第八章 知识图谱

戴洪良

计算机科学与技术学院/人工智能学院

hongldai@nuaa.edu.cn

从文本中抽取知识

- 命名实体识别
- 细粒度实体分类
- **关系抽取**
- 事件抽取

关系抽取 (Relation Extraction)

3月20日消息，微软旗下语音识别子公司 Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

三元组形式的目标输出：

头实体	关系类别	尾实体
Nuance	*是子公司*	微软
Nuance	*有产品*	DAX Express

- 基于预定义好的关系类别，抽取文本中实体间的关系
- 基本实现流程
  - 识别实体
  - 结合上下文将每组实体对分类

注：有时只执行了分类部分也称“关系抽取”

关系抽取

- 关系抽取数据集
- 基于规则的方法
- 基于Bootstrapping的方法
- 普通机器学习设定的关系抽取
  - 传统机器学习、神经网络、预训练模型
- 其他设定的关系抽取
  - 基于远程监督的关系抽取、基于序列的关系抽取
  - 文档级关系抽取、Open Information Extraction

关系抽取数据集

- TACRED
  - 通用领域新闻和网页文本；含*no_relation*的42种关系类别；TACREV和Re-TACRED是该数据集的两个改进版本
    - *per:country_of_death*, *org:founded_by*, *per:children, no_relation* 等
- SemEval
  - 9种关系类别+*other*，除other外的类别有对应的逆关系；其实体为如“princess”，“island”的普通名词
    - *Cause-Effect*, *Component-Whole*, *Component-Whole-R* 等
- FewRel
  - 用于小样本（Few-shot）关系分类的数据集；维基百科文本；100种关系类别
    - *director*, *family*, *capital of* 等
- NYT10
  - 远程监督数据集；New York Times新闻文本；58种关系类别
    - /location/location/contains， /people/person/nationality， /people/person/place_of_birth等
- 其他：
  - 如ACE 2005, 中文数据集FinRE, SanWen等

关系抽取数据集

- TACRED

关系	样本
per:title	Former Jerusalem [Mayor] [Teddy Kollek] , who balanced needs of diverse people , dies at 95.
org:founded_by	Talansky is also the US contact for the [New Jerusalem Foundation] , an organization founded by [Olmert] while he was Jerusalem 's mayor .
per:countries_of_residence	Born in London in 1939 the son of a Greek tycoon , [Negroponte] grew up in Britain , Switzerland and the [United States] .
org:top_members/employees	McCain and his lawyer , former [FEC] chairman [Trevor Potter] , have argued that McCain is entitled to turn down the primary matching funds ...
per:employee_of	[He] was an [Army] veteran of the Korean War .
org:alternate_names	The Commodity Futures Trading Commission settled the charges with [MF Global] , formerly known as [Man Financial] .

关系抽取数据集

- SemEval

关系	样本
Cause-Effect (e2,e1)	Avian [influenza]e1 is an infectious disease caused by type a strains of the influenza [virus]e2
Entity-Origin (e1,e2)	The [mother]e1 left her native [land]e2 about the same time and they were married in that city.
Message-Topic (e2,e1)	Roadside [attractions]e1 are frequently advertised with [billboards]e2 to attract tourists.
Product-Producer (e1,e2)	A child is told a [lie]e1 for several years by their [parents]e2 before he/she realizes that ...
Entity-Destination (e1,e2)	The accident has spread [oil]e1 into the [ocean]e2.
Member-Collection (e2,e1)	The siege started, with a [regiment]e1 of lightly armored [swordsmen]e2 ramming down the gate.
Component-Whole (e2,e1)	The size of a [tree]e1 [crown]e2 is strongly correlated with the growth of the tree.

基于规则的关系抽取

基于规则的关系抽取

- 自然语言在表述两个实体间关系时会有一些常用模式
  - “小明是南航的学生”、“小明，南航的一位学生…”、“小明作为南航的一名学生…”

- 编写或挖掘规则来利用这些模式抽取实体间关系

Hearst Patterns

- Hearst Patterns是一系列用于抽取is-a关系的模板
  - 由学者Marti A. Hearst提出

如：

模板	例
such X as Y	… works by such authors as Herrick, Goldsmith, and Shakespeare.
Y (and|or) other X	Bruises, wounds, broken bones or other injuries …
X including Y	… large countries including Russia, Canada and China …

Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." In COLING 1992 volume 2: The 14th international conference on computational linguistics. 1992.

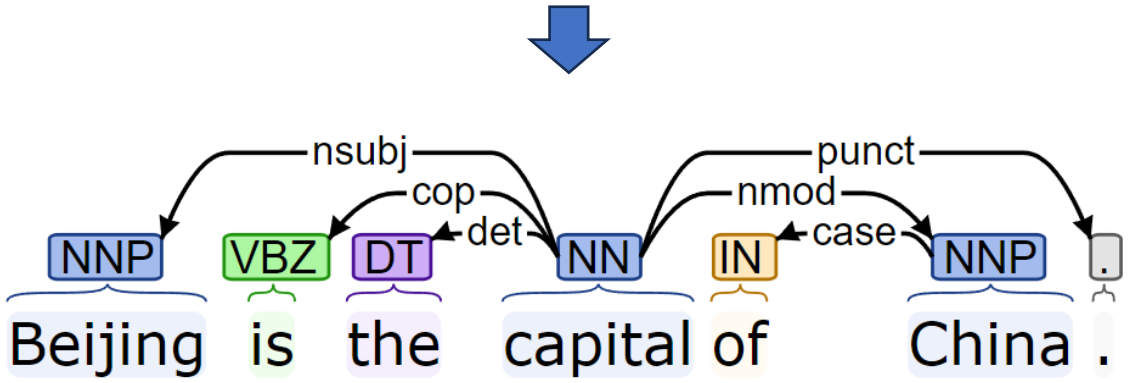
Hearst Patterns

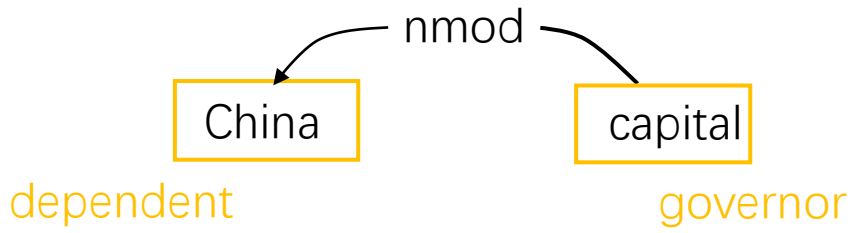
- 形式简单，只匹配词和词性标注（POS tag）
- 只抽取is-a关系
- 准确率（precision）较高，但召回率（recall）低

基于依存句法分析的规则

- 依存句法分析(dependency parsing)将句子解析成由词与词间关系形成的树状结构
  - 其中词与词间的关系是有向的，称作依存关系

Beijing is the capital of China.

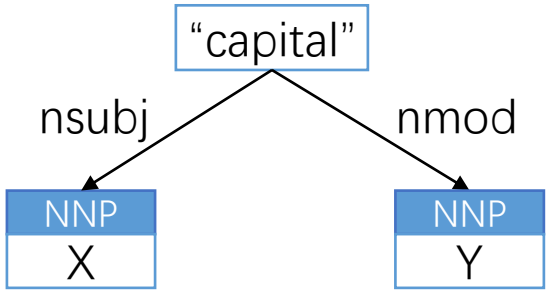




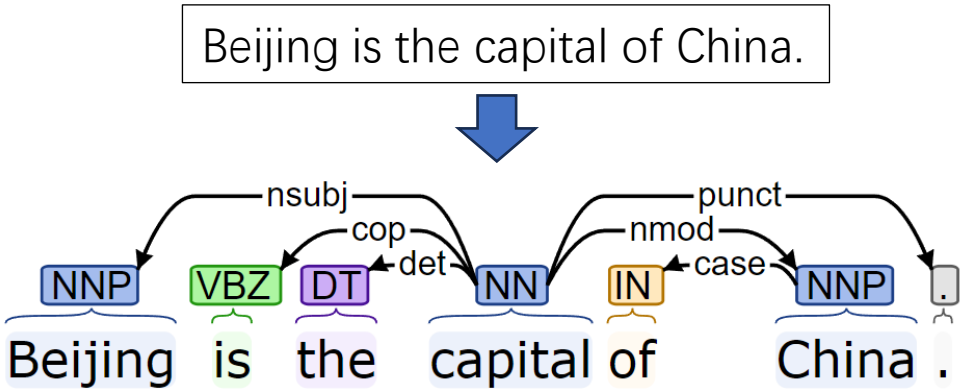
Stanford提供的在线解析工具：https://corenlp.run/

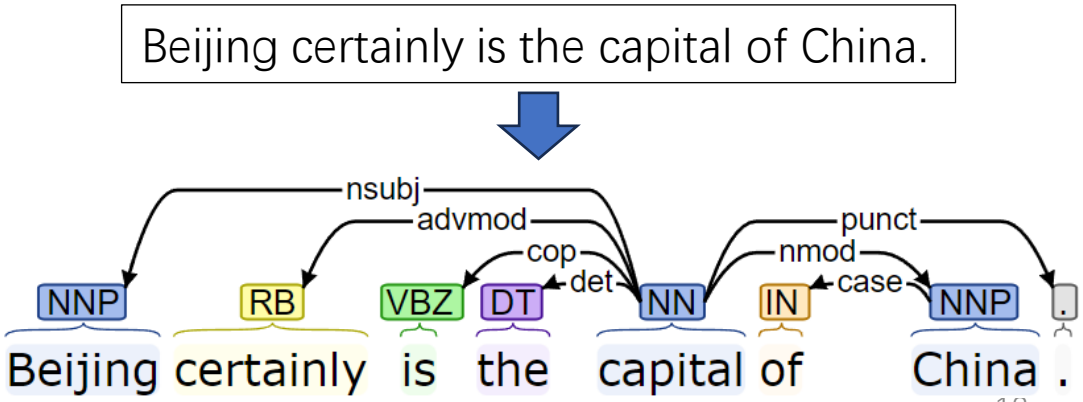
基于依存句法分析的规则

- 例：



规则：匹配以上结构，则X是Y的首都





基于规则的关系抽取

- 优缺点?

基于规则的关系抽取

- 优点
  - 人工编写无需训练数据
  - 人工编写规则可以达到较高的准确率（precision）
  - 结果可解释
- 缺点
  - 召回率（recall）低
  - 编写规则需要对NLP的了解，特定领域可能还需专家知识
  - 如果关系类别多，人工编写规则费时费力；而自动挖掘规则一般又需标注数据
  - 不易维护

基于Bootstrapping的方法

基于Bootstrapping的方法

- 大致流程：
  1. 对于某种实体关系，使用一些已知有该关系的实体对作为种子
  2. 从大量无标注文本句子中匹配这些实体对，从匹配到的句子中挖掘关系抽取模板
  3. 用挖掘到的模板可以找到更多有该关系的实体对，从中选取一些作为新的种子
  4. 迭代上述过程

Agichtein, Eugene, and Luis Gravano. "Snowball: Extracting relations from large plain-text collections." Proceedings of the fifth ACM conference on Digital libraries. 2000.

基于Bootstrapping的方法

- 例：考虑*公司总部*关系

1. 初始时作为种子的有该关系的实体对：

<Microsoft, Redmond>, <IBM, Armonk>, <Intel, Santa Clara>

2. 在大量的无标注文本数据中，找出同时出现上述实体对中两个实体的句子：

… computer servers at Microsoft's headquarter in Redmond …
Santa Clara based Intel Corporation has decided to purchase another …
… …

3. 从找出的句子中，挖掘频繁出现的模式，形成模板：

<ORGANIZATION>'s headquarter in <LOCATION>
<LOCATION> based <ORGANIZATION>
… …

4. 使用模板可以从文本数据中抽取出更多满足该关系的实体对：

<Google, Mountain View>, <Apple, Cupertino>, <Tencent, Shenzhen>, …

5. 从中选出确信度高的作为新种子，重复上述步骤

语义漂移（Semantic Drift）问题

- 找出的模板表达的语义关系可能与目标关系不同
  - 如目标关系是mayor_of，但找出了模板<Person> lives in <Location>

- 缓解该问题的方式：
  - 限制迭代次数
  - 为找出的模板和新的实体对打分，选分数高的进行迭代

语义漂移（Semantic Drift）问题

- 为找出的模板和新的实体对打分，选分数高的进行迭代

例：

模板P的分数： $$Conf(P) = \frac{P.positive}{(P.positive + P.negative)}$$

P.positive: 匹配到的正确实体对数
P.negative: 匹配到的错误实体对数

实体对T的分数： $$Conf(T) = 1 - \prod_{i=0}^{|P|} (1 - (Conf(P_i) \cdot Match(C_i, P_i)))$$

Match(C, P): 实体对所在文本内容与模板P的匹配程度

Agichtein, Eugene, and Luis Gravano. "Snowball: Extracting relations from large plain-text collections." Proceedings of the fifth ACM conference on Digital libraries. 2000.

基于Bootstrapping的方法

- 更适合从大量文本数据找出符合关系的实体，而不适合只给定一段文本识别确定其中两个实体的关系
  - 大量文本：有若个段文本能匹配上找出的模板就行
  - 对给定的一段文本，很可能抽取出的模板匹配不上
- 效果很大程度上取决于关系类别

普通机器学习设定下的关系抽取

- 常规的train/dev/test集划分
- 数据是人工标注的正确样本
- 假设实体已识别好，只需考虑对实体对的关系分类

基于传统机器学习的关系抽取

- 手动设计特征后使用分类模型
  - 如SVM, 最大熵模型（Maximum Entropy Model）等

特征例：

Words
- bag-of-words in M1
- head word of M1
- bag-of-words in M2
- head word of M2

Entity type
- combination of mention entity types

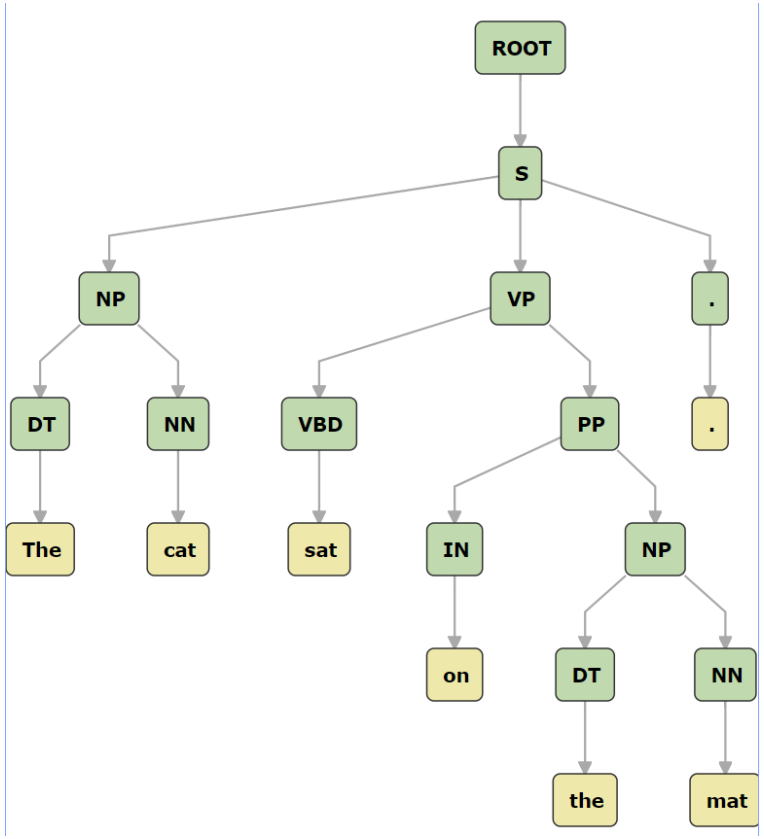
Dependency
- word on which M1 is dependent on
- word on which M2 is dependent on

Kambhatla, Nanda. "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction." In Proceedings of the ACL interactive poster and demonstration sessions, pp. 178-181. 2004.

基于神经网络模型的关系抽取

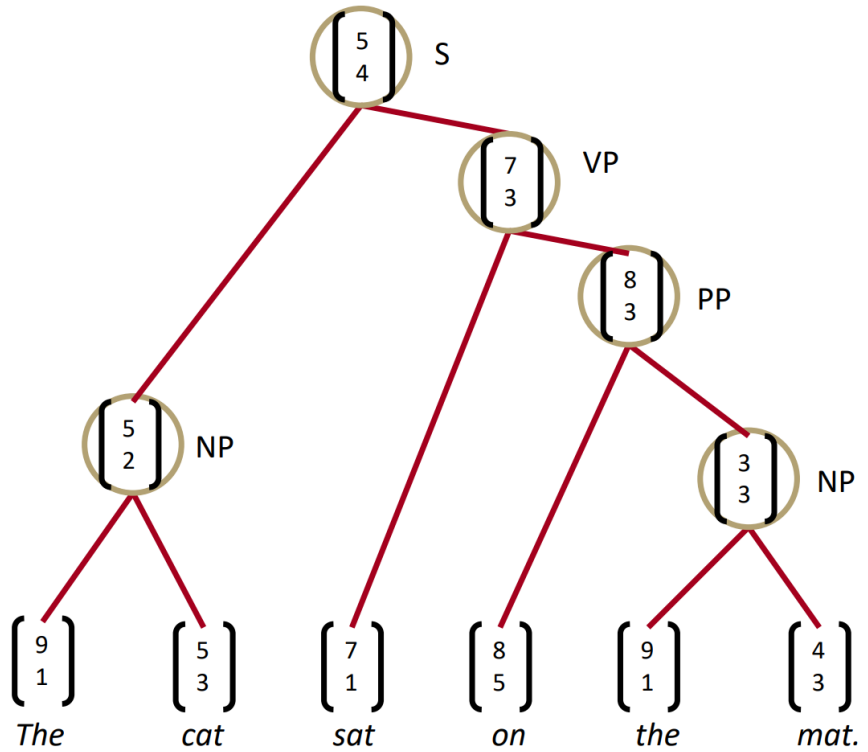
- Recursive Neural Network (RNN)
- 自然语言句子可以被组织成树状结构

对 “The cat sat on the mat”进行Constituency Parsing的结果：



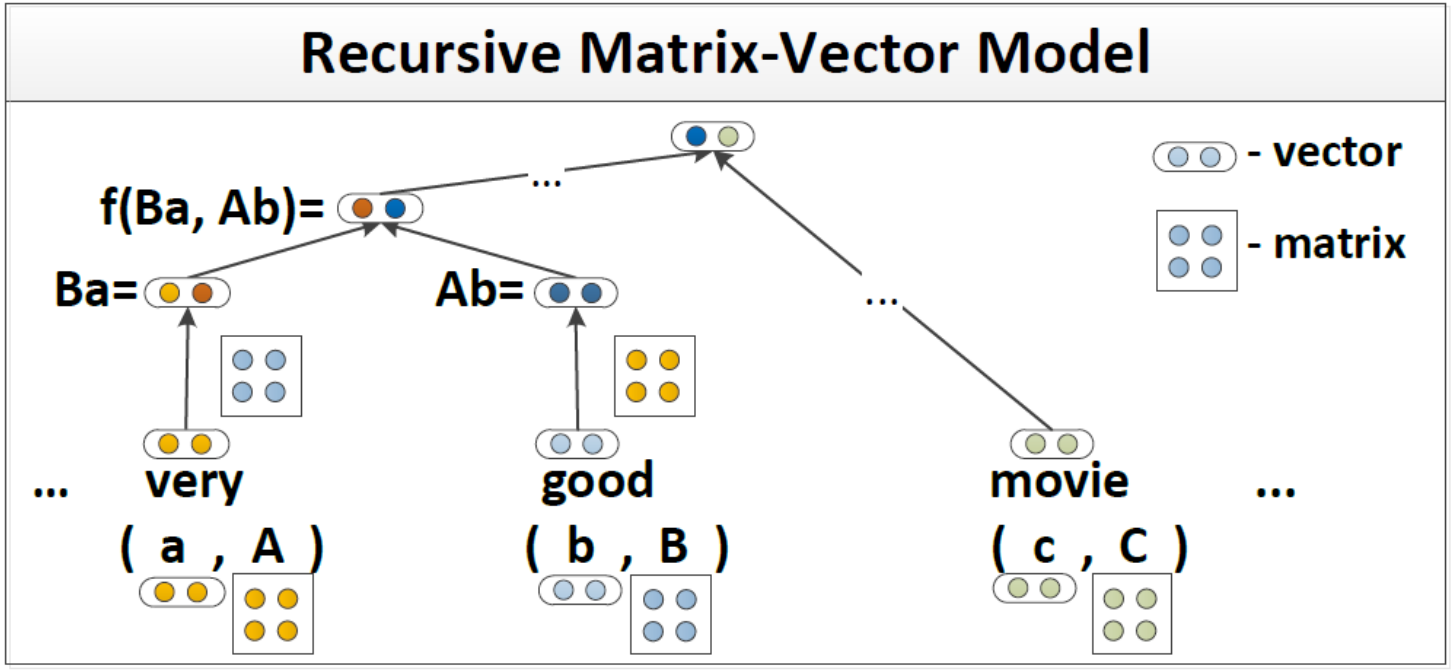
基于神经网络模型的关系抽取

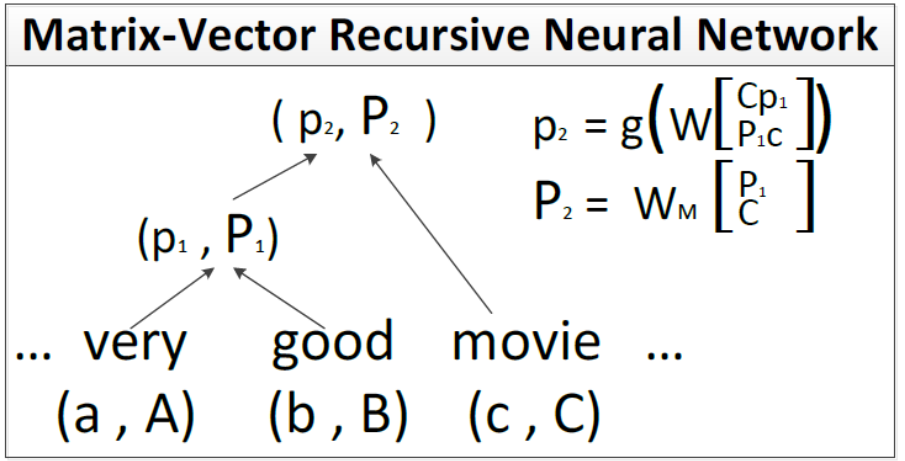
- Recursive Neural Network能够基于将句子解析成的树状结构获得句子和句子中短语的向量表示



基于神经网络模型的关系抽取

- Matrix-vector recursive neural network (MV-RNN)





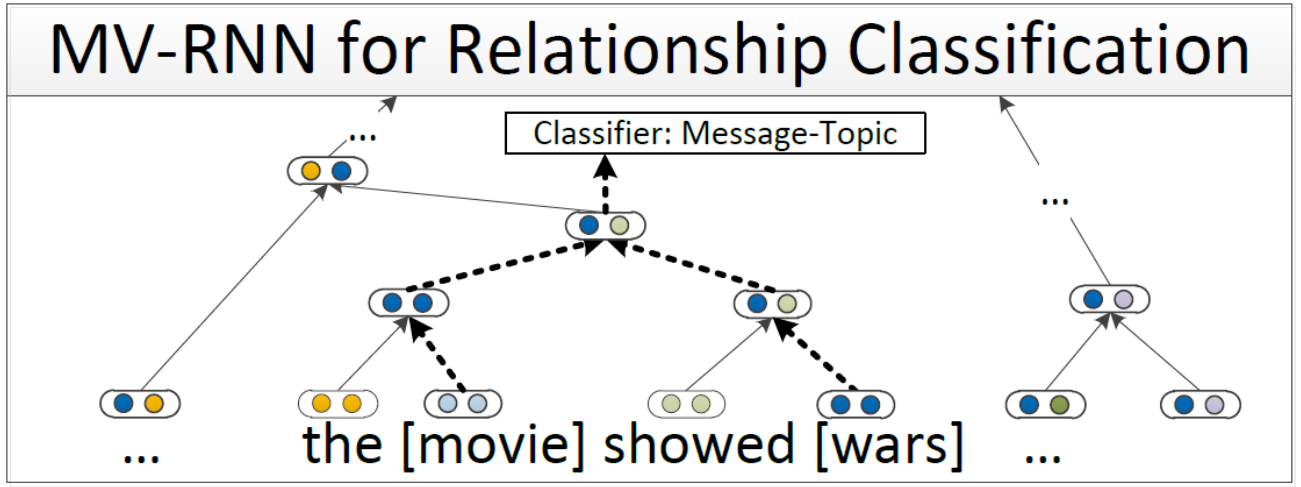
$$f_{A,B}(a,b) = f(Ba, Ab) = g\left(W \begin{bmatrix} Ba \\ Ab \end{bmatrix}\right)$$

$$P = f_M(A,B) = W_M \begin{bmatrix} A \\ B \end{bmatrix}$$

每个子节点（词）赋予一个向量和一个矩阵
每个中间节点通过其子节点计算出一个向量和一个矩阵

基于神经网络模型的关系抽取

- Matrix-vector recursive neural network (MV-RNN)



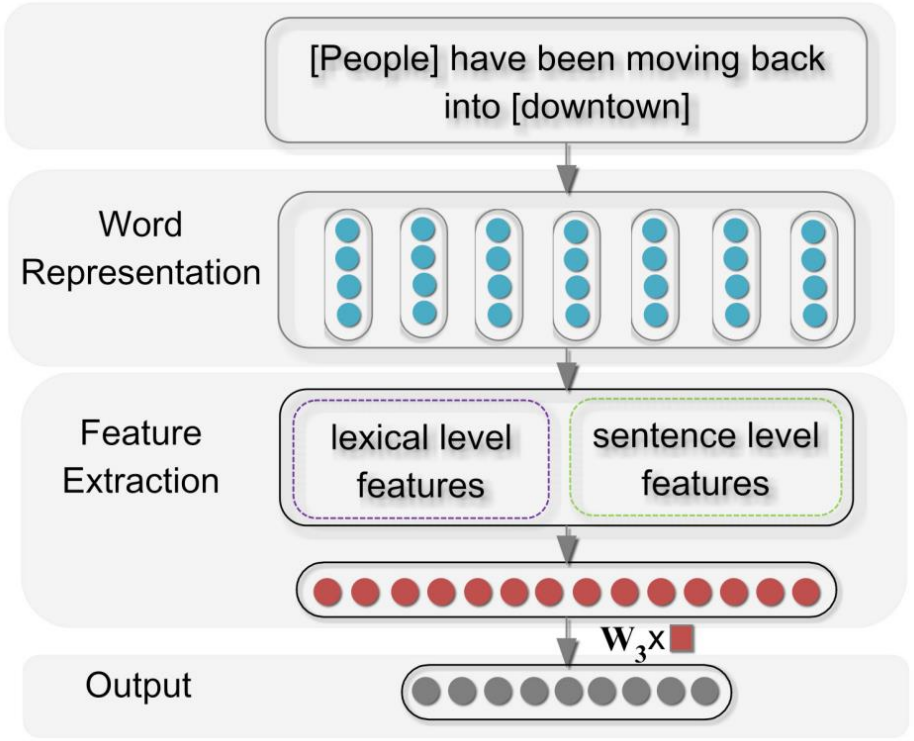
关系类别预测： $d(p) = \mathrm{softmax}(W^{label}p)$

训练loss：cross entropy

实验效果：SemEval数据集上F1为82.4

基于神经网络模型的关系抽取

- 基于卷积神经网络（CNN）的方法

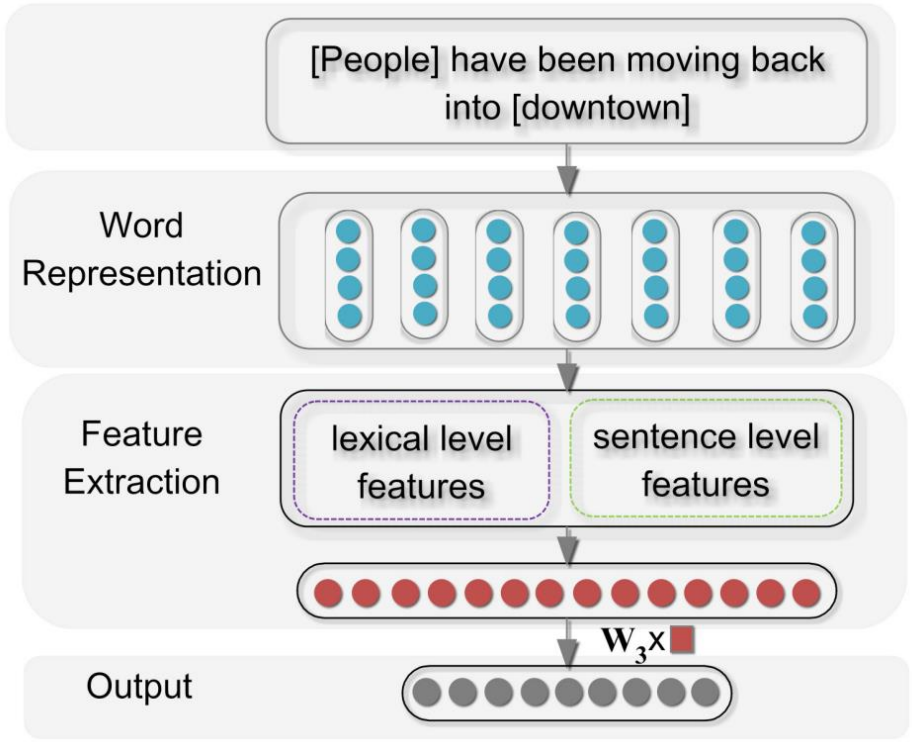


Lexical level features: 基于实体对的词级特征
Sentence level features: 为整个句子获得的特征

Zeng, Daojian, et al. "Relation classification via convolutional deep neural network." Proceedings of COLING. 2014.

基于神经网络模型的关系抽取

- 基于卷积神经网络（CNN）的方法



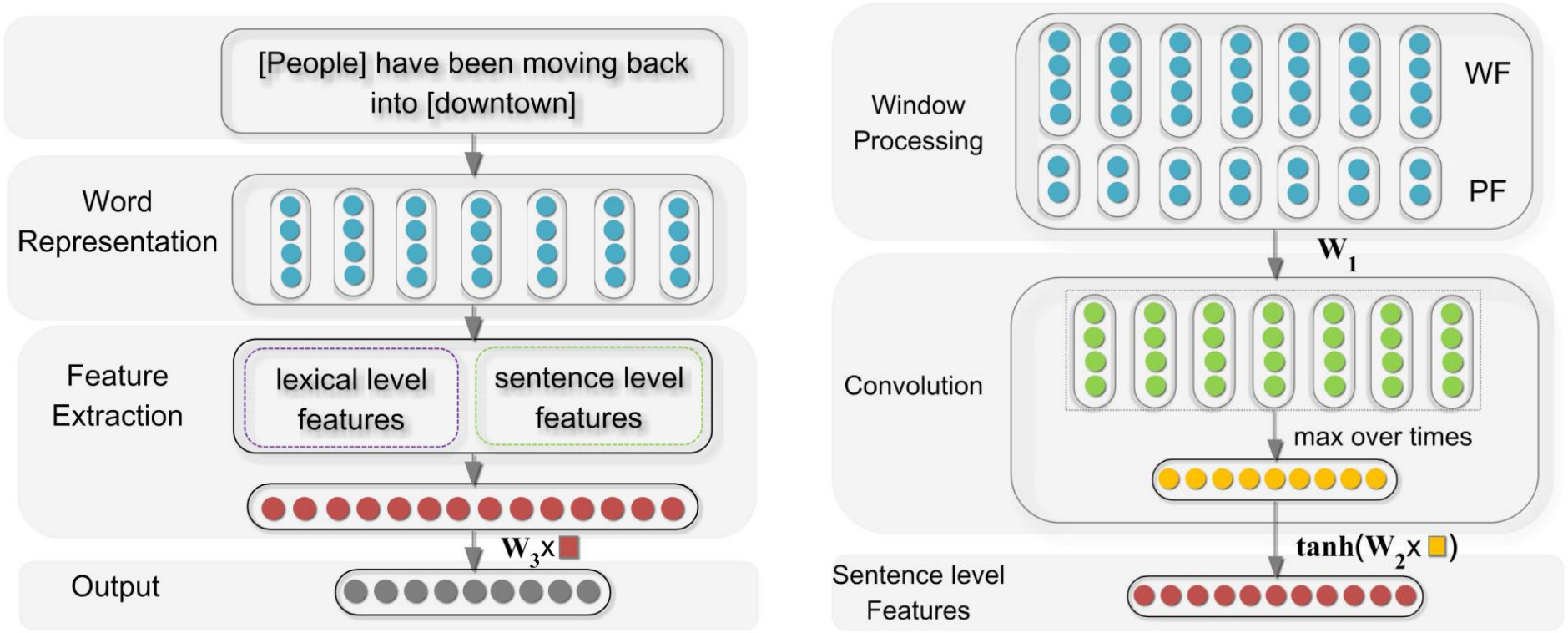
Features	Remark
L1	Noun 1
L2	Noun 2
L3	Left and right tokens of noun 1
L4	Left and right tokens of noun 2
L5	WordNet hypernyms of nouns

Table 1: Lexical level features.

Noun 1, Noun 2: 两个实体名词

基于神经网络模型的关系抽取

- 基于卷积神经网络（CNN）的方法

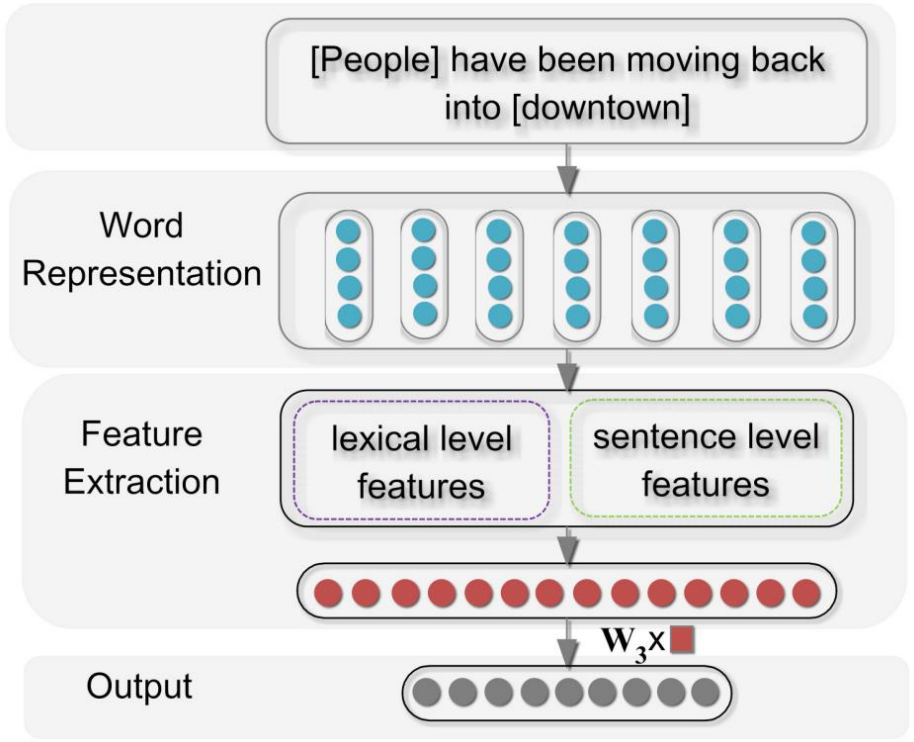


WF: 词向量
PF: 位置向量

Zeng, Daojian, et al. "Relation classification via convolutional deep neural network." Proceedings of COLING. 2014.

基于神经网络模型的关系抽取

- 基于卷积神经网络（CNN）的方法



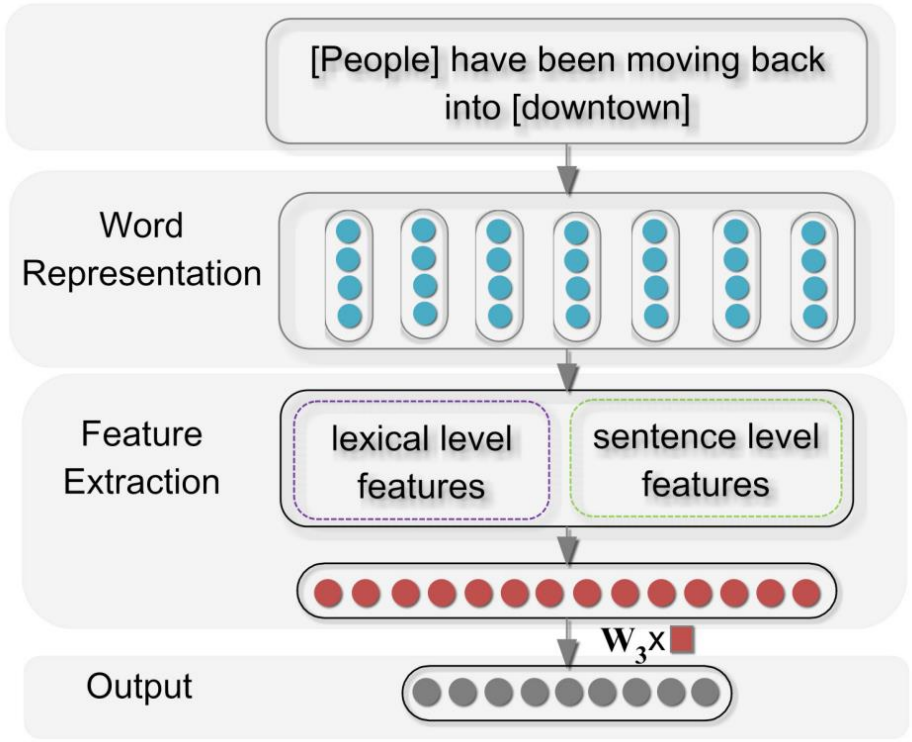
训练目标：

$$J(\theta) = \sum_{i=1}^{T} \log p(y^{(i)}|x^{(i)}, \theta)$$

其中$p(y|x,\theta)$是对模型输出向量做softmax得到的关系类别预测概率

基于神经网络模型的关系抽取

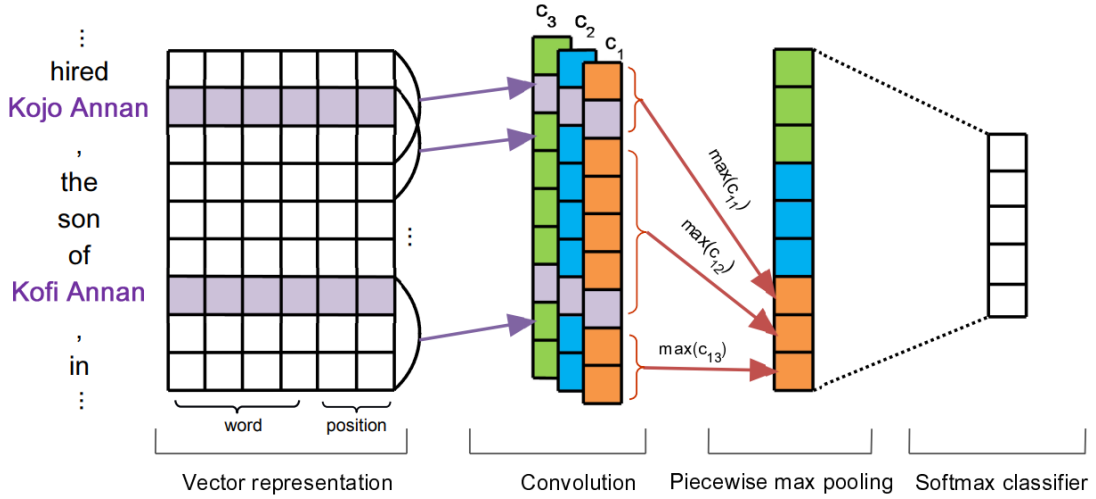
- 基于卷积神经网络（CNN）的方法



方法	F1 on SemEval
MV-RNN	82.4
CNN	82.7

基于神经网络模型的关系抽取

- 基于卷积神经网络（CNN）的方法
- Piecewise CNN
  - 普通CNN在进行pooling时，将用一个filter得到的向量直接变为一个标量值
  - 结果：无法区分信息来自两实体外的上下文还是两实体间的上下文，导致无法捕捉到一些结构信息
  - Piecewise CNN将卷积结果分成三段分别pooling



两个实体将文本分为三段

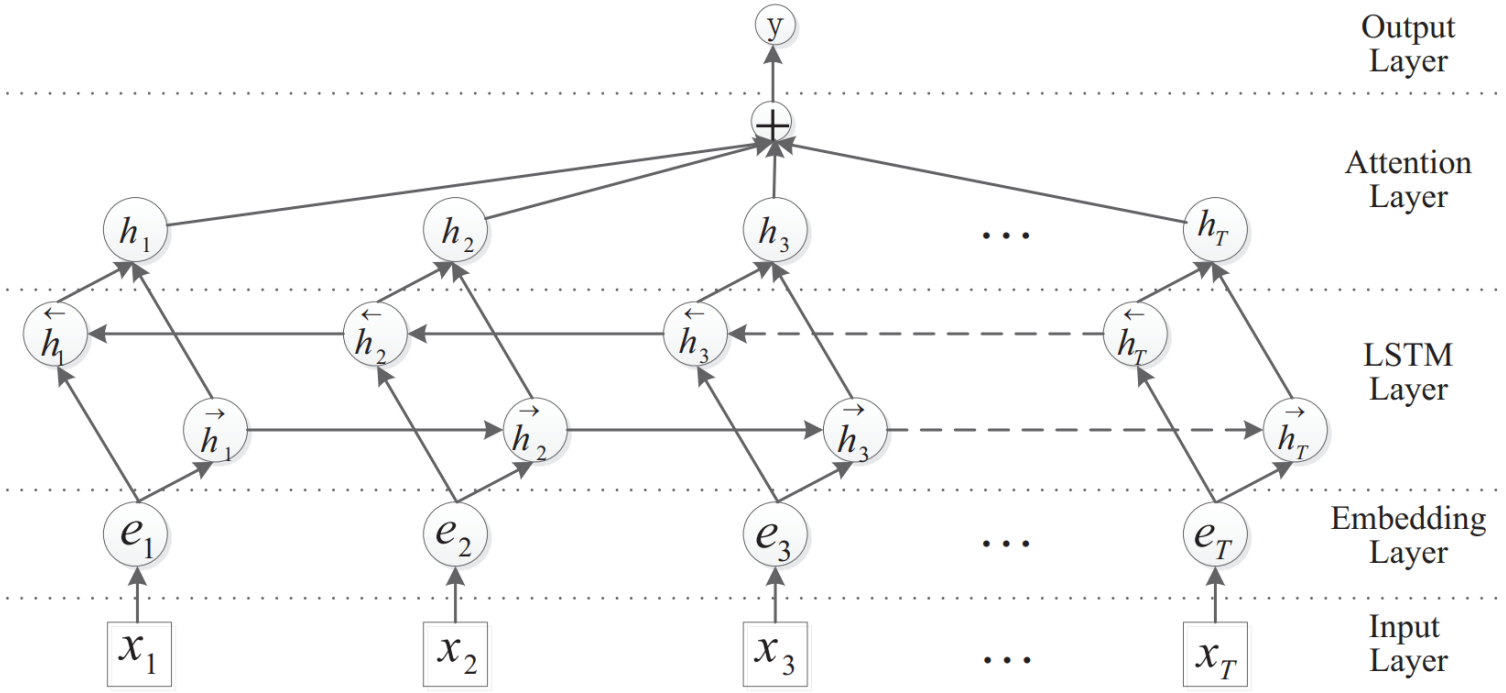
Zeng, Daojian, et al. "Distant supervision for relation extraction via piecewise convolutional neural networks." Proceedings of EMNLP. 2015.

基于神经网络模型的关系抽取

- 基于LSTM和Attention（注意力）的方法 (Zhou et al., 2016)

先用特殊词标记出两个实体的位置：

⟨e1⟩ Flowers ⟨/e1⟩ are carried into the ⟨e2⟩ chapel ⟨/e2⟩ .

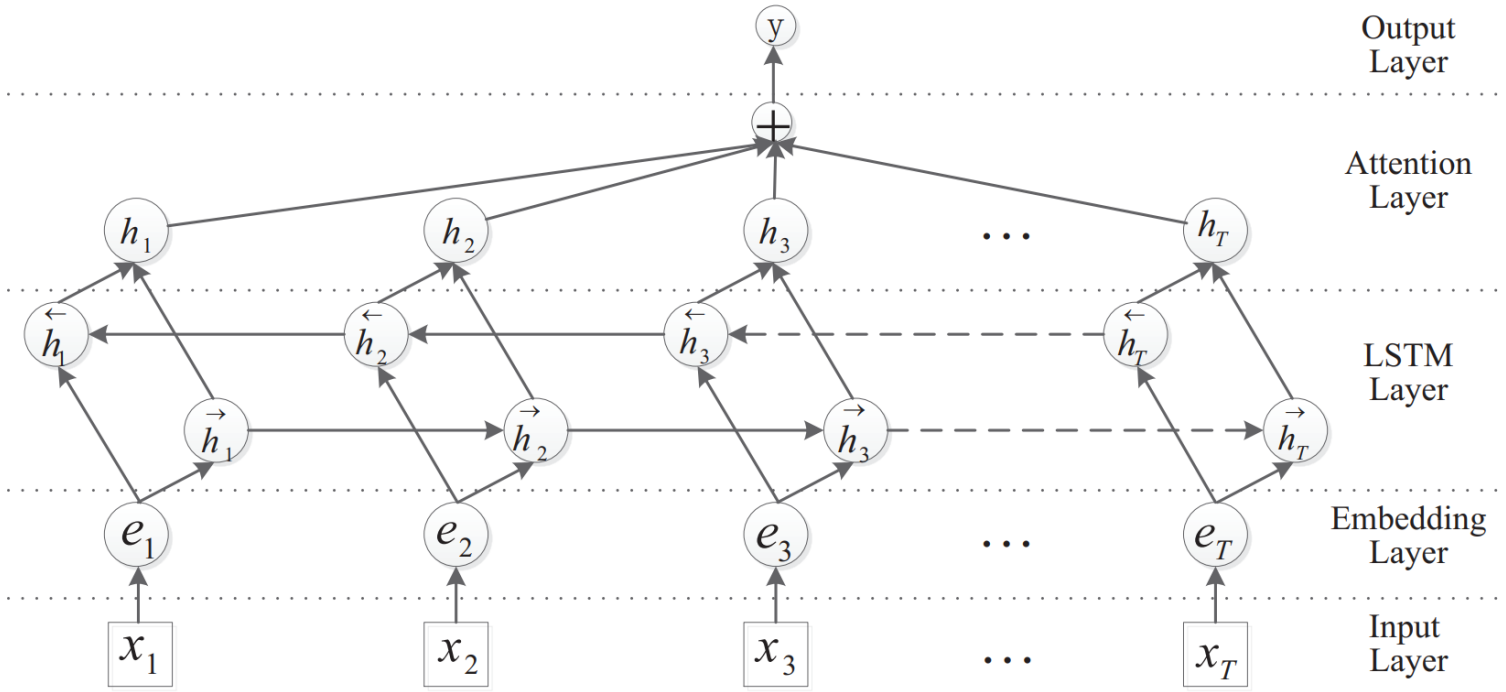


基于神经网络模型的关系抽取

- 基于LSTM和Attention（注意力）的方法
- 模型一共包括5层结构:
  - 输入层: 将句子输入到模型中
  - Embedding层: 将每个词映射到低维空间
  - LSTM层: 使用双向LSTM从Embedding层获取高级特征
  - Attention层: 生成一个权重向量，通过与这个权重向量相乘，使每一次迭代中的词汇级的特征合并为句子级的特征
  - 输出层: 将句子级的特征向量用于关系分类，使用sofmax得到类别分布

基于神经网络模型的关系抽取

- 基于LSTM和Attention（注意力）的方法



Attention Layer计算：

$$M = \tanh(H)$$

$$\alpha = softmax(w^T M)$$

$$r = H\alpha^T$$

其中$H = [h_1, h_2, \ldots h_T]$为LSTM输出

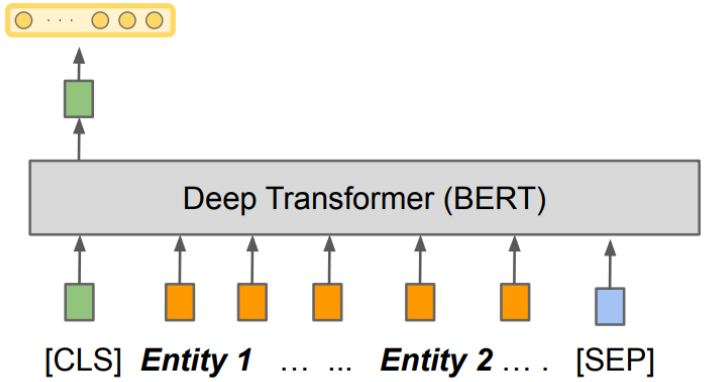
基于神经网络模型的关系抽取

- 基于LSTM和Attention（注意力）的方法

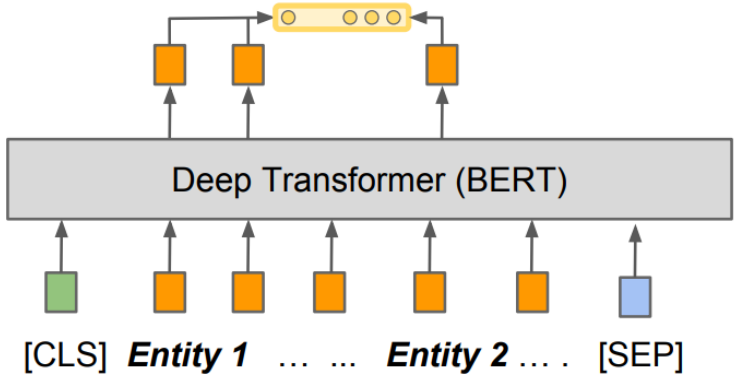
方法	F1 on SemEval
MV-RNN	82.4
CNN	82.7
Bi-LSTM + Attention	84.0

基于预训练模型的方法

- 基于BERT的方法
  - 思路：基于输入用BERT得到一个用于关系分类的向量表示
  - 基于BERT模型，可以有多种不同的方式得到用于关系分类的向量表示



STANDARD – [CLS]

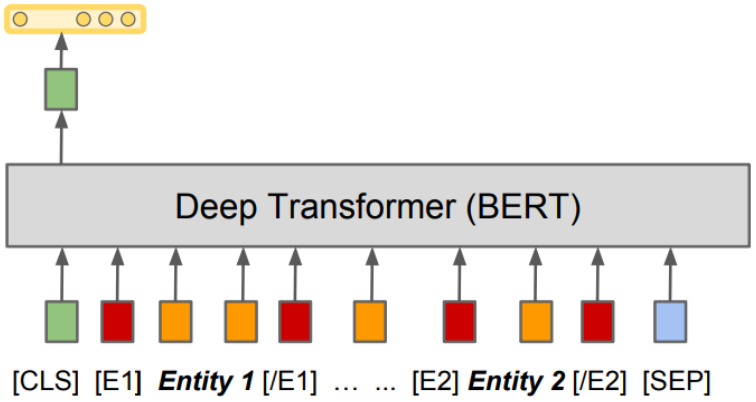


STANDARD – MENTION POOLING

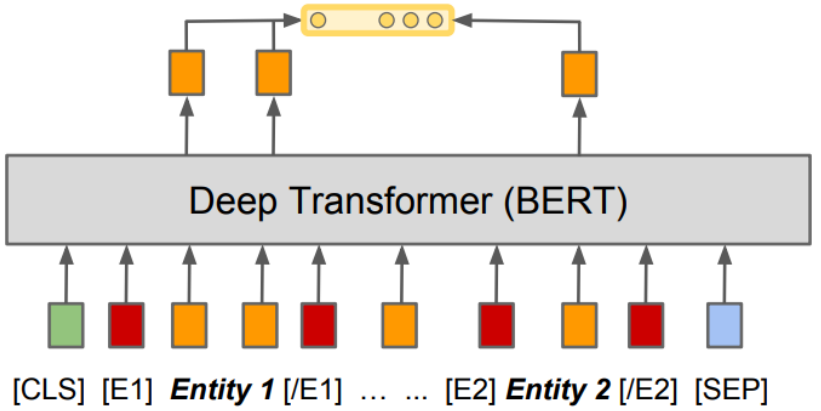
用实体的token对应的输出向量求平均得到表示该实体的向量
将两个实体的表示向量拼接，基于拼接得到的向量进行关系分类

基于预训练模型的方法

- 基于BERT的方法
  - 基于BERT模型，可以有多种不同的方式得到用于关系分类的向量表示



ENTITY MARKERS – [CLS]

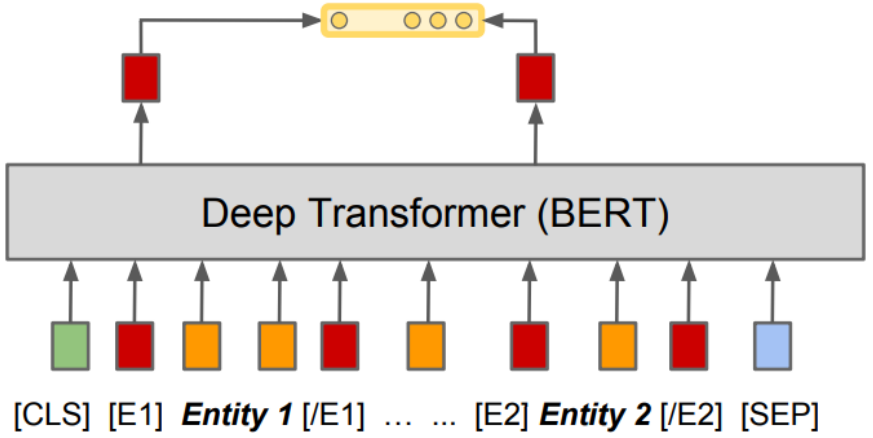


ENTITY MARKERS – MENTION POOL.

用特殊token标记出两个实体

基于预训练模型的方法

- 基于BERT的方法
  - 基于BERT模型，可以有多种不同的方式得到用于关系分类的向量表示



ENTITY MARKERS – ENTITY START

用放在实体前的特殊token对应的向量表示这个实体

基于预训练模型的方法

- 基于BERT的方法

		SemEval 2010 Task 8		KBP37		TACRED		FewRel 5-way-1-shot
# training annotated examples		8,000 (6,500 for dev)		15,916		68,120		44,800
# relation types		19		37		42		100
		Dev F1	Test F1	Dev F1	Test F1	Dev F1	Test F1	Dev Acc.
Wang et al. (2016)*		–	88.0	–	–	–	–	–
Zhang and Wang (2015)*		–	79.6	–	58.8	–	–	–
Bilan and Roth (2018)*		–	84.8	–	–	–	68.2	–
Han et al. (2018)		–	–	–	–	–	–	71.6
Input type	**Output type**							
STANDARD	[CLS]	71.6	–	41.3	–	23.4	–	85.2
STANDARD	MENTION POOL.	78.8	–	48.3	–	66.7	–	87.5
ENTITY MARKERS	[CLS]	81.2	–	68.7	–	65.7	–	85.2
ENTITY MARKERS	MENTION POOL.	80.4	–	68.2	–	69.5	–	87.6
ENTITY MARKERS	ENTITY START	**82.1**	**89.2**	**70**	**68.3**	**70.1**	**70.1**	**88.9**

基于预训练模型的方法

- 基于BERT的方法
- 可将类别信息加入marker
  - 用<E1:TYPE>, </E1:TYPE>和<E2:TYPE>, </E2:TYPE>标记实体，其中TYPE对应实体的类别
    - 引入多个新的token
  - 用@ * e1-type * Entity1 @标记实体1，# ^ e2-type ^ Entity2 #标记实体2。其中e1-type和e2-type分别为实体1和实体2的类别
    - 不引入新的token

“@ * person * Joe Biden @ is the 46th president of the # ^ country ^ United States #”

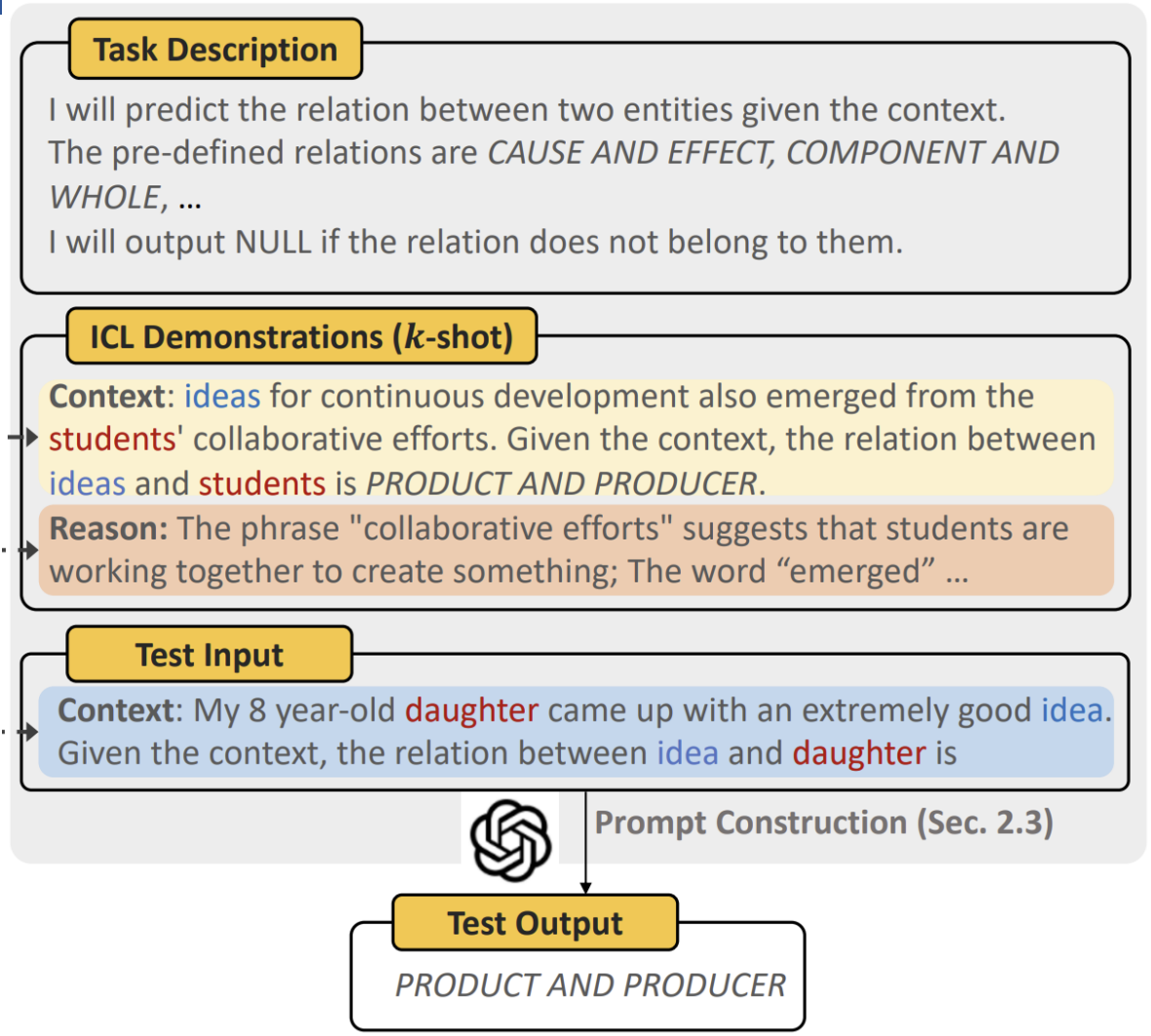
Zhou, Wenxuan, and Muhao Chen. "An Improved Baseline for Sentence-level Relation Extraction." In Proceedings of AACL-IJCNLP, pp. 161-168. 2022.

基于预训练模型的方法

- 基于BERT的方法
- 可将类别信息加入marker
  - Typed entity marker: <E1:TYPE>, </E1:TYPE>和<E2:TYPE>, </E2:TYPE>
  - Typed entity marker (punct): @ * e1-type * e1 @ 和 # ^ e2-type ^ e2 #

Model	TACRED Test F_1	TACREV Test F_1	Re-TACRED Test F_1
Sequence-based Models			
PA-LSTM (Zhang et al., 2017)	65.1	73.3‡	79.4†
C-GCN (Zhang et al., 2018)	66.3	74.6‡	80.3†
Transformer-based Models			
$\text{BERT}_{\text{BASE}}$ + entity marker	68.4	77.2	87.7
$\text{BERT}_{\text{LARGE}}$ + entity marker	69.7	77.9	89.2
$\text{RoBERTa}_{\text{LARGE}}$ + entity marker	70.7	81.2	90.5
SpanBERT (Joshi et al., 2020)	70.8	78.0*	85.3†
KnowBERT (Peters et al., 2019)	71.5	79.3*	-
LUKE (Yamada et al., 2020)	72.7	80.6‡	90.3‡
Improved RE baseline			
$\text{BERT}_{\text{BASE}}$ + typed entity marker	71.5	79.3	87.9
$\text{BERT}_{\text{LARGE}}$ + typed entity marker	72.9	81.3	89.7
$\text{RoBERTa}_{\text{LARGE}}$ + typed entity marker (punct)	**74.6**	**83.2**	**91.1**

基于大模型的关系抽取



- 在给大模型的输入中提供任务样例的方法叫做上下文学习（In-context Learning, ICL）
- 可能的情况：有很多训练样本，但只能给大模型提供若干个上下文学习样例（ICL Demonstrations），如5个、10个
- 选哪些训练样本作为样例提供给大模型是一个重要的问题
- 为每个测试输入（Test Input），可从所有训练样本中找出与之相似的作为样例

Wan, Zhen, et al. "GPT-RE: In-context Learning for Relation Extraction using Large Language Models." The 2023 Conference on EMNLP. 2023.

基于大模型的关系抽取

- GPT-RE效果

Methods	Retriever	Semeval	TACRED	SciERC	ACE05
GPT-3 Baselines (Best k-shot)					
GPT-Random	-	70.04 (30)	32.49 (15)	17.92 (25)	9.04 (25)
GPT-Sent	SimCSE	79.94 (30)	33.45 (15)	20.96 (25)	6.31 (25)
Ours (Best k-shot)					
GPT-RE_SimCSE	SimCSE	81.02 (30)	37.44 (15)	26.46 (25)	8.67 (25)
GPT-RE_SimCSE*	SimCSE	77.49 (15)	31.58 (10)	-	-
+ Reasoning	SimCSE	79.88 (15)	33.18 (10)	-	-
GPT-RE_FT	PURE	**91.90** (25)	72.14 (15)	**69.00** (30)	68.73 (25)
GPT-RE_FT*	PURE	91.11 (15)	70.38 (10)	-	-
+ Reasoning	PURE	91.82 (15)	70.97 (10)	-	-
Fine-tuned RE Baselines					
Cohen et al. (2020)		**91.90**	-	-	-
Wang et al. (2022a)		-	♣**76.80**	-	-
PURE (Zhong and Chen, 2021)		89.90	69.72	68.45	**70.09**

- GPT-Sent: 直接根据包含实体的句子找相似样本
- GPT-RE_SimCSE: 将句子改写后找相似样本：

He has a sister Lisa.

The relation between "He" and "Lisa" in: He has a sister Lisa.

- GPT-RE_FT: 训练一个专为关系抽取找相似样本的模型

其他设定下的关系抽取

- 基于远程监督的关系抽取
- 基于序列标注的关系抽取

基于远程监督的关系抽取

- 远程监督关系抽取的假设：
  - 如果两个实体间有某种关系，那么包含了这两个实体的句子就可能表示了这种关系
- 使用远程监督构建训练数据的方法：
  1. 对于一种关系，先获取一个属于该关系的实体对集合
     - 如为*film_director*关系找到了实体对<Steven Spielberg, Saving Private Ryan>, <Christopher Nolan, Oppenheimer>, …
  2. 对每个实体对，从大量文本数据中找出同时包含它们的句子，将这些句子作为该类关系的训练样本
     - 如通过<Steven Spielberg, Saving Private Ryan>实体对找到了一些句子：

1. Allison co-produced the Academy Award winning [Saving Private Ryan], directed by [Steven Spielberg] …
2. [Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story.
……

基于远程监督的关系抽取

- 问题：有些被作为训练样本的句子中，并没有表示两个实体间有目标关系

  Relation: founderOf Entity 1: Steve Jobs Entity 2: Apple

  1. Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
  2. Steve Jobs passed away day before Apple unveiled iPhone 4S in late 2011.

- 错误的训练样本导致模型更容易做出错误的预测

基于远程监督的关系抽取

- Multi-instance learning (MIL)

直接训练的假设：如果两个实体间有某种关系，那么**每个**包含了这两个实体的句子都表示了这种关系

基于MIL的假设：如果两个实体间有某种关系，那么**至少有一个**包含了这两个实体的句子表示了这种关系

基于MIL的远程监督关系抽取

- Multi-instance learning (MIL)
- 做法：把样本分成bags，每个bag中包含提到同一实体对的句子
- 将实体关系标签赋给bag，而非句子

Bag	Sentence
Bag1 **label**: directorOf	Allison co-produced the Academy Award winning Saving Private Ryan, directed by Steven Spielberg ...
	Steven Spielberg's film Saving Private Ryan is loosely based on the brothers' story.
Bag2 **label**: founderOf	Steve Jobs was the co-founder and CEO of Apple and formerly Pixar
	Steve Jobs passed away day before Apple unveiled iPhone 4S in late 2011.

- 训练时：假设每个bag中至少有一个句子表示了所标注关系
- 预测时：以实体对为单位预测，将实体对对应的句子放在一个bag中，预测bag标签

基于MIL的远程监督关系抽取

- Multi-instance learning (MIL)

假设每个bag中至少有一个句子表示了所标注关系

设定： 设有T个bag $(M_i, y_i), i = 1, \ldots, T$。其中$M_i$为含实体对的句子样本集合；$y_i$为该bag对应实体对的关系标签；$M_i$中的样本数为$q_i$

采用的模型： 普通的关系分类模型，如基于BERT的模型

训练**目标函数**定义为：

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i | m_i^j; \theta)$$

其中j取

$$j^* = \arg\max_j p(y_i | m_i^j; \theta) \quad 1 \le j \le q_i \qquad (9)$$

Zeng, Daojian, et al. "Distant supervision for relation extraction via piecewise convolutional neural networks." Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.

基于MIL的远程监督关系抽取

设定：设有T个bag $(M_i, y_i), i = 1, \ldots, T$。其中$M_i$为含实体对的句子样本集合；$y_i$为该bag对应实体对的关系标签；$M_i$中的样本数为$q_i$

采用的模型：普通的关系分类模型，如基于BERT的模型

训练**目标函数**定义为：$J(\theta) = \sum_{i=1}^{T} \log p(y_i | m_i^j; \theta)$

其中j取 $j^* = \arg\max_j p(y_i | m_i^j; \theta) \quad 1 \le j \le q_i \qquad (9)$

训练流程：

1. 用公式（9）为每个bag找到j
2. 根据找到的j，用对应样本更新模型参数
3. 重复

Zeng, Daojian, et al. "Distant supervision for relation extraction via piecewise convolutional neural networks." Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.

其他设定下的关系抽取

- 基于远程监督的关系抽取
- 基于序列标注的关系抽取

基于序列标注的关系抽取

- 考虑的情况：输入文本都是都对某一实体的介绍性描述
  - 那么，其中提到的大部分其他实体都与该实体有一定关联
- 可只抽取其他实体与被描述实体的关系，这样可将问题转化为序列标注问题
- 同时实现了实体识别和关系分类
- 与NER类似，可使用神经网络+CRF或HMM等模型

George W. Bush

George	is	the	son	of	George	H.	Bush
O	*O*	*O*	*O*	*O*	*B-FATHER*	*I-FATHER*	*I-FATHER*

and	Barbara	Bush
O	*B-MOTHER*	*I-MOTHER*

其他类型的关系抽取问题

- 文档级关系抽取
- Open Information Extraction

文档级关系抽取 (Document Level RE)

- 普通的关系抽取任务只关注从单个句子中抽取
- 但有些实体间关系需要关联多个句子才能推断出

例：

[1] *Rage Against the Machine* is an American rap metal band from Los Angeles, California. [2] Formed in 1991, the group consists of vocalist Zack de la Rocha, guitarist Tom Morello, bassist Tim Commerford and drummer *Brad Wilk*. [3] After a self-issued demo, the band signed with Epic Records and released its debut album *Rage Against the Machine* in 1992. …

Relation: MemberOf **Supporting Evidence:** {1,2}

文档级关系抽取

- 基于BERT的方法

1. 对输入文本，在每个实体提及（entity mention）前加“*”作为标记token
2. 将文本的token序列输入BERT，获得相应向量表示序列，对任意mention m，将其前面的“*”对应的向量作为m的向量表示，记为h_m
3. 对一个实体e_i，记$\{m_j^i\}_{j=1}^{N_{e_i}}$为文本中对它的所有提及（mention），则用下式计算e_i的向量表示：

$$\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\left(\boldsymbol{h}_{m_j^i}\right)$$

注：上式称为logsumexp pooling，类似max pooling和mean pooling

4. 对实体对e_s, e_o，将它们间有关系r的概率建模为：

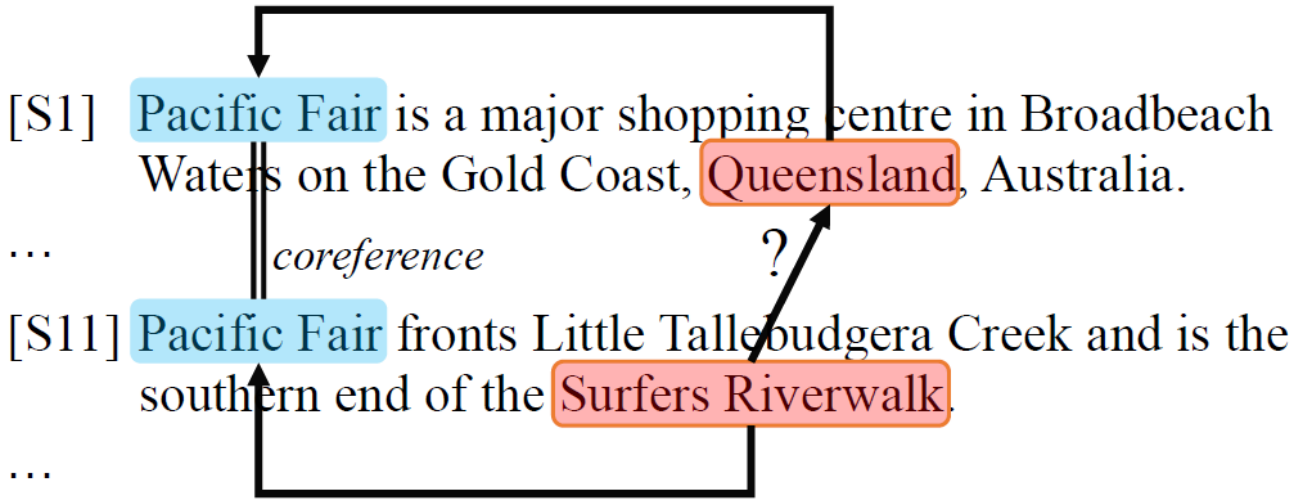
$$\begin{aligned} \boldsymbol{z}_s &= \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right), \\ \boldsymbol{z}_o &= \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right), \\ \mathrm{P}\left(r|e_s, e_o\right) &= \sigma\left(\boldsymbol{z}_s^{\intercal} \boldsymbol{W}_r \boldsymbol{z}_o + b_r\right), \end{aligned}$$

其中σ为sigmoid函数

效果：这种基于BERT的方法在常用的DocRED数据集的Dev集上F1可达约56%

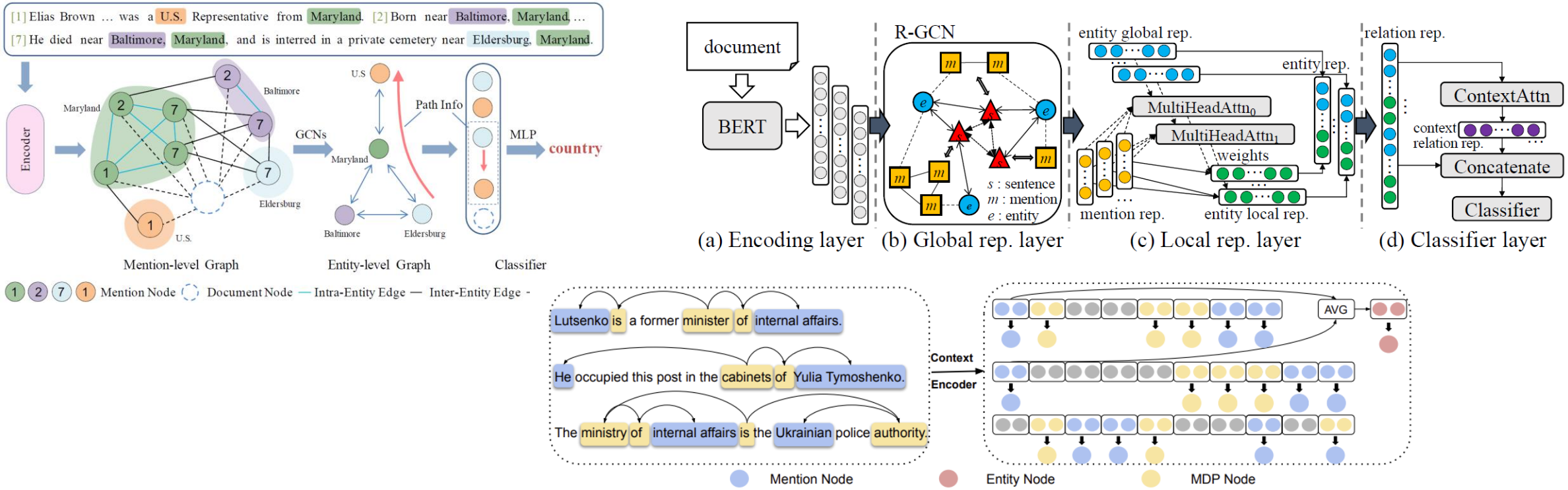
文档级关系抽取

- 基于BERT的方法未对实体提及（mention）间的关联建模，可能导致一些需通过不同句子分析出的实体关系难以预测



文档级关系抽取

- 许多基于图（Graph-based）的方法被提出
- 基于图，可以为只在不同句子中出现的实体也建立关联



如：(Zeng et al., 2020)的方法在DocRED的Dev集上F1可达63.09%

其他类型的关系抽取问题

- 文档级关系抽取
- Open Information Extraction

Open Information Extraction

- 关系抽取任务需预先设定好关系类别
  - 有时可能希望抽取一些未知类别的关系

- Open IE (Open Information Extraction)不预先设定关系类别
  - 直接从文本中识别实体和表示关系的短语得到三元组
  - 知名的Open IE系统有：
    - TextRunner (Banko et al., 2007), ReVerb (Fader et al., 2011), OLLIE (Schmitz et al., 2012)等

Open Information Extraction

- ReVerb (Fader et al., 2011)

- 从句子中抽取关系三元组，其主要步骤：
  1. 抽取关系短语
     - 采用基于词性标注（POS tags）的规则抽取关系短语
     - 如“ invented”，“located in”，“has atomic weight of”
  2. 抽取实体对，形成三元组
     - 找到关系短语左边和右边离它最近的名词，作为对应该关系的实体对

Open Information Extraction

- ReVerb抽取出的三元组示例

<1 billion, is a lot of, money>
<carrots, are a good source of, vitamin A>
<vitamin C, is an essential nutrient for, human>
<a group, made up of, a mother>
<doctors, performed a series of, tests>
<Paul, also made, Pizza>
<Rabin, walks out of, a small room>
<Patrick, is very afraid of, heights>
<Patrick, grew up in, L.A.>
…

Open Information Extraction

- 目前也有基于神经网络模型的方法
- 由于关系类别无限制，应用起来难度更高
- 相对普通关系抽取，对Open IE的研究少很多

END