

低资源事件抽取关键技术研究

(申请清华大学工学博士学位论文)

培养单位：计算机科学与技术系

学 科：计算机科学与技术

研 生：全 美 涵

指 导 教 师：许 磐 副研究员

二〇二二年五月

Low-resource Event Extraction

Thesis Submitted to
Tsinghua University
in partial fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Computer Science and Technology

by

Meihan Tong

Thesis Supervisor: Associate Professor Bin Xu

May, 2022

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

朱小燕	教 授	清华大学
赵 军	研究员	中科院自动化所

答辩委员会名单

主席	赵 军	研究员	中科院自动化所
委员	孙茂松	教 授	清华大学
	李涓子	教 授	清华大学
	孙 乐	研究员	中科院软件所
	许 斌	副研究员	清华大学
秘书	侯 磊	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）按照上级教育主管部门督导、抽查等要求，报送相应的学位论文。

本人保证遵守上述规定。

作者签名：_____

导师签名：_____

日 期：_____

日 期：_____

摘要

事件抽取旨在从非结构化的文本信息中挖掘用户感兴趣的事件，并进行结构化的归纳和整理，使得用户能够快速、准确、全面了解事情发展的全貌。近年来，深度学习模型在事件抽取上取得了不错的效果，但深度学习模型需要大规模标注数据驱动，而获取标注数据的成本很高。因此，研究数据标注稀缺的低资源事件抽取是十分有必要的。本文面向低资源场景下的事件抽取关键技术展开研究，主要研究内容和创新点如下：

针对低资源场景下事件类型消歧难的问题，本文提出了一种基于多模态融合的事件检测方法。该方法通过融合新闻中天然存在的多模态图像信息，为事件消歧提供更多的语义证据，改善了低资源事件检测的消歧效果。相比于传统的单向融合或者拼接融合的方式，本文提出的多模态融合模型能够根据上下文筛选图像模态的重要性信息，提升了融入图像信息的质量。与六个最先进的基线相比，所提方法取得了最佳的效果，证明了所提方法的先进性。

针对低资源场景下事件召回难的问题，本文提出了一种基于知识蒸馏的事件检测方法。该方法通过引入人类常识，为事件识别提供先验，显著的提升了低资源场景下事件的召回率。所提方法利用 WordNet 高效的获取常识知识，之后利用学生拟合教师模型的方式从标注语料和大规模未标注语料上融合常识知识，有很强的常识知识利用能力。实验从定性和定量角度证明了所提的融入外部常识方法对低资源事件检测有效性。

针对低资源场景下事件要素识别难的问题，本文提出了一种基于自标注的事件要素抽取方法。该方法通过挖掘任务相关类，为训练提供更丰富的训练数据，提升低资源事件要素抽取的性能。所提方法借助预定义类的监督信号，可以在没有元数据的情况下从语料中分辨噪音和任务相关类，提升了训练数据的多样性，降低了训练语料单一稀疏带来的事件要素识别过于生硬刻板的问题。实验结果表明，所提方法优于五个最先进的基线模型，证明了所提方法对低资源事件要素抽取的有效性。

最后，本文将研究成果应用于在线新闻挖掘系统 NewsMiner。本文依据现实应用场景重构了事件本体，并构建了大规模中英文事件抽取数据集，在此数据集的基础上搭建了中英文事件抽取系统，体现了本研究的现实应用价值。

关键词：自然语言处理，事件抽取，低资源，多模态，知识蒸馏

Abstract

Event extraction aims to mine events from unstructured text. In recent years, deep learning models have shown superior performance in event extraction. However, deep learning models need to be driven by large-scale labeled data, and obtaining labeled data is time-consuming and expensive. This paper focus on low-resource event extraction. Our contributions can be summarized are as follows:

This paper proposes a multi-modal event detection method to improve event disambiguation performance in low-resource scenarios. Compared with the traditional concatenation fusion or co-attention fusion, the proposed alternative dual attention mechanism show superior performance on heterogeneous information fusion. Compared with the six state-of-the-art baselines, our method achieves the best results. Experiments in low-resource settings also demonstrate that fusing image modal information can effectively alleviate the phenomenon of overfitting.

This paper proposes a knowledge-guided event detection model to improve the recall rate of event detection in low-resource scenarios. The proposed method perform knowledge distillation from both labeled and unlabeled text, which has stronger ability of knowledge utilization. Experiments demonstrate the effectiveness of the proposed model from both qualitative and quantitative perspectives.

This paper proposes a self-labeling model to improve low-resource event argument extraction. Specifically, the proposed model mines multiple task-related classes from the corpus to diversify the training data to alleviate the problem of overfitting. Experimental results show that the proposed model outperforms five state-of-the-art models, demonstrating the effectiveness of the proposed method.

Finally, we apply the proposed methods to an online news mining system News-Miner. To meet the needs of practical applications, we reconstruct the event ontology, and manually annotate a large-scale event extraction dataset. Based on the dataset, we built a Chinese and English event extraction system to show the application value of our research.

Keywords: Natural Language Processing; Event Extraction; Low-resource; Multimodal; Knowledge Distillation

目 录

摘 要	I
Abstract	II
目 录	III
插图和附表清单	VII
第 1 章 绪论	1
1.1 研究背景和意义	1
1.2 存在的挑战	3
1.3 本文的主要研究内容和贡献	5
1.4 论文组织结构	6
第 2 章 相关研究综述	8
2.1 任务定义	8
2.1.1 本文定义	8
2.1.2 相关数据集	9
2.2 事件抽取方法研究框架	13
2.3 基于语义增强的事件抽取方法	16
2.3.1 基于篇章语义增强的事件抽取方法	16
2.3.2 基于句法语义增强的事件抽取方法	17
2.4 基于数据增强的事件抽取方法	19
2.4.1 基于多任务学习的事件抽取方法	19
2.4.2 基于半监督的事件抽取方法	20
2.4.3 基于远程监督的事件抽取方法	21
2.5 其他方法	23
2.6 本章小结	25
第 3 章 基于多模态融合的事件检测方法	27
3.1 引言	27
3.2 相关工作	29
3.3 模型框架	30

目 录

3.4 多模态融合模型 (DRMM)	31
3.4.1 图像数据集构建.....	31
3.4.2 特征提取模块.....	32
3.4.3 多模态融合模块.....	33
3.4.4 事件检测模块.....	35
3.5 实验	35
3.5.1 实验设置.....	35
3.5.2 整体实验效果分析.....	37
3.5.3 低资源场景下模型性能分析.....	38
3.5.4 图像数据集质量分析.....	39
3.5.5 图像知识有效性分析.....	39
3.5.6 多模态融合方法有效性分析.....	40
3.5.7 案例分析.....	41
3.6 本章小结	42
第 4 章 基于外部知识注入的事件检测方法	43
4.1 引言	43
4.2 模型框架	45
4.3 知识蒸馏模型 (EKD)	46
4.3.1 事件检测模块.....	46
4.3.2 知识蒸馏模块.....	46
4.3.3 联合训练.....	49
4.4 实验	49
4.4.1 实验设置.....	49
4.4.2 整体实验效果分析.....	51
4.4.3 低资源场景下模型性能分析.....	53
4.4.4 跨领迁移能力分析.....	53
4.4.5 知识蒸馏框架的通用性分析.....	55
4.4.6 案例分析	56
4.4.7 讨论	57
4.5 本章小结	57
第 5 章 基于自标注的事件要素抽取方法	59
5.1 引言	59
5.2 模型框架	61

5.3 自标注模型 (MUCO)	62
5.3.1 基础原型网络.....	62
5.3.2 任务相关类检测模块.....	63
5.3.3 联合分类模块.....	66
5.4 实验	66
5.4.1 实验设置.....	67
5.4.2 评测指标.....	67
5.4.3 整体实验效果分析.....	68
5.4.4 任务相关类挖掘算法的性能分析.....	69
5.4.5 任务相关类的质量分析.....	70
5.4.6 任务相关类数量分析.....	71
5.4.7 领域迁移能力分析.....	72
5.5 任务通用性分析.....	72
5.6 本章小结	73
第 6 章 事件抽取在 NewsMiner 系统中的应用.....	74
6.1 引言	74
6.2 构建事件抽取数据集	77
6.2.1 事件本体构建.....	77
6.2.2 待标数据收集.....	79
6.2.3 众包标注.....	81
6.3 DocEE 数据统计分析.....	82
6.3.1 总体统计数据.....	82
6.3.2 事件类型统计分析.....	84
6.3.3 事件要素统计分析.....	84
6.4 实验	85
6.4.1 评价指标.....	85
6.4.2 基线系统.....	86
6.4.3 整体实验效果分析.....	86
6.5 应用成果展示	87
6.6 本章小结	88
第 7 章 结论和未来工作	89
7.1 工作总结	89
7.2 未来展望	90

目 录

参考文献.....	92
致 谢.....	102
声 明.....	103
个人简历、在学期间完成的相关学术成果.....	104
指导教师学术评语.....	106

插图和附表清单

图 1.1 事件抽取任务示例	1
图 1.2 DMBERT 在不 ACE 数据集上的表现.....	3
图 1.3 事件类型消歧难	4
图 1.4 事件召回难	4
图 1.5 本文的组织结构	6
图 2.1 DMCNN 模型示意图	16
图 2.2 DLRNN 模型的示意图	17
图 2.3 DEEB-RNN 模型的示意图	17
图 2.4 句法语义对事件抽取的好处	18
图 2.5 dbRNN 模型的示意图	18
图 2.6 EE-GCN 模型的示意图	19
图 2.7 DMBERT+ADV 模型的示意图	21
图 2.8 半监督模型示意图	21
图 2.9 远程监督模型 ANN-FN 示意图	22
图 2.10 ANN-FN 分类体系对齐示意图	22
图 2.11 BERT-QA 模型的示意图	23
图 2.12 BERT-Generation 模型的示意图 ^[78]	24
图 2.13 TBNNAM 模型的示意图 ^[78]	25
图 3.1 事件类型消歧难	27
图 3.2 多模态信息的作用	28
图 3.3 VAD 模型核心思想	30
图 3.4 多模态融合模型 DRMM 的框架示意图	31
图 3.5 交替对偶注意力机制（ADA）的图示	33
图 3.6 不同数据规模下模型性能对比	38
图 4.1 开放域事件触发词的示例	43
图 4.2 多模态融合模型 EKD 架构图	45
图 5.1 任务相关类的示例	60
图 5.2 原型网络中处理其他类的方法对比	60
图 5.3 自标注模型 MUCO 的架构示意图	61
图 5.4 基础原型网络原理示意图 ^[7]	62

图 5.5 不同任务相关类挖掘算法的对比	69
图 5.6 任务相关类数量超参对 MUCO 表现的影响.....	71
图 6.1 NewsMiner 系统中新闻数量统计	74
图 6.2 NewsMiner 热点事件挖掘	75
图 6.3 NewsMiner 现有语义分析功能示意图	76
图 6.4 事件要素初始集合来源示意图	79
图 6.5 DocEE 中 5 类事件的事件要素类型示意图	79
图 6.6 DocEE 中待标数据的两个来源.....	80
图 6.7 DocEE 众包标注系统.....	81
图 6.8 DocEE 中实例数 top10 的事件类型分布.....	84
图 6.9 NewsMiner 应用效果	87
表 1.1 现有知识图谱中事件知识规模	2
表 1.2 DMBERT 过拟合现象	4
表 2.1 事件抽取结果示例	9
表 2.2 主要事件抽取数据集统计	12
表 2.3 JEE 模型特征一览表	14
表 2.4 基于机器学习的事件抽取模型一览表	15
表 3.1 图像数据集统计信息	32
表 3.2 整体实验效果 (%)	37
表 3.3 语言模型有/无图片信息性能对比 (%)	39
表 3.4 图像知识有效性分析 (%)	40
表 3.5 不同多模态融合方法性能比较 (%)	40
表 3.6 案例分析	41
表 4.1 现有方法性能对比表	44
表 4.2 整体实验效果 (%)	52
表 4.3 有/没有开放域触发词知识性能对比 (%)	53
表 4.4 不同资源规模下召回率对比表 (%)	54
表 4.5 领域迁移能力 (%)	54
表 4.6 知识蒸馏框架的通用性分析 (%)	56
表 4.7 案例分析	57
表 5.1 整体实验效果 (%)	69
表 5.2 任务相关类质量定量分析	70

表 5.3 任务相关类质量定性分析	70
表 5.4 领域迁移能力 (%)	72
表 5.5 槽填充任务实验效果 (%)	72
表 5.6 命名实体识别任务实验效果 (%)	73
表 6.1 DocEE 事件分类本体	78
表 6.2 英文 DocEE 和现有的英文事件抽取数据集对比	83
表 6.3 中文 DocEE 和现有的中文事件抽取数据集对比	83
表 6.4 中文 DocEE 中事件要素提及分布	85
表 6.5 事件抽取中文实验设定	85
表 6.6 所提方法在中文事件检测上的整体表现	86
表 6.7 所提方法在中文事件要素抽取上的整体表现	87

第1章 绪论

1.1 研究背景和意义

事件是在特定时间和地点发生的事情，反映了客观事物状态和关系的变化^[1]，是人们认知世界的主要手段^[2]。法国伟大的马克思主义哲学家巴迪欧将“事件”作为巴迪欧认知哲学的核心思想，始终贯穿于他的哲学体系之中^[3]。英国数学家、哲学家和教育理论家怀特海认为“构成现实的终极单位并不是实体而是事件，事件具有第一性，整个宇宙就是由各种事件互连接、互相包涵而形成的有机系统”^[4]。随着计算机的广泛应用以及网络技术（5G）的发展，“网络媒体”逐渐替代电视、收音机和报纸等传统媒体，成为了绝大多数人获取事件信息的主要途径。根据国家信息中心最新发布的《中国网络媒体社会价值白皮书》，截止 2019 年 6 月，网络新闻用户规模达 6.86 亿，较 2018 年底增长 1114 万，占网民整体的 80.3%。虽然网络媒体能让人们更加便捷的获取事件信息，但同时也带来了信息爆炸的问题。面对网络媒体上海量异构、无序、杂乱的文本，如新闻、博客、聊天记录等，仅凭人力从中归纳和整理事件信息是不切实际的。研究如何从海量的文本中高效的获取事件信息是非常有必要的。

事件抽取在这个背景下应运而生。事件抽取旨在从非结构化的文本中抽取结构化的事件信息，是一门融合计算机科学、人工智能、语言学的科学，是自然语言处理领域一个重要的研究方向。如图1.1所示，事件抽取需要从事件提及句“社会民主党人肖尔茨于 12 月 08 日当选德国总理”中，识别事件触发词“当选”，判断事件类型“就职”，并抽取参与事件的相关要素“就职者”、“职位”和“时间”。

The diagram shows the process of extracting events from a sentence. At the top, the sentence is displayed: "社会民主党人肖尔茨于12月08日当选德国总理". An arrow points down to a table below, which is divided into two sections: "事件触发词" (Event Trigger Word) and "事件要素" (Event Elements). The "事件触发词" section contains the word "当选" in red, which is identified as triggering the "就职" (Appointment) event. The "事件要素" section is further divided into three rows: "就职者" (Person Appointed), "职位" (Position), and "时间" (Time). The corresponding extracted elements are "肖尔茨" (Scholz), "德国总理" (German Chancellor), and "12月08日" (December 8th).

社会民主党人肖尔茨于12月08日当选德国总理		
事件触发词		当选 (触发“就职”事件)
事件要素	就职者	肖尔茨
	职位	德国总理
	时间	12月08日

图 1.1 事件抽取任务示例

研究事件抽取不仅具有深刻的理论意义，还具有广泛的应用前景。从理论研究的角度看，事件抽取是评测自然语言处理模型认识世界、理解世界能力的重要手段之一。事件抽取可以提供比实体识别、关系提取更丰富的人物、组织、地点之间的联系，能够帮助自然语言处理模型更深入的了解世界的变化，推动自然语言

模型认知智能的发展。从应用的角度来说，事件抽取对国家、公司和个人都有重要意义。具体来说，事件抽取可以用于：

知识图谱构建。如表1.1中所示，现有的知识图谱仅包含以实体为核心的静态知识，而缺乏以事件为核心的动态知识。事件抽取可以从互联网中获取大量的新闻事件作为知识图谱额外的数据来源，这不仅显著的扩大了知识图谱的知识规模，还保证了知识图谱中知识的时效性。例如，事件抽取可以从俄乌战争的新闻中抽取战争开始时间、伤亡人数等事件知识，帮助知识图谱完成知识补全和动态更新。

表 1.1 现有知识图谱中事件知识规模

名称	创建时间	数据来源	数据规模
OpenCyc ^①	1984	专家知识	23万实体，未定义事件
WordNet ^②	1985	专家知识	15万实体，未定义事件
YAGO ^③	2007	WordNet+Wikipedia	459万实体，未定义事件
DBpedia ^④	2007	Wikipedia+ 专家知识	1694万实体，8万事件
Freebase ^⑤	2008	Wikipedia+ 领域知识 + 集体智慧	5872万实体，2万事件

事件因果关系分析。事件抽取是事件因果关系分析的基础，能够为事件因果关系分析提供更多的语义证据。例如，我们在分析“国会山事件”和“特朗普弹劾事件”这两个事件的因果关系时，由于前一个事件更集中在暴乱的报道上，后一个事件更集中在弹劾过程的报道上，单纯的通过文本相似度无法判断两者的因果关系。但是事件抽取可以发现“国会山事件”和“特朗普弹劾事件”中核心人物都有特朗普，从而为模型识别两个事件间的因果关联提供更多的语义证据。

支撑语义检索。事件抽取能够帮助搜索引擎给出更加精确的回答，提升用户体验。例如，通过对查询问题“汶川地震死了多少人？”的语义分析，事件抽取能够识别这是在查询一个地震类型的事件，并且在查询“伤亡人数”这个事件要素，这样就可以直接返回给用户“14866”这个精确的数字，而不会像传统搜索那样再给用户一大段文本，让用户再从中提炼信息。

风险监控。事件抽取可以帮助公司监控合作组织的不良行为。具体来说，事件抽取可以从财报和公开披露的经营文件中分析合作组织法人变动情况，销售贸易情况，合同纠纷情况等，从而对合作组织的经营状态进行实时的评估，帮助公司规避交易过程中可能发生的交易风险。另外，事件抽取可以通过分析国家政策的变更，国际局势的变化，为投资者预警可能产生的股价波动风险。

1.2 存在的挑战

随着自然语言处理技术的发展，深度学习模型成为目前阶段下事件抽取的主流方法，代表的模型有 Chen 等人在 2015 年提出的 DMCNN 模型^[5]和 Duan 等人在 2017 年提出的 DLRNN 模型^[6]。这些深度学习模型有着强大的语言表示能力，并且极大的提升了事件抽取的性能。但是美中不足的是，这些深度学习模型非常依赖大规模高质量的标注数据。然而，现实情况中，获取大规模标注数据并不容易，人工标注数据的过程费时费力，而且还需要大量的财力支撑。因此，自然语言处理领域学者开始转向低资源研究，学者们希望在数据标注稀缺的情景下，深度学习模型也能达到不俗的性能。其中，比较有代表性的人物有 Google Brain 的联合创始人吴恩达，纽约大学的教授、Facebook 的首席 AI 科学家 Yann LeCun 和 Facebook 元宇宙人工智能研究中心主任 Joelle Pineau。他们都在近两年频繁表达“深度学习模型应该用较少的数据得出准确的结论”的观点以及“希望人工智能系统更加灵活、更加健壮，不需要输入海量的原始数据”的观点。

目前，低资源研究在图像分类、跨语言问答、命名实体识别等任务上都取得了不错的进展^[7-9]，但是在事件抽取任务上研究空白还很大。现有的事件抽取方法在低资源场景下普遍性能表现不佳。如图1.2中所示，目前非常先进的深度学习事件抽取模型 DMBERT^[10]在仅有 10% 标注数据的低资源场景下，准确率只有 31.2%，远低于其在标注丰富的全数据量下的性能表现（准确率 71.7%）。

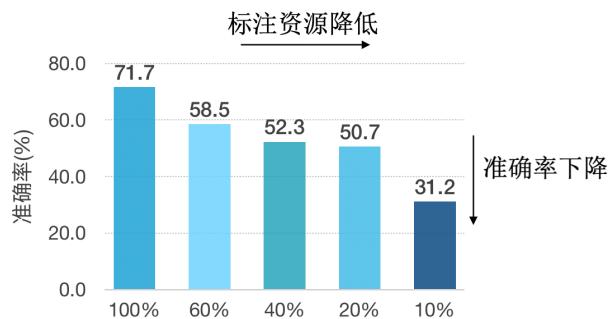


图 1.2 DMBERT 在不 ACE 数据集上的表现

分析原因，这是因为现有的事件抽取方法在低资源场景下容易过拟合训练数据。过拟合会使得现有方法过于关注训练集中出现的事件，而没有理解数据背后的规律，泛化能力差。表1.2中展示了这一点，可以看到，在仅有 10% 标注数据的低资源场景下，现有方法在“训练未见”的数据上表现（10.7%）就比“训练见过”（55.3%）差很多。

因为过拟合的发生，低资源事件抽取面临如下三个挑战：

- 1、事件类型消歧难。事件触发词的歧义性很大，经常面对同词触发不同事件类

表 1.2 DMBERT 过拟合现象

Metrics	训练见过	训练未见
Precision(%)	55.3	10.7

型的情况。现有的事件抽取方法在低资源场景下，容易过度拟合触发词在训练集中触发过的事件类型，而忽略了其他可能的情况。如下图1.3的例子中所示，由于“confront”在训练集中总是触发“开会”事件（S1），现有的方法由于过拟合，认为“confront”只可能触发“开会”事件，而无法识别“confront”触发其他类型的事件的情况（例如 S2 中触发了“攻击”事件），从而导致现有方法面临在低资源场景下事件类型识别不准的问题。

S1: Ford confront_{Meet} members in council chamber
S2: Police confront_{Attack} protesters hurling stones

图 1.3 事件类型消歧难

- 2、**事件召回难**。由于自然语言的灵活性，一个事件往往可以被多个事件触发词触发。现有的事件抽取算法容易过拟合训练数据，只能召回常见的事件触发词触发的事件，而无法召回不常见的事件触发词触发的事件。如下图1.4所示，由于“Attack”事件在训练集中总是被“fire”触发（S1），现有方法会误以为只有“fire”才能触发“Attack”事件，而无法识别“hacked”也可以触发“Attack”事件（S2）的情况，从而导致现有方法面临在低资源场景下事件召回率低的问题。

S1: Iraq terrorist fired_{Attack} towards our position
S2: A man was hacked_{Attack} to death by the criminal

图 1.4 事件召回难

- 3、**事件要素识别难**。自然语言非常灵活，可以用多种说法描述参与一个事件的事件要素的发生，但是现有的方法由于过度拟合训练数据，常常只能识别其中几个固定的说法，这就导致事件要素识别概率低。例如，由于训练集中经常通过“<Place> is the birthplace of <People>”这种语言模式描述“出生”事件的

出生地，当遇到“<People> is born in <Place>”说法的时候，模型由于没见过，就无法识别事件的“出生地”是哪里了。

1.3 本文的主要研究内容和贡献

为了应对低资源事件抽取中存在的这三个挑战，本文主要从以下三个方面展开研究工作。针对低资源场景下事件类型消歧难的问题，本文引入“多模态信息”，为事件类型消歧提供更多的图像模态上的语义证据，提高模型对事件类型识别的准确率。针对低资源场景下事件召回难的问题，本文引入“常识知识”，为事件识别提供先验^[11]，提升模型的泛化性，提高模型对事件的召回率。针对低资源场景下事件要素识别难的问题，本文从语料中“挖掘更多的训练数据”，通过扩大训练数据的规模，让训练数据多样化，解决低资源场景下训练语料单一、稀缺导致的事件要素识别过于生硬刻板的问题。综上所述，本文的主要研究内容和贡献如下，包括三个理论研究工作和一个应用示范验证：

- 1、提出基于多模态融合的事件检测方法。该方法通过深度融合新闻中天然存在的多模态信息的方式，让模型在判断事件类型时不仅可以依赖文本语义信息，还可以依赖图像语义信息，从而提升模型在低资源场景下事件消歧的效果。相比于传统的单向融合或者拼接融合的方式，本文所提的多模态融合方法能够按照重要性对图像模态的信息进行双向筛选，显著的提升了融入图像信息的质量。
- 2、提出基于外部知识注入的事件检测方法。该方法首先利用 WordNet 词汇库高效的获取外部知识，之后采取学生模型拟合教师模型的训练方式，将外部知识融入到深度学习模型的参数中。所提方法不仅能够在标注语料上融合知识，而且能够在大规模未标注语料上融合知识，有很强的知识利用能力。实验从定性和定量两个方面验证了所提知识蒸馏方法能够在低资源场景下高效的召回事件。
- 3、提出基于自标注的事件要素抽取方法。该方法借助预定义类的弱监督信号，可以在没有元数据的情况下从语料中分辨噪音和任务相关类，通过挖掘更多的任务相关类的标注，提升了事件要素训练数据的多样性，有效的降低了事件要素训练语料单一稀疏带来的过拟合的风险。在低资源场景上的实验结果表明，我们的模型优于五个最先进的模型，证明了自标注思路的有效性。
- 4、在新闻挖掘系统 NewsMiner 上展开示范应用。针对现有事件本体和应用需求不匹配的问题，我们基于新闻学理论重新构建了本体，并众包标注了中英双语事件抽取数据集 DocEE。利用该数据集，我们将上述提出的三个低资源

事件抽取方法应用在 NewsMiner 系统上，展现本研究的实际应用价值。

1.4 论文组织结构

在本章中，我们介绍了低资源事件抽取的研究背景和意义，并针对其中存在的挑战介绍了本文的主要研究思路和研究贡献。论文整体的组织结构如图1.5所示。本文的后续章节安排如下：

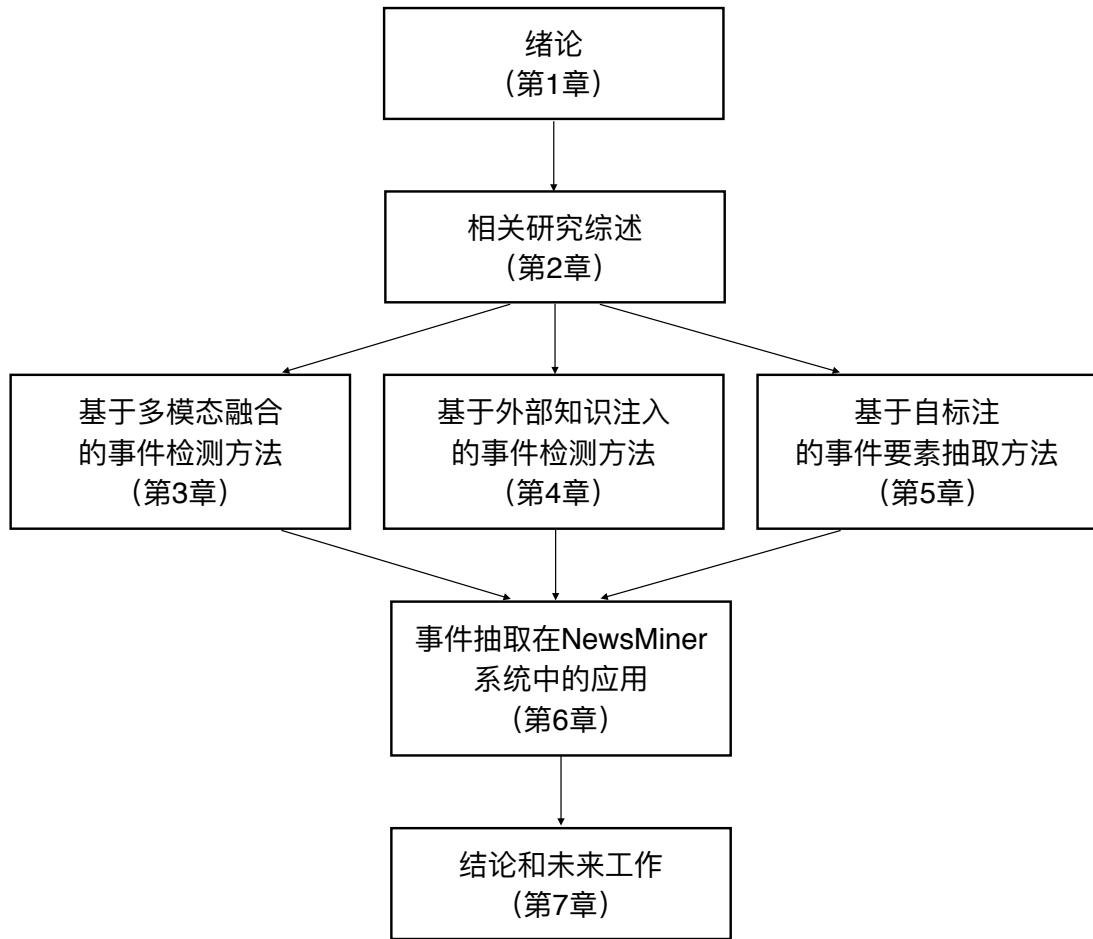


图 1.5 本文的组织结构

第 2 章是综述章节，首先先回顾了事件抽取方法发展的主要阶段，之后从语义增强和数据增强两个角度详细介绍现有的深度学习事件抽取方法。最后，引出本文解决问题的主要思路。

第 3 章介绍基于多模态融合的事件检测方法的学习框架，给出深度融合多张新闻图片的具体计算过程，并在公开数据集 ACE2005 上对所提方法进行系统性的评估。

第 4 章介绍基于知识蒸馏的事件检测方法的学习框架，展示该方法知识获取、

知识编码和知识融合过程，并从定性及定量两方面实验分析所提方法在低资源场景下的性能。

第 5 章介绍基于自标注的事件要素抽取方法的学习框架，展示该方法借助预定义类从语料中挖掘任务相关类标注的过程，并在设计的小样本低资源场景下展开系统性的实验。

第 6 章是示范应用章节，介绍了本文提出的三种事件抽取方法在新闻挖掘 NewsMiner 系统中的应用情况，包括一个线上事件抽取功能的展示。

第 7 章总结了本文的主要贡献，并指出未来的研究方向。

第2章 相关研究综述

本章对事件抽取相关研究进行梳理。首先明确事件抽取任务定义，介绍事件抽取相关语料库。之后，我们在简要回顾传统的事件抽取方法之后，重点关注深度学习事件抽取相关工作，并从语义增强和数据增强两个角度对深度学习事件抽取相关工作进行详细的介绍，包括基于篇章语义增强、基于句法语义增强，基于多任务学习、基于半监督学习和基于远程监督学习。最后，指出现有方法存在的问题，引出本研究的主要思路。

2.1 任务定义

事件抽取在不同人工智能的学科背景和领域场景下，定义大相径庭。为了梳理清楚本文研究的事件抽取的含义，我们首先概述人工智能各个领域对事件抽取的理解，最后阐述本文中对事件抽取的定义。

在计算机视觉研究中，事件抽取是指从图像或者视频中，感知单个物体或者多个物体的行为轨迹，并对行为轨迹进行分析的任务（Activity Recognition^[12]）。比如，从篮球/足球等体育视频中检测进球、犯规等运动行为，从老人病患者等监控视频中预警危险行为，从交通监管视屏中检测公共安全场景下暴力事件等。

在语音识别研究中，事件抽取是指从音频信号中检测声音事件的任务（Sound Event Detection^[13]）。比如，从会议的音频中检测开关门声音，从录音带中检测连绵的下雨声、汽车的鸣笛声、人们的说话声和走路的脚步声。

在常识推理研究中，事件抽取是指基于人的前置行为，推理工人后续可能发生行为的任务（Script Event Prediction^[14]）。例如，依次发生了“顾客进店”，“顾客点菜”、“顾客吃饭”的行为，那么算法应该能推理出接下来大概率会发生“顾客付账”的行为，而不是“顾客直接离店”的行为。

2.1.1 本文定义

本文研究的事件抽取是指根据预定义的事件本体，从非结构化的文本中识别感兴趣的事件的任务。具体来说，本文遵从自动文本抽取竞赛 (Automatic Content Extraction, ACE) 和消息阅读理解评测 (Message Understanding Conference, MUC) 对事件的定义，认为“事件描述特定时间地点发生的事情，是一个有结构的元素组合，包含事件触发词以及多个事件要素”，相关定义如下：

- 「事件触发词」(Event Trigger)：标志事件的发生的短语或短句，通常出现在

动词或者动名词上。

- 「事件类型」(Event Type)：描述事件触发词触发的事件类型的标签。
- 「事件要素」(Event Argument) 指在事件中扮演关键角色的要素集合，如时间、地点、涉事人物等都属于事件要素。
- 「事件要素角色」(Event Argument Role)：描述事件要素类型的标签。

表 2.1 事件抽取结果示例

事件类型	事件触发词	事件要素角色	事件要素
		被捕者	阿坦巴耶夫
逮捕	拘留	执行逮捕的实体	吉尔吉斯斯坦警方
		日期	星期四
		事件发生地	首都比什凯克附近

根据 D.Ahn 在 2006 年的研究工作^[15]，事件抽取可分解为四个子任务：事件触发词检测、事件触发词分类、事件要素检测和事件要素分类，其中前两个任务合并统称为事件检测 (Event Detection)，后两个任务并称为事件要素抽取 (Event Argument Extraction)。

如表2.1中所示，给出事件提及“周四，吉尔吉斯斯坦警方在首都比什凯克附近拘留了前总统阿坦巴耶夫”，事件检测任务需要从中检测事件触发词“拘留”，并判断其触发的事件类型是“逮捕”，事件要素抽取任务需要从中抽取参与该“逮捕”事件的事件要素，包括时间“周四”，地点“首都比什凯克附近”，被捕者“阿坦巴耶夫”和执行逮捕的实体“吉尔吉斯斯坦警方”。

形式化的，事件检测任务的定义如下：给定文本 $S = \langle w_1, w_2, \dots, w_n \rangle$ ，事件检测需要最大化概率 $\frac{1}{n} \sum_{i=1}^n P(y_i | w_i)$ 其中 $y_i \in Y$ ， $Y = \{y_1, y_2, \dots, y_c\}$ 是预定义事件类型。

事件抽取任务的定义如下：给定文本 $S = \langle w_1, w_2, \dots, w_n \rangle$ ，事件触发词 w_i ，事件类型 y_i ，事件抽取任务需要最大化概率 $\frac{1}{n} \sum_{j=1}^n P(a_j | w_j, w_i)$ ，其中 $a_i \in A$ ， $A = \{a_1, a_2, \dots, a_z\}$ 是预定义的事件要素类型，

可以看出，事件检测和事件抽取都是分类任务 (token-level classification)，因此，在后续相关研究工作综述的过程中不再区分。

2.1.2 相关数据集

本节主要介绍事件抽取相关数据集。我们首先按照时间顺序介绍关键事件数据集的发布情况，之后再对各个数据进行详细的介绍。

事件抽取语料库的构建最早可追溯到 1980 年代后期。为了保障国家安全，美国国防高级研究计划局 (DARPA) 在 1987 到 1997 年间先后发布了 7 个事件抽取语料库 MUC-1 MUC-7，用于提升模型在挖掘军事情报，追踪恐怖袭击和暴力武装冲突事件的能力。紧接着，在 1997 年，DARPA 和卡内基梅隆大学以及马萨诸塞大学阿默斯特分校合作，共同创立了“主题检测和跟踪”公共评估项目，并在项目中发布评测数据集 TDT，以促进模型在广播新闻文章流中发现和追踪最新热点事件的能力。后来，美国国家标准与技术研究院 (NIST) 开始关注非结构化文本语义分析，并于 1999 年到 2008 年间多次举办自动内容提取 (ACE) 竞赛，用于推进信息提取技术的发展。其中 ACE2005 语料库的影响力最大，它规范了事件抽取任务的研究框架，将事件抽取细分为了事件触发词检测和事件要素抽取，引领了近二十年主流事件抽取的研究。鉴于 ACE2005 的成功，DARPA 在 2015 年在文本深度探索和过滤 (DEFT) 计划中提出了更简化的事件抽取标准 light ERE，后来又对 light ERE 进行细化和补充，拓展形成了 rich ERE 标准。同一年间，事件抽取也获得了国际知识图谱构建大赛 (KBP) 的关注，KBP 分别在 2015, 2016, 2017 发布了中英双语的事件抽取竞赛数据集，用于评估模型从大型文本语料库中提取信息，补充现有知识图谱的能力。

2.1.2.1 通用域事件抽取数据集

了解了关键数据集发布脉络，我们再具体对通用域的各个数据集进行介绍。通用域事件抽取数据集主要研究新闻事件抽取。我们按照事件抽取发生在句子范围内，还是在篇章范围内，将现有的通用域事件抽取数据集分为句子级事件抽取数据集和篇章级事件抽取数据集分别进行介绍。

句子级事件抽取数据集只从当前句子抽取事件触发词以及事件要素。常用的句子级数据集有：ACE2005, Rich ERE, TAC-KPB2017, CEC, DuEE, MAVEN 和 TimeBANK。

- ACE2005 数据集是一个多语言事件语料库，共标注了 599 篇英文文章，633 篇中文文章以及 403 篇阿拉伯语文章，包含 8 种事件类型和 33 个子事件类型，涵盖了“死亡”、“攻击”等常见事件。
- Rich ERE 语料库共定义了 8 个事件类,18 个事件子类，在事件抽取以及事件共指消歧的标注细节方面，对 ACE2005 进行了细化。例如，RichRER 认为没有任何事件要素支撑的事件触发词也可以被标注出来，这在 ACE 中是不被允许的。
- TAC-KPB2017 是一个包含中文，英文，西班牙语的事件抽取语料库，目前可以从语言数据联盟 (LDC) 下载，包括 158 个训练集文档和 202 测试集文档，

共包括 8 种事件类型和 18 个事件子类型。

- CEC 语料库由上海大学发布，共收录了 332 篇中文新闻报道。它设计了五种事件类型：地震、火灾、交通事故、恐怖袭击和食物中毒，基本囊括常见的公共突发事件。
- DuEE 是目前规模最大的中文句子级事件抽取语料库。DuEE 发布于 2020 年语言与智能技术大赛，发布者是百度公司，标注数据来自百家号新闻，涵盖大量百度搜索的热门话题。DuEE 共定义了 65 个事件类型，以及 121 个事件要素类型，共包含 19,640 个事件标签，41,520 个事件要素标签。
- MAVEN 语料库数据来源自英文维基。不像上述五个事件抽取数据集同时具有事件检测任务的标签和事件要素抽取任务的标签。MAVEN 语料库只有事件检测任务的标签，该数据集共标注了 168 个事件类型和 118,732 个事件触发词。
- TimeBANK 语料库只有事件要素抽取的标签，并且只标注“时间”这一个事件要素。TimeBANK 整个数据集共包含 183 篇文章，有 27592 个时间标签。Factbank 语料库在 TimeBank1.2 之上补充了有关事件真实性的附加信息。

篇章级事件抽取数据集需要模型从整篇文章中抽取事件，在这个设定下，属于同一个事件的多个事件要素可能散落在多个句子中，需要模型做跨句的事件要素抽取，对模型的长文本处理能力提出了更高的要求。常用的篇章级数据集有：MUC，M2E2，WikiEvents，RAMS 和 ASTRE。

- MUC 系列数据集（MUC-1 MUC-7）由美国国防部发布，要求模型提取事件要素填充到预定义的要素槽值结构中。其中，MUC-4 定义了 5 个事件类型以及所有事件类型共享的 4 个事件要素类型，涉及 1,700 篇英文新闻。
- M2E2 语料库由伊利诺伊大学厄巴纳-香槟分校发布，旨在从多媒体文档中提取事件及其要素，共定义了 8 个事件类和 29 个事件要素类型。通过众包标注，该数据集共收集了 245 篇多媒体新闻文章。
- WikiEvents 语料库是同样由伊利诺伊大学厄巴纳-香槟分校发布，标注了 246 篇维基英文文档，定义了 49 个事件类型和 57 个事件要素类型。
- RAMS 语料库标注了 9,124 篇文章，本体定义了 139 个事件类型和 65 个事件要素类型。RAMS 只在离触发词不超过五句的范围内标注事件要素。

2.1.2.2 特定域事件抽取数据集

除了通用域事件抽取数据集，事件抽取数据集还主要集中生物领域，代表性的数据集有：

- GENIA 语料库关注生物领域的蛋白质物理状态改变事件以及属性变化事件，

利用了 MEDLINE 数据库中的摘要数据作为标注数据来源。在数据规模上, GENIAv3.0 总计标注了 2000 篇摘要, 9,372 个句子和 36,114 个事件实例, 是目前比较大的生物领域事件数据集。

- PASBio^[16]语料库重点关注分子生物学 (molecular biology) 中的基因表达事件、分子相互作用事件和信号转导事件, 标注的数据来自于 MEDLINE 数据库中的摘要和 EMBO, PNAS, NAR 和 JV 学术期刊的全文文章。相比于 GENIA 语料库, PASBio 语料库标注数据的来源更加多样化。
- BioNLP^[17]语料库共标注了 1210 篇摘要, 11346 个句子, 13000 多个事件, 任务设定上, BioNLP 除了常规的标注, 还关心次相关的事件要素的识别, 让事件抽取能够捕获更多的事件信息。

这些生物领域事件抽取数据集除了在本体设定上和通用领域事件抽取数据集有较大的差别, 还有一个明显的特点是考虑嵌套事件抽取 (Nest Event Extraction)。由于生物事件的复杂性, 经常会出现一个事件嵌套另一个事件的场景, 为了适应这种情况, 生物领域事件抽取数据集允许将一个事件当做另一个事件的要素这样的标注方式。

除此之外, 事件抽取在金融领域和医疗领域也有不少的应用。例如, Doc2EDAG^[18]语料库关注金融领域财报事件抽取, 总共从 2008-2018 年的 ChFi-nAnn4 财报文档中标注了 32,040 个金融事件, 这些金融事件分布在股权冻结、股权回购、股权减持、股权增持和股权质押上, 基本涵盖了金融领域财报分析时核心关注信息。CCKS^[19]语料库关注肿瘤医疗事件抽取, 要求模型从肿瘤病历文本数据中抽取肿瘤医疗事件的 3 个重要事件要素: 肿瘤大小, 肿瘤原发部位和肿瘤转移部位, 旨在为医生诊断提供便利。

表 2.2 主要事件抽取数据集统计

名称	时间	组织者	事件类型	数据规模	语言
MUC-4	1987-1997	DARPA	1	1700 篇	英文
ACE2005	2005	NIST	33	599 篇	中文, 英文, 阿拉伯语
CEC	2009	上海大学	5	332 篇	中文
TAC-KPB2017	2017	NIST	18	360 篇	中文, 英文, 西班牙语
M2E2	2020	UIUC	8	245 篇	英文

2.2 事件抽取方法研究框架

本节介绍事件抽取方法研究框架的发展脉络。事件抽取方法主要经过三个发展阶段：模式识别，机器学习和深度学习。接下来，我们依次介绍各个阶段的核心思想和代表性的工作。

在模式匹配阶段，事件抽取需要专家设计精细的事件模式（pattern）来指导抽取过程。事件模式通常是从语法树或正则表达式中总结得来的，也可以表现为词典集合。基于模式匹配方法的典型代表工作是 Ellen 在 1993 年开发的 AutoSlog 系统^[20]。AutoSlog 系统定义了 13 种句法树规则，构建了大规模领域事件模式库，用于事件触发词检测和事件要素的抽取。例如，AutoSlog 设计了“died in <np>”，“death of <np>”等事件模式来抽取死亡这个事件的地点、原因等事件要素。为了增强领域短语字典的泛化性，AutoSlog 还考虑了词之间的概念关系，包括上下义关系和总分关系。

除此之外，其他基于模式识别事件抽取模型有：Kilicoglu 等人^[21]使用句法依存规则来执行生物事件提取；Buyko 等人^[22]结合手动归纳的字典来提取事件触发词和缩小有效事件要素出现位置的范围；Yangarber 等人^[23]提出了一种自举事件模式发现方法，从初始种子模板开始，迭代的挖掘更多的事件模式。Borsje 等人^[24]基于词汇语义（如词性标签（POS）、实体信息和形态特征（Lemma））设计丰富的事件抽取规则用于 RSS 金融事件的提取。

基于模式的方法的优点总结如下。首先，这种方法不需要人工标注语料，节省了标注的人力物力。其次，由于其模式是手动设计和维护的，因此具有更好的可解释性。第三，如果模式设计得当，它可以在特定领域取得很高的提取精度。基于模式的方法的缺点总结如下。首先，开发和维护精细的事件模式是相当耗时和费力的。其次，由于模式设计强烈依赖于文本的表达形式，因此将模式从一个领域转移到另一个领域需要付出很大的努力，事件模式的低重用性限制了其泛化性。

为了避免设计精细事件模式，提升事件抽取的泛化性，许多研究人员探索了基于机器学习方法来抽取事件。这类方法将事件抽取划分为两个阶段：特征提取和机器学习。特征提取目的是将离散特征变为可计算向量，机器学习的目的是学习一个好的映射函数以区分不同类型的事件触发词和事件要素。

这类方法的典型代表工作是 Li 在 2013 年提出的 JEE^[25]模型。该模型设计了丰富的文本特征，包括：词法特征，句法特征，实体信息特征和词对特征。表2.3展示了该方法中用到的各个类型的特征信息。可以看到，除了常用的特征，Li 等人还针对事件抽取任务的特点，添加了触发词和要素之间的词对特征，强化了事件触发词和事件要素之间的语义联系。除此之外，Björne 等人^[26]还利用 Charniak-Johnson

句法解析器获得额外的句法特征，辅助支持向量机（SVM）从科研文献中提取生物医学事件。特征除了可以作为分类模型的输入，还可以作为分类结果的筛选器。例如，Huang等人^[27]利用事件触发词和事件要素出现的位置的特征，对条件随机场机器学习方法给出的答案进行了二次过滤，显著的提高分类结果的质量。表2.4中对比了历年来基于机器学习的事件抽取算法。

表 2.3 JEE 模型特征一览表

特征类型	特征描述
词法特征	词干标签（token）
	词性标签（POS）
	同义词标签（synonym）
句法特征	词义聚类的分类标签
	词依存关系标签
实体信息特征	句法分析树中路径位置
	是否是未被引用的代词
词对特征	实体类型
	一阶邻居的实体类型
词对特征	触发词-触发词之间的共现关系
	触发词-要素的距离
词对特征	触发词和要素间的句法依存关系

相比于模式匹配，基于机器学习的事件抽取的优点是减轻了设计精细模式的工作量，具有更好的泛化性和可重用性。缺点在于容易发生误差传递，如果特征设计的不仔细，就会严重影响分类模型的提取精度。

为了缓解误差传递，深度学习的方法引发了广泛的关注。深度学习方法是一个端到端的模型，特征提取过程完全自动化，无须专家再设计精细的事件模式以及设计复杂的特征，彻底解放了人类双手。而多层神经元的连接使得深度神经网络能够拟合更复杂的非线性关系，有更强大的分类能力。这些特性促使了深度学习方法成为目前事件抽取的主流方法。

深度学习方法将事件抽取形式化为一个分类的任务（token-level classification），通过建立单词的嵌入表示到分类空间的映射，达到事件抽取的目的，图2.1中展示了深度学习方法将“cameraman”分类到“fire”事件的“target”事件要素上的典型过程。可以看到，深度学习方法的学习过程一般可以划分为三个阶段：嵌入表示、特征提取和事件分类。嵌入表示用于获取词的向量表示，特征提取用于提取句法和词

表2.4 基于机器学习的事件抽取模型一览表

模型	分类技术	实验数据集	应用领域
Henn et al. ^[28]	SVM	ToMK	general IE
Lu and Roth ^[29]	ME, Markov CRF	ACE2005	general IE
Chen and Ng ^[30]	SVM	ACE2005	general IE
Li et al. ^[31]	ILP, ME, CRF	ACE2005	general IE
Björne and Salakoski ^[26]	SVM	BioNLP'11	biomedical
Björne et al. ^[32]	SVM,CRF	PubMed	biomedical
Miwa et al. ^[33]	SVM	BioNLP'09	biomedical
Satre et al. ^[34]	AkaneRE	BC-IPT, BioNLP'09	biomedical
Li et al. ^[35]	CRF, AdaBoost, SVM	I2B2 ^[36]	biomedical
Liao and Grishman ^[37]	ME	ACE2005	general IE
Ji and Grishman ^[38]	ME	ACE2005, TDT	general IE

法特征，事件分类用于事件触发词和事件要素分类。

深度学习事件抽取方法的一个典型代表是 Chen 等人在 2015 年提出的 DM-CNN (Dynamic Multi-Pooling Convolutional Neural Networks) 模型。该模型^[5]在特征提取的过程中加入了动态池化的操作，让卷积神经网络能更有针对性的捕获特征。如图2.1所示，动态池化会根据当前关心的事件触发词和事件要素的位置（这里是 fire 和 cameraman），动态将句子分割为前中后三段，并用卷积网络分别捕获每一段中的词汇级和句子级特征，来提升卷积神经网络的差异化表示能力，便于模型处理同一句子中有多个事件的场景。形式上来说，给定 $S = w_1, \dots, w_i, \dots, w_j, \dots, w_n$ ，其中 w_i 是事件触发词， w_j 是要判断的事件要素， $w_{i:j}$ 表示 w_i 到 w_j 词汇的拼接，动态池化层表示计算方式为：

$$\begin{aligned} h &= \max(c_{1:i}, c_{i:j}, c_{j:n}) \\ c_{i:j} &= f(W \cdot w_{i:j} + b) \end{aligned} \tag{2.1}$$

由于深度学习事件抽取方法已经成为事件抽取任务的主流方法，因此，本章的后续章节从语义增强和数据增强的角度，进一步的展开介绍现有的深度学习事件抽取方法相关工作，而不再对传统的事件抽取模型展开陈述。

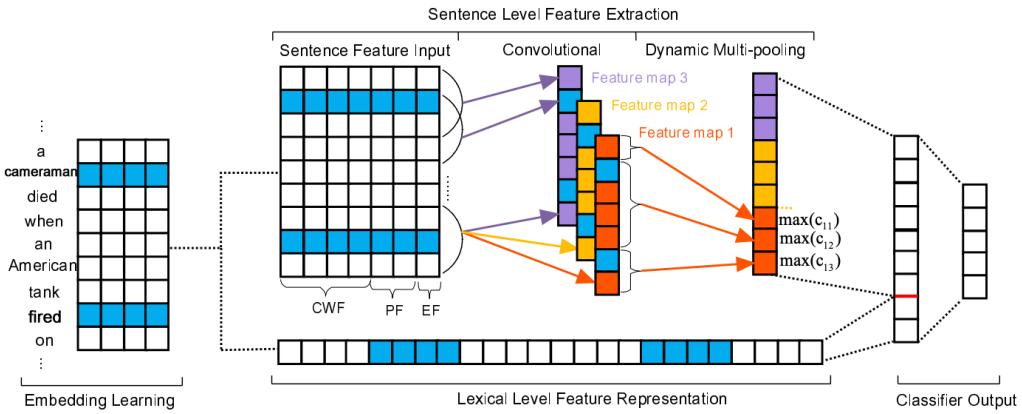


图 2.1 DMCNN 模型示意图

2.3 基于语义增强的事件抽取方法

基于语义增强的事件抽取方法的核心思路是在深度学习模型的特征提取的过程中，融合更多的语义信息，这主要包括篇章语义信息和句法语义信息。

2.3.1 基于篇章语义增强的事件抽取方法

常见的深度学习事件抽取方法仅利用了句子级别的文本信息辅助事件抽取（比如上述提到的 DMCNN 模型），但是忽略篇章中蕴含的丰富语义信息。基于篇章语义增强的事件抽取方法在句子级别的局部特征（local feature）之外，加入了篇章特征（document feature），为事件抽取提供更多的语义证据。

这类方法的典型代表是 Duan 等人在 2017 年提出的 DLRNN^[6]模型（图2.2）。该模型首先利用词表示学习 (PV-DM)^[39]从大规模无标注文本中学习篇章表示，基于的假设是一个好的篇章表示应该能够帮助模型有更好的根据上下文猜测当前词的能力，即最大化如下概率：

$$\frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \dots, w_{t+k}, doc) \quad (2.2)$$

之后，该模型将学习到的篇章表示通过拼接的方式融入到句子特征中，完成篇章信息的注入。

除此之外，现有的工作还会利用注意力机制和记忆力网络来增强模型对篇章语义的利用能力。例如，Zhao 等人^[40]提出了一个基于分层级注意力机制的循环神经网络模型。如图2.3所示，该方法通过先融合句子级信息再融合篇章级信息的方法，赋予了模型更强的篇章信息提取能力。Chen 等人提出了一个基于门控注意力机制的循环卷积神经网络模型^[41]，在融合篇章信息的时候，考虑当前句和篇章之间的关联程度，让模型能够更有重点的融合篇章信息。Liu 等人^[42]提出了一个基

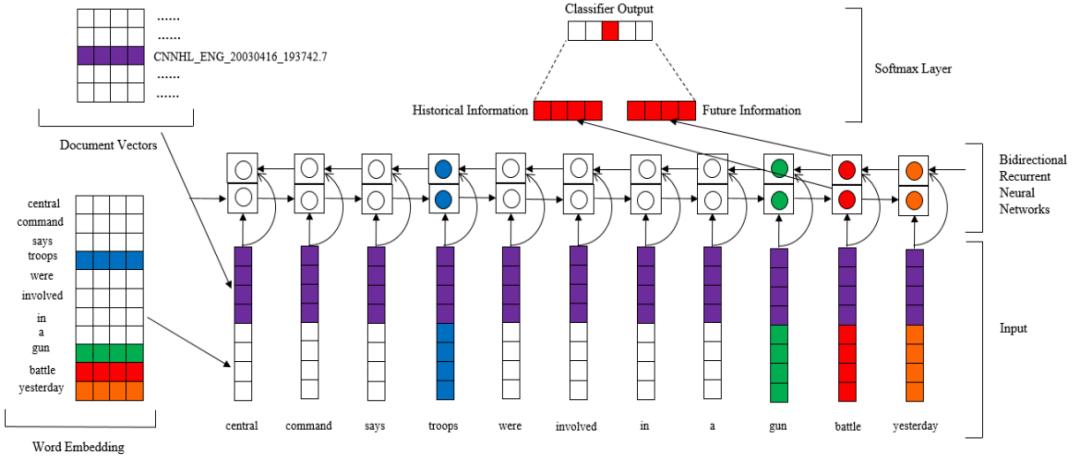


图 2.2 DLRNN 模型的示意图

于动态记忆力机制的循环卷积神经网络模型，通过多次阅读篇章信息，提升模型对篇章信息的利用率。

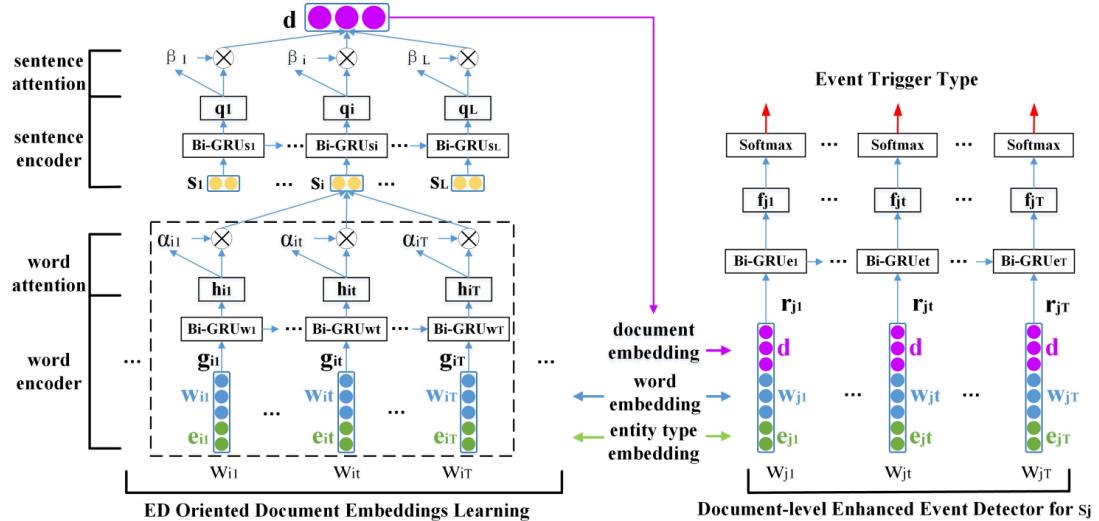


图 2.3 DEEB-RNN 模型的示意图

但是这些方法都受限于预训练模型对输入文本长度的限制（512个词），无法利用更远的篇章语义信息，因此 Veyseh^[43]等人额外的增加了一个句子筛选模块，筛选掉了篇章中不太重要的句子，帮助模型看到更远的有用信息。

基于篇章语义增强的方法的优点是综合了更多的信息，能够利用更多的上下文信息辅助消歧，缺点是没有利用更高层级的语义信息。

2.3.2 基于句法语义增强的事件抽取方法

常见的深度学习事件抽取算法多是按照词语顺序编码上下文信息的，这种扁平化的编码方式导致距离远的关键信息很容易被忽略掉。如图2.4所示，句法依存

关系能够将文本连接成图，有效的拉近关键信息的距离，方便模型捕获长距离语义依赖。基于句法语义增强的事件抽取方法就是在这个思想下应运而生的。

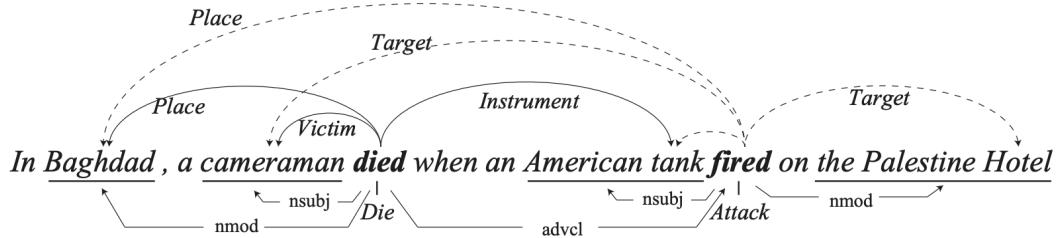


图 2.4 句法语义对事件抽取的好处

早期的事件抽取模型是通过在循环神经网络上添加边的方式来融入句法语义知识的^[44-45]。这类方法的代表工作是 Sha 等人提出的 dbRNN^[44]模型。如图2.5所示，dbRNN 除了 LSTM 单元常规的前向边和后向边之外，还加入了句法分析树中存在的语义边，比如“adv”，“nmod”，“det”等等，将识别事件触发词和事件要素时需要考虑的关键信息直接桥接回来，拉近了有用信息的距离，提升了模型的泛化性。

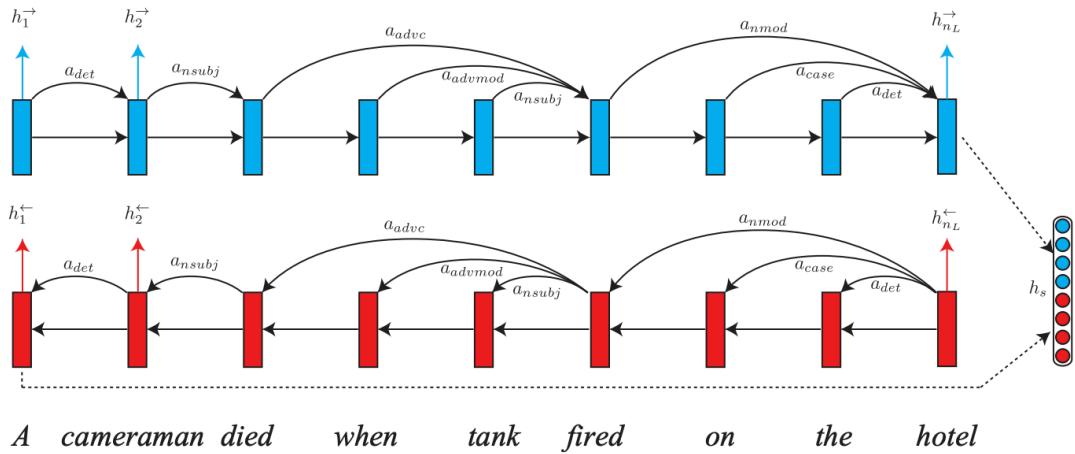


图 2.5 dbRNN 模型的示意图

后来，随着图表示技术的发展，图卷积神经网络成为融入句法语义知识的主流方法^[46-52]。这类方法的典型代表是 Cui 等人在 2021 年提出的 EE-GCN^[49]模型。图2.6展示了该模型的架构。EE-GCN 首先利用句法分析工具将扁平的句子转换为有向无环图，之后，利用卷积神经网络更新图上的节点表示，让图上的节点能够融合全局信息。该过程主要涉及两个核心感知模块：节点感知模块 N 和边感知模块。“节点感知模块”通过聚合节点所有邻接边的信息来更新节点的表示，“边感知模块”通过聚合边所有邻接节点信息来细化边的表示。两个模块更新的计算公式分别

为：

$$\begin{aligned} N_i^L &= \sigma(Pool(N_1^L, N_2^L, \dots, N_p^L)) \\ E_{ij}^L &= W_u[E_{i,j}^{L-1} \oplus N_i^L \oplus N_j^L] \end{aligned} \quad (2.3)$$

最后，EE-GCN 利用全连接网络对图上的节点进行分类，完成事件触发词检测和事件要素抽取。

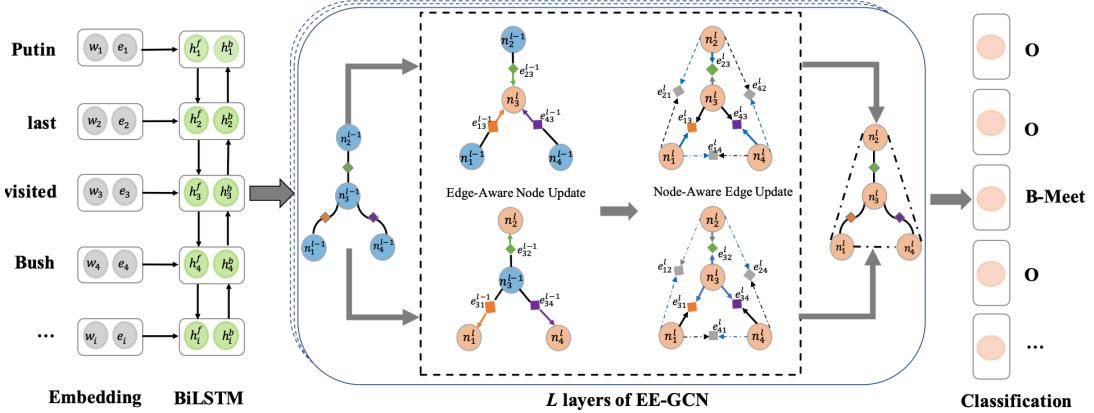


图 2.6 EE-GCN 模型的示意图

相比于数据驱动的方法，基于句法语义的方法的优点是引入了句法信息，缓解了深度学习模型长距离灾难性遗忘的问题。缺点是需要依赖自然语言处理工具的句法分析结果，会产生误差传播的问题。

2.4 基于数据增强的事件抽取方法

基于数据增强的事件抽取方法重点关注如何设计一个合理的网络机制，在利用已有的标注数据以外，同时利用更多的标注数据辅助深度学习模型训练，包括多任务学习，半监督学习和远程监督学习。

2.4.1 基于多任务学习的事件抽取方法

在标注数据比较少的情况下，如果仅在事件抽取单个任务上训练，深度学习模型往往无法取得很好的性能。多任务学习的思路是在训练过程中引入其他任务的语料，通过联合训练的方式，提升深度模型的性能。

基于多任务学习的事件抽取方法通常会引入词语消歧任务（word sense disambiguation）联合学习^[53]，因为词语消歧任务和事件抽取任务都需要处理一词多义的情况，是和事件抽取任务紧密相关的。这类方法的典型代表是 Lu 等人在 2018 年提出的 MATCHING^[54] 模型。MATCHING 有两个关键模块：编码层和分类层。编码层分别独立的编码事件抽取任务和词语消歧任务的数据，分类层通过参数软共

享机制（soft sharing）拉近两个任务编码层表示的距离，借助词语消歧任务的丰富语义提升事件抽取模型的识别能力和泛化能力。参数软共享机制的损失函数的形式化表示如下：

$$L = -P(y|H) + \lambda \frac{1}{d_H} \sum_{i=0}^{d_H} (H_i^{wsd} - H_i^{ed})^2 \quad (2.4)$$

其中， H 是编码层输出的向量表示， d 是 H 的维度， H_i^{wsd} 和 H_i^{ed} 分别是词语消歧任务的向量表示和事件检测任务的向量表示。

此外，多任务事件抽取方法经常会引入命名实体识别（Named Entity Recognition）任务联合学习^[10,55-57]。命名实体识别任务和事件抽取任务是紧密相关的，因为实体类型可以帮助事件消歧，并且事件要素大都是实体。这类方法的代表工作是 Yang 等人在 2019 年提出的 JointEventEntity 模型^[58]，该方法利用概率图模型对事件本身、事件-事件关系、事件-实体关系分别构建因子图，并采取联合推理的方式，综合考虑他们之间的关系，具体计算方式如下：

$$\max_{t,r,a} \sum_{i \in T} E(t_i, r_i, a) + \sum_{i, \hat{i} \in T} R(t_i, t_{\hat{i}}) + \sum_{j \in N} D(a_j) \quad (2.5)$$

其中 T 是事件触发词的集合， N 是实体的集合， E 是事件本身的因子图，事件-事件关系的因子图，事件-实体关系的因子图。

除了引入单个任务相关类，多任务事件抽取模型有时也会同时引入多个任务相关类，比如 Tong 等人^[10]就同时引入了命名实体识别任务和词表示任务增强中文事件抽取的性能。

2.4.2 基于半监督的事件抽取方法

常见的深度学习模型在标注数据少的情况下表现不佳，基于半监督的事件抽取方法^[59-61]的思路是从大规模未标注语料中（通常是新闻中）挖掘更多的训练数据，提升事件抽取的效果。

一种思路是依赖人工设计的规则来挖掘训练数据。例如，Ferguson 等人在 2018 年提出的 ParaphraseClusters 模型^[62]。该模型利用到了 NewsSpike 思想扩充数据^[63]，基于假设是如果两个事件实例（两句话）所在的文章发布的间隔时间很短，并且这两个事件实例同时包含罕见的命名实体，比如“Les Miles”，那么这两个事件实例很可能在讲同一个事件，相似度计算公式如下：

$$S(a_i, a_j) = \sum_{e \in E_{a_i} E_{a_j}} \frac{\text{count}(e, date_{a_i, a_j})}{\text{count}(e, corpus)} \quad (2.6)$$

其中， e 是指实体， a_i 是指事件实例。通过计算相似度，将其中置信度高的事件实

例加入标注数据，达到扩大训练数据的目的。

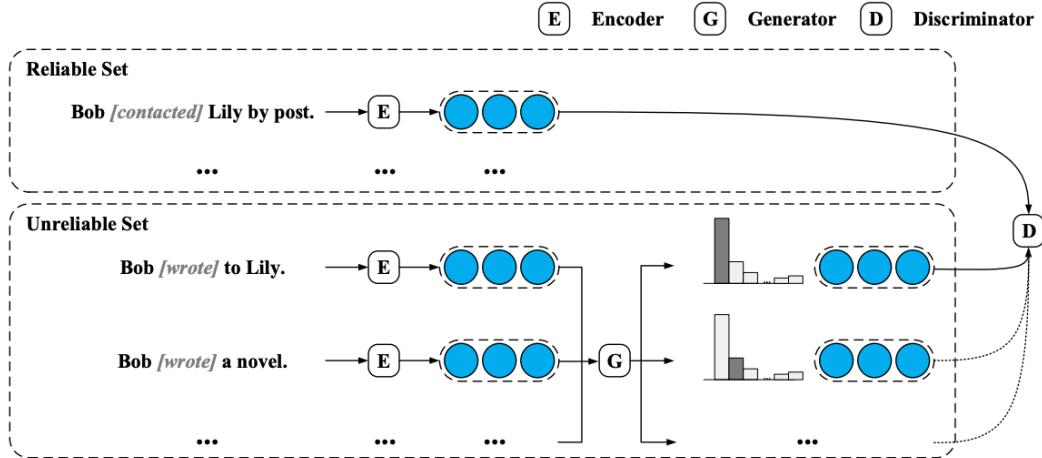


图 2.7 DMBERT+ADV 模型的示意图

另一种思路是利用分类器来挖掘训练数据。例如，Huang 等人在 2012 年提出的模式自举的事件抽取模型^[64]。其基本的框架如图2.8所示。该方法首先利用“种子集”训练一个初始的 SVM 分类器，然后通过分类器打伪标签的方式，从大规模未标注数据中挖掘更多的训练数据。Liao 等人^[65]利用到了三个协同训练的最大熵分类器，从事件触发词、事件要素和事件要素角色相似性的角度来挖掘训练数据。Wang 等人^[66]利用到了对抗生成模型（图2.7），通过生成器挑选可能的事件实例，判别器判断对错的方式挖掘更多的训练数据。

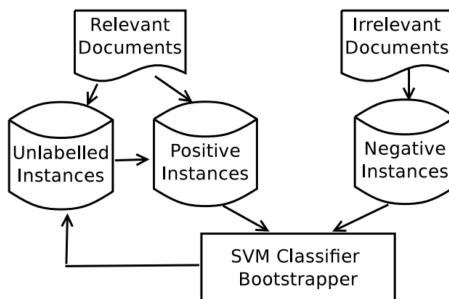


Figure 4: The Bootstrapping Process

图 2.8 半监督模型示意图

2.4.3 基于远程监督的事件抽取方法

有别于半监督的方法从语料中挖掘更多的训练数据，基于远程监督事件抽取方法的思想是利用外部知识库来扩展事件抽取的训练数据。

这类方法的典型代表是 Liu 等人在 2016 年提出的 ANN-FN^[67]模型。图2.9展示了 ANN-FN 模型的架构。ANN-FN 利用本体对齐的方式拓展训练语料。具体来

说，ANN-FN 将训练语料中标记的事件分类体系和 FrameNet 中定义的事件分类体系对齐，通过将 FrameNet 中对齐上的事件实体加入到训练语料，完成数据的拓展。图2.10展示了两个分类体系对齐的情况。

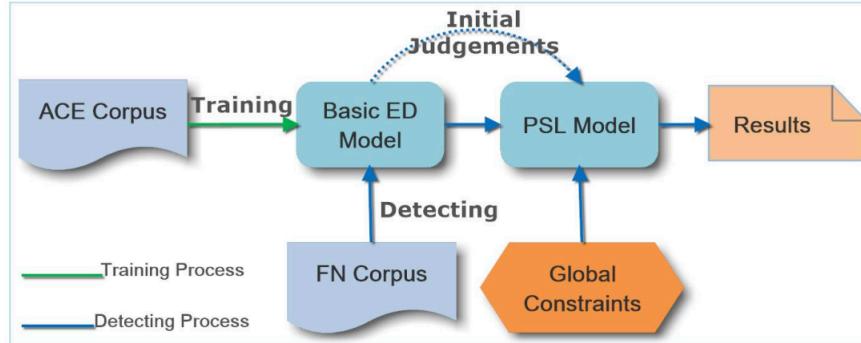


图 2.9 远程监督模型 ANN-FN 示意图

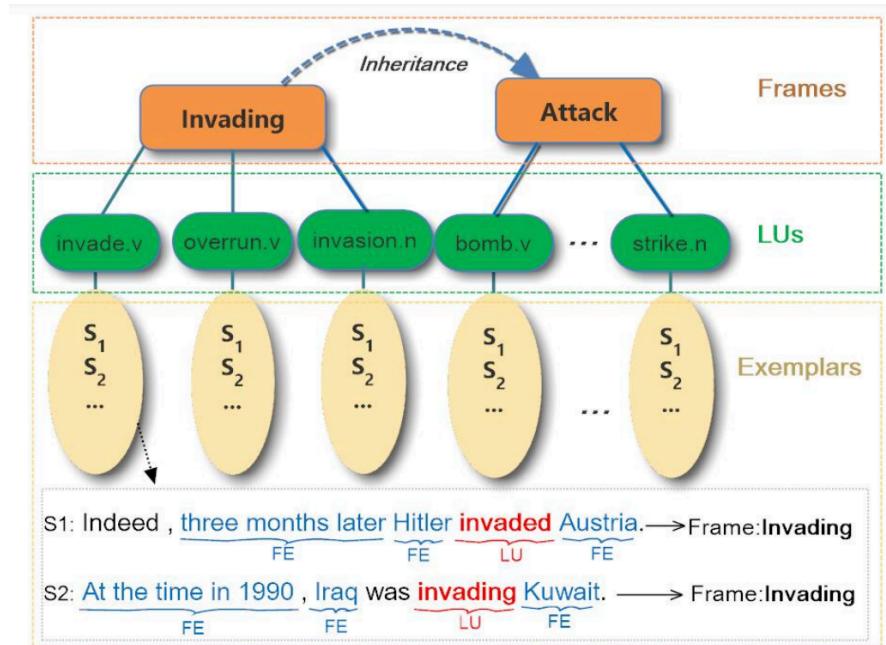


图 2.10 ANN-FN 分类体系对齐示意图

基于远程监督的事件抽取方法除了利用 FrameNet 知识库，还会利用维基百科、WordNet 知识库和一些领域知识库^[68-69]。例如，Reschke 等人^[70]利用维基百科 infobox 作为外部数据源提升了空难事件抽取性能。Araki 等人^[71]利用 WordNet 和基于规则的方法提升了公开域事件抽取的模型效果。Yang 等人^[72]利用从金融知识库中得到的九类事件数据提升了财务公告事件抽取的性能。

2.5 其他方法

上述提到的方法大都是将事件抽取看做是一个序列标注任务，即对句子中的每一个词进行分类，从而完成事件触发词的识别和事件要素的识别。这一节，我们介绍其他方法范式。

基于问答的事件抽取方法。该类方法将事件抽取转化为“问答任务”，以利用阅读理解模型丰富的语义知识，帮助理解事件语义。代表工作是 Du 等人在 2020 年提出的 BERT-QA^[73]模型。图2.11展示了该模型的基本架构。可以看到，该模型设计了两个问答模型分别用于事件检测和事件抽取。事件检测是通过询问问答模型句子中描述的“action”是什么完成的。事件抽取是通过人工设计问题模板，询问问答模型事件的“Who”、“Where”等是什么完成的。

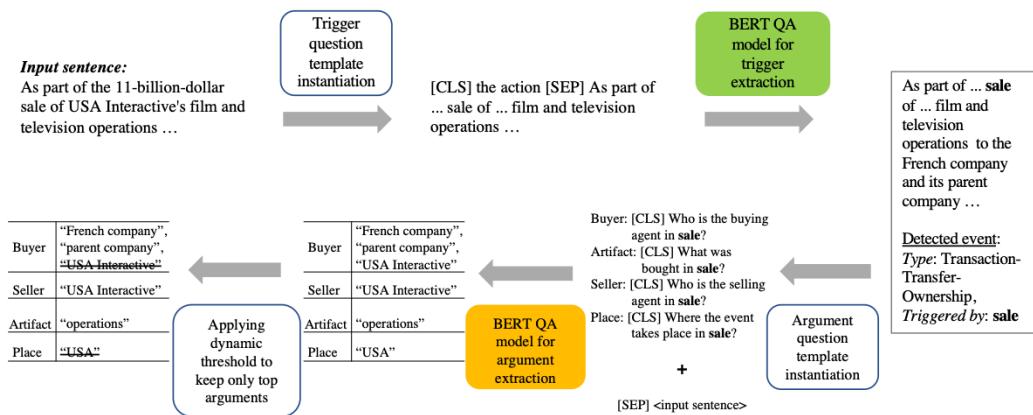
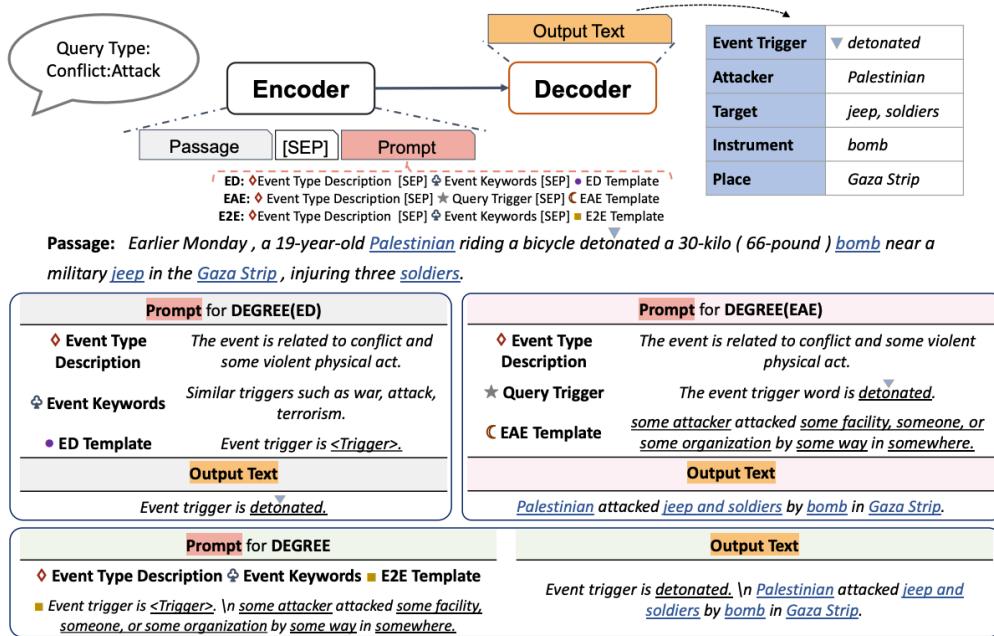


图 2.11 BERT-QA 模型的示意图

除此之外，Liu 等人^[74]提出了一个子问题 + 组装的阅读理解模型。通过对事件类型和事件要素分别设计模板，实际使用时再组装模板的方式，节省了设计自然语言问题的人力成本。Chen 等人^[75]提出一个完形填空式的问答模型，将需要询问的多个事件要素组织成一个需要多处填空的句子，利用问答模型从左到右依次完成填空任务，这种方式能够利用前置的问答结果完成后续的事件要素查询。Wang 等人^[76]提出了一个“询问-抽取”模型，在问题中融入词上下文特征和词性特征，丰富自然语言问题的语义，帮助模型更好的理解问题。Boros 等人^[77]在此基础上提出了一个实体信息增强的问答模型，在问题中融入了多个级别的实体标记，包括实体的位置和实体的类型，帮助模型更好的理解问题。

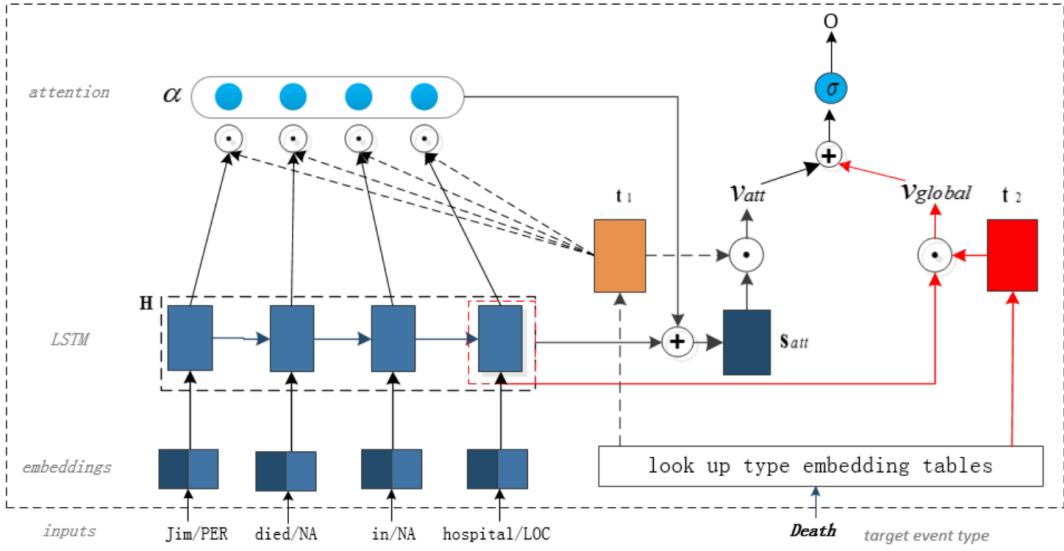
基于生成的事件抽取方法。该类方法将事件抽取转化为“生成任务”，通过编码器编码，解码器生成的方式抽取事件。这类工作代表工作之一是 Hsu 等人在 2021 年提出的 DEGREE^[78]模型。图2.12展示了该模型的基本架构。DEGREE 模型将需要抽取的事件要素转换成一个预先设定好的断言（prompt），和原文一起输入到

图 2.12 BERT-Generation 模型的示意图^[78]

seq2seq 模型，并最终通过序列生成的方式得到答案。

除此之外，Lu 等人^[79]提出了一个基于可控编码的事件抽取模型，通过限制生成过程中只能出现原文中已有的词汇，提升生成答案的质量。Du 等人^[80]提出了一个基于模板填充的事件抽取模型，通过设计事件要素模板，生成模型依次填空的方式完成事件抽取。相比于问答模型需要询问多次才能获得全部的事件抽取结果，基于生成模型的方法仅需一遍就能生成所有的事件抽取结果，效率更高。

基于无触发词的事件抽取方法。该类方法将事件检测任务转化为“多分类任务”。我们知道在 D.Ahn 提出的事件抽取任务划分中将事件抽取分成四个子任务：触发词检测、事件/触发词类型识别、事件论元检测和参数角色识别。按照这种划分，后续的研究大都在抽取事件要素前，先识别触发词，我们称这种先事件触发词检测后事件要素抽取的事件抽取范式为有触发词的事件抽取。目前大都研究都遵从这种范式。然而 Liu 等人^[81]对这个范式提出了质疑，认为在事件抽取中，识别触发词是没有必要的，反而会浪费大量的标注人力。因此 Liu 等人提出了基于类型感知的多标签分类网络 TBNNAM，抛开了事件触发词检测，直接通过对句子进行多元分类（Multi-class Classification），达到从句子中检测多个事件类型的效果。后续的事件要素抽取还是遵循常见的序列标注架构。图2.13展示了这个模型的示意图。

图 2.13 TBNNAM 模型的示意图^[78]

2.6 本章小结

在本章中，我们首先介绍了事件抽取的任务定义以及常用数据集，之后介绍了事件抽取方法发展的主要阶段，并聚焦于目前主流的深度学习事件抽取算法，从语义增强和数据增强的角度上，对相关工作展开介绍。然而，现有这些方法在低资源场景下表现不佳，分析原因有以下几个方面：

- 第一，现有工作在进行事件检测时大都仅仅利用了单一文本模态的信息，而忽略了新闻中天然存在的图像模态的信息。融入多模态信息对低资源场景下事件类型消歧大有裨益，因为图像信息能够提供事件发生时场景、人物动作等额外信息，可以为模型提供更多的事件类型语义指向性证据，从而防止模型过拟合文本数据。我们在第三章提出了一个基于多模态融合的事件抽取模型，通过将高质量的图片信息融入到文本中，提升了模型在低资源场景下对事件类型识别的性能。
- 第二，现有的基于语义增强的方法没有利用与事件检测任务紧密相关的外部知识。他们只利用了篇章、句法、实体等通用语义知识，无法为事件触发词的识别提供非常高质量的先验。针对这一问题，我们在第四章引入了与任务紧密相关的开放域触发词知识，并提出了一个知识蒸馏模型，通过将开放域触发词知识蒸馏到模型中，显著的提升了模型在低资源场景下的事件召回的性能。实验证明我们的方法优于融入实体知识，句法知识等基线模型，取得了更优越的性能。
- 第三，现有的基于数据增强的方法获取的训练数据质量不够高。基于半监督学习的方法获取的训练数据面临同质化的问题，即标注多的类，获得的扩展

数据的规模也会更多，而标注少的类，获得的扩展数据的规模也会比较少，无法从根本上解决标注少的类识别效果差的问题。远程监督方法也受限于外部知识库知识覆盖率不高的问题，无法很好的应对低资源的挑战。针对这一问题，我们在第五章提出了一个基于自标注的事件要素抽取模型。借助预定义类的弱监督信息，从语料中挖掘更多的任务相关类的标注，解决数据稀疏导致的事件要素识别难的问题。

第3章 基于多模态融合的事件检测方法

现有的事件检测方法在低资源场景下面临事件类型消歧难的问题。本章提出了一种基于多模态融合的事件检测方法 DRMM，通过深度融合图像模态的信息，为模型提供更多的语义证据，帮助模型消歧，提升低资源事件检测模型的效果。相比于传统的基于拼接的融合方法和基于互注意力机制的融合方法，DRMM 提出的基于交替对偶注意力机制的融合方法能够对图片中重要信息进行二次筛选，显著的提升了融入图像的质量。与六个最先进的基线相比，我们的方法取得了最佳的效果，证明了所提出多模态融合架构在低资源事件检测任务上有效性。

3.1 引言

事件触发词本身歧义非常大，经常出现同一个触发词触发不同类型事件的情况。现有的事件抽取方法在低资源场景下容易过度拟合训练数据，只能识别触发词在训练集中触发过的事件类型，而忽略了其他的可能性。例如，由于“confront”在训练集中常触发“开会”事件(如图3.1中的 S1 所示)，模型会想当然的认为“confront”一定会触发“开会”事件，而无法识别“confront”触发其他事件的情况(比如在图3.1的 S2 中就触发了“攻击”事件)。

S1: Ford **confront** Meet members in council chamber
S2: Police **confront** Attack protesters hurling stones

图 3.1 事件类型消歧难

为了应对低资源事件触发词歧义性大的挑战，目前的研究工作主要从以下三个方面做了探索：第一个方向是利用篇章信息来消除事件的歧义，代表工作有 Duan 等人的 DLRNN 模型^[6]和 Chen 等人的 HBTNGMA 模型^[41]，他们通过将篇章的上下文信息融入到模型的编码层，帮助模型理解当前事件触发词的语义。第二个方向是使用跨语言资源来增强事件语义理解。例如，Liu 等人^[82]通过门控注意力机制利用中英文语言的互补性，降低触发词歧义性。第三个方向是利用开放域词汇数据库（WordNet、FrameNet）中的常识知识^[67,83]来改进模型的消歧能力。

然而，现有的方法仅利用了单一文本模态的信息，忽略了多模态信息对事件触发词消歧的作用。实际上，利用多模态信息对提升低资源下事件检测模型的消



图 3.2 多模态信息的作用

歧能力是非常“合理”且“有效”的^[84-86]。“合理性”体现在：事件抽取的源数据，例如新闻文章、社交推文，自然是图文并茂的，融合图像信息是有实际数据作为支撑的。“有效性”体现在：（1）图像可以反映文本的核心事件。如图3.2的右图中所示，通过分析新闻中图片反映的核心事件是暴力向的，可以对整篇新闻出现的事件类型定调，在解析事件触发词“confront”的语义的时候，就可以理解它实际上触发了“攻击”事件，而不会过拟合数据认为它触发了“开会”事件。（2）图像可以提供难以用文本描述的信息，例如场景信息、人物动作信息等，可以更好的帮助模型理解事件的语义类型。如图3.2中所示，文本模态的信息无法很直观的反映“confront”触发的是“开会”事件还是“攻击”事件，但是通过观察新闻配图中的场景是会议室还是街道，以及人物的穿着是西装革履还是全副武装，就能够很清晰的分辨出左边的“confront”触发的是“开会”事件，而右边的“confront”触发的是“攻击”事件。

然而，在事件检测中引入多模态信息面临着许多困难。首先，现有语料库在发布的时候未发布相关的图像信息。在 ACE2005 数据集中，标注数据只有新闻文本而缺失原始新闻中的配图。缺失的图片很难用机器自动的还原，需要人工的补全。其次，如何深度融合图像和文本模态的信息仍然是一个开放性的问题。多项研究表明^[87-88]，如果不能够合理的融合异构的图像和文本信息，模型很难从图像中获益，甚至在某些情况下，还会损害模型的性能。第三，我们缺乏新闻图片和待检测事件间明确的对应关系。新闻中的图片往往是对应多段文本的，而事件检测是面向某个句子的，这两者缺乏精准的对齐信息，这为多模态增强的事件检测模型带来了困扰。

为了应对上述挑战，我们首先为通用语料库 ACE2005 手工构建了图像数据集，并提出了一种新颖的双循环多模态融合事件检测方法 DRMM，以提升事件检测在低资源下的性能。具体来说，我们通过搜索原始网站，手动恢复 ACE2005 数据中图像模态信息，并通过四个权威英文网站的图像来扩展我们的图像数据集。通过扩展，我们的数据集能够包含从不同角度描绘事件的图像信息。在此基础上，我们提出了多模态融合的事件检测方法 DRMM，该方法利用循环网络对新闻中的多

张图片依次进行编码，在循环的每一步，我们采用了新颖的交替对偶注意力机制来获取多模态增强的文本表示。交替对偶注意力机制通过双重筛选，能够自动识别事件语义相关的图像，滤除不相关的噪声，从而提升融入的图像特征的质量。

实验表明，我们提出的多模态融合的事件检测方法 DRMM 超越了六个最先进的基线模型。进一步的消融也证明了引入图像模态信息的有效性以及所提出的交替对偶注意力机制在融合图像模态信息时的先进性。

我们的贡献总结如下：

- 我们为事件检测基准 ACE2005 手动恢复了新闻配图，形成了图像数据集，为后续多模态事件检测的研究提供数据支撑。
- 我们提出了一种新颖的多模态融合事件检测方法，提升了事件检测在低资源下的性能。
- 通过在基准 ACE2005 上的实验，证明了图像模态信息有效性和所提的多模态融合架构的先进性。

3.2 相关工作

多模态学习旨在利用来自异构数据源的多种模态的信息提升模型的性能。常见的多模态信息包括图像、视频和音频。近几年来，随着深度学习模型的发展，多模态学习已被广泛用于处理自然语言处理任务，例如命名实体识别任务^[85]和机器翻译任务^[89]。这些方法从补充视觉上下文的角度增强了自然语言模型对短文本和不规范文本的理解，并提出了各种模态融合机制来整合来自不同异构来源的信息。常用的多模态融合机制有：基于矩阵的方法、基于生成模型的方法、基于互注意力机制的方法。

基于矩阵的方法（TFN^[90], LMF^[91]和 PTP^[92]）的核心思想是将来自各个模态的矩阵表示通过矩阵乘的方式融合在一起，缺点是无法剔除各个模态上的冗余信息，忽略了复杂的局部相互关系。基于生成模型的方法（AN^[93]）的核心思路是训练一个编码器 + 解码器，利用其强大的拟合性能将一个模态的向量表示映射到另一个模态的向量表示空间中，从而完成异构模态表示的统一映射，达到异构模态融合的目的。这种方法的缺点是训练不稳定，容易发生模式崩溃的问题。基于互注意力机制的方法（MFN^[94]）的核心思路是在融合各个模态的向量表示时，会基于各个模态信息的重要性动态的分配输入信号的权重，好处是能够剔除冗余的信息，缺点是融合的交互是单次单向的，没有考虑多个模态信息之间的多轮次双向的交互。

聚焦到多模态事件检测任务上，与我们的工作最相关的是张等人在 2017 年的

工作^[95]。如图3.3所示，他们首先通过AMR解析器获取和触发词“confront”语义紧密关联的实体“Mayor Ford”, “members”, “council chamber”，之后为这些实体配上来自于ImageNet的图片，从而为事件抽取提供更多的视觉上下文信息。然而，该方

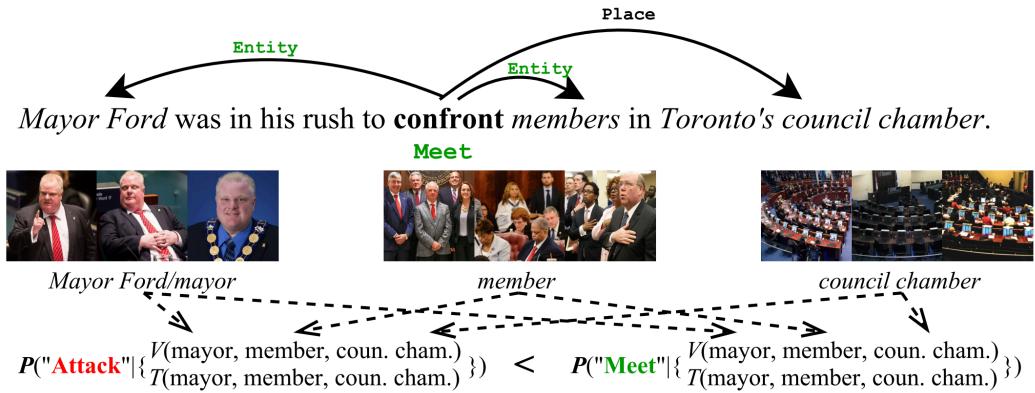


图 3.3 VAD 模型核心思想

法引入的图片信息是对应到实体上的，没有直接反映事件语义，这导致引入的图片信息对事件消歧任务的帮助并不直接，模型的性能也比不过一些纯文本的方法。针对这个问题，我们引入更加直接的新闻配图来做事件检测，这样的好处是能够直接反映文中的核心事件语义，对事件触发词消歧提供更直接的证据。

3.3 模型框架

图3.4展示了我们的多模态融合模型DRMM。模型的输入数据来源自两个方面，一个是文本，一个是新闻中的图片配图。文本和图片都会首先经过DRMM的特征提取模块进行特征化的表示，之后在多模态融合模块完成异构信息的融合。由于新闻配图往往会有许多张，我们在融合图片信息时是采用多次迭代融合的，每输入一张新闻配图，我们都会更新一次文本的表示，直到最后一张图片融合结束。

总的来说，我们提出的多模态融合模型DRMM包括三个模块：特征提取模块、多模态融合模块和事件检测模块。特征提取模块旨在利用预训练模型提取文本模态和图像模态特征。多模态融合模块旨在将异构图像模态特征融入到文本模态特征中，在每一张图片融合时，会使用交替对偶注意机制执行文本和图像模态特征的双向选择，将图像中的要点信息保留到融合后的表示中。事件检测模块旨在对句子中的每个词进行分类，通过将多模态增强后的文本表示映射到事件类型的语言空间中，完成事件类型分类任务。

形式化的，给定新闻中的句子 $S = < w_1, w_2, \dots, w_n >$ 及其该新闻的多张配图 $\Delta = \{p_1, p_2, \dots, p_k\}$ ，事件检测旨在最大化概率 $\frac{1}{n} \sum_{i=1}^n P(y_i | w_i, \Delta)$ ，其中 $y_i \in Y$ ，

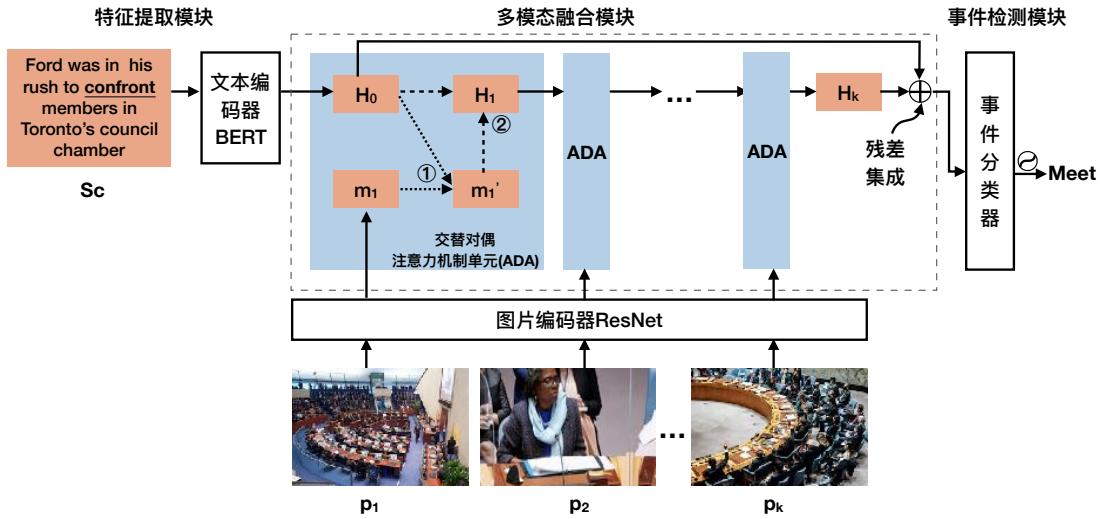


图 3.4 多模态融合模型 DRMM 的框架示意图

$Y = \{y_1, y_2, \dots, y_n\}$ 是预定义事件类型再加上一个其他类的集合

3.4 多模态融合模型 (DRMM)

在本节中，我们介绍提出的多模态增强的事件检测模型 DRMM，旨在利用新闻中天然存在的图片信息，提升模型在事件触发词的消歧能力。我们首先为公开数据集 ACE2005 还原新闻配图，之后从特征提取，多模态融合和事件检测三个方面介绍 DRMM。

3.4.1 图像数据集构建

现有的事件抽取语料库 ACE2005 缺少图像模态信息，在发布时没有将原始新闻中图片一起发布出来。针对这个问题，我们采用众包的方式为 ACE2005 中的数据补充新闻图像。首先，我们利用数据中的 URL 链接到原网站，获取原新闻中的配图作为初始图像集合。由于原网站存在新闻配图缺失或者新闻配图少的问题，我们通过在四个全球新闻网站平台上搜索“相同事件”，并将“相同事件”新闻报道中的图片也包括进来的方式扩展图像集合。四个全球新闻网站包括：CNN，Fox News，NPR 和卫报，这些新闻网站新闻资料全并且能够确保图像质量。

“相同事件”是指共享相同事件类型和事件要素的事件，即主语、宾语、发生地点和发生时间都一致的事件。例如，“野火在 7 月份席卷南加州”与“进入 7 月后，南加州野火失控”是“相同事件”，因为它们共享相同的事件类型“野火”，并且共享相同的事件要素：7 月和南加州。一篇文章的事件要素是利用 AMR^[96] 解析器解析新闻标题获得的。

通过连接“相同事件”的图像，我们获取了多个新闻平台采集的不同的现场图片，大大丰富了图片模态的信息。我们邀请了3名学生参与标注，并采用他们搜索到的图像的并集作为最终的图像集合。最终，我们为ACE2005总共获取了2815张图像。平均每篇文章有4.7张图像，详细信息如表3.1所示。

表3.1 图像数据集统计信息

统计名目	数目
图片总数	2815
每篇文章平均图片数	4.7
每篇文章最多图片数	6
每篇文章最少图片数	3

3.4.2 特征提取模块

在本节中，我们阐述特征提取模块的细节，这包括从文本模态中提取特征和从图像模态中提取特征。

在文本模态特征提取方面，我们采用BERT特征提取器。BERT是一个预训练的语言表示模型^[97]，其通过在大规模无监督语料上接受填字任务的训练获得了不俗的词表示能力，近些年来，在众多自然语言任务（如对话系统、阅读理解和文本摘要）上取得了巨大成功。目前的研究表明^[46,81]，BERT强大的词表示能力同样适用于事件检测任务。

形式化的，给定句子 $S = \langle w_1, w_2, \dots, w_n \rangle$ 经过BERT模型中多层多头注意力机制的编码后输出句子表示为 $H_0 = \langle h_1, h_2, \dots, h_n \rangle$ 。

$$H_0 = BERT(S) \quad (3.1)$$

在图像模态特征提取方面，我们采用ResNet作为特征提取器。ResNet是一个预训练的图像表示模型^[98]，其通过在大规模ImageNet语料上接受图像分类的训练获得了不俗的图像表示能力^[99]，近些年来，在众多计算机视觉任务（如物体检测和物体分割）上取得了巨大成功。

形式化的，给定新闻文章中图片 p_i ，我们采用ResNet最后一个残差块输出作为该图像的向量表示 u_i 。

$$u_i = ResNet(p_i) \quad (3.2)$$

为了将图像映射到与文本相同的纬度空间（从2048到768），我们采用线性函数再

次对图像表示 u_i 进行维度归约化，最终的图像表示 m_i 计算方式如下：

$$m_i = \sigma(W_u u_i + b_u) \quad (3.3)$$

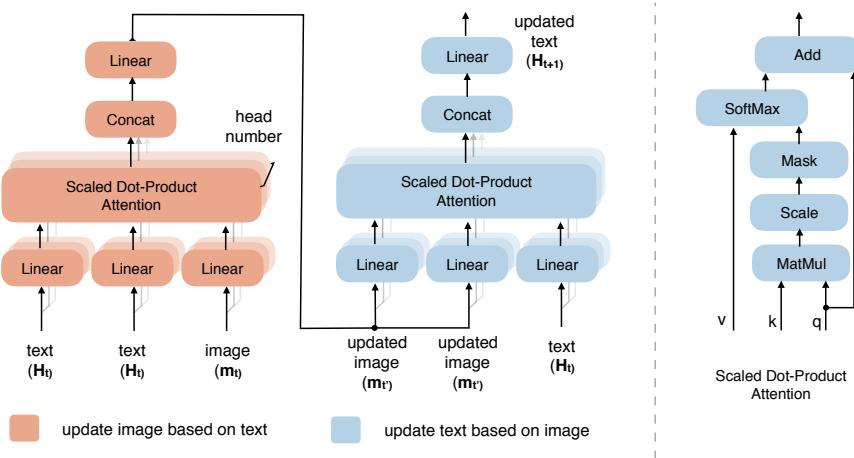
3.4.3 多模态融合模块

多模态信息已被广泛用于处理自然语言问题^[85,89]，能够从视觉的角度增强模型对文本语义理解。在本节中，我们说明了融入多模态信息的过程。与以前仅考虑融合单个图像的方法^[84,89]不同，我们的多模态融合模块能够同时融合多张图像信息辅助事件检测。这更贴合一篇新闻文章含有多张图像信息的实际场景，例如，当我们谈论地震事件时，新闻中会包括道路坍塌的图像也会包括军队救援的图像。不同的图像倾向于从不同的角度描绘一个事件。

多图像编码器的结构展示在图3.4中的虚线框内。多模态融合模块是采用逐步融入多张图片信息的方式动态聚合图像信息。形式化的，在第 t 步，多图像编码器将第 t 张图片信息 m_t 融入到文本表示 H_t 中，将 H_t 更新为新的文本表示 H_{t+1} 。

$$H_{t+1} = ADA(m_t, H_t) \quad (3.4)$$

其中，ADA 指的是我们设计的交替对偶注意力机制，我们在接下来的章节中具体的介绍它的运算过程。



ADA 具有用于深度交互的双重结构，即首先根据文本信息更新图像表示（由红色部分表示），然后反向根据图像信息更新文本表示（由蓝色部分表示）。

图 3.5 交替对偶注意力机制 (ADA) 的图示

交替对偶注意力机制 ADA。交替对偶注意力机制是我们将图片和文本模态的信息进行深度融合的计算单元。图3.5展示了交替对偶注意力机制的具体结构。可以看到，交替对偶注意力机制具有双重结构，它会先基于图像语义对文本重点区域进行筛选，获得文本语义增强的图像表示（第一次筛选），再基于文本语义对图

像重要语义进行筛选，获得图像语义增强的文本表示（第二次筛选）。我们分别介绍两次筛选过程。

图3.5中的红色部分展示了交替对偶注意力机制的第一次筛选过程。该筛选过程是由多头注意力机制实现的。多头注意力机制需要三个输入：查询键值 q ，关键词键值 k 和内容键值 v 。首先，我们用三个全连接网络将文本表示 H_t 分别映射到多头注意力机制的输入 v 和 k 上，将图像表示 m_t 映射到多头注意力机制的输入 q 上。之后，我们采用 q 查询 k 的方式来计算注意力 α 。接下来，我们将计算得到的注意力 α 与 v 进行点积，以获得更新后的图像表示 z 。在计算注意力权重 α 时，我们在数值上额外的惩罚了维度，以解决梯度消失的问题^[100]。

$$\begin{aligned} s &= \frac{\mathbf{q} \cdot \mathbf{k}}{\sqrt{d_k}} \\ \alpha_i &= \frac{s_i}{\sum_{i=1}^L s_i} \\ z &= \alpha \mathbf{v}^T \end{aligned} \tag{3.5}$$

重复上述过程 u 次，并采用线性变换可以得到文本修正图像表示 h

$$\begin{aligned} Z &= [z_1; z_2; \dots; z_u] \\ h &= W_h Z + b_o \end{aligned} \tag{3.6}$$

最后，我们采用残差块将查询信号 q 直接发送到注意力修正后的输出 h 上，以获得最终的图像表示 m'_t 。

$$m'_t = h + q \tag{3.7}$$

我们将公式3.5到公式3.7的运算表示为 Ω ，则文本语义增强的图像表示计算公式如下：

$$m'_t = \Omega(m_t, H_t) \tag{3.8}$$

图3.5中的蓝色部分描述了交替对偶注意力机制的第二个筛选过程。该过程中间的计算方法和第一次筛选相同，但输入的映射上和第一次筛选不同。在该过程中，我们用三个全连接网络将图片表示 m'_t 分别映射到多头注意力机制的输入 v 和 k 上，将图片表示 H_t 映射到多头注意力机制的输入 q 上。最终，图片语义增强的文本表示计算公式如下：

$$H_{t+1} = \Omega(m'_t, H_t) \tag{3.9}$$

至此我们完成了第 t 张图片的融合。对于新闻中的其他图片，我们也依次按照这个方式进行融合。形式化的，给定新闻图集 $\Delta = \langle m_1, m_2, \dots, p_k \rangle$ ，我们依次获得图片语义增强的文本表示 H_1, H_2, \dots, H_k 。

残差集成。我们没有直接采用 H_k 作为最终的多模态表示，而是使用残差块将纯文本表示 H_0 集成回图像增强文本表示 H_k 中。我们希望最终的多模态表示 R 仍然尽可能地保留原始文本语义。我们也从反向传播算法原理的角度来考虑。通过将 BERT 输出 H_0 桥接到最终的多模态表示 R 上，我们可以防止 BERT 中的参数在训练过程中梯度消失。

$$R = H_0 + H_k \quad (3.10)$$

3.4.4 事件检测模块

事件检测模块旨在判断事件触发词的类型，其具体过程如下：如图3.4所示，给定图像增强的文本表示 R ，我们采用全连接网络将 R 的映射到事件类型语义空间上。形式化的，给定第 i 个训练样本 $x_i = \langle S, \Delta \rangle$ ，其中 S, Δ 分别表示新闻文章中的句子以及该新闻中出现图像的集合，事件检测模块输出向量 O ，其中 O_{ijc} 表示 S 中的第 j 个词属于第 c 个事件类的概率。

$$p(y_{(i)}|x_{(i)}, \theta) = \sum_{j=1}^n \frac{\exp(o_{ijc})}{\sum_{c=1}^C \exp(o_{ijc})}/n \quad (3.11)$$

给定输入语料 $D = \{x_i, y_i\}_{i=1}^I$ ，损失函数定义为：

$$J(\theta) = - \sum_{i=1}^I \log p(y_{(i)}|x_{(i)}, \theta) \quad (3.12)$$

我们使用 Adam 作为梯度下降优化器。

3.5 实验

在本节中，我们首先将我们的模型与六种现有的基线方法进行比较，以证明所提出多模态融合方法的优越性。之后的小实验围绕三个大家感兴趣的问题展开：所提方法在低资源场景下的表现如何？构建的图像数据集的质量如何？所提的多模态融合方法的有效性如何？最后，通过案例分析来探究图像在哪些情况下有助于事件检测，哪些情况下又可能损害了事件检测性能。

3.5.1 实验设置

在本节中，我们首先介绍实验用到的数据集，这包括数据集的规模和训练时的数据划分情况，之后介绍实验中超参数的设置情况以及评测模型性能时所用的评测指标。

3.5.1.1 数据集

我们在公开语料库 ACE2005 上展开实验。ACE 语料库包括 599 篇新闻文章，共定义了 8 个事件类型和 33 个事件子类型。我们直接在 33 个事件子类型上进行训练。训练集、验证集和测试集的篇章数量分别为 529/30/40。

3.5.1.2 超参数

我们利用在大规模文本数据上预训练的模型 BERT 对句子进行编码，利用在 ImageNet 图像数据集上预训练的 ResNet 模型对图像进行编码。BERT 采用 8 头注意力机制和 768 嵌入表示维度的 BERT-base 版本。Resnet 采用 50 层卷积神经网络的 Resnet50 版本。我们将 Resnet 的最终池化层从 7*7*1024 特征图扩展为 49*1024 序列以适应交替对偶注意力机制的输入需求。

在训练的过程中，我们采用 Adam 梯度下降算法以保证训练的稳定性。对于超参数，我们采用网格化搜索的方式，将批量大小（batch size）设置的是 32，超参数学习率（learning rate）设为 2e-5。我们选取在验证集上达到最好效果的训练轮数（epoch）作为最终的训练轮数，在本实验中是最终的训练轮数是 4 轮。代码的实现采用了谷歌的 tensorflow 框架。模型所需的算力为一张 3090 显卡。为了体现统计显著性，实验中所有报告的结果都是十次运行结果的平均值。

3.5.1.3 评测指标

我们采用准确率指标、召回率指标和 F1 指标来评测模型的性能。准确率指标计算公式如下：

$$P = \frac{TP}{TP + FP} \quad (3.13)$$

召回率指标计算公式如下：

$$R = \frac{TP}{TP + FN} \quad (3.14)$$

F1 指标计算公式如下：

$$F1 = \frac{2PR}{P + R} \quad (3.15)$$

这里 F1 指标我们采用是 Micro-F1，相比于 Macro-F1，Micro-F1 不区分类别，直接在总体样本上计算。

3.5.1.4 基线

为了验证提出方法的有效性，我们将 DRMM 与以下六个基线模型进行了比较。

- VAD^[95]: 提出一种多模态增强事件检测模型，通过在实体级别结合视觉知识，提升事件检测的性能。
- DLRNN^[6]: 提出一种基于循环神经网络的模型，利用丰富的篇章语义信息，提升事件检测的性能。
- ANN-FN^[67]: 提出一种本体数据增强的事件检测模型，通过将外部知识库 FrameNet 中的本体与 ACE2005 中的本体对齐、拓展训练语料的方式，提升事件检测的性能。
- GM-LATT^[82]: 提出一种基于门控的跨语言事件检测模型，利用中英语言的互补性提升事件检测的性能。
- HBTNGMA^[41]: 提出一种基于分层和偏差标记的事件检测模型，通过改善类别分布的不平衡性，提升事件检测效果。
- AD-DMBERT^[101]: 提出一种基于对抗网络的事件检测模型，通过扩大语料规模，改善事件检测的效果。

3.5.2 整体实验效果分析

表3.2中展示了所提出的方法在 ACE2005 上的整体性能。从表中可以看出，DRMM 优于目前最先进的基线模型，显示了图像模态信息对事件检测的有效性和所提出的多模态融合架构的优越性。

表 3.2 整体实验效果 (%)

Method	Precision	Recall	F1
VAD	75.1	64.3	69.3
DLRNN	77.2	64.9	70.5
ANN-FN	77.6	65.2	70.7
GMLATT	78.9	66.9	72.4
HBTNGMA	77.9	69.1	73.3
AD-DMBERT	77.9	72.5	75.1
DRMM(Our)	77.9	74.8	76.3

与同样融入图片信息的基线方法 VAD 相比，我们的方法具有更加优越的性能（将 F 值提高了 7%）。分析原因，VAD 引入的图片信息是实体图片信息，和事件之间语义关联弱，无法为事件类型消歧提供足够的证据，而我们的方法利用的是新闻中核心事件的图片，和事件的关联性更强，更切合事件检测的场景，因此取得了更好的效果。另外，VAD 仅通过朴素的全连接层融合多模态特征，对来自两个

模态的异构信息融合的不充分，也是导致 VAD 效果不好的原因之一，而我们的方法通过交替对偶注意力机制能够深度融合两种模态的特征，从而达到更好的效果。

和同样融入外部资源的基线方法 ANN-FN、GMLATT、HBTNGMA、TS-DISTILL 和 AD-DMBERT 相比，我们的方法表现出更优秀的性能。分析原因，这些基线模型引入的外部资源仍然限制于单模态的文本资源上，比如 ANN-FN 引入的 FrameNet 资源，GM-LATT 引入的跨语言资源，而我们引入了多模态的图像信息作为额外资源，多模态的图像信息可以提供文本信息难以蕴含的额外知识，比如事件发生时的“场景信息”，“人物动作信息”等，从而帮助我们的方法更好的对事件类型进行消歧。

3.5.3 低资源场景下模型性能分析

在本节中，我们讨论引入图片模态的信息是否能帮助模型在低资源场景下有更好的表现。实验分为三种设定：超低资源、低资源和标注充分。超低资源的实验设定是指每类事件的训练样例数不超过 1 个，低资源的实验设定是指每类事件的训练样例数不超过 10 个，标注充分的实验设定是指每类事件的训练样例数大于 50 个。

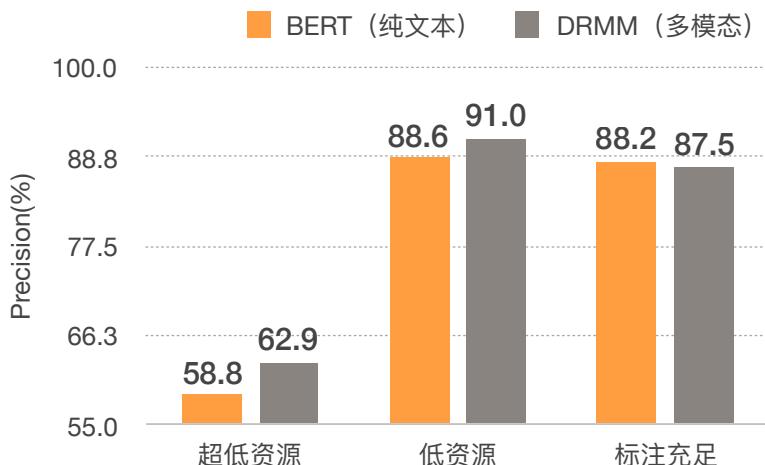


图 3.6 不同数据规模下模型性能对比

图3.6中展示了所提方法 DRMM 和纯文本方法 BERT 在三个场景上的性能表现。实验结果表明，所提的多模态方法 DRMM 在超低资源 (+4.1%) 和低资源 (+2.4%) 场景下的准确率明显高于纯文本方法 BERT，这表明引入图像信息确实有效的提升了低资源场景下事件检测的消歧效果，让模型更容易分辨近似的事件。

3.5.4 图像数据集质量分析

由于图像数据集是本文的贡献之一，我们在这一节评估补充的图像集的质量。为了验证图像数据集的有效性，需要回答两个问题：（1）新闻文本与图像的关联程度。图像必须与其新闻文本密切相关。否则，图像将变成噪音，损害模型的文本理解能力。（2）图像可以为文本的理解提供多少额外的信息。

针对第一个问题，我们训练了一个 *image caption* 模型^[102]。我们将所有的图片放到一个公共池中，*image caption* 模型基于新闻文本从中检索相似图像的，如果模型检索到的图像是正好是数据集中配对的图像，就认为补全的图像确实和新闻文本有较强的相关性。实验表明 *image caption* 模型 top3 的准确率为 75%。很明显，配对的图像与其相应的文章密切相关。

表 3.3 语言模型有/无图片信息性能对比 (%)

Dataset	LM	LM-image
language model	75.8	79.4

针对第二个问题，我们训练了一个语言模型。作为自然语言处理中的基本任务，语言模型性能可以代表文本的理解程度。因此，如果图像信息有助于提升语言模型的性能，就代表图像可以为理解文本提供额外的信息。我们在 BERT^[97]上训练了一个挖空填词的语言模型 (MLM)。由于 ACE2005 是一个小数据集，我们只在 top50 频率的单词上挖空，以将词汇量控制到适当的数量。结果如表3.3所示。我们可以看到图像增强的语言模型的明显优于没有图像增强的语言模型，这表明图像信息对于文本理解的重要性。

3.5.5 图像知识有效性分析

在本节中，我们通过对比有/没有图像知识时模型性能的差异来说明图像知识的有效性。为了更充分的对比，我们在三种编码模型上展开了实验，包括 CNN、RNN 和 BERT。对于 CNN，我们采用滑动窗口大小为 3、4、5 的三层卷积模型。对于 RNN，我们采用两层 Bi-LSTM 模型，嵌入表示的维度为 384。对于 BERT 模型，我们采用 12 层 768 维度的 BERT-base 模型。

如表3.4所示，融合图像知识的模型 (CNN+image, LSTM+image, BERT+image) 始终比不融合图像知识的模型 (CNN, LSTM, BERT) 表现好，这表明引入图像知识确实可以有效的提升模型在事件检测任务上的性能，并且这种有效性是持续存在的，和特定编码模型无关。相比于 BERT 编码模型，CNN 和 LSTM 编码模型在融入图片知识后有更高的性能提升，这反映了当编码模型的拟合能力差的时候，

表3.4 图像知识有效性分析 (%)

Method	Precision	Recall	F1
CNN	72.3	51.2	59.9
CNN+image	74.9	56.1	63.3
improvement	+2.6	+4.9	+3.4
LSTM	71.2	52.2	60.2
LSTM+image	74.3	58.3	64.8
improvement	+3.1	+5.1	+4.6
BERT	76.4	73.8	75.1
BERT+image	77.9	74.8	76.3
improvement	+1.5	+1.0	+1.2

引入图像知识可以带来更大的收益。

3.5.6 多模态融合方法有效性分析

为了证明本文提出的多模态融合方法 DRMM 的有效性，我们将其和三种常见的多模态融合方法进行了比较，包括传统的拼接方法、基于模态注意力机制的方法^[85]和基于互注意力机制的方法^[103]。

表3.5 不同多模态融合方法性能比较 (%)

Fusion Methods	P	R	F1
Concatenation	71.2	67.3	69.2
Modality Attention	78.9	69.4	73.8
Co-Attention	75.3	74.0	74.6
DRMM(our)	77.9	74.8	76.3

表3.5中的实验结果表明 DRMM 比所有常见的融合方法高出 1.5% 以上。传统的拼接方法的性能表现不好，因为拼接方法不分主次，对所有的模态信息一视同仁，这种处理方式不适合文本起主导作用图像起辅助作用的事件检测任务。基于模态注意力机制的方法和基于互注意力机制的方法的表现也不如 DRMM，因为他们没有双向筛选图片中的重要信息。

3.5.7 案例分析

在本节中，我们用案例分析来研究图像在哪些情况下有助于事件检测，哪些情况下损害了事件检测性能。

表 3.6 案例分析

句子	图片	真实标签	纯文本模型 预测结果	多模态模型 预测结果
S1: We do not think that America won , said Dmitry Rogozin.	arm soldier, battlefield, explosion	O	Elect	O
S2: Thousands of Iraq's majority Shiite Muslims march to their main mosque.	protest crowd, chaotic street, shouting	Protest	Transport	Protest
S3: Palestinian forces return before the outbreak of the palestinian uprising .	arm soldier, bloody bus, conflicting	Attack	O	Attack
S4: The EU is set to release 20 million euros in humanitarian aid for iraq	wound people, refugee tents, rescue	Transfer -Money	Transfer -Money	O

表3.6给出了四个案例，用以展示图像模态知识是如何影响事件检测模型预测结果的。

在 S1 中，“won”被错误地认为触发了“Elect”事件，这是由于“won”在训练语料库中的意思大都是选举胜利，纯文本方法过拟合，认为只要出现“won”这个词，都触发了“Elect”事件。图像模态知识“士兵、战场、爆炸”有助于模型分辨这里并没有选举事件发生，使模型正确地将其预测为非触发词。

在 S2 中，事件触发词“march”本身的意思是行军，这使得纯文本方法错误地将其归类为“Transport”事件。然而，考虑到图像模态知识“抗议人群，混乱的街道，喊叫”，“march”在这里更倾向于触发“Protest”事件。

在 S3 中，由于上下文中没有关于骚乱的描述，纯文本方法变得困惑，错误地认为“uprising”没有触发事件。然而，通过图像模态“士兵、血迹墙、冲突”的额外证据，我们的多模态模型成功地认识到“uprising”触发了“攻击”事件。

在少数情况下，图像模态知识会损害事件检测的性能，这主要是图像反映的信息与事件触发词语义无关导致。比如，在 S4 中，“release”会触发的“Transfer-Money”事件，但文章的主要内容是描述伊拉克战争，图片也是如此，因此无法对“Transfer-Money”事件的理解提供语义的帮助。未来，我们将尝试在融入图片信息之前移除不相关或低质量的图像来解决这个问题。

3.6 本章小结

在本章中，我们提出了一个基于多模态增强的事件检测方法 DRMM，利用新闻文章中的天然存在的图像信息来提升低资源事件触发词检测的性能。我们首先为公开事件抽取数据集 ACE2005 手动还原了新闻配图。之后，利用交替对偶注意力机制的双向筛选的机制，将图像和文本这两个异构模态特征进行深度交互，达到将图像中的重要信息注入事件检测模型的目的。我们的方法超越了六个先进的基线模型，验证了所提方法有效性。定性定量的分析也表明构建的图像数据集的质量很高。相关工作 Image Enhanced Event Detection in News Articles^[104] 发表在 AAAI2020 上。

第4章 基于外部知识注入的事件检测方法

上一章中，我们提出了一个基于多模态融合的事件检测方法解决了低资源事件检测中事件类型消歧难的问题。在本章中，我们提出了一个基于知识蒸馏的事件检测方法，通过“引入常识知识”，为模型识别事件触发词提供更多的先验，让模型能够更好的理解数据背后的分布规律，解决模型在低资源场景中事件触发词召回难的问题。所提方法不局限于在标注语料中学习，而是能够同时在大规模未标注语料上学习，有很强的知识利用能力。在公开数据集 ACE2005 上进行的实验表明，我们的模型优于九个强基准模型，证明了所提方法的有效性。

4.1 引言

在低资源场景下，有监督的事件检测方法^[25,105]表现不佳，这具体表现容易漏标事件触发词，从而导致事件召回率偏低。如图4.1中所示，“fire”作为“Attack”事件“训练中常见”的触发词，在预测中就很容易被模型找出来，但是“hacked”作为“Attack”事件“训练中未见”的触发词，在预测中就很难被模型识别。



图 4.1 开放域事件触发词的示例

半监督和远程监督似乎是一种解决方案，他们通过打伪标签（pseudo label）的方式^[56,61,106]或者生成数据的方式^[101]扩展更多的训练数据，为识别触发词提供了更多的监督数据。但实际上，如表4.1所示，半监督和远程监督方法在低资源场景下的表现仍然不令人满意。分析原因，半监督方法依赖训练语料作为初始集，在扩展语料的时候会反复找那些见过的数据，让常见的数据越来越多，无法从根本上改变数据分布。而远程监督的方法也受到知识库覆盖率低的影响。

针对上述问题，在本章中，我们创造性使用开放域触发词知识（Open-Domain Trigger Knowledge）作为额外的常识为模型赋能，提升模型在低资源场景下对事件

表 4.1 现有方法性能对比表

Method	训练未见	训练少见	训练常见
DMBERT _{sup-only}	54.4	72.5	84.1
BOOTSTRAP _{semi-sup}	56.6	73.6	86.9
DGBTBERT _{distant-sup}	54.7	72.8	84.3

触发词的识别能力。开放域触发知识可以认为是一种先验知识，它可以从词义的角度阐释一个词是否触发了一个事件。例如，在图4.1的 S3 中，开放域触发词知识能够基于“hacked”在当前语境下表达了“砍”的词义，将“hacked”识别为事件触发词，因为常识告诉我们“砍”很可能触发了事件。这就为模型识别“训练未见”的事件触发词提供了可能。

直接将开放域触发词知识检测出来的事件当做事件检测结果是不行的，因为开放域触发词知识检测出来的事件类型庞杂，不受预定义事件类型限制，无法和特定数据集上的黄金标签对齐。例如，在图4.1的 S2 中，开放域触发知识认为“trying”, “break”, “stone-throwing”, “use”, “fire”都是触发词，但是 ACE2005 标注中只标记了“fire”是触发词，因为只有“fire”触发的事件处于 ACE2005 预定义的事件类型中。

在本章中，我们提出了一个知识蒸馏 (EKD) 模型，以利用开放域触发知识解决低资源场景下事件召回难的问题。我们首先借助 WordNet 知识库获取开放域触发词知识。在这个过程中，我们先将词消歧到 WordNet 的同义词集合上 (synset)，从同义词集合的角度判断哪些词有触发事件的可能性，这样可以处理一词多义的情景。之后，我们通过边界标记的方式将开放域触发词知识编码到句子中，获得知识增强的训练数据。最后，我们将知识增强的训练数据输入到教师模型中，将原始的训练数据输入到学生模型中，通过让学生模型尽可能的拟合教师模型的预测概率分布的方式，将训练数据中的开放域触发词知识蒸馏到事件检测模型的参数中。由于拟合过程无须黄金标注的监督，我们提出的知识蒸馏模型 EKD 还可以利用大规模未标注语料上蕴含的开放域触发词知识。

我们在 ACE2005 基准上评估我们的模型。我们的方法超过了九个基线，并且对于识别“训练没有见过”的触发词特别有效。

我们的贡献可以概括为：

- 我们创造性的利用丰富的开放域触发知识来提升低资源事件检测的效果。开放域触发知识和事件检测有非常紧密的相关性，能够最大限度的提高模型对事件触发词的理解能力。
- 我们提出了一个新颖的知识蒸馏模型 EKD，通过制造教师模型和学生模型的认知差距，高效的将开放域触发词知识融入到模型的参数中。除了在标记

数据上训练，我们提出的知识蒸馏模型还可以从大规模未标记的数据中持续融合开放域触发知识。

- 在基准 ACE2005 上的实验表明，我们的方法超过了九个同样获得知识增强的强大基线。进一步的研究还表明，我们提出的知识蒸馏模型是知识无关的，还可以适用于融合其他知识到事件检测任务中，例如实体知识、句法知识和事件要素知识。

4.2 模型框架

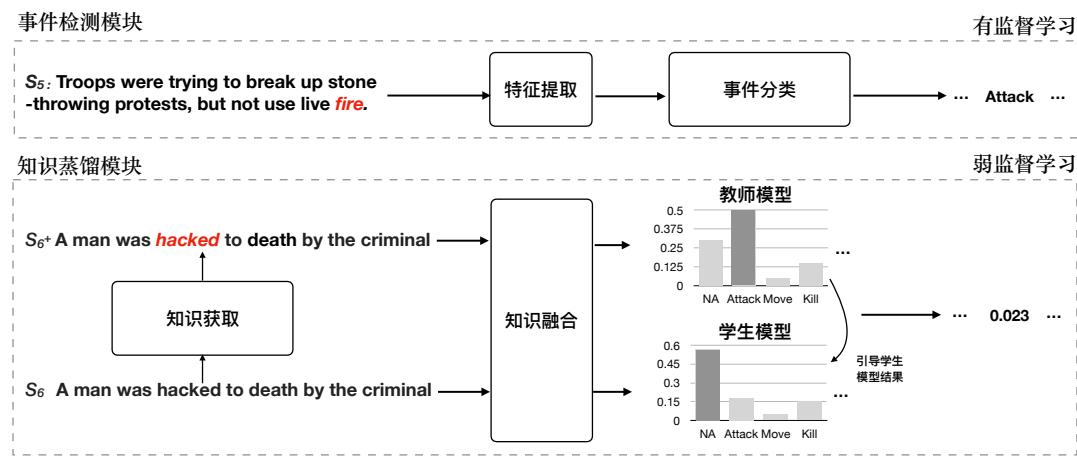


图 4.2 多模态融合模型 EKD 架构图

图 4.2 展示了所提知识蒸馏模型 EKD 的架构。EKD 主要包括两个模块：事件检测模块和知识蒸馏模块。事件检测模块旨在标注数据上训练教师模型，使得教师模型获得很好的事件检测能力，是一个有监督学习过程。知识蒸馏模块旨在从标记语料库和大规模未标记语料库上融合开放域触发词知识，我们通过让学生模型模仿教师模型预测结果的方式，将开放域触发知识提炼到所提模型中，是一个弱监督学习过程。可以看到事件检测模块（前者）是知识蒸馏模块（后者）的基础，只有在教师模型分的足够好的情境下，学生模型才有学习教师模型的必要性。

形式化的，给定标注数据 $L = \{(S_i, Y_i)\}_{i=1}^{N_L}$ ，以及大规模的未标注数据 $U = \{(S_k)\}_{k=N_L+1}^{N_T}$ ，我们的目标是联合优化两个概率函数：(1) 在标注数据 L 上，最大化事件分类概率 $P(Y_i|S_i)$ ，(2) 在标注数据 L 和未标注数据 U 上，最小化教师模型 $P(Y_k|S_k^+)$ 和学生模型 $P(Y_k|S_k^-)$ 之间的预测概率差异，即最小化两个分布的 KL 散度。 S^+ 和 S^- 代表原始句子 S 的知识增强变体和噪音弱化变体，我们将在 4.3.2.1 节中详细解释它们。

4.3 知识蒸馏模型 (EKD)

在本节中，我们具体介绍提出的知识蒸馏模型 EKD 的工作过程，这主要包括事件检测的工作过程以及开放域触发词知识融合的工作过程。

4.3.1 事件检测模块

事件检测模块是一个常规的有监督的学习模块，旨在基于人工标注判断句中的词语触发的事件类型。

4.3.1.1 特征提取

我们采用 BERT 来获得句子的嵌入表示。BERT 是一种预训练的语言表示模型，在问答和语言推理等广泛的任务上取得了最佳的性能。BERT 的强大能力也在事件检测场景中得到了证明^[101]。

形式上，给定句子 S ，我们采用句子编码器 BERT 最后一层的序列输出作为 S 中每个单词的嵌入表示。

$$H = \text{BERT}(S) \quad (4.1)$$

4.3.1.2 事件分类

在获得句子 S 的嵌入表示后，我们采用全连接层来确定句子 S 中每个单词是否触发了某个预定义的事件类型 Y 。我们使用 S_i 和 Y_i 来表示标记语料库 L 中的第 i 个训练句子及其事件类型。我们首先将从第 4.3.1.1 节获得的嵌入表示 H 转换为结果向量 O ，其中 O_{ijc} 表示 S_i 中的第 j 个词属于第 c 个事件类的概率。然后我们通过 softmax 函数对 O 进行归一化，得到分类概率 p ，公式化表达如下：

$$p(Y_i|S_i, \theta) = \sum_{j=1}^n \frac{\exp(O_{ijc})}{\sum_{c=1}^C \exp(O_{ijc})}/n \quad (4.2)$$

给定标注数据 $L = \{(S_i, Y_i)|_{i=1}^{N_L}\}$ ，事件检测模块的损失函数的定义为：

$$J_L(\theta) = - \sum_{i=1}^{N_L} \log p(Y_i|S_i, \theta) \quad (4.3)$$

4.3.2 知识蒸馏模块

知识蒸馏模块旨在将开放域触发知识提炼到我们的模型中，其核心思想是迫使学生模型在标记和未标记数据上生成与教师模型一样好的伪标签。形式上，给

定黄金事件类型标签 Y , 知识蒸馏模块的训练目标是:

$$p(Y|S^+\theta) = p(Y|S^-, \theta) \quad (4.4)$$

其中 $p(Y|S^+\theta)$ 和 $p(Y|S^-, \theta)$ 分别是教师模型和学生模型的预测结果。概率 P 的计算过程和事件检测模块一致。

知识蒸馏模块是一个半监督学习模块, 其不依赖于黄金标注, 能够从标注语料和大规模无标注语料上持续学习。接下来, 我们首先介绍如何获取开放域触发知识。然后, 我们说明如何将开放域触发知识编码到句子中。最后, 我们说明如何融入开放域触发知识。

4.3.2.1 知识获取

开放域触发知识从词义的角度阐述了一个词是否触发了一个事件。例如在图4.1的 S3 中, 开放域触发知识能够基于 hacked 的词义是“砍”, 将 hacked 识别为触发词, 因为常识告诉我们“砍”很可能触发了一个事件。

我们采用轻量级管道方法 Trigger From WordNet (TFW), 收集开放域触发知识^[71]。TFW 方法使用 WordNet 语料库作为中介。在介绍 TFW 算法之前, 我们先对 WordNet 知识库做个简单的介绍。WordNet 是由 Princeton 大学的心理学家, 语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典, 其对每个词进行分组, 分组后的每一组具有相同含义的词称为同义词集 (synset)。每个同义词集都有一个简短、概括的定义, 记录了不同同义词集的语义关系。

TFW 方法有两个步骤: 第一步中, 将词消歧到 WordNet 的同义词集上, 第二步中, 判断同义词集能否触发事件。第一步, 我们采用 It Make Sense 算法^[107]对文本中的词语消歧^[108]。输入到 It Make Sense 算法中的特征是由 Stanford CoreNLP 工具包^[109]中的词性标记器 (POS) 和句法分析器 (Dependence Parser) 中获得的。第二步, 我们采用 Araki 等人^[71]提出的字典映射方法来确定同义词集是否触发事件。TFW 算法是一个通用域的方法, 不限于特定域, 能够从海量的新闻中挖掘开放域触发器知识。在词汇数据库的支持下, TFW 具有很高的效率, 可以应用于大规模的知识收集。

最后, 我们在 2007 年上半年纽约时报语料库 (NYT corpus) 中获得总共 733,848 个带开放域触发词知识的句子。收集到的开放域触发词总数为 265 万, 平均每个句子有 3.6 个触发器。

4.3.2.2 知识嵌入

在本节中, 我们说明如何将开放域触发词知识编码到模型中。

知识增强句 S^+ 我们提出一种边界标记机制将开放域触发知识嵌入到句子中。具体来说，我们引入了两个符号：B-TRI 和 E-TRI 分别标记知识收集模块获得的开放域触发词的开始和结束边界。形式上，给定原始句子 $S = < w_1, w_2, \dots, w_i, \dots, w_n >$ ，不失一般性，假定开放域触发词为 w_i ，那么知识增强句 S^+ 可以表达为 $S^+ = < w_1, w_2, \dots, B-TRI, w_i, E-TRI, \dots, w_n >$ 。标记机制适用于我们的特征提取器 BERT^[110]，它在嵌入知识方面非常灵活，并且可以方便地适应其他类型的知识，而无需进行大量迁移工程工作。

请注意，新添加的符号 B-TRI 和 E-TRI 缺少预训练嵌入表示。我们不能够直接拿随机初始化的值作为 B-TRI 和 E-TRI 的嵌入表示，因为这将破坏引入符号的语义，无法达到我们期望 B-TRI 表示开放域触发词的起始边界，E-TRI 表示开放域触发词的结束边界的语义要求。为了解决这个问题，我们通过预训练语言模型获取 B-TRI 和 E-TRI 的嵌入表示。具体来说，我们采用填词（Masked LM）任务^[97]训练 BERT，基于 Harris 分布假设^[111] 利用周围的词的语义相似性来学习引入符号 B-TRI 和 E-TRI 的语义表示。掩码的概率设置为 0.15，经过训练后 BERT 填词的准确率达到 92.3%，说明 B-TRI 和 E-TRI 的表示已经训练的很充分了。

噪音干扰句 S^- 为了使学生模型和教师模型的认知差距进一步的加大，我们通过屏蔽句子中的词语的方式，进一步干扰学生模型的输入。这样，学生模型必须仅根据周围上下文来判断触发词的事件类型。形式上，给定原始句子 $S = < w_1, w_2, \dots, w_i, \dots, w_n >$ ，以及开放域触发词为 w_i ，噪音干扰句 S^- 可以表达为 $S^- = < w_1, w_2, \dots, [MASK], \dots, w_n >$ 。屏蔽词不是随机选择的，而是限制在知识收集模块找到的开放域触发词之中，这样做好处是可以让学生模型可利用的信息更少，有更强的动机学习教师模型的预测结果。

4.3.2.3 知识融合

在本节中，我们说明知识融合过程。我们将知识增强句 S^+ 输入到教师模型中，噪音干扰句 S^- 输入到学生模型中以分别获取两者的预测概率的表示，具体的计算方式和事件检测模块一样。之后，我们用 KL 散度最小化两个概率分布 $p(Y|S^+, \theta)$ 和 $p(Y|S^-, \theta)$ 之间的差异。一个小细节是我们将添加的符号（如 B-TRI, E-TRI, [MASK]）移动到句子的末尾以确保 S^+ 和 S^- 中的单词严格对齐。

形式化的，给定标记语料库和未标记语料库 $T = \{S_k\}_{k=1}^{N_L+N_U}$ 的集合，知识蒸

馏模块的损失函数为：

$$\begin{aligned} J_T(\theta) &= \text{KL}(p(Y|S^+, \theta) || p(Y|S^-, \theta)) \\ &= \sum_{k=1}^{N_L+N_U} p(Y_k|S_k^+, \theta) \frac{p(Y_k|S_k^+, \theta)}{p(Y_k|S_k^-, \theta)} \end{aligned} \quad (4.5)$$

值得注意的是，KL 散度是个不对称分布拟合函数。在正常的训练中，我们让学生模型拟合教师模型的预测结果。如果我们反过来，让教师模型的预测拟合学生模型的预测，实验结果会显着下降。这样证明了教师模型的预测分布更优，加入开放域触发词确实能改善模型的预测分布。

4.3.3 联合训练

在训练过程中，我们同时优化事件检测模块的损失函数和知识蒸馏模块的损失函数。最终的优化目标形式化表示如下：

$$J(\theta) = J_L(\theta) + J_T(\theta) \quad (4.6)$$

由于未标记数据比标记数据大得多，联合训练导致模型快速过度拟合有限的标记数据，同时仍然欠拟合未标记数据。为了解决这个问题，我们采用了 (Xie et al., 2019) 中提出的训练信号退火 (TSA) 技术，随着训练的进行线性释放标记示例的“训练信号”。我们修正上述损失函数如下：

$$J(\theta) = J_L(\theta) + \lambda J_T(\theta) \quad (4.7)$$

4.4 实验

在本节中，我们将所提方法与九种现有的基线方法进行比较，以证明所提出知识蒸馏方法的优越性。之后的小实验围绕三个大家感兴趣的问题展开：低资源场景下模型性能如何？跨域迁移能力怎么样？提出的知识蒸馏框架是否能利用除了开放域触发词知识以外的其他知识？

4.4.1 实验设置

我们首先介绍实验所用的数据集并给出实验中设置的超参数配置细节，之后，我们给出评测模型性能表现的指标。

4.4.1.1 数据集

我们采用数据集 ACE2005 作为实验的数据集。ACE2005 包含 13,672 个标记句子，这些句子分布在 599 篇文章中。除了预定义的 33 种事件类型外，我们还为

非触发词添加了一个额外的其他类事件类型，以处理某个词不属于任何一个预定义类的情况。跟随 Chen 等人^[112]的工作，我们将 ACE2005 中的 529 篇文章作为训练集，30 篇作为验证集，40 篇作为测试集。

4.4.1.2 超参数

我们采用 BERT-large 的模型提取文本特征，它是一个由 24 个 16 头注意力机制层组成的预训练编码器，嵌入表示的维度为 1024。我们在实验中发现，批量大小这个参数对于我们模型的性能影响很大。经过网格搜索，我们最终在训练事件检测模块时将批量大小设为 32，在训练知识蒸馏模块时将批量大小设为 192，两者比例为 1:6。对于大多数实验，我们将学习率设置为 3e-5，联合训练中的平衡稀疏 λ 设为 1。我们使用 Adam 作为梯度下降优化器，限制输入模型的句子最大长度为 128 个词。训练所提模型需要用到一张 V100 的显卡的算力。模型整体的训练时长大致为 6 个小时。在训练到 12,500 步时，模型在验证集上能够达到最好的预测结果，我们将其作为最终的训练轮数。为了平衡训练效率和模型的精度，除非另有说明，我们实际上使用 40,236 个未标记数据进行知识蒸馏。为了体现统计的显著性，实验中所有报告的结果都是十次运行结果的平均值。

4.4.1.3 评测指标

我们采用准确率指标、召回率指标和 F1 指标来评测模型的性能。准确率指标计算公式如下：

$$P = \frac{TP}{TP + FP} \quad (4.8)$$

召回率指标计算公式如下：

$$R = \frac{TP}{TP + FN} \quad (4.9)$$

F1 指标计算公式如下：

$$F1 = \frac{2PR}{P + R} \quad (4.10)$$

其中 TP 是指真阳例样例，FP 是指假阳例样例，FN 是指假阴例样例。F1 值指标我们具体用的是 Micro-F1，该方式是将所有类的预测结果放到一起计算 F1 值。我们评测是严格的，只有在模型识别出来的事件触发器的类型和边界都与黄金注释是一致的情况下，我们认为该条数模型被正确的预测。

4.4.1.4 基线

我们将我们的方法与两种数据驱动方法和七种最先进的知识增强方法进行比较，包括：

- DMCNN^[5] 提出了一个基于动态池化的卷积神经网络事件抽取模型，可以捕捉同一句子的不同事件的个性化表征，是最早的基于数据驱动的事件检测模型。
- DLRNN^[6] 提出了一个融入更远上下文信息的循环神经网事件抽取模型，该模型同时利用局部特征和句子级上下文特征，有鲁棒的事件检测性能。
- ANN-S2^[113] 提出了一个事件要素信息增强的事件检测模型，通过监督注意机制引导模型在事件触发词识别时关注更为重要的事件要素信息，而不是对所有句子中的所有词语一视同仁。
- GMLATT^[82] 提出了一个基于门控注意力机制的跨语言事件检测模型，来利用不同语言的互补性来对事件触发词消歧提供额外的证据。
- GCN-ED^[46] 提出了一个基于图卷积网络的事件检测模型，通过句法结构树和实体共指图将重要信息的距离拉近，提升模型对重要信息的获取能力。
- Lu's DISTILL^[114] 提出了一种知识蒸馏事件检测方法，旨在从未标记语料中提取通用语义知识，解决模型过拟合的问题。
- TS-DISTILL^[115] 提出了一个基于对比学习的实体知识注入的事件检测模型。
- AD-DMBERT^[101] 提出一个基于对抗生成网络的事件检测模型，采用 BERT 作为句子编码器，通过生成器生成更多的训练数据性来缓解标注数据稀疏的问题。
- DRMM^[104] 提出了一个基于交替对偶注意力机制的多模态事件检测模型。DRMM 采用 BERT 作为句子编码器，ResNet 作为图片编码，通过双向筛选，提升了融入图片的质量。

4.4.2 整体实验效果分析

表4.2展示了我们提出的开放域触发词知识注入模型 EKD 在 ACE2005 上的整体性能。

可以看到，我们提出的 EKD 模型优于多个先进的基线模型，显示了开放域触发知识的优越性和所提出的知识蒸馏模型的有效性。基于 BERT 的模型（AD-DMBERT、DRMM 和 EKD）显着优于基于 CNN 或基于 LSTM 的模型（DMCNN, DLRNN），这是由于 BERT 通过在大规模无监督预料上进行预训练拥有更加优越的词表示能力。在这些基于 BERT 的模型中，我们的方法达到了最好的效果，将

表 4.2 整体实验效果 (%)

Method	P	P	F1
DMCNN	75.6	63.6	69.1
DLRNN	77.2	64.9	70.5
ANN-S2	78.0	66.3	71.7
GMLATT	78.9	66.9	72.4
GCN-ED	77.9	68.8	73.1
Lu's DISTILL	76.3	71.9	74.0
TS-DISTILL	76.8	72.9	74.8
AD-DMBERT	77.9	72.5	75.1
DRMM	77.9	74.8	76.3
EKD (Ours)	79.1	78.0	78.6

F1 分数提高了 3.5% 和 2.3%，这表明注入开放域触发词知识的有效性。

与数据驱动方法 DMCNN 和 DLRNN 相比，知识增强方法 Lu 的 DISTILL、TS-DISTILL 和 EKD（我们的）大大提高了召回率。这是知识增强的方法可以通过引入外部知识，获得分布的先验，防止过度拟合训练数据。在所有的知识增强的方法中，我们的模型取得了最好的性能，这可能是由两个原因造成的：1) 开放域触发知识的优越性。相对于 Lu's DISTILL 中模型使用的通用语言知识和 TS-DISTILL 中使用的实体知识，开放域触发知识能够直接为触发识别提供触发答案，具有更强的任务相关性，蕴含信息量更大。2) 提出的知识蒸馏方法的优越性。我们的方法能够从大规模的未标记数据中持续学习开放域触发知识，而 Lu's DISTILL 和 TS-DISTILL 只能从标记数据中学习外部知识，这种持续学习能力让我们的模型在训练时获得了更多的监督信号，从而达到更高的事件检测性能。

除了召回率的提升，我们的模型在提高事件触发词准确率方面也做的很好。分析原因可能是低频触发词通常是罕见词，而罕见词通常是没有歧义，具有单一的明确定义的含义的。召回这些词之后，模型很容易区分这些词触发的事件类型，从而导致了整体准确率的提高。

为了评估 EKD 是否真的将开放域触发词知识注入到了模型中，我们在表4.3中报告了 EKD 在有知识和没有知识的情况预测的准确率。通过表格中的实验数据可以看出，无论输入数据是否注入了开放领域知识，模型在预测时的性能差异不大 (78.4% vs 78.6%)，这表明 EKD 已经将知识提炼到模型中，也就是说，在预测时，我们的模型不需要额外的工作来收集开放域触发词知识，只需要输入原始文本就可以了，这极大的降低了模型在实际应用时的人工成本。

表 4.3 有/没有开放域触发词知识性能对比 (%)

Test Set	P	R	F
没有知识	78.8	78.1	78.4
有知识	79.1	78.0	78.6

4.4.3 低资源场景下模型性能分析

在本节中，我们利用 ACE2005 数据集，设计了一个低资源场景下的实验，以观察所提的方法是否能够利用开放域触发词知识，为事件识别提供先验，从而提升模型在低资源场景下事件召回率。

实验分为三种设定：超低资源、低资源和标注充分。超低资源的实验设定是指每类事件的训练样例数不超过 1 个，低资源的实验设定是指每类事件的训练样例数不超过 10 个，标注充分的实验设定是指每类事件的训练样例数大于 50 个。我们对比基线包括：

- 监督方法 DMBERT^[5]利用动态池化操作捕获上下文特征的方法提高模型的表征能力。
- 远程监督方法 DGBTBERT^[116]通过和外部知识库 Freebase 中事件本体对齐的方式，将 Freebase 中的事件实例加入到训练数据，拓展现有的数据集。
- 半监督方法 BOOTSTRAP^[117]通过种子集 + 自举挖掘的方式扩展训练数据。

我们将三个基线中的原本的编码器（卷积神经网络以及循环神经网络）替换为更强大的预训练模型 BERT，增强基线的性能，让对比过程更加公平。

如表4.4所示，由于标注数据稀缺，监督的方法，半监督的方法和远程监督的方法在超低资源和低资源这两个场景中性能都有明显的下降，而我们的知识蒸馏方法在超低资源场景 (+6.1%) 和低资源场景 (+2.8%) 上都有着明显的性能提升，这是因为我们的方法通过融合开放域触发词知识，能够从大规模未标记语料库中观察事件触发词的分布，从而在识别事件触发词时有了更多的先验知识，更容易召回罕见的触发词。

4.4.4 跨领迁移能力分析

在本节中，我们使用 ACE2005 来模拟一个跨域的场景，以测试 EKD 模型的领域迁移能力。ACE2005 是一个多领域数据集，其数据采集来源于六个领域域，包括：广播对话 (bc)、广播新闻 (bn)、电话对话 (cts)、电视新闻 (nw)、联合国专线 (usenet) 和网络博客 (wl)。按照研究的惯例^[118-119]，我们采用 bn 和 nw 作为源领域 bc, cts 和 wl 作为目标领域。这样的划分，使得源领域和目标领域之间的

表4.4 不同资源规模下召回率对比表 (%)

Methods	超低资源			低资源			标注充足		
	P	R	F	P	R	F	P	R	F
DMBERT	66.7	45.9	54.4	74.4	70.7	72.5	84.8	83.5	84.1
DGBTERT	76.5	42.6	54.7	75.7	70.1	72.8	85.9	83.8	84.3
BOOTSTRAP	73.7	45.9	56.6	76.0	71.3	73.6	90.6	83.5	86.9
EKD (ours)	79.0	52.0	62.7	80.8	72.4	76.4	92.5	82.2	87.1

表4.5 领域迁移能力 (%)

Methods	In-Domain (bn+nw)			bc			cts			wl		
	P	R	F	P	R	F	P	R	F	P	R	F
MaxEnt	74.5	59.4	66.0	70.1	54.5	61.3	66.4	49.9	56.9	59.4	34.9	43.9
Joint	73.5	62.7	67.7	70.3	57.2	63.1	64.9	50.8	57.0	59.5	38.4	46.7
NguCNN	69.2	67.0	68.0	70.2	65.2	67.6	68.3	58.2	62.8	54.8	42.0	47.5
PLMEE	77.1	65.7	70.1	72.9	67.1	69.9	70.8	64.0	67.2	62.6	51.9	56.7
EKD (ours)	77.8	76.1	76.9	80.8	65.1	72.1	71.7	61.3	66.1	69.0	49.9	57.9

事件类型和词汇分布有很大不同^[118]，加大了领域迁移的难度。

在评估的方面，我们将源域数据拆分为4:1分别作为训练集和测试集。考虑到统计显著性，我们报告十次运行的平均结果作为最终结果。在基线模型选取方面，我们选取了如下几个基线系统：

- MaxEnt 最大熵算法，是一个传统的基于统计分析的事件检测方法。
- Li's Joint^[25] 是一个传统的机器学习算法，利用词汇和全局特征来增强机器学习模型的事件检测能力。
- Nguyen's CNN^[120] 将特征工程中关键特征融入到卷积神经网络中，提升卷积神经网络词表示能力。
- PLMEE^[105] 提出了一个融合预训练模型语义信息的通用事件检测模型。

如表4.5所示，我们的方法在bc和wl目标领域上均超越了现有的基线模型，达到了最佳领域迁移性能，我们在cts目标域上也实现了和现有的基线模型相当的领域迁移性能。EKD拥有优越的跨域迁移能力一个可能的原因是开放领域的触发知识能够在不同领域间建立起一个语义理解的桥梁。开放域触发知识不受特定域的限制，能够召回所有领域的事件的触发词，提供的触发词位置的先验知识也是领域无关的，这让我们的模型在领域迁移时有了额外可以凭借的常识知识，因此在

跨域迁移方面具有更加卓越的性能。

4.4.5 知识蒸馏框架的通用性分析

在本节中，我们验证所提出的知识蒸馏架构是否具有通用性，是否能够将其他知识也注入到事件检测模型中。

为了充分的证明我们提出的知识蒸馏框架的通用性，我们在事件检测场景中常见的三个知识上都展开了实验，包括：(1) 实体知识。实体类型是事件检测消歧中一个重要证据^[121]。(2) 句法知识。句法知识可以从语法的角度为触发词识别提供更多关键上下文。通常来说，句法树上的距离越近的词对触发词越重要^[122]。(3) 事件要素知识。事件要素是指参与事件的关键实体，对事件触发词的语义理解很有帮助。

在获取知识的方面，我们通过 Stanford CoreNLP 获得实体知识的标注，通过 NLP-Cube^[123]获得句法知识的标注，通过 CAMR^[124]获得事件要素知识的标注。标注的具体内容是：(1) 对于实体，我们标记了三种基本实体类型，人员、位置和组织。(2) 对于句法，我们在依赖分析树上取触发词的一阶邻居。我们考虑两个方向的邻居。(3) 对于事件要素，我们重点关注在 AMR 解析结果中充当触发词的 ARG0 和 ARG4 角色的单词^[125]。

在知识编码的方面，我们用知识蒸馏模块（第4.3.2.1节）提到的边界标记机制将实体、句法和事件要素知识在句子中框选出来。为了防止信息泄露（information leaking），我们只在训练过程中使用这些知识，预测过程中不使用这些知识。

我们在三个知识上分别对比的基线方法为：

- TS-DISTILL^[81]，该模型通过对抗师生模型提炼实体类型知识，是目前非常先进的实体知识增强模型。
- GCN-ED^[46]，该模型利用图卷积网络融合句法知识，是目前非常先进的句法知识增强模型。
- ANN-S2^[113]，该模型设计了一个软标签注意力机制利用事件要素知识，是目前非常先进的事件要素知识增强模型。

如表4.6所示，我们在三个知识上模型的性能始终优于基线：EKD-Ent 优于 TS-DISTILL，EKD-Syn 优于 GCN-ED，EKD-Arg 优于 ANN-S2，证明了所提的知识蒸馏模型 EKD 普遍适用于所有类型的知识提炼。EKD 在知识注入上的通用性一方面来自于其灵活的知识编码机制，通过加入两个特殊标记来框处知识的边界，另一方面来自于其优越的知识蒸馏机制，通过增加教师模型和学生模型之间的认知差距，最大限度地将各种知识注入到了模型参数中。

如果我们从融入知识类型的角度观察模型的表现，可以看到融合开放域触发

表 4.6 知识蒸馏框架的通用性分析 (%)

注入知识类型	模型	P	R	F
实体知识	TS-DISTILL	76.8	72.9	74.8
	EKD-Ent	74.5	78.6	76.5
	improvement	-2.3	+4.7	+1.7
句法知识	GCN-ED	77.9	68.8	73.1
	EKD-Syn	76.5	76.3	76.4
	improvement	-1.4	+7.5	+3.3
事件要素知识	ANN-S2	78.0	66.3	71.7
	EKD-Arg	75.8	78.4	77.1
	improvement	-2.2	+12.1	+5.4

知识的模型 (EKD) 优于融合事件要素知识的模型 (EKD-Arg)，优于融入实体知识 (EKD- Ent) 和句法知识的模型 (EKD-Syn)。这说明引入的外部知识与任务相关性越高，引入外部知识的价值就越大。由于开放域触发知识和事件要素知识直接从事件语义单元上考虑重要词，因此它们在事件检测场景下比实体和句法知识更有价值。

4.4.6 案例分析

在本节中，我们举例说明开放域触发词知识是如何增强事件检测模型的召回能力。表4.7给出了三个示例。

在 S1 中，由于 `trek` 是一个在训练过程中从未出现过的触发词，监督方法过度依赖标注数据，对于这种罕见词，监督方法很难识别。开放域触发器知识从 `trek` 意思是“跋涉”的角度，识别出 `trek` 很可能是个事件触发词，帮助模型成功地召回了它。

在 S2 中，`inauguration` 是一个罕见词，在大多数情况下，`inauguration` 不会作为触发词触发“Start-Position（任职）”。监督方法容易过度拟合标记数据从而难以召回 `inauguration`。而开放域触发知识从词义的角度理解什么是触发词，由于 `inauguration` 本身就是“就职典礼”的意思，因此可以成功识别 `inauguration` 为“Start-Position（任职）”事件的触发词。

在 S3 中，`be` 是一个非常模糊的词，在标记数据中，通常不会作为事件的触发词。仅监督的方法容易过度拟合标记数据从而误认为 `be` 在这里也没有触发任何事

表 4.7 案例分析

Sentence	GT	Prediction	
		S	S ⁺
<i>S1: Mr. Caste leaves at 5 A.M. for a train trek to manhattan and does not return util 6 P.M.</i>	Transport	O	Transport
<i>S2: Militants in the region escalate their attacks in the weeks leading up to the inauguration of Nigeria's president.</i>	Start-Position	O	Start-Position
<i>S3: Mr.Mason, who will be president of CBS radio, said that it would play to radio's strengths in delivering local news.</i>	Start-Position	O	Start-Position

件，而开放域触发知识具有词义消歧能力，知道 **be here** 在当前上下文语境下是“占据某个位置”的意思，而不是常见的“有存在的性质”的意思，因此可以成功识别 **be** 为“Start-Position（任职）”事件的触发词。

4.4.7 讨论

获取开放域触发词知识除了可以 WordNet 语义资源库（第4.3.2.1节）以外，也同样可以利用 HowNet 语义资源库。HowNet 是一个常用的中英文双语的语义资源库，其构建了包含 2000 多个义原的精细的语义描述体系，并为十几万个汉语和英语词所代表的概念标注了义原，用于组织词汇知识。前提是需要 HowNet 生态提供词语消歧工具并且建立词和触发词的映射关系集合。之后，就可以直接使用本章所提的知识蒸馏方法 EKD 将从 Hownet 中获取的开放域触发词知识蒸馏到模型的参数中。

4.5 本章小结

针对低资源场景下事件召回难的问题，本章提出了一个基于外部知识注入的事件检测方法，通过引入常识知识，为事件检测提供先验^[11]，有效的提升了事件的召回率。具体来说，我们采用以 WordNet 为媒介的词义消歧算法来进行高效的知识收集，然后我们提出了一个丰富知识蒸馏方法 EKD，通过让学生模型拟合教师模型预测概率分布的弱监督学习方式，从标记数据和大量未标记数据中同时汲取开放域触发知识，有效的将开放域触发词知识融入到模型的参数中。实验表明，我们的方法超过了九个强大的知识增强基线，并且对于低资源事件触发词的召回特别

有效。相关工作 Improving Event Detection via Open-domain Trigger Knowledge^[10] 发表在 ACL2020 上。

第5章 基于自标注的事件要素抽取方法

前两章中，我们从融合“多模态信息”和引入“外部知识”的角度解决了事件检测模型在低资源场景下消歧难和召回难的问题。在这一章中，我们直面低资源下训练语料单一、稀缺的问题，提出了一个基于自标注的事件要素抽取方法，通过“获取更多的标注数据”和已有数据联合训练的方式，提升低资源场景下事件要素抽取的性能。实验结果表明，我们的模型优于五个最先进的模型，证明了所提方法的优越性。进一步的消融实验也从定量和定性的角度上证明了我们获得的自标注数据的质量。

5.1 引言

事件抽取具有高度的领域耦合性，针对不同的领域，事件要素类型差异很大。现实条件不允许我们对每个新出现的事件要素都标注大量数据，因此低资源事件要素引起了广大学者的关注。

对于人类来说，常常只需一个或几个样本，就可以立即识别新定义的事件要素^[126]，然而对于深度学习模型来说，这仍然非常困难^[127]，究其原因，这是因为深度学习参数规模大，而低资源场景中训练数据单一、稀缺，模型容易过拟合训练语料造成的。

为了缓解这种现象，我们希望从语料中获取更多训练数据，从增强训练数据多样性的角度，提升低资源场景下事件要素识别的效果。我们发现其他类（Other Class）中包含丰富的语义，如果我们可以从其他类中挖掘更多的任务相关类出来，就可以获取更多的训练数据，还可以为事件要素识别和消歧提供额外的信息，从而缓解低资源过拟合的现象。如图 5.1 中所示，如果我们可以从其他类中检测到由事件要素的代称组成任务相关类（例如“he”，“professor”），那么根据代称和事件要素间的可替代性^[128]，我们将获得事件要素出现位置的先验知识，为事件要素识别提供更多的助力。此外，如果我们可以从其他类中检测到由“动作”词组成任务相关类（例如“die”，“born”），我们就能够捕获不同事件要素之间的语义关联关系（比如“born”的主语的事件要素类型通常是“人”，宾语的事件要素类型通常是“地点”），从而为事件要素消歧提供更多的语义证据^[129-130]。

基于以上观察，我们提出了基于自标注的事件要素模型 MUCO，旨在从其他类中挖掘更多的任务相关类标注，改进低资源事件要素抽取的性能。然而，从其他类词中检测语义相关的任务相关类具有挑战性，这主要源于两个原因：

● 预定义类 ● 任务相关类

S_1 : Emenya was born in Paris and died in local hospital
 O_1 O_1 O_3

S_2 : Newton is a polymath. He was born in Lincolnshire.
 O_2 O_1

S_3 : The professor from the city studies mathematics
 O_2 O_3 O_1

图 5.1 任务相关类的示例

- 海量的噪音干扰。根据我们的观察，尽管其他类中包含着很多有价值的任务相关类，但其中包含的噪音也很多，例如功能词 the、a 等。这些噪音对事件要素识别任务无法提供任何帮助甚至会造成负面影响。因此，如何免除噪音的干扰，找到高质量的任务相关类是一个很大的挑战。
- 缺少标注数据。我们既缺少任务相关类的标注数据，也缺少任务相关类的元数据描述。在这种情况下，即使是最先进的零样本方法^[131]会失败，因为它们需要类别元数据（类名和类描述）作为已知信息。无监督聚类方法也不能满足质量要求。

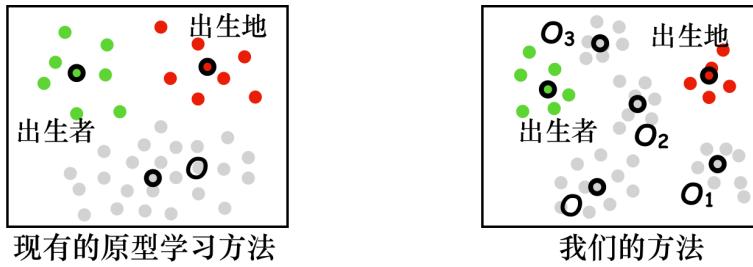


图 5.2 原型网络中处理其他类的方法对比

为了应对上述挑战，我们提出了一种零样本分类方法，利用来自预定义类的弱监督信号，在既没有标注数据，又没有元数据描述的情况下检测任务相关类。该方法的主要思想受到原型网络中迁移学习的启发：如果两个类（A 和 B）是任务相关的，当我们在 A 类上训练原型网络时，即引导 A 类中样本空间上聚集时，B 类中的样本也会自动的在空间上发生聚集，即使我们没有在 B 类上训练^[132]。基于这个观察，我们首先在预定义类上训练原型学习，让预定义类中的单词进行空间上的聚类，然后将其他类中也倾向于聚类的一个个单词群组视为需要挖掘的任务相关类。为了将多个任务相关类相互之间区分开来，我们训练一个二元分类来判断任何两个单词之间是否有聚集的倾向。最后，我们将找到的多个任务相关类标记回句子中，通过联合学习的方式，提升低资源事件要素抽取的性能。

我们的贡献可以总结如下：

- 我们提出了一个新颖的基于自标注的事件要素抽取模型 MUCO，通过从其他类中挖掘多个任务相关类，获得更多训练数据的方式，提升低资源事件要素

抽取的性能。

- 我们提出了一种新颖的零样本分类方法，用于任务相关类挖掘。在既没有标注样本，又没有元数据描述的情况下，我们提出的零样本方法创造性地使用预定义类的弱监督信号来查找任务相关类，实验证明该方法稳定优于无监督聚类的方法。
- 我们在基线数据集 MUC-4 上进行了广泛的实验。与五个最先进的基线相比，我们的方法在低资源的实验设置下达到了最优的性能。进一步的研究表明，我们的自标注方法还具有任务通用性，对事件要素识别和槽填充任务同样有效。

5.2 模型框架

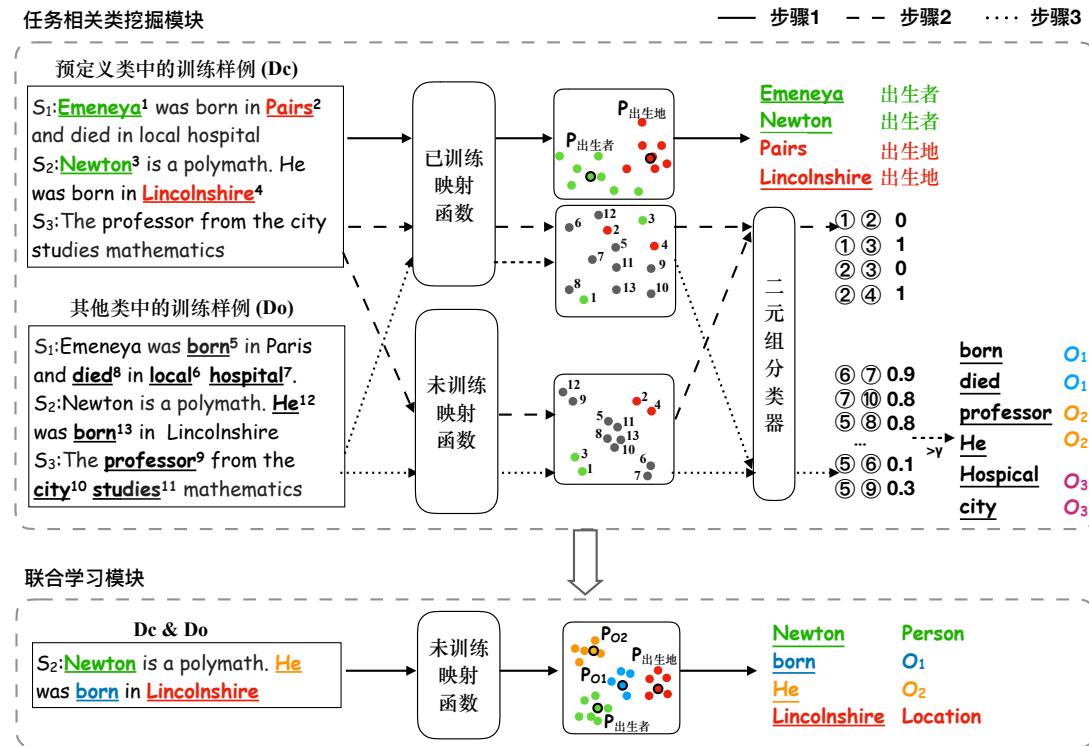


图 5.3 自标注模型 MUCO 的架构示意图

图 5.3 展示了所提出的自标注模型 MUCO 的架构。MUCO 由两个主要模块组成：任务相关类挖掘模块和联合学习模块。任务相关类挖掘模块旨在从其他类中检测任务相关类，以充分利用其他类中丰富的语义。联合学习模块旨在联合学习任务相关类和预定义类，以利用任务相关类中的额外语义知识来增强对预定义类的理解。

我们重点关注任务相关类挖掘模块，这是 MUCO 模型的核心模块。可以看到，

任务相关类挖掘模块的学习过程是一个多步骤的形式（pipeline），我们通过三种不同类型的线段（黑色实线、黑色虚线和灰色实现）来区分这三个步骤的训练过程。在步骤1中，我们通过在预定义类上训练原型网络来学习映射函数。在步骤2中，我们学习一个二元组分类器来判断预定义类中的任何两个点在步骤1训练期间是否倾向于聚类。在步骤3中，我们使用二元组分类器来预测其他类中的样本对，从而得到多个任务相关类的标签。前两个步骤是在为第三个步骤做铺垫。

形式化的，给定训练语料 $D = \{S_i, w_{ij}, y_i\}_{i=1}^{M+N}$ ，其中 $S = < w_1, w_2, \dots, w_n >$ 代表被查询的句子， w_j 代表被查询的事件要素的单词， y 代表事件要素类型的黄金标签， $y \in C \cup O$ ， C 代表预定义的事件要素类， O 代表其他类。我们将 $y \in C$ 的训练样本组成的语料库称为 $D_c = \{x_i, y_i\}_{i=1}^N$ ，将 $y \in O$ 的训练样本组成的语料库称为 $D_o = \{x_i, O\}_{i=1}^M$ 。所提的自标注模型 MUCO 首先在 D_c 上训练原型网络，最大化事件要素分类预测概率 $P(y_i|S_i, w_{ij})$ ， $y_i \in C$ 。之后，从 O 类中检测多个任务相关类，将 O 分解为 o_1, o_2, \dots, or ，为 D_o 中的数据打上细粒度的标签 $D_o = \{x_i, o_i\}_{i=1}^M$ ， $y_i \in O$ 。最后，在训练语料 D_c 和 D_o 上的联合最大化事件要素分类预测概率 $P(y|S_i, w_{ij})$ ，以利用其他类中的丰富语义提升低资源事件要素抽取的性能。

5.3 自标注模型（MUCO）

在本节中，我们首先介绍原型网络的基本思想和相关概念，然后在此基础上介绍我们提出的自标注模型 MUCO。

5.3.1 基础原型网络

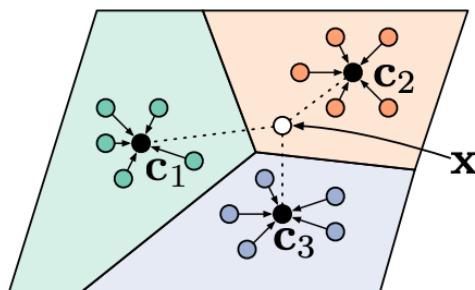


图 5.4 基础原型网络原理示意图^[7]

如图5.4所示，原型网络的基本思想是让同类样本在空间中的距离较近，异类样本在空间中的距离较远，最后通过距离远近划分不同类型的数据。通过在分类前让同类的样本具有更好空间近似性，原型网络可以避免偏离该类语义太多的少

数样本的干扰，提升低资源场景下句子表示的稳定性。

具体来说，原型网络首先通过映射函数将样本映射到一个度量空间。给定输入样本 x ，原型网络映射函数的计算公式是：

$$h = f(x) \quad (5.1)$$

之后，在不断的拟合优化过程中，慢慢的让这个空间中同类样本距离较近，异类样本的距离较远。为了达到这个目的，原型网络会为每个类学习一个原型表示 p 。原型表示相当于给每个类设置了一个锚点，让该类中的样本能够有空间聚集的一个中心点。基础的原型网络^[7]将每类训练样本划分为支撑集 L (support set) 和查询集 Q (query set)。支撑集用来计算每类的原型表示，公式表达如下：

$$p_y = \frac{1}{L_y} \sum_{(x_i, y_i) \in L_y} x_i \quad (5.2)$$

其中 p_y 是类 y 的原型表示， L_y 是类 y 的支撑集。

查询集用来优化样本到原型表示的距离，其计算公式如下：

$$L(\theta) = \frac{1}{Q_y} \sum_{(x_i, y_i) \in Q_y} \frac{\exp(-d(f(x_i), p_{y_i}))}{\sum_{c \in C} \exp(-d(f(x_i), p_c))} \quad (5.3)$$

其中 Q_y 是类 y 的查询集， C 是类别数量。

支撑集和查询集在训练的过程中是动态划分的，某个具体的样例可能在上一步的训练中属于支撑集，下一步的训练中可能就属于查询集了，这导致基础的原型网络面临训练不稳定的问题。针对这个问题，Qi 等人^[133]对基础的原型学习算法做了改进，抛弃了用支撑集计算原型表示的做法，而是在训练的开始随机初始化原型表示，之后通过在训练中拉近该原型表示到类样本间的距离的方式更新原型表示。本文采用这个版本的原型网络作为我们模型的基础架构。

理解了原型网络模型的训练原理之后，我们就可以基于此介绍我们提出的自标注模型 MUCO 中的任务相关类检测模块和联合学习模块。

5.3.2 任务相关类检测模块

任务相关类检测模块的目的是从训练语料中检测更多的任务相关类，从而在没有人工标注的情况下，提升训练语料的规模，解决低资源场景下训练语料单一、稀缺导致的事件要素识别生硬刻板的问题。

从训练语料中检测任务相关类是一项具有挑战性的任务，因为我们缺乏任务相关类的元数据信息以及标注数据信息。更糟糕的是，训练语料中包含的词汇很多，噪音很大，很难区分噪音类和任务相关类。

为了解决这个问题，我们提出了一个零样本分类方法，以利用预定义类的弱

监督信号来检测任务相关类，其基本原理是如果我们在预定义类中训练原型网络，使得其中的样本在空间中产生聚类，那么语义相关的任务相关类中的样本也应该有聚集的迹象^[132]。举例说明，在图5.3中，如果我们引导 Emeneya 和 Newton（绿点1、3）在空间中聚类，语义相关的任务相关类中的样本“Professor”和“He”（灰点9、12）也将倾向于在空间中聚类。

如图5.3所示，我们的零样本分类方法分为三个步骤。在第一步中，我们在预定义类上训练原型网络学习以获得“已训练映射函数”。基于第5.3.1节对原型网络的介绍，样本经过“已训练映射函数”，同类样本将在空间中聚集。在第二步中，我们训练了一个二元组分类器，以判断预定义类中任意两个训练样例是否有空间聚集的倾向。我们通过对比这两个点在“已训练映射函数”和“未训练映射函数”中的位置的变化来做到这一点。在第三步中，我们使用二元组分类器来预测其他类中任意两个训练样例间是否发生聚集，以从其他类中区分多个任务相关类。接下来，我们将依次说明这三个步骤。

5.3.2.1 映射函数学习

我们采用 BERT 模型作为原型网络中的映射函数，这是一种以多头注意力为基本单元的预训练语言表示模型，具有优越的表示能力^[134]。

我们通过在预定义类上训练原型网络的方式来获得“已训练映射函数”。形式化的，给定训练样例 $(x, y) \in D_c$ ，其中 x 由句子 $S = \langle w_1, w_2, \dots, w_n \rangle$ 和查询词 w_j 组成，我们提取 BERT 最后一层的序列输出的第 j 个表示作为该训练样例经过映射函数之后的嵌入表示，形式化的：

$$h = f_\theta(x) \quad (5.4)$$

之后，按照^[133]中提出的原型网络算法，我们在训练开始时随机初始化 y 类的原型表示 p_y ，然后在训练期间缩短 y 类中的训练样例到其原型 p_y 的距离。形式化表达如下：

$$d(x, p_y) = -f_\theta(x)^T p_y \quad (5.5)$$

其中 $f_\theta(x)$ 和 p_y 在计算距离之前，会先通过 L2 距离进行归一化。与传统的原型学习^[7]相比，我们采用的这种算法无须单独的划分支撑集用作原型计算。

最终的训练目标是：

$$L(\theta_1) = -\log \frac{\exp(-d(x, p_y))}{\sum_{p_c \in P_c} \exp(-d(x, p_c))} \quad (5.6)$$

其中 P_c 代表所有预定义类的原型的集合。

5.3.2.2 训练二元组分类器

回想一下，要想从其他类中检测任务相关类，我们需要从海量的其他类单词中找到有聚集趋势的多个单词群组。然而，这里有个棘手的问题，不同任务相关类之间应该怎么划分。针对这个问题，我们的解决思路是判断任何两个点之间是否属于同个任务相关类，其主要思想是，如果我们可以确定任何两个样本是否属于同一个任务相关类，我们就可以将多个任务相关类彼此区分开来。

具体来说，我们基于预定义类的黄金标注语料，训练了一个二元组分类器来判断任何两个点之间是否属于同个任务相关类。具体思路是，通过比较任何两个点在第一步训练开始前的原始位置 h 和训练后位置 \tilde{h} 之间的距离变化，我们可以判断两个点之间是否发生聚合。形式上，给出 D_c 中的一对训练样例 (x_i, y_i) 和 (x_j, y_j) ，他们经过未学习映射函数 $f_\theta(x)$ 后空间位置 h_i, h_j （也是第一步训练前的原始位置），经过已训练映射函数 $\hat{f}_{\theta(x)}$ 后的空间位置 \tilde{h}_i, \tilde{h}_j ，二元组分类器的输入特征计算公式为

$$b_{ij} = W([h_i; h_j; \tilde{h}_i; \tilde{h}_j; |h_i - h_j|; |\tilde{h}_i - \tilde{h}_j|; |h_i - \tilde{h}_i|; |h_j - \tilde{h}_j|]) + b \quad (5.7)$$

二元组分类器的优化目标是

$$L(\theta_2) = \frac{1}{N^2} \sum_i^N \sum_j^N (-y_{ij} * \log(b_{ij}) + (1 - y_{ij}) * \log(1 - b_{ij})) \quad (5.8)$$

其中 N 是预定义类中样本的数量， y_{ij} 指的是这两个训练样例是否属于同一个预定义类。如果 x_i 和 x_j 来自同一个预定义类即 $y_i = y_j$ ，则 y_{ij} 是 1，否则为 0。

5.3.2.3 检测任务相关类

我们利用二元组分类器的预测结果，判断其他类中任何两个点是否属于同一个任务相关类，从而完成从其他类中自动的挖掘任务相关类 $O = \{o_1, o_2, \dots, o_r\}$ 的任务。

形式化的，给定 D_o 中的两个训练样例 x_u 和 x_v ，二元组分类器输出概率 $P(b_{uv}|x_u, x_v)$ 表示 x_u 和 x_v 属于同一个任务相关类的置信度。我们设置了一个阈值 γ 来完成对多个任务相关类的划分：如果 $P(b_{uv}|x_u, x_v)$ 大于设定的阈值 γ ，则 x_u 和 x_v 属于同一个任务相关类。如果连续的词属于同一任务相关类，我们会将这些词视为一个多词实例。其他类中的某些样本可能没有任务相关性。我们假设这些样本来自与任务无关的类，并且没有对这些样本进行进一步分类。

风险控制机制-软标签。二元组分类器的预测可能有错误的地方，为了预防这

种风险，我们加入了软标签机制，用于将置信度高的任务相关类标签和置信度不高的任务相关类标签区分开来。软标签机制的具体做法是：对于每个任务相关类 o_i ，我们将该类下所有样本的均值作为类中心，然后我们将各个样本与其类中心之间归一化后的余弦相似度作为软标签。软标签体现了该样本属于当前任务相关类的可能性有多大。

5.3.3 联合分类模块

在本节中，我们介绍如何使用挖掘出来的任务相关类标签提升低资源事件要素抽取的性能。如图5.3中联合学习模块（Jointly Learning）所示，我们首先将挖掘到任务相关类标签标记回句子中。之后，我们同时在预定义类和找到的任务相关类上展开原型网络训练，以借助任务相关类中的丰富语义提升低资源事件要素抽取的性能。

形式上，给定预定义类 C 和任务相关类 O 中的训练样例 $\{(x, y) \in D_c \cup D_o\}$ ，各类上初始化的原型表示 p_y ，训练目标函数如下：

$$L(\theta_3) = -\log \frac{\exp(-d(x, p_y))}{\sum_{c \in \{C \cup O\}} \exp(-d(x, p_c))} \quad (5.9)$$

与公式5.6相比，我们添加了任务相关类共同进行原型表示学习。

比例因子。在计算训练样本和原型表示之间的距离 $d(f(x), p_y)$ 时， $f(x)$ 和 p_y 被归一化过，也就是说他们的值被限制在 $[-1, 1]$ 之间。这会使得交叉熵损失理论上就不可能降到 0^[133]。举例来说，即使我们的预测结果完全正确：正确类别为 1，错误类别为 -1，输出概率 $p(y|x) = \frac{e^1}{[e^1 + (|CUT|-1)e^{-1}]}$ 仍然无法达到 1。这个问题会随着我们引入更多的任务相关类而变得更加严重。为了缓解这个问题，我们添加一个在所有类中共享的可训练标量 s 来缩放内积^[135]：

$$L(\theta_3) = -\log \frac{\exp(-sd(x, p_y))}{\sum_{p \in \{P_c \cup P_t\}} \exp(-sd(x, p))} \quad (5.10)$$

5.4 实验

在本节中，我们评估提出的自标注算法 MUCO 的性能，我们整体的实验是在小样本这个低资源场景下展开的，主要回答了以下几个问题：所提方法在低资源下的性能如何？所提的任务相关类挖掘算法的有效性如何？挖掘的任务相关类的质量如何？跨域迁移能力如何？

5.4.1 实验设置

我们首先介绍实验所用的数据集，之后，介绍所提方法中使用到的超参数以及实验所用的评测指标。

5.4.1.1 数据集

实验的数据采用的是 MUC-4，MUC-4 数据集包含了 1700 篇恐怖袭击的新闻报道，共预定义了 5 个事件要素角色分别为“恐怖分子”，“恐怖分子所属机构”，“受害者”，“攻击目标”和“攻击方式”。我们将这 5 个事件要素分为两部分：基类和小样本类。基类有“恐怖分子”，“恐怖分子所属机构”，“受害者”，小样本类有“攻击目标”和“攻击方式”。

对于基类，所有标注数据都作为训练数据。对于小样本类，只有 K 个标注数据作为训练数据，其余作为测试数据。也就是说，我们对小样本类采用 N-way K-shot 设置，其中 N 是小样本类的数量， K 是从每个小样本类中采样的标注数据的个数。在我们的实验中， K 设置为 5。需要注意的是，我们不能保证采样时每类训练样例的数量完全等于 K ，因为一句话中会有多个类标签。因此，我们跟随 Alexander 等人^[136]的设定，确保每个小样本类别至少有 K 个标签。

5.4.1.2 超参设置

对于特征提取，我们采用 BERT-base 作为我们的编码网络，它拥有 12 个头注意力层和 768 个隐藏嵌入维度。对于学习率，我们在 $1e-6$ 到 $2e-4$ 的范围内采用网格搜索的方式搜索最优的结果，最终，我们在基类预训练时将学习率设置为 $2e-5$ ，在小样本类微调时将学习率设置为 $5e-6$ 。阈值 γ 设置为 0.68，以确保找到的任务相关类与预定义类充分相关。随机梯度下降算法中的批量大小为 128，输入句子的最大序列长度设置为 128。代码的实现基于 tensorflow 框架。计算的算力需求是一张 32G 内存的 V100 显卡。训练过程持续大约半个小时。我们选取验证集效果最好的训练轮数作为最终的训练轮数，实验中为 100 轮左右。为了体现统计的显著性，实验中所有报告的结果都是十次运行结果的平均值。

5.4.2 评测指标

我们使用精准匹配（Exact Match, EM）和模糊匹配（Head Noun Match Match, HM）这两个指标^[137]来评测模型的性能。精准匹配需要预测答案和标注答案完全一致。模糊匹配更加的宽松，只需要预测答案和标注答案中根名词（head noun）一致就可以。模糊匹配的优点是能够将核心意思正确的预测答案也判为正确，这样

可以减少预测答案由于多标介词或形容词导致整个判错的情况。

5.4.2.1 基线方法

为了更好的评测我们提出的自标注算法 MUCO 在低资源事件要素抽取上的性能，对比的基线模型包括：有监督的方法 BERT，原型网络（PN），语义增强的原型网络（L-TapNet+CDT），预热原型网络（WPN）和元学习网络（MAML）。具体介绍如下：

- BERT^[97]是一个基于预训练模型 BERT 的序列标注的分类器，仅在小样本类上学习。
- PN 原型网络^[7]为每一个类学习一个原型表示，通过聚合同类样本，排斥异类样本的方式提升模型在低资源场景下的鲁棒性。
- L-TapNet+CDT^[138] (LTC) 使用基类和小样本类之间的语义关联来提高原型网络中原型表示的质量。
- WPN 预热原型网络^[136]首先在基类上进行原型网络的预训练，然后在小样本类上进行微调，是原型网络的迁移学习版本。
- MAML 元表示学习模型^[139]，首先在基类上基于能够学习快速适应到新类的训练目标预训练模型的参数，然后在小样本类上微调参数。

5.4.2.2 训练细节

对于我们自己的模型，我们首先在标注数据丰富的基类上预训练，之后再在低资源类上进行微调。这么做的目的是充分利用已有标注类的语料来提高模型在低资源类上的性能。在检测任务相关类时，我们使用来自所有预定义类（包括基类和小样本类）的远程监督信息，以便在预训练和微调之间共享任务相关类的注释，这将提高我们模型的迁移性能。

5.4.3 整体实验效果分析

表5.1展示了该方法在事件要素抽取基准数据集 MUC-4 上的整体性能。可以看到 MUCO（我们的）始终优于最先进的基线模型，显示了其他类中丰富语义知识的有效性以及所提出的自标注模型架构的优越性。

与仅在小样本类上训练的方法（BERT、PN 和 L-TapNet+CDT）相比，基类上预训练的方法（WPN、MAML 和 MUCO）实现了更好的性能，这是因为他们通过利用基类的语义信息，在学习小样本类之前就对小样本类有了一个比较好的初始表示，从而能快速拟合小样本类。基类上预训练的方法中（WPN、MAML 和 MUCO），所提方法 MUCO 取得了最好的性能。以前的方法将其他类视为单个类，忽略了其

表 5.1 整体实验效果 (%)

Methods	EM			HM		
	P	R	F	P	R	F
BERT	84.8	75.1	77.4	86.3	77.2	81.5
PN	81.1	73.6	76.0	83.2	74.7	78.7
L-TapNet+CDT	83.2	72.8	77.7	85.9	76.7	81.0
WPN	86.9	80.0	81.8	87.3	82.4	84.8
MAML	84.6	78.9	81.6	87.6	81.9	84.7
MUCO (ours)	87.5	81.5	83.2	90.5	82.5	86.3

他类中丰富的语义信息。相反，我们的方法从其他类中挖掘更多的任务相关类，从而添加了更多任务相关类的标注数据进行联合训练，这不仅缓解了低资源场景中数据稀缺的困境，而且提供了额外的语义知识来识别并消除事件要素的歧义，从而提高低资源事件要素抽取的性能。

5.4.4 任务相关类挖掘算法的性能分析

在本节中，我们评估所提的零样本分类算法（第5.3.2节）在任务相关类挖掘上的优越性，我们引入基线模型“词聚类算法”（Word-Similarity，WS），将其和我们的零样本分类算法进行性能上的对比。词聚类算法是一个无监督聚类算法，其基本原理是在其他类上执行 KMeans^[140]聚类算法来检测任务相关类，聚类的依据是单词间的语义是否类似，如果类似，就会被分到一个任务相关类中。为了更公平的比较，我们在基线模型词聚类算法上也使用了软标签机制。

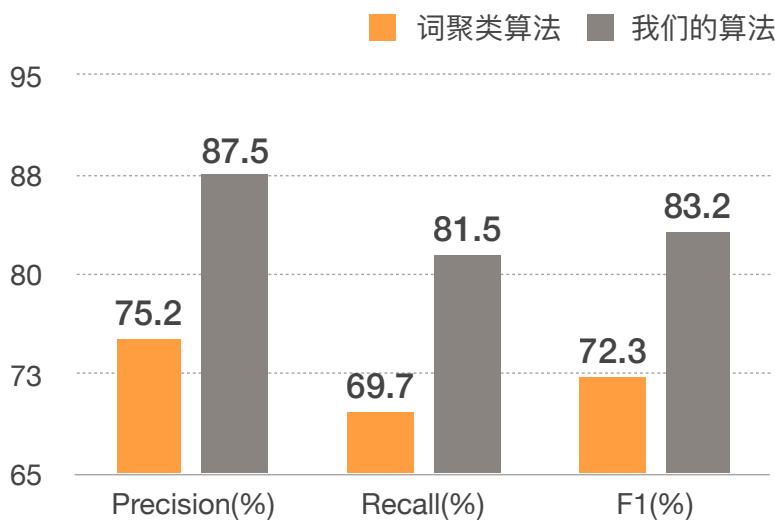


图 5.5 不同任务相关类挖掘算法的对比

如图5.5所示，我们的方法超越了词聚类算法，实现了更好的性能，这显示了我们提出的零样本分类算法在任务相关类挖掘方面的优越性。词聚类算法较差的表现是应为其仅凭借单词自身的语义信息来检测任务相关类，而没有利用来自预定义类的监督信息，因此无法区分嘈杂类（如标点符号）与任务相关类的区别，这不可避免地降低了任务相关类的质量。

5.4.5 任务相关类的质量分析

在本节中，我们从定量和定性的角度评估挖掘的任务相关类的质量。

在定量分析时，我们邀请三名计算机专业的学生手动标记 100 个句子以进行人工评估。评估的指标是“类内相关性”(IC) 和“类间区别度”(ID)。“类内相关性”指标用于评估有多少标签实际上属于声明的类。“类间区别度”指标用于评估有多少标签只属于一个任务相关类，而不是多个类。我们采用多数获胜的投票规则获得标准答案。表5.2报告了人工评估结果。

表 5.2 任务相关类质量定量分析

Metrics	类内相关性	类间区别度
Average Score(%)	49.15	50.85

从表中可以看到，“类内相关性”指标和“类间区分度”指标的得分分别为 49.15% 和 50.85%。考虑到我们在挖掘任务相关类时，是一个零样本场景，既没有类描述的元数据，也没有训练样例，49.15% 和 50.85% 的准确率是可以接受的，也表明找到的任务相关类基本具有一定的类内语义一致性和类间语义差异。

表 5.3 任务相关类质量定性分析

任务相关类	标注的单词
O_1	journalist; president; ambassador; I; he; they; businessmen from;
O_2	the harbour; this land, which; over the river; outsides; the skyline;
O_3	some; a major; the small number; supplied; not only one of the; large;
O_4	believe; comfort; attacked or threatened; arrest; geared; talks
O_5	stop; have; do; discussion; take; seek; sat down; negotiated; think

对于定性分析，我们在表5.3中给出了 MUCO 算法检测到的五个任务相关类的样本。可以看到，任务相关类 O_1 中的词大部分是人称代词、 O_2 中的词大部分是地点代词， O_3 中的词大部分是数字代词。根据语法，事件要素和事件要素代词之

间可以相互替代，检测到事件要素代词组成的任务相关类可以为事件要素识别提供额外的位置先验知识。 O_4 和 O_5 中的词大都是动作，识别动作可以为不同事件要素之间的类型的关联提供更多的证据，并为事件要素消歧提供重要证据。任务相关类中的标注错误主要来自三个方面：(1) 边界划分错误，如 O_1 中，本应该是“businessmen”，但是我们的模型多圈了右边的词，识别成了“businessmen from”。(2) 类内一致性错误。一些奇怪的词降低了类内的一致性，例如 O_3 中的“supplied”。(3) 类间区分度错误，例如， O_4 和 O_5 相似度很高，应该合并成一个类。未来的工作将探索如何提高任务相关类的质量。

5.4.6 任务相关类数量分析

由于我们的模型需要手动设置任务相关类的数量，我们观察模型在不同数量超参设置下的性能。我们通过调整阈值 γ 将任务相关类的数量设置为 1/2/5/10/25/50。

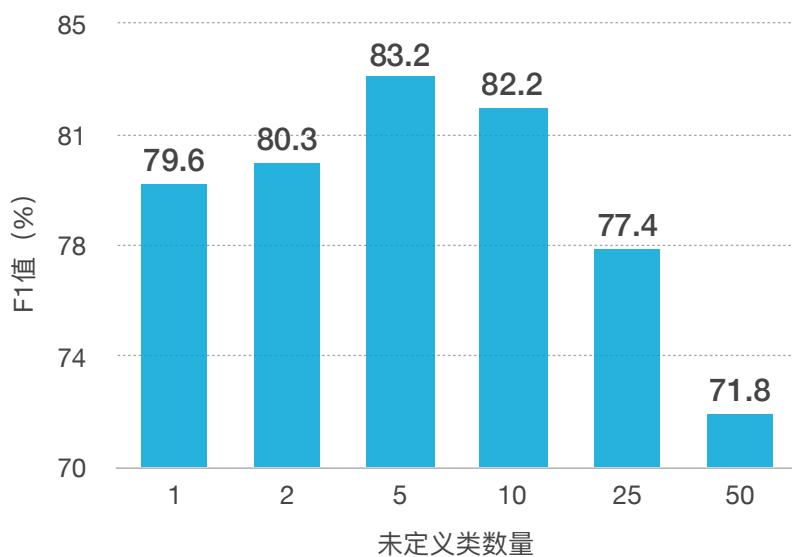


图 5.6 任务相关类数量超参对 MUCO 表现的影响

图5.6说明了我们的模型 MUCO 在各种任务相关类数量超参设置下的 F1 分数。可以看到，数量太大或太小都会影响性能。分析原因，当数量太大时，找到的类会出现重叠问题，导致性能严重下降 (-11.51%)。当数量太少时，模型无法找到足够的与任务相关的类，从而限制了在其他类中捕获细粒度语义的能力。一个经验性的结论是，当任务相关类的数量大约等于小样本类的数量时（这里为 5 类），我们的方法可以达到最佳性能。我们认为预定义类的数量与弱监督中隐藏的信息量成正比。因此，通过更多的预定义类，我们还可以找到更多高质量的任务相关类。

5.4.7 领域迁移能力分析

在本节中，我们将回答我们的模型是否可以在面对不同领域的差异时实现卓越的性能。为了模拟领域迁移场景，我们选择 MUC2004 中的数据作为源域，AnEM^[141]中的数据作为目标域。AnEM 中的类型都是医学学术术语，可以保证与 MUC2004 中数据的差异。

表 5.4 领域迁移能力 (%)

Method	P	R	F
PN	7.3	17.1	7.4
WPN	33.1	32.0	26.9
MUCO (Ours)	34.2	32.8	28.3

如表5.4所示，我们的方法在目标域上实现了最佳适应性能。分析原因，我们在检测任务相关类时，同时使用源域和目标域的预定义类作为远程监督信号，这导致挖掘的任务相关类和源域以及目标域中的类都有很强的任务相关性，我们挖掘的任务相关类的标注可以同时出现在源域以及目标域的训练集中，这提升了两个领域之间数据分布的一致性，提升了我们模型的迁移性能。

5.5 任务通用性分析

在本节中，我们回答我们提出的自标注算法 MUCO 是否对其他自然语言处理任务仍然有效，不局限于事件要素抽取。我们在自然语言处理中两个广泛关注的任务上进行了实验：槽填充任务^[138]和命名实体识别任务。槽填充任务旨在从对话系统中发现用户意图。我们采用 Snips 数据集^[142]进行槽填充任务的实验。命名实体识别任务旨在从句子中识别实体信息。我们采用 Conll2003 数据集进行命名实体识别任务的实验。

表 5.5 槽填充任务实验效果 (%)

Methods	ST		
	P	R	F
PN	61.3	58.2	59.0
WPN	73.6	73.3	70.6
MUCO	75.9	73.7	72.0

表 5.6 命名实体识别任务实验效果 (%)

Methods	NER		
	P	R	F
PN	61.8	58.6	60.1
WPN	65.3	67.7	66.3
MUCO	73.3	69.0	71.1

如表5.5和表5.6所示，所提出的模型在这两个任务上都取得了优异的性能，这证明了我们方法具有任务迁移的能力。无论预定义类属于什么任务，我们的方法总是能够从其他类中挖掘与预定义类相关的任务相关类，这是因为我们的自标注算法完全是数据驱动的，不依赖于任何手动编写任务相关类描述，在不同任务之间迁移的成本是很低的。我们的算法找到的任务相关类能够跟随输入的预定义类的任务类型自动进行更改。

5.6 本章小结

在本文中，我们提出了自标注算法 MUCO，以扩大训练语料的规模，解决低资源场景下事件要素识别生硬刻板的问题。具体来说，我们首先借助预定义类的弱监督信号来从语料中挖掘更多的任务相关类的标注。然后，我们执行联合学习以利用任务相关类中的额外语义知识来增强模型对预定义类的理解。实验表明，我们的方法在公开数据集上取得了最优的结果，相关工作 Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition^[143]发表在 ACL2021 上。

第6章 事件抽取在NewsMiner系统中的应用

在前三章中，我们针对现有的事件抽取方法在低资源下表现不佳的问题，提出了基于多模态融合的事件检测算法，基于外部知识注入的事件检测算法和基于自标注的事件要素抽取算法。在本章中，我们将前三章提到的研究工作在NewsMiner系统中展开了示范应用。我们首先简要介绍NewsMiner系统的情况。之后，针对现有事件分类体系并不能包括用户感兴趣的事件的问题，我们基于新闻学的软硬新闻的理论，重新构建了事件本体，并基于此众包构建了大规模的中英文事件抽取数据集。最后，在此数据集的基础上，我们搭建了中英文事件抽取系统，展现本研究的现实应用价值。

6.1 引言

NewsMiner是面向事件的新闻分析挖掘和搜索系统，由清华大学知识工程实验室发布，致力于利用计算机技术自动化地组织和管理新闻事件，从而帮助普通用户快速了解热点事件的来龙去脉，提高获取信息的效率。如图6.1所示，截止目前，NewsMiner系统中累计收录新闻数据超过1044万条，事件数据超过432万条，实体数据约342万条。

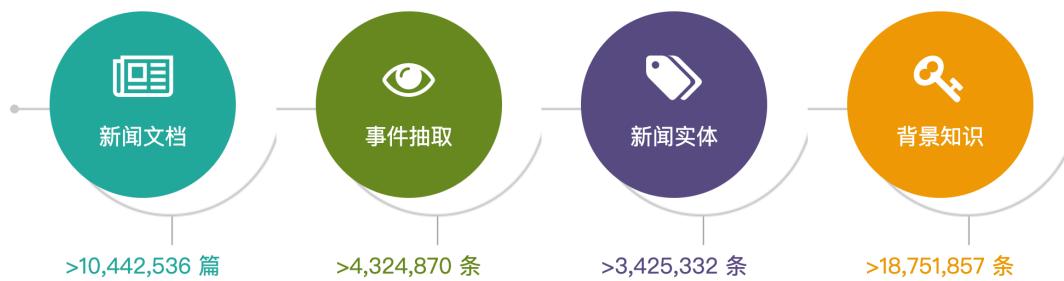


图 6.1 NewsMiner 系统中新闻数量统计

NewsMiner系统主要由四部分组成，包括新闻收集模块、新闻存储模块、事件挖掘模块和语义分析模块。

- **新闻收集模块**负责收集在线新闻文章。该模块监听超过200个国内主流网站，如搜狐新闻、新浪新闻、腾讯新闻、新华网、环球网、人民网等，每日从这些网站收录的新闻规模在8000到12000篇之间，保证主流新闻的高覆盖率，为后续事件挖掘提供了扎实的数据基础。
- **新闻存储模块**负责新闻事件的存储和读写管理，主要用的技术有elastic-

search、mongodb 和 mysql。新闻存储模块每日会将新增的新闻文档定时处理入库，实现自动化存储更新的功能。

- **事件挖掘模块**负责从存储的新闻流中识别热点事件。热点事件往往会被多家媒体报道，出现在多篇新闻文稿中，为此事件挖掘模块提出了一个自适应聚类算法 single pass，以对同类热点事件进行聚类，帮助用户梳理新闻的内容，提高用户阅读新闻的效率。该算法针对适应增量性的聚类场景要求，通过计算两个文档的相似度并设置阈值过滤的方法，能够自动的判断每日新增的新闻是在报道一个新的独立的热点事件，还是是历史热点新闻的后续报道，应该和历史新闻合并。该模块的结果展示在图6.2中。



图 6.2 NewsMiner 热点事件挖掘

- **语义分析模块**主要负责挖掘文本中的语义信息，目前主要包括词云功能和关键词识别功能。词云功能是从词汇的角度对文档进行分析，通过将文档中词汇按照信息量的大小等比例展示成云图的方式，为用户理解文中的核心内容提供帮助。在词云功能中，主要用到的算法是 tf-idf 和最大熵模型。关键词识别是从实体的角度对文档进行分析，通过展示文档中关键人物、地点和组织机构，为用户理解文中的核心内容提供帮助。在关键词识别功能中，主要用到的算法是主题句抽取算法、新闻 5W 要素抽取算法以及 thulac 的词性标注服务和实体类型识别服务。

将事件抽取引入到 NewsMiner 系统是很有必要的。NewsMiner 系统现有的词云和关键词识别等语义分析工具都不足以满足用户对事件信息获取的需求。他们只是简单的罗列文中提到的关键词，人名，地名和组织机构，但并没有更进一步的从事件的角度来组织这些罗列的信息，挖掘这些信息和事件之间具体的语义关联关系。

例如，图6.3所示的新闻介绍了一个印尼地震事件，虽然系统能够将文中提到



图 6.3 NewsMiner 现有语义分析功能示意图

的人物“穆里略”和“林星”识别出来，但是这些人与印尼地震事件的关系如何？他们是地震中的伤者还是扮演了其他的角色？这些语义关联我们都不清楚，简单的罗列两个人名无法为用户理解事件提供足够的信息。同样，“日本气象厅”，“印尼”，“美国”，“尼加拉瓜”地点等的罗列也会让用户感到非常的困惑，这些地点中哪个才是地震发生地，还是都不是地震发生地，这种事件语义的不确定性无法满足用户了解事件全貌的实际诉求。因此，我们很有必要在 NewsMiner 系统中引入事件抽取，从事件的角度对文中的关键信息进行组织，从而满足用户了解事件全貌的要求。

然而，引入事件抽取面临的很多困难。其中一个主要的阻碍就是现有事件抽取数据集中构建的事件分类不符合实际应用需求，无法关注到用户感兴趣的事件。

现有的事件分类体系要不就是太过于泛化，要不就是太过于聚焦到某个特定的领域，都不适合作为新闻核心事件抽取的分类体系。例如，在常用的事件抽取数据集 ACE2005 中，事件类型有“移动 (Transport)”，“死亡 (Die)”和“攻击 (Attack)”等，这些事件类型太过于泛化，无法关注到用户日常阅读新闻会关心的“地震”，“火灾”，“空难”等事件。常用的事件知识库 FrameNet^[144] 和 HowNet^[145] 中，更偏向于关注“吃”、“睡”等琐碎的日常动作事件，语义粒度过细，不适合新闻热点事件这种主题事件的抽取。而特定领域事件抽取数据集定义的事件分类体系又只聚焦在一个小的领域，具有覆盖率低的缺点。例如，Doc2EDAG 只关注金融领域常见的股权冻结、股权回购、股权减持、股权增持和股权质押事件。BioNLP 只关注生物医药领域常见的蛋白质突变、酶转录等事件，他们定义的事件分类体系只适用于领域应用上，无法用在通用域的新闻事件抽取上。

针对现有的事件分类体系不符合实际需求的问题，我们基于新闻学理论重新构建了事件本体，并且基于这个事件本体众包标注了中英双语事件抽取数据集 DocEE，以方便将事件抽取应用到 NewsMiner 系统中。我们提出的 DocEE 数据集有以下三个特点：

- 覆盖面更广的事件分类体系，DocEE关注自然灾害，人为灾害，政治军事事件，经济外交事件和体育娱乐事件，基本覆盖了常见的热点新闻事件的种类，能够满足用户从新闻获取各类事件信息的需求。
- 粒度更细的事件要素体系。除了时间，地点等各个事件类型共享的事件要素，DocEE基于每类的语义，为每类事件设置了更细粒度更专有的事件要素角色，例如，我们为“洪水”事件设置了“水位级别”事件要素角色，为“地震”事件的设置了“震级”事件要素角色。这些细粒度的事件要素角色可以带来更丰富的事件语义，有利于用户全面的了解事件的关键事实点。
- 距离更散的事件要素分布。DocEE去掉了以往数据集对事件要素抽取范围的限制，需要模型在整篇新闻中抽取事件要素。这个设定可以让DocEE直面文章篇幅特别长、事件要素可能出现在文章任意角落的真实新闻场景。

在接下来的章节中，我们详细介绍我们是如何构建DocEE数据集，这包含如何构建事件类型和事件要素类型的本体，如何获取待标数据集以及如何在保证质量的情况下进行众包标注。

6.2 构建事件抽取数据集

在本节中，我们介绍面向应用的事件抽取数据集的构建过程，这包括本体构建和众包标注。

6.2.1 事件本体构建

在构建事件分类本体时，我们的核心思想来自于新闻学的软硬新闻理论。在新闻学的研究认知中，通常将事件分为硬新闻和软新闻^[146-147]。硬新闻是必须立即报告的社会紧急事件，包括（1）社会军事、政治的突变：战争、暴乱、兵变、政权更迭、政策大变动等。（2）重大自然灾害：水灾、旱灾、虫灾、地震、海啸、泥石流等。（3）重大交通事故或其他人为灾害：空难、沉船、撞车等。（4）爆发性流行病：瘟疫、肝炎等局。（5）社会有影响力的新发现或者新成就：卫星上天、登月探险、考古新发现等。软新闻是指更加生活化的新闻，不具有那么强的时效性，更加以人为中心的报道，包括（1）明星琐事：结婚、离婚、参演影视剧等。（2）体育赛事：奥运会、超级碗、世界杯等。基于硬/软新闻理论和Wilzig^[148]给出的事件分类框架，我们共定义了59种事件类型，其中硬新闻事件类型42种，软新闻事件类型17种。详细信息见表6.1。我们定义的事件分类本体基本涵盖了人们阅读新闻时通常关注的有影响力的事情，能够有效的满足用户从新闻中获取事件信息的诉求。

构建好事件分类本体后，我们需要为每类事件匹配事件要素。我们利用了维

表6.1 DocEE事件分类本体

大类	小类
自然灾害	水灾,旱灾,火灾,虫灾,地震,饥荒,海啸 泥石流,暴风雪,火山喷发
人为灾害	空难,沉船,火车碰撞,车祸,银行抢劫 集体中毒,气体爆炸,矿井坍塌
经济事件	组织罚款,组织合并,组织成立,组织倒闭,经济危机,经济援助
外交事件	外交谈判,外交访问,签署协议,撕毁协议,加入组织,退出组织
政治事件	选举,就职,辞职,政府政策变化
暴力冲突事件	罢工,暴动,政权更迭,武装冲突,军事演习,抗议活动
公共卫生事件	环境污染,疾病爆发
科技进步事件	卫星发射,记录突破,考古发现,日食月食
公众人物事件	结婚事件,离婚事件,生病事件,康复事件,死亡事件,调查事件 起诉事件,逮捕事件,判刑事件,释放事件,演讲事件
体育娱乐事件	体育比赛,颁奖典礼

基百科中的信息框来做这个任务。如图6.4所示，维基百科页面包含事件名称，事件描述和事件信息框，其中事件信息框中的键值信息，例如“日期”、“地点”、“乘客”和“死亡”等，可以被认为是当前页面事件的事件要素。基于这一观察，我们为每种事件类型手动收集了20个维基百科页面，并使用这些页面中事件信息框的共享键值作为该类事件的初始事件要素。但是这样收集的事件要素不够全面，有些事件要素维基百科的编辑者们没有给出。因此我们又通过人工总结的方式进一步补充事件要素。具体来说，对于事件类型 e ，我们首先从纽约时报收集20条报道该事件的新闻，然后邀请5名新闻专业的学生总结公众希望从中了解到的关键事件要素 e 。例如，在“地震”事件的新闻中，“震级”是一个关键事件要素，因为它是表达地震灾害程度的关键指标，也是政府救灾时需要考虑到的关键事实点。我们通过合并5个学生提出的关键事件要素，进一步的补充了事件要素。为保证质量，我们在最终形成事件要素体系的时候，会再一次的进行归纳总结，并且去重语义不清楚的事件要素。

最终，我们为59种事件类型定义了356种事件要素类型。平均而言，每个类有5.1个事件要素。图6.5选择了有代表性的五个类的事件要素进行展示。



图 6.4 事件要素初始集合来源示意图



图 6.5 DocEE 中 5 类事件的事件要素类型示意图

6.2.2 待标数据收集

在本节中，我们希望在上一节定义的本体的基础上，构建一个大规模的中英文双语的事件抽取数据集，因此在本节中，我们将分别介绍如何收集英文的待标注数据和如何收集中文的待标注数据。

6.2.2.1 英文待标数据收集

在获取英文待标注数据时，由于 NewsMiner 中监听的英文新闻网站比较少，数据质量也不够高，我们没有选择从 NewsMiner 系统获取待标数据，而是选择英文

维基百科作为我们标注数据的来源。维基百科包含两类英文事件：历史事件和时间线事件^[149]。图 6.6 显示了两个事件的示例。历史事件是指自己有维基页面的事件，例如图 6.6 (a) 中的 *1922 Picardie Air Crash* 事件。时间线事件是指维基编辑者们按时间顺序组织的新闻事件，例如，图 6.6 (b) 中的 *A heatwave 袭击印度和南亚事件*。^①。我们采用这两种事件作为我们的候选数据，因为仅使用历史事件或者仅使用时间线事件都会导致数据在我们的事件体系下分布不均匀，两者可以相互弥补这个缺陷。

<p>1922 Picardie mid-air collision (Title)</p> <p>From Wikipedia, the free encyclopedia</p> <p>The 1922 Picardie mid-air collision took place on 7 April 1922 over Picardie, France, involving British and French passenger-carrying biplanes. The midair collision occurred in foggy conditions. A British aircraft flying Croydon – Paris with only mail on board impacted a French aircraft flying three passengers Paris – Croydon, which resulted in seven deaths.</p> <p>(Article)</p> <p>1922 Picardie mid-air collision</p> <p>Accident</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td>Date</td> <td>7 April 1922</td> </tr> <tr> <td>Summary</td> <td>Mid-air collision in fog</td> </tr> <tr> <td>Site</td> <td>Thieuloy-Saint-Antoine, Picardie, France 49°38'00"N 01°56'49"E</td> </tr> <tr> <td>Total fatalities</td> <td>7 (all)</td> </tr> <tr> <td>Total survivors</td> <td>0</td> </tr> </table> <p>(Infobox)</p>	Date	7 April 1922	Summary	Mid-air collision in fog	Site	Thieuloy-Saint-Antoine, Picardie, France 49°38'00"N 01°56'49"E	Total fatalities	7 (all)	Total survivors	0	<p>June 3, 2007 edit history watch (Sunday)</p> <ul style="list-style-type: none"> • A Paramount Airlines helicopter crashes in Sierra Leone, killing 22 people, with reports of at least one survivor. (BBC)^② (Reuters AlertNet)^② (URL) • 2007 North Lebanon conflict: Soldiers and Islamist militants clash at a second Palestinian refugee camp in Lebanon. (BBC)^② 	<p>Sierra Leone air crash kills 19 (Title)</p> <p>A helicopter ferrying passengers to Freetown airport in Sierra Leone has crashed, killing 19 people, including Togo's Sports Minister Richard Attipoe.</p> <p>The passengers were returning from watching Togo beat Sierra Leone 1-0 in an African Nations Cup qualifier.</p> <p>The helicopter shuttle to the airport takes seven minutes</p> <p>One of the two Ukrainian pilots survived when the helicopter burst into flames as it came into land.</p> <p>Helicopters and ferries are the only way to reach the airport, which is located across a bay from Freetown.</p> <p>The Togolese passengers had chartered the helicopter for the seven-minute flight from the city to the airport.</p> <p>(Article)</p> 
Date	7 April 1922											
Summary	Mid-air collision in fog											
Site	Thieuloy-Saint-Antoine, Picardie, France 49°38'00"N 01°56'49"E											
Total fatalities	7 (all)											
Total survivors	0											

图 6.6 DocEE 中待标数据的两个来源

对于历史事件，我们采用它自身的维基页面作为待标注的文档。对于时间线事件，我们使用 URL 下载原始新闻文章作为待标注事件的文档。因为 22% 的时间线事件有 URL 的问题（维基百科编辑在编辑条目时没有提供 URL），所以我们又额外的使用 Scale SERP^②工具从网易端查找可替代的新闻文章，并手工确认找到的替代新闻文章的有效性。对于历史事件，我们通过检索关键词的方式从维基百科的库中挑选每类的待标事件。检索关键词包括 "List of" + 事件类型、事件类型 + "in" + 年份、"Category:" + 事件类型 + "in" + 国家等。对于时间线事件，我们选择维基工作者在 1980 年到 2021 年之间的整理的新闻事件作为待标数据集。最终，我们从维基百科中收集了 44,000 篇英文待标数据。

6.2.2.2 中文待标数据收集

在获取中文的待标注数据时，我们主要利用中文维基百科数据和 NewsMiner 系统中在线新闻数据。中文维基百科我们仅用有中文维基页面的历史事件，检索数据的方法和英文一样。对于 NewsMiner 系统中在线新闻数据，我们选取了主流的 7 个新闻网站中的新闻作为我们中文事件抽取待标数据的信息来源，包括：搜狐新闻、新浪新闻、腾讯新闻、新华网、环球网和人民网，选取的待标数据的时间范围是 2019-2021 年。在获取 NewsMiner 系统中在线新闻数据待标数据时，为了

① en.wikipedia.org/wiki/Portal:Current_events/June_2010

② <https://app.scaleserp.com/playground>

防止某类上的新闻实例过少，我们按照类别依次查找数据，检索的关键词是类名和 tf-idf 算法计算出来的该类新闻中的高频词。最终，我们收集了 60000 篇中文待标数据。

6.2.3 众包标注

在本节中，我们介绍众包标注中英文事件抽取数据集的过程。

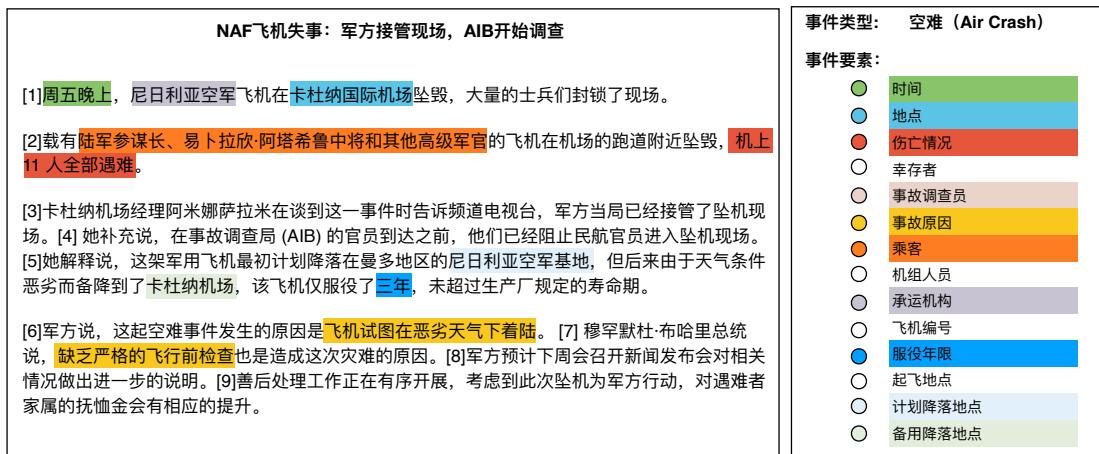


图 6.7 DocEE 众包标注系统

图6.7给出了DocEE中一个标注实例。我们的众包标注过程包括两个阶段：事件检测和事件要素抽取。

6.2.3.1 事件检测

在事件检测标注阶段，标注者需要对新闻中提到的核心事件进行分类。值得注意的是，相比于尽可能的抽取新闻中的事件，我们聚焦于核心事件抽取。核心事件是指新闻标题中反映的并且全文主要在讲述的事件。这个设定符合用户在阅读新闻时的习惯：用户更想了解新闻报道中核心事件的信息，而并不会很在意文中提到的其他的一些旁枝末节的事件。形式上，给定待标新闻 $e = \langle t, a \rangle$ ，其中 t 代表标题， a 代表文章，事件检测众包标注阶段旨在为每篇新闻 e 获取事件分类标签 y 。

在正式标注开始之前，我们首先人工标记了 100 篇新闻作为黄金标注，将标注准确率低于 70% 的标注人员剔除，最终参与该阶段标注的标注员有 48 个人。每篇文章都会由两个独立的标注员分别标注。如果两个标注员的结果不一致（英文有 32.8% 的数据不一致，中文有 23.6% 的数据不一致），我们会让第三个标注员重新标注，并采用第三个标注员的结果作为最终的结果。对于不属于任何预定义类的待标数据，我们将其归入到其他类中。

6.2.3.2 事件要素抽取

在事件要素抽取阶段，标注者需要从新闻全文中查找核心事件的事件要素。我们邀请了 90 名标注者参与该阶段的标注。我们使用“初步标注 + 多轮修正”方法进行众包标注。最开始，每篇文章会先被粗略标注一遍。基于标注结果，我们会总结存在的问题，形式标注手册，培训标注员。之后，我们展开多轮修正的阶段，每篇文章会经过三轮的标注修改，每一轮同一篇文章分配的标注员是不一样，保证每篇文章被至少三个标注员审核过。在每一轮之后，我们都会将问题反馈给标注人员，以便他们在下一轮标注中进行更正。经过多轮的修正，抽检的标注准确率从 56.24%、76.83% 稳步提高到 85.96%。

我们在标注事件要素时，遵循最简标注原则。例如，我们在“一辆损坏的汽车”、“属于受害者的损坏的汽车”和“损坏的汽车”中选择“损坏的汽车”作为我们的标注答案。对于文档中有多个提及的事件要素，我们将标记所有提及以确保标注的完整性。例如，图 6.7 中的 *Cause of the Accident*（事故原因）事件要素有两个提及“trying to land in bad weather”和“the lack of strict pre-flight inspections”，标注时，标注员需要将这两个提及都标注出来。

为了保证标注的质量，所有的英文标注员均是以英语为母语的学生或者托福 100 分以上英语专业学生，所有的中文标注员均为以中文为母语的大学生。

6.2.3.3 标注质量分析

我们采用标注一致性 (Inter-Annotator Agreement, IAA) 指标 kappa^[150-151] 来衡量标注质量。在第一阶段事件检测标注中，英文和中文的 kappa 值分别是 94% 和 95%。在第二阶段事件要素抽取标注中，英文和中文的 kappa 值分别是 81% 和 83%，都达到了较高的分数，这说明在众包标注的过程中，不同标注员标注的结果一致性还是很高的，能够保证数据集的标注质量。在标注花销方面，标注英文数据的成本是 8 元/条，标注中文数据的成本是 2 元/条。

6.3 DocEE 数据统计分析

在本节中，我们展示英文和中文 DocEE 数据统计分析的结果，以方便读者更直观的了解 DocEE 的数据规模以及数据的分布情况。

6.3.1 总体统计数据

英文 DocEE 总共标记了 27,485 篇新闻文章，180,528 个事件要素，平均每篇新闻有 6.6 个事件要素。其中，“离婚”事件上标注的事件要素最多，平均每篇文章

表 6.2 英文 DocEE 和现有的英文事件抽取数据集对比

数据集	事件类型	事件要素角色	篇章数	词汇数	句子数	事件要素实例数
ACE2005	33	35	599	290k	15,789	9,590
KBP2016	18	20	169	94k	5,295	7,919
KBP2017	18	20	167	86k	4,839	10,929
MUC-4	4	5	1,700	495k	21,928	2,641
WikiEvents	50	59	246	190k	8,544	5,536
RAMS	139	65	9,124	957k	34,536	21237
DocEE(en)	59	356	27,485	16,268k	749,568	180,528

表 6.3 中文 DocEE 和现有的中文事件抽取数据集对比

数据集	事件类型	事件要素角色	篇章数	词汇数	句子数	事件要素实例数
CEC	5	5	332	106k	1,385	5,954
DuEE	65	121	-	710k	19,640	41,520
DocEE(zh)	59	356	35,250	30,230k	130,000	218,550

中有 18.1 个事件要素，“政变”事件上标注的事件要素最少，平均每篇文章中有 3.8 个事件要素。中文 DocEE 总共标记了 35,250 篇新闻文章，218,550 个事件要素，平均每篇新闻有 6.2 个事件要素。其中，“空难”事件上标注的事件要素最多，平均每篇文章中有 9.8 个事件要素，“经济危机”事件上标注的事件要素最少，平均每篇文章中有 4.2 个事件要素。

为了对 DocEE 数据集的数据规模有更直观的认知，我们将其和现有的事件抽取数据集进行对比。表 6.2 和表 6.3 展示了对比结果。可以看到，无论是英文 DocEE 还是中文 DocEE 都比现有的事件抽取数据集具有更大的数据规模，这主要体现在标注的新闻篇章多以及标注的事件要素实例多。

以英文 DocEE 为例，与公开语料库 ACE2005、KBP、MUC-4、Wikievents 和 RAMS 相比，我们构建的数据集包含的新闻篇章更多，达到了 27,000 多篇，事件要素标注也更多，达到了 18 万个。同时，我们定义的事件要素类型也远超现有的数据集：我们为 59 个事件类定义了 356 个事件要素类型，精细化的事件要素类型设定能够全面的捕获用户想要了解的事件信息，提升 DocEE 应用时的用户体验。

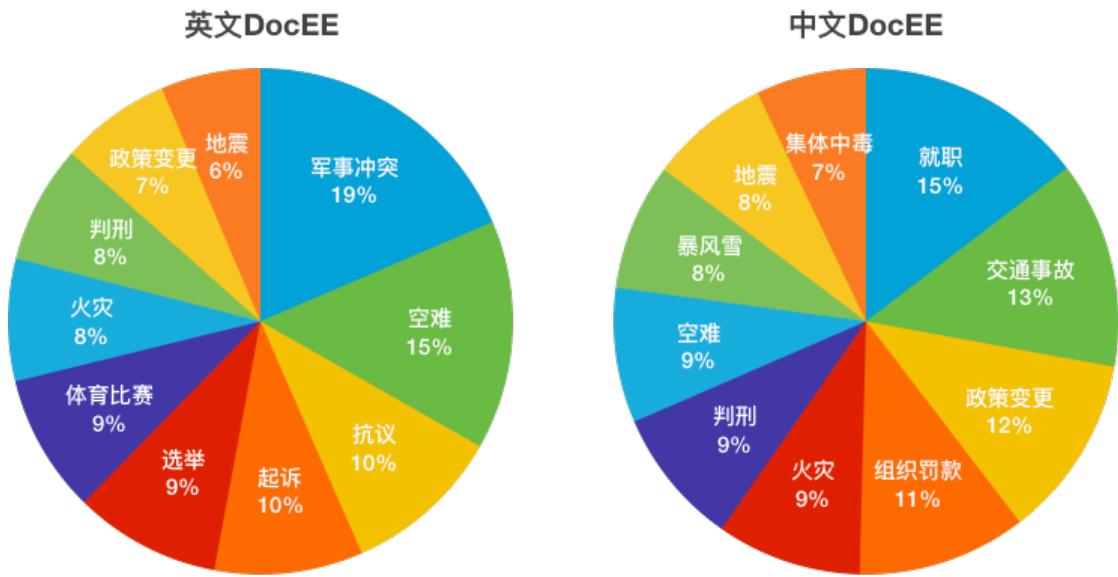


图 6.8 DocEE 中实例数 top10 的事件类型分布

6.3.2 事件类型统计分析

图 6.8 中显示了 DocEE 中实例最多的前 10 个事件类型的统计数据。可以看到，英文 DocEE 中标注实例较多的事件主要有“军事冲突”、“空难”、“抗议”、“起诉”、“选举”和“体育比赛”。而中文 DocEE 中标注实例较多的事件主要有“就职”、“交通事故”、“政策变更”和“组织罚款”，这也一定程度上体现了国情的不同。从图 6.8 还可以看出，DocEE 数据遵循长尾分布，这是由于各类事件在真实新闻中分布不均导致的。据统计，英文 DocEE 中标注实例数量大于 500 的事件类型占比 30.5%，标注实例数量大于 200 的事件类型占比 83.1%。中文 DocEE 中标注实例数量大于 500 的事件类型占比 42.3%，标注实例数量大于 200 的事件类型占比 93%。

6.3.3 事件要素统计分析

我们从中文 DocEE 中随机抽取了 100 篇文章来进行事件要素统计分析，这 100 篇文章中总共包含 831 个事件要素实例。我们首先根据事件要素提及的个数对事件要素实例进行分类。如表 6.4 所示，67% 的事件要素有一个提及，33% 的事件要素有多个提及，需要从文中的多句话中查找该事件要素的答案。多提及的事件要素对模型的召回能力提出了很大的挑战。

此外，我们根据事件要素提及的长度对事件要素进行统计分析。在这 831 个事件要素实例中，42% 的事件要素提及的长度不超过 3 个单词，这类事件要素角色主要有“时间”、“地点”等。30% 的事件要素提及的长度在 3 个词和 10 个词之间，这类事件要素角色主要有“伤亡情况”、“经济损失”等。28% 的事件要素提及的单词长度超过 10 个词，这类事件要素角色主要有“原因”、“调查结果”等。

表 6.4 中文 DocEE 中事件要素提及分布

事件要素提及	比例	例子
单提及 事件要素	67%	下午五时，身穿黑色连帽衫的蒙面男子拿了把自动步枪抢劫了银行，并从向东边的沃伦街逃跑 事件类型：银行抢劫 事件要素角色：时间
多提及 事件要素	33%	一架比奇 1900D 客机撞向飞机维修库，全机 21 人罹难。失事的原因可能是燃料缸盖子的零件的损坏。... 另一个可能原因是可能飞进了早前在同一跑道起飞的一架庞巴迪 CRJ 的尾波乱流导致失事。 事件类型：空难 事件要素角色：原因

6.4 实验

我们在构建的中文 DocEE 数据集上展开实验。为了模拟低资源场景，我们从 DocEE 中选取了 5 个类，每个类挑选 10 条数据作为训练集，剩下的数据按照 1:1 分为验证集和测试集。数据集划分情况如表 6.5 中所示。

表 6.5 事件抽取中文实验设定

事件类型	训练集	验证集	测试集
空难	10	112	110
暴风雪	10	102	100
船难	10	98	100
颁奖典礼	10	98	110
集体中毒	10	101	108

6.4.1 评价指标

针对事件检测任务，我们使用准确率（Precision）、召回率（Recall）和 F 值（Micro-F1）作为评测指标。针对事件要素抽取任务，我们使用精准匹配（Exact Match, EM）和模糊匹配（Head Noun Match Match, HM)^[137]作为评测指标。精准匹配需要预测答案和标注答案完全一致。模糊匹配更加的宽松，只需要预测答案和标注答案中根名词（head noun）一致就可以。模糊匹配可以将核心意思正确的预测答案也判为正确，这样可以减少预测答案由于多标介词或形容词导致整个判错的情况。

表6.6 所提方法在中文事件检测上的整体表现

Methods	P	R	F
Rich-C	54.1	49.0	51.4
CMS	71.4	56.3	63.0
C-BiLSTM	76.6	73.0	74.8
BCNN	79.7	74.2	76.9
HNN	80.3	76.8	78.5
CAEE	88.1	90.9	89.5
DRMM (多模态方法)	92.3	93.4	92.8
EKD (知识注入方法)	92.2	95.5	93.8
MUCO (自标注方法)	94.5	96.4	95.4

6.4.2 基线系统

我们将前三章提出的算法和六个常用中文事件抽取方法进行比较：

- Rich-C^[30]利用字符和篇章中丰富特征的中文事件抽取模型。
- CMS^[152]提出了融合中文词语的形态结构和义原信息的中文事件抽取模型。
- C-BiLSTM^[153]在字符级别同时考虑 n-gram 信息和序列信息。
- BCNN^[154]提出了一个基于注意力和语义特征融合的中文事件抽取模型。
- HNN^[155]结合双向循环神经网络和卷积神经网络来学习事件抽取任务的语言无关特征。
- CAEE^[156]利用注意力增强的图卷积网络融合句法结构信息，增强预训练 BERT 模型在中文事件抽取的性能。

在实验中，我们将本文中提出的三个低资源事件抽取方法同时应用在中文事件检测和中文事件要素抽取任务上。特别的，我们将本文中提出的三个低资源事件抽取方法的底层编码器由 BERT 换为 BERT-Chinese，以适应中文事件抽取的场景。在做事件检测任务时，我们将第四章所提方法用到的 WordNet 从英文版本（普林斯顿大学研发）切换为 Open Multilingual Wordnet 版本，以适应中文事件检测。在做事件要素抽取任务时，我们将第四章所提方法使用的开放域触发词知识转变为开放域命名实体知识，以更好的匹配事件要素抽取任务。

6.4.3 整体实验效果分析

从表6.6和表6.7中可以看到，我们的模型超过了六个基线模型，取得了更好的性能表现，这证明了本文所提的三个低资源事件抽取算法在中文事件检测和中文

表6.7 所提方法在中文事件要素抽取上的整体表现

Methods	EM			HM		
	P	R	F	P	R	F
Rich-C	12.3	10.4	11.3	17.3	14.5	15.8
CMS	13.2	11.6	12.3	18.2	13.8	15.7
C-BiLSTM	17.8	12.4	14.6	19.3	12.5	15.2
BCNN	23.4	16.2	19.1	25.3	16.6	20.0
HNN	24.1	17.2	20.1	26.2	18.6	28.0
CAEE	48.4	42.1	45.0	50.3	42.7	46.2
DRMM（多模态方法）	48.9	46.7	47.7	53.1	47.2	48.6
EKD（知识注入方法）	47.6	47.3	47.4	52.8	48.5	50.6
MUCO（自标注方法）	48.3	48.5	48.5	57.3	49.2	52.9

事件要素抽取任务上的有效性。

相比于仅利用文本模态信息的基线模型 Rich-C, CMS, C-BiLSTM, BCNN, HNN 和 CAEE, 我们的多模态增强的事件抽取方法 DRMM 能够额外的利用新闻文档中的图片模态信息, 为缓解过拟合提供了更多视觉证据。相比于纯数据驱动的基线模型 Rich-C, C-BiLSTM, HNN 和 CAEE, 我们的知识蒸馏模型 EKD 能够利用外部的常识信息, 提升了事件要素识别的鲁棒性和泛化性。相比于仅能利用标注数据的基线模型 Rich-C, C-BiLSTM, HNN 和 CAEE, 我们提出的自标注事件抽取方法 MUCO 能够自动的挖掘更多的任务相关类的标注, 为模型提供更多的监督信息, 从而缓解模型在低资源场景下容易过拟合的问题。

6.5 应用成果展示

我们在本节中展示事件抽取在 NewsMiner 系统中的应用效果。

The screenshot displays the NewsMiner application interface. At the top, there is a news headline: "巴基斯坦南部地震造成至少15人死亡". Below the headline, there are several small text snippets and labels indicating the source and type of information. On the right side, there is a detailed "事件信息" (Event Information) panel. This panel has a header "事件信息" with an icon. Below it, there is a summary sentence: "巴基斯坦南部地震造成至少15人死亡". Underneath this summary, there is a table with two columns: "Type" and "Value". The table contains the following data:

Type	Value
受影响地区	巴基斯坦南部、巴基斯坦
人员伤亡	15人死亡
日期	10月7日、7日凌晨、10月07日06时01分
震级	5.8级

图 6.9 NewsMiner 应用效果

如图6.9所示，“事件信息”栏展示了我们将事件抽取应用到 NewsMiner 中的效果。系统用户通过阅读事件抽取得到的结构化信息，就能够快速了解新闻中报道的事件的大致内容，极大的提升系统用户的阅读效率。相比于文本摘要，事件抽取的结果更有条理，对信息的归纳程度更高，更容易帮助人们记住关键信息，有自己独特的优势。相比于系统中已有的关键词识别算法，事件抽取的优势在于能够展示更重要的信息。我们是从事件角度出发挖掘新闻文本语义的，这样得到的每一个实体或者文本描述都一定是和核心事件有很强的语义相关性，比如我们会展示“10月7号”，是因为它是地震发生的时间，我们会展示“5.8”这个数字，是因为它代表着地震的等级。但是关键词识别算法只是尽可能的罗列文本中出现的所有人名、地点或者机构，即使他们和新闻核心内容并不相关，这导致许多不重要的信息都被展示了出来，比如“中国地震台网”，降低了人们在使用系统时的用户体验。

6.6 本章小结

在本章中，我们主要介绍了事件抽取研究在新闻挖掘系统 NewsMiner 中的应用情况。我们首先简要概述了 NewsMiner 系统现状，指出了现有的语义分析工具关键词算法的不足。之后，针对现有事件本体不能反映用户感兴趣事件的问题，我们依据新闻学的软硬新闻理论，重新构建了事件本体，并基于此众包构建了大规模的中英文事件抽取数据，作为事件抽取应用到 NewsMiner 系统中的数据基础。最后，我们将前三章提到的研究工作应用到 NewsMiner 系统中的低资源事件抽取场景中，实现研究的落地。我们也展示了事件抽取最终在 NewsMiner 系统中的应用效果，分析了事件抽取对提升 NewsMiner 系统用户体验带来的贡献。相关工作 DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction^[157] 发表在 NAACL2022 上。

第 7 章 结论和未来工作

7.1 工作总结

事件抽取任务旨在从非结构化的文本中识别事件触发词并抽取参与事件的要素，例如时间，地点，人物等。事件抽取得到的结构化信息可以用于构建事件知识图谱，梳理事件因果关系以及支撑语义检索，因而近些年来吸引了很多研究者的兴趣。近些年来，深度学习模型在事件抽取上取得了不俗的成果。但是深度学习模型需要大规模的标注数据驱动，而现实场景中，获取大规模的标注数据并不是一件很容易的事情。因此，本文面向低资源事件抽取关键技术开展研究。为了缓解低资源场景中模型过拟合的现象，本文从三个思路上开展研究：

- 本文提出了一个基于多模态融合的事件检测方法 D，旨在利用新闻中图像模态的信息，解决低资源场景下事件类型消歧难的问题。所提方法利用交替对偶注意力机制逐步将新闻中多张配图信息融入到事件检测模型，通过文本和图像双向筛选的方式，显著的提升了融入图像的质量。
- 本文提出了一个基于知识蒸馏的事件检测方法，旨在利用开放域触发词知识，为事件识别提供先验，解决低资源场景下事件召回难的问题。所提方法不局限于在标注语料中学习，而是能够同时在大规模未标注语料上学习，有很强的知识利用能力。
- 本文提出了一个基于自标注的事件要素抽取方法，旨在没有人工标注的情况下，挖掘更多训练数据，解决低资源场景下事件要素识别难的问题。所提方法借助预定义类的监督信号，可以在没有元数据的情况下从语料中分辨噪音和任务相关类，通过标注任务相关类提升了训练数据的多样性，降低了训练语料单一稀疏带来的事件要素识别生硬刻板的问题。

最后，我们在新闻挖掘系统 NewsMiner 上展开示范应用。针对现有的事件本体不符合实际应用需求的问题，我们基于新闻学理论重新构建了事件本体，关注用户真正感兴趣的事件，例如地震，空难，军事冲突等，并在此基础上众包标注了中英双语事件抽取数据集 DocEE。基于此数据集，我们将前三章的研究成果应用在 NewsMiner 系统上，实现研究的落地。相比于 NewsMiner 系统中已有的关键词功能，我们提供的事件抽取功能更有条理，对信息的归纳程度更高，更容易帮助人们记住新闻中的关键信息，极大的提升了用户体验。

7.2 未来展望

在本节中，我们探讨事件抽取未来的发展方向。我们主要从面向弱监督的事件抽取研究，面向常识约束的事件抽取研究和面向跨篇章的事件抽取研究这三个方面展开描述。

1、面向弱监督的事件抽取研究。现有的深度学习事件抽取模型非常依赖监督数据，模型要想应用到新的事件类型上还需要大量的标注数据。因此，研究如何在弱监督场景下提升事件抽取模型的性能是未来一个重要的方向。自监督学习是一个可能的思路。自监督学习能够自动从大规模无标注语料中构建监督信号，让模型“自我学习”，通过“同一事件提及的表示更加接近，不同事件提及的表示更加远离”的预训练过程，让事件抽取模型在微调训练（fine-tuning）之前就拥有更好的文本表示，从而提升模型在少量标注的低资源情境下快速拟合的能力。另一个可能的思路是零样本学习。零样本学习在应用到新的事件类型时只需要给出类的元数据描述（类名，类语义解释）即可，无须费时费力的标注大量数据，极大的减轻了领域迁移的工作量。第三个可能的解决思路是及时调整训练方式（prompt-tuning），这种思路通过设置前置网络的方式，减少了训练中需要优化的参数量，提升了深度学习模型在少量的样本下的学习能力。

2、面向常识约束的事件抽取研究。虽然深度学习技术在事件抽取特征学习上展现出来了强大的表征能力，但这些神经网络模型通常具有过多超参数和网络配置，如果超参数设置不当，会极大的损害模型的性能。另一方面，知识图谱中有很多高质量的人类知识。一个未来方向可能是如何在外部知识图谱指导下提升深度学习模型的鲁棒性。虽然目前已有一些方法尝试加入外部知识的约束，但是他们大都集中在深度学习模型 + 注意力机制的框架下，并没有将二者更深层次的融合。未来可以尝试通过融入整数逻辑规划或者符号学习的方式，让深度学习模型能够更好利用常识信息。

3、面向跨篇章的事件抽取研究。目前的事件抽取多局限在句子级和篇章级，也就是在一句话或者一篇文章的语义范围内抽取事件，而忽略跨篇章事件抽取的研究。实际上，推动跨篇章事件抽取是很有必要的。新闻报道有及时性、碎片化的特点，而事件有持续性的特点，读者往往需要查看多天的新闻报道，才能够获知事件的全貌。比如“特朗普二次弹劾事件”，从刚开始的“议员提出弹劾”，到“司法取证”，到“参议院众议院投票表决”，再到最后“判决弹劾无效”跨越了近一个月的多篇新闻报道。只有从多个篇章中抽取事件信息，才能够获知事件的全貌。因此，为了推动事件抽取系统的真正落地，面向跨篇章的事件抽取工作是未来工作中不可或缺的一步。面向跨篇章的事件抽取工作的未来研究方向可能会集中在如下三个方面：

面：（1）构建大规模高质量的跨篇章的事件抽取数据集。现有的数据集多停留在事件类型聚类的层面上，缺乏跨篇章的事件要素标注，未来需要更高质量的数据为跨篇章的事件抽取研究提供扎实的数据基础。（2）定义更具普适意义的事件本体，以应对现实世界中事件描述灵活多变的挑战。（3）提出具有更优秀的长文本理解能力的跨篇章事件抽取方法。跨篇章事件抽取意味着同一件事的事件要素出现的位置更分散，而无论是卷积神经网络还是预训练模型 BERT，都面临着长距离灾难性遗忘的问题，这导致现有的事件抽取模型的长距离性能不佳，因此，未来学者需要研究更优秀的长文本处理算法来应对这个困境。

参考文献

- [1] Kim J. Philosophy of mind[M]. Routledge, 2018.
- [2] Peng Y. The event-domain cognitive model perspective on terminology: A case study of atmospheric environment terms[J]. Lexicography, 2019: 43-67.
- [3] Pluth E. Badiou: A philosophy of the new[M]. Polity, 2010.
- [4] Whitehead A N. Process and reality: An essay in cosmology[M]. 1929.
- [5] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 167-176.
- [6] Duan S, He R, Zhao W. Exploiting document level information to improve event detection via recurrent neural networks[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017: 352-361.
- [7] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//Proceddings of the 31th Annual Conference on Neural Information Processing Systems. 2017: 4077-4087.
- [8] Nooralahzadeh F, Bekoulis G, Bjerva J, et al. Zero-shot cross-lingual transfer with meta learning [J]. arXiv preprint arXiv:2003.02739, 2020.
- [9] Yang Y, Katiyar A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics, 2020: 6365-6375. <https://aclanthology.org/2020.emnlp-main.516>. DOI: 10.18653/v1/2020.emnlp-main.516.
- [10] Tong M, Xu B, Wang S, et al. Improving low-resource chinese event detection with multi-task learning[C]//Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management. 2020: 421-433.
- [11] Carbonell J G, Michalski R S, Mitchell T M. An overview of machine learning[J]. Machine learning, 1983: 3-23.
- [12] Jhuang H, Gall J, Zuffi S, et al. Towards understanding action recognition[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision. 2013: 3192-3199.
- [13] Mesaros A, Heittola T, Virtanen T, et al. Sound event detection: A tutorial[J]. IEEE Signal Processing Magazine, 2021: 67-83.
- [14] Li Z, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction [J]. arXiv preprint arXiv:1805.05081, 2018.
- [15] Ahn D. The stages of event extraction[C]//Proceedings of the Workshop on Annotating and Reasoning about Time and Events. 2006: 1-8.
- [16] Wattarujeekrit T, Shah P K, Collier N. Pasbio: predicate-argument structures for event extraction in molecular biology[J]. BMC bioinformatics, 2004: 1-20.

- [17] Kim J D, Ohta T, Pyysalo S, et al. Overview of bionlp'09 shared task on event extraction[C]// Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. 2009: 1-9.
- [18] Zheng S, Cao W, Xu W, et al. Doc2edag: An end-to-end document-level framework for chinese financial event extraction[J]. arXiv preprint arXiv:1904.07535, 2019.
- [19] 余杰, 纪斌, 刘磊, 等. 面向中文医疗事件的联合抽取方法[J]. 计算机科学, 2021: 287-293.
- [20] Riloff E, et al. Automatically constructing a dictionary for information extraction tasks[C]// Proceedings of the 11th National Conference on Artificial Intelligence. Citeseer, 1993: 2-1.
- [21] Kilicoglu H, Bergler S. Syntactic dependency based heuristics for biological event extraction [C]//Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. 2009: 119-127.
- [22] Buyko E, Faessler E, Wermter J, et al. Event extraction from trimmed dependency graphs[C]// Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. 2009: 19-27.
- [23] Yangarber R, Grishman R, Tapanainen P, et al. Automatic acquisition of domain knowledge for information extraction[C]//Proceedings of the 18th International Conference on Computational Linguistics. 2000.
- [24] Borsje J, Hogenboom F, Frasincar F. Semi-automatic financial events discovery based on lexico-semantic patterns[J]. Journal of Web Engineering and Technology, 2010: 115-140.
- [25] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 73-82.
- [26] Björne J, Salakoski T. Generalizing biomedical event extraction[C]//Proceedings of BioNLP Shared Task 2011 Workshop. 2011: 183-191.
- [27] Huang R, Riloff E. Modeling textual cohesion for event extraction[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. 2012: 1664-1670.
- [28] Henn S, Sticha A, Burley T, et al. Visualization techniques to enhance automated event extraction[J]. arXiv preprint arXiv:2106.06588, 2021.
- [29] Lu W, Roth D. Automatic event extraction with structured preference modeling[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012: 835-844.
- [30] Chen C, Ng V. Joint modeling for chinese event extraction with rich linguistic features[C]// Proceedings of the 24th International Conference on Computational Linguistics. 2012: 529-544.
- [31] Li P, Zhu Q, Diao H, et al. Joint modeling of trigger identification and event type determination in chinese event extraction[C]//Proceedings of the 24th International Conference on Computational Linguistics. 2012: 1635-1652.
- [32] Björne J, Ginter F, Pyysalo S, et al. Complex event extraction at pubmed scale[J]. Bioinformatics, 2010: i382-i390.
- [33] Miwa M, Sætre R, Kim J D, et al. Event extraction with complex event classification using rich features[J]. Journal of bioinformatics and computational biology, 2010: 131-146.
- [34] Ananiadou S, Pyysalo S, Tsujii J, et al. Event extraction for systems biology by text mining the literature[J]. Trends in biotechnology, 2010: 381-390.

-
- [35] Li Z, Liu F, Antieau L, et al. Lancet: a high precision medication event extraction system for clinical text[J]. Journal of the American Medical Informatics Association, 2010: 563-567.
 - [36] Sakaki T, Matsuo Y, Yanagihara T, et al. Real-time event extraction for driving information from social sensors[C]//Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems. 2012: 221-226.
 - [37] Liao S, Grishman R. Using document level cross-event inference to improve event extraction [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 789-797.
 - [38] Ji H, Grishman R. Refining event extraction through cross-document inference[C]//Proceedings of ACL-08: Hlt. 2008: 254-262.
 - [39] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 34th International Conference on Machine Learning. 2014: 1188-1196.
 - [40] Zhao Y, Jin X, Wang Y, et al. Document embedding enhanced event detection with hierarchical and supervised attention[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 414-419.
 - [41] Chen Y, Yang H, Liu K, et al. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1267-1276.
 - [42] Liu S, Cheng R, Yu X, et al. Exploiting contextual information via dynamic memory network for event detection[J]. arXiv preprint arXiv:1810.03449, 2018.
 - [43] Veyseh A P B, Van Nguyen M, Trung N N, et al. Modeling document-level context for event detection via important context selection[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 5403-5413.
 - [44] Sha L, Qian F, Chang B, et al. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.
 - [45] Orr W, Tadepalli P, Fern X. Event detection with neural networks: A rigorous empirical evaluation[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 999-1004. <https://aclanthology.org/D18-1122>. DOI: 10.18653/v1/D18-1122.
 - [46] Nguyen T, Grishman R. Graph convolutional networks with argument-aware pooling for event detection[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.
 - [47] Liu X, Luo Z, Huang H. Jointly multiple events extraction via attention-based graph information aggregation[J]. arXiv preprint arXiv:1809.09078, 2018.
 - [48] Yan H, Jin X, Meng X, et al. Event detection with multi-order graph convolution and aggregated attention[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2019: 5766-5770. <https://aclanthology.org/D19-1582>. DOI: 10.18653/v1/D19-1582.

- [49] Cui S, Yu B, Liu T, et al. Edge-enhanced graph convolution networks for event detection with syntactic relation[J]. arXiv preprint arXiv:2002.10757, 2020.
- [50] Lai V D, Nguyen T N, Nguyen T H. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks[J]. arXiv preprint arXiv:2010.14123, 2020.
- [51] Veyseh A P B, Nguyen T N, Nguyen T H. Graph transformer networks with syntactic and semantic structures for event argument extraction[J]. arXiv preprint arXiv:2010.13391, 2020.
- [52] Balali A, Asadpour M, Campos R, et al. Joint event extraction along shortest dependency paths using graph convolutional networks[J]. Knowledge-Based Systems, 2020: 106492.
- [53] Yang M, Wang Y, Chen F. Mbced: Multi-task learning event detection method based on pre-training[C/OL]//Proceedings of the 4th International Conference on Algorithms, Computing and Artificial Intelligence. Association for Computing Machinery, 2021: 6. <https://doi.org/10.1145/3508546.3508605>.
- [54] Lu W, Nguyen T H. Similar but not the same: Word sense disambiguation improves event detection via neural representation matching[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 4822-4828. <https://aclanthology.org/D18-1517>. DOI: 10.18653/v1/D18-1517.
- [55] Yang B, Mitchell T. Joint extraction of events and entities within a document context[J]. arXiv preprint arXiv:1609.03632, 2016.
- [56] Zhang T, Ji H, Sil A. Joint entity and event extraction with generative adversarial imitation learning[J]. Data Intelligence, 2019: 99-120.
- [57] Xiang Y, Li C. A trigger-aware multi-task learning for chinese event entity recognition[C]// Proceedings of the 20th China National Conference on Chinese Computational Linguistics. 2021: 341-354.
- [58] Nguyen T M, Nguyen T H. One for all: Neural joint modeling of entities and events[C]// Proceedings of the 2019 AAAI Conference on Artificial Intelligence. 2019: 6851-6858.
- [59] Liu J, Zhao S, Wang G. Ssel-ade: a semi-supervised ensemble learning framework for extracting adverse drug events from social media[J]. Artificial intelligence in medicine, 2018: 34-49.
- [60] Zhou D, Zhong D. A semi-supervised learning framework for biomedical event extraction based on hidden topics[J]. Artificial intelligence in medicine, 2015: 51-58.
- [61] Ferguson J, Lockard C, Weld D S, et al. Semi-supervised event extraction with paraphrase clusters[J]. arXiv preprint arXiv:1808.08622, 2018.
- [62] Ferguson J, Lockard C, Weld D, et al. Semi-supervised event extraction with paraphrase clusters [C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2018: 359-364. <https://www.aclweb.org/anthology/N18-2058>. DOI: 10.18653/v1/N18-2058.
- [63] Zhang C, Soderland S, Weld D S. Exploiting parallel news streams for unsupervised event extraction[J]. Transactions of the Association for Computational Linguistics, 2015: 117-129.
- [64] Huang R, Riloff E. Bootstrapped training of event extraction classifiers[C/OL]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 286-295. <https://www.aclweb.org/anthology/E12-1029>.

- [65] Liao S, Grishman R. Can document selection help semi-supervised learning? a case study on event extraction[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 260-265.
- [66] Wang X, Han X, Liu Z, et al. Adversarial training for weakly supervised event detection[C/OL]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 998-1008. <https://aclanthology.org/N19-1105>. DOI: 10.18653/v1/N19-1105.
- [67] Liu S, Chen Y, He S, et al. Leveraging framenet to improve automatic event detection[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 2134-2143.
- [68] Zuo X, Chen Y, Liu K, et al. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision[J]. arXiv preprint arXiv:2010.10833, 2020.
- [69] Alrashdi R, O'Keefe S. Automatic labeling of tweets for crisis response using distant supervision[C]//Proceedings of the Web Conference 2020. 2020: 418-425.
- [70] Reschke K, Jankowiak M, Surdeanu M, et al. Event extraction using distant supervision[C]// Proceedings of the 9th International Conference on Language Resources and Evaluation. 2014: 4527-4531.
- [71] Araki J, Mitamura T. Open-domain event detection using distant supervision[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 878-891.
- [72] Yang H, Chen Y, Liu K, et al. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 50-55.
- [73] Du X, Cardie C. Event extraction by answering (almost) natural questions[J]. arXiv preprint arXiv:2004.13625, 2020.
- [74] Liu J, Chen Y, Liu K, et al. Event extraction as machine reading comprehension[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 1641-1651.
- [75] Chen Y, Chen T, Ebner S, et al. Reading the manual: Event extraction as definition comprehension[C/OL]//Proceedings of the 4th Workshop on Structured Prediction for NLP. Association for Computational Linguistics, 2020: 74-83. <https://aclanthology.org/2020.spnlp-1.9>. DOI: 10.18653/v1/2020.spnlp-1.9.
- [76] Wang S, Yu M, Chang S, et al. Query and extract: Refining event extraction as type-oriented binary decoding[J]. arXiv preprint arXiv:2110.07476, 2021.
- [77] Boros E, Moreno J G, Doucet A. Event detection as question answering with entity information [J/OL]. CoRR, 2021. <https://arxiv.org/abs/2104.06969>.
- [78] Hsu I, Huang K H, Boschee E, et al. Degree: A data-efficient generative event extraction model [J]. arXiv preprint arXiv:2108.12724, 2021.
- [79] Lu Y, Lin H, Xu J, et al. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction[J]. arXiv preprint arXiv:2106.09232, 2021.

- [80] Du X, Rush A M, Cardie C. Template filling with generative transformers[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021: 909-914.
- [81] Liu S, Li Y, Zhang F, et al. Event detection without triggers[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 735-744.
- [82] Liu J, Chen Y, Liu K, et al. Event detection via gated multilingual attention mechanism[J]. Statistics, 2018: 1250.
- [83] Lu W, Nguyen T H. Similar but not the same: Word sense disambiguation improves event detection via neural representation matching[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4822-4828.
- [84] Zhang K, Lv G, Wu L, et al. Image-enhanced multi-level sentence representation net for natural language inference[C]//Proceedings of the 2018 IEEE International Conference on Data Mining. IEEE, 2018: 747-756.
- [85] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts[J]. arXiv preprint arXiv:1802.07862, 2018.
- [86] Elliott D, Frank S, Hasler E. Multi-language image description with neural sequence models [J]. CoRR, abs/1510.04709, 2015.
- [87] Varni G, Castagné N. Multimodal (multisensory) integration, in technology[M]. Enactive Systems Books, 2007.
- [88] Noda K, Arie H, Suga Y, et al. Multimodal integration learning of robot behavior using deep neural networks[J]. Robotics and Autonomous Systems, 2014: 721-736.
- [89] Heo Y, Kang S, Yoo D. Multimodal neural machine translation with weakly labeled images[J]. IEEE Access, 2019.
- [90] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.
- [91] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint arXiv:1806.00064, 2018.
- [92] Hou M, Tang J, Zhang J, et al. Deep multimodal multilinear fusion with high-order polynomial pooling[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [93] Pérez-Rúa J M, Vielzeuf V, Pateux S, et al. Mfas: Multimodal fusion architecture search[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6966-6975.
- [94] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.
- [95] Zhang T, Whitehead S, Zhang H, et al. Improving event extraction via multimodal integration [C]//Proceedings of the 25th ACM International Conference on Multimedia. ACM, 2017: 270-278.

- [96] Banarescu L, Bonial C, Cai S, et al. Abstract meaning representation for sembanking[C]// Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. 2013: 178-186.
- [97] 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.
- [98] Jung H, Choi M K, Jung J, et al. Resnet-based vehicle classification and localization in traffic surveillance systems[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 61-67.
- [99] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [100] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.
- [101] Wang X, Han X, Liu Z, et al. Adversarial training for weakly supervised event detection[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 998-1008.
- [102] Wang L, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5005-5013.
- [103] Qian C, Zhu X, Ling Z H, et al. Neural natural language inference models enhanced with external knowledge[J]. arXiv preprint arXiv:1711.04289, 2017.
- [104] Tong M, Wang S, Cao Y, et al. Image enhanced event detection in news articles[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020: 9040-9047.
- [105] Yang S, Feng D, Qiao L, et al. Exploring pre-trained language models for event extraction and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5284-5294.
- [106] Cao Y, Hu Z, Chua T s, et al. Low-resource name tagging learned with weakly labeled data[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.
- [107] Zhong Z, Ng H T. It makes sense: A wide-coverage word sense disambiguation system for free text[C/OL]//Proceedings of the ACL 2010 System Demonstrations. Association for Computational Linguistics, 2010: 78-83. <https://www.aclweb.org/anthology/P10-4014>.
- [108] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to wordnet: An on-line lexical database [J]. International journal of lexicography, 1990: 235-244.
- [109] Manning C, Surdeanu M, Bauer J, et al. The stanford corenlp natural language processing toolkit[C]//Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014: 55-60.
- [110] Soares L B, Fitzgerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning[J]. arXiv preprint arXiv:1906.03158, 2019.
- [111] Harris Z S. Distributional structure[J]. <i>WORD</i>, 1954: 146-162.

- [112] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C/OL]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2015: 167-176. <https://aclanthology.org/P15-1017>. DOI: 10.3115/v1/P15-1017.
- [113] Liu S, Chen Y, Liu K, et al. Exploiting argument information to improve event detection via supervised attention mechanisms[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1789-1798.
- [114] Lu Y, Lin H, Han X, et al. Distilling discrimination and generalization knowledge for event detection via delta-representation learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4366-4376.
- [115] Liu J, Chen Y, Liu K. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection[C]//Proceedings of the 2019 AAAI Conference on Artificial Intelligence. 2019: 6754-6761.
- [116] Chen Y, Liu S, Zhang X, et al. Automatically labeled data generation for large scale event extraction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 409-419.
- [117] He H, Sun X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017.
- [118] Plank B, Moschitti A. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: volume 1. 2013: 1498-1507.
- [119] Nguyen T H, Grishman R. Employing word representations and regularization for domain adaptation of relation extraction[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 68-74.
- [120] Nguyen T H, Grishman R. Event detection and domain adaptation with convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language. 2015: 365-371.
- [121] Zhang K, Zi J, Wu L G. New event detection based on indexing-tree and named entity[C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2007: 215-222.
- [122] McClosky D, Surdeanu M, Manning C D. Event extraction as dependency parsing[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2011: 1626-1635.
- [123] Boroş T, Dumitrescu S D, Burtica R. NLP-cube: End-to-end raw text processing with neural networks[C]//Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. 2018: 171-179.
- [124] Wang C, Xue N, Pradhan S. A transition-based algorithm for amr parsing[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics. 2015: 366-375.

- [125] Huang L, Ji H, Cho K, et al. Zero-shot transfer learning for event extraction[J]. arXiv preprint arXiv:1707.01066, 2017.
- [126] Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction[J]. Science, 2015: 1332-1338.
- [127] Li X, Sun X, Meng Y, et al. Dice loss for data-imbalanced nlp tasks[J]. arXiv preprint arXiv:1911.02855, 2019.
- [128] Katz J J, Fodor J A. The structure of a semantic theory[J]. Language, 1963: 170-210.
- [129] Ghosh S, Maitra P, Das D. Feature based approach to named entity recognition and linking for tweets[C]//Proceedings of the 6th workshop on Making Sense of Microposts. 2016.
- [130] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv preprint arXiv:1706.05075, 2017.
- [131] Pushp P K, Srivastava M M. Train once, test anywhere: Zero-shot learning for text classification [J]. arXiv preprint arXiv:1712.05972, 2017.
- [132] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition [C]//Proceedings of the 32nd International Conference on Machine Learning: volume 2. Lille, 2015.
- [133] Qi H, Brown M, Lowe D G. Low-shot learning with imprinted weights[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 5822-5830.
- [134] Geng R, Li B, Li Y, et al. Few-shot text classification with induction network[J]. arXiv preprint arXiv:1902.10482, 2019.
- [135] Wang F, Xiang X, Cheng J, et al. Normface: L2 hypersphere embedding for face verification [C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 1041-1049.
- [136] Fritzler A, Logacheva V, Kretov M. Few-shot classification in named entity recognition task [C]//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 2019: 993-1000.
- [137] Du X, Cardie C. Document-level event role filler extraction using multi-granularity contextualized encoding[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 8010-8020. <https://aclanthology.org/2020.acl-main.714>. DOI: 10.18653/v1/2020.acl-main.714.
- [138] Hou Y, Che W, Lai Y, et al. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network[J]. arXiv preprint arXiv:2006.05702, 2020.
- [139] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017: 1126-1135.
- [140] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: Analysis and implementation[J]. Proceedings of the IEEE transactions on Pattern Analysis and Machine Intelligence, 2002: 881-892.
- [141] Ohta T, Pyysalo S, Tsujii J, et al. Open-domain anatomical entity mention detection[C]// Proceedings of the workshop on Detecting Structure in Scholarly Discourse. 2012: 27-36.

- [142] Coucke A, Saade A, Ball A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[J]. arXiv preprint arXiv:1805.10190, 2018.
- [143] Tong M, Wang S, Xu B, et al. Learning from miscellaneous other-class words for few-shot named entity recognition[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021.
- [144] Baker C F. FrameNet: A knowledge base for natural language processing[C/OL]//Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014). 2014: 1-5. <https://aclanthology.org/W14-3001>. DOI: 10.3115/v1/W14-3001.
- [145] Dong Z, Dong Q. Hownet - a hybrid language and knowledge resource[C/OL]//International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003. 2003: 820-824. DOI: 10.1109/NLPKE.2003.1276017.
- [146] Reinemann C, Stanyer J, Scherr S, et al. Hard and soft news: A review of concepts, operationalizations and key findings[J/OL]. Journalism, 2012: 221-239. <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-15496-9>.
- [147] Tuchman G. Making news by doing work: Routinizing the unexpected[J]. American journal of Sociology, 1973: 110-131.
- [148] Lehman-Wilzig S N, Seletzky M. Hard news, soft news, ‘general’news: The necessity and utility of an intermediate classification[J/OL]. Journalism, 2010: 37-56. <https://doi.org/10.1177/1464884909350642>.
- [149] Hienert D, Luciano F. Extraction of historical events from wikipedia[J/OL]. CoRR, 2012. <http://arxiv.org/abs/1205.4138>.
- [150] Artstein R, Poesio M. Inter-coder agreement for computational linguistics[J]. Computational Linguistics, 2008: 555-596.
- [151] McHugh M L. Interrater reliability: the kappa statistic[J]. Biochimia medica, 2012: 276-282.
- [152] Li P, Zhou G. Employing morphological structures and sememes for chinese event extraction [C]//Proceedings of the 24th International Conference on Computational Linguistics. 2012: 1619-1634.
- [153] Zeng Y, Yang H, Feng Y, et al. A convolution bilstm neural network model for chinese event extraction[M]//Natural Language Understanding and Intelligent Applications. Springer, 2016: 275-287.
- [154] Wu Y, Zhang J. Chinese event extraction based on attention and semantic features: A bidirectional circular neural network[J]. Future Internet, 2018: 95.
- [155] Feng X, Qin B, Liu T. A language-independent neural network for event detection[J]. Science China Information Sciences, 2018: 092106.
- [156] Wu X, Wang T, Fan Y, et al. Chinese event extraction via graph attention network[J]. Transactions on Asian and Low-Resource Language Information Processing, 2022: 1-12.
- [157] Tong M, Xu B, Wang S, et al. Docee: A large-scale and fine-grained benchmark for document-level event extraction[C]//Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2021.

致 谢

衷心感谢我的导师许斌老师对我本人的精心指导。能够来到清华大学进一步的学习，离不开许老师的帮助，在此衷心感谢许老师给我这个机会，让我能够圆梦清华。许老师坚持认真的科研态度也深深感染了我，让我以后在做任何事情的时候都能够沉得下心，沉得住气，不会过于浮躁。

其次，非常感谢李涓子教授对我的帮助。李老师每次都能够在我科研的关键阶段给与我方向，在博士论文写作过程中，也给我提出了很多建设性的意见。

我在初次步入清华大学求学时，对科研是什么没有一个很清晰的认知，是侯磊老师帮助我从零开始了解应该怎样做科研，应该去关注哪些领域专家，应该怎样开展自己的第一篇工作，在此由衷的感谢侯老师的指导。

同时特别感谢身在新加坡的曹艺馨师兄对我的热心指导与帮助。在每一次论文投稿的过程中，曹艺馨师兄都能对我的写作给出建设性的意见，让我受益良多。

在这里也要由衷的感谢我的父母在我读博期间给予我的鼓励和帮助。没有他们的帮助，我很难度过读博最开始那段艰难的时光。没有他们的帮助，我也很难有现在的成就。

特别的，感谢实验室王帅师兄在读博期间对我科研的帮助。感谢金海龙师兄、史佳欣师兄、吕鑫对我科研的帮助以及实验室韩美奂、刘明辉、赵博文、齐济、王晓智等同窗的热情帮助和支持！

最后，要感谢自己在读博期间坚持不懈的努力。回顾这段读博的经历，其实我对自己还是很敬佩的，特别是最后的博士论文学写阶段，虽然遇到了重重困难，但是我还是坚持了下来，这很不容易。在这里，隆重地表达对自己的感谢，也鼓励自己再接再厉，好好过好今后的每一天。

声 明

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

个人简历、在学期间完成的相关学术成果

个人简历

1995 年 05 月 05 日出生于河南三门峡市。

2013 年 9 月考入北京邮电大学计算机系网络工程专业，2017 年 7 月本科毕业并获得工学学士学位。

2017 年 9 月免试进入清华大学计算机系攻读工学博士至今。

在学期间完成的相关学术成果

学术论文：

- [1] **Meihan Tong**, Shuai Wang, Bin Xu, Yixin Cao, Minghui Liu, Lei Hou, Juanzi Li. Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021. (CCF 推荐 A 类会议, 长文)
- [2] **Meihan Tong**, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li and Jun Xie. Improving Event Detection via Open-domain Trigger Knowledge. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. (CCF 推荐 A 类会议, 长文)
- [3] **Meihan Tong**, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, Tat-Seng Chua. Image Enhanced Event Detection in News Articles. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. (CCF 推荐 A 类会议, 长文)
- [4] **Meihan Tong**, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, Juanzi Li. DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction. In Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL 2022. (TH-CPL 推荐 B 类会议, 长文)
- [5] **Meihan Tong**, Bin Xu, Shuai Wang, Lei Hou, and Juanzi Li. Improving Low-Resource Chinese Event Detection with Multi-Task Learning. In Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management, KSEM 2020.
- [6] **Meihan Tong**, Bin Xu, Lei Hou, Juanzi Li, Shuai Wang. Leveraging Multi-head Attention Mechanism to Improve Event Detection. In Proceedings of the 18th China National Conference on Computational Linguistics, CCL 2019.

- [7] Shuai Wang, Lei Hou, **Meihan Tong**. Unsupervised Cross-Lingual Sentence Representation Learning via Linguistic Isomorphism. In Proceedings of the 12th International Conference on Knowledge Science, Engineering and Management, KSEM 2019.

专利：

- [8] 许斌, 全美涵, 李涓子, 侯磊. 一种多模态事件检测方法及装置: 中国, 专利号: ZL 2020 1 0076960.1.
- [9] 许斌, 全美涵, 李涓子, 侯磊. 事件检测深模型的构建方法、装置、电子设备及存储介质: 中国, 专利号: ZL 2020 1 0548917.0.
- [10] 许斌, 全美涵, 李涓子, 侯磊. 用于命名实体识别的模型训练方法、识别方法及装置: 中国, 专利号: ZL 2021 1 0621275.7.

指导教师学术评语

事件抽取是知识图谱研究中一个重要研究方向，该论文以深度学习模型中的低资源事件抽取技术作为研究内容，选题具有十分重要的理论意义和应用价值。

论文的主要创新性成果包括：

提出了一个多模态融合的事件检测方法，通过融合新闻中天然存在的多模态信息，为事件消歧提供更多的语义证据，提升模型在低资源事件检测上的效果。

提出了一个基于知识蒸馏的事件检测方法，通过引入人类常识，为模型提供先验，从定性和定量角度证明了引入外部知识对低资源事件检测的有效性。

提出了一个基于自标注的事件要素抽取方法，通过挖掘任务相关类，为模型提供更多训练数据，提升模型在低资源事件要素抽取上的效果。

最后，将上述成果应用于在线新闻挖掘系统 NewsMiner，构建了大规模中英文事件抽取数据集，在此基础上搭建了中英文事件抽取系统。

论文结构合理，叙述清晰，书写规范。论文工作表明，论文作者已经具有本学科坚实宽广的基础理论知识和系统全面深入的专门知识，具有独立从事科研工作的能力。论文达到工学博士学位水平，同意并建议安排该生博士论文答辩。