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**

事件抽取 (Event Extraction)

Trigger

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。

目标输出：

触发词	发布
类别	*产品发布*
发布公司	Nuance
发布产品	DAX Express
时间	3月20日
地点	NULL

- ***事件***是由一个**事件触发词（Event Trigger）**和一些对应的**事件元素（Event Arguments）**所组成的结构。
  - 事件触发词: 最明确地表示了事件的发生的词。
  - 事件元素：与事件相关的实体。

事件抽取 (Event Extraction)

Trigger

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。

- 两个子任务
  - 事件识别（Event Detection）
    - 识别事件触发词（Event Trigger）
    - 事件分类
  - 事件元素（Event Argument）抽取
    - 元素识别
    - 元素角色分类

触发词	发布
类别	*产品发布*
发布公司	Nuance
发布产品	DAX Express
时间	3月20日
地点	NULL

元素角色
(Argument role)

事件元素
(Event argument)

事件抽取数据集

- ACE 2005
  - 英文、中文、阿拉伯语；通用领域文本（broadcast news, newsgroups等）；8个事件类别，33个子类；共36种元素角色；约6000个事件
- ERE
  - 英文、中文、西班牙语；通用领域（新闻、论坛）；38个事件类型；约30,000个事件
- 其他
  - MAVEN（只含事件识别子任务）、GENIA（生物医疗领域）、DuEE（中文新闻文本）、DuEE-Fin（中文经济新闻文本）等

事件抽取数据集

- ACE 2005中的事件类别

SN	Event Type	SN	Event subtype
1	Life	1-5	Be-Born, Marry, Divorce, Injure, Die
2	Movement	6	Transport
3	Contact	7-8	Meet, Phone-write
4	Conflict	9-10	Attack, Demonstrate
5	Business	11-14	Merge-org, Declare-bankruptcy, Start-Org, End-org
6	Transaction	15-16	Transfer-money, Transfer-ownership
7	Persosnnel	17-20	Elect, Start-position, End-position, Nominate
8	Justice	21-33	Arrest-jail, Execute, Pardon, Release-parole, Fine, Convict, Charge-indict, Trial-hearing, Acquite, Sentence, Sue, Extradite, appeal

事件抽取数据集

- ACE 2005数据 – 例

At least 19 people were killed and 114 people were wounded in Tuesday's southern Philippines airport blast, officials said, but reports said the death toll could climb to 30.

Type: *Life*
Subtype: *Die*

Trigger: “killed”
Victim: “At least 19 people”
Place: “southern Philippines airport”
Time-Within: “Tuesday”

基于规则的事件抽取

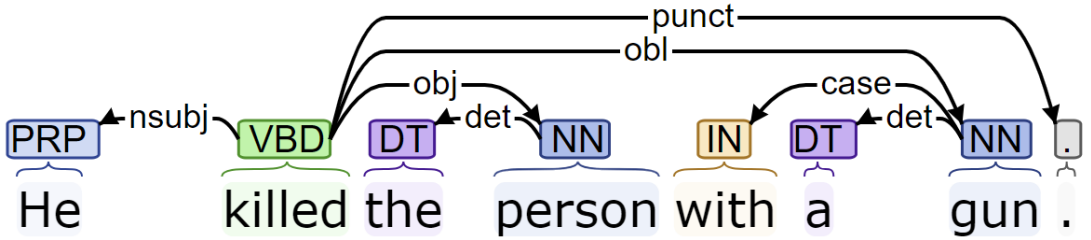
- 基于词典、字符串匹配、语法解析结果的规则抽取触发词和元素

例：(Riloff et al., 1993)使用的抽取事件触发词和对应的一个元素的模板：

<victim> was murdered	killing <victim>
<perpetrator> attempted to kill	fatality was <victim>
killed <victim>	bomb against <target>
threatened to attack <target>	killed with <instrument>

Riloff, Ellen. "Automatically constructing a dictionary for information extraction tasks." AAAI. Vol. 1. No. 1. 1993.

例：基于依存句法解析的规则



nsubj一般关联了施动者，obj一般关联了受动者

基于机器学习的事件抽取

- Pipeline方法
  1. Event detection（事件识别）
     - 识别event trigger（事件触发词），并确定事件类别
  2. Event argument extraction（事件元素抽取）
     - 识别事件元素及对应角色

- Joint方法
  - 同时识别事件触发词和事件元素

Pipeline方法

- 识别event trigger（事件触发词）
  - 思路：将输入文本中每个词分类，看它是不是trigger，如果是，属于哪种类型的事件
    - 可使用单个分类器：
      - 分为*NA*, *Life-Die*, *Movement-Transport*, *Conflict-Attack*等类别，其中*NA*表示不是trigger
    - 也可使用两个分类器：
      - 先分为两类：是trigger, 不是trigger
      - 如果是trigger，再分类到事件类别：*Life-Die*, *Movement-Transport*, *Conflict-Attack*等
- 识别event argument（事件元素）
  - 思路：将已识别的trigger作为额外输入，对每个实体（用NER提前找出）分类
    - 如类别*NA*表示不是该事件的元素，*victim*表示在该事件中是受害人的角色
- 具体方法：手工设计特征、卷积神经网络、预训练模型、基于机器阅读理解等

手工设计特征的方法

(Ahn 2006)使用的触发词识别特征:

Lexical features: full word, lowercase word, lemmatized word, POS tag, depth of word in parse tree

Left context (3 words): lowercase, POS tag

Right context (3 words): lowercase, POS tag

Dependency features: if the candidate word is the dependent in a dependency relation, the label of the relation is a feature value, as are the dependency head word, its POS tag, and its entity type

...

(Ahn 2006)使用的元素识别特征:

Trigger word of event mention: full, lowercase, POS tag, and depth in parse tree

Event type of event mention

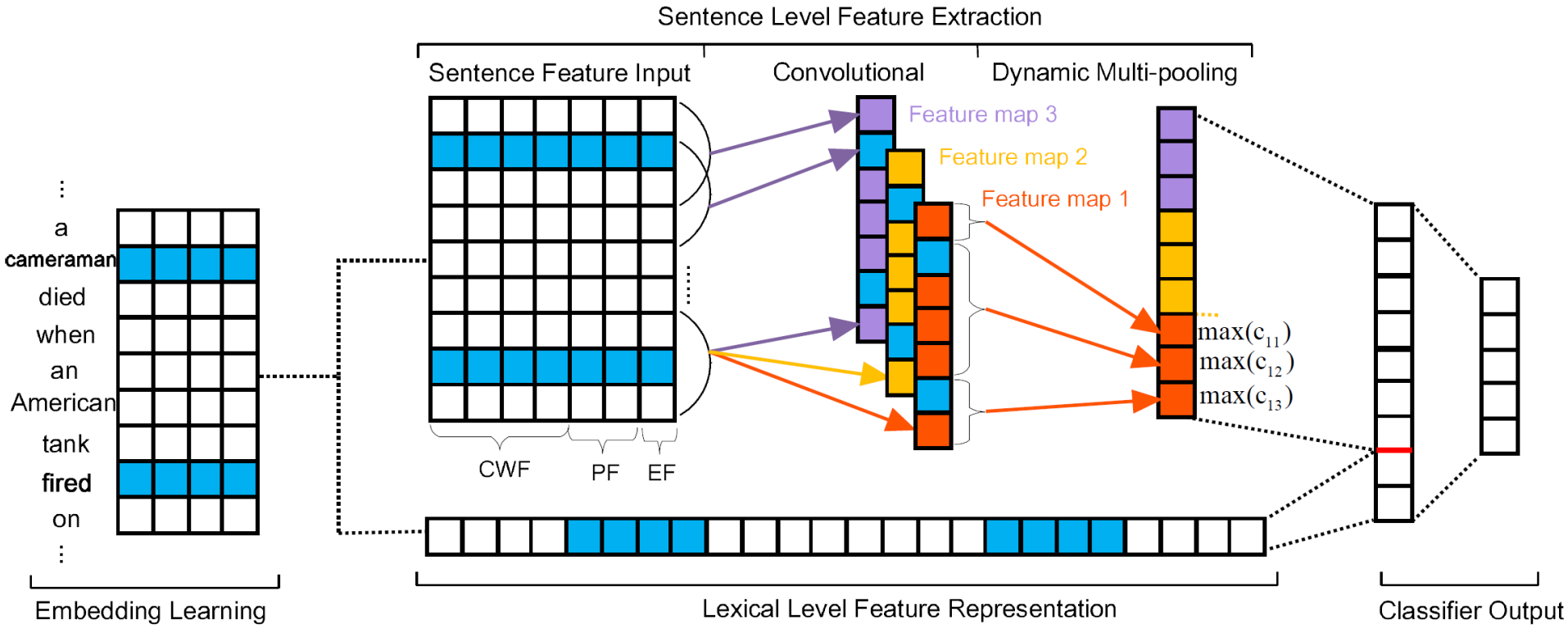
Dependency path between trigger word and constituent head word of entity mention, expressed as a sequence of labels, of words, and of POS tags

...

Ahn, David. "The stages of event extraction." Proceedings of the Workshop on Annotating and Reasoning about Time and Events. 2006.

事件识别方法

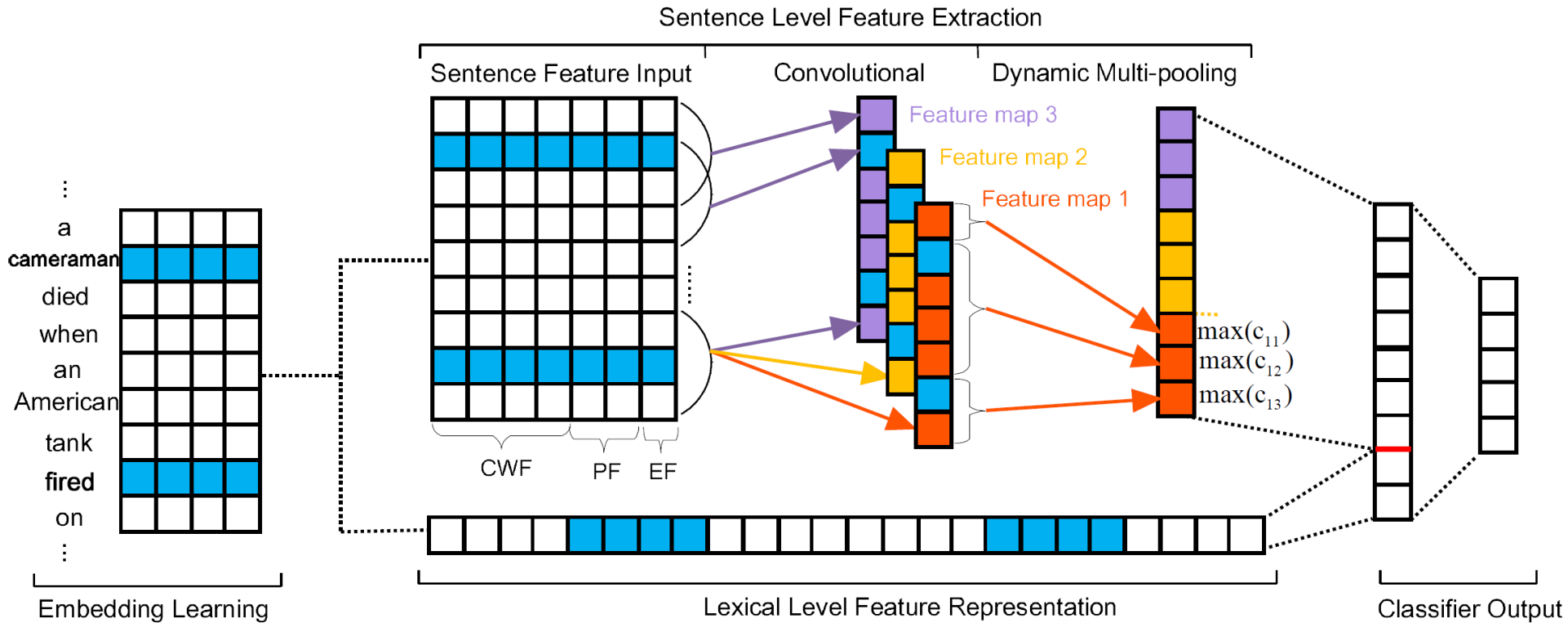
- 基于普通神经网络的方法
- 基于CNN的事件触发词/元素识别 （Chen et al., 2015）



Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of ACL-IJCNLP. 2015.

基于CNN的事件触发词/元素识别

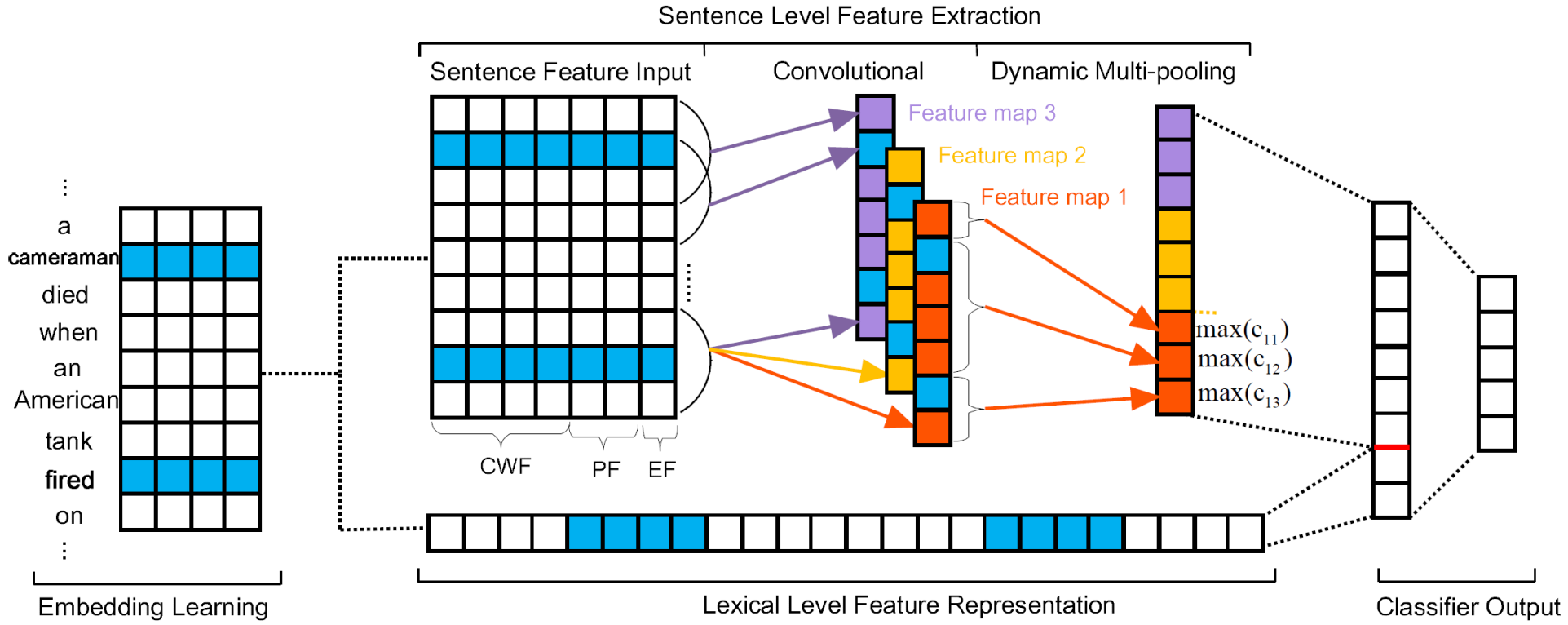
- 用于已预测触发词后识别事件元素（即pipeline第二步）的模型结构:



1) 基于CNN为获取句子级特征向量表示
2) 基于当前触发词和所考虑实体获取的词级向量表示
3) 将两者拼接得到完整向量表示，基于该向量分类

Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of ACL-IJCNLP. 2015.

基于CNN的事件触发词/元素识别

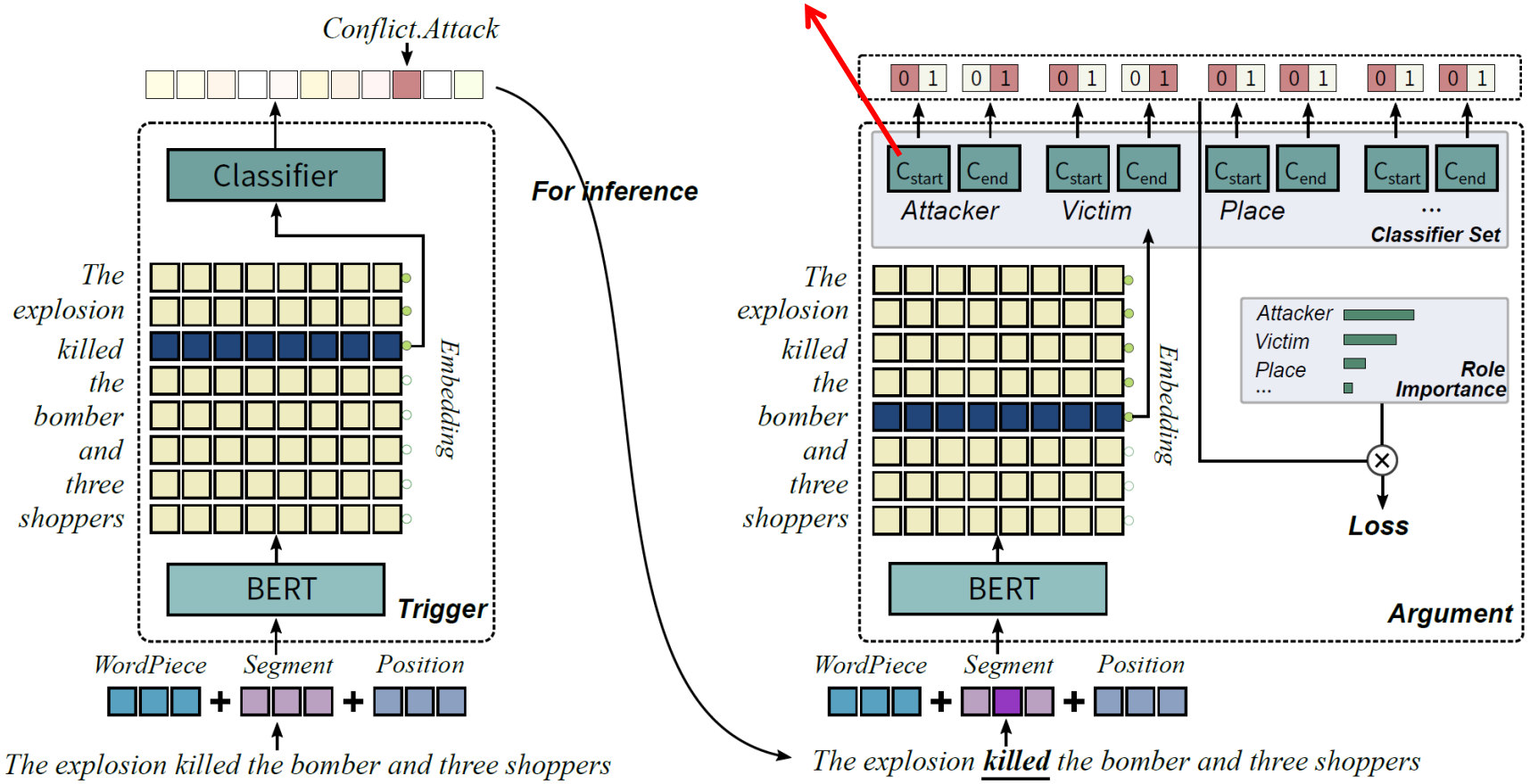


- 将该模型简化后可用于触发词识别:
  - Sentence Feature Input部分去除EF，即事件类别的向量表示
  - Dynamic Multi-pooling部分根据当前考虑的词分两段，而非三段
  - Lexical Level Feature包含基于前考虑的词获取的特征

Chen, Yubo, et al. "Event extraction via dynamic multi-pooling convolutional neural networks." Proceedings of ACL-IJCNLP. 2015.

基于预训练模型的方法

- PLMEE



事件元素识别部分：
为触发词使用与其他词不同的segment embedding，以进行区分

Yang, Sen, et al. "Exploring pre-trained language models for event extraction and generation." Proceedings of ACL. 2019.

基于预训练模型的方法

- 实验效果(ACE 2005)

Model \ Phase	Trigger Identification(%)			Trigger Calssfication(%)			Argument Identification(%)			Argument Calssfication(%)		
	P	R	F	P	R	F	P	R	F	P	R	F
Cross Event		N/A		68.7	68.9	68.8	50.9	49.7	50.3	45.1	44.1	44.6
Cross Entity		N/A		72.9	64.3	68.3	53.4	52.9	53.1	51.6	45.5	48.3
Max Entropy	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
DMCNN	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5
JRNN	68.5	75.7	71.9	66.0	73.0	69.3	61.4	64.2	62.8	54.2	56.7	55.4
DMCNN-DS	79.7	69.6	74.3	75.7	66.0	70.5	71.4	56.9	63.3	62.8	50.1	55.7
ANN-FN		N/A		79.5	60.7	68.8		N/A			N/A	
ANN-AugATT		N/A		78.0	66.3	71.7		N/A			N/A	
PLMEE(-)	84.8	83.7	**84.2**	81.0	80.4	**80.7**	71.5	59.2	64.7	61.7	53.9	57.5
PLMEE							71.4	60.1	**65.3**	62.3	54.2	**58.0**

基于机器阅读理解（MRC）的方法

- 机器阅读理解（Machine Reading Comprehension, MRC）
  - MRC: 给定一段文本和一个问题，根据文本对问题进行回答。

- 基于MRC的事件抽取：
  - 先通过分类方法识别事件触发词
  - 基于识别出的触发词和事件类别，将事件元素抽取转化为对输入文本的提问，以MRC方式实现事件元素抽取

Liu, Jian, et al. "Event extraction as machine reading comprehension." Proceedings of EMNLP. 2020.

基于机器阅读理解（MRC）的方法

输入文本： On Sunday, a protester stabbed an officer with a paper cutter.

Trigger识别：

Trigger: stabbed
Event type: *Attack*

事件元素问题构建：

对目标元素角色*Instrument* 使用模板　　对原输入文本转换形式

[What is the instrument] [that a protester use to stab an officer?]

使用MRC识别事件元素：

Q: What is the instrument that a protester use to stab an officer?

A: a paper cutter

基于机器阅读理解（MRC）的方法

- 实验效果

	Trigger Ex.			Argument Ex.			Argument Ex.(O)		
Method	P	R	F1	P	R	F1	P	R	F1
JointBeam (Li et al., 2013)	73.7	62.3	67.5	64.7	44.4	52.7	-	-	-
DMCNN (Chen et al., 2015)	75.6	63.6	69.1	62.2	46.9	53.5	59.0†	54.8†	56.8†
JRNN (Nguyen et al., 2016)	66.0	73.0	69.3	54.2	56.7	55.4	57.5†	58.2†	57.9†
dbRNN (Sha et al., 2018)	74.1	69.8	71.9	66.2	52.8	58.7	58.4†	64.2†	61.2†
JMEE (Liu et al., 2018b)	**76.1**	71.3	73.7	66.8	54.9	60.3	59.8†	64.2†	62.0†
BERTEE	74.8†	73.9†	74.3†	70.5†	52.2†	60.6†	66.8†	62.6†	64.7†
RCEE_ER (ours)	75.6	**74.2**	**74.9***	63.0	**64.2**	**63.6***	**71.2**	**69.1**	**70.1***
RCEE_ER *w/o* DA (ours)	-	-	-	61.8	63.6	62.7	69.6	68.4	69.0

Liu, Jian, et al. "Event extraction as machine reading comprehension." Proceedings of EMNLP. 2020.

基于机器学习的事件抽取

- Pipeline方法
  1. Event detection（事件识别）
     - 识别event trigger（事件触发词），并确定事件类别
  2. Event argument extraction（事件元素抽取）
     - 识别事件元素及对应角色

- Joint方法
  - 同时识别事件触发词和事件元素

Joint方法

- (Li et al., 2013)的方法
- 采用结构化预测（structured prediction）同时预测出事件触发词和元素

设句子含s个词，其中有m个可作为事件元素的实体，则其预测输出形式：

$$y = (t_1, a_{1,1}, \ldots, a_{1,m}, \ldots, t_s, a_{s,1}, \ldots, a_{s,m})$$

其中t_i为第i个词的触发词标签，$a_{i,j}$为当第i个词为触发词时，第j个实体的事件元素标签

Li, Qi, Heng Ji, and Liang Huang. "Joint event extraction via structured prediction with global features." Proceedings of ACL. 2013.

(Li et al., 2013)的方法

设句子含s个词，其中有m个可作为事件元素的实体，则其预测输出形式：

$$y = (t_1, a_{1,1}, \ldots, a_{1,m}, \ \ldots, \ t_s, a_{s,1}, \ldots, a_{s,m})$$

其中t_i为第i个词的触发词标签，$a_{i,j}$为当第i个词为触发词时，第j个实体的事件元素标签

例 **输入x: 句子为** Jobs founded Apple **实体为** [Jobs, Apple]

则： $y = (\perp, \perp, \perp, \underbrace{Start\_Org}_{t_2}, \underbrace{Agent, \ Org}_{\text{args for } founded}, \perp, \perp, \perp)$

$\perp$ 表示不是触发词或不是事件元素

(Li et al., 2013) 的方法

设句子含s个词，其中有m个可作为事件元素的实体，则其预测输出形式：

$$y = (t_1, a_{1,1}, \ldots, a_{1,m}, \ \ldots, \ t_s, a_{s,1}, \ldots, a_{s,m})$$

其中t_i为第i个词的触发词标签，$a_{i,j}$为当第i个词为触发词时，第j个实体的事件元素标签

- 模型（感知机, perceptron）：

$$z = \underset{y' \in \mathcal{Y}(x)}{\mathbf{argmax}} \quad \mathbf{w} \cdot \mathbf{f}(x, y')$$

其中w为可训练参数向量，$f(x, y')$为特征向量

- **预测时**，使用搜索算法找出近似最优的z
- **训练时**，对一个训练样本，如果找出的z错的，则更新参数：$\mathbf{w} = \mathbf{w} + \mathbf{f}(x, y) - \mathbf{f}(x, z)$

总结

- 事件抽取目标
  - 事件识别和事件元素抽取
- 基于规则的方法
- 基于机器学习的方法
  - Pipeline方法（事件识别->事件元素识别）
  - Joint方法