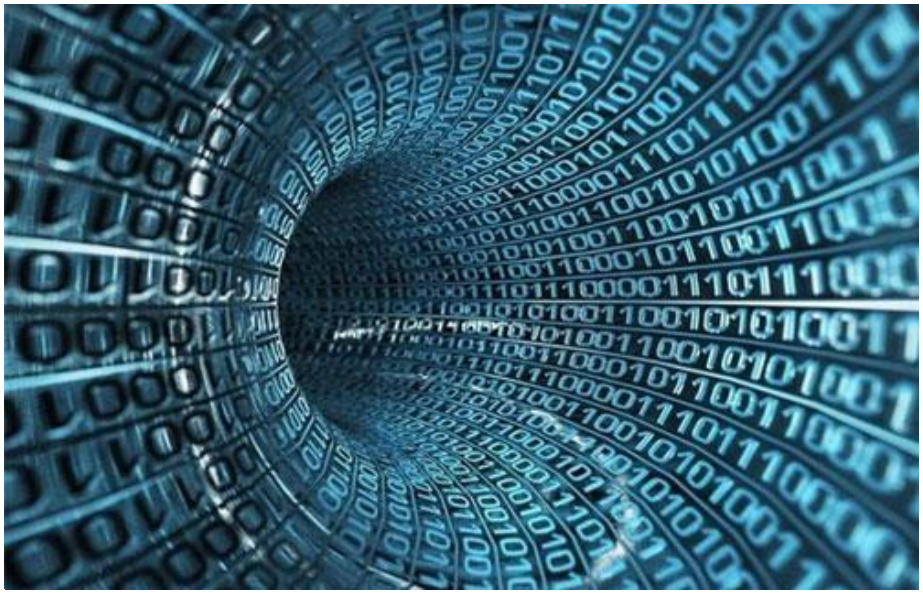
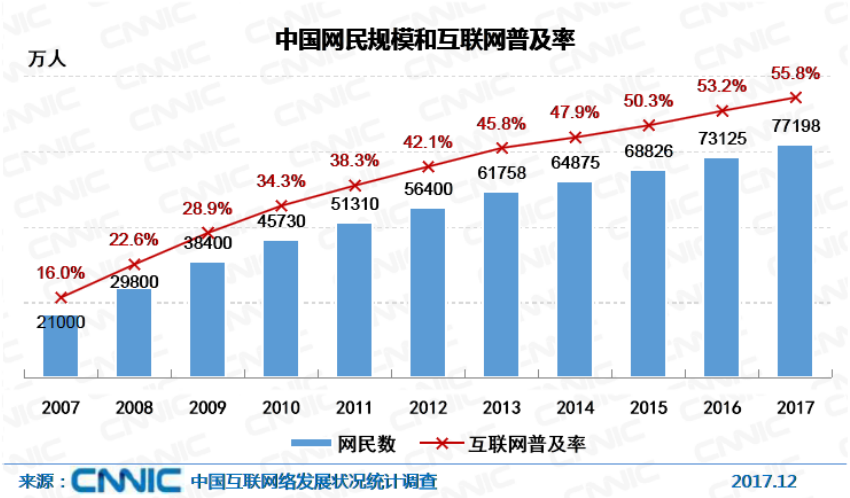
事件抽取与金融事件图谱构建

陈玉博
中国科学院自动化研究所
模式识别国家重点实验室

背景

互联网信息呈
爆炸性增长





中国网民规模达7.72亿
网站总数达到533万个
网页总数超过2604亿个

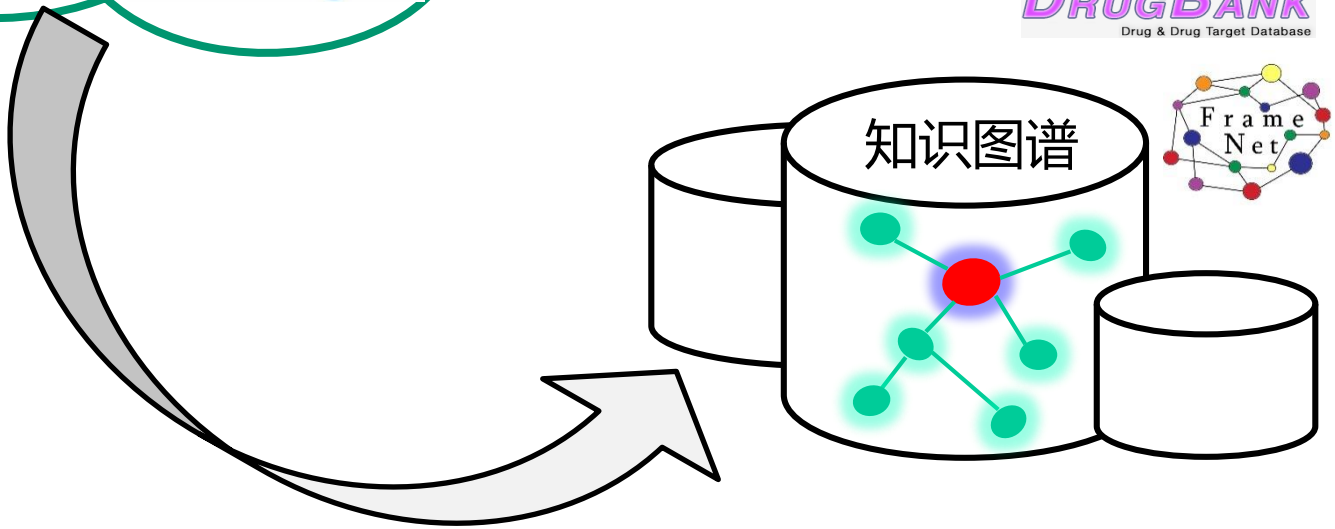
——中国互联网络发展状况统计报告2018

背景

DBpedia
Freebase
Cycorp
yago
select knowledge
DrugBank
Drug & Drug Target Database
Frame Net

实体抽取
关系抽取
事件抽取

信息抽取技术



事件图谱的意义

- **丰富现有的知识图谱**

名称	创建时间	数据来源	数据规模
OpenCyc	1984	专家知识	23 万实体，未定义事件
WordNet	1985	专家知识	15 万实体，未定义事件
YAGO	2007	WordNet+Wikipedia	459 万实体，未定义事件
DBpedia	2007	Wikipedia+ 专家知识	1694 万实体，8 万事件
Freebase	2008	Wikipedia+ 领域知识 + 集体智慧	5872 万实体，2 万事件

- **支撑其它信息获取引擎**

语义搜索

事件监控

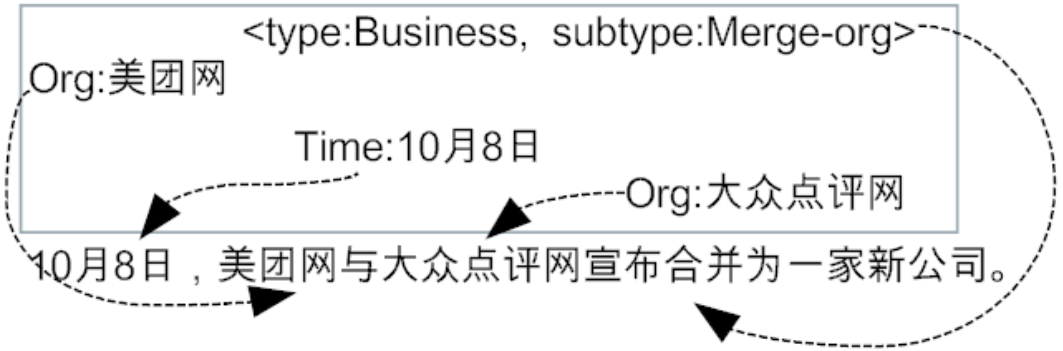
事件图谱构建的关键技术：事件抽取

- 事件抽取

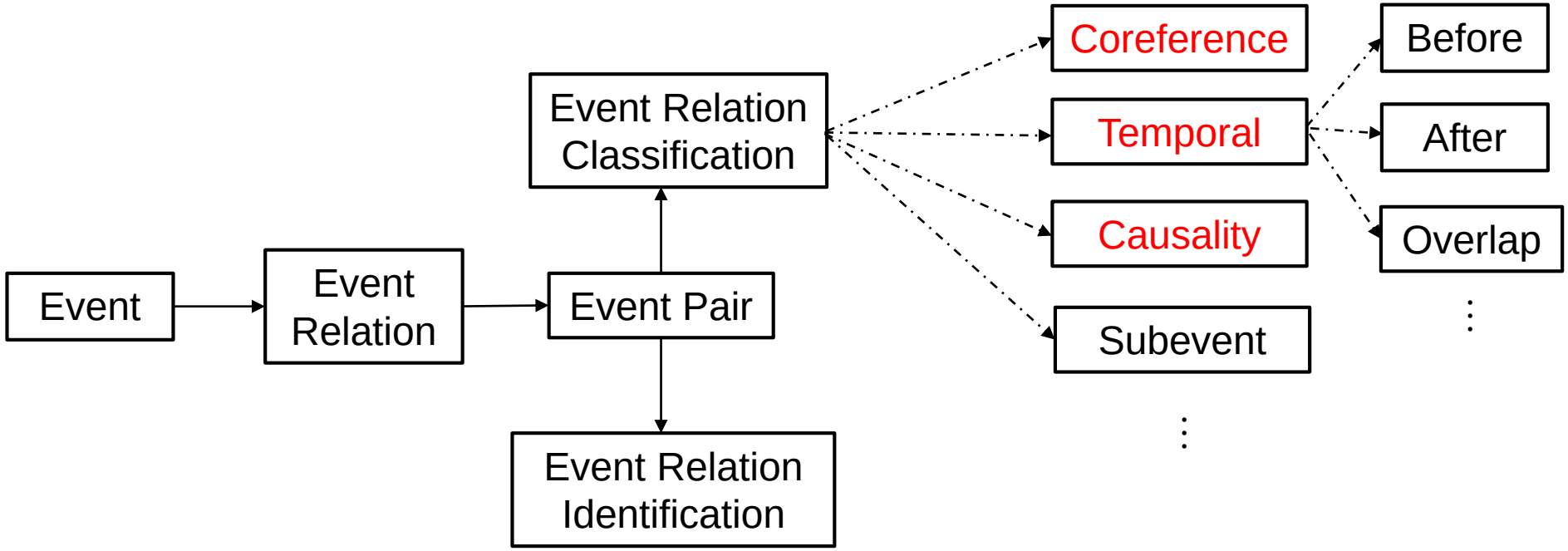
  从自然语言文本中抽取出用户感兴趣的事件信息并以结构化的形式呈现出来，如什么人/组织，什么时间，在什么地方，做了什么事

- 事件抽取相关的任务

事件发现（Event Detection）：从文本中发现事件触发词（Event Trigger），事件元素抽取（Argument Extraction）：从文本中识别事件元素（Event Argument）并判断元素扮演的角色（Argument role）



事件图谱构建的关键技术：事件关系抽取



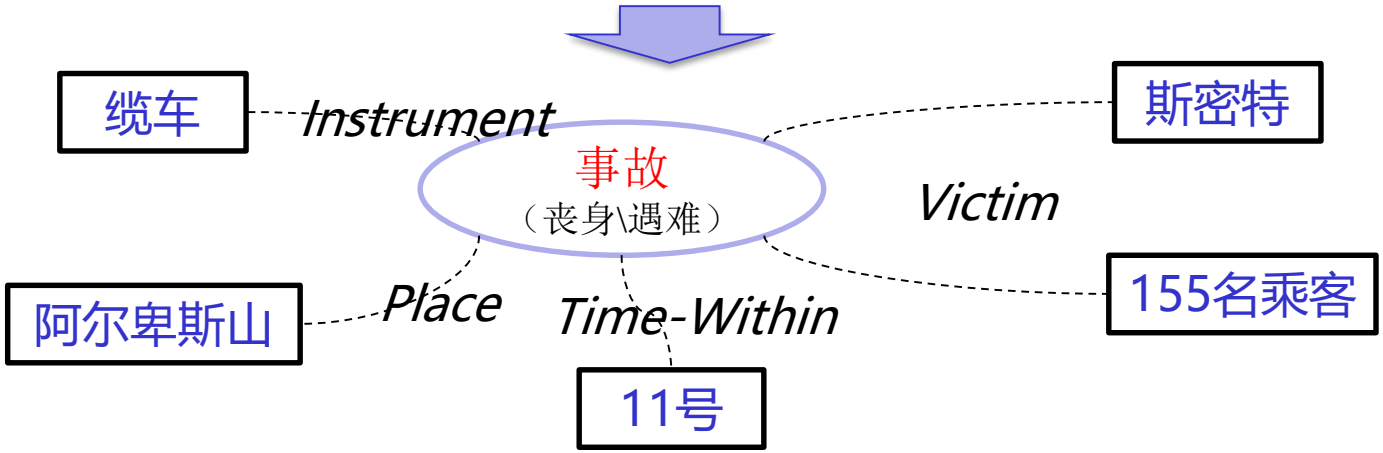
Coreference Relation

- 数据来源：ACE-Chinese

S1:根据奥地利救灾组织的统计，在阿尔卑斯山登山缆车失火惨剧中丧生的155名乘客中包括有1999年世界女子花式滑雪冠军施密特。

S2:调查单位仍然无法断定事故发生的原因。但是指出，乘客的滑雪服装和设备都是易燃材料。

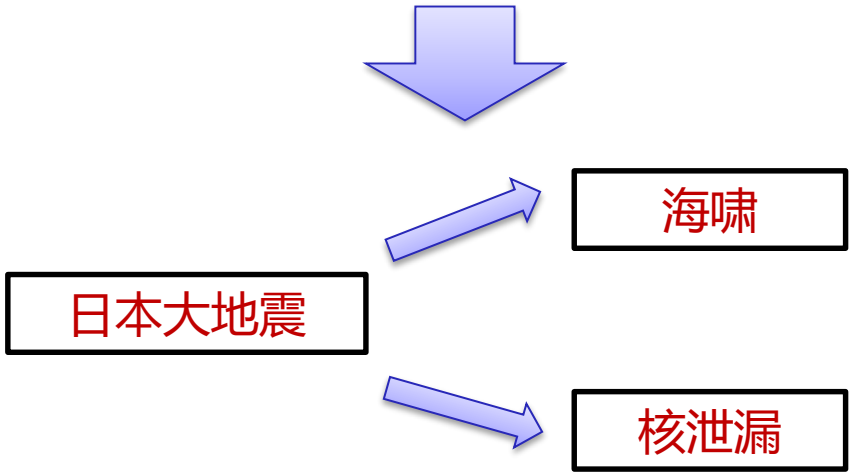
S3:奥地利一处滑雪胜地的登山缆车11号在隧道发生缆车失火惨剧。事发后有18名乘客及时逃脱存活，有155人不幸遇难。



Causality Relation

- 数据来源：百度百科

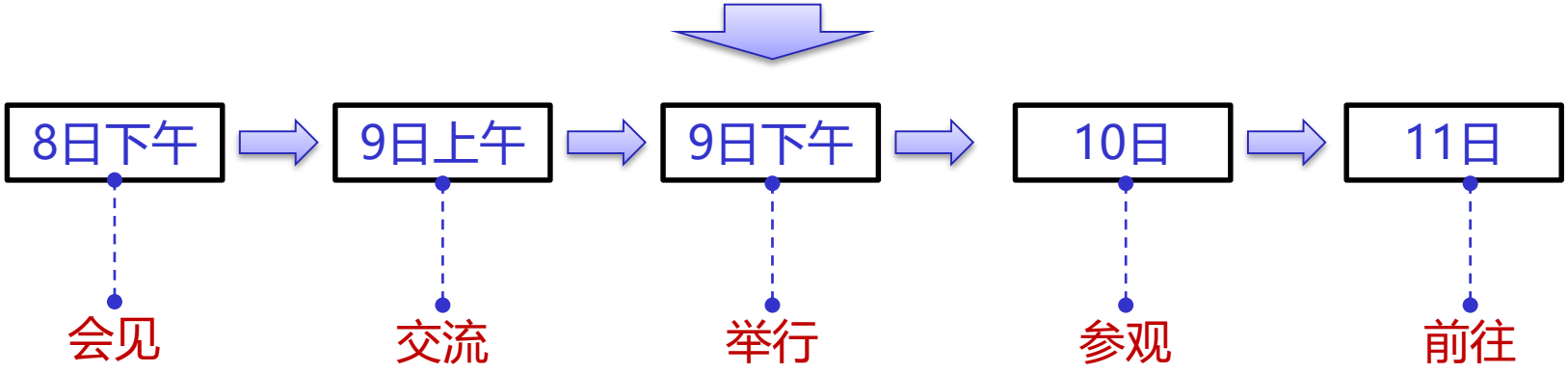
2011年3月11日13时46分，日本发生里氏9.0级大地震，地震随即引发了海啸，后来更使福岛核电站发生核泄漏危机。地震震中位于宫城县以东太平洋海域，震源深度20公里，东京有强烈震感。



Temporal Relation

- 数据来源：ACE-Chinese

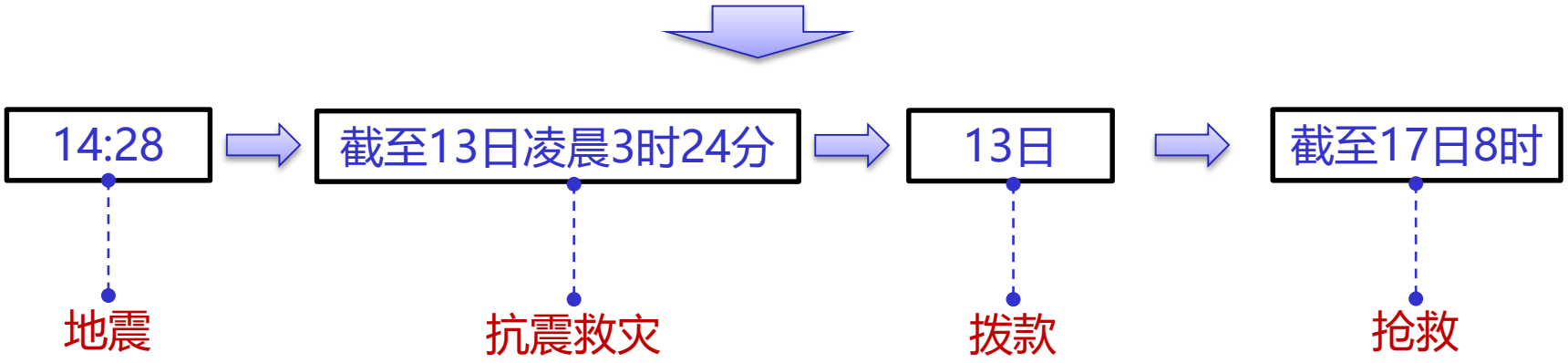
据了解，代表团将在乌鲁木齐停留 3 天。活动暂定为 8 日下午，新疆维吾尔自治区领导会见代表团成员； 9 日上午在新疆师范大学与新疆 6 所高校师生交流，下午在新疆人民会堂举行报告会和体育表演； １０日参观乌鲁木齐市容。“奥运健儿祖国西部行”代表团在结束乌鲁木齐的活动后，将于１１日前往南疆喀什。



Temporal Relation

- 数据来源：百度百科

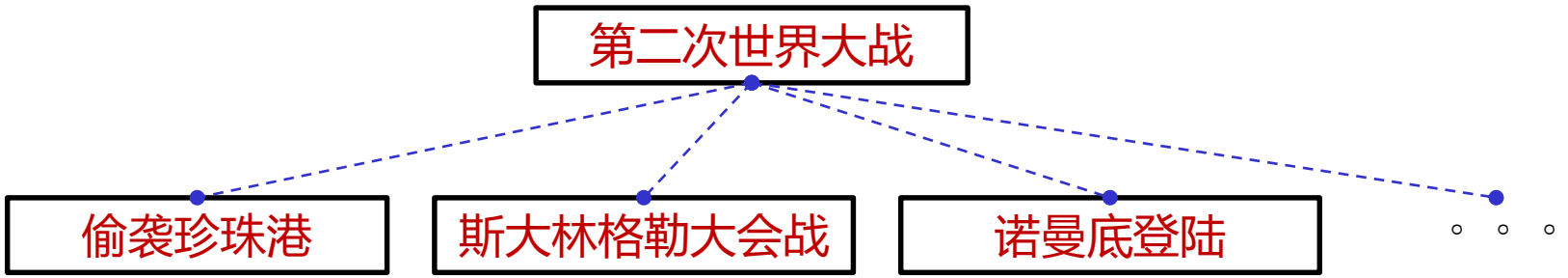
5月12日14:28，四川汶川发生7.8级地震。截至13日凌晨3时24分，武警部队已出动13000余名官兵急赴灾区抗震救灾。13日，为帮助地震灾区开展紧急救灾工作，中央财政紧急下拨款地震救灾资金8.6亿元。截至17日8时，参加救援的民兵预备役部队人员已从灾区废墟中一共抢救出1352名幸存者，发现和掩埋遇难者遗体2756人，救治伤员7600多人。



Subevents Relation

- 数据来源：百度百科

偷袭珍珠港使第二次世界大战的规模迅速扩大,发展成世界大战。斯大林格勒大会战是苏联军队在苏联卫国战争中对德国军队的一次决定性战役，也是欧洲东线战场的转折。苏军在这场保卫战中所取得的胜利具有巨大的战略意义，不仅扭转了苏德战场的整个形势，而且成为第二次世界大战的根本转折点。而诺曼底登陆是规模最大,最成功的战役.是苏联免于单独作战,同时让德国成为两线作战.开辟第 二战场，为二战结束奠定基础.



事件抽取评测及语料

	MUC	TDT	ACE(KBP)	ECB+
Name	Message Understanding Conference	Topic Detection and Tracking	Automatic Content Extraction	EventCorefBank
Organizer	DARPA	DARPA	NIST	UA
Time	1987-1997	1998-2004	ACE：2000-2008 KBP：2014-2018	2014
Task	Event Extraction	Event Extraction	Event Extraction & Coreference Relation	Event Extraction & Coreference Relation

我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

特征层面

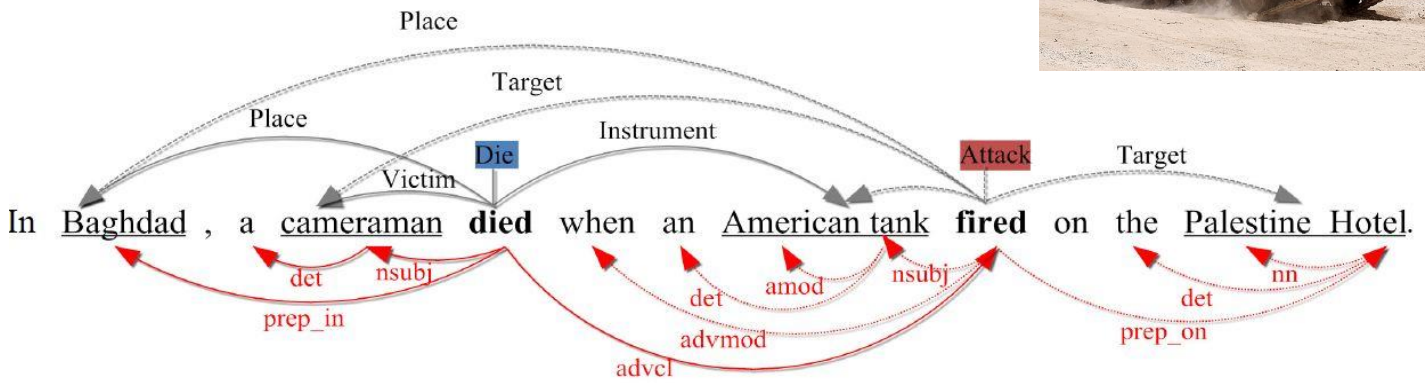
- 词汇级特征
  - 歧义性
  - 人工设计、独热表示

    S1: Obama **beats** McCain. (Elect 事件)

    S2: Tyson **beats** his opponent. (Attack 事件)

特征层面

- 句子级特征
  - 传统的方法依赖于句法分析等复杂的NLP工具
    - 很多语言缺少上述自然语言处理工具
    - 导致处理过程中的误差累积

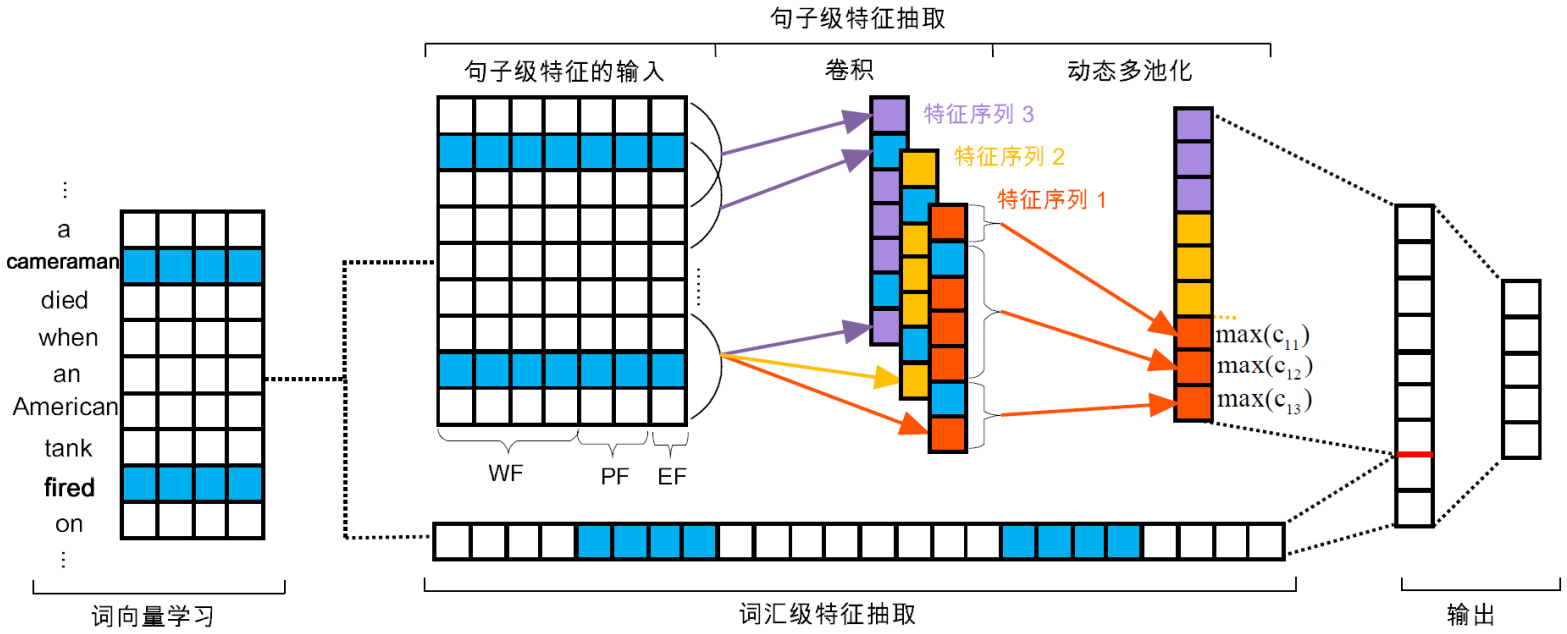


nsubj -> (cameraman 扮演 Victim 角色 在 Die 事件)

????-> (cameraman 扮演 Target 角色 在 Attack 事件)

特征层面

- 基于动态多池化卷积神经网络的事件抽取方法 (ACL2015)



主要由四部分组成：

词向量学习、词汇级特征抽取、句子级特征抽取、分类输出

特征层面

- 实验
  - 测试集：来自ACE2005英文数据集的 40篇新闻
  - 开发集：来自ACE2005英文数据集的 30篇标注文档
  - 训练集：剩下的529篇标注文档

  - 结果

方法	触发词识别 (%)			触发词分类 (%)			元素识别 (%)			元素分类 (%)		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Li's baseline	76.2	60.5	67.4	74.5	59.1	65.9	74.1	37.4	49.7	65.4	33.1	43.9
Liao's cross-event	N/A			68.7	68.9	68.8	50.9	49.7	50.3	45.1	44.1	44.6
Hong's cross-entity	N/A			72.9	64.3	68.3	53.4	52.9	53.1	51.6	45.5	48.3
Li's structure	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 model	80.4	67.7	**73.5**	75.6	63.6	**69.1**	68.8	51.9	**59.1**	62.2	46.9	**53.5**

与传统方法的对比

阶段	方法	1/1	1/N	All
		F_1	F_1	F_1
触发词分类	Embedding+T	68.1	25.5	59.8
	CNN	72.5	43.1	66.3
	DMCNN	**74.3**	**50.9**	**69.1**
元素分类	Embedding+T	37.4	15.5	32.6
	CNN	51.6	36.6	48.9
	DMCNN	**54.6**	**48.7**	**53.5**

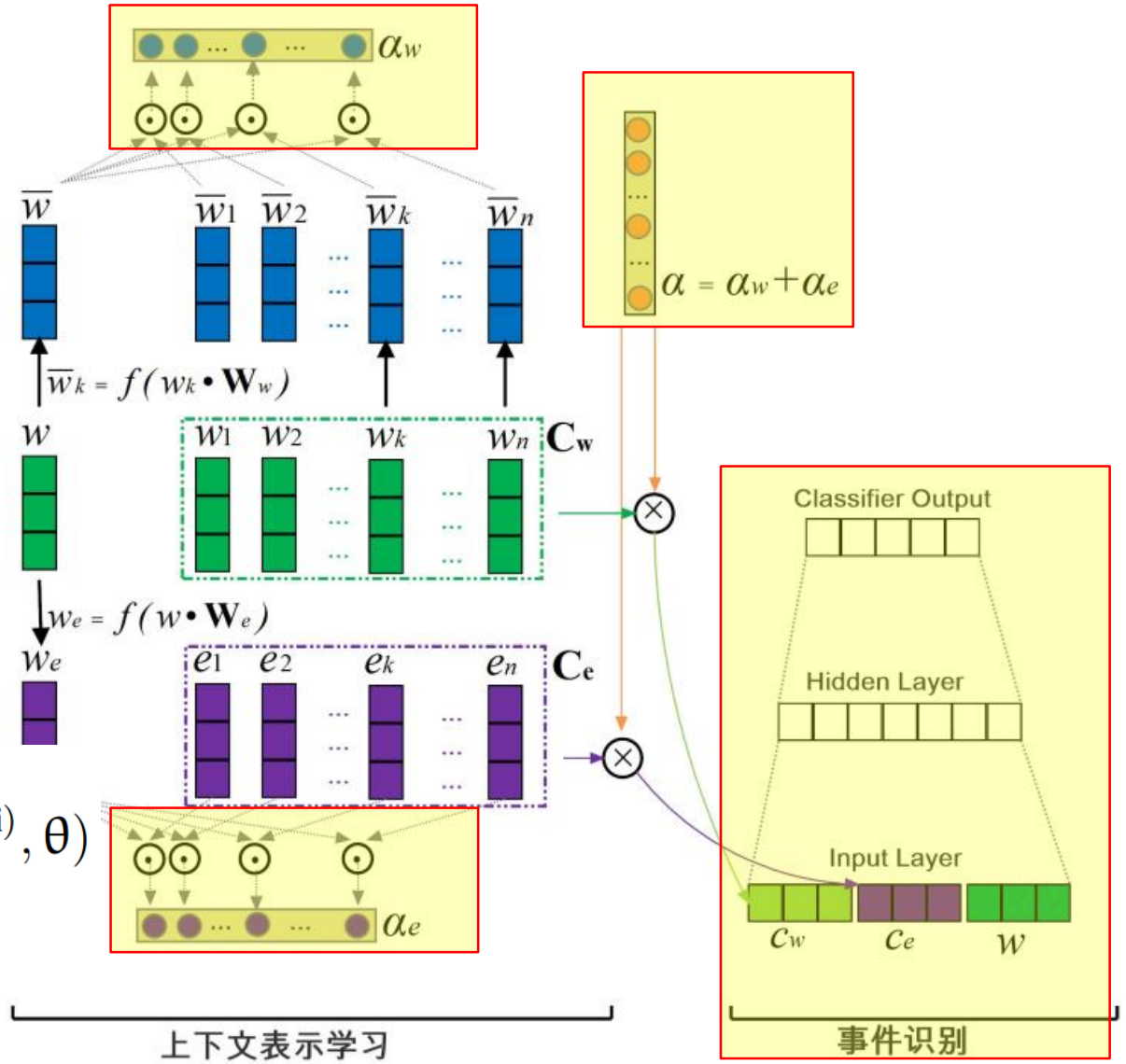
与CNN等方法的对比

我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

特征层面

基于有监督关注机制的事件识别（ACL2017）

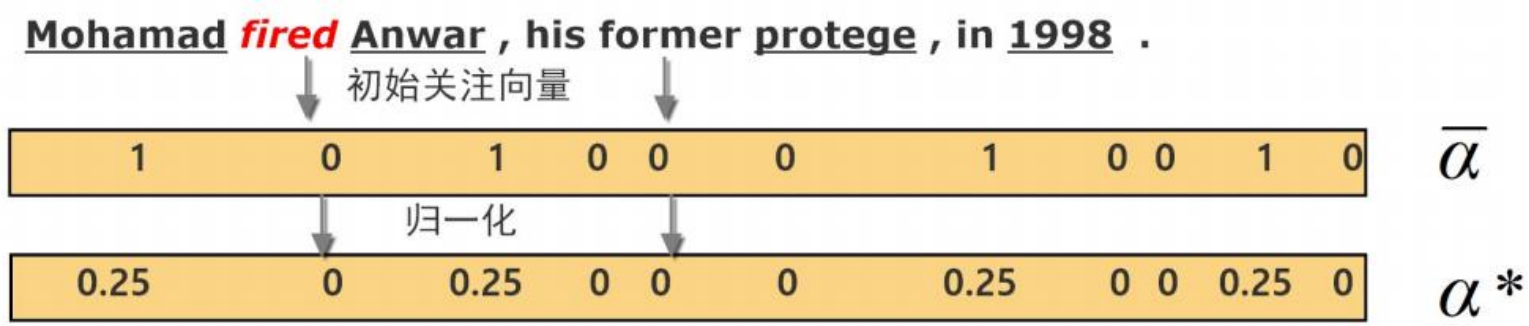


$$J(\theta) = -\sum_{i=1}^{T} \log p(y^{(i)}|\mathbf{x}^{(\mathbf{i})}, \theta)$$

特征层面

构建标准关注向量

策略1：只关注角色词



策略2：同时关注角色词及其周围的词，距离角色词越近的词受到的关注度越高

特征层面

关注机制的损失函数

$$D(\theta) = \sum_{i=1}^{T} \sum_{j=1}^{n} (\alpha^{*i}_{j} - \alpha^{i}_{j})^2$$

关注机制和事件识别模型的融合

$$J'(\theta) = J(\theta) + \lambda D(\theta)$$

特征层面

- 实验

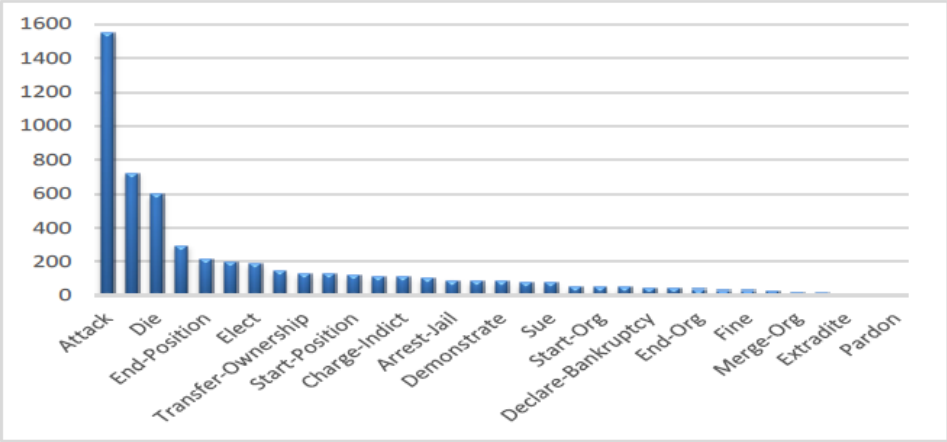
方法	正确率（%）	召回率（%）	F_1 值（%）
Li's joint model (2013)	73.7	62.3	67.5
Liu's PSL (2016)	75.3	64.4	69.4
Liu's FN-Based (2016)	77.6	65.2	70.7
Ngyuen's joint (2016)	66.0	**73.0**	69.3
Nguyen's Skip-CNN (2016)	N/A		71.3
ANN	73.2	57.9	64.6
ANN-S1†	**81.4**	62.4	70.8
ANN-S2†	78.0	66.3	**71.7**

我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

语料层面

ACE2005　33类事件 599篇标注文档



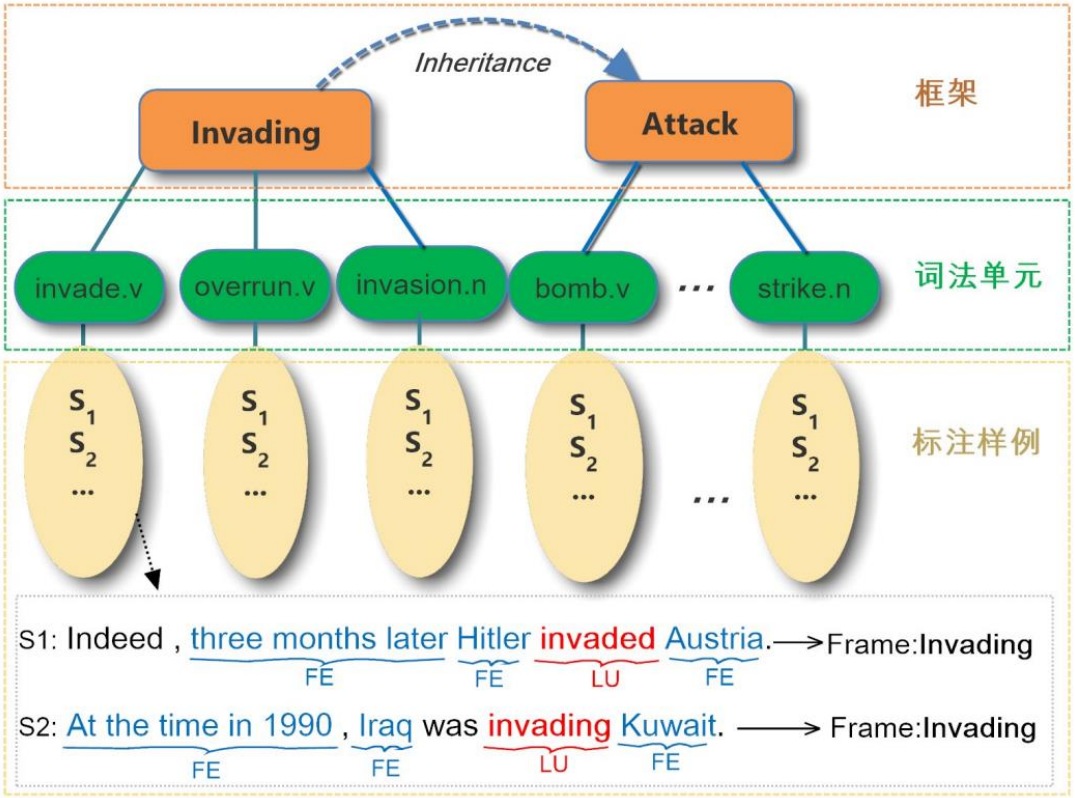
事件类型	标注个数
机构合并（Merge-Org）	14
人事提名（Nominate）	12
司法引渡（Extradite）	7
无罪释放（Acquit）	6
假释出狱（Release-Parole）	2

语料层面

运用FrameNet扩展语料（ACL2016）

- FrameNet
  - 语言学家定义及标注的语义框架资源
  - 层级的组织结构

- FrameNet规模
  - 1000+ 框架
  - 10000+ 词法单元
  - 150000+ 标注例句



语料层面

- 结构相似性
  - 框架：一个词法单元和若干框架元素
  - 事件： 一个触发词和若干事件角色

- 含义相似性

框架名称	FrameNet 中的例句	事件类型
Attack	*Aeroplane **bombed** London.*	Attack
Rape	*Girl, 16, **raped** by gang.*	Attack
Fining	*The court **fined** her $40.*	Fine
Execution	*He was **executed** yesterday.*	Execute

语料层面

- 系统框架图

语料层面

- 实验

词法单元	事件类型	$N_e/\|\|S_l\|\|$	ψ
gunfight.n	Attack	14/14	1.0
injure.v	Injure	14/14	1.0
divorce.n	Divorce	11/11	1.0
decapitation.n	Die	5/5	1.0
trial.n	Trial-Hearing	25/25	1.0
assault.v	Attack	21/21	1.0
fight.v	Attack	12/12	1.0
arrest.n	Arrest-Jail	38/38	1.0
divorce.v	Divorce	35/35	1.0
shoot.v	Attack	2/2	1.0

框架名称	事件类型	$N_e/\|\|S_f\|\|$	ϕ
Hit_target	Attack	2/2	1.0
Relational_natural_features	Meet	1/1	1.0
Invading	Attack	120/121	0.99
Fining	Fine	26/27	0.96
Being_born	Be-Born	32/36	0.88
Rape	Attack	104/125	0.83
Sentencing	Sentence	57/70	0.81
Attack	Attack	99/129	0.77
Quitting	End-Position	102/137	0.74
Notification_of_charges	Charge-Indict	73/103	0.71

语料层面

- 实验

方法	正确率 (%)	召回率 (%)	F_1 值 (%)
Nguyen's CNN(2015)	71.8	**66.4**	69.0
Chen's DMCNN(2015)	75.6	63.6	69.1
Liu's Approach(2016)	75.3	64.4	69.4
ANN (Ours)	**79.5**	60.7	68.8
ANN-FN (Ours)	77.6	65.2	**70.7**

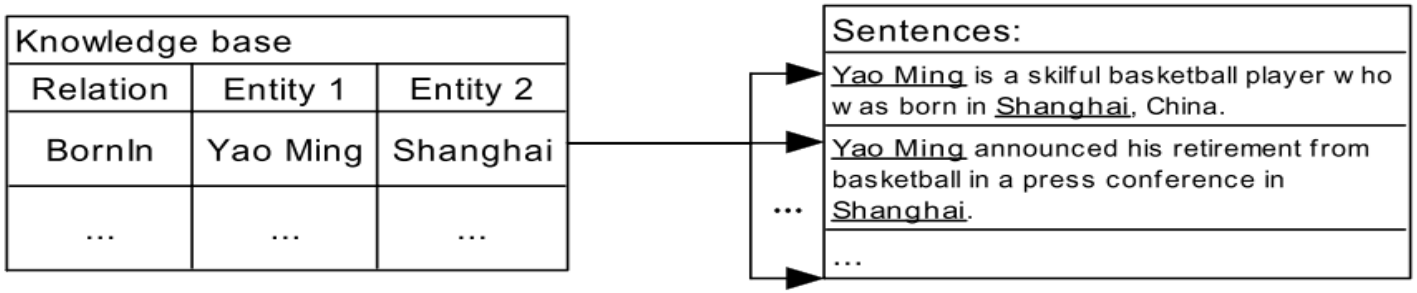
我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

语料层面

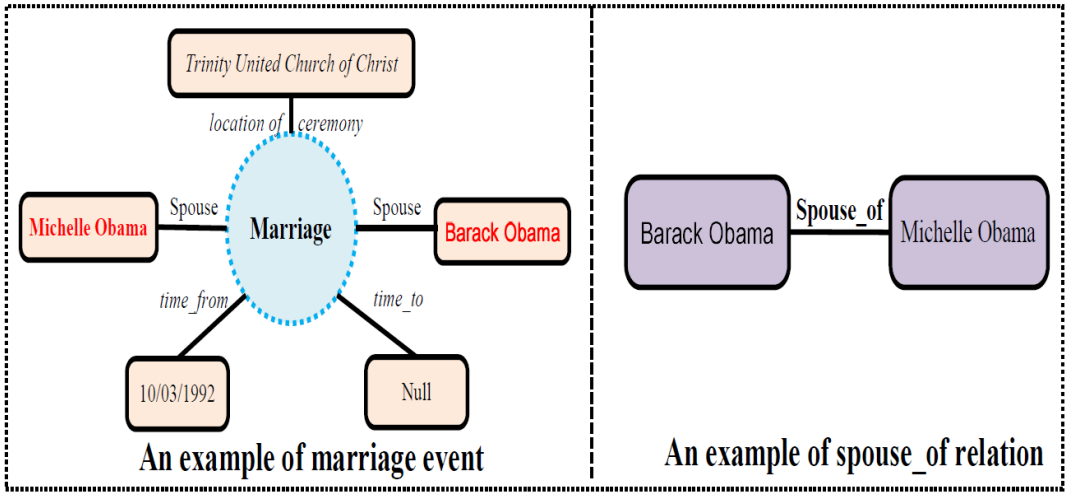
运用结构化的知识库自动生成语料（**ACL2017**）

- 远距离监督的方法在关系抽取中取得成功



语料层面

- 现有事件知识库中缺少触发词信息



– 关系抽取：（实体1，关系, 实体2)

可以利用Michelle Obama 和 Barack Obama回标

– 事件抽取：(事件实例，事件类型；角色1，事件元素1，角色2，事件元素2;...; 角色N, 事件元素N)

无法利用m.02nqglv 和 Barack Obama

– ACE中，用触发词代表一个事件实例

语料层面

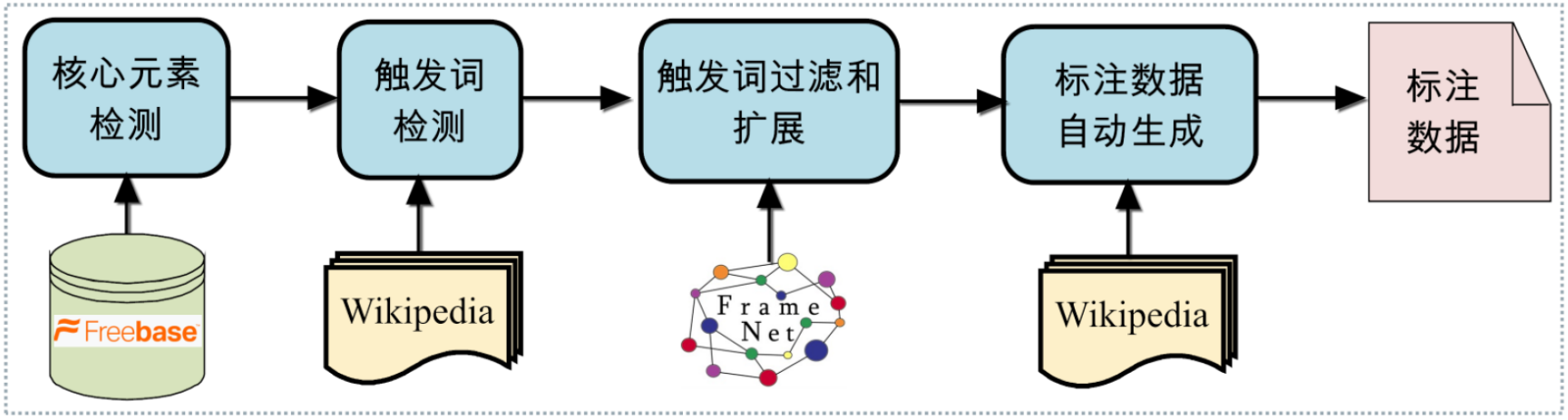
- 一个句子的多个元素会出现在不同的句子中

事件类型	EI#	A#	S#
education.education	530,538	8	0
film.film_crew_gig	252,948	3	8
people.marriage	152,276	5	0
...	...	...	...
military.military_service	27,933	6	0
olympics.olympic_medal_honor	20,790	5	4
sum of the selected 21 events	**3,870,492**	**100**	**798**

  - 知识库中只有60%的事件实例包含所有的事件元素
  - 只有0.02%的事件实例能在一句话中找到所有的事件元素

语料层面

- 利用世界知识和语言学知识
- 自动生成大规模事件语料



语料层面

- 核心元素检测
  - 角色显著性（RS）：反映了一个事件元素区分同一事件类型下不同事件实例的能力

$$RS_{ij} = \frac{Count(A_i, ET_j)}{Count(ET_j)}$$

  - 事件相关性（ER）：反映了一个事件元素区分不同类型的事件的能力

$$ER_i = \log \frac{Sum(ET)}{1 + Count(ETCi)}$$

  - 核心率（KR）：反映了一个事件元素在一个事件中的重要程度

$$KR_{ij} = RS_{ij} * ER_i$$

语料层面

- 事件触发词检测
  - 我们利用所有的核心元素去Wikipedia中回标
  - 触发率（TR）：

$$TR_{ij} = TCF_{ij} * TETF_i$$

  - 触发词频率（TCF）：

$$TCF_{ij} = \frac{Count(V_i, ETS_j)}{Count(ETS_j)}$$

  - 触发词事件频率（TETF）：

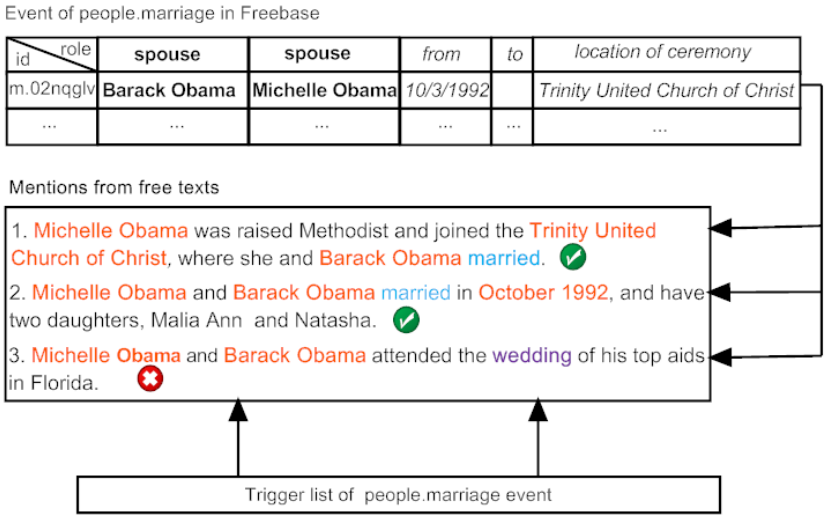
$$TETF_i = \log \frac{Sum(ET)}{1 + Count(ETI_i)}$$

语料层面

- 事件触发词过滤和扩展
  - 上一模块得到的触发词词表只有动词性触发词，缺少名词性触发词而且其中存在噪声
  - 利用语言学知识FrameNet过滤动词性触发词中的噪声并扩展名词性触发词

$$frame(i) = \arg\max_{j}(similarity(e_i, e_{j,k}))$$

- 标注数据的自动生成



语料层面

- 自动生成的数据

Event Type	Freebase Size	Sentences (KA)	Sentences (KA+T)	Examples of argument roles sorted by KR	Examples of triggers
people.marriage	152,276	56,837	26,349	spouse, spouse, from, to, location	marriage, marry, wed, wedding, couple,..., wife
music.group_membership	239,813	90,617	20,742	group, member, start, role, end	musician, singer, sing, sang, sung, concert,..., play
education.education	530,538	26,966	11,849	student, institution, degree,..., minor	educate, education, graduate, learn, study,..., student
organization.leadership	43,610	5,429	3,416	organization, person, title,..., to	CEO, charge, administer, govern, rule, boss,..., chair
olympics.olympic_medal_honor	20,790	4,056	2,605	medalist, olympics, event,..., country	win, winner, tie, victor, gold, silver,..., bronze
...	...	...	...	...	...
sum of 21 selected events	3,870,492	421,602	72,611	argument1, argument2 ,..., argumentN	trigger1, trigger2, trigger3, ... , triggerN

  - 当仅利用两个核心元素回标时，生成421,602个标注数据，但是这个数据中没有标注触发词信息
  - 当同时利用核心元素和事件触发词回标时，生成72,611个标注数据
  - 与ACE人工标注的将近6,000个的标注数据相比，我们提出的方法能自动生成大规模训练数据

语料层面

- 标注数据的人工评价

##001 He is the uncle of [**Amal Clooney**], [**wife**] of the actor [**George Clooney**].
Trigger: wife **Event Type:** Marriage **MannalAnotate[Y/N]:**
Argument: Amal Clooney **Role:**Spouse **MannalAnotate[Y/N]:**
Argument: George Clooney **Role:**Spouse **MannalAnotate[Y/N]:**

##002 She was [**married**] to the cinematographer [**Theo Nischwitz**] and was sometimes credited as [**Gertrud Hinz-Nischwitz**].
Trigger: married **Event Type:** Marriage **MannalAnotate[Y/N]:**
Argument: Theo Nischwitz **Role:**Spouse **MannalAnotate[Y/N]:**
Argument: Gertrud Hinz-Nischwitz **Role:**Spouse **MannalAnotate[Y/N]:**

- 随机地从标注数据中选择500 个样例，重复三次以平均正确率作为人工评价的结果
- 评判中每个句子都由三个标注者评价，最终投票决定

阶段	平均正确率 (%)
触发词标注	88.9
元素标注	85.4

语料层面

- 标注数据的自动评价
  - 数据：ACE，ED only， ACE+ ED
  - 评价指标：同ACE一样
  - 实验结果：

Methods	Trigger Identification(%)			Trigger Identification + Classification(%)			Argument Identification(%)			Argument Role(%)		
	P	R	F	P	R	F	P	R	F	P	R	F
Li's structure trained with ACE	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
Chen's DMCNN trained with ACE	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5
Nguyen's JRNN trained with ACE	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 trained with ED Only	77.6	67.7	72.3	72.9	63.7	68.0	64.9	51.7	57.6	58.7	46.7	52.0
DMCNN trained with ACE+ED	79.7	69.6	**74.3**	75.7	66.0	**70.5**	71.4	56.9	**63.3**	62.8	50.1	**55.7**

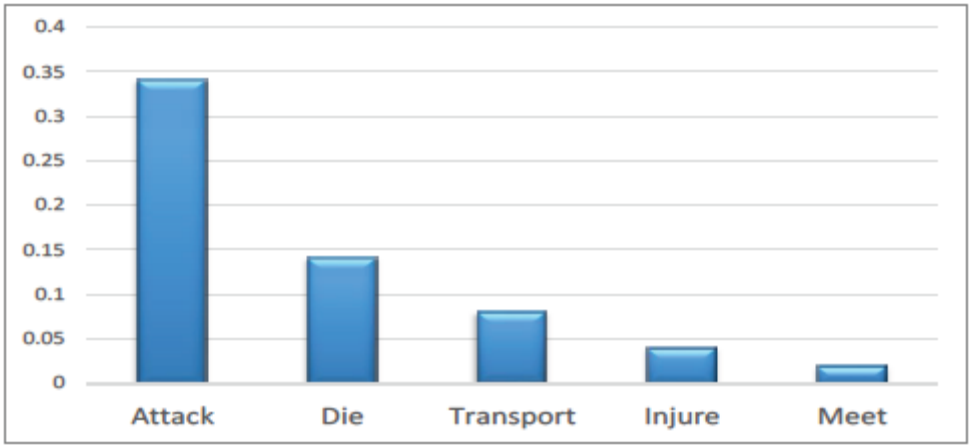
ACE+ED取得到最好的效果，证明自动生成的数据可以有效扩展人工标注的数据

我们的工作

- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

多事件协同抽取

- 一句话的多个事件之间具有依存关系



Die

S1:*In Baghdad, a cameraman* ***died*** *when an American tank* ***fired*** *on the Palestine hotel.*

Attack or End-position ?

S2: *The project leader was* ***fired*** *for the* ***bankruptcy*** *of the subsidiary company.*

Bankruptcy

多事件协同抽取

- 句子级信息 v.s. 篇章级信息

篇章级信息更重要的情况:

Transport

He planned to go shopping before he went home, because he got off work early today

S3: He **left** the company.

End-position

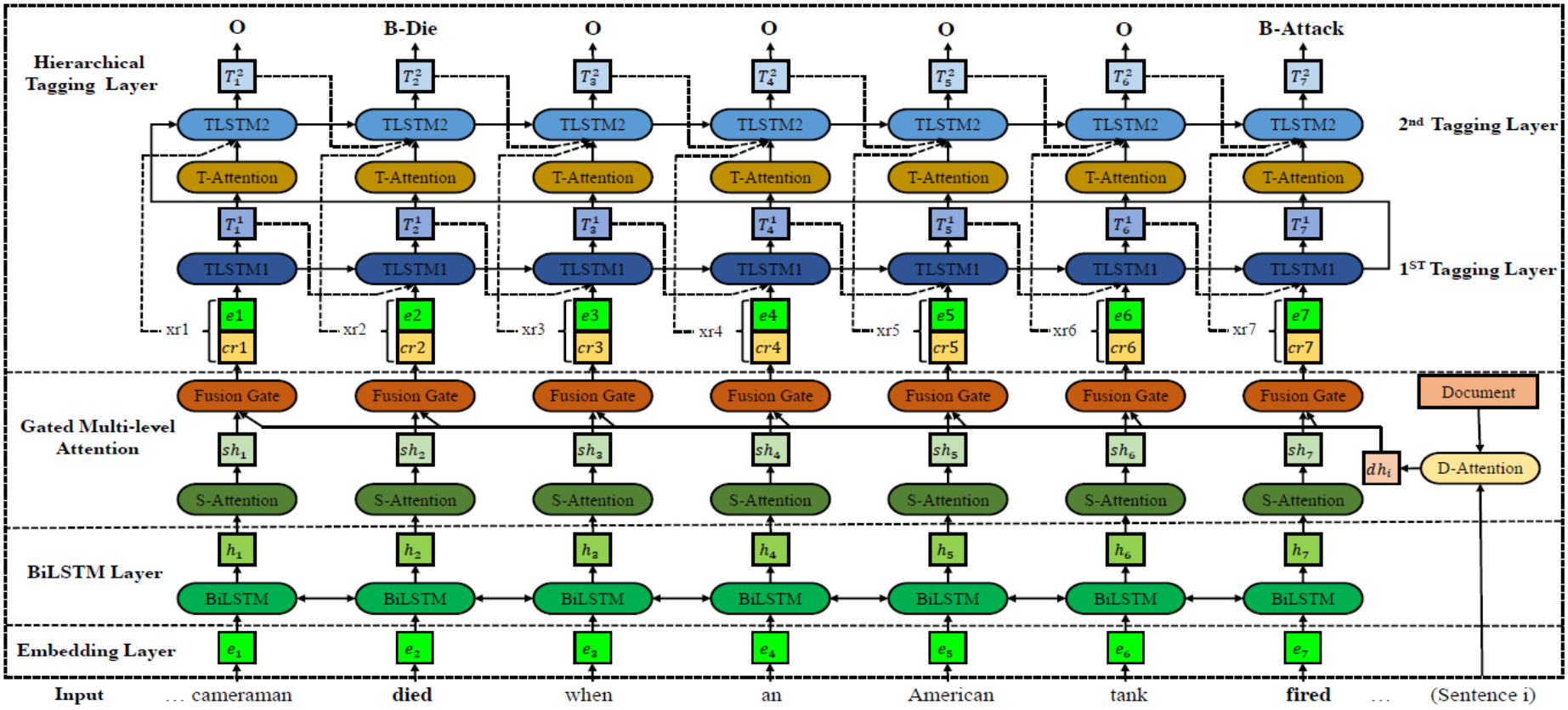
They held a party for his retirement.

句子级信息更重要的情况:

S1:*In Baghdad, a cameraman **died** when an American tank **fired** on the Palestine hotel.*

多事件协同抽取

- 基于层次偏置标注网络和多粒度门控关注机制的多事件协同抽取



多事件协同抽取

- 实验结果

Methods	P	R	F_1
Li's MaxEnt (2013)	74.5	59.1	65.9
Liao's CrossEvent (2010)	68.7	68.9	68.8
Hong's CrossEntity (2011)	72.9	64.3	68.3
Chen's DMCNN (2015)	75.6	63.6	69.1
Chen's DMCNN+ † (2017)	75.7	66.0	70.5
Liu's FrameNet † (2016a)	77.6	65.2	70.7
Liu's ANN-Aug † (2017)	76.8	67.5	71.9
Li's Structure (2013)	73.7	62.3	67.5
Yang's JointEE (2016)	75.1	63.3	68.7
Nguyen's JRNN (2016)	66.0	73.0	69.3
Liu's PSL (2016b)	75.3	64.4	69.4
Ours HBTNGMA	**77.9**	**69.1**	**73.3**

Method	1/1	1/N	all
LSTM+Softmax	74.7	44.6	66.8
LSTM+CRF	75.1	49.5	68.5
LSTM+TLSTM	76.8	51.2	70.2
LSTM+HTLSTM	77.9	57.3	72.4
LSTM+HTLSTM+Bias	**78.4**	**59.5**	**73.3**

Table 2: Performance of different ED systems. 1/1 means one sentence that only has one event and 1/N means that one sentence has multiple events.

Method	P	R	F_1
Word Only	70.1	63.4	66.6
Word+SA	75.6	68.2	71.7
Word+DA	73.1	65.8	69.3
Word+Average MA	76.5	68.7	72.4
Word+Gated MA	**77.9**	69.1	**73.3**

Table 3: Performance of gated multi-level attention.

我们的工作

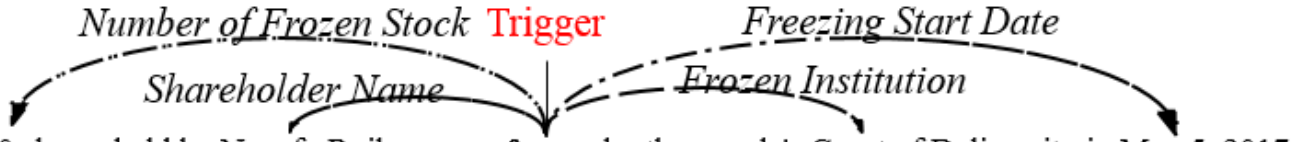
- **特征表示**
  - **Dynamic CNN**: Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks (ACL2015)
  - **Argument Attention**: Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms (ACL2017)
- **训练数据生成和扩展**
  - **External Resources:** Exploring Leveraging FrameNet to Improve Automatic Event Detection (ACL2016)
  - **Distant Supervision:** Automatically Labeled Data Generation for Large Scale Event Extraction (ACL2017)
  - **Multilingual Resources:** Event Detection via Gated Multilingual Attention Mechanism （AAAI2018）
- **多事件协同抽取**
  - **HBTNGMA:** Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention (EMNLP-2018)
- **篇章级事件抽取**
  - **DCFEE:** A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)
- **事件关系抽取**
  - **ATT-ERNN:** Attention-based Event Relevance Model for Stock Price Movement Prediction (CCKS-2017 Best Paper Award)
  - **MLNN:** Event Coreference Resolution via Multi-loss Neural Network without Arguments （CCKS-2018）

任务定义：篇章级金融事件抽取

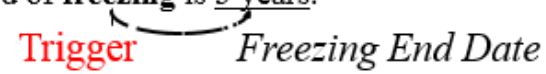
- **股权冻结事件示例：**

证券代码：600747股票简称：大连控股编号：临2017-04大连大福控股股份有限公司关于大股东股份冻结的公告本公司董事会及全体董事保证本公告内容不存在任何虚假记载、误导性陈述或者重大遗漏，并对其内容的真实性、准确性和完整性承担个别及连带责任。
公司于近日收到通知，公司第一大股东长富瑞华持有的上市公司520,000股被大连市人民法院于2017年5月5日冻结。冻结期限为3年。自转为正式冻结之日起计算。
本次轮候冻结包括孳息（包括派发的送股、转增股及现金红利），其效力从登记在先的冻结证券解除冻结且本次轮候冻结部分或全部生效之日起产生。
此次轮候冻结股数占公司总股本的35.51%,截止本公告日，长富瑞华持有本公司股份520,000,000股，占公司总股本35.51%；此次股份冻结后累计股份冻结的数量520,000,000股，占公司总股本的35.51%，经公司向大股东长富瑞华了解，此次股份冻结事项不会对公司的控制权造成影响，也不影响公司正常经营,长富瑞华将与相关方积极协商妥善处理解决相关事宜,公司将密切关注该事项的进展并及时履行信息披露义务,特此公告。
大连大福控股股份有限公司董事会二〇一七年一月十三日2

C1:长富瑞华持有的上市公司520,000股被大连市人民法院于2017年5月5日**冻结**。
C2:**冻结**期限为3年。



E1:The 520,000 shares held by Nagafu Ruihua were **frozen** by the people's Court of Dalian city in May 5, 2017.
E2:The period of **freezing** is 3 years.

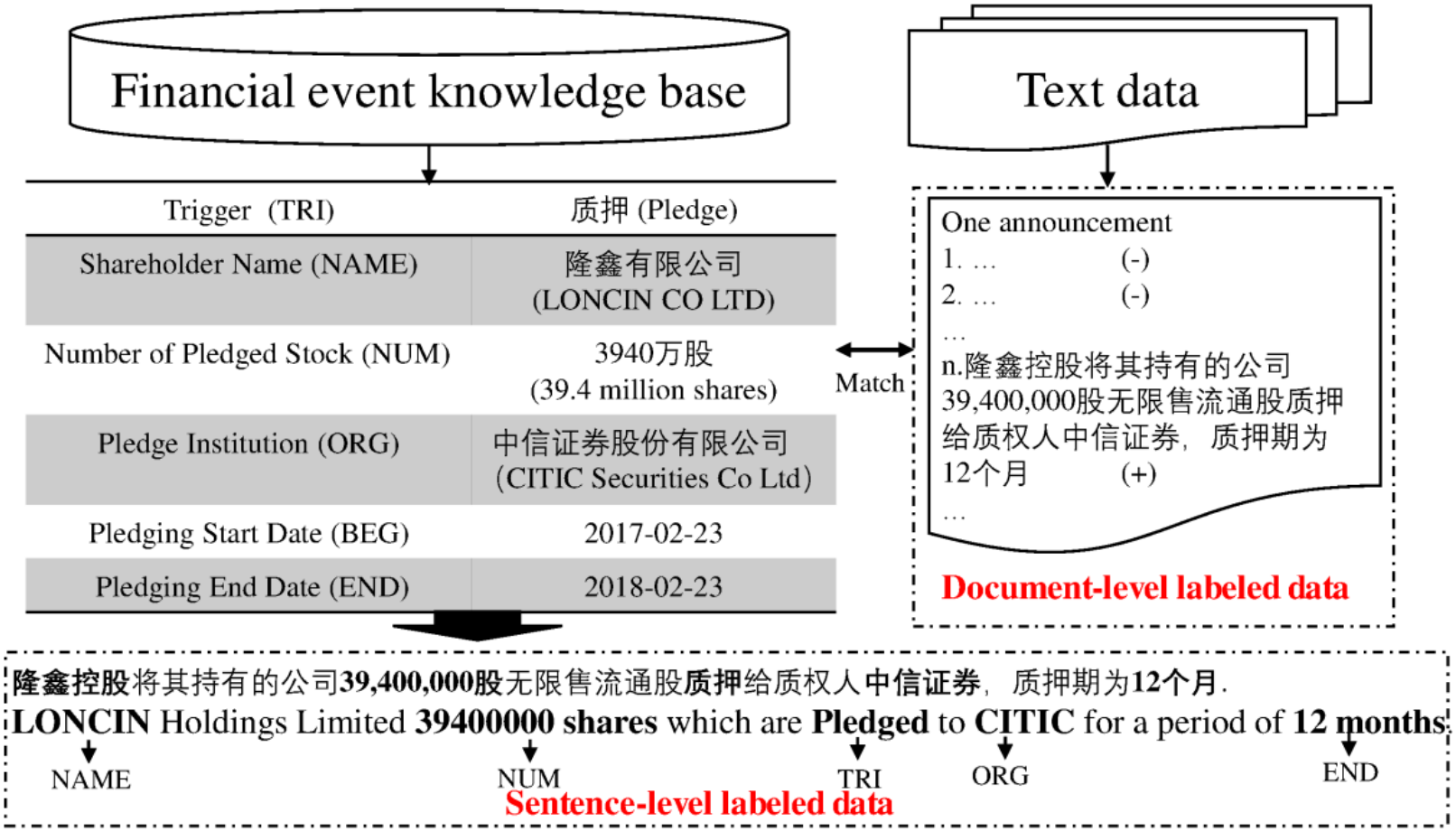


挑战

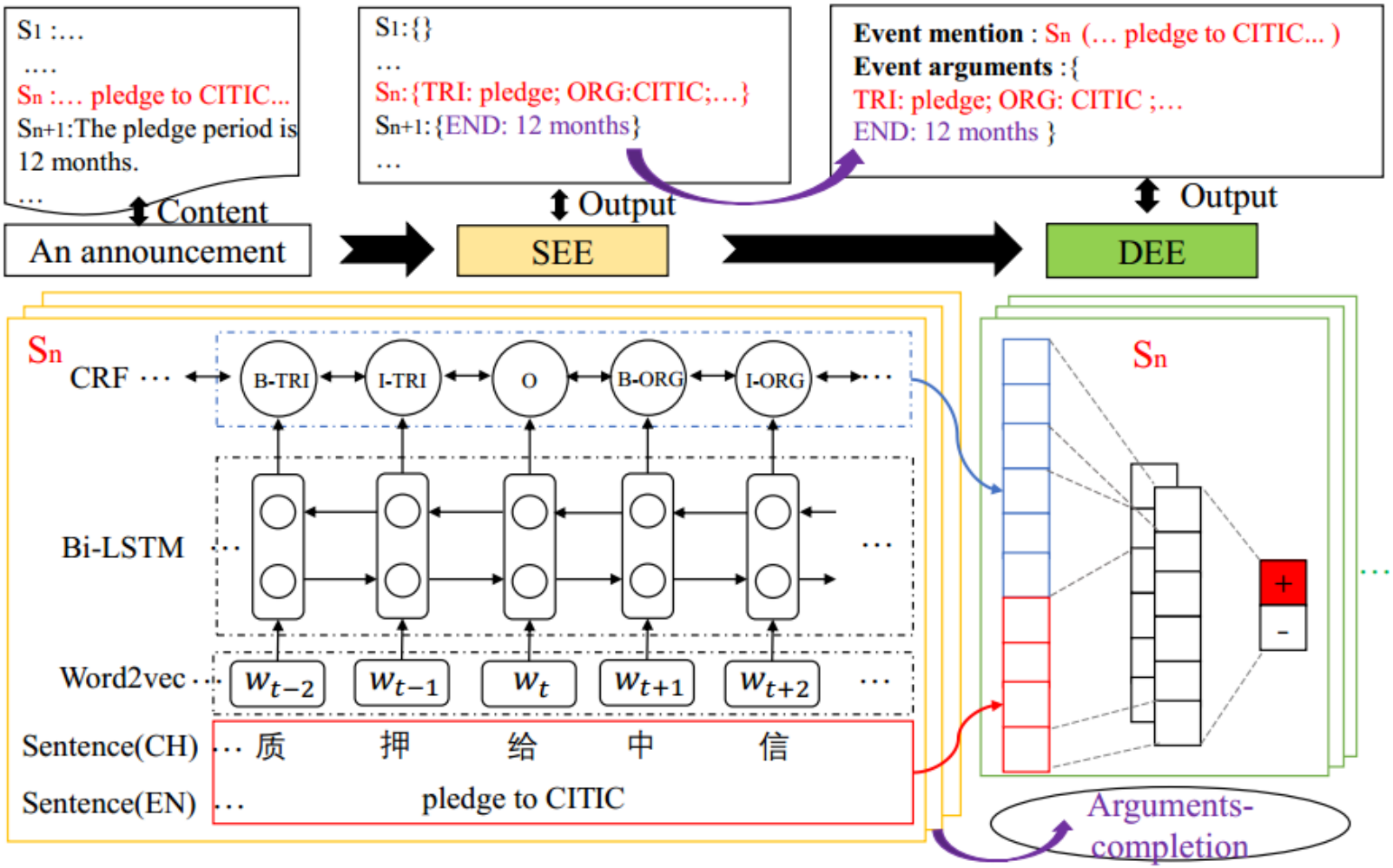
- **标注数据的缺失**
  - 现有的事件抽取系统性能都依赖于人工标注数据
  - 人工标注数据耗时费力，成本高昂，金融领域缺乏大规模高质量的标注数据

- **篇章级事件抽取**
  - 目前大多数的事件抽取系统都是针对一个句子进行抽取
  - 通常由多个句子描述一个事件，一个事件的多个元素分布在不同的句子中

自动生成标注数据


Financial event knowledge base
Text data
Trigger (TRI)
质押 (Pledge)
Shareholder Name (NAME)
隆鑫有限公司 (LONCIN CO LTD)
Number of Pledged Stock (NUM)
3940万股 (39.4 million shares)
Pledge Institution (ORG)
中信证券股份有限公司 (CITIC Securities Co Ltd)
Pledging Start Date (BEG)
2017-02-23
Pledging End Date (END)
2018-02-23
Match
One announcement
1. … (-)
2. … (-)
…
n.隆鑫控股将其持有的公司39,400,000股无限售流通股质押给质权人中信证券，质押期为12个月 (+)
…
Document-level labeled data
隆鑫控股将其持有的公司39,400,000股无限售流通股质押给质权人中信证券，质押期为12个月.
LONCIN Holdings Limited 39400000 shares which are Pledged to CITIC for a period of 12 months
NAME
NUM
TRI
ORG
END
Sentence-level labeled data

篇章级事件抽取


S1 :…
….
Sn :… pledge to CITIC...
Sn+1:The pledge period is 12 months.
…
Content
An announcement
S1:{}
…
Sn:{TRI: pledge; ORG:CITIC;…}
Sn+1:{END: 12 months}
…
Output
SEE
Event mention : Sn (… pledge to CITIC...)
Event arguments :{
TRI: pledge; ORG: CITIC ;…
END: 12 months }
Output
DEE
Sn
CRF
B-TRI
I-TRI
O
B-ORG
I-ORG
Bi-LSTM
Word2vec
w_{t-2}
w_{t-1}
w_t
w_{t+1}
w_{t+2}
Sentence(CH) 质 押 给 中 信
Sentence(EN) pledge to CITIC
Sn
+
-
Arguments-completion

实验结果

- 数据集

Dataset	NO.BUL	NO.POS	NO.NEG
EF	526	544	2960
EP	752	775	6392
EB	1178	1192	11590
EI	520	533	11994
Total	2976	3044	32936

- 系统性能

Stage	SEE			DEE		
Type	$P(\%)$	$R(\%)$	$F_1(\%)$	$P(\%)$	$R(\%)$	$F_1(\%)$
EF	90.00	90.41	90.21	80.70	63.40	71.01
EP	93.31	94.36	93.84	80.36	65.91	72.30
EB	92.79	93.80	93.29	88.79	82.02	85.26
EI	88.76	91.88	90.25	80.77	45.93	58.56

金融领域的应用


概览
全部分析结果
冻结事件 1
事件触发词 3
冻结股数 3
冻结起始日期 2
冻结截止日期 1
执行冻结机构 2
被冻结股东名称 8
文本
返回
长久物流：关于控股股东股份冻结的公告","原始公告":"证券代码：603569证券简称：长久物流公告编号：2016-【002】北京长久物流股份有限公司关于控股股东股份冻结的公告本公司董事会及全体董事保证本公告内容不存在任何虚假记载、误导性陈述或者重大遗漏，并对其内容的真实性、准确性和完整性承担个别及连带责任。北京长久物流股有限公司（以下简称“长久物流”或“公司”）于近日收到上海证券交易所通知，公司控股股东吉林省长久实业集团有限公司（以下简称“长久集团”）持有公司的部分股份被冻结，具体情况如下：一、股份冻结的具体情况因长久集团与中国光大银行股份有限公司长春分行（以下简称“光大长春”）及第三人大理鸿迪汽车销售服务有限公司（以下简称“大理鸿迪”）合同纠纷一案，光大长春向吉林省长春市中级人民法院（以下简称“长春中院”）申请冻结长久集团在银行的存款人民币4,307万元或查封、扣押其相应价值的财产。 长春中院根据其作出的（2016）吉-01民初字第682-1号《民事裁定书》和《协助执行通知书》，对长久集团持有的公司4,307万股股份（限售流通股）进行司法冻结。 此次冻结包括孳息（包括派发的送股、转增股利及现金股利），冻结期限为两年，从2016年8月12日至2018年8月11日止。截止本公告出具日，长久集团持有公司30,463.618万限售股，占公司总股本的76.16%。本次冻结股份后，长久集团所

金融领域的应用


概览
全部分析结果
冻结事件 1
事件触发词 1
冻结股数 1
冻结起始日期 1
冻结截止日期 1
执行冻结机构 1
被冻结股东名称 1
文本
返回
长久物流：关于控股股东股份冻结的公告","原始公告":"证券代码：603569证券简称：长久物流公告编号：2016-【002】北京长久物流股份有限公司关于控股股东股份冻结的公告本公司董事会及全体董事保证本公告内容不存在任何虚假记载、误导性陈述或者重大遗漏，并对其内容的真实性、准确性和完整性承担个别及连带责任。北京长久物流股有限公司（以下简称“长久物流”或“公司”）于近日收到上海证券交易所通知，公司控股股东吉林省长久实业集团有限公司（以下简称“长久集团”）持有公司的部分股份被冻结，具体情况如下：一、股份冻结的具体情况因长久集团与中国光大银行股份有限公司长春分行（以下简称“光大长春”）及第三人大理鸿迪汽车销售服务有限公司（以下简称“大理鸿迪”）合同纠纷一案，光大长春向吉林省长春市中级人民法院（以下简称“长春中院”）申请冻结长久集团在银行的存款人民币4,307万元或查封、扣押其相应价值的财产。 长春中院根据其作出的（2016）吉-01民初字第682-1号《民事裁定书》和《协助执行通知书》，对长久集团持有的公司4,307万股股份（限售流通股）进行司法冻结。 此次冻结包括孳息（包括派发的送股、转增股利及现金股利），冻结期限为两年，从2016年8月12日至2018年8月11日止。截止本公告出具日，长久集团持有公司30,463.618万限售股，占公司总股本的76.16%。本次冻结股份后，长久集

总结

- 事件知识不可或缺
  - 企业信息监控
  - 风险信用控制
  - 智能投顾
- 通用领域的事件抽取很难
  - 大规模、高质量的训练数据
  - 鲁棒的特征表示
- 限定域的事件抽取有可能取得不错的性能
  - 文本类型受限
  - 语言表示规律性较强，知识密集
- 未来工作
  - 篇章级事件抽取
  - 事件关系抽取

谢谢！

Questions?