据模型输出的预测类别得分对输入图像各个像素的梯度值，获得输入与对应预测类别的相关性程度<sup>[759]</sup>，计算公式如下所示：

$$G_i(x) = \nabla_x F_i(x) \quad (14.2)$$

其中  $x$  是模型输入的一幅图像， $F_i(x)$  是模型将  $x$  归为类别  $i$  的预测得分， $G_i(x)$  则表示该模型输入  $x$  对输出类别  $i$  的显著图，大小与  $x$  一致。图14.4给出了使用 ConvNet 神经网络，通过 ILSVRC-2013 数据集进行训练，针对模型输出得分最高的类别给出的显著图示例。通过对比输入图片和所对应的显著图，我们可以看到模型分类所依赖的信息是否与人的认知一致，从而可以分析和改进模型结构。

为了解决直接使用梯度所产生的梯度饱和等问题，许多工作在此基础上提出了改进。其中一类直接修改显著图的计算方式。SmoothGrad 方法<sup>[760]</sup>提出对输入加上随机的高斯噪声，以减少基于梯度的显著图中的视觉噪声，生成更平滑的显著图。具体做法是对特定的输入图像随机采样多



图 14.4 基于梯度的显著图生成方法效果<sup>[759]</sup>

个高斯噪声以生成多个模型输入，然后生成的多个显著图进行平均，计算方式如下：

$$G_i(x) = \frac{1}{N} \sum_i^N \nabla_{(x+\epsilon)} F_i(x + \epsilon) \quad (14.3)$$

其中  $N$  是随机采样的样本数， $\epsilon \sim \mathcal{N}(0, \sigma^2)$  表示高斯噪声。尽管该方法可以生成视觉上更清晰的显著图，但梯度饱和的问题并未解决。

集成梯度（Integrated Gradients, IG）<sup>[761]</sup> 方法则对输入进行线性插值，然后将其梯度沿直线进行积分。它将输入从基础值到当前值的梯度积分看作相关性得分，公式如下：

$$G_i(x) = (x - \tilde{x}) \times \int_{\alpha=0}^1 \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x} \quad (14.4)$$

其中  $\tilde{x}$  表示  $x$  的基础值。因为积分梯度的计算与输入相关，所以它可以解决当输入到达某些值之后造成的梯度饱和的问题。基于集成梯度的显著图方法也应用于多模态问题回答（Visual Question Answer, VQA）模型分析<sup>[762]</sup>。图14.5给出了一个分析的示例，红色的单词表示对问题回答有正面贡献，蓝色的单词表示对问题回答有负面贡献，灰色的单词表示对问题回答基本没有贡献。



问题: How symmetrical are the white bricks on either side of the building ?  
预测结果: very  
正确结果: very

图 14.5 基于集成梯度的显著图方法在 VQA 任务分析结果示例<sup>[762]</sup>

从这个示例中，可以看到虽然模型对该问题给出了正确的答案，但是模型所依赖的分类依据是“how”、“are”等这种对于问题语义表达并不重要的词语。相反，“white”、“either”等对语义有重要影响的词语还对本问题的正确回答起到了负面作用。文献[762]中还对具有非重要作用的单词进行了替换，发现这些非重要单词替换之后确实并不影响分类结果。例如，将问句替换为“how spherical are the white bricks on either side of the building”，“how soon are the bricks fading on either side of the building”等句子后，模型所给出的结果依然是“very”。

### 3. 沙普利值局部分析算法

沙普利值（Shapley Value）是一种来自于联合博弈论的方法，用于根据玩家对总支出的贡献来分配支出<sup>[763]</sup>。在机器学习中，则是将不同的特征对最终预测的总贡献分配到各个特征上。沙普利值可以看做边际效益的均值，比如，当 A 单独工作时产生效益  $v(A)$ ，B 加入后则收益变成  $v(A, B)$ ，那么 A 的边际效益则为  $v(A, B) - v(A)$ 。A 的沙普利值就是在所有可能的工作排列组合中边际效益的加权求和。

给定模型  $f$ ，要计算特征  $i$  对模型的贡献  $\phi_i$ ， $F$  为模型  $f$  中所使用的全部特征集合，计算公式如下：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (14.5)$$

其中， $S$  是模型中使用的特征的子集， $x$  是要解释的实例， $f_S(x_S)$  则表示利用特征子集  $S$  训练的模型  $f_S$  对使用相同特征子集的实例  $x_S$  的预测。因此该算法需要对不同的特征子集都分别训练一个模型，所以只能应用于小数据、小模型。为解决上述问题，文献[764]等方法提出使用蒙特卡洛采样的方法来近似计算。

沙普利值是唯一满足有效性、对称性、冗余性和可加性的归因方法：有效性指的是各个特征贡献值之和等于总贡献；对称性指的是如果两个特征对所有可能的特征集合贡献都相同，那这两个特征的沙普利值相同；冗余性指的是如果一个特征不管加到任意特征集中产生的贡献都为 0，那么它的沙普利值为 0；可加性指的是某个特征的总贡献值是多个特征组合的累计贡献。但由于沙普利值需要计算总特征  $F$  的所有子集，而当特征的数量增加时，特征集  $S$  的数量会随之指数增长。

沙普利值通过枚举特征的所有排列组合，来计算特征在所有排列组合上收益的加权平均值。相比之下，LIME 通过采样来获取特征的排列组合，采样的数量并不能保证公平性。沙普利值的优点在于可以公平的将贡献分到特征值上，因此沙普利值的忠实性相较于 LIME 更好，最后的输出和 LIME 一样也是所有特征的重要值，因此可读性和 LIME 一样较好。但是沙普利值的缺点在于计算速度很慢，需要一些近似方法来加速计算，这样也会损失一部分公平性。除此之外，和 LIME 不同的一点是，沙普利值需要所有的训练数据来计算每个特征的重要性，这也在一定程度上限制了模型的可用范围，而 LIME 不存在这个问题。

### 4. 神经元激活局部分析算法

激活最大化（Activation Maximization）目标是获得一个可以最大化某些神经元激活值的输入。在正常的神经网络训练过程中，通过反复调整网络的权重，从而最小化神经网络在训练集上的损失。而激活最大化是在神经网络训练完成之后，在固定神经网络参数的条件下，通过基于梯度的方法优化输入，使某一个神经元的激活值最大<sup>[765]</sup>。

假设一个已经训练好的分类器，其参数为  $\theta$ ，可以将输入  $x$  映射到一个多类别的概率分布上。激活最大化方法的目标就是寻找一个能够最大化该分类器网络第  $l$  层第  $i$  个神经元激活值的输入  $x^*$ ，可以形式化表示为：

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}} \{a_i^l(\theta, \mathbf{x})\} \quad (14.6)$$

其中， $a_i^l$  是一个单独的神经元的激活值，但也可以扩展成一组神经元的激活值，也就是说希望找到一个输入  $x^*$ ，其可以最大化一组神经元的激活值。通过激活最大化获得的输入  $x^*$  被认为是针对神经网络中的某一小部分神经元的解释。通过分析  $x^*$ ，可以得知这一小部分神经元对什么输入内容更为敏感。为了简化表示，本节以下部分使用  $a(\cdot)$  代替  $a_i^l(\cdot)$ 。

最大化激活是一个非凸优化问题，可以通过基于梯度的方法找到一个局部最优点。在优化过程中，神经网络模型的参数是已知的，可以通过梯度上升更新输入：

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma \frac{\partial a(\theta, \mathbf{x}_t)}{\partial \mathbf{x}_t} \quad (14.7)$$

其中， $x_0$  是一个随机初始化的起始输入，通过不断的进行迭代，期望在输入空间中找到一个能够使目标神经元激活值最大的输入  $x^*$ ， $\gamma$  是根据经验选择的学习率。这个过程中神经网络的参数是固定的。对输入的优化过程通常在到达一个合适的阈值或者一定的步数后停止。图14.6给出了基于激活最大化方法，使用 ConvNet 神经网网络，通过 ILSVRC-2013 数据集进行训练后，在分类层“washing machine”、“computer keyboard”以及“kit fox”所对应的神经元所对应的激活最大化输入。

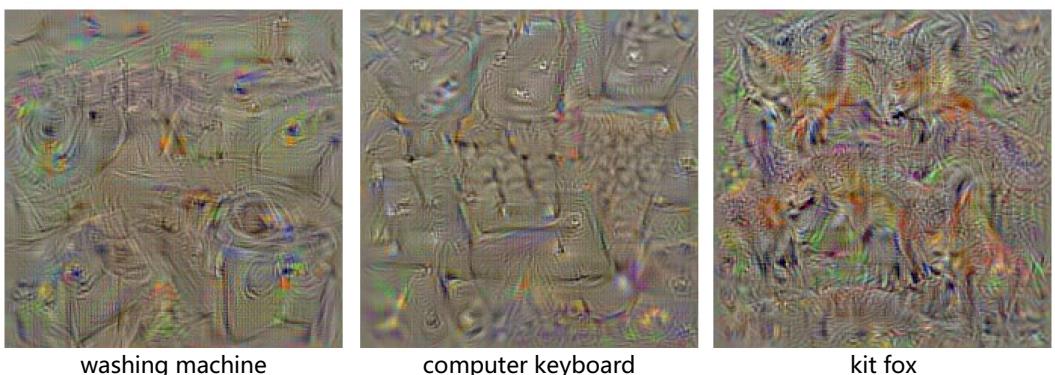


图 14.6 基于激活最大化方法示例<sup>[766]</sup>

直接从完全随机的输入出发，在没有任何约束的情况下通过最大化激活值的方法进行优化，得到的结果往往难以理解。如图14.6所示，虽然在相关输入中能够看到一定的类别图像特征，但是还是很难让人理解。因此，可以对搜索空间进行一定程度的限制。例如，可以从一张真实的图片出发，使得所得到的结果和真实的图像或者训练集中的图片比较相像，进而有较强的可解释性。也可以在目标函数中加入自然图片的一些先验特征来限制搜索范围，改善最大化激活图像的可识别性<sup>[767][768]</sup>。比如，为了增强激活最大化图像的光滑程度，可以定义一个函数  $R$  计算图片的总变差 (Total Variation)。然后，沿着同时满足最大化神经元激活和最小化总变差损失两个条件的梯度方向更新。

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma_1 \frac{\partial a(\theta, \mathbf{x}_t)}{\partial \mathbf{x}_t} - \gamma_2 \frac{\partial R(\mathbf{x}_t)}{\partial \mathbf{x}_t} \quad (14.8)$$

根据先验函数  $R$  的选择不同，最大激活图像会有不同的特点。如图14.7所示，使用 Jitter 函数作为正则化项与不使用正则化项之间还是存在一定的区别，引入正则化项可以更好的进行解释和理解<sup>[767]</sup>。

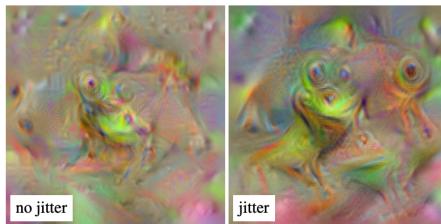


图 14.7 使用和不使用 Jitter 正则化项得到的激活最大化图像对比示例<sup>[767]</sup>

## 14.2.2 全局分析方法

局部分析方法可以帮助我们理解模型的单个预测。但是，除了单个预测之外，在模型真正应用到真实场景中前还需要对模型进行整体的评估。全局分析方法则可以从全局角度上提供对模型的解释。对拥有大规模训练集合的模型而言，通过局部分析方法逐个检查模型对训练和测试数据预测是否合理，在时间和成本上通常是不可接受的。全局分析方法可以通过建议检查特定样例，大大缩小需要检查的数据范围。

### 1. SP-LIME 全局分析法

SP-LIME<sup>[757]</sup> 是基于 LIME 的一种全局分析方法。LIME 是在局部找到对当前预测影响较大的特征，从而提供单个预测的解释，而 SP-LIME 则是通过选取一组具有代表性且多元化的实例来表示模型的整体行为。SP-LIME 将这个选取样例的问题转化为次模优化 (Submodular Optimization) 问题。次模 (Submodular) 是经济学上边际效益递减的形式化描述，即往集合  $A$  中增加一个元素的增益要小于等于往  $A$  的子集中增加一个元素的增益。因此，可以借鉴次模的想法不断的往集合中

添加增益最大的元素，来寻找最具代表性的集合。

SP-LIME 首先需要根据 LIME 算法得到  $n$  样本对应的特征的重要性，从而得到一个  $n \times d$  的重要性矩阵  $\mathbf{W}$ ，其中  $n$  表示样本的数量， $d$  表示特征的数量。如图 14.8 所示，重要性矩阵  $\mathbf{W}$ ， $\mathbf{W}_{ij}$  表明第  $i$  个样本的第  $j$  个特征的重要性。图中将  $\mathbf{W}$  简化为二元矩阵。为了选取有代表性且多元化的样本，需要选择覆盖尽可能多特征且彼此重合小的样本，例如，第 2 个样本和第 3 个样本具有相同的特征值，则只需要选取一个样本。对重要性矩阵  $\mathbf{W}$  的每一列求和得到解释空间中不同特征的全局重要性  $I_j = \sqrt{\sum_{i=1}^n \mathbf{W}_{ij}}$ 。在图 14.8 中，特征 f2 覆盖的样本数最多，因此  $I_2$  最大。

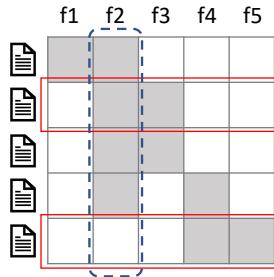


图 14.8 SP-LIME 中使用的  $n \times d$  的矩阵<sup>[757]</sup>

选取样例的标准是尽可能的覆盖所有的特征，因此可以用贪心的方法来选择样例。首先初始化样例选择集合  $V$  为空集，然后不断地添加使集合  $V$  的覆盖率提高最大的样本点。集合的覆盖率定义为：

$$c(V, \mathbf{W}, I) = \sum_{j=1}^d \mathbf{1}_{[\exists i \in V : W_{ij} > 0]} I_j \quad (14.9)$$

即集合  $V$  覆盖的特征的个数。当集合  $V$  的大小达到预设的挑选数量则停止添加。上述算法迭代地增加有最高边际覆盖增益的样本  $i$ ，并以常数  $1 \frac{1}{e}$  的速度近似到最优。集合  $V$  中的样本就是所选取具有代表性的实例。

## 2. 模型蒸馏全局分析法

现在神经网络模型通常通过引入大量参数提升模型预测性能，在提升性能的同时，参数数量的提升也为模型的行为分析和解释带来了很大的挑战。如果能将大模型学到的知识通过某种方式转移到一个相对简单的、更可解释的小模型中，就可以认为小模型可以在一定程度上反应大模型的决策过程，通过对小模型进行解释性分析，进而得到在大模型上全局解释。

模型蒸馏就是一种将大模型的知识迁移到小模型中的常见技术。虽然大模型往往拥有非常大量的可学习参数，需要更大的存储空间和更长的推理时间。但是，与此同时很多参数并没有得到充分的利用。如果能将大模型中的知识迁移到小模型中，那么可以约束参数数量对解释性的影响，同时最大限度的保留大模型的性能优势。这也是蒸馏一词的来源，意味着去除大模型中的“杂质”。

在知识蒸馏中，大模型被形象地称为教师网络，小模型被称为学生网络，学生网络被要求去拟合教师网络的输出。

文献 [769] 中提出使用高度结构化的、更容易进行解释性分析的决策树来近似一个黑盒模型，将得到的决策树作为黑盒模型的全局解释代理。该方法使用坐标轴对齐 (Axis-aligned) 的决策树，树中的非叶子结点都包含一个坐标轴对齐条件  $C = (x_i < t)$ ，其中  $i \in [1 \cdots d]$ ,  $t \in \mathbf{R}$ ,  $d$  是输入空间  $\mathcal{X}$  的维度，记条件  $C$  的可行集为  $F(C) = \{x \in X | x \text{ 满足 } C\}$ 。决策树  $T$  是一个二叉树，其中一个内部节点  $N = (N_L, N_R, C)$  拥有左节点  $N_L$  和右节点  $N_R$ ，以及一个条件  $C = (x_i < t)$ 。叶子节点  $N = (y)$  则和某个标签  $y \in Y$  绑定。记  $N_T$  为  $T$  的根结点。对于一个节点  $N \in T$ ，记  $C_N$  为根结点  $T$  到节点  $N$  路径上的条件的交集。

对于一个训练集  $X_{train} \subseteq \mathcal{X}$  和一个黑盒模型  $f : \mathcal{X} \rightarrow \mathcal{Y}$ ，该方法的目标是学习一个决策树  $T : \mathcal{X} \rightarrow \mathcal{Y}$  去近似  $f$ 。首先使用  $X_{train}$  估计  $\mathcal{X}$  的分布  $\mathcal{P}$ ，然后贪心的构造决策树  $T$ :  $T$  初始化为一个根结点，然后不断迭代分割其叶子结点。当分割叶子结点  $N \in T$  时，使用动态采样策略得到一个新的输入  $x \sim \mathcal{P}$ ，并且  $x \in F(C_N)$ ，使用黑盒模型  $f$  计算其对应的标签  $y = f(x)$ ，并用这些数据验证划分的好坏。

首先使用 EM 算法得到一个拟合  $X_{train}$  的混合坐标轴对齐的高斯分布  $\mathcal{P}$ 。用类似 CART<sup>[770]</sup> 的方法，构造一棵大小为  $k$  的贪心决策树  $T^*$ 。初始化  $T^*$  为单节点  $N_{T^*} = (y)$  树， $y$  是分布  $\mathcal{P}$  中的出现次数最多的标签。然后进行  $k - 1$  次迭代划分  $T^*$  中的叶子结点：在每次迭代时，我们选择一个叶子结点  $N = (y)$ ，然后用一个内部节点  $N' = (N_L, N_R, C)$  替代它，其中  $N_L = (y_L)$ 、 $N_R = (y_R)$ ， $C = (x_{i^*} \leq t^*)$ ：

$$(i^*, t^*) = \underset{i \in [d], t \in \mathbf{R}}{\operatorname{argmax}} G(i, t)$$

其中划分的收益  $G$  使用基尼杂质 (Gini Impurity)  $H$  表示为：

$$\begin{aligned} G(i, t) = & -H(f, C_N \wedge (x_i \leq t)) \\ & -H(f, C_N \wedge (x_i > t)) + H(f, C_N) \end{aligned}$$

$$H(f, C) = \left( 1 - \sum_{y \in \mathcal{Y}} \Pr_{x \sim \mathcal{P}}[f(x) = y | C]^2 \right) \cdot \Pr_{x \sim \mathcal{P}}[C]$$

划分之后，叶子节点的标签为：

$$\begin{aligned} y_L &= \arg \max_{y \in \mathcal{Y}} \Pr_{x \sim \mathcal{P}}[f(x) = y | C_N \wedge (x_i \leq t)] \\ y_R &= \arg \max_{y \in \mathcal{Y}} \Pr_{x \sim \mathcal{P}}[f(x) = y | C_N \wedge (x_i > t)] \end{aligned}$$

文献 [769] 中通过问卷调查的方式评估抽取出来的决策树的准确性和可解释性。在一个糖尿

病相关的数据集上，作者征集了 46 名有机器学习相关背景的本科生，让他们以抽取出来的决策树为依据，完成一个糖尿病诊断相关的问卷，用这些志愿者的得分来衡量模型的可解释性。

## 14.3 自然语言处理算法解释分析方法

前一节介绍的通用可解释性算法应用到具体的自然语言处理任务时，还需要针对具体任务的特点来调整。通用方法往往假设模型的输入空间或者隐空间是连续的（例如：数字图像信号、音频信号），但是自然语言处理任务处理的大都是单词，模型往往需要处理离散信号。因此在运用可解释算法时需要考虑模型不可微、搜索空间大等挑战。另一方面，针对自然语言处理任务本身的特点，研究人员也探索了许多聚焦于文本任务的可解释性方法。本节将从模型解释性分析算法，数据解释分析方法，以及可解释评估三个角度介绍一些自然语言处理任务中常用的解释性分析方法。

### 14.3.1 模型解释性分析算法

通用的解释性分析方法通过稍加改造，大都可以应用于自然语言处理模型，主要需要处理的问题在于，通用解释分析算法通常是计算输入的每个维度的重要程度。但是自然语言处理任务的输入是由若干个单词组成，每个单词由包含若干维度的向量表示，因此如何将对每个维度重要性的解释转换到单词级别是需要研究的问题。此外，注意力机制以及探针任务是针对自然语言处理领域算法特性而设计的可解释模型。本节将从上述三个方面分别介绍针对自然语言处理算法的可解释模型。

#### 1. 显著图分析方法

本章第 14.2.1 节介绍了显著图分析方法，可以通过基于梯度、传播或者遮挡的方法来衡量神经网络中特定单元或者输入数据特定维度的重要性。文献 [771] 给出了基于一阶导数的显著图方法，衡量输入中每个单元对最终决策的贡献，采用一阶导数来近似。假设对于一个分类任务，输入  $e$  所对应的正确分类结果为  $c$ ， $S_c(e)$  表示模型针对输入  $E$  在类别  $c$  上的得分。显著图分析方法的目标是获取输入中的每个单元对于最终分类结果  $c$  的得分  $S_c(e)$  的贡献进行评价。

对于  $E$  施加微小噪音可以得到  $e$ ，同样可以通过模型得到  $S_c(e)$ ，由于在深度网络条件下  $S_c(e)$  是高度非线性函数，可以采用一阶泰勒展开进行近似，从而可以将其转换为线性表示：

$$S_c(e) \approx w(e)^T e + b \quad (14.10)$$

其中  $w(e)$  表示  $S_c$  关于输入  $e$  的导数：

$$w(e) = \frac{\partial(S_c)}{\partial e} |_e \quad (14.11)$$

导数的绝对值表示某一特定维度对最终决策的贡献度大小，因此显著性得分  $S(e)$  定义为：

$$S(e) = |w(e)| \quad (14.12)$$

图14.9给出了使用一阶导数显著图方法的分析示例。模型采用句子级情感倾向分析语料进行训练，对于语句“我喜欢这部电影”，分类为“褒义”类别的显著图进行了可视化。每个单词的嵌入表示由 100 维向量表示，每个维度的显著性归一化到 0 到 1，颜色由浅至深进行表示。从显著性分析结果上，该句子分类着重依赖了“喜欢”和“电影”两个单词。

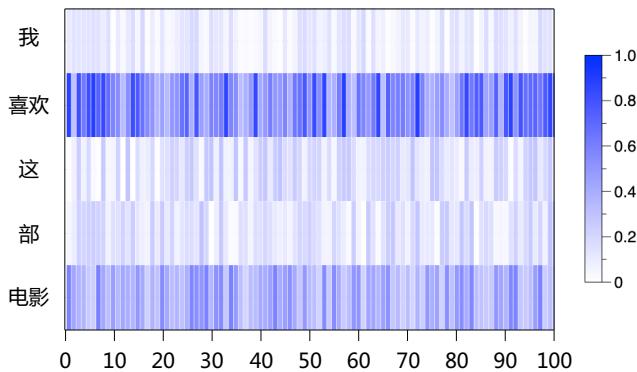


图 14.9 显著图分析方法结果样例

通过一阶导数显著图方法可以得到输入单词每个维度的重要性，但是文本输入是离散的，每个单词的嵌入表示由多个维度组成，因此如何通过每个维度的重要性确认单词的重要性也是需要研究的问题。通过对单词每个维度的重要性求平均、中值、最大值等方法得到单词的显著性值。

## 2. 注意力分析

注意力机制是基于神经网络自然语言处理模型中重要的一个部件：对于输入文本的不同部分，神经网络赋予它们不同的权重，并基于这些权重构建文本处理任务的预测结果。以自注意力机制为基础的 Transformer 模型，以及基于其构建的预训练语言模型，在当前自然语言处理任务上都取得了非常好的效果。因此，对注意力进行分析可以提供理解自然语言处理模型决策过程的一种重要途径。本节我们将以预训练模型 Transformer 结构（如 BERT<sup>[29]</sup>, ALBERT<sup>[772]</sup> 等）中的注意力为例，简介如何通过注意力分析解释模型的内在运行机理。

预训练模型在大量无标注的语料上进行自监督训练获得文本表示。它们在众多自然语言处理任务中获取了良好的表现，但这些模型的表示学习过程仍是黑盒状态：我们并不清楚在自监督学习的过程中模型学习到了什么类型的知识，以及这些知识如何被使用到具体的下游任务中。为更好的理解预训练模型，需要开发针对它们的可解释性分析工具，而其中重要的一类工具是探究预训练模型的注意力中是否蕴含了与语言学结构相匹配的知识。

文献 [773] 尝试探索预训练语言模型 (BERT<sup>[29]</sup>) 是否学习到了语法结构。它主要通过详细分析 BERT 模型内部注意力分布，观察模型从输入中学到的结构信息。具体来看，为了研究注意力头中包含的语言学信息，将 Transformer 结构中的每个注意力头看作简单的分类器，通过标准评测数据集观察它们在预测语法结构的效果。

以识别依存句法分析中的“nsubj”关系为例介绍算法流程。对于注意力头  $h$ ，为句子中每个位置  $1 \leq j \leq n$ ，寻找最相关的词  $w_{\arg \max_i \alpha(h)_i^j}$ ，其中  $\alpha(h)$  表示注意力头  $h$  注意力分数分布， $\alpha(w, h)_i^j$  表示  $w_j$  与  $w_i$  之间的注意力权重。在标注数据中对于输入词序列  $w_1, \dots, w_n$ ，若词对  $(w_i, w_j)$  存在 nsubj 关系，则记  $l(w_i, w_j) = 1$ ，否则为 0。将得到的关系词对集合记为  $S_l(w) = \{j : \sum_{i=1}^n l(w_i, w_j) > 0\}$ 。那么评估注意力头  $h$  是否学到了 nsubj 关系，可以统计所有词对  $(w_{\arg \max_i \alpha(h)_i^j}, w_j)$  在  $S_l(w)$  中出现的频率，从而得到其精度：

$$\text{Precision}(h) = \frac{1}{N} \sum_{w \in \text{corpus}} \sum_{j \in S_l(w)} l(w_{\arg \max_i \alpha(h)_i^j}, w_j)$$

其中  $N$  表示语料库中所有关系词对集合的总词数。

使用这种方法，文献 [773] 观察到 BERT 的注意力头在 Penn Treebank 依存关系标注数据集和 CoNLL-2012 指代关系数据集上均达到了较高的准确率，证明了 BERT 算法在自监督过程中确实学到了一些语法结构特征，如图 14.10 所示。

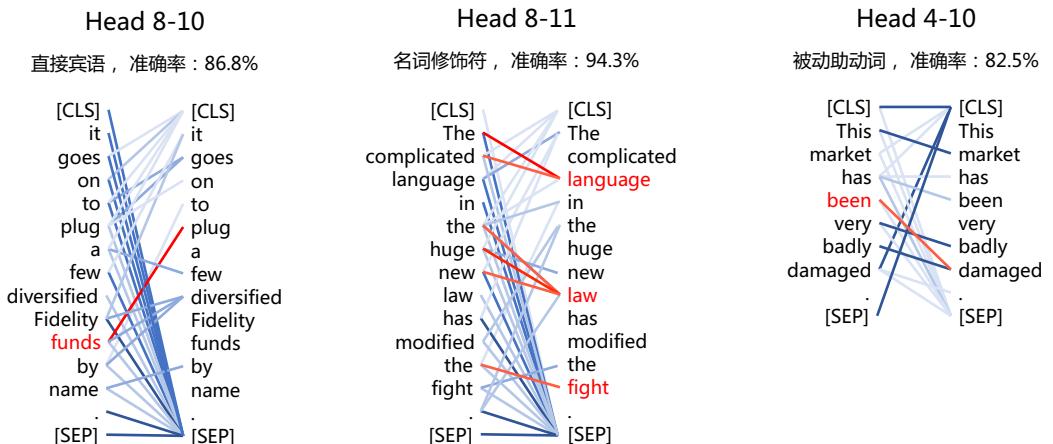


图 14.10 BERT 中不同注意力头在语法结构识别上的结果示例<sup>[773]</sup>

### 3. 探针任务

设计探针任务 (Probe Tasks) 也是一类可以面向隐藏表示的解释性方法。比如，希望探究预训练语言模型 (如 BERT) 的某隐层表示是否包含词性信息，可以通过构建“探针分类器”来尝试根

据一个词的隐层向量预测该词在句子中的词性。具体构建过程可以通过以下几步完成：

- (1) 选择带有词性标注的数据集  $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=0}^{|\mathcal{D}|}$ , 其中  $\mathbf{x}_i$  为句子,  $\mathbf{x}_i = \{w_{i1}, w_{i2}, \dots, w_{i|x_i|}\}$ ,  $y$  为词性类别  $y = \{p_1, p_2, \dots, p_{|y|}\}$ ;
- (2) 选定所关心的预训练模型隐层  $l$ ;
- (3) 构建探针分类器的训练（测试）数据集;  $\mathcal{D}' = \{(\mathbf{g}_i, y_i)\}$ , 其中  $\mathbf{g}_i = \{h_{i1}^{(l)}, h_{i2}^{(l)}, \dots, h_{i|x_i|}^{(l)}\}$ ,  $h_{i*}^{(l)}$  为输入  $x_i$  在第  $l$  层的隐层表示;
- (4) 利用构建的数据集合  $\mathcal{D}'$  训练并测试模型精度。

探针模型对于词性的预测准确程度可以作为一种对隐层向量的解释：若从一个隐层能够很好的预测词性信息，说明该隐层较好的包含了词性信息。通常情况下，为了更好的解释隐层（而不是预测词性），探针分类器通常选择较简单的网络结构（如线性分类器，单层神经网络等方法）。现有的探针任务主要包括句子级别<sup>[774]</sup> 和词级别<sup>[775]</sup> 两大类。

句子级别的探针任务是根据模型上训练得到的句子的向量表示，来探究模型是否学习到了句子级别的语义信息，主要包括以下任务：

- 浅层信息的探针任务
  - 长度探测(Sentence Length), 预测句子长度, 该任务用于测试句向量(Sentence Embedding)是否保留了句子长度的相关信息。
  - 词成分探测 (Word Content)，预测一组中频单词是否在句子中出现过，该任务用于测试句向量是否学习到了单个单词的信息。
- 句法信息的探针任务
  - 语序探测 (Bigram Shift), 调换输入中两个单词的位置，然后对扰动后的输入做二分类判断是否调换过位置，可以用于探测句向量是否保留了词序信息。
  - 句法树深度探测 (Tree Depth), 预测句子语法树的深度，可以用于探测句向量是否包含了句子的层次结构信息。
  - 浅层成分块探测 (Top Constituent), 预测句子的浅层成分（即，句子成分句法树第二层 (S 结点的下一层) 的语法标签）。将一个句子的浅层成分拼接构成待探测的标签。例如：一个句子由一个形容词短语 (ADVP)、一个名词短语 (NP)、一个动词短语 (VP) 构成，则对应的预测标签为“ADVP NP VP”。为减少标签数量，可以选取出现频率最高的几个做为标签集合。
- 语义信息的探针任务
  - 时态探测 (Tense)，预测句子时态。
  - 主语单复数探测 (Subject Number)，预测主语的单复数。
  - 宾语单复数探测 (Object Number)，预测宾语的单复数。
  - 动名替换探测 (Semantic Odd Man Out)，随机将句子中的动词或名词替换为其他的动词或者名词，二分类判断是否进行过替换。

- 主从顺序探测（Coordination Inversion），随机交换两个并列的分句的前后顺序（如以“and”连接的两个句子），用于探测句向量是否学习到了语言逻辑相关的信息。

词级别的探针任务作为对句子级别的探针任务的补充，关注单词级别的语义信息，主要包括以下任务：

- 词汇语义相似性（Lexical Semantic Similarity），评测人工标注的单词对的语义相似度评分和词向量的余弦相似度评分的相关性（Spearman’s Rank Correlation）。
- 词类比（Word Analogy），对于单词间的类比关系对  $w_a : w_b = w_c : w_d$ （如：“男人”：“国王”=“女人”：“女王”），该任务需要在给定  $w_a, w_b, w_c$  的情况下，预测  $w_a : w_b = w_c : x$  中的  $x$ 。

探针任务一定程度上解释了模型对于不同语料以及不同任务学习到了哪些相关的知识，但是探测任务的准确性高就一定证明了模型学习到了我们所探测的性质吗？也很有可能模型只是去拟合了数据的分布，文献[776]提出可以采用控制任务的方法来一定程度上评测探针任务是否有效。最基础的控制任务就是随机打乱探针任务中的标签后重新进行预测。若打乱后的预测结果依旧很好，那可能就说明了模型拟合了数据，而并不是真正学习了句子的语义表示。

### 14.3.2 数据解释分析方法

统计机器学习方法不仅依赖网络结构和损失函数等算法结构，训练数据也起到了非常重要的作用，对最终模型的结果产生重要影响，因此如何衡量训练语料对于模型预测结果的影响，也是可解释性研究中重要的研究内容。文献[718]针对数据对模型预测结果的影响开展了研究，该论文获得机器学习领域重要国际会议 ICML 2017（International Conference on Machine Learning）最佳论文奖。

文献[718]提出将训练语料中某个数据对模型某个预测的影响拆解为两个问题：(1) 如果将训练语料中的某个样本去掉，重新训练得到的新模型，利用该模型做出的预测，会发生什么样的变化？(2) 如果对训练语料中的某个样本进行微小的扰动，重新训练得到新模型的预测结果会有什么样的变化？上述问题可以形式化地定义为：输入样本空间为  $\mathcal{X}$ ，输出目标空间为  $\mathcal{Y}$ ，给定训练语料集合  $Z = \{z_1, z_2, \dots, z_n\}$ ，其中  $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ 。给定一个样本  $z$  和模型参数  $\theta \in \Theta$ ，损失函数为  $\mathcal{L}(z, \theta)$ ，相应的  $\frac{1}{n} \sum_{i=1}^{n=1} \mathcal{L}(z_i, \theta)$  为经验损失。给定训练样本和损失函数，可以通过经验风险最小化准则训练模型参数，即  $\hat{\theta} \stackrel{\text{def}}{=} \arg \min_{\theta \in \Theta} \sum_{i=1}^{n=1} \mathcal{L}(z_i, \theta)$ 。

针对在训练语料中去除某个训练点  $z$ ，模型针对某个测试样本发生变化的问题，可以通过训练包含和不包含  $z$  的数据集合，得到参数  $\hat{\theta}$  和  $\hat{\theta}_{-z}$ ，通过参数变化  $\hat{\theta}_{-z} - \hat{\theta}$  来衡量。但是这种方法需要重新对模型进行训练，对于需要大规模计算的深度学习模型来说，所需要的计算量和时间过多。影响函数（Influence function）方法提供了通过对  $z$  进行微小加权来近似计算的方法。假设对目标训练数据  $z$  给于一个非常小的加权  $\epsilon$ ，则  $\hat{\theta}_{\epsilon, z} \stackrel{\text{def}}{=} \arg \min_{\theta \in \Theta} \sum_{i=1}^{n=1} \mathcal{L}(z_i, \theta) + \epsilon \mathcal{L}(z, \theta)$ 。通过文

献 [777] 中分析结论, 可以得到:

$$\mathcal{I}_{\text{up, params}}(z) \stackrel{\text{def}}{=} \frac{d\hat{\theta}_{\epsilon, z}}{d\epsilon} \Big|_{\epsilon=0} = -\mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\theta} \mathcal{L}(z, \hat{\theta}) \quad (14.13)$$

其中  $\mathbf{H}_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \nabla_{\theta}^2 L(z_i, \hat{\theta})$  为海森矩阵 (Hessian)。由于当  $\epsilon = -\frac{1}{n}$  时相当于移除  $z$ , 可以线性近似移除  $z$  后的参数变化  $\hat{\theta}_{-z} - \hat{\theta} \approx -\frac{1}{n} \mathcal{I}_{\text{up, params}}(z)$ 。

基于  $\mathcal{I}_{\text{up, params}}(z)$ , 通过如下解析解衡量在对训练语料  $z$  进行提权后, 对测试点  $z_{\text{test}}$  的影响:

$$\begin{aligned} \mathcal{I}_{\text{up, loss}}(z, z_{\text{test}}) &\stackrel{\text{def}}{=} \frac{d\mathcal{L}(z_{\text{test}}, \hat{\theta}_{\epsilon, z})}{d\epsilon} \Big|_{\epsilon=0} \\ &= \nabla_{\theta} \mathcal{L}(z_{\text{test}}, \hat{\theta})^T \frac{d\hat{\theta}_{\epsilon, z}}{d\epsilon} \Big|_{\epsilon=0} \\ &= -\nabla_{\theta} \mathcal{L}(z_{\text{test}}, \hat{\theta})^T \mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\theta} \mathcal{L}(z, \hat{\theta}) \end{aligned} \quad (14.14)$$

对训练语料中某个数据  $z = (x, y)$  进行微小改动, 对模型预测的影响的问题。首先定义改动后的数据  $z_{\delta} \stackrel{\text{def}}{=} (x + \delta, y)$ ,  $\hat{\theta}_{z_{\delta}, -z}$  表示通过使用  $z_{\delta}$  代替  $z$  后的训练语料, 根据经验风险最小化训练得到的参数。为了估计该影响, 定义  $\epsilon$  权重下从  $z$  更换为  $z_{\delta}$  的模型参数为:

$$\hat{\theta}_{\epsilon, z_{\delta}, -z} \stackrel{\text{def}}{=} \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(z_i, \theta) + \epsilon \mathcal{L}(z_{\delta}, \theta) - \epsilon \mathcal{L}(z, \theta) \quad (14.15)$$

类似公式14.13, 可以得到:

$$\begin{aligned} \frac{d\hat{\theta}_{\epsilon, z_{\delta}, -z}}{d\epsilon} \Big|_{\epsilon=0} &= \mathcal{I}_{\text{up, params}}(z_{\delta}) - \mathcal{I}_{\text{up, params}}(z) \\ &= -\mathbf{H}^{-1} (\nabla_{\theta} \mathcal{L}(z_{\delta}, \hat{\theta}) - \nabla_{\theta} \mathcal{L}(z, \hat{\theta})) \end{aligned} \quad (14.16)$$

与前面的计算类似, 也可以采用线性近似得到  $\hat{\theta}_{z_{\delta}, -z} - \hat{\theta} \approx \frac{1}{n} (\mathcal{I}_{\text{up, params}}(z_{\delta}) - \mathcal{I}_{\text{up, params}}(z))$ 。如果  $x$  是连续的, 并且  $\delta$  足够小, 还可以进一步的对公式14.16进行估计。假设输入样本空间  $\mathcal{X} \in \mathbb{R}^d$ , 参数空间  $\Theta \in \mathbb{R}^p$ , 并且  $L$  对  $\theta$  和  $x$  可导。因为  $\|\delta\| \rightarrow 0$ , 因此  $\nabla_{\theta} \mathcal{L}(z_{\delta}, \hat{\theta}) - \nabla_{\theta} \mathcal{L}(z, \hat{\theta}) \approx [\nabla_x \nabla_{\theta} L(z, \hat{\theta})] \delta$ , 其中  $\nabla_x \nabla_{\theta} L(z, \hat{\theta}) \in \mathbb{R}^{p \times d}$ 。代入公式14.16, 可以得到:

$$\frac{d\hat{\theta}_{\epsilon, z_{\delta}, -z}}{d\epsilon} \Big|_{\epsilon=0} \approx -\mathbf{H}_{\hat{\theta}}^{-1} [\nabla_x \nabla_{\theta} \mathcal{L}(z, \hat{\theta})] \delta \quad (14.17)$$

因此,  $\hat{\theta}_{z_\delta, z} - \hat{\theta} \approx -\frac{1}{n} \mathbf{H}_{\hat{\theta}}^{-1} [\nabla_x \nabla_{\theta} \mathcal{L}(z, \hat{\theta})] \delta$ , 并由此可以得到相应的影响函数:

$$\begin{aligned}\mathcal{I}_{\text{pert,loss}}(z, z_{\text{test}}) &\stackrel{\text{def}}{=} \nabla_{\delta} \mathcal{L}(z_{\text{test}}, \hat{\theta}_{z_\delta, -z}) \Big|_{\delta=0} \\ &= -\nabla_{\theta} \mathcal{L}(z_{\text{test}}, \hat{\theta})^T \mathbf{H}_{\hat{\theta}}^{-1} \nabla_x \nabla_{\theta} \mathcal{L}(z, \hat{\theta})\end{aligned}\quad (14.18)$$

在定义了上述两种影响函数后, 直接按照上述公式进行计算, 所需的计算量很大。主要原因在于需要对  $n$  个训练样本计算海森矩阵并取平均和求逆。同时还需要对训练语料中的所有样本, 针对测试样本都计算影响函数。使用隐式海森向量积 (Hessian-vector Product) 来近似计算, 定义  $s_{\text{test}} \stackrel{\text{def}}{=} \mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\theta} \mathcal{L}(z_{\text{test}}, \hat{\theta})$ 。由此  $\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}}) = \nabla_{\theta} \mathcal{L}(z_{\text{test}}, \hat{\theta})^T \mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\theta} \mathcal{L}(z, \hat{\theta})$  可以转化为  $\mathcal{I} = -s_{\text{test}} \cdot \nabla_{\theta} \mathcal{L}(z, \hat{\theta})$ 。文献 [718] 提出了两种近似计算方法, 详细过程可以参考文献内容。

### 14.3.3 可解释评估

目前自然语言处理的评估方法通常基于公开数据集合, 使用统计机器学习中常用的准确率、精度、召回、F1 值等指标进行评价。虽然这种评价方法极大地推动了自然语言处理的高速发展, 但是近年来也逐渐暴露出了单一的粗粒度指标无法很好的区分不同系统之间在细粒度任务维度上的优势和劣势等问题。可解释评估 (Interpretable Evaluation) 旨在通过对特定任务设计多个不同的可解释属性, 细粒度地评估模型在一个或多个数据集上不同属性类的性能, 并对模型偏差、数据集偏差及二者之间的相关性进行评价。可解释评估的主要流程包括:

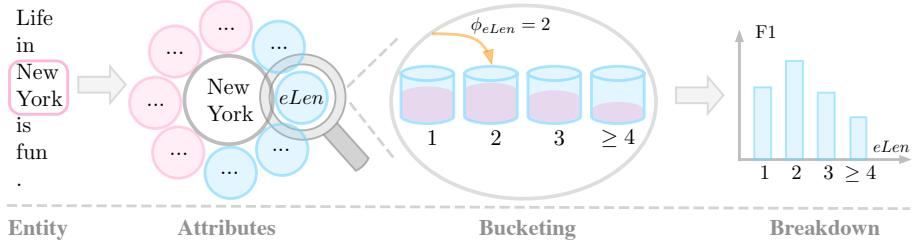
- (1) 属性定义: 针对特定任务设计多个不同的可解释属性 (例如, 针对命名实体识别任务设置实体类型、实体长度等属性);
- (2) 样本分桶: 计算待测试的样本的属性值, 并将样本放入符合相关属性的桶中;
- (3) 分桶性能评估: 评估每个分桶中的样本的性能, 进行细粒度评估。

文献 [778] 中针对命名实体识别任务的可解释评估如图14.11所示。针对命名实体识别任务定义了包括实体长度、标签一致性、实体密度、句子长度等在内的属性。针对预先定义的命名实体识别任务属性, 计算测试样例“Life in New York is fun.”中的“New York”的所对应的属性值。本例中, 针对实体长度属性, “New York”所对应的属性值为 2, 因此将本测试样例放入实体长度为 2 的分桶中。当将整个测试集中的实体分类放入对应的存储桶之后, 分别计算每个分桶中的实体识别性能。

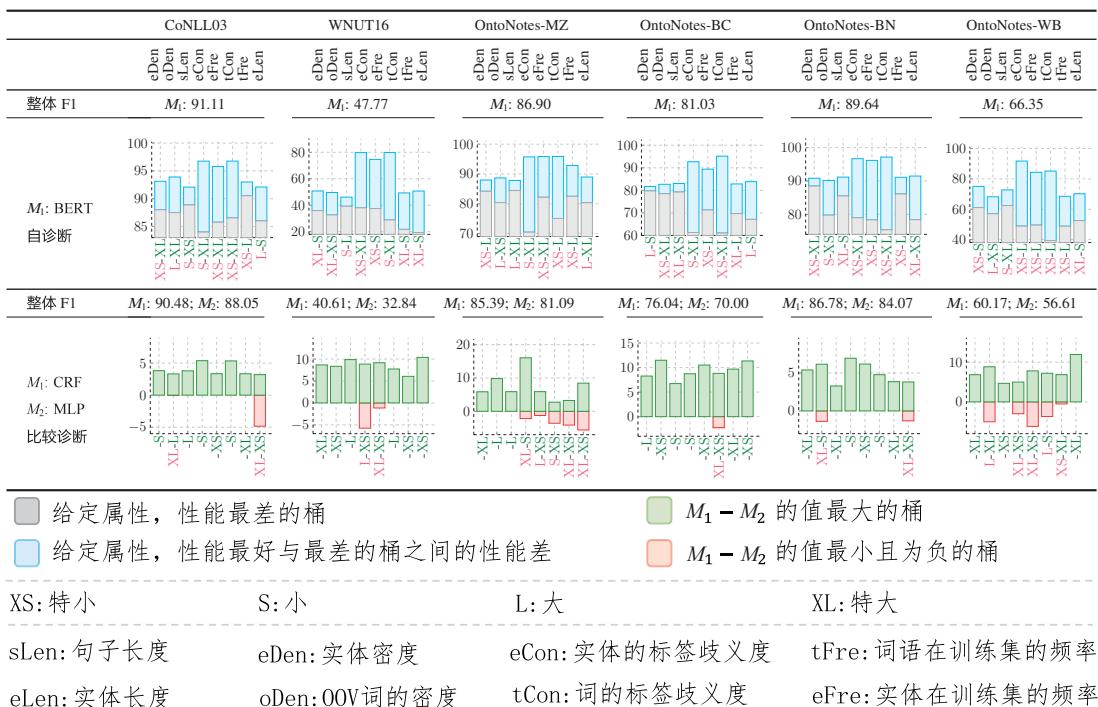
在可解释评估基础上, 文献 [778] 中还提出了两种模型诊断 (Model Diagnosis) 方法:

- (1) 自我诊断: 给定模型和特定的评估属性 (例如, 实体长度), 获得测试样本的性能在其中获得最高值和最低值的桶。可以帮助诊断特定模型在哪些条件下表现良好或较差;
- (2) 比较诊断: 给定两个模型  $M_1$  和  $M_2$  以及特定属性, 对比两个系统之间的性能差距达到最高值和最低值的桶。可以指示系统之间在哪些条件下可能优势和劣势。

图14.12给出了在 6 个命名实体识别数据集中的模型诊断结果, 其中  $M_1$  和  $M_2$  表示两个模型。属性值被分为四类: 特小 (XS)、小 (S)、大 (L) 和特大 (XL)。在自我诊断直方图中, 绿色 (红色) 的 X 轴刻度标签表示系统在该桶上获得最佳 (最差) 性能。灰色柱子表示性能最差, 蓝色的

图 14.11 命名实体识别任务可解释评估示例<sup>[778]</sup>

柱子表示最佳性能和最差性能之间的差距。通过图14.12所给出的基于BERT的命名实体模型的自我诊断的结果可以观察到，对于实体在训练集和测试集上的标签一致程度（eCon, tCon）较低的情况下或者实体频率（eFre）较低的情况下，模型的结果较差。通过CRF和MLP的对比诊断，还可以发现CRF在长实体上相较于MLP有更好的性能。

图 14.12 命名实体识别任务模型诊断结果示例<sup>[778]</sup>

可解释评估的分析结果相较于传统的单一评价指标，可以更好的进行模型错误类型统计、归类与分析，从而对模型的性能来源做出解释。错误样本分类的依据可以是语法、语义或任务相关

的特征。具有能够发现某个任务的困难样本类型（例如，未登录词、标签不一致）；能够对比不同模型的优缺点（例如，模型 A 比模型 B 在某类型样本上错误更少）；还能够通过对比发现不同结构所起的作用（例如，基于 LSTM 的模型比基于 CNN 的模型在句子较长的样本上错误更少，说明 LSTM 能更好地建模长距离依赖关系）。在此基础上，文献 [779] 还提出了可解释排行榜 Explaina Board，使研究人员可以采用人机交互的评估方式，利用模型自我诊断、系统辅助诊断和数据偏差分析等可解释评估方法，对模型优势和劣势以及数据集的偏差进行更细粒度的分析。

## 14.4 延伸阅读

随着人工智能技术的发展，数据驱动的算法对经济社会发展以及人们日常生活都带来了深远的影响。为了更好的控制与理解模型，关于透明性，可解释性等问题的研究近年来得到了广泛的关注。本章概述了可解释人工智能以及可解释自然语言处理的基本方法。近年来可解释人工智能更详尽的综述可参看<sup>[750]</sup>。除此之外，可解释技术与以下方向的发展也紧密相关。

- 文本偏见分析。做为可解释性分析的一个角度，研究发现自然语言处理模型的预测结果会受到训练数据偏差的影响。例如在词向量表示中发现“men”与“engineer”的相似度显著的超过“women”与“engineer”的距离 [780]。这样的偏见影响了包括共指消解 [781]，机器翻译 [782] 等系统的预测结果。如何定义文本中的偏见 [783]，探索文本偏见与现实世界的关系 [784] 等问题都值得更进一步的研究。
- 算法公平性是另一个与可解释性密切相连的研究课题。算法公平性主要关注机器决策过程对属于不同类别样本的偏差。定位偏差，提升模型公平性相关工作可以参见 [785]。
- 隐私与安全技术。对自然语言模型的可解释性的探索可以帮助理解模型中暗含的隐私问题：提升模型的可解释性能够提升模型保护隐私与对抗攻击的能力。与隐私与安全相关的工作可以参见 [786, 787]。

## 14.5 习题

- (1) 请尝试分析使用注意力的可解释性分析可能存在的缺陷，并设计实验证明。
- (2) 在使用探针分类器来解释隐层变量是否包含某种信息时，一个可变因素为探针分类器的容量——参数量越大，设计越精细的模型往往带来更好的分类性能，但又偏离了解释的目的。请设计实验探索探针分类器容量与解释性分析可靠性的关系。
- (3) 请分析影响力函数计算的复杂度。
- (4) 当删除多个样本点时探索影响力函数应该如何计算？请设计实验探索删除多样本点情况下影响力函数的可靠性。
- (5) 请尝试使用 ExplainaBoard 分析一个你所熟悉的自然语言处理模型，并尝试通过分析结论进行模型改进。

## 参考文献

- [1] 吴立德. 大规模中文文本处理[M]. 复旦大学出版社, 1997.
- [2] Turing A M. Computing machinery and intelligence[M]//Parsing the turing test. Springer, 2009: 23-65.
- [3] Kupiec J. Augmenting a hidden markov model for phrase-dependent word tagging[C]//Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989. 1989.
- [4] Oshika B, Machi F, Evans B, et al. Computational techniques for improved name search[C]//Second Conference on Applied Natural Language Processing. 1988: 203-210.
- [5] Sahami M. Learning limited dependence bayesian classifiers.[C]//KDD: volume 96. 1996: 335-338.
- [6] Yang Y. Feature selection in statistical learning of text categorization[C]//Proc. 14th International Conference on Machine Learning. 1997: 412-420.
- [7] Vapnik V. The nature of statistical learning theory[M]. Springer science & business media, 1999.
- [8] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification[C]//IJCAI-99 workshop on machine learning for information filtering: volume 1. Stockholm, Sweden, 1999: 61-67.
- [9] Nakamura M, Maruyama K, Kawabata T, et al. Neural network approach to word category prediction for english texts[C]//COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics. 1990.
- [10] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. 2001: 282-289.

- [11] Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C]//Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002). 2002: 1-8.
- [12] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786):504-507.
- [13] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- [14] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE):2493-2537.
- [15] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.
- [16] Nallapati R, Zhou B, dos Santos C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. CoNLL 2016, 2016:280.
- [17] Qiu M, Li F L, Wang S, et al. Alime chat: A sequence to sequence and rerank based chatbot engine[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017: 498-503.
- [18] Lei W, Jin X, Kan M Y, et al. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1437-1447.
- [19] Konstas I, Iyer S, Yatskar M, et al. Neural amr: Sequence-to-sequence models for parsing and generation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 146-157.
- [20] Yin J, Jiang X, Lu Z, et al. Neural generative question answering[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2972-2978.
- [21] Shi X, Huang H, Jian P, et al. Neural chinese word segmentation as sequence to sequence translation [C]//Chinese National Conference on Social Media Processing. Springer, 2017: 91-103.
- [22] Mikolov T, Karafiat M, Burget L, et al. Recurrent neural network based language model.[C]// Interspeech: volume 2. Makuhari, 2010: 1045-1048.

- [23] Sundermeyer M, Schlüter R, Ney H. Lstm neural networks for language modeling[C]//Thirteenth annual conference of the international speech communication association. 2012.
- [24] Irsoy O, Cardie C. Deep recursive neural networks for compositionality in language[J]. Advances in neural information processing systems, 2014, 27.
- [25] Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks[C/OL]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015: 103-112. <https://aclanthology.org/N15-1011>. DOI: 10.3115/v1/N15-1011.
- [26] Velivcković P, Cucurull G, Casanova A, et al. Graph attention networks[C]//International Conference on Learning Representations. 2018.
- [27] Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for chinese ner[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 1040-1050.
- [28] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers): volume 1. 2018: 2227-2237.
- [29] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.
- [30] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8):9.
- [31] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
- [32] Zhang Z, Han X, Liu Z, et al. Ernie: Enhanced language representation with informative entities [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1441-1451.
- [33] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

- [34] Kess J F, Hoppe R A. Ambiguity in psycholinguistics[M]. Benjamins Amsterdam, 1981.
- [35] 中国社会科学院语言研究所词典编辑室. 现代汉语词典（第7版）[M]. 商务印书馆, 2019.
- [36] Miller G A. Wordnet: a lexical database for english[J]. Communications of the ACM, 1995, 38(11):39-41.
- [37] 冯志伟. 论歧义结构的潜在性[J]. 中文信息学报, 1995, 9(4):14-24.
- [38] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013.
- [39] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8):1798-1828.
- [40] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33:1877-1901.
- [41] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
- [42] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. arXiv preprint arXiv:2210.11416, 2022.
- [43] Fromkin V, Rodman R, Hyams N. An introduction to language[M]. Cengage Learning, 2018.
- [44] Francis W N. A tagged corpus—problems and prospects[J]. Studies in English linguistics for Randolph Quirk, 1980:192-209.
- [45] Jurafsky D, Martin J H. Speech and language processing: An introduction to speechrecognition, natural language processing and computational linguistics[M]. 2nd ed. Pearson, 2008.
- [46] Porter M F. An algorithm for suffix stripping[J]. Program, 1980.
- [47] 俞士汶, 段慧明, 朱学锋, 等. 北大语料库加工规范: 切分·词性标注·注音[M]. 北京大学计算语言学研究所, 2003.
- [48] 黄昌宁, 李玉梅, 朱晓丹. 中文文本标注规范(5.0版)[M]. 微软亚洲研究院, 2006.
- [49] 刘开瑛. 中文文本自动分词和标注[M]. 商务印书馆, 2000.
- [50] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3):8-19.
- [51] 李航. 统计学习方法(第二版) [M]. 清华大学出版社, 2019.

- [52] Zhang Y, Clark S. Chinese segmentation with a word-based perceptron algorithm[C]//Proceedings of the 45th annual meeting of the association of computational linguistics. 2007: 840-847.
- [53] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.
- [54] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with lstm[J]. Neural computation, 2000, 12(10):2451-2471.
- [55] 邱锡鹏. 神经网络与深度学习[M/OL]. 北京: 机械工业出版社, 2020. <https://nndl.github.io/>.
- [56] Rabiner L R. A tutorial on hidden markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2):257-286.
- [57] 张虎, 郑家恒, 刘江. 语料库词性标注一致性检查方法研究[J]. 中文信息学报, 2004, 18(5): 12-17.
- [58] Brill E. A simple rule-based part of speech tagger[C/OL]//ANLC '92: Proceedings of the Third Conference on Applied Natural Language Processing. USA: Association for Computational Linguistics, 1992: 152–155. <https://doi.org/10.3115/974499.974526>.
- [59] Bahl L, Brown P, De Souza P, et al. Maximum mutual information estimation of hidden markov model parameters for speech recognition[C]//ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing: volume 11. IEEE, 1986: 49-52.
- [60] Petrov S, Das D, McDonald R. A universal part-of-speech tagset[C]//Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012: 2089-2096.
- [61] Yang F, Vozila P. Semi-supervised chinese word segmentation using partial-label learning with conditional random fields[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 90-98.
- [62] Fujii R, Domoto R, Mochihashi D. Nonparametric bayesian semi-supervised word segmentation [J]. Transactions of the Association for Computational Linguistics, 2017, 5:179-189.
- [63] Zeng X, Wong D F, Chao L S, et al. Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 770-779.

- [64] Zhang L, Wang H, Sun X, et al. Exploring representations from unlabeled data with co-training for chinese word segmentation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 311-321.
- [65] Zhang Q, Liu X, Fu J. Neural networks incorporating dictionaries for chinese word segmentation [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.
- [66] Liu J, Wu F, Wu C, et al. Neural chinese word segmentation with lexicon and unlabeled data via posterior regularization[C]//The World Wide Web Conference. 2019: 3013-3019.
- [67] Zhao X, Yang M, Qu Q, et al. Improving neural chinese word segmentation with lexicon-enhanced adaptive attention[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1953-1956.
- [68] Chen X, Shi Z, Qiu X, et al. Adversarial multi-criteria learning for chinese word segmentation[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1193-1203.
- [69] Gong J, Chen X, Gui T, et al. Switch-lstms for multi-criteria chinese word segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 6457-6464.
- [70] Zhang M, Zhang Y, Fu G. Transition-based neural word segmentation[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 421-431. <https://www.aclweb.org/anthology/P16-1040>. DOI: 10.18653/v1/P16-1040.
- [71] Yang J, Zhang Y, Liang S. Subword encoding in lattice lstm for chinese word segmentation[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 2720-2725.
- [72] Fu J, Liu P, Zhang Q, et al. Rethinkcws: Is chinese word segmentation a solved task?[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 5676-5686.
- [73] Wang X, Liu Q, Gui T, et al. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021: 347-355.

- [74] Mann G S, McCallum A. Simple, robust, scalable semi-supervised learning via expectation regularization[C]//Proceedings of the 24th international conference on Machine learning. 2007: 593-600.
- [75] Hajic J, Raab J, Spousta M, et al. Semi-supervised training for the averaged perceptron pos tagger [C]//Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). 2009: 763-771.
- [76] Sogaard A. Semi-supervised condensed nearest neighbor for part-of-speech tagging[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 48-52.
- [77] Gadde A, Anis A, Ortega A. Active semi-supervised learning using sampling theory for graph signals[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 492-501.
- [78] Schnabel T, Schütze H. Flors: Fast and simple domain adaptation for part-of-speech tagging[J]. Transactions of the Association for Computational Linguistics, 2014, 2:15-26.
- [79] Song Y, Klassen P, Xia F, et al. Entropy-based training data selection for domain adaptation[C]// Proceedings of COLING 2012: Posters. 2012: 1191-1200.
- [80] Finkel J R, Manning C D. Hierarchical bayesian domain adaptation[C]//Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics. 2009: 602-610.
- [81] Gui T, Zhang Q, Huang H, et al. Part-of-speech tagging for twitter with adversarial neural networks [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2411-2420.
- [82] Liu M, Song Y, Zou H, et al. Reinforced training data selection for domain adaptation[C]// Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1957-1968.
- [83] Ng H T, Low J K. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based?[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 277-284.
- [84] Jiang W, Huang L, Liu Q, et al. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging[C]//Proceedings of ACL-08: HLT. 2008: 897-904.

- [85] Jiang W, Mi H, Liu Q. Word lattice reranking for chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). 2008: 385-392.
- [86] Sun W. A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 1385-1394.
- [87] Tian Y, Song Y, Ao X, et al. Joint chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8286-8296.
- [88] 维多利亚·弗罗姆金(著), 罗伯特·罗德曼(著), 王大惟(译), 等. 语言引论(第八版)[M]. 北京大学出版社, 2017.
- [89] 崔应贤. 现代汉语语法学习与研究入门[M]. 清华大学出版社有限公司, 2004.
- [90] Chomsky N. Syntactic structures[M]. De Gruyter Mouton, 2009.
- [91] Lucien T. Eléments de syntaxe structurale[J]. Paris, Klincksieck, 1959:25.
- [92] Carroll J, Briscoe T, Sanfilippo A. Parser evaluation: a survey and a new proposal[C]//Proceedings of the 1st International Conference on Language Resources and Evaluation: volume 32. Granada, 1998.
- [93] De Marneffe M C, MacCartney B, Manning C D, et al. Generating typed dependency parses from phrase structure parses.[C]//Lrec: volume 6. 2006: 449-454.
- [94] Cocke J. Programming languages and their compilers: Preliminary notes[M]. New York University, 1969.
- [95] Younger D H. Recognition and parsing of context-free languages in time n<sup>3</sup>[J]. Information and control, 1967, 10(2):189-208.
- [96] Kasami T. An efficient recognition and syntax-analysis algorithm for context-free languages[J]. Coordinated Science Laboratory Report no. R-257, 1966.
- [97] Ullman J D. The theory of parsing, translation, and compiling[M]. Prentice-Hall, 1972.
- [98] Lari K, Young S J. The estimation of stochastic context-free grammars using the inside-outside algorithm[J]. Computer speech & language, 1990, 4(1):35-56.

- [99] Manning C, Schütze H. Foundations of statistical natural language processing[M]. MIT press, 1999.
- [100] Black E, Abney S, Flickinger D, et al. A procedure for quantitatively comparing the syntactic coverage of English grammars[C]//Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991. 1991.
- [101] Kubler S, McDonald R, Nivre J. Dependency parsing[M]. Morgan & Claypool Publishers, 2009.
- [102] Chu Y J. On the shortest arborescence of a directed graph[J]. Scientia Sinica, 1965, 14:1396-1400.
- [103] Edmonds J. Optimum branchings[J]. Journal of Research of the National Bureau of Standards, B, 1967, 71:233-240.
- [104] McDonald R, Pereira F, Ribarov K, et al. Non-projective dependency parsing using spanning tree algorithms[C]//Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005: 523-530.
- [105] Pei W, Ge T, Chang B. An effective neural network model for graph-based dependency parsing [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 313-322.
- [106] Dozat T, Manning C D. Deep biaffine attention for neural dependency parsing[C/OL]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. <https://openreview.net/forum?id=Hk95PK9le>.
- [107] Ji T, Wu Y, Lan M. Graph-based dependency parsing with graph neural networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2475-2485.
- [108] Nivre J. Algorithms for deterministic incremental dependency parsing[J]. Computational Linguistics, 2008, 34(4):513-553.
- [109] Nivre J, Hall J, Nilsson J. Maltparser: A data-driven parser-generator for dependency parsing.[C]//LREC: volume 6. 2006: 2216-2219.
- [110] Nivre J, Hall J, Nilsson J. Memory-based dependency parsing[C]//Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. 2004: 49-56.

- [111] Che W, Li Z, Hu Y, et al. A cascaded syntactic and semantic dependency parsing system[C]//CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning. 2008: 238-242.
- [112] Chen D, Manning C D. A fast and accurate dependency parser using neural networks[C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 740-750.
- [113] Dyer C, Ballesteros M, Ling W, et al. Transition-based dependency parsing with stack long short-term memory[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 334-343.
- [114] Yamada H, Matsumoto Y. Statistical dependency analysis with support vector machines[C]// Proceedings of the eighth international conference on parsing technologies. 2003: 195-206.
- [115] De Marneffe M C, Manning C D. The stanford typed dependencies representation[C]//Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation. 2008: 1-8.
- [116] De Marneffe M C, Dozat T, Silveira N, et al. Universal stanford dependencies: A cross-linguistic typology.[C]//LREC: volume 14. 2014: 4585-4592.
- [117] Zeman D. Reusable tagset conversion using tagset drivers[C]//Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008.
- [118] Che W, Shao Y, Liu T, et al. Semeval-2016 task 9: Chinese semantic dependency parsing[C]// Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 2016: 1074-1080.
- [119] Grover C, Tobin R. Rule-based chunking and reusability[C]//Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC' 06). 2006.
- [120] Koeling R. Chunking with maximum entropy models[C]//Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop. 2000.
- [121] Kudo T, Matsumoto Y. Chunking with support vector machines[C]//Second Meeting of the North American Chapter of the Association for Computational Linguistics. 2001.
- [122] Sha F, Pereira F. Shallow parsing with conditional random fields[C]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 213-220.

- [123] Henderson J. Inducing history representations for broad coverage statistical parsing[C]// Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 103-110.
- [124] Stern M, Andreas J, Klein D. A minimal span-based neural constituency parser[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 818-827.
- [125] Gaddy D, Stern M, Klein D. What's going on in neural constituency parsers? an analysis[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 999-1010.
- [126] Vinyals O, Kaiser L, Koo T, et al. Grammar as a foreign language[J]. Advances in neural information processing systems, 2015, 28.
- [127] Liu L, Zhu M, Shi S. Improving sequence-to-sequence constituency parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.
- [128] Kitaev N, Klein D. Constituency parsing with a self-attentive encoder[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2676-2686.
- [129] Tian Y, Song Y, Xia F, et al. Improving constituency parsing with span attention[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1691-1703.
- [130] Kitaev N, Cao S, Klein D. Multilingual constituency parsing with self-attention and pre-training [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3499-3505.
- [131] Matsuzaki T, Miyao Y, Tsujii J. Probabilistic cfg with latent annotations[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL' 05). 2005: 75-82.
- [132] Petrov S, Barrett L, Thibaux R, et al. Learning accurate, compact, and interpretable tree annotation[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006: 433-440.
- [133] Huang Z, Harper M. Self-training pcfg grammars with latent annotations across languages[C]// Proceedings of the 2009 conference on empirical methods in natural language processing. 2009: 832-841.

- [134] Liu J, Zhang Y. In-order transition-based constituent parsing[J]. Transactions of the Association for Computational Linguistics, 2017, 5:413-424.
- [135] Goto I, Utiyama M, Sumita E, et al. Preordering using a target-language parser via cross-language syntactic projection for statistical machine translation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2015, 14(3):1-23.
- [136] Tiedemann J, Agić vZ, Nivre J. Treebank translation for cross-lingual parser induction[C]// Proceedings of the Eighteenth Conference on Computational Natural Language Learning. 2014: 130-140.
- [137] Zeman D, Resnik P. Cross-language parser adaptation between related languages[C]//Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. 2008.
- [138] Shen Y, Lin Z, Huang C w, et al. Neural language modeling by jointly learning syntax and lexicon [C]//International Conference on Learning Representations. 2018.
- [139] Drozdov A, Verga P, Yadav M, et al. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 1129-1141.
- [140] Jin L, Doshi-Velez F, Miller T, et al. Unsupervised learning of pcfgs with normalizing flow[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2442-2452.
- [141] Klein D, Manning C D. Corpus-based induction of syntactic structure: Models of dependency and constituency[C]//Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04). 2004: 478-485.
- [142] Headden III W P, Johnson M, McClosky D. Improving unsupervised dependency parsing with richer contexts and smoothing[C]//Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics. 2009: 101-109.
- [143] Spitkovsky V I, Alshawi H, Jurafsky D. Punctuation: Making a point in unsupervised dependency parsing[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. 2011: 19-28.

- [144] Spitkovsky V I, Alshawi H, Jurafsky D, et al. Viterbi training improves unsupervised dependency parsing[C/OL]//Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Uppsala, Sweden: Association for Computational Linguistics, 2010: 9-17. <https://aclanthology.org/W10-2902>.
- [145] Pate J K, Goldwater S. Unsupervised dependency parsing with acoustic cues[J]. Transactions of the Association for Computational Linguistics, 2013, 1:63-74.
- [146] Cai J, Jiang Y, Tu K. Crf autoencoder for unsupervised dependency parsing[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1638-1643.
- [147] Koo T, Carreras X, Collins M. Simple semi-supervised dependency parsing[C/OL]//Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics, 2008: 595-603. <https://aclanthology.org/P08-1068>.
- [148] Sogaard A, Rishoj C. Semi-supervised dependency parsing using generalized tri-training[C]// Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). 2010: 1065-1073.
- [149] Chen W, Zhang M, Zhang Y. Semi-supervised feature transformation for dependency parsing[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1303-1313.
- [150] Li Z, Zhang M, Chen W. Ambiguity-aware ensemble training for semi-supervised dependency parsing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 457-467.
- [151] Kiperwasser E, Goldberg Y. Semi-supervised dependency parsing using blexical contextual features from auto-parsed data[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1348-1353.
- [152] Wang G, Tu K. Semi-supervised dependency parsing with arc-factored variational autoencoding [C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 2485-2496.
- [153] Ahmad W, Zhang Z, Ma X, et al. Cross-lingual dependency parsing with unlabeled auxiliary languages[C]//Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019: 372-382.

- [154] Huang K H, Ahmad W, Peng N, et al. Improving zero-shot cross-lingual transfer learning via robust training[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 1684-1697.
- [155] Ji T, Jiang Y, Wang T, et al. Word reordering for zero-shot cross-lingual structured prediction[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 4109-4120.
- [156] Schuster T, Ram O, Barzilay R, et al. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 1599-1613.
- [157] Ma X, Xia F. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 1337-1348.
- [158] Wang Y, Che W, Guo J, et al. Cross-lingual bert transformation for zero-shot dependency parsing [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 5721-5727.
- [159] Kong L, Schneider N, Swayamdipta S, et al. A dependency parser for tweets[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1001-1012.
- [160] Zhang M, Zhang Y, Fu G. Cross-lingual dependency parsing using code-mixed treebank[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 997-1006.
- [161] Li Z, Cai J, He S, et al. Seq2seq dependency parsing[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3203-3214.
- [162] Zhang Y, Li Z, Zhang M. Efficient second-order treecrf for neural dependency parsing[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3295-3305.
- [163] 何三本, 王玲玲. 现代语义学[M]. 台北: 三民书局, 1995.

- [164] Wierzbicka A. Semantic primitives[J]. 1972.
- [165] Dong Z, Dong Q. Hownet-a hybrid language and knowledge resource[C]//International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003. IEEE, 2003: 820-824.
- [166] Fillmore C J, Atkins B T. Toward a frame-based lexicon: The semantics of risk and its neighbors [J]. Frames, fields and contrasts: New essays in semantic and lexical organization, 1992, 75:102.
- [167] Baker C F, Fillmore C J, Lowe J B. The berkeley framenet project[C]//COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics. 1998.
- [168] 李福印. 语义学概论[M]. 北京大学出版社, 2006.
- [169] 梅德明. 语言学与应用语言学百科全书[M]. 北京大学出版社;, 2017.
- [170] Sundheim B M. Overview of the third message understanding evaluation and conference[C]// Proceedings of the 3rd conference on Message understanding. 1991: 3-16.
- [171] Hendrix G G. Expanding the utility of semantic networks through partitioning[C]//Proceedings of the 4th international joint conference on Artificial intelligence-Volume 1. 1975: 115-121.
- [172] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.
- [173] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model[J]. Advances in neural information processing systems, 2000, 13.
- [174] Almeida F, Xexéo G. Word embeddings: A survey[J]. arXiv preprint arXiv:1901.09069, 2019.
- [175] Sahlgren M. The distributional hypothesis[J]. Italian Journal of Disability Studies, 2008, 20:33-53.
- [176] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American society for information science, 1990, 41(6):391-407.
- [177] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the association for computational linguistics, 2015, 3:211-225.
- [178] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [C]//ICLR. 2013.

- [179] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
- [180] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL), 2016: 1715-1725.
- [181] Agirre E, Alfonseca E, Hall K, et al. A study on similarity and relatedness using distributional and wordnet-based approaches[J]. 2009.
- [182] Hill F, Reichart R, Korhonen A. SimLex-999: Evaluating semantic models with (genuine) similarity estimation[J/OL]. Computational Linguistics, 2015, 41(4):665-695. <https://aclanthology.org/J15-4004>.
- [183] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C/OL]//Cortes C, Lawrence N, Lee D, et al. Advances in Neural Information Processing Systems: volume 28. Curran Associates, Inc., 2015. <https://proceedings.neurips.cc/paper/2015/file/f442d33fa06832082290ad8544a8da27-Paper.pdf>.
- [184] Pagliardini M, Gupta P, Jaggi M. Unsupervised learning of sentence embeddings using compositional n-gram features[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 528-540. <https://aclanthology.org/N18-1049>. DOI: 10.18653/v1/N18-1049.
- [185] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2016.
- [186] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[C/OL]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 427-431. <https://aclanthology.org/E17-2068>.
- [187] Gale W A, Church K W, Yarowsky D. A method for disambiguating word senses in a large corpus [J]. Computers and the Humanities, 1992, 26(5):415-439.
- [188] Melamud O, Goldberger J, Dagan I. context2vec: Learning generic context embedding with bidirectional lstm[J]. conference on computational natural language learning, 2016.

- [189] Loureiro D, Jorge A. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5682-5691. <https://aclanthology.org/P19-1569>. DOI: 10.18653/v1/P19-1569.
- [190] Blevins T, Zettlemoyer L. Moving down the long tail of word sense disambiguation with gloss-informed biencoders[J]. arXiv preprint arXiv:2005.02590, 2020.
- [191] Huang L, Sun C, Qiu X, et al. GlossBERT: BERT for word sense disambiguation with gloss knowledge[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 3509-3514. <https://aclanthology.org/D19-1355>. DOI: 10.18653/v1/D19-1355.
- [192] Levine Y, Lenz B, Dagan O, et al. SenseBERT: Driving some sense into BERT[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 4656-4667. <https://aclanthology.org/2020.acl-main.423>. DOI: 10.18653/v1/2020.acl-main.423.
- [193] Loureiro D, Jorge A. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation[J]. arXiv preprint arXiv:1906.10007, 2019.
- [194] Miller G A, Chodorow M, Landes S, et al. Using a semantic concordance for sense identification [C]//Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994. 1994.
- [195] Petrolito T, Bond F. A survey of WordNet annotated corpora[C/OL]//Proceedings of the Seventh Global Wordnet Conference. Tartu, Estonia: University of Tartu Press, 2014: 236-245. <https://aclanthology.org/W14-0132>.
- [196] Taghipour K, Ng H T. One million sense-tagged instances for word sense disambiguation and induction[C]//Proceedings of the nineteenth conference on computational natural language learning. 2015: 338-344.
- [197] Raganato A, Camacho-Collados J, Navigli R. Word sense disambiguation: A unified evaluation framework and empirical comparison[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017: 99-110.

- [198] He L, Lee K, Levy O, et al. Jointly predicting predicates and arguments in neural semantic role labeling[C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 364-369. <https://aclanthology.org/P18-2058>. DOI: 10.18653/v1/P18-2058.
- [199] Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling[C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1506-1515. <https://aclanthology.org/D17-1159>. DOI: 10.18653/v1/D17-1159.
- [200] Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles [J]. Computational linguistics, 2005, 31(1):71-106.
- [201] Meyers A, Reeves R, Macleod C, et al. The nombank project: An interim report[C]//Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004. 2004: 24-31.
- [202] Carreras X, Màrquez L. Introduction to the CoNLL-2004 shared task: Semantic role labeling[C/OL]//Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004: 89-97. <https://aclanthology.org/W04-2412>.
- [203] Zhang Z, Strubell E, Hovy E. Comparing span extraction methods for semantic role labeling [C/OL]//Proceedings of the 5th Workshop on Structured Prediction for NLP (SPNLP 2021). Online: Association for Computational Linguistics, 2021: 67-77. <https://aclanthology.org/2021.spnlp-1.8>. DOI: 10.18653/v1/2021.spnlp-1.8.
- [204] Hajic J, Ciaramita M, Johansson R, et al. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages[C/OL]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. Boulder, Colorado: Association for Computational Linguistics, 2009: 1-18. <https://aclanthology.org/W09-1201>.
- [205] Li Z, Zhao H, Zhou J, et al. Dependency and span, cross-style semantic role labeling on propbank and nombank[J/OL]. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 2022. <https://doi.org/10.1145/3526214>.
- [206] Preda S, Emerson G. Using dependency parsing for few-shot learning in distributional semantics [C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Dublin, Ireland: Association for Computational Linguistics, 2022: 461-466. <https://aclanthology.org/2022.acl-srw.38>. DOI: 10.18653/v1/2022.acl-srw.38.

- [207] Munir K, Zhao H, Li Z. Neural unsupervised semantic role labeling[J/OL]. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 2021, 20(6). <https://doi.org/10.1145/3461613>.
- [208] Zhang Z, Strubell E, Hovy E. On the benefit of syntactic supervision for cross-lingual transfer in semantic role labeling[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 6229-6246. <https://aclanthology.org/2021.emnlp-main.503>. DOI: 10.18653/v1/2021.emnlp-main.503.
- [209] Conia S, Bacciu A, Navigli R. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources[C/OL]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 338-351. <https://aclanthology.org/2021.naacl-main.31>. DOI: 10.18653/v1/2021.naacl-main.31.
- [210] Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: What we know about how bert works [J]. Transactions of the Association for Computational Linguistics, 2021, 8:842-866.
- [211] Tenney I, Das D, Pavlick E. Bert rediscovers the classical nlp pipeline[J]. arXiv preprint arXiv:1905.05950, 2019.
- [212] Hewitt J, Manning C D. A structural probe for finding syntax in word representations[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4129-4138.
- [213] Clark K, Khandelwal U, Levy O, et al. What does bert look at? an analysis of bert's attention[J]. arXiv preprint arXiv:1906.04341, 2019.
- [214] Yenicelik D, Schmidt F, Kilcher Y. How does bert capture semantics? a closer look at polysemous words[C]//Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020: 156-162.
- [215] Zhang Z, Wu Y, Zhao H, et al. Semantics-aware bert for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 9628-9635.
- [216] Bai J, Wang Y, Chen Y, et al. Improving pre-trained transformers with syntax trees[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Kiev, Ukraine. 2021: 21-23.

- [217] Sachan D, Zhang Y, Qi P, et al. Do syntax trees help pre-trained transformers extract information? [C/OL]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 2021. <https://aclanthology.org/2021.eacl-main.228>. DOI: 10.18653/v1/2021.eacl-main.228.
- [218] Sun K, Li Z, Zhao H. Multilingual pre-training with universal dependency learning[J]. Advances in Neural Information Processing Systems, 2021, 34:8444-8456.
- [219] McDonald R, Nivre J, Quirkbach-Brundage Y, et al. Universal Dependency annotation for multilingual parsing[C/OL]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Sofia, Bulgaria: Association for Computational Linguistics, 2013: 92-97. <https://aclanthology.org/P13-2017>.
- [220] De Beaugrande R A, Dressler W U. Introduction to text linguistics: volume 1[M]. longman London, 1981.
- [221] Halliday M A K, Hasan R. Cohesion in english[M]. Routledge, 2014.
- [222] 苗兴伟, 张蕾. 汉语语篇分析[M]. 外语教学与研究出版社, 2021.
- [223] Van Dijk T A. News analysis[J]. Case Studies of International and National News in the Press. New Jersey: Lawrence, 1988.
- [224] Mann W C, Thompson S A. Rhetorical structure theory: A theory of text organization[M]. University of Southern California, Information Sciences Institute Los Angeles, 1987.
- [225] Hoey M. Textual interaction: An introduction to written discourse analysis[M]. Psychology Press, 2001.
- [226] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information [C/OL]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 228-235. <https://aclanthology.org/N03-1030>.
- [227] Magerman D M. Statistical decision-tree models for parsing[J]. arXiv preprint cmp-lg/9504030, 1995.
- [228] Carlson L, Marcu D, Okurovsky M E. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory[C/OL]//Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. 2001. <https://aclanthology.org/W01-1605>.

- [229] Wang Y, Li S, Yang J. Toward fast and accurate neural discourse segmentation[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 962-967. <https://aclanthology.org/D18-1116>. DOI: 10.18653/v1/D18-1116.
- [230] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 2227-2237. <https://aclanthology.org/N18-1202>. DOI: 10.18653/v1/N18-1202.
- [231] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse TreeBank 2.0.[C/OL]//Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008. [http://www.lrec-conf.org/proceedings/lrec2008/pdf/754\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf).
- [232] Hernault H, Prendinger H, du Verle D A, et al. Hilda: A discourse parser using support vector machine classification[J]. *Dialogue & Discourse*, 2010, 1(3):1-33.
- [233] Feng V W, Hirst G. Text-level discourse parsing with rich linguistic features[C/OL]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics, 2012: 60-68. <https://aclanthology.org/P12-1007>.
- [234] Li J, Li R, Hovy E. Recursive deep models for discourse parsing[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 2061-2069. <https://aclanthology.org/D14-1220>. DOI: 10.3115/v1/D14-1220.
- [235] Webber B. D-ltag: extending lexicalized tag to discourse[J]. *Cognitive Science*, 2004, 28(5):751-779.
- [236] Pitler E, Nenkova A. Using syntax to disambiguate explicit discourse connectives in text[C/OL]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore: Association for Computational Linguistics, 2009: 13-16. <https://aclanthology.org/P09-2004>.
- [237] Marcinkiewicz M A. Building a large annotated corpus of english: The penn treebank[J]. *Using Large Corpora*, 1994, 273.

- [238] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations[C/OL]//Coling 2008: Companion volume: Posters. Manchester, UK: Coling 2008 Organizing Committee, 2008: 87-90. <https://aclanthology.org/C08-2022>.
- [239] Rutherford A, Demberg V, Xue N. A systematic study of neural discourse models for implicit discourse relation[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017: 281-291.
- [240] Zhang B, Su J, Xiong D, et al. Shallow convolutional neural network for implicit discourse relation recognition[C/OL]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 2230-2235. <https://aclanthology.org/D15-1266>. DOI: 10.18653/v1/D15-1266.
- [241] Ji Y, Haffari G, Eisenstein J. A latent variable recurrent neural network for discourse-driven language models[C/OL]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 332-342. <https://aclanthology.org/N16-1037>. DOI: 10.18653/v1/N16-1037.
- [242] Shi W, Demberg V. Next sentence prediction helps implicit discourse relation classification within and across domains[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 5790-5796. <https://aclanthology.org/D19-1586>. DOI: 10.18653/v1/D19-1586.
- [243] Ji Y, Cohn T, Kong L, et al. Document context language models[J]. arXiv preprint arXiv:1511.03962, 2015.
- [244] Bengtson E, Roth D. Understanding the value of features for coreference resolution[C/OL]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics, 2008: 294-303. <https://aclanthology.org/D08-1031>.
- [245] Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases[J/OL]. Computational Linguistics, 2001, 27(4):521-544. <https://aclanthology.org/J01-4004>.

- [246] Ng V, Cardie C. Improving machine learning approaches to coreference resolution[C/OL]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 104-111. <https://aclanthology.org/P02-1014>. DOI: 10.3115/1073083.1073102.
- [247] Clark K, Manning C D. Improving coreference resolution by learning entity-level distributed representations[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 643-653. <https://aclanthology.org/P16-1061>. DOI: 10.18653/v1/P16-1061.
- [248] Denis P, Baldridge J. A ranking approach to pronoun resolution.[C]//IJCAI: volume 158821593. 2007.
- [249] Lee K, He L, Lewis M, et al. End-to-end neural coreference resolution[C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 188-197. <https://aclanthology.org/D17-1018>. DOI: 10.18653/v1/D17-1018.
- [250] Rahman A, Ng V. Supervised models for coreference resolution[C/OL]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009: 968-977. <https://aclanthology.org/D09-1101>.
- [251] Wiseman S, Rush A M, Shieber S M. Learning global features for coreference resolution[C/OL]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 994-1004. <https://aclanthology.org/N16-1114>. DOI: 10.18653/v1/N16-1114.
- [252] Wiseman S, Rush A M, Shieber S, et al. Learning anaphoricity and antecedent ranking features for coreference resolution[C/OL]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015: 1416-1426. <https://aclanthology.org/P15-1137>. DOI: 10.3115/v1/P15-1137.
- [253] Durrett G, Klein D. Easy victories and uphill battles in coreference resolution[C/OL]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013: 1971-1982. <https://aclanthology.org/D13-1203>.

- [254] Li J, Liu M, Qin B, et al. A survey of discourse parsing[J]. *Frontiers of Computer Science*, 2022, 16(5):165329.
- [255] Marcu D. The automatic construction of large-scale corpora for summarization research[C]// Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999: 137-144.
- [256] Feng V W, Hirst G. Text-level discourse parsing with rich linguistic features[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2012: 60-68.
- [257] Feng V W, Hirst G. A linear-time bottom-up discourse parser with constraints and post-editing [C/OL]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014: 511-521. <https://aclanthology.org/P14-1048>. DOI: 10.3115/v1/P14-1048.
- [258] Ji Y, Eisenstein J. Representation learning for text-level discourse parsing[C/OL]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014: 13-24. <https://aclanthology.org/P14-1002>. DOI: 10.3115/v1/P14-1002.
- [259] Li Q, Li T, Chang B. Discourse parsing with attention-based hierarchical neural networks[C/OL]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 362-371. <https://aclanthology.org/D16-1035>. DOI: 10.18653/v1/D16-1035.
- [260] Wang Y, Li S, Wang H. A two-stage parsing method for text-level discourse analysis[C/OL]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 184-188. <https://aclanthology.org/P17-2029>. DOI: 10.18653/v1/P17-2029.
- [261] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information[C]// Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 228-235.
- [262] Joty S, Carenini G, Ng R T. CODRA: A novel discriminative framework for rhetorical analysis [J/OL]. *Computational Linguistics*, 2015, 41(3):385-435. <https://aclanthology.org/J15-3002>.

- [263] Kobayashi N, Hirao T, Kamigaito H, et al. Top-down rst parsing utilizing granularity levels in documents[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05):8099-8106. <https://ojs.aaai.org/index.php/AAAI/article/view/6321>. DOI: 10.1609/aaai.v34i05.6321.
- [264] Zhang L, Xing Y, Kong F, et al. A top-down neural architecture towards text-level parsing of discourse rhetorical structure[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6386-6395. <https://aclanthology.org/2020.acl-main.569>. DOI: 10.18653/v1/2020.acl-main.569.
- [265] Koto F, Lau J H, Baldwin T. Top-down discourse parsing via sequence labelling[C/OL]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 2021: 715-726. <https://aclanthology.org/2021.eacl-main.60>. DOI: 10.18653/v1/2021.eacl-main.60.
- [266] Braud C, Coavoux M, Sogaard A. Cross-lingual RST discourse parsing[C/OL]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 292-304. <https://aclanthology.org/E17-1028>.
- [267] Mabona A, Rimell L, Clark S, et al. Neural generative rhetorical structure parsing[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 2284-2295. <https://aclanthology.org/D19-1233>. DOI: 10.18653/v1/D19-1233.
- [268] Zhang L, Kong F, Zhou G. Adversarial learning for discourse rhetorical structure parsing[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3946-3957.
- [269] Lin X, Joty S, Jwalapuram P, et al. A unified linear-time framework for sentence-level discourse parsing[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 4190-4200. <https://aclanthology.org/P19-1410>. DOI: 10.18653/v1/P19-1410.
- [270] Guz G, Carenini G. Coreference for discourse parsing: A neural approach[C/OL]//Proceedings of the First Workshop on Computational Approaches to Discourse. Online: Association for

- Computational Linguistics, 2020: 160-167. <https://aclanthology.org/2020.codi-1.17>. DOI: 10.18653/v1/2020.codi-1.17.
- [271] Yu N, Zhang M, Fu G, et al. RST discourse parsing with second-stage EDU-level pre-training [C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 4269-4280. <https://aclanthology.org/2022.acl-long.294>. DOI: 10.18653/v1/2022.acl-long.294.
- [272] Nguyen T T, Nguyen X P, Joty S, et al. RST parsing from scratch[C/OL]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 1613-1625. <https://aclanthology.org/2021.naacl-main.128>. DOI: 10.18653/v1/2021.naacl-main.128.
- [273] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text [C/OL]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009: 683-691. <https://aclanthology.org/P09-1077>.
- [274] Rutherford A, Xue N. Improving the inference of implicit discourse relations via classifying explicit discourse connectives[C/OL]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015: 799-808. <https://aclanthology.org/N15-1081>. DOI: 10.3115/v1/N15-1081.
- [275] Wu C, Shi X, Chen Y, et al. Improving implicit discourse relation recognition with discourse-specific word embeddings[C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 269-274. <https://aclanthology.org/P17-2042>. DOI: 10.18653/v1/P17-2042.
- [276] Lan M, Xu Y, Niu Z. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition[C/OL]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, Bulgaria: Association for Computational Linguistics, 2013: 476-485. <https://aclanthology.org/P13-1047>.
- [277] Zhou Z M, Xu Y, Niu Z Y, et al. Predicting discourse connectives for implicit discourse relation recognition[C/OL]//Coling 2010: Posters. Beijing, China: Coling 2010 Organizing Committee, 2010: 1507-1514. <https://aclanthology.org/C10-2172>.

- [278] Qin L, Zhang Z, Zhao H, et al. Adversarial connective-exploiting networks for implicit discourse relation classification[C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 1006-1017. <https://aclanthology.org/P17-1093>. DOI: 10.18653/v1/P17-1093.
- [279] Chen J, Zhang Q, Liu P, et al. Implicit discourse relation detection via a deep architecture with gated relevance network[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1726-1735. <https://aclanthology.org/P16-1163>. DOI: 10.18653/v1/P16-1163.
- [280] Bai H, Zhao H. Deep enhanced representation for implicit discourse relation recognition[C/OL]// Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 571-583. <https://aclanthology.org/C18-1048>.
- [281] Shi W, Demberg V. Next sentence prediction helps implicit discourse relation classification within and across domains[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 5790-5796.
- [282] Aralikatte R, Lent H, Gonzalez A V, et al. Rewarding coreference resolvers for being consistent with world knowledge[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 1229-1235. <https://aclanthology.org/D19-1118>. DOI: 10.18653/v1/D19-1118.
- [283] Liu L, Song Z, Zheng X. Improving coreference resolution by leveraging entity-centric features with graph neural networks and second-order inference[J]. arXiv preprint arXiv:2009.04639, 2020.
- [284] Wu W, Wang F, Yuan A, et al. CorefQA: Coreference resolution as query-based span prediction [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6953-6963. <https://aclanthology.org/2020.acl-main.622>. DOI: 10.18653/v1/2020.acl-main.622.
- [285] Kantor B, Globerson A. Coreference resolution with entity equalization[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 673-677. <https://aclanthology.org/P19-1066>. DOI: 10.18653/v1/P19-1066.

- [286] Moosavi N S, Strube M. Using linguistic features to improve the generalization capability of neural coreference resolvers[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 193-203. <https://aclanthology.org/D18-1018>. DOI: 10.18653/v1/D18-1018.
- [287] Subramanian S, Roth D. Improving generalization in coreference resolution via adversarial training[C/OL]//Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM 2019). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 192-197. <https://aclanthology.org/S19-1021>. DOI: 10.18653/v1/S19-1021.
- [288] Joshi M, Levy O, Zettlemoyer L, et al. BERT for coreference resolution: Baselines and analysis[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 5803-5808. <https://aclanthology.org/D19-1588>. DOI: 10.18653/v1/D19-1588.
- [289] Joshi M, Chen D, Liu Y, et al. SpanBERT: Improving pre-training by representing and predicting spans[J/OL]. Transactions of the Association for Computational Linguistics, 2020, 8:64-77. <https://aclanthology.org/2020.tacl-1.5>.
- [290] Xu L, Choi J D. Revealing the myth of higher-order inference in coreference resolution[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 8527-8533. <https://aclanthology.org/2020.emnlp-main.686>. DOI: 10.18653/v1/2020.emnlp-main.686.
- [291] Liu F, Zettlemoyer L, Eisenstein J. The referential reader: A recurrent entity network for anaphora resolution[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5918-5925. <https://aclanthology.org/P19-1593>. DOI: 10.18653/v1/P19-1593.
- [292] Xia P, Sedoc J, Van Durme B. Incremental neural coreference resolution in constant memory [C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 8617-8624. <https://aclanthology.org/2020.emnlp-main.695>. DOI: 10.18653/v1/2020.emnlp-main.695.
- [293] Toshniwal S, Wiseman S, Ettlinger A, et al. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Compu-

- tational Linguistics, 2020: 8519-8526. <https://aclanthology.org/2020.emnlp-main.685>. DOI: 10.18653/v1/2020.emnlp-main.685.
- [294] Pham N Q, Kruszewski G, Boleda G. Convolutional neural network language models[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 1153-1162.
- [295] Sukhbaatar S, Weston J, Fergus R, et al. End-to-end memory networks[C]//Advances in neural information processing systems. 2015: 2440-2448.
- [296] Good I J. The population frequencies of species and the estimation of population parameters[J]. Biometrika, 1953, 40(3-4):237-264.
- [297] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE transactions on acoustics, speech, and signal processing, 1987, 35(3):400-401.
- [298] Jelinek F. Interpolated estimation of markov source parameters from sparse data[C]//Proc. Workshop on Pattern Recognition in Practice, 1980. 1980.
- [299] WITTEN I, BELL T. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression[J]. IEEE transactions on information theory, 1991, 37(4):1085-1094.
- [300] Kneser R, Ney H. Improved backing-off for m-gram language modeling[C]//1995 international conference on acoustics, speech, and signal processing: volume 1. IEEE, 1995: 181-184.
- [301] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
- [302] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7871-7880.
- [303] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications[J]. arXiv preprint arXiv:2201.08239, 2022.
- [304] Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization [J]. arXiv preprint arXiv:2110.08207, 2021.
- [305] Fu H, Yao; Peng, Khot T. How does gpt obtain its ability? tracing emergent abilities of language models to their sources[J/OL]. Yao Fu's Notion, 2022. <https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0>

- [306] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.
- [307] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1):5485-5551.
- [308] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
- [309] Gao L, Biderman S, Black S, et al. The pile: An 800gb dataset of diverse text for language modeling [J]. arXiv preprint arXiv:2101.00027, 2020.
- [310] Baumgartner J, Zannettou S, Keegan B, et al. The pushshift reddit dataset[C]//Proceedings of the international AAAI conference on web and social media: volume 14. 2020: 830-839.
- [311] Artetxe M, Bhosale S, Goyal N, et al. Efficient large scale language modeling with mixtures of experts[J]. arXiv preprint arXiv:2112.10684, 2021.
- [312] Shoeybi M, Patwary M, Puri R, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism[J]. arXiv preprint arXiv:1909.08053, 2019.
- [313] Scao T L, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv:2211.05100, 2022.
- [314] Smith S, Patwary M, Norick B, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model[J]. arXiv preprint arXiv:2201.11990, 2022.
- [315] Rasley J, Rajbhandari S, Ruwase O, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3505-3506.
- [316] Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.
- [317] Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners[C]//International Conference on Learning Representations. 2022.
- [318] Wang Y, Mishra S, Alipoormolabashi P, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks[J]. arXiv preprint arXiv:2204.07705, 2022.

- [319] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint arXiv:2203.02155, 2022.
- [320] Ziegler D M, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences[J]. arXiv preprint arXiv:1909.08593, 2019.
- [321] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020, 33:3008-3021.
- [322] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. arXiv preprint arXiv:2107.13586, 2021.
- [323] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 255-269.
- [324] Lin J, Nogueira R, Yates A. Pretrained transformers for text ranking: Bert and beyond[J]. Synthesis Lectures on Human Language Technologies, 2021, 14(4):1-325.
- [325] Bragg J, Cohan A, Lo K, et al. Flex: Unifying evaluation for few-shot nlp[J]. Advances in Neural Information Processing Systems, 2021, 34:15787-15800.
- [326] Cui L, Wu Y, Liu J, et al. Template-based named entity recognition using bart[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1835-1845.
- [327] Ma R, Zhou X, Gui T, et al. Template-free prompt tuning for few-shot NER[C/OL]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, 2022: 5721-5732. <https://aclanthology.org/2022.naacl-main.420>. DOI: 10.18653/v1/2022.naacl-main.420.
- [328] Ram O, Kirstain Y, Berant J, et al. Few-shot question answering by pretraining span selection[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3066-3079.
- [329] Zhong W, Gao Y, Ding N, et al. Proqa: Structural prompt-based pre-training for unified question answering[J]. arXiv preprint arXiv:2205.04040, 2022.

- [330] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.
- [331] Yao X, Zheng Y, Yang X, et al. Nlp from scratch without large-scale pretraining: A simple and efficient framework[C]//International Conference on Machine Learning. PMLR, 2022: 25438-25451.
- [332] Zhang X, Jiang Y, Wang X, et al. Domain-specific ner via retrieving correlated samples[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2398-2404.
- [333] Hu D, Hou X, Du X, et al. Varmae: Pre-training of variational masked autoencoder for domain-adaptive language understanding[J]. arXiv preprint arXiv:2211.00430, 2022.
- [334] Yin Y, Chen C, Shang L, et al. Autotinybert: Automatic hyper-parameter optimization for efficient pre-trained language models[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5146-5157.
- [335] Ding N, Qin Y, Yang G, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models[J]. arXiv preprint arXiv:2203.06904, 2022.
- [336] Zhou X, Ma R, Zou Y, et al. Making parameter-efficient tuning more efficient: A unified framework for classification tasks[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 7053-7064.
- [337] Shi H, Zhang R, Wang J, et al. Layerconnect: Hypernetwork-assisted inter-layer connector to enhance parameter efficiency[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 3120-3126.
- [338] Sarawagi S, Cohen W W. Semi-markov conditional random fields for information extraction[J]. Advances in neural information processing systems, 2004, 17.
- [339] Yan H, Deng B, Li X, et al. Tener: adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv:1911.04474, 2019.
- [340] Zhang Y, Yang J. Chinese ner using lattice lstm[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1554-1564.
- [341] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for

- Computational Linguistics, 2020: 5951-5960. <https://aclanthology.org/2020.acl-main.528>. DOI: 10.18653/v1/2020.acl-main.528.
- [342] Sohrab M G, Miwa M. Deep exhaustive model for nested named entity recognition[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2843-2849.
- [343] Finkel J R, Manning C D. Nested named entity recognition[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. 2009: 141-150.
- [344] Xu M, Jiang H, Watcharawittayakul S. A local detection approach for named entity recognition and mention detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1237-1247.
- [345] Tan C, Qiu W, Chen M, et al. Boundary enhanced neural span classification for nested named entity recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 9016-9023.
- [346] Yan H, Gui T, Dai J, et al. A unified generative framework for various ner subtasks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5808-5822.
- [347] Sang E T K, De Meulder F. Introduction to the conll-2003 shared task: Language-independent named entity recognition[C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. 2003: 142-147.
- [348] Weischedel R, Palmer M, Marcus M, et al. Ontonotes release 5.0 ldc 2013t19[J]. Linguistic Data Consortium, Philadelphia, PA, 2013, 23.
- [349] Levow G A. The third international chinese language processing bakeoff: Word segmentation and named entity recognition[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.
- [350] Peng N, Dredze M. Named entity recognition for chinese social media with jointly trained embeddings[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 548-554.
- [351] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation.[C]//Lrec: volume 2. Lisbon, 2004: 837-840.

- [352] Mani I, Hitzeman J, Richer J, et al. Ace 2005 english spatialml annotations[J]. Linguistic Data Consortium, Philadelphia, 2008.
- [353] Kim J D, Ohta T, Tateisi Y, et al. Genia corpus—a semantically annotated corpus for bio-textmining [J]. Bioinformatics, 2003, 19(suppl\_1):i180-i182.
- [354] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008: 1247-1250.
- [355] Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 2205-2215. <https://aclanthology.org/D18-1244>. DOI: 10.18653/v1/D18-1244.
- [356] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//EMNLP. 2015.
- [357] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C/OL]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 2124-2133. <https://aclanthology.org/P16-1200>. DOI: 10.18653/v1/P16-1200.
- [358] Etzioni O, Banko M, Soderland S, et al. Open information extraction from the web[J/OL]. Commun. ACM, 2008, 51(12):68–74. <https://doi.org/10.1145/1409360.1409378>.
- [359] Downey D, Etzioni O, Soderland S. A probabilistic model of redundancy in information extraction [C]//Proceedings of the 19th international joint conference on Artificial intelligence. 2005: 1034-1041.
- [360] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction[C/OL]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011: 1535-1545. <https://aclanthology.org/D11-1142>.
- [361] Hu X, Wen L, Xu Y, et al. Selfore: Self-supervised relational feature learning for open relation extraction[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3673-3682.

- [362] Zhao J, Gui T, Zhang Q, et al. A relation-oriented clustering method for open relation extraction [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9707-9718.
- [363] Hendrickx I, Kim S N, Kozareva Z, et al. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C/OL]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics, 2010: 33-38. <https://aclanthology.org/S10-1006>.
- [364] Zhang Y, Zhong V, Chen D, et al. Position-aware attention and supervised data improve slot filling [C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). 2017: 35-45. <https://nlp.stanford.edu/pubs/zhang2017tacred.pdf>.
- [365] Han X, Zhu H, Yu P, et al. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4803-4809.
- [366] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2010: 148-163.
- [367] Ahn D. The stages of event extraction[C]//Proceedings of the Workshop on Annotating and Reasoning about Time and Events. 2006: 1-8.
- [368] Daelemans W, Zavrel J, Van Der Sloot K, et al. Timbl: Tilburg memory-based learner[J]. Tilburg University, 2004.
- [369] Daumé III H. Notes on cg and lm-bfgs optimization of logistic regression[J]. Paper available at <http://pub.hal3.name#daume04cg-bfgs>, implementation available at <http://hal3.name/megam>, 2004, 198:282.
- [370] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 167-176.
- [371] Nguyen T H, Cho K, Grishman R. Joint event extraction via recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 300-309.

- [372] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998: 28-36.
- [373] Liu X, Huang H Y, Zhang Y. Open domain event extraction using neural latent variable models[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2860-2871.
- [374] Klein D, Manning C D. Accurate unlexicalized parsing[C]//Proceedings of the 41st annual meeting of the association for computational linguistics. 2003: 423-430.
- [375] Fritzler A, Logacheva V, Kretov M. Few-shot classification in named entity recognition task[C]// Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 2019: 993-1000.
- [376] Gao F, Cai L, Yang Z, et al. Multi-distance metric network for few-shot learning[J]. International Journal of Machine Learning and Cybernetics, 2022:1-12.
- [377] Huang Y, He K, Wang Y, et al. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2515-2527.
- [378] Li J, Chiu B, Feng S, et al. Few-shot named entity recognition via meta-learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2020.
- [379] de Lichy C, Glaude H, Campbell W. Meta-learning for few-shot named entity recognition[C]// Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing. 2021: 44-58.
- [380] Huang J, Li C, Subudhi K, et al. Few-shot named entity recognition: An empirical baseline study [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 10408-10423.
- [381] Rahimi A, Li Y, Cohn T. Massively multilingual transfer for ner[C]//ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. Association for Computational Linguistics-ACL, 2019: 151-164.
- [382] Jia C, Zhang Y. Multi-cell compositional lstm for ner domain adaptation[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 5906-5917.

- [383] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in chinese ner[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5951-5960.
- [384] Peng M, Ma R, Zhang Q, et al. Toward recognizing more entity types in ner: an efficient implementation using only entity lexicons[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 678-688.
- [385] Wu S, Song X, Feng Z. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1529-1539.
- [386] Chen X, Li L, Deng S, et al. Lightner: A lightweight tuning paradigm for low-resource ner via pluggable prompting[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2374-2387.
- [387] Lai P, Ye F, Zhang L, et al. Pcbert: Parent and child bert for chinese few-shot ner[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2199-2209.
- [388] Chen X, Zhang N, Xie X, et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]//Proceedings of the ACM Web Conference 2022. 2022: 2778-2788.
- [389] Zhang H, Liang B, Yang M, et al. Prompt-based prototypical framework for continual relation extraction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2801-2813.
- [390] Yang S, Song D. Fpc: Fine-tuning with prompt curriculum for relation extraction[C]//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. 2022: 1065-1077.
- [391] Zhang Q, Fu J, Liu X, et al. Adaptive co-attention network for named entity recognition in tweets [C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [392] Yu J, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28:429-439.
- [393] Zheng C, Feng J, Fu Z, et al. Multimodal relation extraction with efficient graph alignment[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 5298-5306.

- [394] Zhang D, Wei S, Li S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 14347-14355.
- [395] Liu X, Gao F, Zhang Q, et al. Graph convolution for multimodal information extraction from visually rich documents[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). 2019: 32-39.
- [396] Yu W, Lu N, Qi X, et al. Pick: processing key information extraction from documents using improved graph learning-convolutional networks[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 4363-4370.
- [397] Lockard C, Shiralkar P, Dong X L, et al. Zeroshotceres: Zero-shot relation extraction from semi-structured webpages[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8105-8117.
- [398] Li Y, Qian Y, Yu Y, et al. Structext: Structured text understanding with multi-modal transformers [C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 1912-1920.
- [399] Lu Y, Liu Q, Dai D, et al. Unified structure generation for universal information extraction[C/OL]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 5755-5772. <https://aclanthology.org/2022.acl-long.395>.
- [400] Kan Z, Feng L, Yin Z, et al. A unified generative framework based on prompt learning for various information extraction tasks[J]. arXiv preprint arXiv:2209.11570, 2022.
- [401] Forcada M L, Ginestí-Rosell M, Nordfalk J, et al. Apertium: A free/open-source platform for rule-based machine translation[J]. Machine Translation, 2011, 25(2):127-144.
- [402] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C/OL]//NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. USA: Association for Computational Linguistics, 2003: 48-54. <https://doi.org/10.3115/1073445.1073462>.
- [403] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C/OL]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Prague, Czech Republic: Association for Computational Linguistics, 2007: 177-180. <https://aclanthology.org/P07-2045>.

- [404] Chrisman L. Learning recursive distributed representations for holistic computation[J]. Connection Science, 1991, 3(4):345-366.
- [405] Allen R B. Several studies on natural language and back-propagation[C]/Proceedings of the IEEE First International Conference on Neural Networks: volume 2. IEEE Piscataway, NJ, 1987: 341.
- [406] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]/Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1700-1709.
- [407] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[Z]. 2016.
- [408] Akhbardeh F, Arkhangorodsky A, Biesialska M, et al. Findings of the 2021 conference on machine translation (wmt21)[C]/Proceedings of the Sixth Conference on Machine Translation. 2021: 1-88.
- [409] Khashabi D, Stanovsky G, Bragg J, et al. Genie: A leaderboard for human-in-the-loop evaluation of text generation[J]. arXiv preprint arXiv:2101.06561, 2021.
- [410] 肖桐, 朱靖波. 机器翻译: 基础与模型[M]. 北京: 电子工业出版社, 2021.
- [411] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing [J]. ieee Computational intelligenCe magazine, 2018, 13(3):55-75.
- [412] Song F, Croft W B. A general language model for information retrieval[C]/Proceedings of the eighth international conference on Information and knowledge management. 1999: 316-321.
- [413] Wallach H M. Topic modeling: Beyond bag-of-words[C/OL]/ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2006: 977–984. <https://doi.org/10.1145/1143844.1143967>.
- [414] Britz D, Goldie A, Luong M T, et al. Massive exploration of neural machine translation architectures [J]. arXiv preprint arXiv:1703.03906, 2017.
- [415] Galley M, Manning C D. A simple and effective hierarchical phrase reordering model[C/OL]/Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics, 2008: 848-856. <https://aclanthology.org/D08-1089>.
- [416] Chiang D, Knight K, Wang W. 11,001 new features for statistical machine translation[C]/Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics. 2009: 218-226.

- [417] Green S, Wang S I, Cer D, et al. Fast and adaptive online training of feature-rich translation models[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 311-321.
- [418] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C/OL]//Guyon I, Luxburg U V, Bengio S, et al. Advances in Neural Information Processing Systems: volume 30. Curran Associates, Inc., 2017. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf>.
- [419] Voita E, Talbot D, Moiseev F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5797-5808. <https://aclanthology.org/P19-1580>. DOI: 10.18653/v1/P19-1580.
- [420] Li J, Tu Z, Yang B, et al. Multi-head attention with disagreement regularization[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 2897-2903. <https://aclanthology.org/D18-1317>. DOI: 10.18653/v1/D18-1317.
- [421] Hao J, Wang X, Shi S, et al. Multi-granularity self-attention for neural machine translation[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 887-897. <https://aclanthology.org/D19-1082>. DOI: 10.18653/v1/D19-1082.
- [422] Setiawan H, Sperber M, Nallasamy U, et al. Variational neural machine translation with normalizing flows[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 7771-7777. <https://aclanthology.org/2020.acl-main.694>. DOI: 10.18653/v1/2020.acl-main.694.
- [423] Beltagy I, Peters M E, Cohan A. Longformer: The long-document transformer[J/OL]. CoRR, 2020, abs/2004.05150. <https://arxiv.org/abs/2004.05150>.
- [424] Choromanski K M, Likhoshesterov V, Dohan D, et al. Rethinking attention with performers[C/OL]//International Conference on Learning Representations. 2021. <https://openreview.net/forum?id=Ua6zuk0WRH>.

- [425] Guo Q, Qiu X, Liu P, et al. Star-transformer[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1315-1325. <https://aclanthology.org/N19-1133>. DOI: 10.18653/v1/N19-1133.
- [426] Wu L, Wang Y, Xia Y, et al. Exploiting monolingual data at scale for neural machine translation[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 4207-4216. <https://aclanthology.org/D19-1430>. DOI: 10.18653/v1/D19-1430.
- [427] Zhang J, Zong C. Exploiting source-side monolingual data in neural machine translation[C/OL]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 1535-1545. <https://aclanthology.org/D16-1160>. DOI: 10.18653/v1/D16-1160.
- [428] Wang X, Pham H, Dai Z, et al. SwitchOut: an efficient data augmentation algorithm for neural machine translation[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 856-861. <https://aclanthology.org/D18-1100>. DOI: 10.18653/v1/D18-1100.
- [429] Dou Z Y, Anastasopoulos A, Neubig G. Dynamic data selection and weighting for iterative back-translation[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 5894-5904. <https://www.aclweb.org/anthology/2020.emnlp-main.475>. DOI: 10.18653/v1/2020.emnlp-main.475.
- [430] Wang S, Liu Y, Wang C, et al. Improving back-translation with uncertainty-based confidence estimation[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 791-802. <https://aclanthology.org/D19-1073>. DOI: 10.18653/v1/D19-1073.
- [431] Peters M E, Ruder S, Smith N A. To tune or not to tune? adapting pretrained representations to diverse tasks[C/OL]//Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence, Italy: Association for Computational Linguistics, 2019: 7-14. <https://aclanthology.org/W19-4302>. DOI: 10.18653/v1/W19-4302.

- [432] Sun C, Qiu X, Xu Y, et al. How to fine-tune BERT for text classification?[J/OL]. CoRR, 2019, abs/1905.05583. <http://arxiv.org/abs/1905.05583>.
- [433] Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation[C/OL]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015: 1723-1732. <https://aclanthology.org/P15-1166>. DOI: 10.3115/v1/P15-1166.
- [434] Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism[C/OL]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 866-875. <https://aclanthology.org/N16-1101>. DOI: 10.18653/v1/N16-1101.
- [435] Al-Shedivat M, Parikh A. Consistency by agreement in zero-shot neural machine translation [C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1184-1197. <https://aclanthology.org/N19-1121>. DOI: 10.18653/v1/N19-1121.
- [436] Gu J, Neubig G, Cho K, et al. Learning to translate in real-time with neural machine translation [C/OL]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics, 2017: 1053-1062. <https://aclanthology.org/E17-1099>.
- [437] Grissom II A, He H, Boyd-Graber J, et al. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1342-1352. <https://aclanthology.org/D14-1140>. DOI: 10.3115/v1/D14-1140.
- [438] Ma M, Huang L, Xiong H, et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3025-3036. <https://aclanthology.org/P19-1289>. DOI: 10.18653/v1/P19-1289.

- [439] Zheng R, Ma M, Zheng B, et al. Speculative beam search for simultaneous translation[C/OL]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 1395-1402. <https://aclanthology.org/D19-1144>. DOI: 10.18653/v1/D19-1144.
- [440] Jean S, Cho K. Context-aware learning for neural machine translation[J/OL]. CoRR, 2019, abs/1903.04715. <http://arxiv.org/abs/1903.04715>.
- [441] Sugiyama A, Yoshinaga N. Data augmentation using back-translation for context-aware neural machine translation[C/OL]//Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019). Hong Kong, China: Association for Computational Linguistics, 2019: 35-44. <https://aclanthology.org/D19-6504>. DOI: 10.18653/v1/D19-6504.
- [442] WIEBE J M. Learning subjective adjectives from corpora[C]//Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000). 2000.
- [443] Das S, Chan M. Extracting market sentiment from stock message boards[J]. Asia Pacific Finance Association, 2001, 2001.
- [444] Tong R M. An operational system for detecting and tracking opinions in on-line discussion[C]// Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification: volume 1. 2001.
- [445] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.
- [446] Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing [C]//Proceedings of the 2nd international conference on Knowledge capture. 2003: 70-77.
- [447] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th international conference on World Wide Web. 2003: 519-528.
- [448] Wiebe J. Identifying subjective characters in narrative[C]//COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics. 1990.
- [449] Wiebe J M. Tracking point of view in narrative[J]. Computational Linguistics, 1994, 20(2):233-287.

- [450] Wiebe J, Bruce R, O' Hara T P. Development and use of a gold-standard data set for subjectivity classifications[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics. 1999: 246-253.
- [451] Liu B. Sentiment analysis: Mining opinions, sentiments, and emotions[M]. Cambridge university press, 2020.
- [452] Liu B. Web data mining: exploring hyperlinks, contents, and usage data: volume 1[M]. Springer, 2011.
- [453] Jindal N, Liu B. Mining comparative sentences and relations[C]//Aaai: volume 22. 2006: 9.
- [454] Parrott W G. Emotions in social psychology: Essential readings[M]. psychology press, 2001.
- [455] Plutchik R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American scientist, 2001, 89(4): 344-350.
- [456] Mehrabian A, Russell J A. An approach to environmental psychology.[M]. the MIT Press, 1974.
- [457] Russell J A. Evidence of convergent validity on the dimensions of affect.[J]. Journal of personality and social psychology, 1978, 36(10).
- [458] Pang B, Lee L, et al. Opinion mining and sentiment analysis[J]. Foundations and Trends® in Information Retrieval, 2008, 2(1–2):1-135.
- [459] Liu B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1):1-167.
- [460] Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web[C]// Proceedings of the 14th international conference on World Wide Web. 2005: 342-351.
- [461] Behdenna S, Barigou F, Belalem G. Document level sentiment analysis: a survey[J]. EAI Endorsed Transactions on Context-aware Systems and Applications, 2018, 4(13).
- [462] Behdenna S, Barigou F, Belalem G. Sentiment analysis at document level[C]//International Conference on Smart Trends for Information Technology and Computer Communications. Springer, 2016: 159-168.
- [463] Tsytarau M, Palpanas T. Survey on mining subjective data on the web[J]. Data Mining and Knowledge Discovery, 2012, 24(3):478-514.

- [464] Wu Y, Zhang Q, Huang X J, et al. Phrase dependency parsing for opinion mining[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. 2009: 1533-1541.
- [465] Moraes R, Valiati J F, Neto W P G. Document-level sentiment classification: An empirical comparison between svm and ann[J]. Expert Systems with Applications, 2013, 40(2):621-633.
- [466] Tripathy A, Anand A, Rath S K. Document-level sentiment classification using hybrid machine learning approach[J]. Knowledge and Information Systems, 2017, 53(3):805-831.
- [467] Drucker H, Burges C J, Kaufman L, et al. Support vector regression machines[J]. Advances in neural information processing systems, 1996, 9.
- [468] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.
- [469] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011: 142-150.
- [470] Diao Q, Qiu M, Wu C Y, et al. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars)[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 193-202.
- [471] Huang X, Paul M J. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 4113-4123. <https://aclanthology.org/P19-1403>. DOI: 10.18653/v1/P19-1403.
- [472] Tan S, Zhang J. An empirical study of sentiment analysis for chinese documents[J]. Expert Systems with applications, 2008, 34(4):2622-2629.
- [473] Bu J, Ren L, Zheng S, et al. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction[C/OL]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 2069-2079. <https://www.aclweb.org/anthology/2021.naacl-main.167>.
- [474] Bongirwar V K. A survey on sentence level sentiment analysis[J]. International Journal of Computer Science Trends and Technology (IJCST), 2015, 3(3):110-113.

- [475] Jagtap V, Pawar K. Analysis of different approaches to sentence-level sentiment classification[J]. International Journal of Scientific Engineering and Technology, 2013, 2(3):164-170.
- [476] Li Z, Zou Y, Zhang C, et al. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 246-256.
- [477] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics, 2011, 37(2):267-307.
- [478] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 168-177.
- [479] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.
- [480] Tian H, Gao C, Xiao X, et al. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4067-4076.
- [481] Turney P. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.
- [482] Demszky D, Movshovitz-Attias D, Ko J, et al. Goemotions: A dataset of fine-grained emotions[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4040-4054.
- [483] Li C, Xu B, Wu G, et al. Recursive deep learning for sentiment analysis over social data[C]//2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT): volume 2. IEEE, 2014: 180-185.
- [484] Fu X, Xu Y. Recursive autoencoder with hownet lexicon for sentence-level sentiment analysis[M]// Proceedings of the ASE BigData & SocialInformatics 2015. 2015: 1-7.
- [485] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation [J]. Computational linguistics, 2011, 37(1):9-27.

- [486] Li X, Bing L, Li P, et al. Aspect term extraction with history attention and selective transformation [C/OL]//Lang J. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, 2018: 4194-4200. <https://doi.org/10.24963/ijcai.2018/583>.
- [487] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs [C]//Proceedings of the 16th international conference on World Wide Web. 2007: 171-180.
- [488] Fan F, Feng Y, Zhao D. Multi-grained attention network for aspect-level sentiment classification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3433-3442.
- [489] Yan H, Dai J, Ji T, et al. A unified generative framework for aspect-based sentiment analysis[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 2416-2429. <https://aclanthology.org/2021.acl-long.188>. DOI: 10.18653/v1/2021.acl-long.188.
- [490] Manandhar S. Semeval-2014 task 4: Aspect based sentiment analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014.
- [491] Pontiki M, Galanis D, Papageorgiou H, et al. Semeval-2015 task 12: Aspect based sentiment analysis[C]//Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015: 486-495.
- [492] Pontiki M, Galanis D, Papageorgiou H, et al. Semeval-2016 task 5: Aspect based sentiment analysis [C]//Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). 2016: 19-30.
- [493] Dong L, Wei F, Tan C, et al. Adaptive recursive neural network for target-dependent twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers): volume 2. 2014: 49-54.
- [494] Saeidi M, Bouchard G, Liakata M, et al. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 1546-1556.
- [495] Ganu G, Elhadad N, Marian A. Beyond the stars: Improving rating predictions using review text content.[C]//WebDB: volume 9. Citeseer, 2009: 1-6.

- [496] Deng L, Wiebe J. Mpqa 3.0: An entity/event-level sentiment corpus[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1323-1328.
- [497] Bu J, Ren L, Zheng S, et al. Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2069-2079.
- [498] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 129-136.
- [499] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]//Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011: 151-161.
- [500] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]//EMNLP. Association for Computational Linguistics, 2012: 1201-1211.
- [501] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [502] Tang D, Qin B, Liu T. Learning semantic representations of users and products for document level sentiment classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1014-1023.
- [503] Chen H, Sun M, Tu C, et al. Neural sentiment classification with user and product attention[C]//EMNLP. 2016.
- [504] Dodge J, Ilharco G, Schwartz R, et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping[J]. arXiv preprint arXiv:2002.06305, 2020.
- [505] Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification?[C]//China National Conference on Chinese Computational Linguistics. Springer, 2019: 194-206.
- [506] Xu H, Liu B, Shu L, et al. Bert post-training for review reading comprehension and aspect-based sentiment analysis[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 2324-2335.

- [507] Song Y, Wang J, Jiang T, et al. Attentional encoder network for targeted sentiment classification[J]. arXiv preprint arXiv:1902.09314, 2019.
- [508] Sun C, Huang L, Qiu X. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019.
- [509] He R, Lee W S, Ng H T, et al. Exploiting document knowledge for aspect-level sentiment classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2018: 579-585.
- [510] Ziser Y, Reichart R. Pivot based language modeling for improved neural domain adaptation[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1241-1251.
- [511] Li Z, Wei Y, Zhang Y, et al. Hierarchical attention transfer network for cross-domain sentiment classification[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [512] Li Z, Zhang Y, Wei Y, et al. End-to-end adversarial memory network for cross-domain sentiment classification.[C]//IJCAI. 2017: 2237-2243.
- [513] Du C, Sun H, Wang J, et al. Adversarial and domain-aware bert for cross-domain sentiment analysis [C]//ACL. 2020.
- [514] Qu X, Zou Z, Cheng Y, et al. Adversarial category alignment network for cross-domain sentiment classification[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 2496-2508.
- [515] Zhang Z, Wu Y, Zhao H, et al. Semantics-aware bert for language understanding[C]//AAAI. 2020.
- [516] Zhou J, Tian J, Wang R, et al. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis[C]//Proceedings of the 28th international conference on computational linguistics. 2020: 568-579.
- [517] Levine Y, Lenz B, Dagan O, et al. Sensebert: Driving some sense into bert[J]. arXiv preprint arXiv:1908.05646, 2019.
- [518] Ke P, Ji H, Liu S, et al. SentiLr: Linguistic knowledge enhanced language representation for sentiment analysis[J]. arXiv preprint arXiv:1911.02493, 2019.

- [519] Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm[C]//Proceedings of AAAI. 2018: 5876-5883.
- [520] Zhou J, Huang J X, Hu Q V, et al. Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification[J]. Knowledge-Based Systems, 2020, 205:106292.
- [521] Nguyen T H, Shirai K. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 2509-2514.
- [522] He R, Lee W S, Ng H T, et al. Effective attention modeling for aspect-level sentiment classification [C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1121-1131.
- [523] Gu S, Zhang L, Hou Y, et al. A position-aware bidirectional attention network for aspect-level sentiment analysis[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 774-784.
- [524] Li X, Bing L, Lam W, et al. Transformation networks for target-oriented sentiment classification [C]//ACL. 2018: 946-956.
- [525] Kiritchenko S, Zhu X, Cherry C, et al. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014: 437-442.
- [526] Tang D, Qin B, Feng X, et al. Effective lstms for target-dependent sentiment classification[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 3298-3307.
- [527] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 214-224.
- [528] Chen P, Sun Z, Bing L, et al. Recurrent attention network on memory for aspect sentiment analysis [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 452-461.
- [529] Fan C, Gao Q, Du J, et al. Convolution-based memory network for aspect-based sentiment analysis [C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2018: 1161-1164.

- [530] Zhang M, Zhang Y, Vo D T. Gated neural networks for targeted sentiment analysis.[C]//AAAI. 2016: 3087-3093.
- [531] Wang J, Li J, Li S, et al. Aspect sentiment classification with both word-level and clause-level attention networks.[C]//IJCAI. 2018: 4439-4445.
- [532] Ma D, Li S, Zhang X, et al. Interactive attention networks for aspect-level sentiment classification [C]//IJCAI. 2017: 4068-4074.
- [533] Wang Y, Huang M, Zhu X, et al. Attention-based lstm for aspect-level sentiment classification[C]// Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606-615.
- [534] Zhou J, Chen Q, Huang J X, et al. Position-aware hierarchical transfer model for aspect-level sentiment classification[J]. Information Sciences, 2020, 513:1-16.
- [535] Green Jr B F, Wolf A K, Chomsky C, et al. Baseball: an automatic question-answerer[C]//Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. 1961: 219-224.
- [536] Voorhees E M, et al. The trec-8 question answering track report[C]//Trec: volume 99. 1999: 77-82.
- [537] Woods W A. Progress in natural language understanding: an application to lunar geology[C]// Proceedings of the June 4-8, 1973, national computer conference and exposition. 1973: 441-450.
- [538] Winograd T. Procedures as a representation for data in a computer program for understanding natural language[R]. MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971.
- [539] Warren D H, Pereira F C. An efficient easily adaptable system for interpreting natural language queries[J]. American journal of computational linguistics, 1982, 8(3-4):110-122.
- [540] Wilensky R. The berkeley unix consultant project[M]//Wissensbasierte Systeme. Springer, 1987: 286-296.
- [541] Lehnert W G. A conceptual theory of question answering[C]//Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1. 1977: 158-164.
- [542] Katz B. Using english for indexing and retrieving[R]. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1988.

- [543] Katz B. From sentence processing to information access on the world wide web[C]//AAAI Spring Symposium on Natural Language Processing for the World Wide Web: volume 1. Stanford University Stanford, CA, USA, 1997: 997.
- [544] Katz B, Felshin S, Lin J, et al. Viewing the web as a virtual database for question answering.[C]// New Directions in Question Answering. 2004: 215-226.
- [545] Ittycheriah A, Franz M, Roukos S. Ibm's statistical question answering system - trec-10[C]//TREC. 2001.
- [546] Ferrucci D, Brown E, Chu-Carroll J, et al. Building watson: An overview of the deepqa project[J]. AI magazine, 2010, 31(3):59-79.
- [547] 许静芳. 搜狗汪仔的“大梦想” [J]. 中国计算机学会通讯, 2007, 13(5).
- [548] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.
- [549] Yang Z, Qi P, Zhang S, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2369-2380.
- [550] Reddy S, Chen D, Manning C D. Coqa: A conversational question answering challenge[J]. Transactions of the Association for Computational Linguistics, 2019, 7:249-266.
- [551] Saha A, Pahuja V, Khapra M, et al. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.
- [552] Nan L, Hsieh C, Mao Z, et al. Fetaqa: Free-form table question answering[J]. Transactions of the Association for Computational Linguistics, 2022, 10:35-49.
- [553] 段楠, 周明. 智能问答[M]. 北京: 高等教育出版社, 2018.
- [554] Hirschman L, Light M, Breck E, et al. Deep read: A reading comprehension system[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics. 1999: 325-332.
- [555] Xu K, Meng H. Using verb dependency matching in a reading comprehension system[C]//Asia Information Retrieval Symposium. Springer, 2004: 190-201.

- [556] Pasupat P, Liang P. Compositional semantic parsing on semi-structured tables[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1470-1480.
- [557] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 189-198.
- [558] Riloff E, Thelen M. A rule-based question answering system for reading comprehension tests[C]//ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems. 2000.
- [559] Riloff E, Phillips W. An introduction to the sundance and autoslog systems[R]. Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.
- [560] Xu K, Meng H, Weng F. A maximum entropy framework that integrates word dependencies and grammatical relations for reading comprehension[C]//Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. 2006: 185-188.
- [561] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension[J]. arXiv preprint arXiv:1611.01603, 2016.
- [562] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387, 2015.
- [563] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8:64-77.
- [564] Nguyen T, Rosenberg M, Song X, et al. Ms marco: A human generated machine reading comprehension dataset[C]//CoCo@ NIPs. 2016.
- [565] Pan F, Canim M, Glass M, et al. Cltr: An end-to-end, transformer-based system for cell level table retrieval and table question answering[J]. arXiv preprint arXiv:2106.04441, 2021.
- [566] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.
- [567] Chen Q, Zhu X, Ling Z H, et al. Enhanced lstm for natural language inference[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1657-1668.

- [568] Zhu F, Lei W, Wang C, et al. Retrieving and reading: A comprehensive survey on open-domain question answering[J]. arXiv preprint arXiv:2101.00774, 2021.
- [569] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1870-1879.
- [570] Karpukhin V, Ouguz B, Min S, et al. Dense passage retrieval for open-domain question answering [J]. arXiv preprint arXiv:2004.04906, 2020.
- [571] Lazaridou A, Gribovskaya E, Stokowiec W, et al. Internet-augmented language models through few-shot prompting for open-domain question answering[J]. arXiv preprint arXiv:2203.05115, 2022.
- [572] Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model?[J]. arXiv preprint arXiv:2002.08910, 2020.
- [573] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7:453-466.
- [574] Liu J, Lin Y, Liu Z, et al. Xqa: A cross-lingual open-domain question answering dataset[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2358-2368.
- [575] Dunn M, Sagun L, Higgins M, et al. Searchqa: A new q&a dataset augmented with context from a search engine[J]. arXiv preprint arXiv:1704.05179, 2017.
- [576] Fang Y, Sun S, Gan Z, et al. Hierarchical graph network for multi-hop question answering[J]. arXiv preprint arXiv:1911.03631, 2019.
- [577] Ding M, Zhou C, Chen Q, et al. Cognitive graph for multi-hop reading comprehension at scale[J]. arXiv preprint arXiv:1905.05460, 2019.
- [578] Jiang Y, Bansal M. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa[J]. arXiv preprint arXiv:1906.07132, 2019.
- [579] Nishida K, Nishida K, Nagata M, et al. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction[J]. arXiv preprint arXiv:1905.08511, 2019.
- [580] Dalvi B, Jansen P, Tafjord O, et al. Explaining answers with entailment trees[J]. arXiv preprint arXiv:2104.08661, 2021.

- [581] Ribeiro D, Wang S, Ma X, et al. Entailment tree explanations via iterative retrieval-generation reasoner[J]. arXiv preprint arXiv:2205.09224, 2022.
- [582] Hong R, Zhang H, Yu X, et al. Metgen: A module-based entailment tree generation framework for answer explanation[J]. arXiv preprint arXiv:2205.02593, 2022.
- [583] Kembhavi A, Seo M, Schwenk D, et al. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. 2017: 4999-5007.
- [584] Huang Z, Liu F, Wu X, et al. Audio-oriented multimodal machine comprehension via dynamic inter- and intra-modality attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 13098-13106.
- [585] Yagcioglu S, Erdem A, Erdem E, et al. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes[J]. arXiv preprint arXiv:1809.00812, 2018.
- [586] Mai S, Hu H, Xu J, et al. Multi-fusion residual memory network for multimodal human sentiment comprehension[J]. IEEE Transactions on Affective Computing, 2020, 13(1):320-334.
- [587] Luhn H P. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal of research and development, 1957, 1(4):309-317.
- [588] Luhn H P. The automatic creation of literature abstracts[J]. IBM Journal of research and development, 1958, 2(2):159-165.
- [589] Neto J L, Freitas A A, Kaestner C A. Automatic text summarization using a machine learning approach[C]//Brazilian symposium on artificial intelligence. Springer, 2002: 205-215.
- [590] Alami N, Meknassi M, En-nahnabi N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning[J]. Expert systems with applications, 2019, 123:195-211.
- [591] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.
- [592] Nguyen K, Daumé III H. Global voices: Crossing borders in automatic news summarization[C]//Proceedings of the 2nd Workshop on New Frontiers in Summarization. 2019: 90-97.

- [593] Zhu J, Li H, Liu T, et al. Msmo: Multimodal summarization with multimodal output[C]// Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 4154-4164.
- [594] Mihalcea R, Tarau P. TextRank: Bringing order into text[C/OL]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004: 404-411. <https://aclanthology.org/W04-3252>.
- [595] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[J]. Computer networks and ISDN systems, 1998, 30(1-7):107-117.
- [596] Carbonell J, Goldstein J. The use of mmr, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998: 335-336.
- [597] Narayan S, Cohen S B, Lapata M. Ranking sentences for extractive summarization with reinforcement learning[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1747-1759.
- [598] Zhou Q, Yang N, Wei F, et al. Neural document summarization by jointly learning to score and select sentences[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 654-663.
- [599] Nallapati R, Zhai F, Zhou B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]//Thirty-first AAAI conference on artificial intelligence. 2017.
- [600] Liu Y, Lapata M. Text summarization with pretrained encoders[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3730-3740.
- [601] Jing H. Using hidden markov modeling to decompose human-written summaries[J]. Computational Linguistics, 2002, 28(4):527-543.
- [602] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//ICLR (Poster). 2015.
- [603] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1073-1083.

- [604] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.
- [605] Ranzato M, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks[J]. arXiv preprint arXiv:1511.06732, 2015.
- [606] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
- [607] Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[C]// International Conference on Learning Representations. 2018.
- [608] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[J]. computer vision and pattern recognition, 2017.
- [609] Liu C W, Lowe R, Serban I V, et al. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation[J]. empirical methods in natural language processing, 2016.
- [610] Yu L, Zhang W, Wang J, et al. Seqgan: Sequence generative adversarial nets with policy gradient [C]//Proceedings of the AAAI conference on artificial intelligence: volume 31. 2017.
- [611] Liu L, Lu Y, Yang M, et al. Generative adversarial network for abstractive text summarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.
- [612] Gehrmann S, Deng Y, Rush A M. Bottom-up abstractive summarization[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4098-4109.
- [613] Chen Y C, Bansal M. Fast abstractive summarization with reinforce-selected sentence rewriting[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 675-686.
- [614] Peyrard M. A simple theoretical model of importance for summarization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1059-1073.
- [615] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
- [616] Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert[C]//International Conference on Learning Representations. 2019.

- [617] Hermann K M, Kočiský T, Grefenstette E, et al. Teaching machines to read and comprehend[C]// Advances in Neural Information Processing Systems. 2015: 1693-1701.
- [618] Hu B, Chen Q, Zhu F. Lcsts: A large scale chinese short text summarization dataset[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1967-1972.
- [619] Cohan A, Dernoncourt F, Kim D S, et al. A discourse-aware attention model for abstractive summarization of long documents[C]//NAACL-HLT (2). 2018.
- [620] Fabbri A R, Li I, She T, et al. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1074-1084.
- [621] Liu P J, Saleh M, Pot E, et al. Generating wikipedia by summarizing long sequences[C]// International Conference on Learning Representations. 2018.
- [622] Gliwa B, Mochol I, Biesek M, et al. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization[C]//Proceedings of the 2nd Workshop on New Frontiers in Summarization. 2019: 70-79.
- [623] Chen Y, Liu Y, Chen L, et al. Dialogsum: A real-life scenario dialogue summarization dataset[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 5062-5074.
- [624] Carletta J, Ashby S, Bourban S, et al. The ami meeting corpus: A pre-announcement[C]// International workshop on machine learning for multimodal interaction. Springer, 2005: 28-39.
- [625] Janin A, Baron D, Edwards J, et al. The icsi meeting corpus[C]//2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).: volume 1. IEEE, 2003: I-I.
- [626] Sanabria R, Caglayan O, Palaskar S, et al. How2: A large-scale dataset for multimodal language understanding[C]//NeurIPS. 2018.
- [627] Zhu J, Wang Q, Wang Y, et al. Ncls: Neural cross-lingual summarization[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3054-3064.
- [628] Ladhak F, Durmus E, Cardie C, et al. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 4034-4048.

- [629] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [630] Pasunuru R, Bansal M. Multi-reward reinforced summarization with saliency and entailment[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 646-653.
- [631] Zhong M, Liu P, Chen Y, et al. Extractive summarization as text matching[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6197-6208.
- [632] Liu Y, Liu P. Simcls: A simple framework for contrastive learning of abstractive summarization [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 1065-1072.
- [633] Narayan S, Cohen S B, Lapata M. Don' t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1797-1807.
- [634] Kryściński W, Keskar N S, McCann B, et al. Neural text summarization: A critical evaluation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 540-551.
- [635] Feigenbaum E A. The art of artificial intelligence: Themes and case studies of knowledge engineering[C]//Proceedings of the Fifth International Joint Conference on Artificial Intelligence: volume 2. Boston, 1977.
- [636] Powers D M. Robot intelligence[J]. Electronics Today International (Australia), 1983:15-18.
- [637] Sowa J F. Semantic networks[J]. 1987.
- [638] Gruber T R. Toward principles for the design of ontologies used for knowledge sharing?[J]. International journal of human-computer studies, 1995, 43(5-6):907-928.
- [639] Berners-Lee T, Hendler J, Lassila O. The semantic web[J]. Scientific american, 2001, 284(5):34-43.
- [640] Sullivan D. A reintroduction to our knowledge graph and knowledge panels[J]. The Keyword, Google, May, 2020, 20.

- [641] Yamada I, Asai A, Shindo H, et al. Luke: Deep contextualized entity representations with entity-aware self-attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6442-6454.
- [642] Angles R. The property graph database model.[C]//AMW. 2018.
- [643] Miller J J. Graph database applications and concepts with neo4j[C]//Proceedings of the southern association for information systems conference, Atlanta, GA, USA: volume 2324. 2013.
- [644] Consortium W W W, et al. Rdf 1.1 concepts and abstract syntax[J]. 2014.
- [645] Hitzler P, Krötzsch M, Parsia B, et al. Owl 2 web ontology language primer[J]. W3C recommendation, 2009, 27(1):123.
- [646] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26.
- [647] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.
- [648] Wyłot M, Hauswirth M, Cudré-Mauroux P, et al. Rdf data storage and query processing schemes: A survey[J]. ACM Computing Surveys (CSUR), 2018, 51(4):1-36.
- [649] Francis N, Green A, Guagliardo P, et al. Cypher: An evolving query language for property graphs [C]//Proceedings of the 2018 International Conference on Management of Data. 2018: 1433-1445.
- [650] Sirin E, Parsia B. Sparql-dl: Sparql query for owl-dl.[C]//OWLED: volume 258. Citeseer, 2007.
- [651] Rodriguez M A. The gremlin graph traversal machine and language (invited talk)[C]//Proceedings of the 15th Symposium on Database Programming Languages. 2015: 1-10.
- [652] Xu H, Wang W, Mao X, et al. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5214-5223.
- [653] Hoffart J, Suchanek F M, Berberich K, et al. Yago2: exploring and querying world knowledge in time, space, context, and many languages[C]//Proceedings of the 20th international conference companion on World wide web. 2011: 229-232.
- [654] Lehmann J, Isele R, Jakob M, et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2):167-195.

- [655] Bollacker K, Tufts P, Pierce T, et al. A platform for scalable, collaborative, structured information integration[C]//Intl. Workshop on Information Integration on the Web (IIWeb' 07). 2007: 22-27.
- [656] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]// Proceedings of the 2012 ACM SIGMOD international conference on management of data. 2012: 481-492.
- [657] Kolitsas N, Ganea O E, Hofmann T. End-to-end neural entity linking[J]. CoNLL 2018, 2018:519.
- [658] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2619-2629.
- [659] Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 1511-1517.
- [660] Wang Z, Lv Q, Lan X, et al. Cross-lingual knowledge graph alignment via graph convolutional networks[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 349-357.
- [661] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs [J]. arXiv preprint arXiv:1312.6203, 2013.
- [662] Tsarkov D, Horrocks I. Fact++ description logic reasoner: System description[C]//International joint conference on automated reasoning. Springer, 2006: 292-297.
- [663] Haarslev V, Möller R. Racer system description[C]//International Joint Conference on Automated Reasoning. Springer, 2001: 701-705.
- [664] Sirin E, Parsia B, Grau B C, et al. Pellet: A practical owl-dl reasoner[J]. Journal of Web Semantics, 2007, 5(2):51-53.
- [665] Glimm B, Horrocks I, Motik B, et al. Hermit: an owl 2 reasoner[J]. Journal of Automated Reasoning, 2014, 53(3):245-269.
- [666] Eiter T, Gottlob G, Mannila H. Disjunctive datalog[J]. ACM Transactions on Database Systems (TODS), 1997, 22(3):364-418.
- [667] Gebser M, Kaminski R, Kaufmann B, et al. Clingo= asp+ control: Preliminary report[J]. arXiv preprint arXiv:1405.3694, 2014.

- [668] Nenov Y, Piro R, Motik B, et al. Rdfox: A highly-scalable rdf store[C]//International Semantic Web Conference. Springer, 2015: 3-20.
- [669] Forgy C L. Rete: A fast algorithm for the many pattern/many object pattern match problem[M]// Readings in Artificial Intelligence and Databases. Elsevier, 1989: 547-559.
- [670] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[J]. Advances in neural information processing systems, 2013, 26.
- [671] Sun Z, Deng Z H, Nie J Y, et al. Rotate: Knowledge graph embedding by relational rotation in complex space[J]. arXiv preprint arXiv:1902.10197, 2019.
- [672] Purington A, Taft J G, Sannon S, et al. " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo[C]//Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems. 2017: 2853-2859.
- [673] Kaplan A, Haenlein M. Siri, siri, in my hand: Who' s the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence[J]. Business Horizons, 2019, 62(1):15-25.
- [674] Hoy M B. Alexa, siri, cortana, and more: an introduction to voice assistants[J]. Medical reference services quarterly, 2018, 37(1):81-88.
- [675] Wong Y W, Mooney R. Learning synchronous grammars for semantic parsing with lambda calculus[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007: 960-967.
- [676] Liang P. Lambda dependency-based compositional semantics[J]. arXiv preprint arXiv:1309.4408, 2013.
- [677] Yao X, Durme B V. Information extraction over structured data: Question answering with freebase [J]. meeting of the association for computational linguistics, 2014.
- [678] Callan J, Hoy M, Yoo C, et al. Clueweb09 data set[Z]. 2009.
- [679] Gabrilovich E, Ringgaard M, Subramanya A. Facc1: Freebase annotation of clueweb corpora[M]. Version, 2013.
- [680] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers): volume 1. 2015: 260-269.

- [681] Saxena A, Tripathi A, Talukdar P P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[J]. meeting of the association for computational linguistics, 2020.
- [682] Unger C, Forascu C, Lopez V, et al. Question answering over linked data (qald-4)[C]//Working Notes for CLEF 2014 Conference. 2014.
- [683] Usbeck R, Gusmita R H, Ngomo A N, et al. 9th challenge on question answering over linked data (QALD-9)[C/OL]//Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th - 9th, 2018. 2018: 58-64. <https://svn.aksw.org/papers/2018/QALD9/public.pdf>.
- [684] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]// Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1533-1544.
- [685] Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks [J]. arXiv preprint arXiv:1506.02075, 2015.
- [686] Zhang Y, Dai H, Kozareva Z, et al. Variational reasoning for question answering with knowledge graph[C]//Thirty-second AAAI conference on artificial intelligence. 2018.
- [687] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the AAAI conference on artificial intelligence: volume 28. 2014.
- [688] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers). 2015: 687-696.
- [689] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]// International conference on machine learning. PMLR, 2016: 2071-2080.
- [690] Chao L, He J, Wang T, et al. Pairre: Knowledge graph embeddings via paired relation vectors[J]. arXiv preprint arXiv:2011.03798, 2020.
- [691] García-Durán A, Dumančić S, Niepert M. Learning sequence encoders for temporal knowledge graph completion[J]. arXiv preprint arXiv:1809.03202, 2018.

- [692] Goel R, Kazemi S M, Brubaker M, et al. Diachronic embedding for temporal knowledge graph completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 3988-3995.
- [693] Xu C, Chen Y Y, Nayyeri M, et al. Temporal knowledge graph completion using a linear temporal regularizer and multivector embeddings[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2569-2578.
- [694] Li Z, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 4201-4207.
- [695] Wang X, Ye Y, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6857-6866.
- [696] Kampffmeyer M, Chen Y, Liang X, et al. Rethinking knowledge graph propagation for zero-shot learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 11487-11496.
- [697] Liu L, Zhou T, Long G, et al. Attribute propagation network for graph zero-shot learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 4868-4875.
- [698] Guo S, Wang Q, Wang L, et al. Knowledge graph embedding with iterative guidance from soft rules [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.
- [699] Wang P, Dou D, Wu F, et al. Logic rules powered knowledge graph embedding[J]. arXiv preprint arXiv:1903.03772, 2019.
- [700] Cheng K, Yang Z, Zhang M, et al. Uniker: A unified framework for combining embedding and definite horn rule reasoning for knowledge graph inference[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9753-9771.
- [701] He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention[C]// International Conference on Learning Representations. 2020.
- [702] Wang A, Pruksachatkun Y, Nangia N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems[J]. Advances in neural information processing systems, 2019, 32.
- [703] Xing X, Jin Z, Jin D, et al. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3594-3605.

- [704] Lin H, Lu Y, Tang J, et al. A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land?[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7291-7300.
- [705] Si C, Yang Z, Cui Y, et al. Benchmarking robustness of machine reading comprehension models[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 634-644.
- [706] Zhang X Y, Liu C L, Suen C Y. Towards robust pattern recognition: A review[J]. Proceedings of the IEEE, 2020, 108(6):894-922.
- [707] 周志华. 机器学习[M]. 北京: 清华大学出版社有限公司, 2016.
- [708] Gardner M, Artzi Y, Basmov V, et al. Evaluating models' local decision boundaries via contrast sets [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 1307-1323.
- [709] Sakaguchi K, Le Bras R, Bhagavatula C, et al. Winogrande: An adversarial winograd schema challenge at scale[C/OL]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 8732-8740. <https://ojs.aaai.org/index.php/AAAI/article/view/6399>.
- [710] Trichelair P, Emami A, Cheung J C K, et al. On the evaluation of common-sense reasoning in natural language understanding[J]. arXiv preprint arXiv:1811.01778, 2018.
- [711] Zellers R, Bisk Y, Schwartz R, et al. SWAG: A large-scale adversarial dataset for grounded commonsense inference[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 93-104. <https://aclanthology.org/D18-1009>. DOI: 10.18653/v1/D18-1009.
- [712] Ebrahimi J, Rao A, Lowd D, et al. Hotflip: White-box adversarial examples for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 31-36.
- [713] Ren S, Deng Y, He K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1085-1097.
- [714] Jin D, Jin Z, Zhou J T, et al. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]//Proceedings of the AAAI conference on artificial intelligence: volume 34. 2020: 8018-8025.

- [715] Zhao Z, Dua D, Singh S. Generating natural adversarial examples[C]//International Conference on Learning Representations. 2018.
- [716] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11):139-144.
- [717] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.
- [718] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//International conference on machine learning. PMLR, 2017: 1885-1894.
- [719] Chen X, Salem A, Backes M, et al. Badnl: Backdoor attacks against nlp models[C]//ICML 2021 Workshop on Adversarial Machine Learning. 2021.
- [720] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pretrained models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2793-2806.
- [721] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [722] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features[J]. Advances in neural information processing systems, 2019, 32.
- [723] Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy[C]// International Conference on Learning Representations: number 2019. 2019.
- [724] Tishby N, Pereira F C, Bialek W. The information bottleneck method[J]. arXiv preprint physics/0004057, 2000.
- [725] Alemi A A, Fischer I, Dillon J V, et al. Deep variational information bottleneck[J]. arXiv preprint arXiv:1612.00410, 2016.
- [726] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6382-6388.
- [727] Chen J, Yang Z, Yang D. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2147-2157.

- [728] Mozes M, Stenetorp P, Kleinberg B, et al. Frequency-guided word substitutions for detecting textual adversarial examples[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 171-186.
- [729] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2021-2031.
- [730] Ribeiro M T, Wu T, Guestrin C, et al. Beyond accuracy: Behavioral testing of nlp models with checklist[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 4902-4912.
- [731] Delobelle P, Winters T, Berendt B. RobBERT: a Dutch RoBERTa-based Language Model[C/OL]// Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020. <https://www.aclweb.org/anthology/2020.findings-emnlp.292>. DOI: 10.18653/v1/2020.findings-emnlp.292.
- [732] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[C]//International Conference on Learning Representations. 2019.
- [733] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C/OL]// Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. <http://arxiv.org/abs/1412.6572>.
- [734] Gilmer J, Metz L, Faghri F, et al. Adversarial spheres[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=SkthILkPf>.
- [735] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=rJzIBfZAb>.
- [736] Stutz D, Hein M, Schiele B. Disentangling adversarial robustness and generalization [C/OL]//IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 2019: 6976-6987. [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Stutz\\_Disentangling\\_Adversarial\\_Robustness\\_and\\_Generalization\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Stutz_Disentangling_Adversarial_Robustness_and_Generalization_CVPR_2019_paper.html). DOI: 10.1109/CVPR.2019.00714.

- [737] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features[C/OL]// Wallach H M, Larochelle H, Beygelzimer A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 125-136. <https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html>.
- [738] Hauser J, Meng Z, Pascual D, et al. BERT is robust! A case against synonym-based adversarial examples in text classification[J/OL]. CoRR, 2021, abs/2109.07403. <https://arxiv.org/abs/2109.07403>.
- [739] Morris J X, Lifland E, Lanchantin J, et al. Reevaluating adversarial examples in natural language [C/OL]// Cohn T, He Y, Liu Y. Findings of ACL: volume EMNLP 2020 Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Association for Computational Linguistics, 2020: 3829-3839. <https://doi.org/10.18653/v1/2020.findings-emnlp.341>.
- [740] Li L, Ma R, Guo Q, et al. BERT-ATTACK: adversarial attack against BERT[C/OL]// Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 6193-6202. <https://doi.org/10.18653/v1/2020.emnlp-main.500>.
- [741] Garg S, Ramakrishnan G. BAE: bert-based adversarial examples for text classification[C/OL]// Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 6174-6181. <https://doi.org/10.18653/v1/2020.emnlp-main.498>.
- [742] Qi F, Li M, Chen Y, et al. Hidden killer: Invisible textual backdoor attacks with syntactic trigger [C/OL]// Zong C, Xia F, Li W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 443-453. <https://doi.org/10.18653/v1/2021.acl-long.37>.
- [743] Yang W, Lin Y, Li P, et al. Rethinking stealthiness of backdoor attack against NLP models[C/OL]// Zong C, Xia F, Li W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 5543-5557. <https://doi.org/10.18653/v1/2021.acl-long.431>.

- [744] Zheng R, Rong B, Zhou Y, et al. Robust lottery tickets for pre-trained language models[C/OL]// Muresan S, Nakov P, Villavicencio A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 2211-2224. <https://doi.org/10.18653/v1/2022.acl-long.157>.
- [745] Fu Y, Yu Q, Zhang Y, et al. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks[C/OL]// Ranzato M, Beygelzimer A, Dauphin Y N, et al. Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. 2021: 13059-13072. <https://proceedings.neurips.cc/paper/2021/hash/6ce8d8f3b038f737cefcdafcf3752452-Abstract.html>.
- [746] Xi Z, Zheng R, Gui T, et al. Efficient adversarial training with robust early-bird tickets[C/OL]// Goldberg Y, Kozareva Z, Zhang Y. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics, 2022: 8318-8331. <https://aclanthology.org/2022.emnlp-main.569>.
- [747] Gardner M, Artzi Y, Basmova V, et al. Evaluating models' local decision boundaries via contrast sets[C/OL]// Cohn T, He Y, Liu Y. Findings of ACL: volume EMNLP 2020 Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Association for Computational Linguistics, 2020: 1307-1323. <https://doi.org/10.18653/v1/2020.findings-emnlp.117>.
- [748] Mishra S, Arunkumar A. How robust are model rankings : A leaderboard customization approach for equitable evaluation[C/OL]// Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021: 13561-13569. <https://ojs.aaai.org/index.php/AAAI/article/view/17599>.
- [749] Kiela D, Bartolo M, Nie Y, et al. Dynabench: Rethinking benchmarking in NLP[C/OL]// Toutanova K, Rumshisky A, Zettlemoyer L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, 2021: 4110-4124. <https://doi.org/10.18653/v1/2021.naacl-main.324>.

- [750] 杨强等. 可解释人工智能导论[M]. 电子工业出版社, 2022.
- [751] Council of European Union. (eu)2016/679 general data protection regulation(gdpr)[Z]. 2018.
- [752] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning. PMLR, 2018: 2668-2677.
- [753] Blunsom P, Camburu O M, Foerster J, et al. Can i trust the explainer? verifying post- hoc explanatory methods[J]. CoRR, 2019.
- [754] Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps[J]. Advances in neural information processing systems, 2018, 31.
- [755] Yeh C K, Hsieh C Y, Suggala A, et al. On the (in) fidelity and sensitivity of explanations[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [756] Warnecke A, Arp D, Wressnegger C, et al. Evaluating explanation methods for deep learning in security[C]//2020 IEEE european symposium on security and privacy (EuroS&P). IEEE, 2020: 158-174.
- [757] Ribeiro M T, Singh S, Guestrin C. "why should I trust you?": Explaining the predictions of any classifier[C/OL]//Krishnapuram B, Shah M, Smola A J, et al. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, 2016: 1135-1144. <https://doi.org/10.1145/2939672.2939778>.
- [758] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences[C/OL]//Precup D, Teh Y W. Proceedings of Machine Learning Research: volume 70 Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. PMLR, 2017: 3145-3153. <http://proceedings.mlr.press/v70/shrikumar17a.html>.
- [759] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[C/OL]//Bengio Y, LeCun Y. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings. 2014. <http://arxiv.org/abs/1312.6034>.
- [760] Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise[J/OL]. CoRR, 2017, abs/1706.03825. <http://arxiv.org/abs/1706.03825>.

- [761] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C/OL]//Precup D, Teh Y W. Proceedings of Machine Learning Research: volume 70 Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. PMLR, 2017: 3319-3328. <http://proceedings.mlr.press/v70/sundararajan17a.html>.
- [762] Mudrakarta P K, Taly A, Sundararajan M, et al. Did the model understand the question?[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1896-1906.
- [763] Shapley L S. A value for n-person games, contributions to the theory of games, 2, 307–317[M]. Princeton University Press, Princeton, NJ, USA, 1953.
- [764] Strumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions[J/OL]. Knowl. Inf. Syst., 2014, 41(3):647-665. <https://doi.org/10.1007/s10115-013-0679-x>.
- [765] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[J]. 2009.
- [766] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint arXiv:1312.6034, 2013.
- [767] Mahendran A, Vedaldi A. Visualizing deep convolutional neural networks using natural pre-images [J]. International Journal of Computer Vision, 2016, 120(3):233-255.
- [768] Nguyen A, Dosovitskiy A, Yosinski J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[J]. Advances in neural information processing systems, 2016, 29.
- [769] Bastani O, Kim C, Bastani H. Interpreting blackbox models via model extraction[J/OL]. CoRR, 2017, abs/1705.08504. <http://arxiv.org/abs/1705.08504>.
- [770] Breiman L, Friedman J, Olshen R, et al. Classification and regression trees—crc press[J]. Boca Raton, Florida, 1984.
- [771] Li J, Chen X, Hovy E, et al. Visualizing and understanding neural models in nlp[C]//Proceedings of NAACL-HLT. 2016: 681-691.
- [772] Lan Z, Chen M, Goodman S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=H1eA7AEtvS>.

- [773] Manning C D, Clark K, Hewitt J, et al. Emergent linguistic structure in artificial neural networks trained by self-supervision[J/OL]. Proc. Natl. Acad. Sci. USA, 2020, 117(48):30046-30054. <https://doi.org/10.1073/pnas.1907367117>.
- [774] Conneau A, Kruszewski G, Lample G, et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties[J/OL]. CoRR, 2018, abs/1805.01070. <http://arxiv.org/abs/1805.01070>.
- [775] Vulic I, Ponti E M, Litschko R, et al. Probing pretrained language models for lexical semantics [C/OL]/Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 7222-7240. <https://doi.org/10.18653/v1/2020.emnlp-main.586>.
- [776] Hewitt J, Liang P. Designing and interpreting probes with control tasks[C/OL]//Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, 2019: 2733-2743. <https://doi.org/10.18653/v1/D19-1275>.
- [777] Cook R D, Weisberg S. Residuals and influence in regression[M]. New York: Chapman and Hall, 1982.
- [778] Fu J, Liu P, Neubig G. Interpretable multi-dataset evaluation for named entity recognition[J]. arXiv preprint arXiv:2011.06854, 2020.
- [779] Liu P, Fu J, Xiao Y, et al. Explainaboard: An explainable leaderboard for nlp[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021: 280-289.
- [780] Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[C]//Advances in neural information processing systems. 2016: 4349-4357.
- [781] Zhao J, Wang T, Yatskar M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 15-20.

- [782] Stanovsky G, Smith N A, Zettlemoyer L. Evaluating gender bias in machine translation[C/OL]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 1679-1684. <https://aclanthology.org/P19-1164>. DOI: 10.18653/v1/P19-1164.
- [783] Du Y, Zheng Q, Wu Y, et al. Understanding gender bias in knowledge base embeddings[C/OL]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 1381-1395. <https://aclanthology.org/2022.acl-long.98>. DOI: 10.18653/v1/2022.acl-long.98.
- [784] Garg N, Schiebinger L, Jurafsky D, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes[J]. Proceedings of the National Academy of Sciences, 2018, 115(16):E3635-E3644.
- [785] Barocas S, Hardt M, Narayanan A. Fairness and machine learning: Limitations and opportunities [M/OL]. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [786] Xu R, Baracaldo N, Joshi J. Privacy-preserving machine learning: Methods, challenges and directions[J/OL]. CoRR, 2021, abs/2108.04417. <https://arxiv.org/abs/2108.04417>.
- [787] Cristofaro E D. An overview of privacy in machine learning[J/OL]. CoRR, 2020, abs/2005.08679. <https://arxiv.org/abs/2005.08679>.

# 索引

- n* 元文法, 187
- n* 元语法, 187
- n* 元语法单元, 187
- Accountability, 486
- Activation Maximization, 493
- Additive Smoothing, 188
- Adjective, 19
- Adverb, 20
- Adversarial Attack, 459
- Affix, 17
- Algorithm Decomposability, 486
- Algorithm Transparency, 486
- Anaphor, 171
- Anaphora, 171
- Antecedent, 171
- Applicable border, 486
- Argument, 136
- Article, 21
- Aspect Term Extraction, ATE, 321
- Aspect-level Sentiment Analysis, ABSA, 317
- Backdoor Attack, 465
- Bidirectional LSTM, BiLSTM, 34
- Black-Box Attack, 459
- Blind Attack, 459
- Chinese Word Segmentation, 24
- Chomsky Normal From, 54
- Close Class Words, 19
- Community Question Answering, CQA, 340
- Comparative Sentiment, 298
- Completeness, 253
- Componental Analysis, 96
- Conditional Random Field, CRF, 28
- Conjunction, 21
- Constituency Parsing, 53
- Constituent, 49
- Constituent Grammar, 48
- Content Words, 19
- Contextualized Word Embedding, 195
- Coreference, 171
- Coreference Resolution, 171
- Cross-entropy, 214
- Demonstrative Ambiguity, 8
- Dependency Grammar, 49
- Dependency Parsing, 70
- Dependency Tree, 71
- Discourse, 149
- Discourse Segmentation, 157
- Discourse Semantics, 95
- Distance Model, 410
- Distant Supervision, 244
- Distributed Representation, 107
- Document-level Sentiment Analysis, 305
- Double Propagation, 318
- Dynamic Word Embedding, 195
- Ellipsis Ambiguity, 8
- Emotion, 299
- Emotion Classification, 303
- Emotional Sentiment, 299
- Empiricism, 3
- Entailment, 101
- Entity Alignment, 427
- Entity Linking, 423
- Event Extraction, 254
- Explainable Artificial Intelligence, XAI, 484
- Explicit Discourse Relation, 166
- Explicit Sentiment, 298
- Exposure Bias, 386
- Intrinsic Evaluation, 393
- Fertility, 279
- Fine-grained Sentiment Analysis, 318
- Fleiss 卡帕系数, 394
- Form Relations, 97

- Function Words, 19
- Generalization Ability, 455
- Grammar, 48
- Hidden Markov Model, HMM, 39
- Homogeneity, 252
- Implicit Discourse Relation, 166
- Implicit Sentiment, 299
- Inconsistency, 101
- Information Overloading, 370
- Interjection, 21
- Interpretability, 484
- Interpretable Evaluation, 503
- Intrinsic Evaluation, 392
- Knowledge based Question Answering, KBQA, 340
- Knowledge Base Question Answering, KBQA, 440
- Knowledge Graph, KG, 402
- Knowledge reasoning, 432
- Knowledge Representation Learning, 410
- Language Model, LM, 185
- Lemma, 17
- Lemmatization, 23
- Lexcial Primitives, 96
- Lexical Semantics, 95
- Long Short-Term Memory, LSTM, 33
- Machine Reading Comprehension, MRC, 339
- Machine Translation, 267
- Meronymy, 98
- Model Robustness, 454
- Morphological Parsing, 23
- Morphology, 17
- Named Entity, 218
- Named Entity Recognition, NER, 218
- Natural Language Processing, NLP, 1
- Negative Sampling, 111
- Nested Named Entities, 219
- Nested Named Entity, 228
- Non-nested Named Entities, 219
- Noun, 19
- Numeral, 20
- Object Relations, 97
- One-hot Representation, 107
- Online Event Extraction, 259
- Ontology, 404
- Open Class Words, 19
- Open Domain Event Extraction, 259
- Open Relation Extraction, ORE, 248
- Open-domain Question Answering, ODQA, 341
- Opinion, 298
- Opinion Collocation Extraction, 303
- Opinion Collocation Polarity Classification, 303
- Opinion Elements Extraction, 303
- Opinion Holder Extraction, 303
- Opinion Identification, 302
- Opinion Mining, 297
- Opinion Summarization, 304
- opinion target extraction, 303
- Opinion Word Extraction, 303
- Out Of Vocabulary, OOV, 27
- Pair-wise, 394
- Part of Speech, POS, 18
- Part-of-speech Tagging, POS Tagging, 37
- Perplexity, 214
- Phonetic Ambiguity, 6
- Point-wise, 394
- Polarity Classification, 302
- Post-hoc Explanation, 486
- Pragmatic ambiguity, 9
- Preposition, 21
- Presupposition, 101
- Pronoun, 20
- Property Graph, 407
- Questing Answering, QA, 337
- Rational Sentiment, 299
- RDFS, 409
- Relation Extraction, 238
- Resource Description Framework, RDF, 408
- Retrospective Event Extraction, 259
- Robust Machine Learning, 455
- Robustness, 455
- Saliency Map, 490
- Semantic Analysis, 94
- Semantic Case, 99
- Semantic Field, 95
- Semantic Network, 105, 404
- Semantic Representation, 94

- Semantic Role Labeling, SRL, 136  
 Semantic Web, 404  
 Sense Relations, 97  
 Sentential Semantics, 95  
 Sentiment analysis, 297  
 Sentiment Classification, 302  
 Sentiment Information Extraction, 303  
 Sentiment Strength Detection, 303  
 Shapley Value, 492  
 Spam Review Detection, 305  
 Stance Detection, 304  
 Stemming, 23  
 Structural Ambiguity, 8  
 Subjective Classification, 302  
 Submodular Optimization, 494  
 Subword, 115  
 Superstructure, 154  
 Synonym, 101  
 Synonymy, 97  
 Syntax, 48
- Table based Question Answering, TBQA, 340  
 Text Summarization, 370  
 Textual Pattern, 156  
 Theory of Lexcial Primitives, 96  
 Token, 22  
 Transition System, 82  
 Translational Model, 410  
 Truth-conditional Semantics, 99
- V-measure, 253  
 Verb, 19
- Web Ontology Language, OWL, 409  
 White-Box Attack, 459  
 Word Distributed Representation, 107  
 Word Normalization, 22  
 Word Segmentation Ambiguity, 7  
 word Sense Ambiguity, 7  
 Word Sense Disambiguation, WSD, 123  
 Word Tokenization, 22
- 上下文相关的词向量, 195  
 下指照应, 151  
 中文分词, 24  
 主客观分类, 302  
 义元, 96  
 义元理论, 96
- 乔姆斯基范式, 54  
 事件抽取, 254  
 交叉熵, 214  
 介词, 21  
 代词, 20  
 依存句法分析, 70  
 依存树, 71  
 依存语法, 49  
 信息抽取, 216  
 信息过载, 370  
 修辞结构理论, 155
- 先行词, 171  
 全局解释, 488  
 共指, 171  
 关系抽取, 238  
 内在评价, 392  
 内在语义特征, 96  
 内指照应, 151  
 冠词, 21  
 分布式表示, 107  
 切分歧义, 26  
 副词, 20  
 功能词, 19  
 加一平滑, 188  
 加法平滑, 188  
 动态词向量, 195  
 单向语言模型, 199  
 单词分布式表示, 107  
 双向传播, 318  
 双向长短期记忆网络, 34  
 反义关系, 101  
 古德-图灵估计法, 188  
 句子级情感分析, 312  
 句子语义学, 95  
 句法, 48  
 可解释人工智能, 484  
 可解释性, 484  
 可解释评估, 503  
 同义关系, 97, 101  
 同质性, 252  
 名词, 19  
 后门攻击, 465  
 命名实体, 218  
 命名实体识别, 218  
 回指, 171  
 回指照应, 151  
 回顾事件抽取, 259

- 困惑度, 214
- 在线事件抽取, 259
- 垃圾评论检测, 305
- 外在评价, 393
- 外指照应, 151
- 子词, 115
- 完整性, 253
- 实义词, 19
- 实体关系, 97
- 实体对齐, 427
- 实体链接, 423
- 对抗攻击, 459
- 对抗样本检测, 473
- 局部解释, 488
- 属性图, 407
- 属性级情感分析, 317
- 属性词抽取, 321
- 嵌套命名实体, 219, 228
- 常规型观点, 298
- 平滑, 188
- 平移模型, 410
- 开放关系抽取, 239, 248
- 开放域事件抽取, 259
- 开放领域问答, 341
- 开类词, 19
- 归纳推理, 432
- 形体关系, 97
- 形容词, 19
- 形态学, 17
- 情感信息抽取, 303
- 情感分析, 297
- 情感分类任务, 302
- 情感强度判别, 303
- 情绪, 299
- 情绪分类, 303
- 意义关系, 97
- 感受性语义特征, 96
- 感叹词, 21
- 感性情感, 299
- 成分, 49
- 成分句法分析, 53
- 成分语法, 48
- 抽取式文本摘要, 373
- 指代歧义, 8
- 指代消解, 171
- 搭配关系, 150
- 数词, 20
- 文本摘要, 370
- 文档级情感分析, 305
- 显式篇章关系, 166
- 显式观点, 298
- 显著图, 490
- 智能问答, 337
- 曝光偏差问题, 386
- 替代, 151
- 最大熵模型, 239
- 未登录词, 27
- 本体, 404
- 机器翻译, 267
- 条件随机场, 28
- 极性分类, 302
- 框架语义学, 96
- 模型稳健性, 454
- 次模优化, 494
- 比较型观点, 298
- 沙普利值, 492
- 泛化能力, 455
- 溯因推理, 432
- 演绎推理, 432
- 激活最大化, 493
- 照应, 150
- 照应词, 171
- 独热表示, 107
- 理性情感, 299
- 白盒攻击, 459
- 盲攻击, 459
- 省略, 151
- 省略歧义, 8
- 真值条件语义学, 99
- 知识图谱, 402
- 知识图谱问答, 340, 440
- 知识推理, 432
- 知识表示学习, 410
- 社区问答, 340
- 稳健性, 455
- 算法事后解释, 486
- 算法可担责性, 486
- 算法可解构性, 486
- 算法适用边界, 486
- 算法透明度, 486
- 篇章, 149
- 篇章级情感分析, 305
- 篇章超级结构, 154
- 繁衍率, 279

- 细粒度情感分析, 318  
经验主义, 3  
结构歧义, 8  
网络本体语言, 409  
  
自然语言处理, 1  
自然语言理解, 1  
自然语言生成, 1  
蕴含关系, 101  
衔接, 150  
表格问答, 340  
表示学习, 13  
表述发现, 171  
观点, 298  
观点持有者抽取, 303  
观点挖掘, 297  
观点摘要, 304  
观点识别, 302  
论元, 136  
论域, 102  
评价对象抽取, 303  
评价搭配抽取, 303  
评价搭配极性判别, 303  
评价词抽取, 303  
词, 17  
词义歧义, 7  
词义消歧, 123  
词对齐, 272  
词干提取, 23  
词形, 22  
词形分析, 23  
词形还原, 23  
词性, 18  
词性标注, 37  
词根, 17  
词汇语义学, 95  
词缀, 17  
词语切分, 22  
词语切分歧义, 7  
词语规范化, 22  
话语分割, 157  
话语语义学, 95  
语义分析, 94  
语义场理论, 95  
语义学, 94  
语义成分分析, 96  
语义格, 99  
语义网, 404  
  
语义网络, 105, 404  
语义表示, 94, 101  
语义角色标注, 136  
语法, 48  
语法-语义特征, 96  
语用歧义, 9  
语篇模式, 156  
语素, 17  
语言模型, 185  
语音歧义, 6  
负采样, 111  
资源描述框架, 408  
资源描述框架模式, 409  
距离模型, 410  
转移系统, 82  
辩论立场检测, 304  
远程监督, 244  
连接, 151  
连词, 21  
连贯, 152  
  
逐对评估, 394  
逐点评估, 394  
部分整体关系, 98  
重述关系, 150  
长短期记忆网络, 33  
闭类词, 19  
阅读理解, 339  
隐式观点, 299  
隐马尔可夫模型, 39  
非嵌套命名实体, 219  
预定关系抽取, 239  
预设关系, 101  
鲁棒机器学习, 455  
黑盒攻击, 459