文本表征学习

1 自然语言处理绪论

EE1513

宋彦

提纲

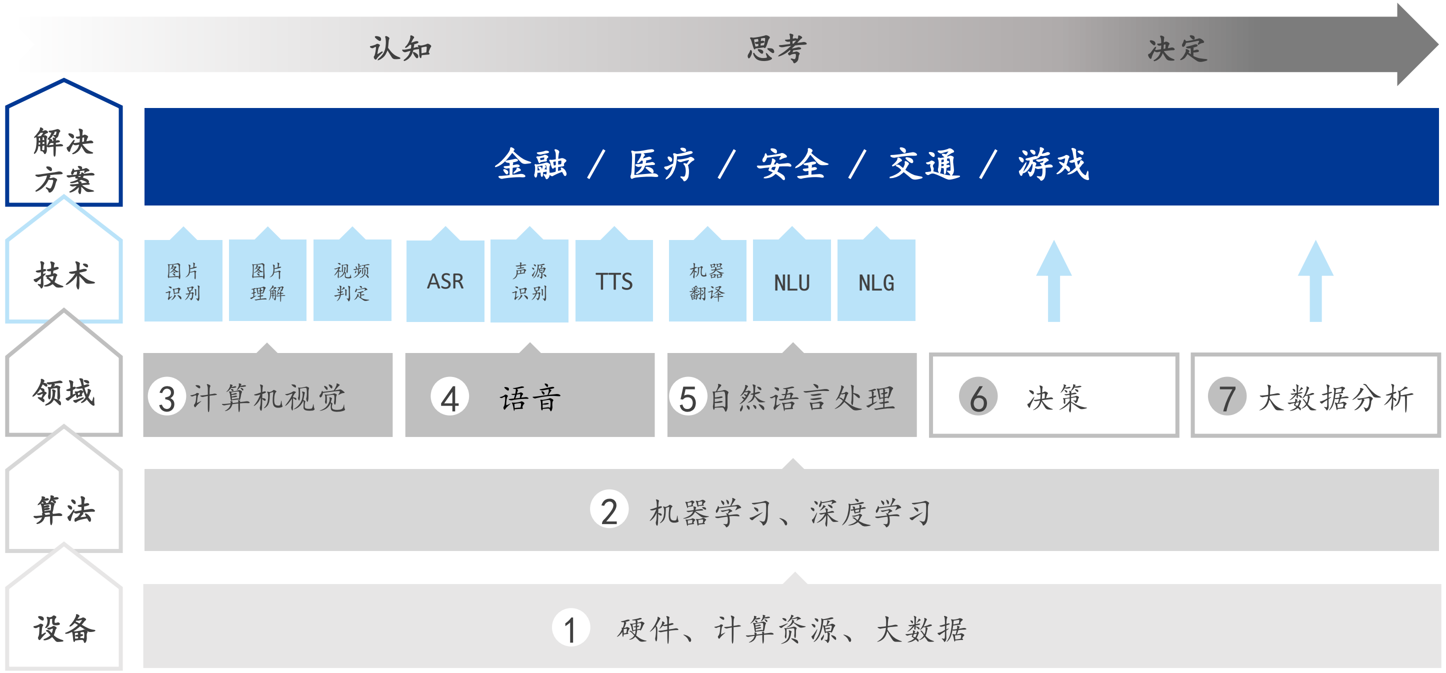
- 自然语言处理简介
- 为什么需要自然语言处理

- 计算机的编码
- 自然语言的不同编码方式

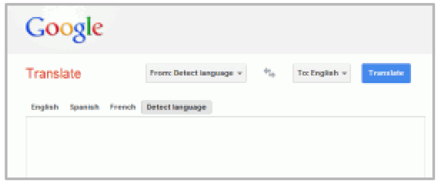
- 预备知识

自然语言处理简介

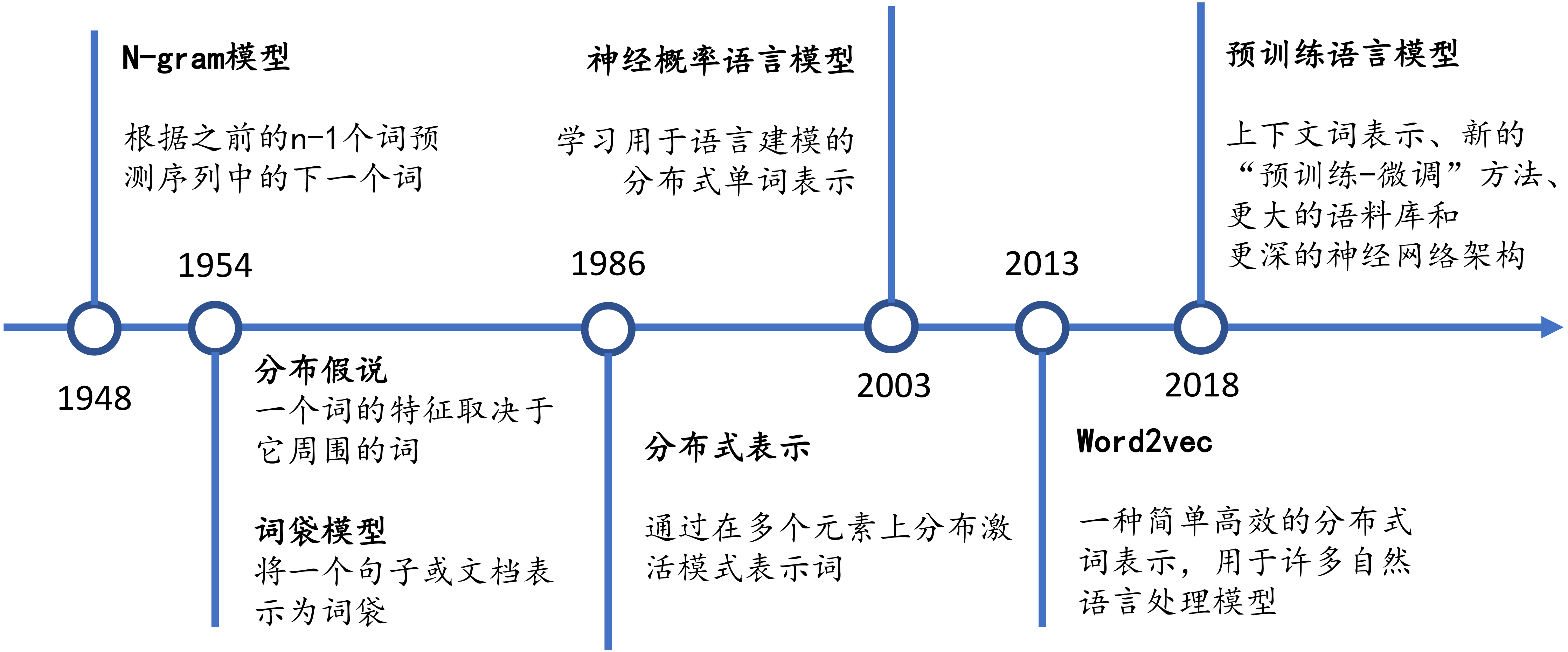
- 自然语言处理是语言学、计算机科学、信息工程、人工智能的交叉学科，研究计算机如何处理和分析自然语言。
- 自然语言处理（NLP）在AI中的位置



自然语言处理历史

	起源	低谷与复苏	发展	优化
	1950s–1960s	1970s–1980s	1990s–2000s	2010s以后
		Météo Média La douceur persiste et signe Jeudi Vancouver Winnipeg Montréal	 Google Translate From: Detect language To: English Translate English Spanish French Detect language	
应用	IBM-701 首次将俄语机器翻译成英语	ALPAC报告； TAUM-METEO 用于天气预报翻译；	NLP进入语音处理、搜索引擎等多个领域	Watson参加Jeopardy并获胜； 个人助理、聊天机器人；
技术	基于规则的方法； 基于概率的方法；	篇章分析	数据驱动； 数据来自互联网	浅层+深度学习
基础	计算机的发明		现代电脑成为主流； 互联网驱动NLP需求；	大数据

表征学习历史



1948

N-gram模型

根据之前的n-1个词预
测序列中的下一个词

1954

分布假说

一个词的特征取决于
它周围的词

词袋模型

将一个句子或文档表
示为词袋

1986

分布式表示

通过在多个元素上分布激
活模式表示词

2003

神经概率语言模型

学习用于语言建模的
分布式单词表示

2013

Word2vec

一种简单高效的分布式
词表示，用于许多自然
语言处理模型

2018

预训练语言模型

上下文词表示、新的
“预训练-微调”方法、
更大的语料库和
更深的神经网络架构

为什么自然语言处理很难？

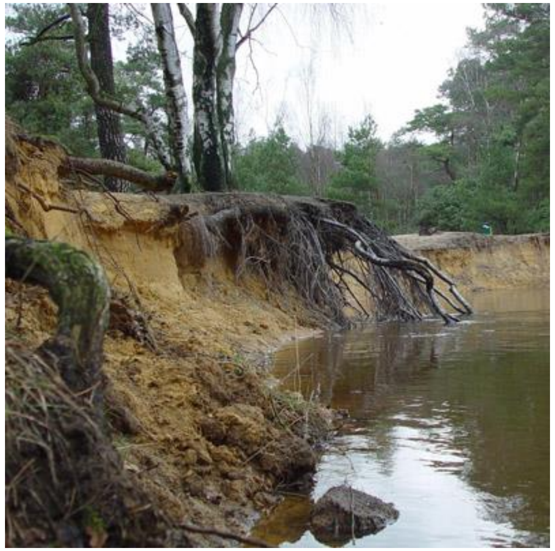
- 让我们考虑下面的英语句子

I saw a man with a telescope.

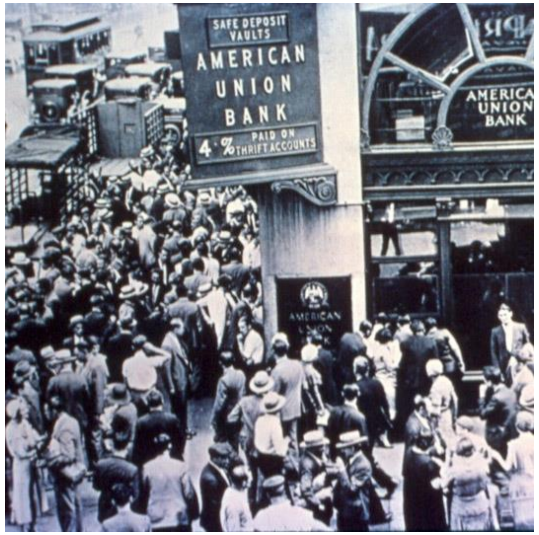
- 由一组符号组成
- 以 Unicode 或其他编码方式按顺序（字符串）出现
- 可以被拆分为一些“词” —— 词的歧义
- 组成了一个句子 —— 句子结构的歧义

词的歧义

The bank is crumbling.



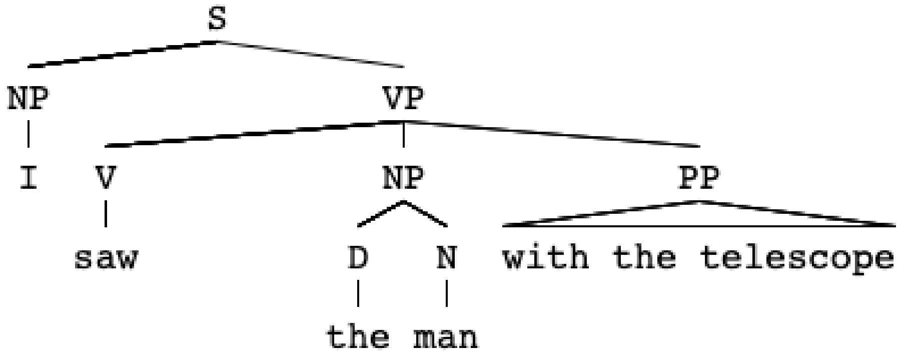
?

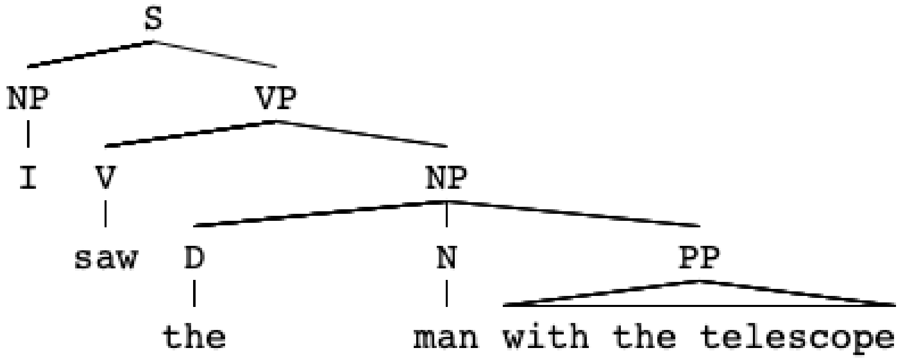


句子结构的歧义

I saw a man with a telescope.

?





为什么需要自然语言处理:
自然语言处理很难

- （书面）语言是一种符号系统
  - 语言极其复杂，尽管我们很容易使用它
  - 歧义是语言的固有特征
  - 推断一种语言的句法规则是困难的
- 机器通常处理数字（0,1）
- 保留尽可能多的信息
- 语言无关
- 效率和稳健性
- ...

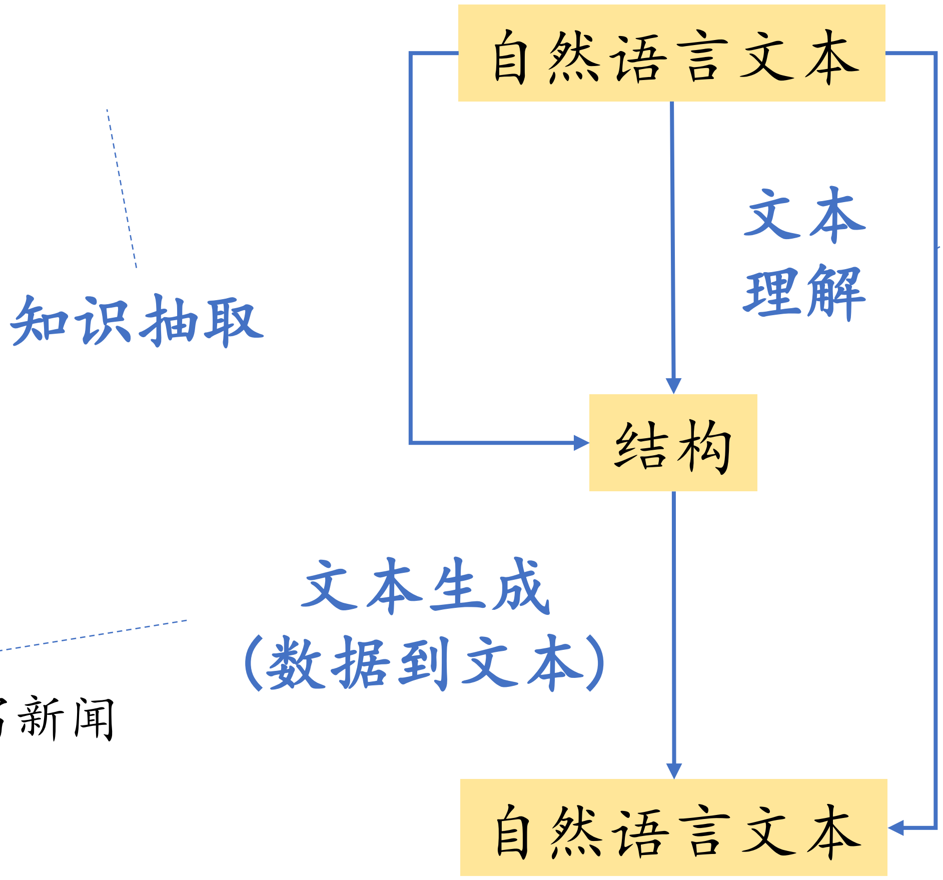
自然语言处理的方面

- 词、词库：词法分析
  - 词法、分词
- 句法
  - 句子结构、短语、语法、……
- 语义学
  - 意义
  - 执行命令
- 话语分析
  - 文本的含义
  - 句子之间的关系（例如照应）

主要的自然语言处理任务

- 关系抽取
- 特征抽取
- 词向量
- …

- 文本分类
- 词性标注
- 命名实体识别
- 句法分析
- 语义分析
- …



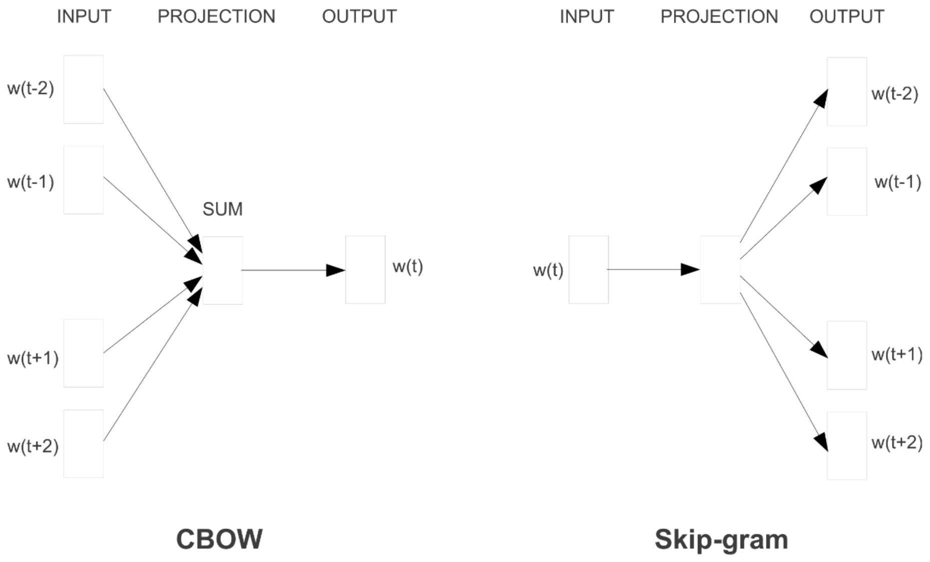
文本生成
（文本到文本）

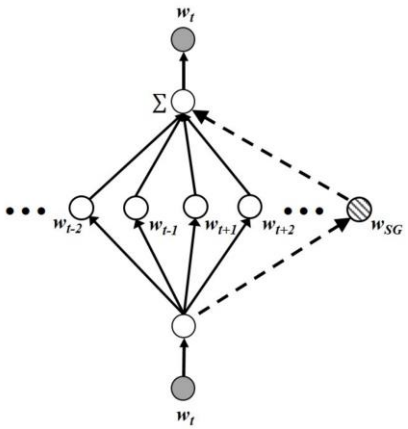
- 根据体育数据写新闻
- …

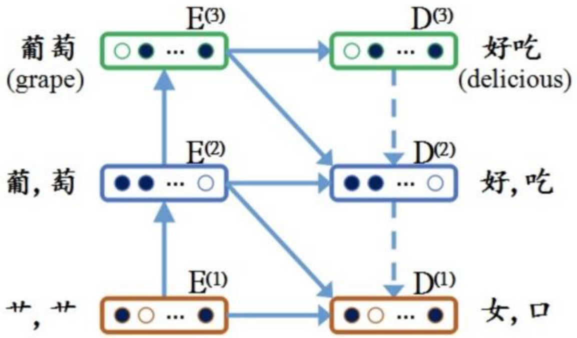
- 机器翻译
- 聊天机器人
- …

词向量

- The way to represent text for deep learning
  - Vectors
  - Trainable
- Unsupervised learning
  - No annotation required
- High density and low dimension
  - Reduce complexity
  - Fit for neural networks
- Semantic driven
  - Similarity and relatedness computation



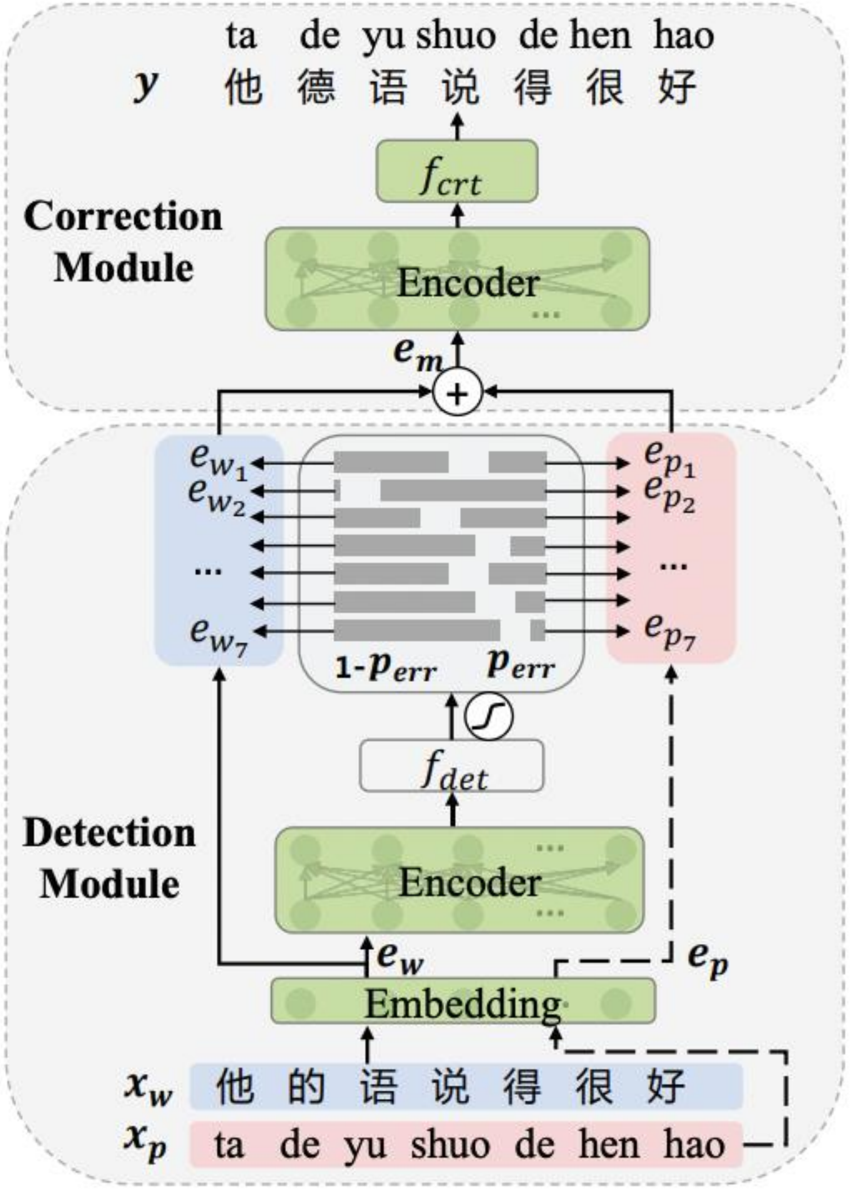




Linguistic Regularities in Continuous Space Word Representations (NAACL, 2013)
Distributed Representations of Words and Phrases and their Compositionality (NIPS, 2013)
Joint learning embeddings for Chinese words and their components via ladder structured networks (IJCAI 2018)
Complementary Learning of Word Embeddings (IJCAI 2018)

拼写检查

- Generally contains two steps
  - Error detection
  - Error correction
- Approaches
  - Language models
  - Lexicons
- Challenges
  - Out of vocabulary words
  - Multilingual support



Correcting Chinese Spelling Errors with Phonetic Pre-training (ACL: Findings, 2021)

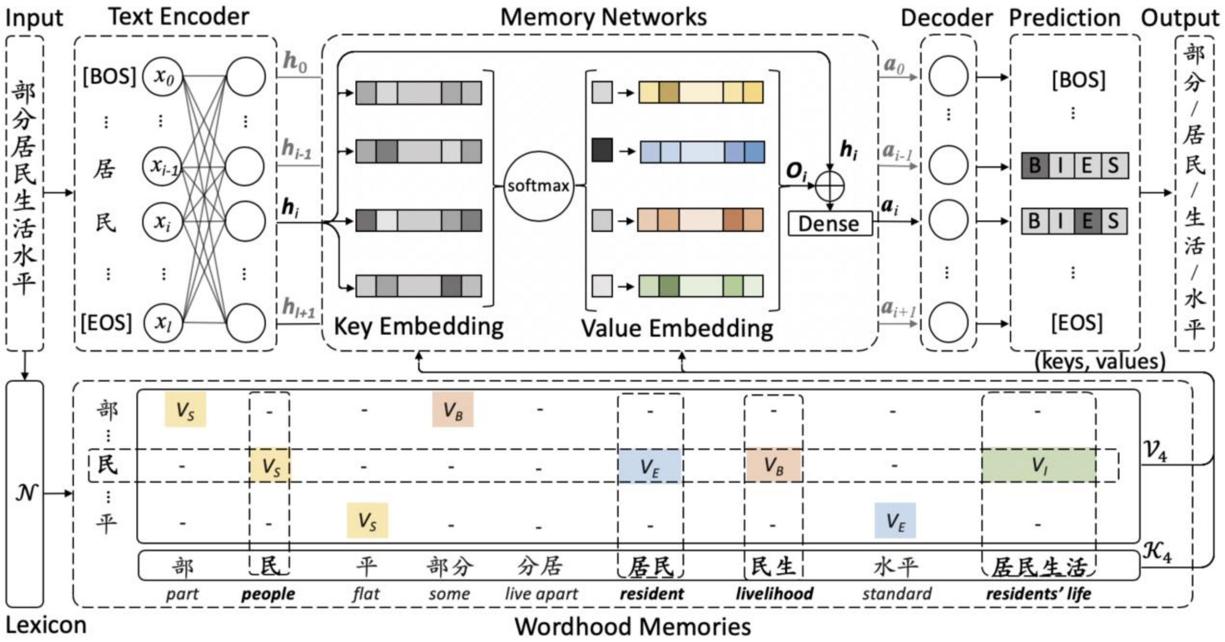
分词

- Basic processing tasks for Chinese

  当/原子/结合/成/分子/时

  乒乓球/拍卖/完了

- Different approaches
  - Word-based
  - Character-based（mainstream）
    - Sequence labeling based on neural networks
- Challenges
  - Ambiguities
  - Out-of-vocabulary words
  - Data sparseness



Improving Chinese Word Segmentation with Wordhood Memory Networks (ACL, 2020)

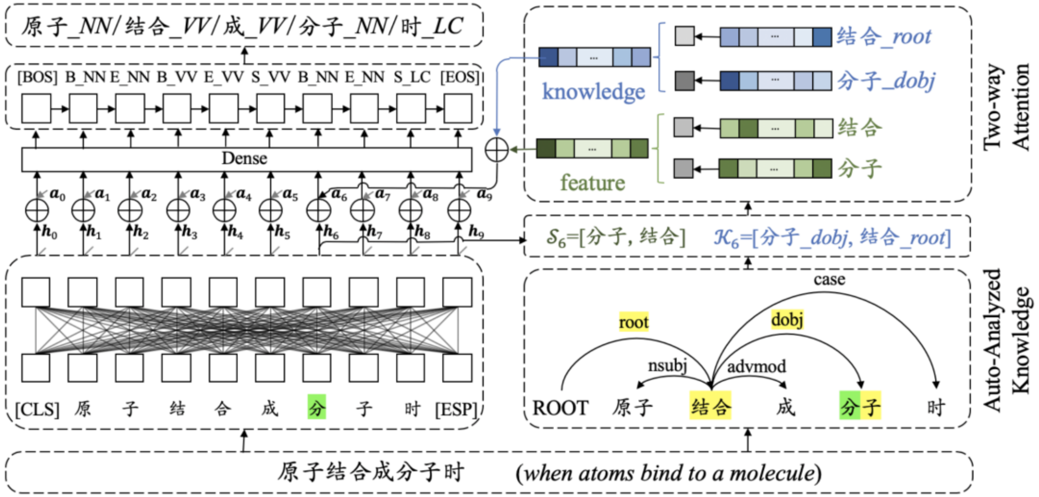
词性标注

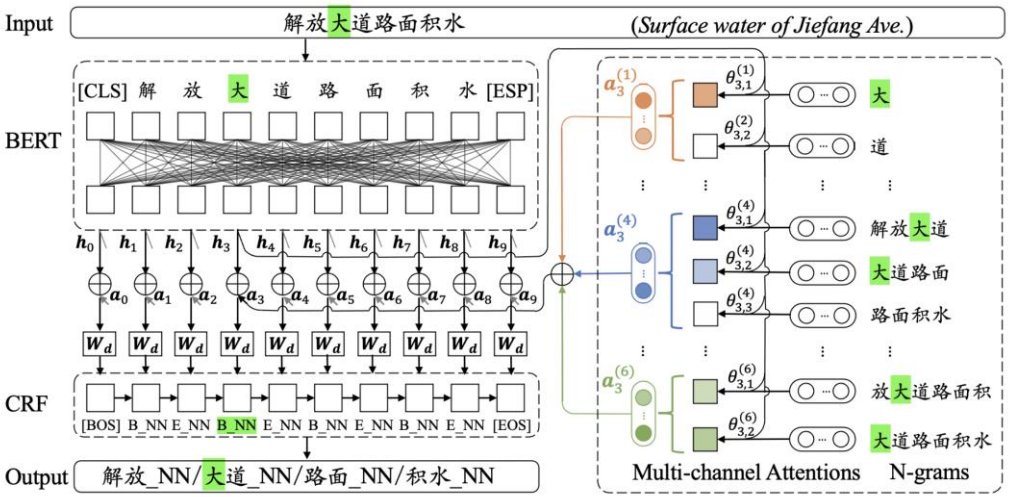
- Annotate POS tags

  报告/VV 书/NN 上/LC 的/DEG 内容/NN

  发表/VV 一/CD 篇/M 报告/NN

- For Chinese, it is generally performed jointly with CWS
- It is a typical sequence labeling task
- Challenges
  - Ambiguities
  - Out-of-vocabulary words

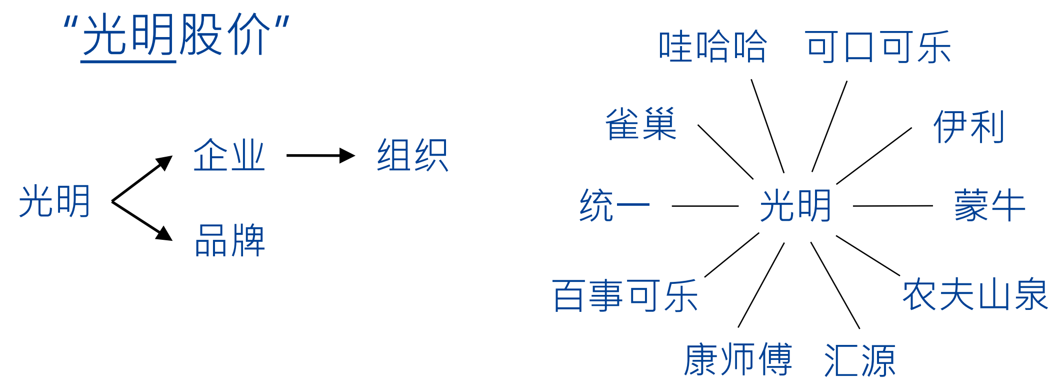

原子_NN/结合_VV/成_VV/分子_NN/时_LC
[BOS] B_NN E_NN B_VV E_VV S_VV B_NN E_NN S_LC [EOS]
Dense
a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9
h_0 h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9
[CLS] 原 子 结 合 成 分 子 时 [ESP]
knowledge
feature
结合_root
分子_dobj
结合
分子
Two-way Attention
$\mathcal{S}_6$=[分子, 结合]
$\mathcal{K}_6$=[分子_dobj, 结合_root]
case
root
dobj
nsubj
advmod
ROOT 原子 结合 成 分子 时
Auto-Analyzed Knowledge
原子结合成分子时 (when atoms bind to a molecule)



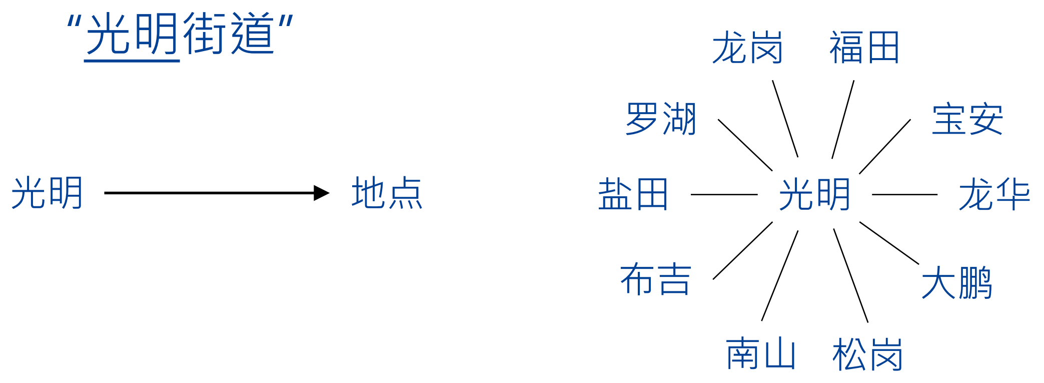
Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge (ACL, 2020)
Joint Chinese Word Segmentation and Part-of-speech Tagging via Multi-channel Attention of Character N-grams (COLING, 2020)

词义消歧

- A semantic analysis task
  - Context driven: word meaning is decided by its context
  - Sometimes require to extend the current word with more features

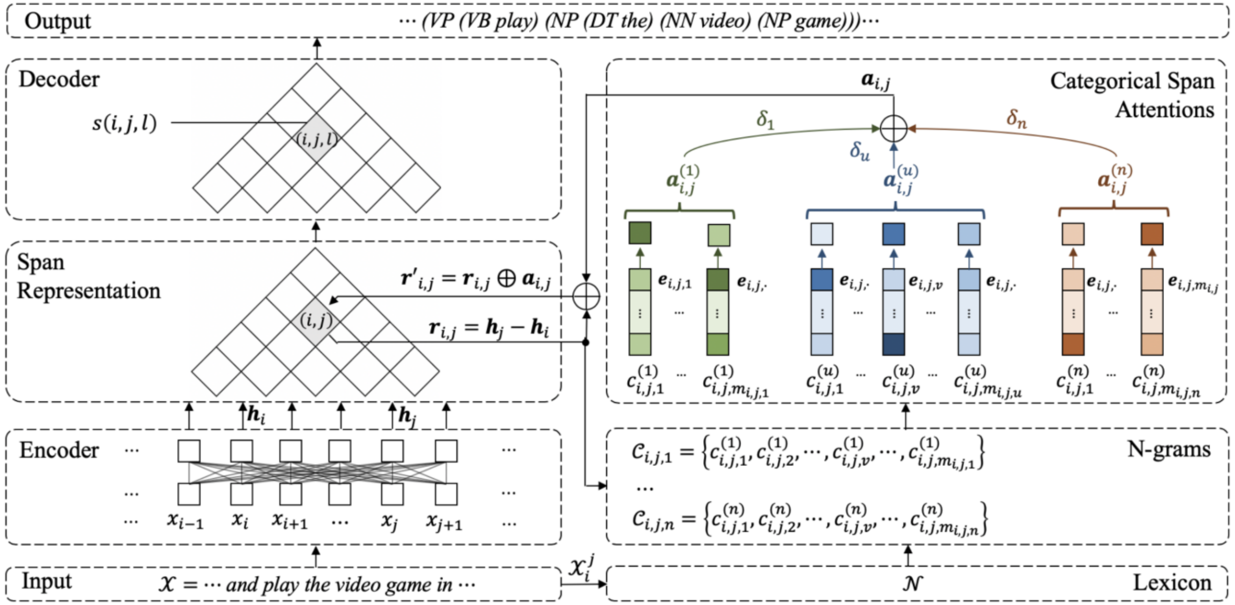


- Methods
  - Knowledge induced: information extraction and clustering
  - Better representation from knowledge and context

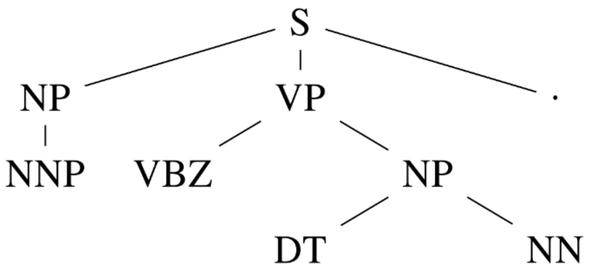


句法分析

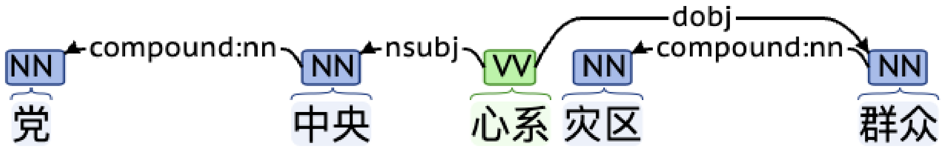
- Parsing text with structures
  - Top-down and bottom-up
  - Syntax and dependency

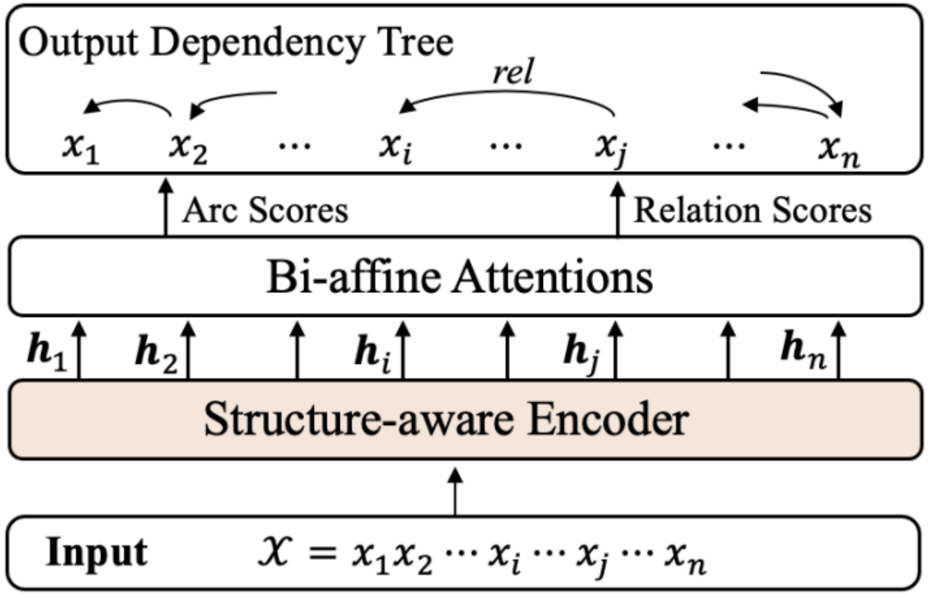

Output
··· (VP (VB play) (NP (DT the) (NN video) (NP game)))···
Decoder
$s(i,j,l)$
(i,j,l)
Span Representation
(i,j)
$r'_{i,j} = r_{i,j} \oplus a_{i,j}$
$r_{i,j} = h_j - h_i$
h_i
h_j
Encoder
x_{i-1} x_i x_{i+1} ··· x_j x_{j+1}
Input
$\mathcal{X}$ = ··· and play the video game in ···
x_i^j
Categorical Span Attentions
$a_{i,j}$
δ_1 δ_u δ_n
$a_{i,j}^{(1)}$ $a_{i,j}^{(u)}$ $a_{i,j}^{(n)}$
$e_{i,j,1}$ $e_{i,j,\cdot}$ $e_{i,j,v}$ $e_{i,j,m_{i,j}}$
$c_{i,j,1}^{(1)}$ $c_{i,j,m_{i,j,1}}^{(1)}$ $c_{i,j,1}^{(u)}$ $c_{i,j,v}^{(u)}$ $c_{i,j,m_{i,j,u}}^{(u)}$ $c_{i,j,1}^{(n)}$ $c_{i,j,m_{i,j,n}}^{(n)}$
N-grams
$\mathcal{C}_{i,j,1} = \{c_{i,j,1}^{(1)}, c_{i,j,2}^{(1)}, \cdots, c_{i,j,v}^{(1)}, \cdots, c_{i,j,m_{i,j,1}}^{(1)}\}$
$\mathcal{C}_{i,j,n} = \{c_{i,j,1}^{(n)}, c_{i,j,2}^{(n)}, \cdots, c_{i,j,v}^{(n)}, \cdots, c_{i,j,m_{i,j,n}}^{(n)}\}$
$\mathcal{N}$
Lexicon

John has a dog . →



John has a dog . → (S (NP NNP)$_{\text{NP}}$ (VP VBZ (NP DT NN)$_{\text{NP}}$)$_{\text{VP}}$.)$_{\text{S}}$




Output Dependency Tree
rel
x_1 x_2 ··· x_i ··· x_j ··· x_n
Arc Scores
Relation Scores
Bi-affine Attentions
h_1 h_2 h_i h_j h_n
Structure-aware Encoder
Input
$\mathcal{X} = x_1 x_2 \cdots x_i \cdots x_j \cdots x_n$

Improving Constituency Parsing with Span Attention (EMNLP: Findings, 2020)
Enhancing Structure-aware Encoder with Extremely Limited Data for Graph-based Dependency Parsing (COLING, 2022)

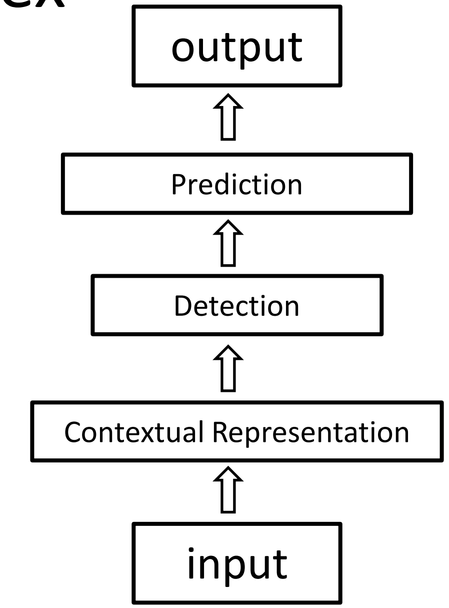
指代消解

- Finding the relations among different nouns/pronouns
- From simple to complex
  - Neural models
  - Add Knowledge

Rule and features

Entity-centric coreference resolution with model stacking. (ACL, 2015)

Deep reinforcement learning for mention ranking coreference models. (EMNLP, 2016)



output

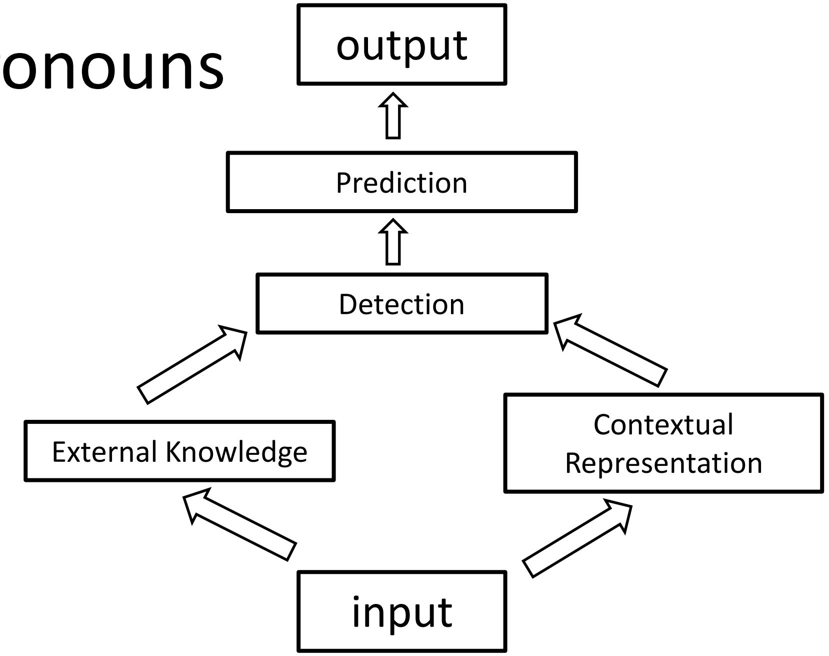
Prediction

Detection

Contextual Representation

input

End-to-end coreference resolution (EMNLP, 2017)



output

Prediction

Detection

External Knowledge

Contextual Representation

input

Incorporating context and external knowledge for pronoun coreference resolution. (NAACL, 2019)

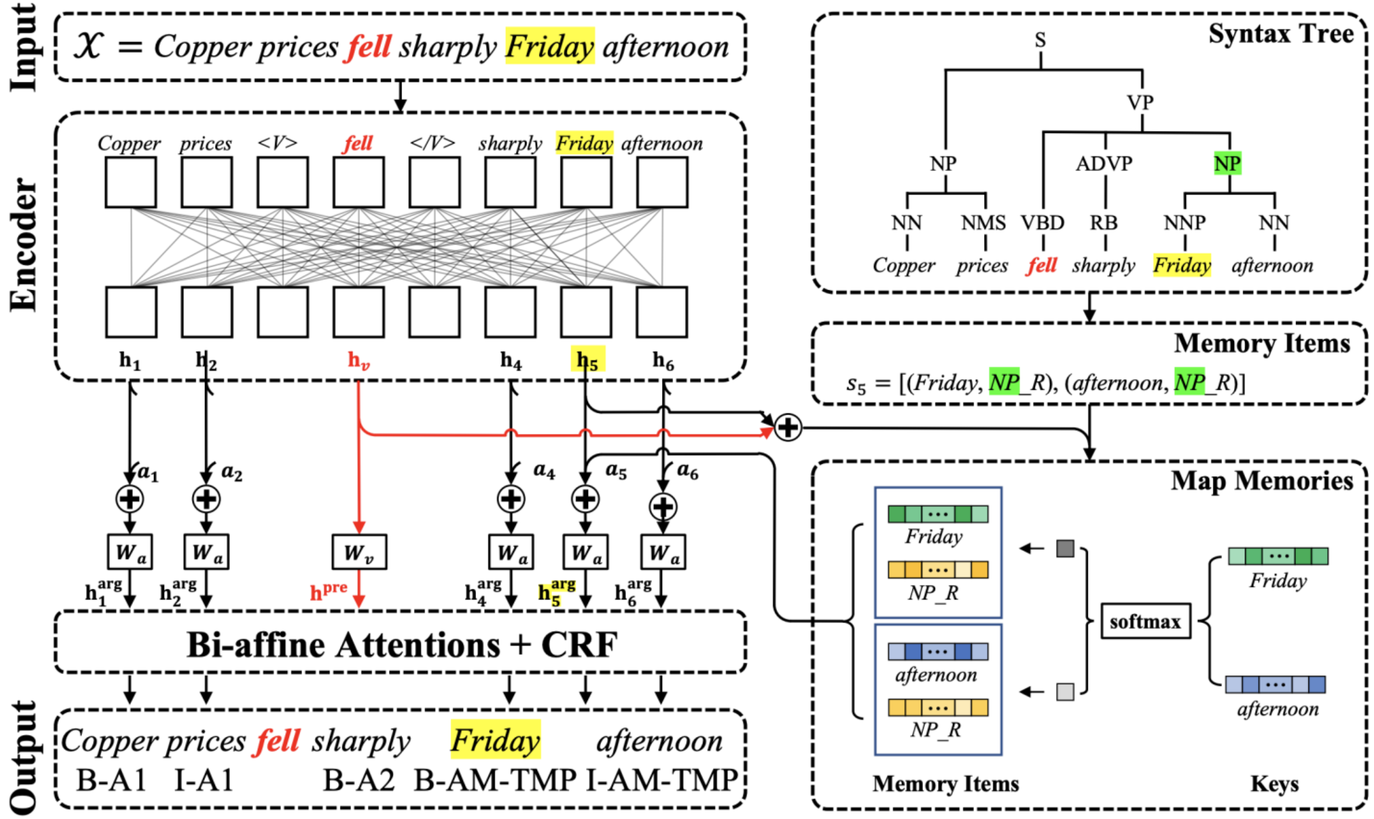
Knowledge-aware Pronoun coreference resolution. (ACL, 2019)

语义角色标注

- Identify the semantic roles played by each constituent in a sentence

他	买	书
施事	谓语	受事

- Two types of SRL tasks
  - Span-style
  - Dependency style
- Challenges:
  - Implicit arguments
  - Multilingual SRL


Input
$\mathcal{X}$ = Copper prices fell sharply Friday afternoon
Encoder
Copper prices <V> fell </V> sharply Friday afternoon
h_1 h_2 h_v h_4 h_5 h_6
a_1 a_2 a_4 a_5 a_6
W_a W_v
h_1^{arg} h_2^{arg} h^{pre} h_4^{arg} h_5^{arg} h_6^{arg}
Bi-affine Attentions + CRF
Output
Copper prices fell sharply Friday afternoon
B-A1 I-A1 B-A2 B-AM-TMP I-AM-TMP
Syntax Tree
S
VP
NP
ADVP
NP
NN NMS VBD RB NNP NN
Copper prices fell sharply Friday afternoon
Memory Items
s_5 = [(Friday, NP_R), (afternoon, NP_R)]
Map Memories
Friday
NP_R
afternoon
NP_R
softmax
Friday
afternoon
Memory Items
Keys

Syntax-driven Approach for Semantic Role Labeling (LREC, 2022)

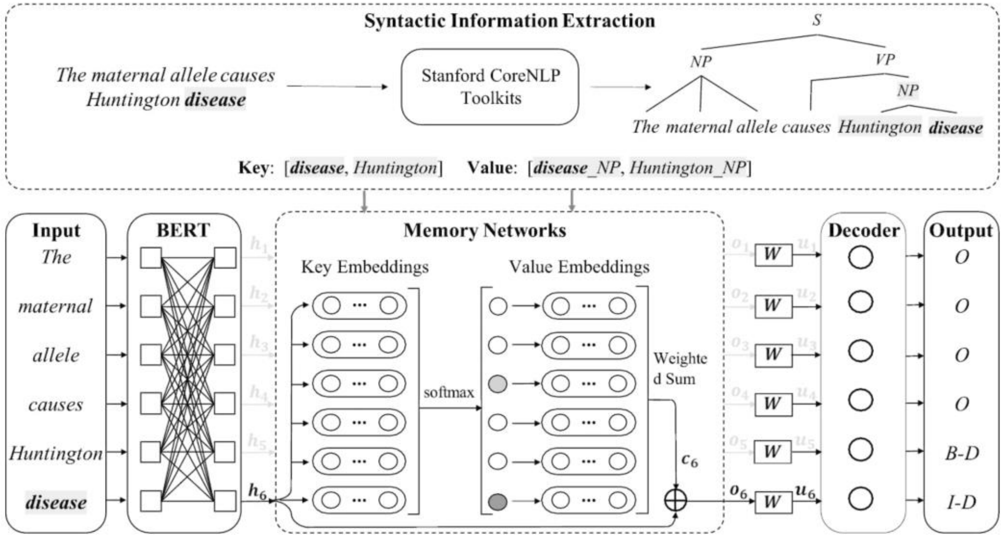
实体识别

- Recognize the named entities in the text

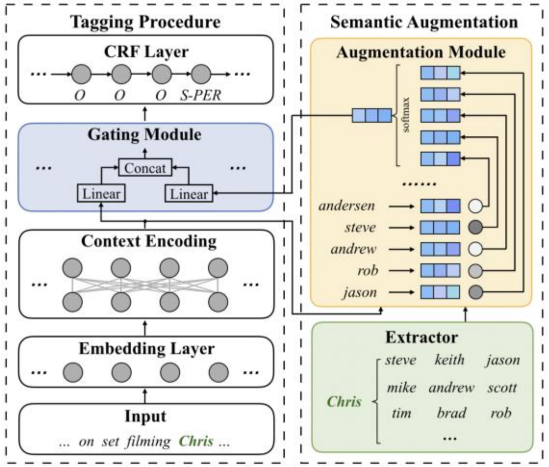
  张三 在 法国 旅游

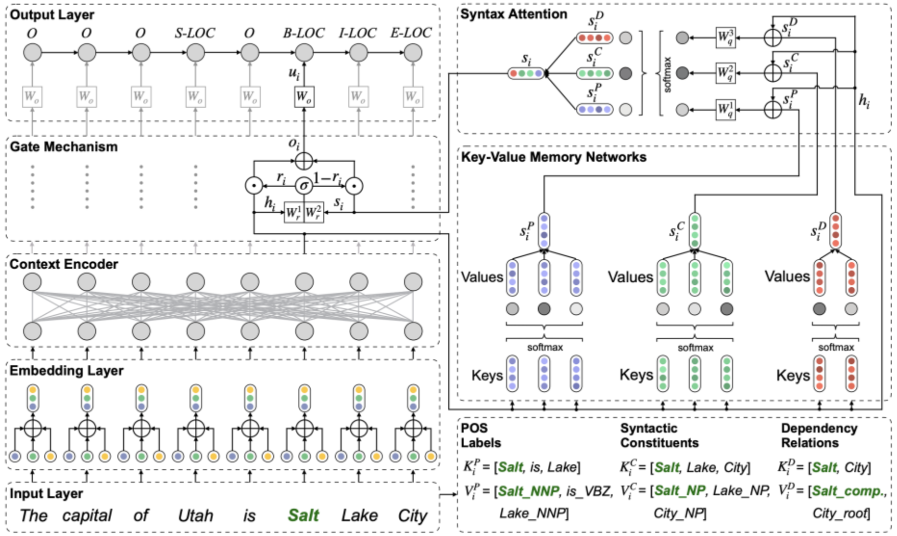
  [人名] [地名]

- Generally performed as a sequence labeling task
- Challenges
  - Overlapping entities
  - Novel entities



Syntactic Information Extraction
The maternal allele causes Huntington disease
Stanford CoreNLP Toolkits
S
NP
VP
NP
The maternal allele causes Huntington disease
Key: [disease, Huntington]
Value: [disease_NP, Huntington_NP]
Input
The
maternal
allele
causes
Huntington
disease
BERT
Memory Networks
Key Embeddings
Value Embeddings
softmax
Weighted Sum
h_1
h_6
c_6
o_6
u_6
W
Decoder
Output
O
O
O
O
B-D
I-D





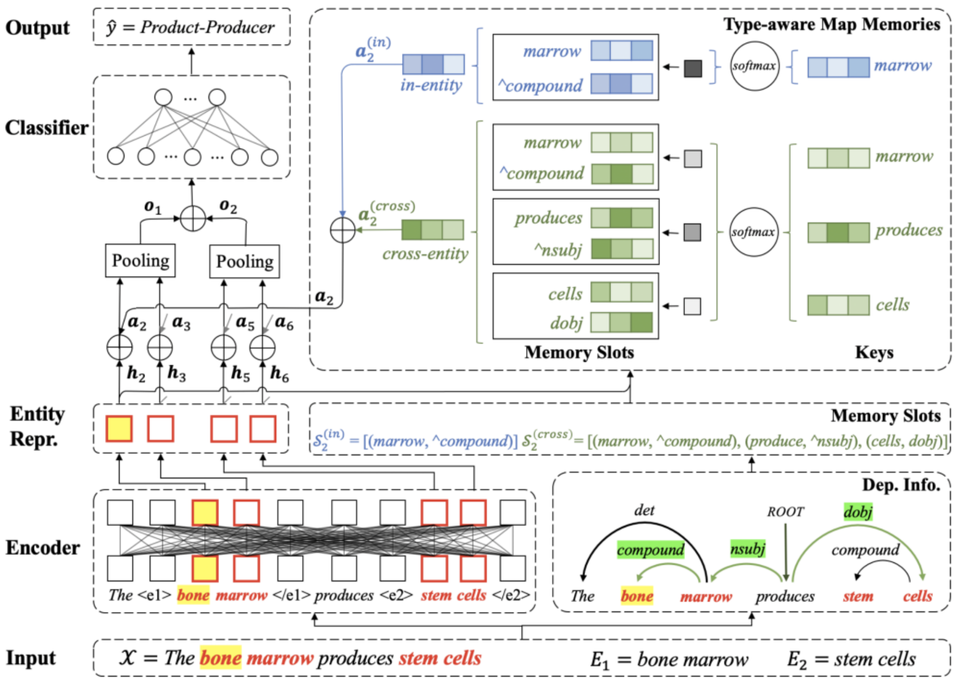
Improving biomedical named entity recognition with syntactic information (BMC Bioinformatics, 2020)
Named Entity Recognition for Social Media Texts with Semantic Augmentation (EMNLP, 2020)
Improving Named Entity Recognition with Attentive Ensemble of Syntactic Information (EMNLP: Findings, 2020)

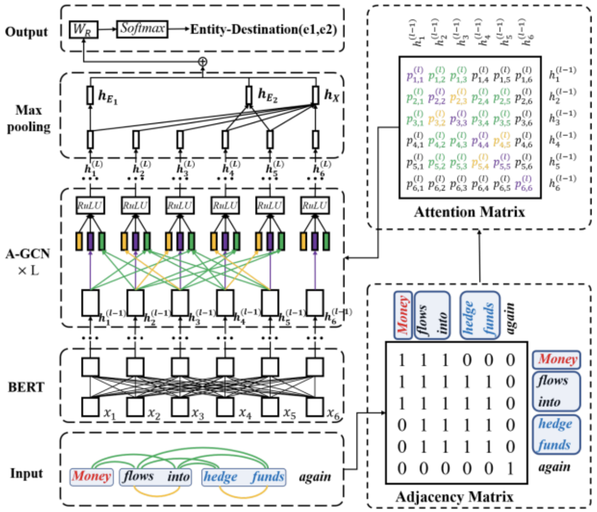
关系抽取

- Predict the relation between two entities

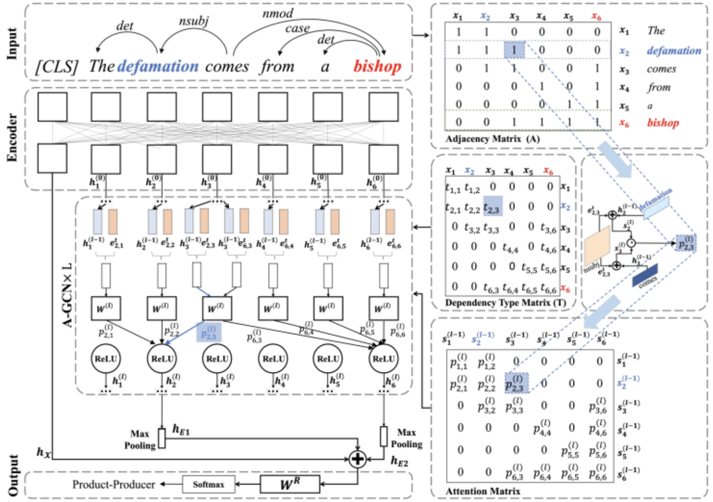
  工厂生产零件

  （工厂，零件，生产者-产品）
- Classification task on entity pairs
- Challenges
  - Unseen entities
  - Ambiguity





Output
W_R
Softmax
Entity-Destination(e1,e2)
Max pooling
h_{E_1} h_{E_2} $h_\mathcal{X}$
A-GCN × L
ReLU
BERT
Input
Money flows into hedge funds again
Attention Matrix
Adjacency Matrix

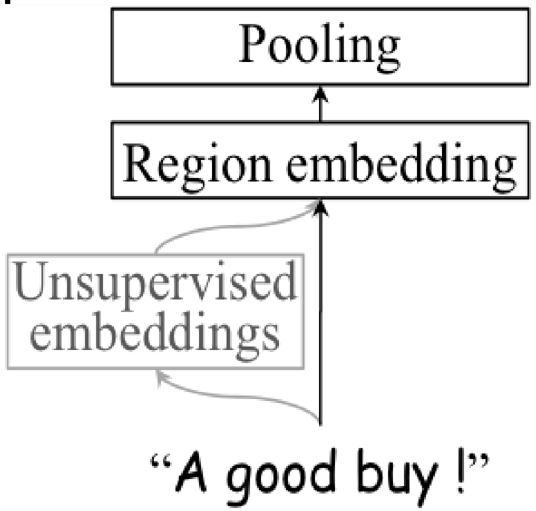


Relation Extraction with Type-aware Map Memories of Word Dependencies (ACL: Findings, 2021)
Dependency-driven Relation Extraction with Attentive Graph Convolutional Networks (ACL, 2021)
Relation Extraction with Word Graphs from N-grams (EMNLP, 2021)

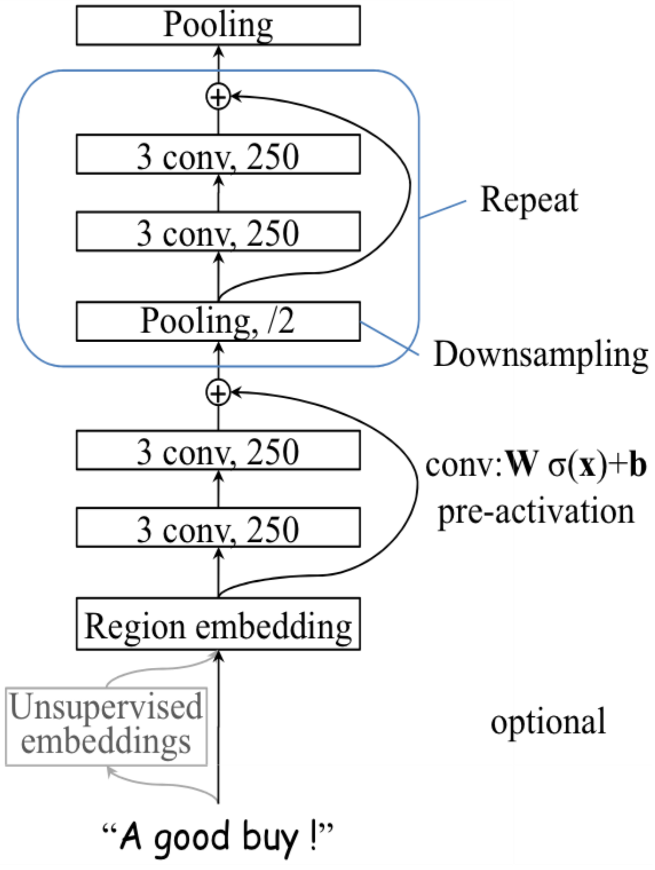
文本分类

- Using different neural models
  - DNN
  - CNN
  - LSTM

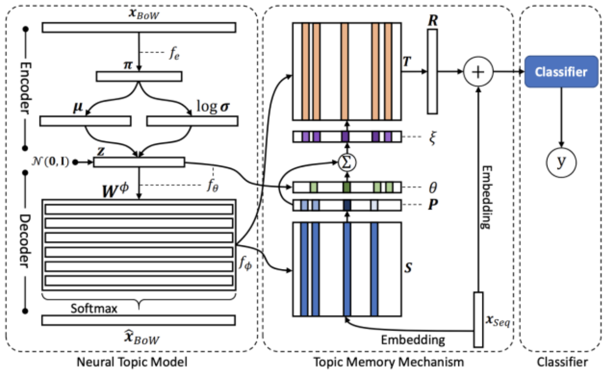
Linear Model



Effective Use of Word Order for Text Categorization with Convolutional Neural Networks (NAACL, 2015)



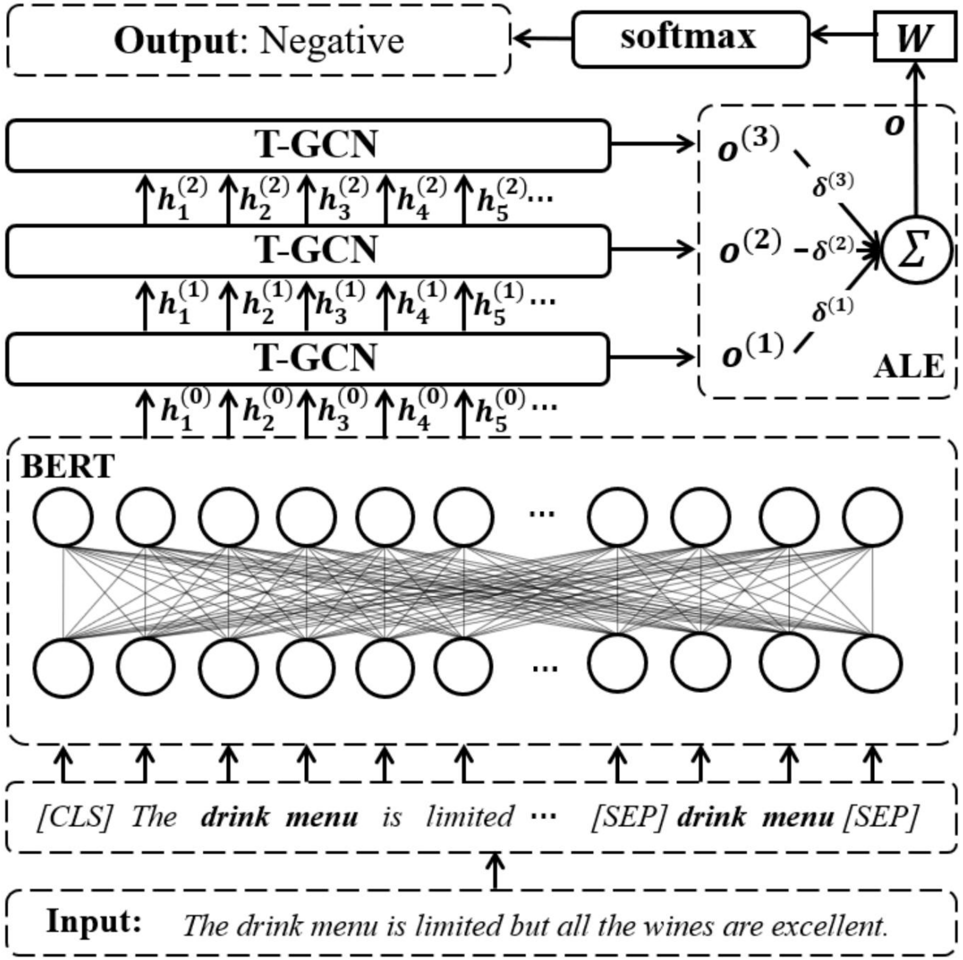
Deep Pyramid Convolutional Neural Networks for Text Categorization (ACL, 2017)



Topic Memory Networks for Short Text Classification (EMNLP, 2018)

情感分析

- A classification task
  - Any text categorization model can apply
  - So as deep learning models
- Early solutions use sentiment vocabulary
  - Limited coverage
  - No dependencies considered among words
- Current trend
  - Fine-grained and mixed granularity
  - Aspect-based and joint task

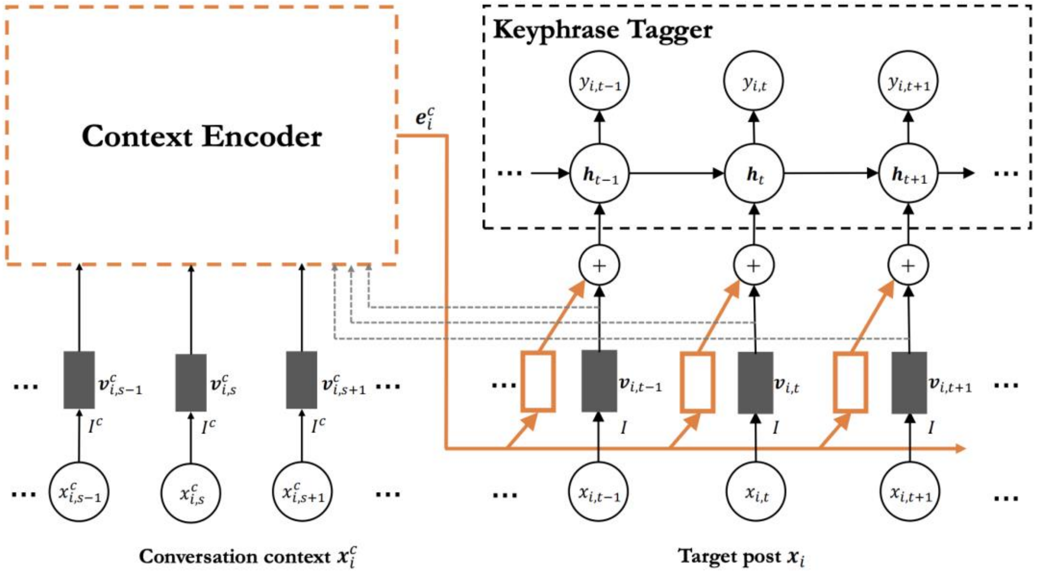


Output: Negative
softmax
W
o
T-GCN
$h_1^{(2)}$ $h_2^{(2)}$ $h_3^{(2)}$ $h_4^{(2)}$ $h_5^{(2)}$ ...
T-GCN
$h_1^{(1)}$ $h_2^{(1)}$ $h_3^{(1)}$ $h_4^{(1)}$ $h_5^{(1)}$ ...
T-GCN
$h_1^{(0)}$ $h_2^{(0)}$ $h_3^{(0)}$ $h_4^{(0)}$ $h_5^{(0)}$ ...
$o^{(3)}$
$\delta^{(3)}$
$o^{(2)}$
$\delta^{(2)}$
Σ
$\delta^{(1)}$
$o^{(1)}$
ALE
BERT
...
...
[CLS] The **drink menu** is limited ··· [SEP] **drink menu** [SEP]
Input: The drink menu is limited but all the wines are excellent.

Aspect-based Sentiment Analysis with Type-aware Graph Convolutional Networks and Layer Ensemble (NAACL, 2021)

关键词抽取和生成

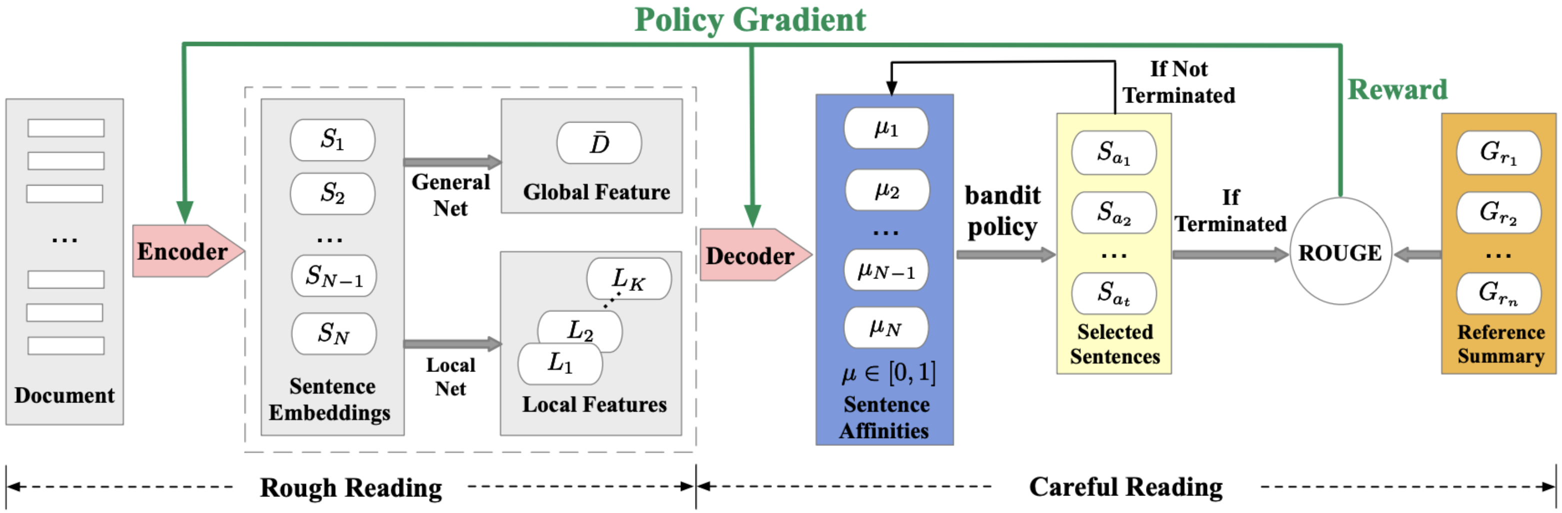
- Extraction: Extracting words to form keyphrases
  - Sequence labeling approach
  - Give each word a tag
- Generation: Generating keyphrases that are not present in the text
  - Encoding-decoding paradigm
  - One keyphrase per time or a keyphrase sequence



Encoding Conversation Context for Neural Keyphrase Extraction from Microblog Posts (NAACL, 2018)

文本摘要

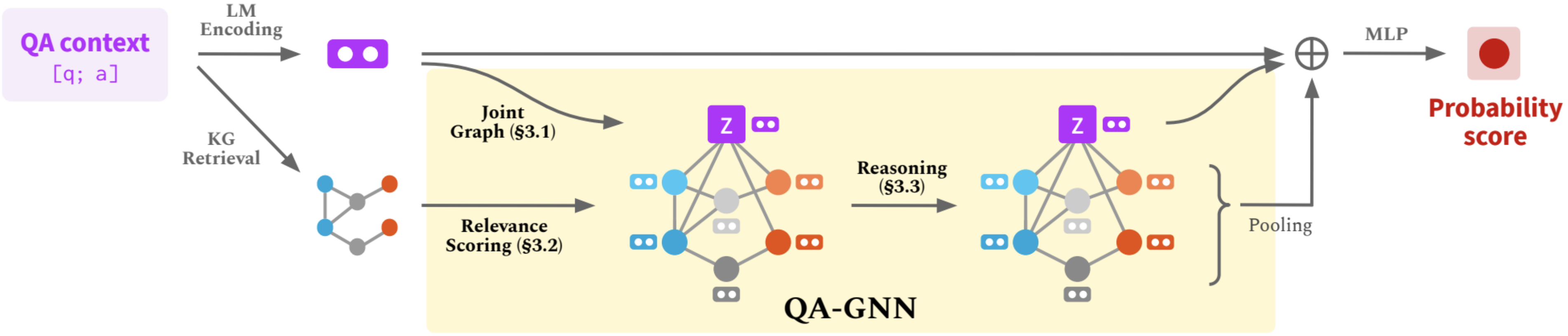
- Using a short text to describe a long text (sentences/document)
  - Extractive
  - Generative
- Neural models
  - Encoding-decoding
  - With knowledge
  - Require to model sentence relations



Reading Like HER: Human Reading Inspired Extractive Summarization (EMNLP, 2019)
Summarizing Medical Conversations via Identifying Important Utterances (COLING, 2020)

问答

- Generally has different forms:
  - Select the best answer from a candidate list
  - Open domain QA that requires to generate answers
  - Etc.
- Require knowledge to obtain plausible answers



QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering (NAACL 2021)
ChiMed: A Chinese Medical Corpus for Question Answering (BioNLP, 2020)

计算机编码

- 01编码
- ASCII
- Extended ASCII
- Unicode
  - UTF
  - UTF-8

文本表征

- One-hot表征
- TF-IDF
- 基于神经网络的表征
- 基于词典的表征
  - WordNet
  - FrameNet
  - HowNet
- 评价

预备知识

- 先验概率：$P(A)$
- 联合概率：$P(A,B)$
- 条件概率： $P(A|B)$
- 条件概率的链式法则：
  $$P(A,B,C,D,\ldots) = P(A)P(B|A)P(C|A,B)P(D|A,B,C\ldots)$$
- 独立性： A,B独立，$P(A,B) = P(A)P(B)$
- 贝叶斯公式
  $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

预备知识

- 随机变量：X
- 随机变量的期望： $E(X)=\sum_x x\cdot p(x)$
- 随机变量的方差：$Var(X)=E(\left(X-E(X)\right)^2)$

- 熵： $H(X)=-\sum_{x\in X}p(x)\log_2 p(x)$
- 联合熵： $H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(X,Y)$
- 条件熵： $H(Y|X)=-\sum_{x\in X}p(x)\sum_{y\in Y}p(y|x)\log(y|x)$
- 互信息：$I(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$