自然语言处理与生物识别

Natural Language Processing and Biometric Identification

2020 年 3 月 2 日

Springer

目录

Chapter 1
NLP 发展现状与应用领域

1.1 定义简介

自然语言处理（Natural Language Processing，简称 NLP），属于计算机科学与语言学的交叉学科，所以又称计算语言学；它是用计算机来理解、处理、运用人类语言的学科。人类通过自然语言进行沟通协作，可以说如果没有语言人类的智能将无从谈起，它是人类区别于动物的重要标志。也可以说，只有当计算机具备了准确的自然语言的理解处理能力时，才算真正实现了人工智能。

- 研究内容：NLP 研究内容主要包括词法分析、句法分析、语义分析、篇章理解、机器翻译等。
- 应用场景：NLP 广泛应用信息系统方方面面。例如：手写体识别、光学符号识别、语音识别、语音合成、信息检索、机器翻译、对话系统等。
- 关联学科：NLP 紧密相关的研究领域包括机器学习、数据挖掘、知识图谱等；紧密相关的学科包括信息论、语言学、计算机科学等。

NLP 研究范围涉及自然语言的形态学、语法学、语义学和语用学等几个层次。

- 形态学（morphology）：研究词的内部结构，包括屈折变化和构词法两个部分。
- 语法学（syntax）：研究句子结构成分之间的相互关系、句子序列的组成规则。

- 语义学（semantics）：研究各级语言单位（词素、词组、句子、段落、片等）的意义，以及与语音、语法、语境的关系等等，其重点在探明符号与其所指对象之间的关系。
- 语用学（pragmatics）：研究在不同上下文下的语句应用，以及上下文对语句理解所产生的影响。大概来说，语用学研究范围问题是很广，重点在于研究包括直指、会话隐含、预设、语言行为、话语结构等。

NLP 面临的两大难题是歧义消解、未知语言现象。

- 歧义消解：在自然语言的词法、句法、语义等各个层次中存在大量的歧义现象。比如“什么是一个词”，这就是 NLP 面临的一个难题。因为不同的人对词语粒度、标准有不同的理解。再比如，语言中存在大量一词多义的现象，在上下文语境中如何准确找到对应的词义，这些都是 NLP 研究面临的实实在在的难题。
- 未知语言现象：未知语言现象主要由两个方面的原因导致。第一点，人类语言一直处于不断演化过程中，同一个语言表达，在新的时空环境下，可能已经不再是以前的含义；而且，由于信息网络的发达，人们构造、传播新语言现象的能力大大增强，比如互联网上每天都在涌现新的语言词汇。第二点，在 NLP 研究中，由于整理收录能力、知识表达能力等现实因素的制约，实际中并没有一种可以准确、全面表达人类语言知识工程或工具。

1.2 发展历史

NLP 发展历史中存在两种不同的的研究方法：基于规则的理性主义；基于统计的经验主义。它们对语言的不同理解，体现了它们不同的哲学思想。

- 理性主义：认为自然语言是由语言规则来产生和描述的；因此他们相信，只要能够用适当的形式将人类语言规则表示出来，就能够理解人类语言。
- 经验主义：认为语言知识可以从语言数据中获取，只要建立有效的统计模型就可以理解语言；因此他们相信，如果有足够多语言数据用于统计，就能够理解人类语言。

NLP 发展历史可以总结为 5 个时期：

1. 经验主义萌芽时期：时间大约到 20 世纪 50 年代。这个时期 NLP 或多或少具有经验主义色彩。例如，1913 年马尔科夫提出马尔科夫模型 [1] 的时候，就曾经计算过长诗中元音与辅音出现的频度概率；再比如，1948 年，香农把离散马尔科夫的概率模型应用于语言自动机 [2] 的时候，也曾统计过英语字母的频率。
2. 经验主义低谷时期：时间从 1956 年 20 世纪 90 年代。1956 年，乔姆斯基首先提出使用有限状态机来刻画自然语言 [3]。具体来说，就是使用数学的代数、集合论为基础核心，将各种语言现象统一抽象为代数、集合上的运算规则。形式语言理论影响深远，在此后很长一段时期，很多学者逐步完善并扩展了形式语言理论。这段时期，NLP 领域的研究方法几乎完全被理性主义主导，经验主义被打入谷底。
3. 经验主义复苏时期：时间从 20 世纪 50 年代末到 90 年代初期。虽然这期间，例行理性主义占据主流，但是有学者已经开始思考引入基于语料库的统计方法到 NLP 研究中；这其中的代表是 1967 年诞生的联机语料库 Brown Corpus[4]。另外，和 NLP 紧密相关的机器学习方法得到了较大的发展；比如这段时期，诞生了贝叶斯模型、最大熵 [5]、维特比算法 [6]、隐马尔可夫模型 [7] 等等，NLP 研究有了更多理论工具可供使用。
4. 经验主义爆发时期：时间 20 世纪 90 年代中期 2010 年左右。时间处于 20 世纪 90 年代前期，此刻经验主义已经处于全面复苏的前夜。一方面，由于机器学习领域诞生了很多新理论和方法，推动了 NLP 快速发展；另一方面，计算机的存储容量、计算能力已经极大提升，使得很多计算量偏大的机器学习方法逐渐实用。受益于这两方面，经验主义终于开始全面复苏，迎来了一个前所未有的黄金时期。20 世纪 90 年代以后，以语料库和统计学习为基础，基于机器学习的词法分析、句法分析、机器翻译、语音识别等研究不断涌现。
5. 经验主义现代时期：时间从 2003 年左右至今。这段时期，经验主义再上高峰，具有更加鲜明的特色，神经网络、深度学习的是目前 NLP 研究的关键词。Ruder 在博文 [8] 在中总结了这段时期里 NLP 的研究趋势，其中以神经语言模型（Nerual Language Model，NLM）、多任务学习（Multi-task Learning）、词嵌入（Word Embedding）、Seq2Seq（Sequence To Sequence Model）模型、注意力（Attention）机制、基于

记忆的网络（Memory-based Networks）、预训练语言模型（Pretrained Language Models）等这个时期重要的里程碑。

1.3 应用领域

自然语言处理研究的内容非常广泛，应用范围也非常广泛，如下举例一些常见的应用场景：

- 机器翻译：利用计算机实现自然语言（英语、汉语等）之间的自动翻译。
- 自动摘要：利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文。
- 文本分类：在预定义分类体系下，根据文本的特征，将给定文本于一个或多个类别相关联的过程。
- 情感分类：根据文本所表达的含义和情感将文本划分为褒扬或者贬义的两种或几种类型，是对作者倾向性、观点、态度的划分，因此也称倾向性分析。
- 信息抽取：从非结构化或半结构的自然语言文本中提取出于某个主题相关的实体、关系。事件等事实信息，并且形成结构化信息输出。
- 信息检索：用户输入一个表述需求信息的查询字段，系统回复一个包含所需要信息的文档列表。其核心技术在于索引构建和相关性计算。
- 问答系统：接受用户自然语言形式描述的问题，从大量异构数据中查找或者推断用户问题答案的信息检索系统。

Chapter 2
语言模型与主题模型

本章简单介绍自然语言处理中的语言模型和主题模型的概念，并其常见的算法，以及在语音识别中的应用。

2.1 语言模型

2.1.1 基本定义

语言模型（Language Model）用于计算语言序列 $w_1, w_2, \cdots, w_n$ 的概率，数学表示为 $P(w_1, w_2, ..., w_n)$，它是对语句的概率分布的建模。其最直接的应用就是判断一句话来自于人生成的语句的概率，例如在我们自然语言中，句子 “我去吃饭” 相比于 “吃饭去我” 的出现的概率更高，因此 $P(\text{“我去吃饭”}) > P(\text{“吃饭去我”})$。讲到这里，最直接的一个问题就是，如何计算 $P(w_1, w_2, ..., w_n)$ 呢? 我们下面介绍一种最基本的语言模型: n-gram 语言模型。

2.1.2 n-gram 语言模型

n-gram 语言模型是一种最基础的语言模型。根据链式法则 (Chain Rule)，公式 $P(w_1, w_2, ..., w_n)$ 可以得到:

$$P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, \cdots, w_{n-1})$$

其中的每一项 $P(w_i|w_1,\cdots,w_{i-1})$，可以用以下公式来估计，即:

$$P(w_i|w_1,\cdots,w_{i-1}) = \frac{C(w_1,\cdots,w_{i-1},w_i)}{C(w_1,\cdots,w_{i-1})}$$

其中，$C(\cdot)$ 表示该序列在训练语料中出现的次数。但是，当序列长度很长时候，计算 $P(w_i|w_1,\cdots,w_{i-1})$ 比较困难，一种常见的处理方式是引入马尔可夫假设 (Markov Assumption)，即假设当前词出现的概率只依赖于前 $n-1$ 个词，也就是:

$$P(w_i|w_1,\cdots,w_{i-1}) = P(w_i|w_{i-n+1},\cdots,w_{i-1})$$

根据 n 取值的不同，我们可以得到不同的 n-gram 语言模型:

- Unigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i)$
- Bigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i|w_{i-1})$
- Trigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i|w_{i-1},w_{i-2})$

2.1.3 n-gram 语言模型中的平滑技术

在计算 n-gram 时候，一个很重要的问题就是测试集中出现了训练集中未出现过的词而导致语言模型计算出的概率为零，我们称这些词为未登录词 (OOV)。平滑 (Smoothing) 技术就是为了缓解这类问题，常见的平滑技术有：拉普拉斯平滑 (Laplace Smoothing),、古德图灵法（good-turing)、线性减值法（Linear Discounting）等，感兴趣的读者可以深入阅读相关论文。

2.1.4 语言模型在语音识别中的应用

自动语音识别（Automatic Speech Recognition，ASR）是一种将人的语音转换为文本的技术，它是目前很多互联网产品如语音助手，语音搜索引擎等中必不可少的一部分。图 2-1给出了常见的语音识别系统的基本工作流程。其中基本可以分为以下几个模块:

- 数据预处理: 典型的预处理包含静音处理（Voice Activity Detection，VAD）等，用于去除其中的静音片段。

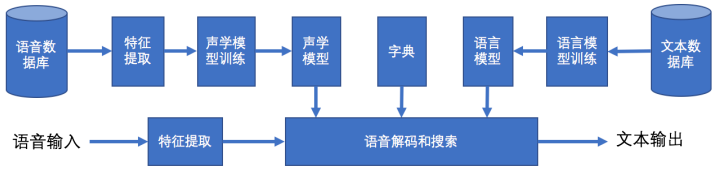


图 2-1 语音音识别基本流程

- 特征提取：将声音转换成包含声音信息的多维向量，常见的有 MFCC 等。
- 声学模型：主要是通过语音数据训练得到，其输出是音素等信息。
- 词典：字/词和音素之间的对应关系。
- 语言模型：也就是上文提到的语言模型部分，主要用于评估字或者词序列的概率。

语音系统首先将语音信号做特征提取工作，转化成诸如 MFCC 等特征来表示，然后使用语言模型和声学模型来解码，解码过程会产生很多候选 (Candidates)，最终最优的候选会被输出成为最终的结果。语言模型是其中很重要的一部分，它用于从根据语言统计规律评估声学模型给出的句子序列候选的概率，决定了最终输出的结果。

2.2 主题模型

2.2.1 基本概念

主题模型 (Topic Models) 是近些年来非常重要的一项技术，它被广泛应用于工业和学术界。在主题模型中，我们一般用 d 来表示要分析的文档，例如一篇文章或者一个网页等，而一个文档 d 通常由一系列词 $(w_1, w_2, ..., w_n)$ 组成，其中 w_n 是文档中的第 n 个词。多份文档共同构成了我们要分析的语料集，我们用 $\mathscr{D}$ 来表示，$\mathscr{D} = (d_1, d_2, ..., d_m)$ 组成，其中 d_m 是语料库中的第 m 个文档。主题一般用 z 来表示，它由一些词组成，同时也有该词在这个主题下的概率。主题模型泛指由一类可以从语料库

中抽取主题并利用这些主题表示文档的模型，常见的主题模型有 PLSA，LDA，以及各种 LDA 的变种，例如 SentenceLDA 等。在熟悉了这些基本概念之后，我们通过一种常见的主题模型 Latent Diriclet Allocation (LDA) 来认识主题模型。

2.2.2 常见的主题模型：LDA

2003 年 Blei 等人在《Latent Dirichlet Allocation》 [9] 一文中提出了 LDA 模型。如图2-2所示，其中空心节点表示隐藏变量，实心变量表示客观测变量，整个模型具有 K 个主题，M 个文档和 N 个词。LDA 将文档的主题分布 $P(z|d)$ 看做随机变量 θ，并且假设 θ 从一个狄利克雷先验中产生。同时，由于训练数据之外的文档对应的主题分布 θ 可以从上述狄利克雷分布中产生，训练数据之外的文档的 θ 可以更自然地进行计算。

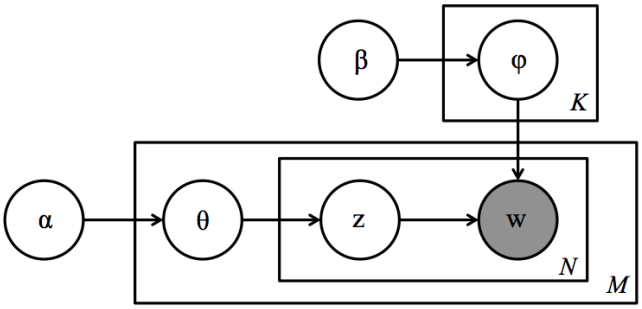


图 2-2 LDA 图模型

读者可以使用常见的开源主题模型工具如 LightLDA[10]、Familia[11] 来尝试训练 LDA 模型。近年来，出现了各种对主题模型的扩展工作，例如联邦学习场景下的主题模型 [12] 工作等。

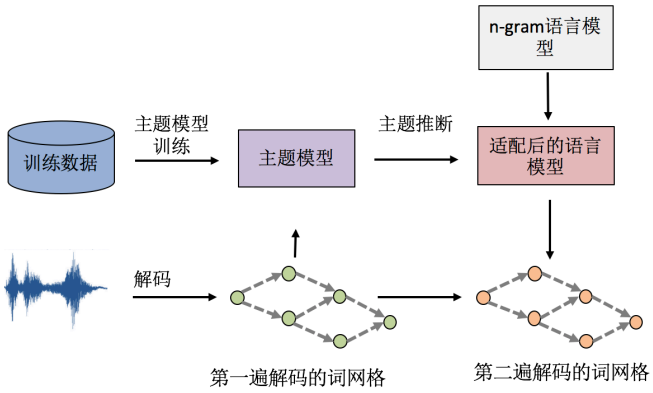


图 2-3 语言模型适配

2.2.3 主题模型在语音识别中的应用：语言模型适配

语音识别系统中一个常见的问题就是，我们训练语言模型的语料和它实际线上应用的语料之间存在不一致，这种情况下，除了重新训练模型，有一种代价更小的方法就是语言模型适配（Language Model Adaptation）。语言模型适配指的是用实际应用的语料相关的信息，对语言模型做适配。图 2-3给出了其常见的工作流程，采用预先训练好的主题模型，我们对语音识别系统第一遍识别出来的词网格 (Word Lattice) 做主题推断，可以发现其语义级别的内容，同时它也可以作为一个 unigram 的语言模型 $P_{LDA}(w|\theta_d) = \sum_{k \in K} \varphi_{kw} \theta_{dk}$，对之前的 n-gram 语言模型就行适配：

$$P_d(w|C) = \lambda P_{LDA}(w|\theta_d) + (1-\lambda) P_{n-gram\ LM}(w|C) \tag{2.1}$$

其中 C 代表当前词 w 的上下文，λ 是一个权重参数，$P_{n-gram\ LM}(w|C)$ 是基础的 ngram 语言模型给出来的评估分数。这个新适配过的语言模型，可以用于语音识别系统，重新解码出新的词网格以及最终的结果。

Chapter 3
词法，语法与语义分析

3.1 词法分析

3.1.1 问题定义

词法分析是自然语言处理的第一步，要做 NLP 深层次分析，比如句法分析、语义分析，甚至 NLP 复杂应用的先决条件，就是首先进行词法分析。词法分析的核心是，将自然语言解析为一个个词的序列，并判断每个词的词性、专名信息，为后续分析做好准备。总的来说，在中文这种孤立语中，词法分析主要由分词、词性标注、命名实体识别 3 个子任务组成。在英语、阿拉伯语等屈折语中，词法分析一般还包括词根还原（Word Stemming）任务。忽略词根还原，词法分析可以归纳为 3 个具体的子任务：

- 自动分词（Word Segmentation）：是将连续的自然语言文本，切分成具有语义合理性和完整性的词汇序列的过程。
- 词性标注（Part-of-Speech Tagging）：是指为自然语言文本中的每个词汇赋予一个词性的过程。
- 命名实体识别（Named Entity Recognition，也称专名识别）：是指识别自然语言文本中具有特定意义的实体，主要包括人名、地名、机构名、时间日期等。

如图3-1所示，将输入句子切分成一个个词汇，然后给每个词汇标记出名词、动词、介词等词性；并且识别出“2003 年 10 月 15 日”是一个时间专名，识别“杨利伟”是一个人名专名等。

图 3-1 词法分析示例

因为词法分析的自动分词、词性标注、专名识别本质上是类似的，所以它们的发展历史方也是类似的。总结起来，都大致经历词典匹配、机器学习、深度学习这 3 个发展阶段。

3.1.2 实现方案

3.1.2.1 词典匹配

基于词典匹配的实现步骤：

1. 词典构建：根据具体子任务的语言学知识构建词典。如果是自动分词，则收录常见的词条短语；如果是词性标注、专名识别，则收录当前常见词条对应的名词、动词、介词、专名类型等属性，并且保留相应词性、专名属性的概率。词典构建，通常需要人工收集、整理、离线更新，维护成本较高。
2. 词典匹配：扫描输入的所有子序列，如果当前子序列能够匹配词典中某个词条，则当前子序列属一个可能的候选。如果是分词，则当前子序列就可能为一个分词的词汇；如果是词性标注、专名识别，则当前子序列以相应的概率取词典中的词条属性。
3. 歧义消解：由于候选子序列之间存在歧义，所以基于词典匹配之上，需要加入一些启发式规则以解决歧义问题。常用规则主要包含最大前向匹配（Forward Maximum Matching，FMM）、最大后向匹配（Backward Maximum Matching，BMM）、最少切分、双向最大匹配、长片段优先等策略 [13][14]。

中文分词如图3-2所示。输入序列进行词典匹配之后，得到对应的 DAG 图，图中每条边都是词典词条，边的权重都为 1，图中每一种首尾

贯通的路径都是一种候选的歧义切分结果。只要求解 DAG 图的最优路径，则可以得到输入序列对应的分词结果。

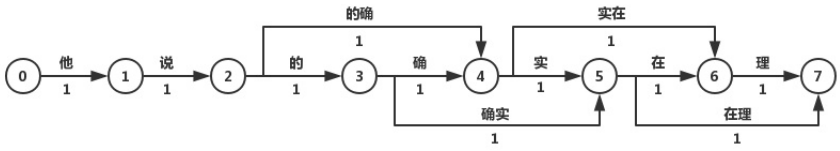


图 3-2 中文分词示例

很长一段时间内研究者都在对基于词典匹配方法进行优化，比如最大长度设定、字符串存储和查找方式以及对于词表的组织结构，比如采用 TRIE 索引树 [15]、哈希索引 [16] 等、AC 自动机 [17] 等结构方便快速查找。

基于词典匹配的优缺点：

- 优点：实现简单、运行速度快
- 缺点：面临词典收录更新困难、未登录词难处理；同时由于消除歧义策略过于简单，通常效果不甚理想。

3.1.2.2 机器学习

基于机器学习的实现步骤：

1. 转换为序列标注任务：词法分析 3 个子任务，都通过定义标注空间标签集，将具体任务转换为标准序列标注任务。以中文分词为例：转换方法为，标注每个字在其所属词中的位置。因为对于任何一个字来说，它可以是一个词的开始（Begin），一个词的中间（Inside），一个词的结尾（End），或者本身就是一个单字的词（Singleton），这也就是在分词序列标注中常用的 BIES 的分类。只需将输入序列的每个字标上 BIES 标签中的一个，就可以转换得到对应的分词结果。这种标注空间（模型状态空间）的划分在词性标注和专名识别任务上也很常用，也会有一些类似的变种，比如专名识别中常用 BIO 标签集。

2. 求解序列标注任务：传统序列标注模型主要包括隐马尔科夫模型（Hidden Markov Model，HMM）[18][19][20][21]、最大熵马尔科夫模型（Maximum Entropy Markov Model，MEMM）[22][23][24]、条件随机场（Conditional Random Field，CRF）[25][26][27]、结构化感知机（Structural Perception Machine）[28][29] 等浅层模型。这些浅层模型的区别主要在于如何对待输入字序列和标签序列之间的概率，训练目标是最大联合概率似然、最大条件概率似然，还是最小化风险等。

总结来说：传统序列标注模型中，CRF 是集大成者。相比于 HMM，CRF 去除了输出独立性要求，对于整个序列内部的信息和外部观测信息都可以有效利用，可以更加有效建模上下文。相比于 MEMM，CRF 通过全局归一化（global normalization），避免了 MEMM locally normalized 导致的 label bias 缺陷。

以自动分词任务为例，则其序列标注任务定义为：定义标签集为 $L = \{B, I, E, S\}$，给定输入文本序列 $X = \{x_1, x_2, ..., x_n\}$，目标是求解最优标注序列 $Y^* = \{y_1^*, y_2^*, ..., y_n^*\}$：

$$y^* = \arg\max_{Y \in L^n} p(\boldsymbol{Y}|\boldsymbol{X})$$

预测时，使用维特比算法求解最优标注序列 Y^*；训练时，使用最大化条件似然来训练模型，其中，f_i 为人工定义的特征函数：

$$p_w(Y|X) = \frac{1}{Z_w(x)} \exp(\sum_i w_i f_i(y_i, y_{i-1}, x))$$

$$Z_w(x) = \sum_{y \in Y} \exp(\sum_i w_i f_i(y_i, y_{i-1}, x))$$

基于机器学习的优缺点：

- 优点：通过人工设计的特征工程，充分地挖掘了序列的上下文信息。模型的歧义消解胜过词典匹配的方法；同时具有很强的泛化能力，能够很好地处理未登录词问题。
- 缺点：特征工程不经需要耗费大量人力，而且需要大量语言学知识，设计和寻找有效特征存在较高门槛；另外，这些浅层模型，通常使用离散的 binary 特征，无法表达复杂先验，比如没法利用词向量。

3.1.2.3 深度学习

随着大数据、神经网络、深度学习的快速发展，很多研究提出利用前馈神经网络来解决词法分析 [30][31] 任务。总的来说，与前面基于机器学习的方法类似，也是把词法分析任务作为序列标注问题进行求解，只是把人工设计特征函数，改成了使用多层前馈神经网络进行自动特征抽取。

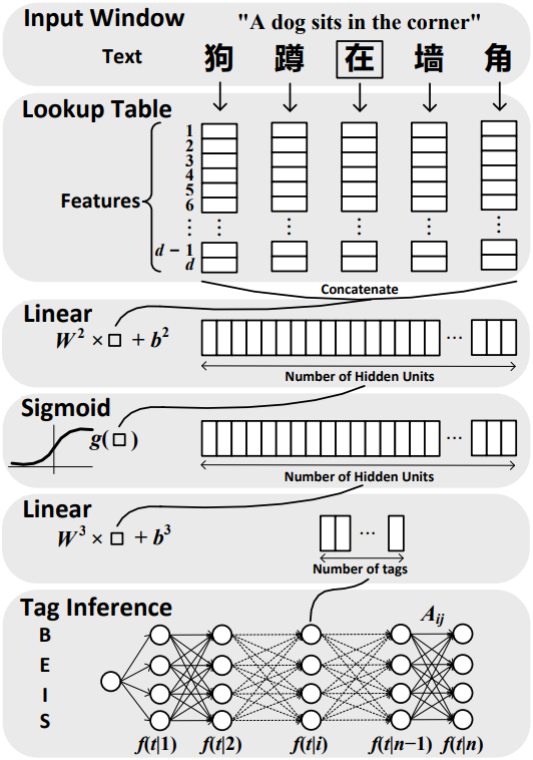


图 3-3 深度学习分词示例

如图3-3所示，网络的第一层输入句子中每个字的字向量，第二层将一个固定长度的字向量进行拼接，然后输入到标准前馈神经网络结构中，神经网络输出在标注集合上的 lattice，最后利用维特比算法进行解码，就可以得到最优标注序列。

后续有很多研究，深度学习框架基础上，对前馈神经网络进行改进，进一步探索了卷积神经网络 [32]、循环神经网络 [33]、递归神经网络 [34] 等复杂结构对词法分析任务的影响。

另外，传统的词法分析通常会把分词、词性标注、命名实体识别当成 pipeline 形式的进行处理，这样容易导致错误传播问题。比如，如果分词任务出现错误，将会导致后续词性标注、专名识别、句法分析、语义分析出现错误。所以在学术界也有很多联合建模 [29][35][36] 的工作，通过多个任务联合学习，以期望避免错误传播。联合建模的优点：各个子任务可以充分共享有用信息，通常可以取得比单任务更好的结果。联合建模的缺点：解码空间变大，搜索的复杂度提高，需要更有效的搜索剪枝机制在降低搜索复杂度的同时，尽量保持搜索结果的质量。

值得一提的是，2018 年诞生了 ELMO[37]、GPT[38]、BERT[39] 这 3 个杰出代表。它们通过将语言模型嵌入作为特征，使用目标任务数对语言模型对进行微调，使得众多 NLP 任务都获得了大幅度提升，这其中就包括了自动分词、词性标注、专名识别。

总结起来，基于深度学习方法的优点是通过深度神经网络自动学习多层特征抽象，避免了复杂的特征工程；模型的歧义消解、泛化能力通常都很好。只是，深度学习模型需要训练充分，通常对数据量、计算量都有较高的要求。

3.1.3 应用案例

通常 NLP 中深层次的语法语义分析通常都是以词作为基本单位，所以词法分析都是这些深层次分析的基础和先决条件。词法分析作为自然语言处理的第一步，是下游所有分析任务的基础。这些下游应用小句法分析、语义分析，大到对话系统、文本分类、自动摘要、机器翻译 [40]、信息检索、搜索引擎、语音合成等等。几乎只要有利用到自然语言处理技术的地方，词法分析都是不可或缺的基础技术。

比如在搜索引擎中，用户输入一个表述需求信息的查询字段，系统回复一个包含所需要信息的文档列表。其核心技术在于索引构建和相关性计算。

索引构建：首先需要对互联网海量文档进行，分词、词性标注、专名识别；利用词法分析的结果，以词为粒度，为所有文档建立倒排索引。

相关性计算：首先需要对用于查询字段进行词法分析，找出表达用户需求的核心词汇；然后以这些核心词汇为 key，去拉取相应的倒排索引，获取初步筛选后的文档；接着，为了要提高返回结果的相关性，需要计算用户查询和初筛文档之间的相似度，取相似度较好的文档返回。其中计算用户查询和文档之间的相似度，通常利用 BM25、语义计算等技术，而它们也都是以词为单位进行的，由此看出词法分析在搜索引擎中的重要性、基础性。

在比如在文本分类和情感分类中，在预定义分类体系下，输入给定文本，抽取文本特征，将给定文本于一个或多个类别相关联的过程。其中，抽取文本特征的时候，通常也都是先进行词法分析，先获得给定文本的词汇、词性、专名等信息，然后以词汇为单位，以词汇的词向量、词性专名等信息为特征，利用机器学习、深度学习的分类模型进行预测。

3.2 语法分析

3.2.1 问题定义

短语结构语法（Constituency Structure Grammar）和依存关系语法（Dependency Grammar）是现在常见的两种语法关系。短语结构语法又叫上下文无关文法 (Context-Free Grammars，CFGs)，它从一个特殊的初始符号出发，不断的应用一些产生式规则，从而生成出一个字串的集合（如句子）。产生式规则指定了某些符号组合如何被另外一些符号组合替换。它呈现一个树分类关系，句法根据一定的规则进行转换分析。每一个词的转换都是需要按照设定的树值规则进行目的性的转换。

依存语法 (从属关系语法) 是由法国的语言家 Lucien Tesnière 提出的 [41]，它将句子各个词语之前的搭配关系描述成预先定义好的依存关系。它基于一个基本假设：句法结构本质上包含词和词之间的依存（修饰）关系。一个依存关系连接两个词，分别是核心词（head）和依存词（dependent）。依存关系可以细分为不同的类型，表示两个词之间的具体句法关系。比如主语依赖于谓语（SBV），宾语也依赖于谓语（VOB）以

及定语依赖于名词性短语（ATT）等。依存句法认为“谓语”中的动词是一个句子的中心，其他成分与动词直接或间接地产生联系。

除了以上介绍的两种句法体系外，国内外都开展了对句法分析的研究。不论是国外的链语法 (Link Grammar)、组合范畴语法（Combinatory categorial grammar，CCG）等，还是国内黄曾阳教授提出的 HNC 理论 (Hierarchical Network of Concept)[42] 都是目前行业内常用的语法，只是由于设定区域的不同，所以使用有一定的局限性。

短语结构分析的语法集是由固定的语法集组成，较为固定和呆板，依存语法则更加的自由。另外依存语法树标注简单且 parser 准确率更高，再加之通用依存数据集（Universal Dependencies Treebanks）的发展，依存语法分析受到专家学者普遍的青睐，得到越来越多的应用。这里也将着重介绍依存语法分析。

图 3-4 LTP 依存句法分析样例

例如，从上述例子中我们可以看到，句子的核心谓词为“强调”，主语（SBV）是“他”，强调的宾语（VOB）是“紧迫性”，“紧迫性”的修饰语（ATT）是“武汉疫情控制”。

3.2.2 实现方案

依存句法分析方法主要可以分为两种，一种是基于图的方法（Graph Based），一种是基于转移（Transition Based）的方法。基于图的方法先建立句子中所有词语的全连接图，然后求图中的最大生成树。两种方法中更主流的算法是基于转换的依存句法分析，基于转移的方法将依存树的构成过程建模为一个动作序列，将依存分析问题转化为寻找最优动作序列

的问题。基于转移的解析过程是线性的，动作步骤随句子长度线性增长，而基于图的方法需要在全图上做搜索，所以时间复杂度上基于转移的方法会有优势。但是基于转移的方法在解析的每一步都只是利用局部信息，会导致错误传播，性能比基于图的略差。

近几年，分别出现了针对这两种不同方法的神经网络模型。比如基于图的 [43, 44, 45]，直接用神经网络来预测每两个词之间存在依存关系的概率，得到一个全连接图，图上每个边代表了节点 a 指向节点 b 的概率，然后使用 MST 等方法来来将图转换为一棵树。概率的计算可以简单的使用节点 a 和节点 b 的 embedding 向量做向量运算，也可以使用复杂的多层 GNN 网络迭代更新。基于转移的如 [46, 47, 48]，通过两个 LSTM 来分别建模 stack 状态、buffer 状态，使用第三个 LSTM 网络或者 Pointer 网络来建模动作序列。

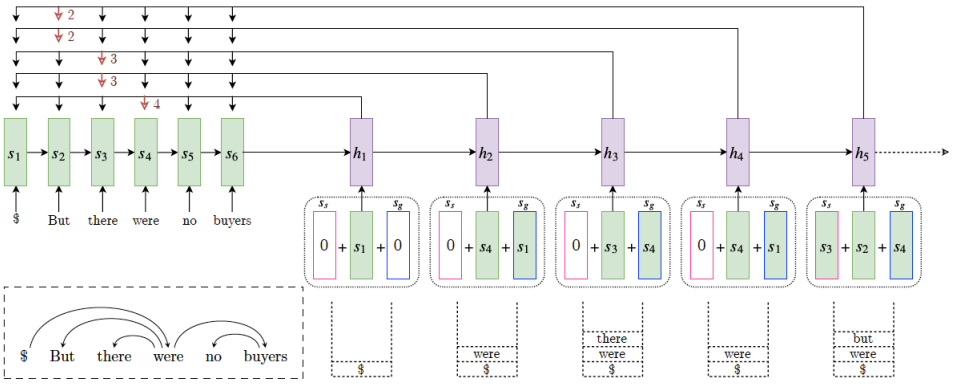


图 3-5 基于转移的算法 Stack-Pointer Networks for Dependency Parsing 网络结构图

3.2.3 应用案例

依存语法分析在信息检索，评价抽取和情感分析等 NLP 任务上都有很多应用。比如“谢霆锋的儿子是谁”和“谢霆锋是谁的儿子”两句话如果不使用依存语法分析，很有可能就返回了一样的结果。依存句法分析能够

更直接地通过语法结构的规则约束筛选出可能正确结果，提升相应任务的准确性。又比如“深圳大学非常漂亮，学生都很聪明”，这里“漂亮”形容的是“深圳大学”，“聪明”形容的是“学生”，通过依存句法分析，就可以抽取出对应的搭配。再比如“我家音响声音很大”和“我家洗衣机声音很大”，两者在情感上前者是正评价，后者是负评价，需要使用依存句法分析来识别“声音很大”的修饰对象。

常用的中文依存句法分析的工具有复旦大学 fnlp[1]，斯坦福大学 Stanford CoreNLP[2]，Hanlp[3]和哈工大 LTP[4]。

3.3 语义分析

在自然语言处理领域，语义分析涉及在某种程度上理解单词、短语、句子或文档的意义。传统狭义语义分析主要包括语义消歧（word sense disambiguation）[49] 和语义角色标注（semantic role labeling）[50, 51]。语义消歧指在给定文本上下文中确定多义词语的含义，例如，“他买了一台新苹果，用来修图更方便了”中，苹果一次指代苹果电脑，而非水果。语义角色标注是给词语和短语标注其在上下文文本中的含义的过程，典型标注标签包括主体、意图、结果等，如图3-6。语义角色标注是面向任务型对话系统中核心组件自然语言解析模块的基础技术之一。

随着深度神经网络在自然语言处理领域的应用和研究，基于神经网络的语义分析得到了越来越多的应用和发展，并成为了驱动神经机器翻译 [52, 53]、阅读理解 [54]、对话系统 [55, 56] 的最基础和核心的技术。基于神经网络的语义分析广义上可分为语义表示和语义匹配。其中语义表示任务将词语和短语镶嵌到高维向量空间中，称作词向量，作为 CNN[57]、RNN[58] 和 Transformer[59] 等模型的底层输入，根据任务需要，可以灵活的实现文本分类 [60, 61]、机器翻译、阅读理解、对话系统、文本摘要 [62, 63] 等自然语言处理任务，逐步替代或部分替代了传统以 ngram one-hot 向量作为底层表示的方法。语义匹配任务更多关注句子和篇章层级的语义的相似性，例如，寻找给定语料库中与查询语句语义层面最相

[1] https://github.com/FudanNLP/fnlp

[2] https://stanfordnlp.github.io/CoreNLP/

[3] https://github.com/hankcs/HanLP

[4] https://github.com/HIT-SCIR/ltp

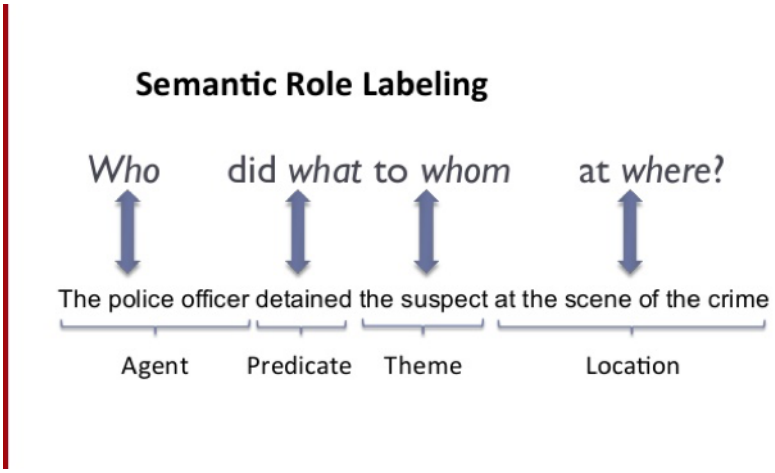


图 3-6 语义角色标注示例

似的句子。语义匹配任务在信息抽取、对话系统、问答系统都有广泛的应用。

3.3.1 语义表示

语义表示中词向量的概念可以追溯到 Bengio 的著作 [64]。作者在文中提出了一种基于神经网络的语言模型，创新的提出将每一个词表示成一个高维的向量，而后使用神经网络计算给定上文词语后，下一个词出现的概率分布，最大化训练语料出现的概率。训练语料是单词序列 $w_1,\cdots,w_t, w_t \in V$，其中 V 是全体单词的集合。语言模型的目标是找到一个模型 $f(w_t,\cdots,w_{t-n+1}) = P(w_t|w_1^{t-1})$，即在给定前 $t-1$ 个词的情况下，对第 t 个词的概率分布建模。Bengio 提出把词表示成实数向量，$C(i) \in \mathbb{R}^m$，即每一个词对应一个维度为 m 的实数向量，映射 C 可以用一个 $|V| \times m$ 的实数矩阵表示。在第二步中，作者提出使用神经网络 g 来建模给定前缀序列下一个词的概率。模型结构如图 3-7。其中，将词表示成高维实数向量的方法是词向量以及一系列神经网络在自然语言处理应用的开端。通常，我们会设计一个仅与文本结构或上下文相关的（无监督）任务作为目标，拟合训练数据，得到词向量表示。在 Bengio 之后，Word2Vec[65], Glove[66]，ELMo[37] 等一系列词向量方法出现并逐渐成为自然语言处理

的标准工具。语义表示不仅限于词的层面，广义的语义表示在现代基于神经网络的自然语言处理中应用广泛。例如，神经机器翻译通常采用编码器译码器（Encoder-Decoder）[67] 结构，其中编码器的输出可以看做是翻译模型对输入文本的语义表示，译码器基于该表示，生成目标语言的文本。再如，闲聊型对话系统中，需要针对聊天的上文给出合适的回答。闲聊对话系统可以分为生成式对话系统（Generation-based）[68] 和选择式对话系统（Selection-based）[56]。选择式对话系统依赖语义匹配方法。典型的生成式闲聊对话系统也采用编码器译码器结构，其中编码器的输出是对聊天上文的总结性表示，可以认为是广义上的聊天上文的语义表示，用于在译码器中生成合适的回答语句。此外，随着自然语言处理预训练模型的发展，Bert[39]，GPT-2[69]，MASS[70] 等预训练模型（Pretrained models）在阅读理解等任务上大放异彩。这些模型可以看做是多个联合任务共享底层语义表示的学习，通常使用与文本结构相关的目标作为训练目标，设计神经网络结构表示文本。在应用时，这些预训练获得的表示仅需在少量标注数据上进行优化，即可获得优秀的子任务模型。通常，编码器和预训练模型采用 CNN、RNN、Transformer 等结构处理文本序列，在 CNN 和 RNN 中通常还会使用注意力（Attention）机制，得到文本片段的语义表示。

3.3.2 语义匹配

语义匹配任务在语义表示的基础上，对文本片段之间的相似度或相关度给出量化指标，语义越相似的片段匹配分数越高。语义匹配可大致分为基于表示的匹配（representation-based matching）[71] 和基于交互的匹配（interaction-based matching）[72, 73]。基于表示的匹配方法注重对表示层的构建。常见的匹配度计算方式包括 cosine 函数，和神经网络匹配，如图 3-8。Cosine 函数直接计算待匹配对的语义表示向量间的 cosine 值，通常 cosine 值越大代表待匹配对的相似度或相关度越高。这种方法不需要额外的训练数据，实现简单并且高效，在工业中广泛应用。神经网络匹配方式使用一个额外的神经网络结构，将待匹配对的语义表示向量作为输入，计算语义匹配分数。这种方法更加灵活，可根据数据定制匹配结构。但需要额外的标注数据进行训练，才能得到可用的匹配模型。

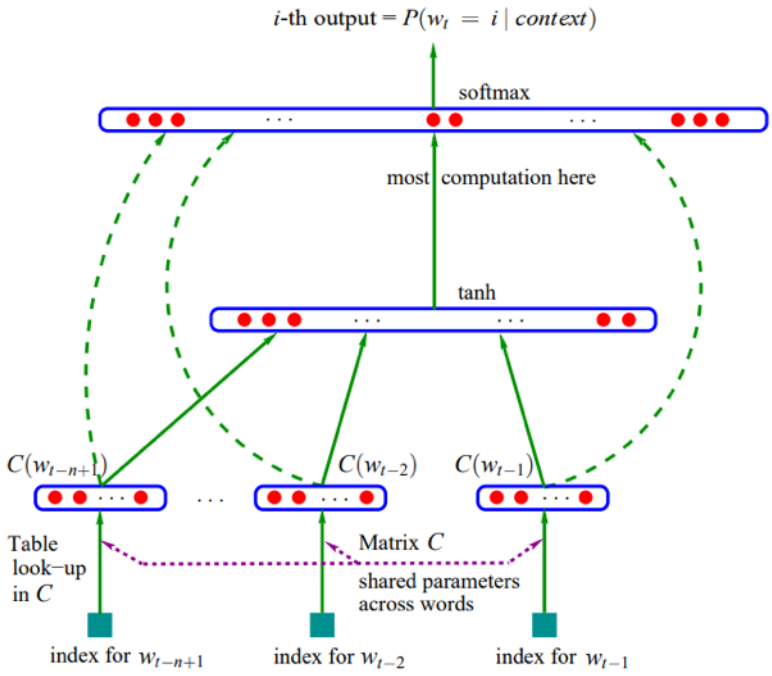


图 3-7 一种使用词向量的神经语言模型结构

图 3-8 基于表示的匹配

基于交互的匹配方法通常会保留待匹配对的序列信息，不会在表示层将文本转换成唯一的一个整体向量表示，而保留一个向量序列，用于接下来的交互匹配过程。例如将句子表示成一个与句子等长的向量序列。该向量序列可以使用语义表示的方法得到，例如使用 RNN 对文本序列进行建模。得到待匹配的序列对后，可以对序列中每个位置的向量计算与待匹配序列中向量的相关度。对每一个位置使用相同的方式计算相关度，可以得到一个匹配矩阵（matching matrix）。匹配矩阵包含了更细致的局

部文本交互信息，在交互矩阵的寄出上，我们可以构建神经网络结构计算最终匹配得分，去拟合目标得分。图 3-9展示了一个典型的基于交互的语义匹配方法在选择式对话系统中的结构。其中 Representation module 对应语义表示部分，matching block 对应交互匹配部分。

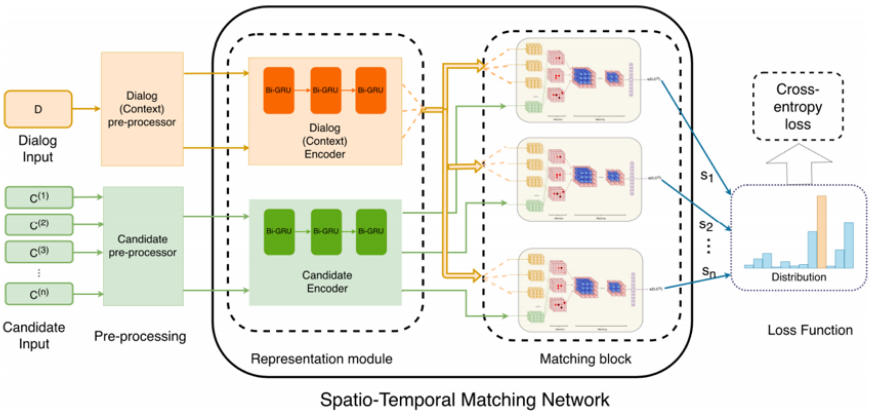


图 3-9 基于交互的匹配

Chapter 4
智能客服系统

4.1 简介

现在越来越多的交易和交流都可以通过网上平台完成，如果顾客在使用产品中碰到问题，他们可以在随时在网上提问，跟客服交流。传统公司会雇佣客服专员来回应客服的问题，然而提问高峰期的时候因为人太多，顾客不得不等待很长时间；深夜凌晨的时候客服下班，顾客的问题得不到解答。如果传统公司雇佣一大批客服专员 24 小时值班，那导致的人工成本会非常高昂。智能客服系统使用人工智能技术进行问题语义分析，答案查找，可以自动完成问答，能提高顾客的用户体验，大大降低公司所需客服专员数量。智能客服系统的速度很快，可以同时回答很多顾客的问题，而且可以 7*24 小时在线不会下班，价格也比人工客服更经济实惠。

4.2 问题形式化定义

此处客服系统假设客人和客服专员通过文字进行交流。智能客服系统会根据用户输入的问题和当前聊天语境，输出对应的文字解答。我们用 $\mathbf{X}$ 来代表用户的输入，用 $\mathbf{Y}$ 来代表系统的输出。客服系统可以表示为

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}).$$

根据客服系统的功能不同，一般分为检索型对话系统，任务型对话系统 [74] 和生成型对话系统 [67, 68, 55, 75]。工业应用中，一个完整的对话系统可能会同时包含以上三种对话系统。对于一个用户的提问，任务型

对话系统首先要判断用户是否要开始一个任务，如果是任务则交由任务式对话系统负责；然后检索性对话系统判断用户是否在询问某个知识点，如果是则回复对应的知识点；如果不是以上两种意图，则使用生成式对话系统来进行闲聊回复。

4.3 检索型对话系统

顾名思义，检索式模型就是查找问题库中跟当前问题最相似的问答对，输出答案。检索型对话系统适用于能一句话就能描述清楚的问题，而且答案比较固定不会经常变动的情况。这样的例子包括“你们公司的地址是哪里？”和“你们的咨询电话是多少？”这类简单问题。

检索模型可以形式化地表现为

$$\mathbf{Y} = \text{argmax}_{\mathbf{Y}} \mathbf{s}(\mathbf{X}, \mathbf{X_i}), \forall < \mathbf{X_i}, \mathbf{Y_i} > \in \mathscr{D}$$

. 检索式对话系统一般包括问题库 $\mathscr{D}$，问题匹配模型 $s(\mathbf{X_1}, \mathbf{X_2})$ 两个部分。问题库记录了常见问题和标准的回答，而匹配模型会计算用户的输入问句 $\mathbf{X}$ 跟问题库中的所有问题 $\mathscr{D} = \{< \mathbf{X_i}, \mathbf{Y_i} >\}_{\mathbf{i=0}}$ 之间的匹配度，系统输出最高的问题的标准答案。

检索式对话系统的核心在于问题匹配模型，就是如何计算用户输入问题跟标准问题库之间的相似度。

读者可能很容易就可以想到的一种方法就是对所有的输入问题，经过分词，提取 TF-IDF 特征，然后使用两个问题的特征向量来计算问题之间的相似度。由于自然语言中，问题本身的表述形式非常多样，难以穷尽。这种人工特征的性能有很大改进空间，经常对同一种意思的几种不同说法不是太鲁棒。

基于深度学习的句子语义匹配，可以在训练集模型上学习把意思相近，可是说法不一样的一对句子映射到相似的语义空间中去。常见的深度学习匹配模型包括基于语义表示的匹配模型，基于交互的语义表示模型。经过训练之后的深度语义匹配模型可以把测试集中相同意思可是不同说法的两个句子也表示成相似，增加了方法的泛化能力。基于深度学习的句子匹配模型复杂，一般来说需要大量的训练数据才能训练好。有一些工

作为了减少所需要的训练数据，他们提出了利用大数据上预训练的 NLP 模型来对深度语义模型进行初始化，取得了良好的效果。

检索性对话系统的缺点包括问题库更新困难，标准答案不能实时变动等。

4.4 多轮任务型对话系统

多轮任务型对话系统 [74]，又称填槽型对话系统，会通过提问，来引导用户提供完成当前任务所需的多个关键信息 “槽位”，进而帮用户完成复杂任务。多轮任务型对话系统适用于不能用一句话描述清楚的，需要多轮交流才能完成的问题，这样的例子包括购买机票，预定酒店等。

第 n 个用户输入可以表示为 $\mathbf{X_n}$, 第 n 个系统回复可以表示为 $\mathbf{Y_n}$，第 n 轮对话的对话记录可以标记为 $\mathbf{H_n} = \{\{\mathbf{X_i}, \mathbf{Y_i}\}_{\mathbf{i=1}}^{\mathbf{n-1}}, \mathbf{X_n}\}$。多轮对话系统可以被形式化地表示为

$$\mathbf{Y_n} = \mathbf{f}(\mathbf{H_n}).$$

多轮任务型对话系统跟单轮系统不一样的是，任务型系统需要考虑对话的状态。对于一个任务，设计者需要指定完成要获取哪些槽位信息 $\{s_i\}$，而在对话过程中，每一个信息的获取情况，就可以认为是对话的状态。例如在定机票任务中，需要获取的槽位包括 “起飞时间”,“起飞城市”,“到达城市” 等，而对话状态 $\tilde{\mathbf{H}}_n$ 就是这些槽位的获取状态，假设每个槽位都只有获取完成和未获取两种状态，那订机票任务的对话状态一共有 $2^3 = 8$ 个。

多轮任务型对话系统包括四个模块，自然语言理解模块，对话状态跟踪模块，对话策略模块和自然语言生成模块。口语理解模块自然语言理解模块会根据对话的每一轮输入 $\mathbf{X_n}$，从用户的问题中识别出用户的意图 a_n 和槽位值 $\mathbf{s_n} = \{\mathbf{s_i} = \mathbf{v_i}\}$。对话状态跟踪模块会根据当前的对话记录 $\mathbf{H_n}$，计算最新的对话状态 $\tilde{\mathbf{H}}_n$。对话策略模块会根据最新的对话状态 $\tilde{\mathbf{H}}_n$ 计算系统的回复动作 a'_n 和回复槽位值 $\mathbf{s'_n} = \{\mathbf{s_i} = \mathbf{v_i}\}$。自然语言生成模块会根据回复动作 a'_n 和回复槽位值 $\mathbf{s'_n} = \{\mathbf{s_i} = \mathbf{v_i}\}$ 生成回复的句子 $\mathbf{Y_n}$。

这里介绍一下四个模块的常见实现方法。其中自然语言理解模块可以分为口语预处理，语义切割，语义去噪，意图识别和词槽识别四个主要部分。用户在与智能客服系统，尤其是语音智能客服交流问题的时候，其

输入带有很强的口语化特征，例如在用户问句“你好我那个那个问个问题啊我的还款逾期逾期了要咋整哦”中，口语化特征包括招呼语“你好”，重复“那个那个”和“逾期逾期”，无关短语“问个问题啊”以及口语“要咋整哦”。口语化处理部分会先处理这些招呼语，重复和一些无用的语气词“哦”。使得整个句子更加干净。

口语输入还有个问题是用户会说很长一段话或敲很长一段文字，所以在初步口语处理过后需要有一个语义切割的模块来将句子切割成小片段。实践中一般会按照句号或者逗号两种尺度来切割句子。句号分割的会有完整的转折或者因果关系，比如“我的还款逾期了要咋整”。逗号分隔的会把语言切割成比较小的粒度，比如“问个问题”，“我的还款逾期了”和“要咋整”，小的粒度比较好理解。两种尺度结合起来实现更全面的语言理解。

接下来语义去噪部分会把“问个问题”这样与任务无关的短语识别出来，词槽识别会把“我的还款逾期了”识别成“还款状态：逾期”，意图识别会把“要咋整”识别成意图“如何操作”。意图识别部分和语义去噪部分都可以使用文本分类模型，常见的文本分类深度学习模型有 TextCNN[76]，RCNN[77] 和 HAN[78] 等。词槽识别和语义切割都可以使用序列标注模型，现在比较流行是 BERT+LSTM+CRF。

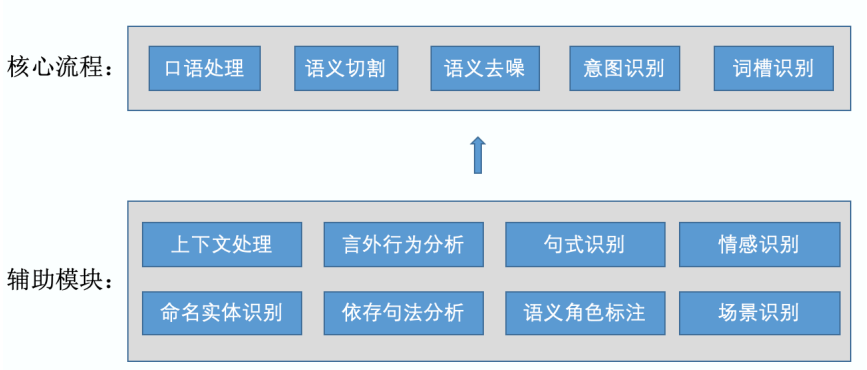


图 4-1 自然语言理解模块主要组成部分

自然语言理解部分还有一些比较重要的模块比如句式识别，用户言外行为识别，上下文处理模块，场景识别模块，情感识别模块，依存语法分析模块和语义角色识别模块。其中句式识别主要用来识别是陈述句还是疑问句，疑问句会被优先输入意图识别模块，陈述句会被优先输入词

槽抽取模块。用户言外行为识别可以用来识别用户是否在发出请求或者表达感谢，然后系统据此作出相应的回应。比如用户说”我要还款“，该句为陈述句，可能会被意图识别模块漏掉，但是一旦言外行为识别模型识别出这是用户请求，意图识别模块就会介入处理。

上下文处理模块在识别出当前问句语义成分不全的情况下用来改写当前问句得到完整的问句，然后再输入意图识别模块识别意图。场景识别模块用来识别当前聊天所处场景，用来辅助意图识别。情感识别可以有效的识别出用户当前的情绪，如果用户急躁或者抱怨的时候及时转人工处理以免影响用户体验。依存语法分析模块和语义角色识别模块会把用户问句转换成比较规范的中间语义表示格式，比如主谓宾，可以给包括意图识别在内的其他模块提供辅助。

对话状态跟踪模块 [79, 80, 81]，常见做法是把任务所需的所有槽位指定一个顺序，然后对每个槽位有一个状态变量标志着这个槽位是否已经被获取。另外一种做法是把对话状态定义为高维向量，然后联合训练口语理解模块和对话跟踪模块，使得联合模块读入 $\mathbf{H_n}$ 后直接预估 $\tilde{\mathbf{H}}_n$。这种方法的好处是可以把两个模块进行联合训练，不需要人工设计规则。坏处是输出的高维向量 $\tilde{\mathbf{H}}_n$ 人工难以解释，训练需要大量的标注数据，而且在标注数据缺乏的情况下不鲁棒。

对话策略的实现方法分为基于规则的策略和基于强化学习 [82] 的对话策略。一个最简单的基于规则的策略就是，对于对话状态中还未获取到的槽位，按顺序对用户进行提问，并且期望用户进行回答。基于强化学习的对话策略把策略进行数学建模，策略以完成某种对话任务为目标，例如成功引导客人预订了酒店，或者成功销售某个商品。对话策略的目标由回报函数指定，如果策略成功引导对话状态完成了某种任务，那么这个策略会得到正的回报分数。根据对策略的建模方法的不一样，强化学习分为基于价值函数的框架，基于动作函数的框架，和 actor-critic 框架。其中基于价值函数的强化学习框架中的 Q-learning 比较常用。具体地，Q-learning 的策略 [83] 可以表示为

$$a'_n, \mathbf{s'_n} = \text{argmax}_{\mathbf{a'},\mathbf{s'}} \mathbf{Q}(\tilde{\mathbf{H}}_\mathbf{n}, \mathbf{a'}, \mathbf{s'}),$$

其中 $Q(\tilde{\mathbf{H}}_n, a', \mathbf{s'})$ 代表在对话状态 $\tilde{\mathbf{H}}_n$ 下，系统回复 $a'_n, \mathbf{s'_n}$ 的情况下，整个对话的预估回报分数。$Q()$ 的训练方法可以表示为以下公式

$$Q(\tilde{\mathbf{H}}_n, a'_n, \mathbf{s'_n}) = \mathbf{Q}(\tilde{\mathbf{H}}_\mathbf{n}, \mathbf{a'_n}, \mathbf{s'_n}) + \alpha(\mathbf{r_n} + \gamma \text{max}_{\mathbf{a'},\mathbf{s'}}\{\mathbf{Q}(\tilde{\mathbf{H}}_{\mathbf{n+1}}, \mathbf{a'}, \mathbf{s'})\} - \mathbf{Q}(\tilde{\mathbf{H}}_\mathbf{n}, \mathbf{a'_n}, \mathbf{s'_n})),$$

其中 r_n 是当前训练数据第 n 步获得的即时回报分数，而 $\max_{a',s'}\{Q(\tilde{\mathbf{H}}_{n+1},a',\mathbf{s}')\}$ 代表当前训练数据第 $n+1$ 步以后可能获得的最高回报分数，$r_n+\gamma\max_{a',s'}\{Q(\tilde{\mathbf{H}}_{n+1},a',\mathbf{s}')\}$ 可以看做 $Q(\tilde{\mathbf{H}}_n,a'_n,\mathbf{s}'_\mathbf{n})$ 的更新目标。γ 是一个未来回报分数的衰减记分参数，α 是学习率。$Q()$ 可以根据情况选很多不同的模型，例如线性回归，深度神经网络等。强化学习的方法可以应用于难以设计对话规则的复杂场景，可是需要大量的训练数据。

自然语言生成模块有基于模板的方法和基于生成模型的方法。基于模板的回复生成方法，就是根据对话策略的输出 $a'_n,\mathbf{s}'_\mathbf{n}$，选择合适的回复模板，然后把 $\mathbf{s}'_\mathbf{n}$ 填在回复模板中的对应位置，这种方法不需要训练，实现简单。基于生成模型 [84, 85] 的方法则会根据 $a'_n,\mathbf{s}'_\mathbf{n}$ 一个词接一个词地生成回复，直到生成完整句回复。大部分基于生成模型的方法都是基于递归神经网络 RNN 以及其变种，其特点在于比较灵活，不过需要大量的训练数据进行训练，而且生成的模型难以被人理解。

4.5 生成型对话系统

生成型对话系统 [67, 68, 55, 75]，会根据用户的提问一个词接一个词地生成合适的回复。生成型系统最灵活，多用于对用户的问题进行闲聊式回复。

生成型对话系统中，生成一句回复 $\mathbf{Y}=\{\mathbf{y_1},\mathbf{y_2},...,\mathbf{y_m}\}$ 的过程，可以被定义为，递归地根据用户问题 $\mathbf{X}$ 和当前已经生成的 $\mathbf{Y}$ 的前半句回复 $\{y_1,y_2,...y_{i-1}\}$，生成第 i 个单词 y_i 的过程，即

$$y_i=\text{argmax}_{y'}P(y'|\mathbf{X},\{\mathbf{y_1},\mathbf{y_2},...\mathbf{y_{i-1}}\}),$$

其中 y_i 是回复 $\mathbf{Y}$ 中的第 i 个字符。

目前比较常见的方法是编码器解码器模型，模型的输入是一个问题的单词序列，模型的输出是回复的单词序列。模型首先会用编码器逐个字逐个字地读取用户问题 $\mathbf{X}$，把用户的问题压缩成一个隐向量 $\mathbf{h^c}$；然后模型用解码器根据 $\mathbf{h^c}$ 逐个字逐个字生成整个回复 $\mathbf{Y}$。其中编码器的在第 i 步的隐变量 $\mathbf{h^e_{i-1}}$ 为

$$\mathbf{h^e_i}=\mathbf{h^e_{i-1}W^e}+\mathbf{x_iU^e}$$

其中 $\mathbf{x_i}$ 是用户问题的第 i 个单词的词向量，$\mathbf{h^e_{i-1}}$ 是第 $i-1$ 步的隐变量，$\mathbf{W^e}$ 和 $\mathbf{U^e}$ 是参数矩阵。用户的问题压缩的隐向量 $\mathbf{h^c}$ 等于编码器处理完用户问题的最后一个单词之后的隐变量 $\mathbf{h_m}$。解码器的初始化隐变量为 $\mathbf{h^d_0}$ 等于 $\mathbf{h^c}$，编码器在第 i 步的隐变量 $\mathbf{h_{i-1}}$ 为

$$\mathbf{h^d_i} = \mathbf{h^d_{i-1}}\mathbf{W^d} + \mathbf{y_{i-1}}\mathbf{U^d}$$

其中 $\mathbf{y_{i-1}}$ 是生成的用户回复的第 $i-1$ 个单词的词向量，$\mathbf{h^d_{i-1}}$ 是第 $i-1$ 步的隐变量，$\mathbf{W^d}$ 和 $\mathbf{U^d}$ 是参数矩阵。根据 $\mathbf{h^d_i}$，我们可以解码出对应的单词 y_i。除了基本的编码器解码器模型外，很多新的方法使用了记忆模块，注意力模块等技术来改进生成型模型。

生成型对话很灵活，不过由于模型参数多，一般需要比较多的训练数据来进行训练。后面有一些使用预训练模型来减少模型所需训练数据的工作。除此之外，生成型对话系统容易产生对所有问题都合适，可是不包含任何信息的 “安全回复”，例如 “我不知道” 等。

Chapter 5
对话理解与智能质检

本节首先给出对话理解任务的定义，然后介绍对话理解的主要方法。接下来以智能质检为例，讲述对话理解是怎么落地和实现的。

5.1 对话理解

5.1.1 什么是对话理解

对话理解是指希望计算机跟人一样，具备自然语言理解的能力，从对话内容中挖掘对话意图，理解对话意图，用户情绪识别等。例如，在客服与用户交互的对话中，用户询问“今天的天气如何”，这里就是一个“询问天气”的意图。在实际对话中，由于自然语言的多义同义问题，语言的词序问题，我们不能只停留在字面理解层面，更需要语义层面的理解。

这里对话可以包含语音对话和文本对话，如果是语音对话，我们一般可以利用语音自动识别技术将语音转为文本。后续我们要讨论的内容是文本的对话理解。

5.1.2 技术路线简介

一般而言，文本的对话理解从技术角度上可以分为两类：文本匹配和文本分类。

文本匹配 文本匹配的目标是得到 $f(text_1, text_2)$ 的语义匹配得分，其中 $text_1$ 和 $text_2$ 是输入的文本，f 是文本匹配模型。传统文本匹配技术主要按照字面上词汇重合度来进行文本匹配，比如传统信息检索中的向量空间模型 VSM、BM25 等算法。这种基于词汇重合度的匹配算法存在一定的局限性。例如“出租车”和“的士”语义是一致的，但是字面上却完全不匹配。近年来，随着深度学习技术的发展，我们通过多层神经网络对文本语义进行建模，在语义匹配效果上有了很大的提升。

深度学习的模型主要包含 Representation-based Match 和 Interaction-based Match 两种，如下图所示。

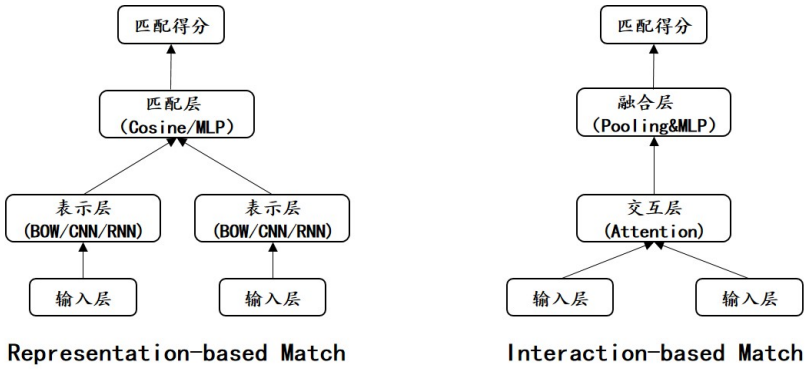


图 5-1 Representation-based Match 和 Interaction-based Match 框架图

输入层一般是将词通过 look-up table 变成词向量，常见的方法有 word2vec, 其后陆续出现了一些文本向量化方法, 如 GloVe, FastText, EMLo 等，不过 word2vec 几乎还是最适用的方法之一。词到向量的转化使得词与词之间的语义相似性可以通过向量相似度的方法来度量。为了减少数据稀疏，我们可以加入 subword information，即词内的结构信息，中文里面就是字的信息，英文就是字母的信息。

Representation-based Match 的核心思想是先将句子转为向量，然后再匹配。怎么获得句子向量呢？最简单的方式有 Bag of Words（BOW），句子中词向量求和作为句子向量表示，向量求和可以将不定长的句子转化为定长的向量，但 BOW 的方式忽略了句子中词序对语义带来的影响。当然，句子向量更高级的获取方式可以使用卷积神经网络（CNN）或者循环神经网络（RNN）来得到。有了句子向量，我们通过匹配层去计算两个

句子向量的匹配得分，这里匹配层可以是用余弦距离（Cosine Distance）或者把两个句子向量拼接（concatenation）再经过一个多层感知机（MLP）来实现。经典的方法有 Siamese CNN，Siamese LSTM，DSSM，ARC-I 等。Representation-based Match 的优势是可以离线计算好句子向量的表示，易于建立索引，执行效率高，非常适合文本匹配的粗召回。但其存在的问题是容易丢失语义焦点，词的上下文重要性较难捕捉。

Interaction-based Match 则是构造了两个句子之间的语义单元（e.g., term，n-gram，part-of-speech）的交互矩阵，然后再经过一个融合层将细粒度的语义匹配信息合并，较好的把握句子的语义焦点和保留重要的句子间相似信息。主流的方法包含 ARC-II，DeepMatch 和 MatchPyramid 等。Interaction-based Match 不能离线预处理，需要在线匹配，适合精排序。

文本分类 文本分类的目标则是得到 $g(text)$ 的语义标签，其中 $text$ 是输入的文本，g 是文本分类模型。文本分类问题需要数据标注其语义标签，在任务型对话中，这个语义标签就是意图。文本分类模型训练好后，我们就可以对新数据进行分类了。具体内容可以参考对话系统的章节。

5.2 应用案例：智能质检

对话理解在智能客服，智能质检有着广泛的应用。下面以智能质检为例，阐述对话理解相关技术是怎么应用的。

5.2.1 什么是智能质检

智能质检使用人工智能算法，分析坐席呼叫场景下人工客服与客户的对话，实现全量质量检查，提高人工客服的服务质量和客户的满意度。智能质检系统的输入是一通人工客服和客户对话的录音，输出是质检报表，显示该录音在不同质检项的合格情况。质检项的重要性通过质检项的分数来决定。智能质检无需人工介入，节省质检人力，覆盖率高（100%），提升质检效率，降低漏检错检率。

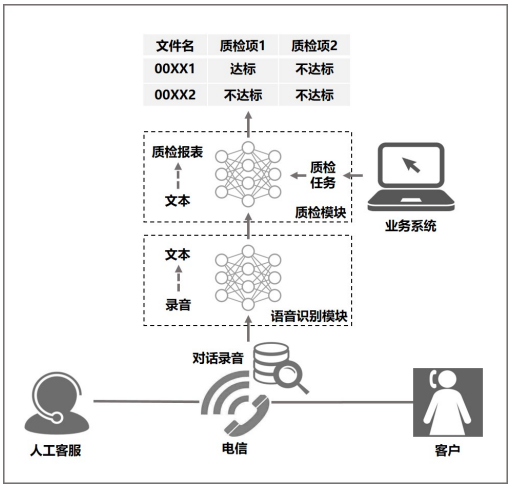


图 5-2 智能质检基本流程图

从智能质检的基本流程图中我们可以发现，对话录音经过语音识别模块之后，我们得到了客户和人工客服之间的对话文本。质检员配置了质检项之后，我们将对话文本输入质检模型，最后得到了质检报表。

5.2.2 实现方案与应用状况

实现方案 假设我们定义了质检项要求客服在对话中“核实用户的工作地址”，比如“你的公司地址在哪里”。一种简单的实现方案是通过规则配置“算子 + 逻辑操作符”或者正则表达式，如果录音文本中满足匹配条件，则命中该质检项。我们可以配置“(公司 | 工作)& 地址 &(哪 | 地方)”这样识别出“你的公司地址在哪里”和“您的工作地址在什么地方”这两种说法。这里“公司”就是一个关键词算子，“&”是一个“与”的逻辑操作符，“|”是一个“或”的逻辑操作符。这种本质上还是一种基于词汇重合度的匹配算法，适合冷启动，在没有标注数据的情况下就能构建一个简单的质检系统。但它也存在诸多弊端：第一、配置规则需要一定的专业知识，有培训成本；第二、当质检项和表达方式增多时，不同质检项容易出现规则冲突，维护成本高；第三，对于一些比较复杂的质检项，很难通过配置规则进行质检，容易出现漏检和误检的问题。

如果我们有标注数据，就可以使用文本匹配和文本分类的方法。计算 f(录音文本, 质检例句) 的语义匹配得分或者 g(录音文本) 的质检项标签。为了减少对数据的依赖和利用大量的无标签数据，一种有效的做法是在 BERT（Bidirectional Encoder Representation from Transformers）模型上去 fine-tune 分类器，BERT 用于上层模型的特征提取，作为上层模型的输入，它能够较好的捕捉对话片段的高层语义，兼容少量的漏字情况。我们通过数据驱动的方法让模型越来越聪明，业务方只需要提供标注数据就能进行质检，不需要人工定义规则，模型具有一定的泛化能力。但遇到 bad case 没有基于规则的方法容易修复，另外需要标注数据积累到一定规模才能发挥模型的优势。

应用状况 当前智能质检的应用可以包含离线质检和坐席实时质检。离线质检是指结合语音识别和自然语言处理技术，对海量录音数据进行智能化分析。离线质检可对全量录音质检，质检过程无需人工介入，可以提供内容质检，敏感词识别，语速分析等质检结果。坐席实时质检是指在人工客服和客户对话过程中，提供实时质检能力，辅助人工客服判断客户情绪和实时分析对话过程的信息，及时提醒人工客服从而使客户获得更好的服务。现在智能质检的产品形态包含 SaaS 云服务和私有化部署。SaaS 级产品部署，让中小企业也能够享受智能质检带来的高效与便捷，克服了采购费用高部署周期长的问题。

未来随着多方业务的使用，可以基于联邦学习进行智能质检，在满足数据安全和私隐保护的前提下，通过模型的参数梯度共享，获得了把所有数据放在一起训练的效果，使得不同的业务方合作共赢，建立更准确的数据模型。

Chapter 6
自然语言处理技术展望

随着算法的持续改进以及数据和算力的持续增长，自然语言处理技术在最近十年间取得了长足的进步，基于深度神经网络的模型在多个自然语言处理任务中刷新了历史记录。这些新兴的自然语言处理技术在改善了传统自然语言处理任务的效果的同时，也极大地促进了一些相关领域的发展。例如，在信息检索领域，自然语言理解技术能够帮助搜索引擎更准确地识别用户的潜在需求；文本摘要技术能够帮助搜索引擎更好地刻画文本中的重要信息；自然语言对话技术有极大的潜力带来搜索引擎和用户交互方式的深刻变革。在语音识别领域，最新的语言模型技术 (例如 BERT、XLNet 等 [39, 86, 87]) 能够帮助语音识别系统更好地控制识别结果在语法和语义层面的合理性；通过自然语言理解技术和语音识别系统的无缝结合，智能音箱等设备正在逐步走进人们的日常生活，也重构了 AIOT 行业的生态版图。

然而，当前的自然语言处理技术还面临着诸多挑战:

- 基于深度神经网络的自然语言处理模型往往是 Model-blind 的，其效果严重依赖大量的数据和超强的算力，这通常会导致极高的模型成本，也极大地限制了自然语言处理技术的公平性和普惠性。因此，以贝叶斯程序学习 (Bayesian Program Learning) [88] 为代表的 Model-based 小样本学习正在逐步引起研究人员的关注。通过引入领域相关的先验知识，贝叶斯程序学习等技术可以有效降低模型训练中对于大量训练数据的依赖，节省数据标注成本和算力开销。
- 当前的自然语言处理技术通常只能发现数据之间的关联 (Association) 而不能发现数据之间的因果关系 (Causality)。例如，当前的对话系统

技术虽然可以生成一些似是而非的语句，但是从多轮的语境来看，这些机器生成的对话往往存在非常严重的逻辑问题。导致这种现象的重要原因是神经网络模型通常只是对训练数据做曲线拟合，缺乏对数据背后的复杂关系的深刻理解。最近日益受到科研人员重视的因果推理 (Causal Inference) [89] 技术有助于人们对数据进行因果关系分析，该技术和传统自然语言处理技术的深度结合有潜力使自然语言处理在认知层面更上一层楼。

- 随着诸多类似欧盟 General Data Protection Regulation （GDPR）法规的陆续出台，数据监管正在变得日益严格，单一研发机构获取大量的训练语料变得越发困难。如何在合规前提下统筹利用行业内的数据助力自然语言处理的研发任务是当前需要研究的重要课题。以联邦学习 (Federated Learning) [90, 12] 为代表的技术可以在保护各个研发机构数据隐私的前提下，提供一种挖掘跨机构数据背后的价值的计算范式。通过应用联邦学习这种计算范式，整个行业可以有效解决数据孤岛问题，打造更强大的自然语言处理模型，实现一个互利共赢的业界生态。

综上所述，自然语言处理技术的发展同时面临着机遇与挑战。我们在体验着近年来自然语言处理的飞速发展带来的技术红利的同时，也期待在可见的将来该领域能在上述三个方面取得可喜的进展，从而带来自然语言处理相关行业的又一次腾飞。

Chapter 7
人脸识别

7.1 问题定义

人脸识别，是指对输入的图像和视频，检测其中存在的人脸，依据人脸的面部特征，完成身份识别的过程，属于生物特征识别技术。整个流程包含人脸检测、人脸对齐、人脸特征提取、人脸匹配几个阶段，如图 7-1 所示。目前人脸识别已经广泛应用于安防、金融、军事等领域。

人脸识别具有以下优点：

自然性：所谓自然性，是指人脸识别技术所利用的生物特征，与人类进行人脸识别时所利用的生物特征是一致的，与之相比，虹膜识别、指纹识别等技术，则不具备自然性。

非接触性：在人脸识别技术中，用户不会与识别设备发生任何接触，对于用户来说体验较好。而指纹识别则需要用户进行按压设备。

使用便捷：用户使用人脸识别技术时非常方便，基本上无需做特殊的配合。

人脸识别也具有一些缺点，比如易受光照条件的影响，易受人脸遮挡物的影响，跨年龄识别难度较高等。但总的来说，人脸识别是目前一种可靠的，实用的，便捷的身份核验技术。

图 7-1 人脸识别流程图

7.2 实现方案

7.2.1 人脸检测及对齐

人脸检测是人脸识别的第一步，属于目标检测的子方向。其目的是找出图像中的人脸以及对应的位置。可能还会包含一些人脸的额外信息，比如人脸的关键点，姿态角度等。

典型的人脸检测是基于以下的流程：由于人脸可能出现的图像的任何位置，因此需要通过滑动窗口（sliding windows）来获取可能包含人脸的子图像。获取到的子图像，需要通过一个二分类的分类器，来判断图像中是否包含人脸，如果还需要确定人脸的精确位置，还需要加上一个回归人脸框的操作。同一个人脸可能会检测出多个人脸框，因此需要使用非极大值抑制 (Non-Maximum suppression, NMS) 来进行合并去重。接下来本文介绍一些具有代表性的人脸检测方法。

Viola-jones[91] 使用 Haar-like 小波特征，并通过级联的 AdaBoost 分类器构造检测器。该方法具有检测效率高，并且能够保持较好的精度的特点，是第一个具有实用意义的人脸检测算法。MTCNN[92] 将人脸分类、人脸框回归以及人脸关键点定位在同一个任务内完成了，是一个多任务 (multi-task) 的检测方法，这种思路在后续的很多方法里也得到了使用。

anchor 的思想在目标检测方法 Faster-rcnn[93] 中首先被提出，在人脸检测中也经常使用到。anchor 提出的目的是为了解决目标在图像中可能以不同的形状存在，比如不同的长宽比，所以加入人工的先验信息，预先定义不同比例的 anchor 来进行候选目标框的获取。Face r-cnn[94], Pyramidbox[95], Retinaface[96] 这些方法，都用到了 anchor 的思想。另一个在人脸检测中经常使用的思想是特征金字塔网络 (feature pyramid network)，为了自适应不同尺度人脸的检测，一般有两种做法，一种做法是图像金字塔，这种方法需要对输入图像做不同尺度的缩放，缺点是耗时较高；另一种更好的做法则是特征金字塔，其思想是在不同分辨率的特征图 (feature map) 上检测对应尺度的目标，同时将不同分辨率的特征图与更高层的特征图进行特征融合，保证每一层的特征图都具有足够的表达能力。Pyramidbox,Retinaface 都用到了特征金字塔，SSH[97] 虽然没有直接用到特征金字塔，但其也是对网络 3 个不同尺度的特征图进行分别预测，来解决多尺度的人脸检测问题。

做完人脸检测后，一般需要进行人脸对齐。通过对人脸进行关键点定位，以及预先定义好的关键点模板，进行仿射变换，通过旋转、平移、缩放等操作，进行人脸对齐，对齐后的人脸能够更好的进行人脸特征提取。目前常见的关键点个数，有 5 个关键点、68 个关键点、90 个关键点以及 106 个关键点等。

7.2.2 人脸特征提取

特征提取是人脸识别的关键步骤，它将人脸图像映射都某个特征空间中，使得映射后的特征能够很好地区分不同人之间的差异点。经过特征提取得到人脸的特征表示之后，可以进行特征匹配。如果是对两个特征进行比对，我们一般称为人脸比对或者人脸验证 (verification)，如果是将一个特征与一组特征进行匹配，我们一般称为人脸检索或者人脸识别 (identification)，如图 7-2 所示。

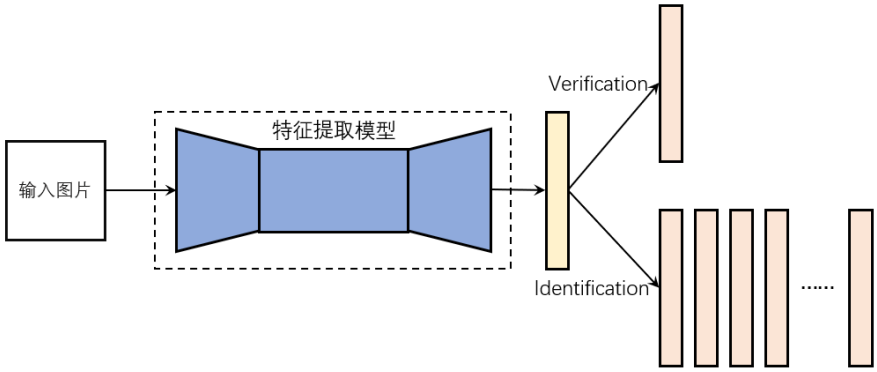


图 7-2 人脸特征抽取

传统的特征提取算法，通过一些降维方法，得到一系列降维后的特征，用来表示人脸。比如使用 PCA 进行降维的 EigenFace，基于 LDA 进行降维的 FisherFace 等，都是早期人脸识别中非常经典的算法。但这些方法存在一些缺点，对光照、表情、姿态敏感，泛化能力不足，因此在实际使用中的准确度不高。

随着深度学习的广泛应用，越来越多具有实用价值的方法被提出，人脸识别的研究得到了极大的发展。基于深度学习的特征提取方法可以分为两大类：

度量学习 (metric learning)

通过一个度量函数，来衡量相同人或者不同人的特征表示之间的距离，从而学习到每个人每个照片的特征表示，基本思路是同一个人的特征表示之间的距离尽可能小，不同人的特征表示的距离尽可能大。这一个方向的典型方法包含 2014 年的 Contrastive loss[98] 和 2015 年 google 提出的 Triplet loss[99]。

DeepID2 是基于 Contrastive loss 的模型，它在训练的时候，同时训练 classification 和 verification 两个信号，其中的 verification 信号，就是用特征表示之间的 Contrastive loss 来构造的。Contrastive loss 是基于 pairwise 的思想，模型训练时，需要输入两张图片，如果两个图片是同一个人，则 verification 的标签为 1，如果不是同一个人，则标签为 0。

google 于 2015 年提出的 Facenet 中，则用到了 Triplet loss。其思想是以三元组的形式来训练模型，每次输入需要三张图片，其中两张图片是同一个人，以及一张其他人的图片，要求同一个人的照片之间的距离要小于不同人之间的距离，且要超过一个 margin。

基于 margin 的分类方法 (margin based classification)

第二类思想是基于分类的思想来进行特征提取，根据训练集中的数据，同一个人的照片属于同一类，训练集一共包含多少个 id，则总共需要分多少类。由于用的分类的思想，所以自然而然可以使用分类的损失函数。而在此基础上，又提出了一系列的方法，用以最小化类内间距或者最大化类间间距。比较有代表性的方法有 Center loss[100]，SphereFace[101]，CosFace[102] 以及 ArcFace[103] 等。

由于是多分类任务，所以最基本的损失函数形式是 softmax loss。但是直接用 softmax loss 训练出来的特征，往往效果不理想，某些类别的类内间距甚至比类间间距大，导致人脸识别的时候出现错误。Center loss

引入了类内中心，为每个类别提供一个类内中心，最小化训练集中每个样本与其类内中心的距离，从而达到减少类内间距的效果。

基于 SphereFace 的训练方式，是在此基础上做了改进，对权重进行了归一化，且增加了角度裕量，在 cos 函数上对角度乘上因子 m，加大分类难度。CosFace 和 ArcFace 更进一步，对特征表示也做了归一化，并分别引入了不同的 margin 形式，取得了更好的效果。

以上方法都是通过不同方式去减少同一个人的类内间距以及增大不同人之间的类间间距。但除了损失函数以及网络的设计之外，更为重要的是训练的数据的分布，比较好的训练数据是同一个人包含多张不同的照片，这些照片覆盖此人不同年龄段，不同姿态角度，不同遮挡程度，不同妆容情况等，这样的数据能够学习到鲁棒性更强，通用性更好的模型。

7.3 应用案例

人脸识别目前在安防、金融等领域都得到了广泛应用，下面介绍一些常见的应用案例。

门禁闸机

这是人脸识别的典型应用场景，属于人脸检索（1：N）的应用。门禁闸机在初始化的时候，会要求录入一个人脸库，该人脸库经过特征提取后，作为识别的底库。当有人通过闸机的时候，会拍摄来人的照片，通过特征提取转化为特征表示之后，与底库中的特征集合进行对比，找出该人员是否存在于底库中。

金融核身业务

目前几乎所有的金融核身业务都支持人脸核身功能，属于人脸比对（1:1）的应用。当用户的办理某些业务的时候，会被要求进行人脸核身，系统会通过摄像头采集用户的照片，与用户留底的另一张照片进行比对，以确定用户是否为本人。这种方式大大减少了金融业务中进行业务审核的人员数量及审核时间，节省了用户时间，提升了用户体验。

Chapter 8
声纹识别

本章简单介绍声纹识别的概念，并其常见的算法，以及在实际中的应用。

8.1 问题定义

8.1.1 基本定义

声纹识别（Voice Print Recognition），也称作说话人识别（Speaker Recognition），是一种生物识别技术，能够根据说话人的声音特征提供精准、高效、便捷的身份识别服务。从感官直觉上来说，声纹虽然不像人脸、指纹的个体差异那样直观可见，但由于人在讲话时使用的发声器官–舌、牙齿、喉头、肺、鼻腔在尺寸和形态方面每个人的差异很大，因此反映到任何两个人的声纹图谱都存在有差异。最直观的感受是当我们打电话给认识的人的时候，通过很短一句话甚至一声“喂？”，就能准确地分辨出接电话的是谁。这种语音中承载的说话人身份信息的唯一性使得声纹也可以像人脸、指纹那样作为生物信息识别技术的生力军，可广泛应用于金融安全，公共安防，智能家居等领域。

8.1.2 分类

声纹识别通常分为两大类，说话人确认和说话人辨别，也就是常说的声纹 1：1 识别和声纹 1：N 识别。声纹 1：1 识别是指确认某段语音是否是指定的某个人所说的，而声纹 1：N 识别是判断某段语音是若干人中的哪一个所说的。不同的任务和应用会使用不同的声纹识别技术，如诈骗电话需要缩小人员范围时可能需要声纹 1：N 技术进行辨别，而银行金融交易时则需要声纹 1：1 识别技术进行确认。

从另一方面，声纹识别也分为文本相关（Text-Dependent）和文本无关（Text-Independent）两类。与文本相关的声纹识别要求用户按规定内容进行朗读，这样能更加准确的建立模型，在识别的时候也要求用户按规定内容朗读。文本无关的声纹识别并不要求用户根据指定文本进行朗读，这样也能建立模型，验证的时候同样不需要用户根据指定文本进行朗读。一般来说，文本相关的声纹识别效果会更好，安全性更高，但是用户体验较差和使用场景就相对较窄，通常用于安全要求比较高的场景，如金融核实身份。文本无关的声纹识别会和文本依赖比较弱，因此能进行跨语种使用，就算我们没有别的语种的语料，也是能应用到那语种上面去。文本无关的使用场景很宽广，在金融客服上就能建立声纹库，后面的用户在金融上快速验证和关联。

8.1.3 挑战和机遇

声纹识别技术的应用会有很多挑战，比如同一个人的声音具有易变性，可以控制不同部分造成不同发音；同时会受到身体状况、年龄、情绪等的影响，对识别性能有影响；又比如不同的麦克风、信道、环境噪音等对识别性能有比较大的干扰；又比如混合说话人的情形下人的声纹通常不易分离和识别等等。当然这些挑战也会被利用转化为优势，比如声音会受身体状况，年龄，情绪的影响，因此会单独训练出利用声纹识别情绪，或者利用声纹识别是否生病目前有应用到养猪场识别病猪等等。

尽管如此，与其他生物特征相比，声纹识别的应用有一些特殊的优势：

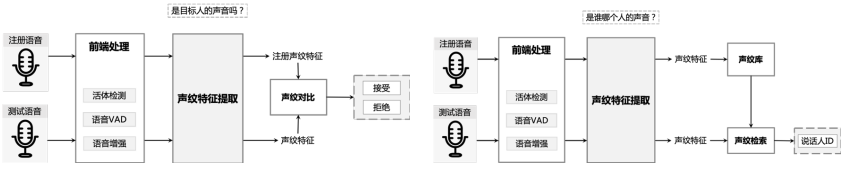
- 获取声纹数据成本较低

- 用户使用接受程度较高
- 使用成本较低，适合远程身份确认
- 随着文本无关技术提升，能在小语种、跨种族等方向都有应用场景

这些优势受到系统开发者和用户青睐，从而使得声纹识别的应用越来越受到欢迎。

8.2 实现方案

声纹识别中通常的流程如图 8-1所示。不管是 1：1 还是 1：N 大致都是分为三部分：前置处理，特征抽取和声纹匹配。前置处理通常是 VAD 检测、反欺诈活体检测、声音增强等。VAD 检测处理是把有声音部分和静音部分区分出来，把有声音部分送到后续处理中去。反欺诈活体检测是主要应对声纹识别是否被攻击，提高安全性。这些前置处理不在这里详细展开细说，匹配算法我们可以用最普通的计算向量距离如欧式距离或 cos 值，也可以用深度学习网络进行计算相似度。接下来我们重点介绍声纹的特征抽取。



(a) 声纹 1：1 识别　　(b) 声纹 1：N 识别

图 8-1 声纹识别流程

8.2.1 特征抽取

声纹识别的特征抽取大致经历了三代算法，GMM 模型到 ivector 最后到深度学习网络。我们重点介绍前面两代，深度学习网络的方法和人脸特征抽取方法类似，也是分为两大类：度量学习 (Metric Learning) 和

基于 margin 的分类方法 (Margin Based Classification)，将会在人脸识别中重点介绍。

8.2.1.1 GMM-UBM 模型

GMM 模型 [104] 即高斯混合模型，是由多个高斯函数进行加权求和进行拟合复杂的函数，如

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{8.1}$$

其中，π_k 表示第 k 个高斯函数的权重，μ_k 和 Σ_k 表示第 k 个高斯的均值和方差。

GMM 模型的参数求解，一般使用 EM 算法进行求解。通常情况下，GMM 模型可以平滑地逼近任意形状的函数，具备对实际数据极强的表征力。而声纹识别实际上就是从不同语音中抽取出相同的表征特征来。GMM 模型同时还具备比较好的泛化能力。因此 GMM 模型在声纹识别初期获得比较好的效果。随着 k 的增大，所需要的训练数据也就更加大了，否则获取不到泛化能力较好的模型。

在实际使用过程中，每个人语音数据有限，很难获取到比较通用的声纹识别模型。为了解决这个问题，DA Reynolds 提出了通用背景模型 (Universal Background Model) [105]，简称 UBM。先使用大量和说话人无关的语音数据训练一个 GMM 模型，然后再使用少量的说话人数据，通过自适应算法 (如最大后验概率 MAP、最大似然线性回归 MLLR 等) 获取到说话人的个性特征的模型叫做 UBM 模型。这个思想有点像现在深度学习的 finetune 思想。这个模型就是 GMM-UBM 模型。该模型参数可以减半并有更快收敛的特点。

随着实际应用，GMM-UBM 的存在问题：参数仍然很大和受信道的干扰比较大。学术界提出了 GMM-SVM 模型 [106]、JFA 模型 [107] 等等去优化解决。

8.2.1.2 ivector 模型

基于 GMM-UBM 的模型，基本是基于特征声纹空间与特征信道空间的独立假设，但是在现实使用中，数据之间都是具有相关性的。之前的假设更多是方便了公式推导同时也限制了模型的泛化能力。N.Dehak 认为既然声纹信息与信道信息不能做到完全独立，那就用一个一段低维度的定长向量同时描述声纹信息和信道信息，从而提出了 ivector 模型 [108]。

对于每一段语音都有高斯均值向量 M 表示如下：

$$M = m + \omega T \tag{8.2}$$

其中，m 表示通用背景模型 (UBM) 的高斯均值向量，该值和声纹信息、信道信息无关，ω 是全局差异空间因子，即为 ivector 向量，它的先验服从标准正态分布 $N(0,1)$，T 表示全局的差异空间矩阵。接下来只需要估计 ω 和 T 值即可。

对于 ω 和 T 的参数估计，我们基于假设：每一段语音都来自不同的说话人。首先计算训练数据中每个说话人所对应的 Baum-Welch 统计量，随后随机产生 T 的初始值。后续采取 EM 算法估算得到相关的参数。

8.3 应用案例

8.3.1 声纹 1：1 识别应用案例

8.3.1.1 电话客服核身

在电话客服中应用声纹核身，可以节省核身时间，降低运营成本；并减少核身问题，提升客户体验；更重要的是，即使犯罪分子掌握了用户的所有信息，也能通过声音判断是否为本人，是否存在欺诈风险。例如，部分银行的服务热线目前已接入声纹核身技术，大致流程如图 8-2。

8.3.1.2 社保核查

声纹验证能够解决参保人员面临的远程和现场身份核查及生存验证的问题，避免了指纹验证和人脸识别等需要现场办理、不易采集、伪造等

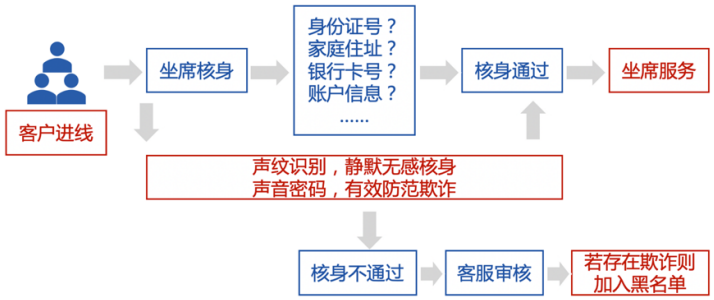


图 8-2 电话客服核身基本流程

问题，有效杜绝冒领养老金的可能性，节约社保资金和人力成本。例如，印尼新一代养老金认证系统已通过声纹验证技术，使其 250 万离退休人员在领取养老金时可通过电话或手机 app 进行远程身份认证，不仅节省了大量人力投入，还显著降低了传统骗保率。

8.3.1.3 声纹锁

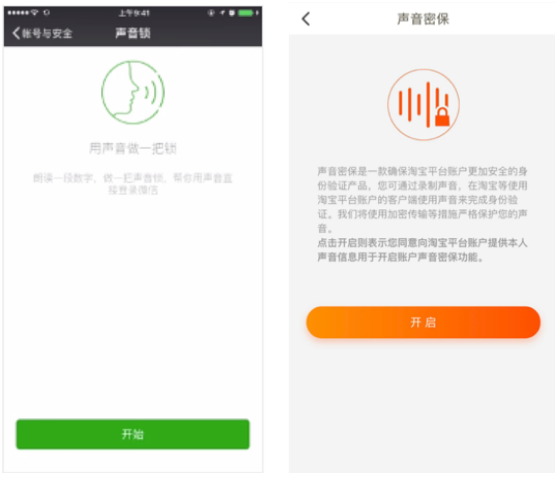
如图 8-3所示声纹验证可作为登录、密保、修改账户信息的一种验证方式，也可以应用于门禁/闸机等。

8.3.1.4 声纹支付

央行发布声纹识别安全应用技术标准，认定声纹技术可适用于手机银行、第三方支付，通过与语音交互硬件结合，能够解决无屏或远场进行身份验证的痛点。

8.3.1.5 声纹唤醒

通过声纹识别“主人”身份，只允许主人唤醒设备，或定制个性化服务，可适用于手机助手、智能音箱、智能家居、车载助手、服务机器人等智能设备。如图 8-5所示，在车联网应用中用户可以提前注册声纹信息并添加



(a) 声纹锁（一） (b) 声纹锁（二）

图 8-3 声纹锁样例

(a) 声纹支付 (b) 声纹支付界面

图 8-4 声纹支付样例

个性化配置，车机将通过声纹识别确认当前的驾驶人身份，可快速切换至对应的用户配置，令行车体验更加轻松。

图 8-5 汽车声纹唤醒

8.3.2 声纹 1: N 识别应用案例

8.3.2.1 公安

以破案、追逃为导向，利用声纹识别技术公安可进行“案查人”、“人查案”、“案查案”与“人查人”等多种排查方式：

- 案查人：如电信诈骗，主要线索只有语音的情况下，将该语音进行声纹库大库检索，快速锁定嫌疑人。
- 人查案：公安抓捕到可疑人员后，提取出该人的声纹特征，将其放入尚未侦破的语音案件中，排查该人是否为在逃人员。
- 案查案：公安人员可使用声纹识别技术将尚未侦破的语音案件以及语音线索归纳整理，从中排查是否有多起案件是同一人所为，帮助侦察人员获得更多线索，提高排查效率。
- 人查人：公安机关在抓捕到可疑人员后，提取出该人的声纹特征，为避免该人使用伪造身份，可将其声纹特征放入已知人员的声纹库，查询其真实身份。

声纹识别技术还能应用于重点人员监管、反电信诈骗、反恐、刑事案件侦破、身份查询与核验，助力公安有效遏制与打击犯罪，构建和强化安全的社会公众环境。例如：

- 反电信网络欺诈：在通信系统或安全监测系统中嵌入声纹识别技术，能够对黑名单人员语音对话实时预警，提示重点人员可疑行为；语音内容关键词识别动态预警，提示可疑案件与犯罪意图。

- 动态声纹布控：通过声纹识别和声纹大数据技术，进行对重点人员和关键卡口的布控监管，在第一时间完成举报人或嫌疑人身份鉴定，辅助刑事案件侦破和案情分析。

图 8-6 公安布控监管系统

8.3.2.2 金融黑名单识别

将信贷黑名单、风险等级高、不良中介、金融欺诈等用户声纹加入黑名单库，当其再次办理业务时，匹配到黑名单库的用户，直接给出风险预警。例如，车险业务能够针对报假案、修理厂、黑中介等不良用户建立黑名单声纹库，当不良用户再次报案时，业务员端能够及时给出预警。

8.3.2.3 客户定制化服务

对于某些小群体用户，如银行 VIP 客户电话呼入时，可通过声纹 1:N 匹配 VIP 声纹库，识别用户身份，从而进行定制化服务。

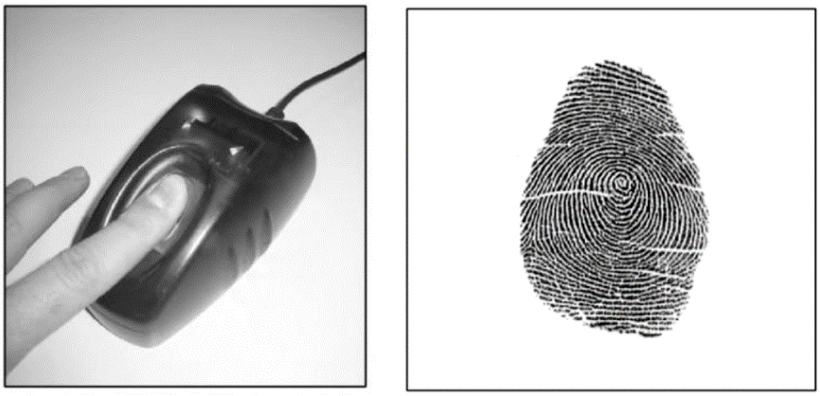
Chapter 9
其他生物特征识别

随着移动互联网、智能移动终端设备的快速发展，由于具备便捷的使用体验、可靠的安全保障，以人脸识别为代表的生物特征识别技术 (biometrics) 得到了迅速应用和推广。生物特征识别技术是利用固有的生物特征进行身份认证的一类技术，由于生物特征通常具备唯一性，如果具备可测量和可验证，那么利用生物识别技术进行身份认证往往安全、可靠和准确。除了前面介绍的人脸识别之外，下面将介绍几种在金融行业应用的生物识别技术以及应用场景。

9.1 指纹/掌纹识别

作为生物识别技术在金融领域应用最早的一项技术，指纹识别早在上世纪 90 年代就大规模进军金融行业。指纹是人类手指末端由凹凸的皮肤所形成的纹路，如图9-1(b)所示，每个个体指纹的形状不会随着年龄发生改变，而且每个人的指纹都是不同的。指纹识别技术通过分析所采集的指纹图像中可测量的特征点并提取特征值，然后进行比对认证。指纹识别目前也早已在消费电子、安防等领域广泛应用，相关针对指纹的国家标准也已陆续制定和发布，指纹识别的性能已得到明显提高，技术也最为成熟，目前已处于应用成熟的平台期。

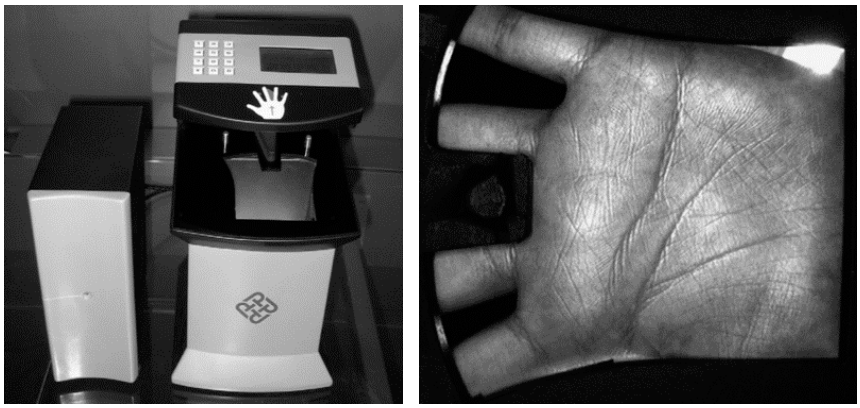
掌纹识别是近些年提出的一种相对较新的生物特征识别技术。掌纹一般指手指末端到手腕之间这一区域的手掌表面的各种纹理特征，如图9-2(b)所示。与指纹识别类似，每个人的掌纹纹理都不一样，掌纹中具有的



(a) 指纹采集设备 (b) 指纹图像

图 9-1 指纹采集设备和图像 [109]

很多特征可以进行测量并提取特征值，进而进行身份认证。指纹识别和掌纹识别都是非侵犯性的识别方法，实际应用中用户接受度较高。

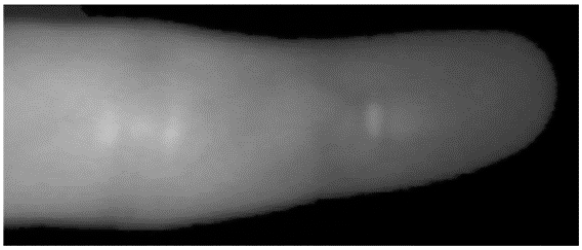


(a) 掌纹采集设备 (b) 掌纹图像

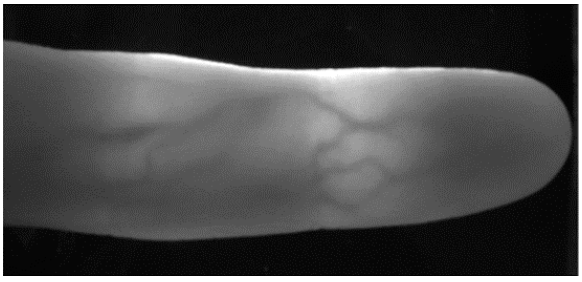
图 9-2 掌纹采集设备和图像 [110]

9.2 静脉识别

静脉识别利用的是静脉中的血红蛋白相对于肌肉、骨骼等其他生理组织对近红外光的吸收率更高，当有近红外光照射在手指或者手掌上时，通过红外摄像头获取手指或手掌的图像，静脉血管会呈现深色，肌肉组织则为浅色，呈现出黑白对比分明的图像特点，如图9-3所示，静脉血管结构可以清晰的得到呈现。静脉识别技术通过算法对图像进行分析，提取特征值进行身份认证。静脉采集设备按照近红外光源和图像传感器的相对位置不同，主要分为透射式和反射式两种，如图9-4所示。



(a) 指静脉反射式成像



(b) 指静脉透射式成像

图 9-3 指静脉图像 [111]

由于静脉属于人体内部特征，相比人脸和指纹来说极难复制和盗取，且只有在活体上才能采集到，因此这项技术的安全性更好，更难以被盗取和假冒。静脉识别过程中一般受外界环境因素（例如温度、湿度等）以及个体皮肤表面状态（如粗糙程度、是否磨损等）影响较小，可靠性较高。同时在使用过程中，用户手无需与设备进行接触，这种非接触式的使用更加卫生、易于用户接受。目前静脉识别技术发展时间较短，识别准确率

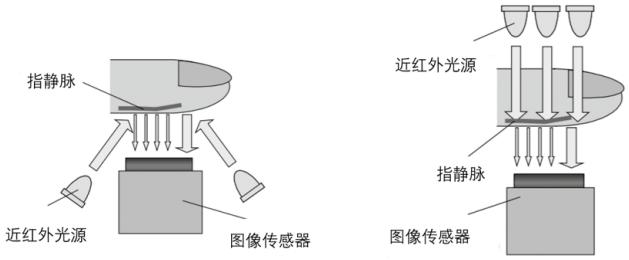


图 9-4 两种指静脉图像采集方式示意图 [111]

有望进一步得到提升。由于在活体鉴定方面的优势，与指纹融合可以较好地预防假指纹攻击，并提高识别准确率。静脉识别在金融行业的应用也已开始了探索。2016 年 1 月，工农中建四大银行委托广电运通起草指静脉在金融行业的应用标准。2016 年 11 月，广东省社保基金管理局也已经制定指静脉在社保行业的应用标准，计划全省开始推广指静脉养老金发放的认证工作。

9.3 虹膜识别

虹膜是位于人眼表面黑色瞳孔和白色巩膜之间的圆环状区域，在近红外光下能够呈现出丰富的纹理，如图9-5所示。而且虹膜在胎儿发育阶段形成后的整个生命历程中是保持不变的。这些特征决定的虹膜特征以及用于身份识别的唯一性。虹膜识别属于非接触式识别，通过专门的虹膜图像采集装置采集清晰的虹膜图像提取特征进行身份认证，识别过程高效且准确率高。虹膜识别技术被认为是生物特征识别技术中准确率最高的技术之一，在金融领域一般应用于金库管理、押运管理的较多，通过虹膜识别确认出入和押运人员身份，确保财产安全；同时，也有部分银行在尝试将虹膜识别和指静脉识别集成于自助终端中，实现更高安全级别的身份认证，以帮助用户完成自助贷款、自助理财等业务的办理。在其他行业，虹膜识别技术因眼镜、光线干扰和特征部位与采集方式等因素，目前主要用于煤矿工人等其他种类生物特征难以采集和识别的人群。

以上是目前较为常见的几种生物特征识别技术。随着智能终端设备以及生物特征传感器的快速普及和优化，生物特征识别技术已经进入大

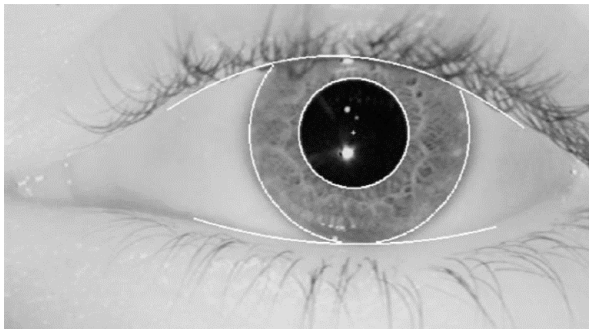


图 9-5 虹膜图像 [112]

规模应用阶段。单一的人脸识别或指纹识别已经难以满足金融机构的多样化需求，多模态生物识别是金融科技不可更改的趋势。

Chapter 10
反欺诈：声纹与人脸识别的抗攻击

这是里关于主题模型和语言模型的介绍。

10.1 声纹识别中的抗攻击

随着声纹识别的应用越来越广泛，声纹识别的安全性要求也越来越高。因此声纹识别研究中，必须了解声纹识别的攻击技术，才能进一步进行防范。在声纹识别应用中，主要的攻击方法分为四大类：模拟 (Impersonation)、录播 (Replay)、语音合成 (Speech Synthesis) 和语音转换 (Voice Conversion)。

模拟攻击通常来自两种情况：双胞胎和专业模仿人员。双胞胎主要是因为生物信息相对比较接近，所以存在声纹接近的可能。曾经有报道说有双胞胎弟弟用自己声音破解了哥哥在银行的声纹密码。有论文 [113] 提到这个问题，从他们的实验结果来看，就算是双胞胎还是存在一定的差异性，能进行区分。对应专业模仿人员，也有数据 [114] 显示具有声纹攻击性，但是差异比双胞胎大。这一类攻击最大的问题就是可供训练和测试的数据很有限，小数据上能进行防范，但是对于其真正的攻击情况目前认为是不确定的。

录播攻击，一般是用录音机对特定人进行录音，然后播放出来从而达到声纹认证通过的效果。在 ASVspoof2017 [115] 比赛中，通过参赛人的比赛结果以及对往届的比赛结果来看，录播攻击识别难度比语音合成、

语音转换攻击识别难度更高。录播攻击其中涉及到录音环境、播放环境、录音设备等干扰，会一定程度对识别是否是攻击造成一定难度。

语音合成攻击，一般分为两类，一个是拼接方式，一个是 TTS 语音生成方式。针对 TTS，机器识别准确一般比较高，因此不会造成比较大的威胁。但是语音拼接，这种技术方式，针对我们人耳能准确识别，但是对于机器而已，他们比较大可能认为是真实说话声音。

语音转换是把一个人说话的录音改变音色，频率等转换为具有另外一个人说话特点的技术 [116]。一般会用语音转换和 TTS 技术进行结合进行攻击，这样能比较方便有效，相对攻击成本比较低。这种攻击方式通常都能识别出来，机器生成的毕竟存在一定模式规律。但是也会存在一种风险，这种防止攻击的鲁棒性不一定很高。

针对这一系列的攻击方法，通用的抗攻击框架是在声纹识别前增加一个前置模块，用来专门识别是否是攻击。目前业界对这些攻击方式能较为准确识别，但是需要不断积累足够大的样本，从而提高更高的防护能力。

10.2 人脸识别中的抗攻击

人脸活体检测（Face Anti-Spoofing）技术是人脸识别系统中，用以确认待认证对象是否为真实生物活体的一项技术。一方面，人脸识别技术的商业化愈加成熟和广泛，极大改善和推进了社会金融活动的智能化和便捷性；另一方面，由于人脸照片、视频数据相对容易获取和复制，若无活体检测这一环节，那么使用被盗取的合法用户的照片、视频或者 3D 面及头套等即可入侵人脸识别系统，由此将带来极大危害 [117]。在目前的人脸识别系统中，常见的活体检测技术包括动作活体、3D 活体、红外活体、光线活体等。下面将逐一简单介绍。

动作活体是通过利用人脸关键点和人脸跟踪等技术，检测用户眨眼、张嘴、摇头等多种动作及其组合，可有效抵御照片、换脸、面具、遮挡以及屏幕翻拍等常见的攻击手段，但较难抵御视频回放攻击。

3D 活体通过专用硬件（例如 3D 结构光、ToF 等）获取人脸部 3D 结构信息，可以有效防御如手机、电脑等屏幕显示和打印照片等 2D 攻击手段，但是需要配合其他方法抵御近几年出现的高质量 3D 面具攻击 [118]。

红外活体检测一般利用人体皮肤对近红外光的反射率较高，相比于其他材质有明显区别的特性，通过专用红外设备获取人脸部红外图像判断是否为活体。实际应用中通常使用主动红外摄像，即通过红外 LED 照射人脸，利用红外摄像头获取人脸部图像，分类判断是否为活体。红外活体检测技术对于常见攻击手段具有较好的防御效果，缺点在于需要特定红外设备 [118, 119]。

光线活体是近两年出现的一种活体检测技术。由于 3D 活体、红外活体需要特殊设备，在已有系统中部署较为困难。光线活体技术利用屏幕发出不同颜色和强度的光线照射在人脸上，由于人脸自身的三维结构以及皮肤等生理组织对于不同颜色光线的反射率不同，从获取的视频中提取相应的活体信息，如图10-1所示。这项技术由于无需特殊硬件设备、且具有较高的准确率，在手机等移动端使用较为方便。其缺点在于要求视频拍摄过程稳定，闪光带来的用户体验需要得到提升，同时户外强光也会带来较大干扰 [120]。

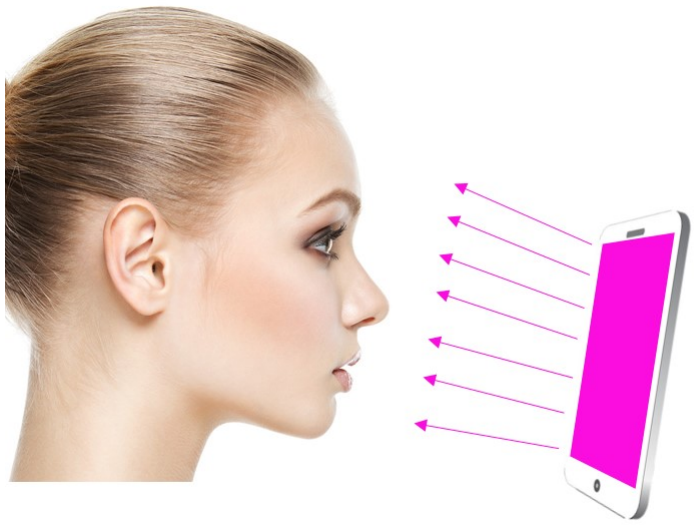


图 10-1 光线活体

References

[1] MARKOV A A. An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. Classical text in translation[J]. Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 1913, 23 : 591 – 600.

[2] SHANNON C E. A mathematical theory of communication[J]. Bell system technical journal, 1948, 27(3) : 379 – 423.

[3] CHOMSKY N. Three models for the description of language[J]. IRE Transactions on information theory, 1956, 2(3) : 113 – 124.

[4] KUČERA H, FRANCIS W N. Computational analysis of present-day American English[M]. [S.l.] : Dartmouth Publishing Group, 1967.

[5] JAYNES E T. Information theory and statistical mechanics[J]. Physical review, 1957, 106(4) : 620.

[6] VITERBI A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. IEEE transactions on Information Theory, 1967, 13(2) : 260 – 269.

[7] STRATONOVICH R L. Conditional markov processes[G] // Nonlinear transformations of stochastic processes. [S.l.] : Elsevier, 1965 : 427 – 453.

[8] RUDER S. A Review of the Neural History of Natural Language Processing[K/OL]. 2018. `https://ruder.io/a-review-of-the-recent-history-of-nlp/index.html#nonneuralmilestones`.

[9] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan) : 993 – 1022.

[10] YUAN J, GAO F, HO Q, et al. Lightlda: Big topic models on modest computer clusters[C] // Proceedings of the 24th International Conference on World Wide Web. 2015 : 1351 – 1361.

[11] JIANG D, SONG Y, LIAN R, et al. Familia: A configurable topic modeling framework for industrial text engineering[J]. arXiv

preprint arXiv:1808.03733, 2018.

[12] JIANG D, SONG Y, TONG Y, et al. Federated Topic Modeling[C] // Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019 : 1071 – 1080.

[13] WONG P-K, CHAN C. Chinese word segmentation based on maximum matching and word binding force[C] // Proceedings of the 16th conference on Computational linguistics-Volume 1. 1996 : 200 – 203.

[14] WEICHUN H, JIANJIAN J. Research on Longest Backward Segmentation for Context[C] // 1st International Workshop on Cloud Computing and Information Security. 2013.

[15] YANG L, XU L, SHI Z. An enhanced dynamic hash TRIE algorithm for lexicon search[J]. Enterprise Information Systems, 2012, 6(4) : 419 – 432.

[16] LI Q-H, CHEN Y-J, SUN J-G. A New Dictionary Mechanism for Chinese Word Segmentation [J][J]. Journal of Chinese Information Processing, 2003, 4 : 001.

[17] NG H I, LUA K T. A word finding automation for Chinese sentence tokenization[J]. submitted to ACM Transaction of Asian Languages Processing, .

[18] ZHANG H-P, LIU Q, CHENG X-Q, et al. Chinese lexical analysis using hierarchical hidden markov model[C] // Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. 2003 : 63 – 70.

[19] KUPIEC J. Robust part-of-speech tagging using a hidden Markov model[J]. Computer speech & language, 1992, 6(3) : 225 – 242.

[20] YU H-K, ZHANG H-P, LIU Q, et al. Chinese named entity identification using cascaded hidden Markov model[J]. Journal-China Institute Of Communications, 2006, 27(2) : 87.

[21] MORWAL S, JAHAN N, CHOPRA D. Named entity recognition using hidden Markov model (HMM)[J]. International Journal on Natural Language Computing (IJNLC), 2012, 1(4) : 15 – 23.

[22] MCCALLUM A, FREITAG D, PEREIRA F C. Maximum Entropy Markov Models for Information Extraction and Segmentation.[C] // Icml : Vol 17. 2000 : 591 – 598.

[23] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging[C] // Conference on Empirical Methods in Natural Language Processing. 1996.

[24] BORTHWICK A, GRISHMAN R. A maximum entropy approach to named entity recognition[D]. [S.l.] : Citeseer, 1999.

[25] ZHAO H, HUANG C, LI M. An improved Chinese word segmentation system with conditional random field[C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006 : 162 – 165.

[26] CONSTANT M, SIGOGNE A. MWU-aware part-of-speech tagging with a CRF model and lexical resources[C] // Proceedings of the workshop on multiword expressions: from parsing and generation to the real world. 2011 : 49 – 56.

[27] EKBAL A, HAQUE R, BANDYOPADHYAY S. Named entity recognition in Bengali: A conditional random field approach[C] // Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II. 2008.

[28] 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.

[29] ZHANG Y, CLARK S. Joint word segmentation and POS tagging using a single perceptron[C] // Proceedings of ACL-08: HLT. 2008 : 888 – 896.

[30] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(Aug) : 2493 – 2537.

[31] ZHENG X, CHEN H, XU T. Deep learning for Chinese word segmentation and POS tagging[C] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 : 647 – 657.

[32] WU F, LIU J, WU C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C] // The World Wide Web Conference. 2019 : 3342 – 3348.

[33] CHEN X, QIU X, ZHU C, et al. Long short-term memory neural networks for chinese word segmentation[C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 : 1197 – 1206.

[34] CHEN X, QIU X, ZHU C, et al. Gated recursive neural network for Chinese word segmentation[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 : 1744 – 1753.

[35] HATORI J, MATSUZAKI T, MIYAO Y, et al. Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese[C] // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012 : 1045 – 1053.

[36] SHAO Y, HARDMEIER C, TIEDEMANN J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[J]. arXiv preprint arXiv:1704.01314, 2017.

[37] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[38] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[39] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[40] CHANG P-C, GALLEY M, MANNING C D. Optimizing Chinese word segmentation for machine translation performance[C]

// Proceedings of the third workshop on statistical machine translation. 2008 : 224 – 232.

[41] TESNIÈRE L. Eléments de syntaxe structurale[J], 1959.

[42] 黄曾阳. HNC 理论概要 [J/OL]. 中文信息学报, 1997, 11(4) : 12. http://jcip.cipsc.org.cn/CN/abstract/article_674.shtml.

[43] DOZAT T, MANNING C D. Deep biaffine attention for neural dependency parsing[J]. arXiv preprint arXiv:1611.01734, 2016.

[44] 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.

[45] MA X, HOVY E. Neural probabilistic model for non-projective mst parsing[J]. arXiv preprint arXiv:1701.00874, 2017.

[46] FERNÁNDEZ-GONZÁLEZ D, GÓMEZ-RODRÍGUEZ C. Left-to-right dependency parsing with pointer networks[J]. arXiv preprint arXiv:1903.08445, 2019.

[47] KIPERWASSER E, GOLDBERG Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations[J]. Transactions of the Association for Computational Linguistics, 2016, 4 : 313 – 327.

[48] MA X, HU Z, LIU J, et al. Stack-pointer networks for dependency parsing[J]. arXiv preprint arXiv:1805.01087, 2018.

[49] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C] // 33rd annual meeting of the association for computational linguistics. 1995 : 189 – 196.

[50] CARRERAS X, MÀRQUEZ L. Introduction to the CoNLL-2005 shared task: Semantic role labeling[C] // Proceedings of the ninth conference on computational natural language learning (CoNLL-2005). 2005 : 152 – 164.

[51] BJÖRKELUND A, HAFDELL L, NUGUES P. Multilingual semantic role labeling[C] // Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. 2009 : 43 – 48.

[52] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[53] LUONG M-T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.

[54] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C] // Advances in neural information processing systems. 2015 : 1693 – 1701.

[55] SERBAN I V, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C] // Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16). 2016.

[56] CHEN H, LIU X, YIN D, et al. A survey on dialogue systems: Recent advances and new frontiers[J]. Acm Sigkdd Explorations Newsletter, 2017, 19(2) : 25 – 35.

[57] LECUN Y, BENGIO Y, OTHERS. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10) : 1995.

[58] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8) : 1735 – 1780.

[59] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in neural information processing systems. 2017 : 5998 – 6008.

[60] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.

[61] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C] // Advances in neural information processing systems. 2015 : 649 – 657.

[62] TAN J, WAN X, XIAO J. Abstractive document summarization with a graph-based attentional neural model[C] // Proceedings of

the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 : 1171 – 1181.

[63] YAO J-G, WAN X, XIAO J. Recent advances in document summarization[J]. Knowledge and Information Systems, 2017, 53(2) : 297 – 336.

[64] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb) : 1137 – 1155.

[65] 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.

[66] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C] // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014 : 1532 – 1543.

[67] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C] // Advances in neural information processing systems. 2014 : 3104 – 3112.

[68] SHANG L, LU Z, LI H. Neural responding machine for short-text conversation[J]. arXiv preprint arXiv:1503.02364, 2015.

[69] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8) : 9.

[70] SONG K, TAN X, QIN T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[71] FENG M, XIANG B, GLASS M R, et al. Applying deep learning to answer selection: A study and an open task[C] // 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2015 : 813 – 820.

[72] ZHOU X, LI L, DONG D, et al. Multi-turn response selection for chatbots with deep attention matching network[C] // Proceedings

of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 : 1118 – 1127.

[73] LU J, XIE Z, LING G, et al. Spatio-Temporal Matching Network for Multi-Turn Responses Selection in Retrieval-Based Chatbots[J], .

[74] YOUNG S, GAŠIĆ M, THOMSON B, et al. POMDP-based statistical spoken dialog systems: A review[J]. Proceedings of the IEEE, 2013, 101(5) : 1160 – 1179.

[75] SERBAN I V, SORDONI A, LOWE R, et al. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues[J]. arXiv preprint arXiv:1605.06069, 2016.

[76] KIM Y. Convolutional Neural Networks for Sentence Classification[C/OL] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar : Association for Computational Linguistics, 2014 : 1746 – 1751. https://www.aclweb.org/anthology/D14-1181.

[77] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C] // Twenty-ninth AAAI conference on artificial intelligence. 2015.

[78] 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.

[79] YOUNG S, SCHATZMANN J, WEILHAMMER K, et al. The hidden information state approach to dialog management[C] // 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 : Vol 4. 2007 : IV – 149.

[80] HENDERSON M, THOMSON B, YOUNG S. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation[C] // Spoken Language Technology Workshop (SLT), 2014 IEEE. 2014 : 360 – 365.

[81] HENDERSON M, THOMSON B, YOUNG S. Word-based dialog state tracking with recurrent neural networks[C] // Proceedings of

the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). 2014 : 292 – 299.

[82] SUTTON R S, BARTO A G. Reinforcement learning: An introduction : Vol 1[M]. [S.l.] : MIT press Cambridge, 1998.

[83] WATKINS C J C H. Learning from delayed rewards[D]. [S.l.] : University of Cambridge England, 1989.

[84] WEN T-H, GASIC M, KIM D, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking[J]. arXiv preprint arXiv:1508.01755, 2015.

[85] WEN T-H, GASIC M, MRKSIC N, et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems[J]. arXiv preprint arXiv:1508.01745, 2015.

[86] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C] // Advances in neural information processing systems. 2019 : 5754 – 5764.

[87] SONG Y, DI JIANG X W, XU Q, et al. Topic-aware dialogue speech recognition with transfer learning[J]. Interspeech, Austria, 2019.

[88] LAKE B M, SALAKHUTDINOV R, TENENBAUM J B. Human-level concept learning through probabilistic program induction[J]. Science, 2015, 350(6266) : 1332 – 1338.

[89] PEARL J. An introduction to causal inference[J]. The international journal of biostatistics, 2010, 6(2).

[90] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2) : 1 – 19.

[91] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C] // Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 : Vol 1. 2001 : I – I.

[92] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10) : 1499 – 1503.

[93] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C] // Advances in neural information processing systems. 2015 : 91 – 99.

[94] WANG H, LI Z, JI X, et al. Face r-cnn[J]. arXiv preprint arXiv:1706.01061, 2017.

[95] TANG X, DU D K, HE Z, et al. Pyramidbox: A context-assisted single shot face detector[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018 : 797 – 813.

[96] DENG J, GUO J, ZHOU Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv preprint arXiv:1905.00641, 2019.

[97] NAJIBI M, SAMANGOUEI P, CHELLAPPA R, et al. Ssh: Single stage headless face detector[C] // Proceedings of the IEEE International Conference on Computer Vision. 2017 : 4875 – 4884.

[98] SUN Y, CHEN Y, WANG X, et al. Deep learning face representation by joint identification-verification[C] // Advances in neural information processing systems. 2014 : 1988 – 1996.

[99] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 815 – 823.

[100] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C] // European conference on computer vision. 2016 : 499 – 515.

[101] LIU W, WEN Y, YU Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 212 – 220.

[102] WANG H, WANG Y, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018 : 5265 – 5274.

[103] DENG J, GUO J, XUE N, et al. Arcface: Additive angular margin loss for deep face recognition[C] // Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. 2019 : 4690 – 4699.

[104] REYNOLDS D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech communication, 1995, 17(1-2) : 91 – 108.

[105] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3) : 19 – 41.

[106] CAMPBELL W M, STURIM D E, REYNOLDS D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation[C] // 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings : Vol 1. 2006 : I – I.

[107] KENNY P. Joint factor analysis of speaker and session variability: Theory and algorithms[J]. CRIM, Montreal,(Report) CRIM-06/08-13, 2005, 14 : 28 – 29.

[108] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4) : 788 – 798.

[109] MALTONI D, MAIO D, JAIN A K, et al. Handbook of fingerprint recognition[M]. [S.l.] : Springer Science & Business Media, 2009.

[110] KONG A, ZHANG D, KAMEL M. A survey of palmprint recognition[J]. pattern recognition, 2009, 42(7) : 1408 – 1418.

[111] HASHIMOTO J. Finger vein authentication technology and its future[C] // 2006 Symposium on VLSI Circuits, 2006. Digest of Technical Papers.. 2006 : 5 – 8.

[112] DAUGMAN J. How iris recognition works[G] // The essential guide to image processing. [S.l.] : Elsevier, 2009 : 715 – 739.

[113] PATIL H A. Speaker recognition in Indian languages: A feature based approach[J]. Indian Institute of Technology Kharagpur (IIT-K), Ph. D Thesis, 2005.

[114] PATIL H A, BASU T K. LP spectra vs. Mel spectra for identification of professional mimics in Indian languages[J]. International

Journal of Speech Technology, 2008, 11(1) : 1 – 16.

[115] KINNUNEN T, SAHIDULLAH M, DELGADO H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[J], 2017.

[116] JIA Y, ZHANG Y, WEISS R, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C] // Advances in neural information processing systems. 2018 : 4480 – 4490.

[117] CHINGOVSKA I, ANJOS A, MARCEL S. On the effectiveness of local binary patterns in face anti-spoofing[C] // 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). 2012 : 1 – 7.

[118] ZHANG S, WANG X, LIU A, et al. A dataset and benchmark for large-scale multi-modal face anti-spoofing[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019 : 919 – 928.

[119] YI D, LEI Z, ZHANG Z, et al. Face anti-spoofing: Multi-spectral approach[G] // Handbook of Biometric Anti-Spoofing. [S.l.] : Springer, 2014 : 83 – 102.

[120] LIU Y, TAI Y, LI J, et al. Aurora guard: Real-time face anti-spoofing via light reflection[J]. arXiv preprint arXiv:1902.10311, 2019.