

事件抽取中的共性数据依赖研究

2024 年 6 月

中图分类号：TP391.1

UDC分类号：004.8

事件抽取中的共性数据依赖研究

作者姓名	***
学院名称	计算机学院
指导教师	****
答辩委员会主席	****
申请学位	工学博士
学科/类别	计算机科学与技术
学位授予单位	北京理工大学
论文答辩日期	2024 年 6 月

Research on Common Data Dependency in Event Extraction

Candidate Name:	***
School or Department:	School of Computer Science
Faculty Mentor:	****
Chair, Thesis Committee:	****
Degree Applied:	Doctor of Philosophy
Major:	Computer Science and Technology
Degree by:	Beijing Institute of Technology
The Date of Defence:	June, 2024

事件抽取中的共性数据依赖研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下独立完成的研究成果。文中所撰写内容符合以下学术规范（请勾选）：

☒ 论文综述遵循“适当引用”的规范，全部引用的内容不超过 50%。

☒ 论文中的研究数据及结果不存在篡改、剽窃、抄袭、伪造等学术不端行为，并愿意承担因学术不端行为所带来的一切后果和法律责任。

☒ 文中依法引用他人的成果，均已做出明确标注或得到许可。

☒ 论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

☒ 与本人一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此声明。

签 名：

日 期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：

①学校有权保管、并向有关部门送交学位论文的原件与复印件；

②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；

③学校可允许学位论文被查阅或借阅；

④学校可以学术交流为目的，复制赠送和交换学位论文；

⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签 名：

日期：

导师签名：

日期：

摘要

事件抽取作为信息抽取领域的关键组成部分，其提供的结构化语义信息广泛应用在事实核查、信息检索、智能问答、推荐系统等多样化自然语言场景中。此外，以事件抽取为基础构建的事件知识图谱同样具备了广阔的应用前景。随着深度学习技术的快速发展，现有方法利用特定神经网络架构研究事件抽取的不同阶段任务。尽管这些方法取得了优异的性能进步，仍存在如下的共性数据挑战：1) 不同事件检测方法存在的数据不平衡依赖，限制了其识别事件的能力，进一步降低了整体检测性能；2) 不同事件要素抽取方法存在的实体类型过度依赖，阻碍了其学习实体类型以外的更多语义信息，影响了事件要素类型的正确预测；3) 当事件要素抽取任务扩展到多级别文本时，不同事件要素抽取方法存在的内在局限和跨事件信息限制，阻碍了其从不同级别文本中有效构建跨事件依赖关系，限制了多事件的要素抽取能力。为此，本文围绕事件抽取中的共性数据依赖展开研究，主要研究内容和创新成果如下：

(1) 提出基于分类器自适应知识蒸馏的数据不平衡依赖消解方法。针对事件检测中存在的数据不平衡依赖，首先研究其影响事件检测性能的方式，并相应定义和引入句子级别识别信息，以主动消解数据不平衡依赖导致的性能损失。在此基础上，提出一种基于分类器自适应知识蒸馏的事件检测方法。具体地，该方法在事件识别增强网络中引入根据训练实例标签转化的句子级别识别信息，并将其作为输入的一部分参与训练。进一步，设计额外的训练任务，通过事件识别增强网络的分类器引导事件检测网络从原始文本中自动学习句子级别识别信息，从而实现事件检测性能的提升。实验评估表明，提出的方法可以通用有效地消解数据不平衡依赖对于不同事件检测模型的性能影响，并展现自主适应不同程度数据不平衡和跨任务扩展的能力。

(2) 提出基于多角度对比学习的实体类型过度依赖消解方法。针对事件要素抽取中存在的实体类型过度依赖，首先定义并研究该依赖在不同事件要素抽取模型架构中的存在性。在此基础上，提出一种基于多角度对比学习的事件要素抽取方法。具体地，构建两个监督对比学习模块，在正样本角度提升具有相同要素类型但具有不同实体类型的实例特征表示间的相似性，而在负样本角度降低具有不同要素类型但具有相同实体类型的实例特征表示间的相似性。进一步，设计循环训练策略提升多角度消解的效率。实验评估表明，提出的方法可以通用有效地消解不同事件要素抽取架构的实体类型过度依赖，并基于预训练语言模型架构取得当前最优性能。

(3) 提出基于分离-融合范式的跨事件信息构建与利用方法。首先总结目前通用文本级别事件要素抽取方法存在的内在局限, 提出基于多词元链接的建模方法。进一步, 为了从不同级别文本中高效利用跨事件依赖信息, 提出分离-融合抽取范式, 其将跨事件信息获取和事件要素抽取先分开建模再进行融合, 从而充分利用不同抽取范式的优势。遵循该范式, 提出一种新的多词元链接模型。具体地, 引入一个链接模块连接不同事件的触发词和对应的要素类型以提供跨事件信息, 同时构建另一个链接模块抽取目标事件的要素信息。此外, 提出一个两阶段融合模块将获取的跨事件信息融合到目标事件的要素抽取中。实验结果表明, 提出的方法在不同的句子级别和文档级别数据集上均显著超越目前性能最优的事件要素抽取基线。

本文面向事件抽取系统中的数据不平衡依赖、实体类型过度依赖和跨事件依赖三类共性数据挑战, 相应研究分类器自适应知识蒸馏、多角度对比学习和分离-融合范式三种通用性方法, 覆盖不同模型架构、不同文本范围和不同事件抽取阶段, 以有力提升语言智能化和事件知识图谱构建自动化水平。

关键词: 事件抽取; 共性数据依赖; 数据不平衡依赖; 实体类型过度依赖; 跨事件依赖

Abstract

As a key component in information extraction, event extraction provides structured semantic information that is widely used in diverse natural language scenarios such as fact checking, information retrieval, intelligent question answering, and recommendation systems. In addition, the event knowledge graph based on event extraction also holds broad application prospects. With the rapid development of deep learning technology, existing methods use specific neural network architectures to study different stages of event extraction. Although these methods have made excellent performance progress, there are still the following common data challenges: 1) The data imbalance dependency in different event detection methods limits the ability to identify events and further reduces the overall detection performance; 2) The entity type overdependency in different event argument extraction methods prevents these methods from learning more semantic information besides entity types, which affects the correct predictions of event argument roles; 3) When extending event argument extraction to multi-level text, the inherent limitations and cross-event information restrictions of different event argument extraction methods prevent them from effectively constructing cross-event dependency from text of different levels, which limits the ability of multi-event argument extraction. Therefore, this paper focuses on the common data dependency in event extraction. The main research contents and innovative achievements are as follows:

(1) A data imbalance dependency resolution method based on classifier-adaptation knowledge distillation is proposed. For the data imbalance dependency in event detection, this paper first studies how it affects event detection performance, and defines and introduces sentence-level identification information accordingly to eliminate the performance degradation caused by the data imbalance dependency. On this basis, an event detection method based on classifier-adaptation knowledge distillation is proposed. Specifically, this method introduces the sentence-level identification information transformed according to the training example labels into the event identification enhancement network, and takes it as part of the input to participate in the training. Furthermore, an additional training task is designed to guide the event detection network to automatically learn sentence-level identification information from the original text through the classifier of the event identification enhancement

network to improve event detection performance. Experimental evaluation shows that the proposed method can generally and effectively eliminate the performance impact of the data imbalance dependency on different event detection models, and demonstrate the ability to adapt to varying degrees of data imbalance and cross-task expansion.

(2) An entity type overdependency resolution method based on multi-view contrastive learning is proposed. This paper first defines entity type overdependency in event argument extraction and studies its existence in different event argument extraction models. On this basis, an event argument extraction method based on multi-view contrastive learning is proposed. Specifically, two supervised contrastive learning modules are constructed to increase the similarity between the learned feature representations of instances with the same role type but having different entity types from the perspective of positive samples and reduce the similarity between the learned feature representations of instances with different role types but having the same entity type from the perspective of negative samples. Furthermore, a cyclic training strategy is designed to improve the efficiency of multi-view resolution. Experimental evaluation shows that the proposed method can generally and effectively eliminate the entity type overdependency of different event argument extraction architectures and achieve the state-of-the-art performance based on the pre-trained language model.

(3) A cross-event information construction and utilization method based on separation-and-fusion paradigm is proposed. Firstly, the inherent limitations of the current general text-level event argument extraction methods are summarized, and a modelling method based on multiple token linking is proposed. Furthermore, a separation-and-fusion extraction paradigm is proposed to efficiently utilize cross-event dependency information from the text of different levels. It models cross-event information acquisition and event argument extraction separately and then fuses them, thus fully using the advantages of different extraction paradigms. Following this paradigm, a novel multiple token linking model is proposed. Specifically, a linking module is introduced to connect the triggers of different events and the corresponding role types to provide cross-event information, and another linking module is constructed to extract the argument information of the target event. In addition, a two-fold fusion module is proposed to fuse the acquired cross-event information into the argument extraction of the target event. The experimental results show that the proposed method significantly sur-

passes the state-of-the-art event argument extraction baselines on different sentence-level and document-level datasets.

For the three common data challenges in event extraction systems: data imbalance dependency, entity type overdependency and cross-event dependency, this paper studies three general methods, including classifier-adaptation knowledge distillation, multi-view contrastive learning and separation-and-fusion paradigm, to cover different model structures, different text ranges and different event extraction stages, thus effectively improving language intelligence and event knowledge graph automatic construction.

Key Words: event extraction; common data dependency; data imbalance dependency; entity type overdependency; cross-event dependency

目录

摘要	I
Abstract	III
第 1 章 绪论	1
1.1 研究背景与研究意义	1
1.2 研究内容与主要贡献	2
1.3 论文组织结构	4
第 2 章 国内外研究工作概述	7
2.1 事件检测	7
2.1.1 基于卷积神经网络的方法	8
2.1.2 基于循环神经网络的方法	9
2.1.3 基于图神经网络的方法	11
2.1.4 基于预训练语言模型的方法	14
2.2 给定实体提及的事件要素抽取	18
2.2.1 基于卷积神经网络的方法	20
2.2.2 基于循环神经网络的方法	20
2.2.3 基于预训练语言模型的方法	21
2.3 未给定实体提及的事件要素抽取	22
2.3.1 基于分类的方法	23
2.3.2 基于机器阅读理解的方法	25
2.3.3 基于生成式的提示学习方法	26
2.3.4 基于抽取式的提示学习方法	27
2.3.5 其他方法	28
2.4 现有研究存在的问题	29

第 3 章	基于分类器自适应知识蒸馏的数据不平衡依赖消解	30
3.1	引言	30
3.2	方法设计	32
3.2.1	事件识别增强网络	32
3.2.2	事件检测网络	35
3.2.3	训练和推理	36
3.3	实验评估	37
3.3.1	实验设置	37
3.3.2	性能结果	39
3.3.3	与基于预训练语言模型的基线性能比较	40
3.3.4	事件触发词识别能力分析	41
3.3.5	数据不平衡程度分析	41
3.3.6	案例研究	42
3.3.7	其他信息抽取任务上的扩展实验	43
3.4	本章小结	44
第 4 章	基于多角度对比学习的实体类型过度依赖消解	46
4.1	引言	46
4.2	实体类型过度依赖定义与验证	48
4.2.1	定义	48
4.2.2	预实验验证	50
4.3	方法设计	52
4.3.1	编码器	52
4.3.2	正样本处理	53
4.3.3	负样本处理	56
4.3.4	循环训练策略	56
4.4	实验评估	57
4.4.1	实验设置	58

4.4.2	性能结果	59
4.4.3	消融研究	60
4.4.4	实体类型信息分析	61
4.4.5	超参数敏感性分析	63
4.4.6	计算复杂度分析	64
4.4.7	案例研究	65
4.5	本章小结	67
第 5 章	基于分离-融合范式的跨事件信息构建与利用	68
5.1	引言	68
5.2	方法设计	71
5.2.1	多事件链接构造	71
5.2.2	目标事件链接构造	73
5.2.3	两阶段融合	75
5.2.4	训练和推理	76
5.3	实验评估	77
5.3.1	实验设置	77
5.3.2	性能结果	80
5.3.3	不同多词元链接模型性能比较	81
5.3.4	单事件/多事件实例性能分析	84
5.3.5	消融研究	84
5.3.6	与资源增强模型的性能比较	86
5.3.7	案例研究	86
5.3.8	训练权重超参数分析	87
5.4	本章小结	88
结论	89
参考文献	92
攻读学位期间发表论文与研究成果清单	106

插图

图 1.1 事件抽取实例	2
图 1.2 研究内容关联图	3
图 1.3 论文组织结构示意图	5
图 2.1 DMCNN 模型示意图	9
图 2.2 RNN 示意图	10
图 2.3 GCN 示意图	12
图 2.4 Transformer 编码器示意图	15
图 2.5 四类事件要素抽取方法示意图	23
图 3.1 基于 DMBERT 的 CAKD 方法架构图	33
图 3.2 不同程度数据不平衡下不同方法在 F1 (%) 指标上的结果对比	42
图 4.1 实体类型过度依赖示例图	50
图 4.2 METOR 模型架构图	52
图 4.3 不同模型的实体类型过度依赖程度比较	63
图 4.4 METOR 模型参数敏感性分析	64
图 5.1 基于不同建模方式的多词元链接事件要素抽取示意图	69
图 5.2 Sep2F 模型架构图	71
图 5.3 标签矩阵 L_M 示例图	72
图 5.4 标签矩阵 L_S 示例图	74
图 5.5 要素类型串接长度与 Arg-C F1 指标的关联性	83
图 5.6 三种数据集上 Sep2F 与资源增强模型的性能结果对比	86
图 5.7 WikiEvents 数据集上的案例分析	87

表格

表 3.1	ACE2005 数据集上引入句子级别识别信息在 F1 (%) 指标上的结果对比	31
表 3.2	ACE2005 数据集上的事件检测性能结果	40
表 3.3	ACE2005 数据集上基于预训练语言模型的事件检测性能结果对比	41
表 3.4	ACE2005 数据集上的事件触发词识别任务 F1 (%) 指标结果	42
表 3.5	ACE2005 数据集上的案例分析	43
表 3.6	TACRED 数据上的关系抽取性能结果	44
表 4.1	实体类型过度依赖示例	47
表 4.2	测试集中不同类型实例簇的数目及性能结果	51
表 4.3	符合实体类型过度依赖的簇中实例统计信息	51
表 4.4	ACE2005 数据集上的事件要素抽取性能结果	60
表 4.5	ACE2005 数据集上的消融研究结果	61
表 4.6	ACE2005 数据集上不同变体模型的性能结果对比	62
表 4.7	ACE2005 数据集上的案例分析	66
表 5.1	数据集详细统计信息	78
表 5.2	超参数设置	79
表 5.3	四种数据集上的事件要素抽取性能结果	81
表 5.4	四种数据集上不同多词元链接模型在 Arg-C F1 (%) 指标上的结果对比 .	82
表 5.5	四种数据集上单事件/多事件实例对应的 Arg-C F1 (%) 指标结果	84
表 5.6	四种数据集上关于 Arg-C F1 (%) 指标的消融研究结果	85
表 5.7	四种数据集上不同 α 取值对应的 Arg-C F1 (%) 指标结果	87

第 1 章 绪论

1.1 研究背景与研究意义

随着数字化技术的高速发展，以互联网为主要传播载体的各种信息资源呈现出过载体化和爆炸化发展趋势。在这些形式多样的信息资源中，以新闻报道、电子刊物和社交内容等为代表的海量文本信息占据了不可或缺的地位，然而其主要以非结构方式存在，信息密度相对不高，不利于知识的进一步存储、转化与应用。为了提升对于文本信息的利用效率，大量的信息抽取（Information Extraction）^[1-4] 研究工作聚焦于自动化和智能化地获取非结构文本内容中的结构化知识数据。

近年来，事件抽取（Event Extraction）成为了信息抽取领域新的研究热点。事件抽取旨在从非结构的纯文本中抽取状态发生改变的事物，可表示为在特定时间、特定地点发生涉及若干参与者（物体）动作的结构化语义信息，其广泛应用于语言理解的不同领域，如事实核查^[5,6]、信息检索^[7,8]、智能问答^[9,10]、推荐系统^[11,12] 等。此外，利用事件抽取技术以及事件指代消解^[13,14]、事件关系抽取^[15,16] 等后续任务构建的事件知识图谱区别于以实体为中心的传统静态知识描述^[17]，反映了逻辑社会的动态事件性知识，可以进一步增益相应的下游应用^[18,19]。而且，以 ChatGPT 为代表的大语言模型虽然在多种自然语言处理任务中表现出类人的性能，但仍然受限于幻觉和缺乏显式知识建模^[20]。事件知识图谱作为幻觉检测和知识注入技术应用发展的重要基础之一^[21]，为探索解决大语言模型的此类局限提供了坚实支撑。

现有研究工作通常将事件抽取形式化处理为两阶段任务：事件检测和事件要素抽取。其中，事件检测任务识别出文本中触发事件的单词或者短语，并将其归类到正确的预定义事件类型上。而事件要素抽取任务旨在抽取出参与事件的实体提及以及对应的预定义事件要素类型。图 1.1 展示了一个具体的事件抽取实例，事件检测任务首先识别出文本中的短语“辞去”触发了“辞职”事件，而事件要素抽取任务基于事件检测的结果进一步抽取出参与事件的实体提及“埃隆·马斯克”、“推特”和“首席执行官”以及对应的要素类型“人物”、“公司”和“职位”。

随着深度学习技术的发展，这两阶段任务均取得相当优异的性能提升。然而，不管是针对事件检测还是事件要素抽取的研究，主要聚焦在特定的模型架构上进行，如循环神经网络、卷积神经网络、图神经网络和预训练语言模型等，忽略了任务中不同

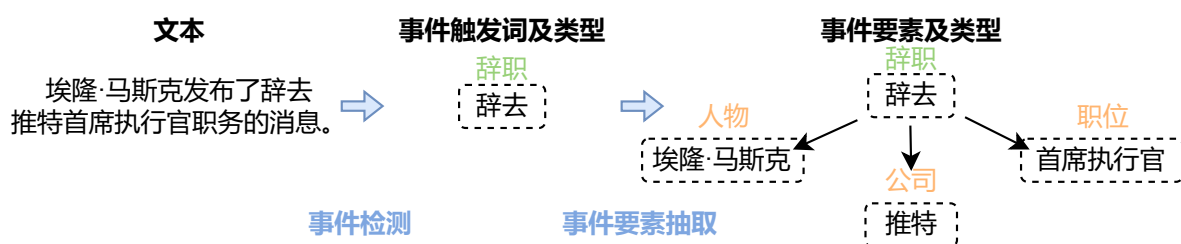


图 1.1 事件抽取实例

模型架构均普遍存在的数据依赖：(1) 事件检测数据集中不存在事件的实例比例过高，这种数据不平衡依赖影响了不同模型架构识别事件的能力，从而进一步导致事件检测分类的错误；(2) 事件要素抽取数据集中实体提及的类型与要素类型呈现出的强关联依赖，阻碍了不同模型架构更好地建模理解文本中的语义信息。此外，近年来随着各类文本信息资源呈现复杂化的演变趋势，越来越多的事件要素以跨越多个句子的方式出现，其对事件抽取中的事件要素抽取建模方式提出了新要求。因此，研究同时适用于句子级别和文档级别的事件要素抽取方法成为新的研究热点和挑战。然而，(3) 现有研究无法在避免共有内在局限性的前提下，实现对不同级别文本中均存在的跨事件依赖关系进行有效构建与利用。

因此，研究事件抽取中的共性数据依赖，可以进一步增强基于不同类型神经网络架构和不同级别文本的现有事件抽取方法的性能表现，提升其作为结构化语义信息应用到不同语言智能理解领域的质量，支撑事件知识图谱的构建。此外，本研究面向共性数据依赖的特点，赋予其适应新的事件抽取模型架构和建模范式的潜力，具备良好的扩展性。更为重要的是，本研究提出的共性数据依赖是各自所在子领域广泛存在但尚未引起重视的新挑战，具有广阔的深入研究前景。

1.2 研究内容与主要贡献

面对事件抽取系统中不同阶段存在的共性数据依赖，本文从不同的模型架构和文本范围出发，研究通用性的应对方法。其中，针对事件检测阶段普遍存在的数据不平衡依赖问题，提出了基于分类器自适应知识蒸馏的方法；针对事件要素抽取阶段普遍存在的实体类型过度依赖问题，提出了基于多角度对比学习的方法；针对将事件要素抽取扩展到多级别文本时普遍缺乏有效利用跨事件依赖关系的问题，提出了基于分离-融合范式的方法。如图1.2所示，上述研究内容覆盖了当前事件抽取系统的关键阶段和应用范围。接下来，具体阐述本文的研究内容和主要贡献如下：

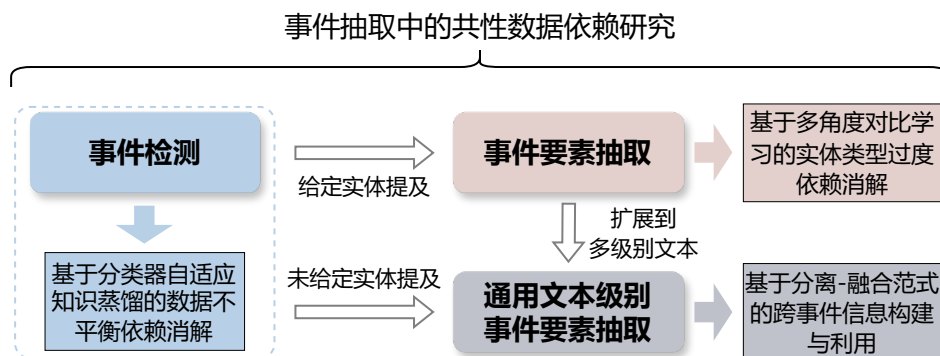


图 1.2 研究内容关联图

(1) 基于分类器自适应知识蒸馏的数据不平衡依赖消解

问题阐述：当前事件检测数据集存在严重的数据不平衡，限制了已有不同事件检测模型系统识别事件的能力，进一步降低了模型整体检测性能。针对该问题，相关研究特定于具体的模型架构和任务设计，依赖超参数以适应动态变化的数据不平衡程度，阻碍了可扩展性。

研究内容：针对上述问题，本文工作首先定义和引入句子级别识别信息，并验证该信息在消解数据不平衡依赖导致的性能下降方面的优势。基于此，研究利用知识蒸馏原理自动捕捉句子级别识别信息，提出在事件识别增强网络输入中增加对于句子级别识别信息的构建，并通过分类器参数引导事件检测网络从原始文本中自动学习该信息，实现事件检测性能的增强。

主要贡献：该研究为消解事件检测中存在的数据库不平衡依赖问题提供了新颖的角度。与现有的消解方法基于具体数据库不平衡程度被动地增强正样本或减弱负样本的训练权重不同，本文工作从数据库不平衡依赖影响事件检测性能的方式出发，以此引入了句子级别识别信息进行主动地消解数据库不平衡依赖导致的性能损失，并针对性地设计了知识蒸馏框架自动学习该信息。因此，这种主动方式实现了不依赖超参数而自主适应不同模型架构和数据库不平衡程度的多维度通用数据库不平衡依赖消解。

(2) 基于多角度对比学习的实体类型过度依赖消解

问题阐述：不同的事件要素抽取模型在建模事件触发词与实体提及的复杂交互时，受益于实体提及类型信息，忽略了其与要素类型间的强关联性。该强关联性使得不同的事件要素抽取模型在建模中对于实体类型过度依赖，影响了事件要素类型的正确预测。

研究内容：针对上述问题，本文工作首先定义实体类型过度依赖问题并通过实验

研究该问题在不同事件要素抽取架构中的存在性。基于此，研究利用监督对比学习分别从正样本和负样本两个角度进行实体类型过度依赖消解，并构建循环训练策略，以实现不同角度消解的高效协作。

主要贡献：该研究提供了在事件要素抽取中利用实体类型信息的新思路。与现有方法关注实体类型信息的**正面增益**不同，本文工作从实体类型与要素类型的数据特点出发，清晰地定义了实体类型过度依赖，并验证了其在不同模型架构中存在的普遍性，评估了其对于要素类型预测性能产生的**负面影响**。为了解决该问题，本文工作提出了两种新的监督对比学习方法，聚焦从不同角度减少实例特征表示对于实体类型的依赖性，提升整体文本语义信息的建模质量，通用有效地消解了实体类型过度依赖导致的性能损失。

(3) 基于分离-融合范式的跨事件信息构建与利用

问题阐述：当前适用于不同文本级别的事件要素抽取方法存在各自的局限性，而基于多词元链接的建模方法能够有效避免这些局限性并表现出良好的性能潜力。然而，在现有的多词元链接方法中，单一事件建模方式缺乏利用跨事件依赖关系，多事件建模方式需要增加编码序列和链接操作，导致了更多预测错误的积累。

研究内容：针对上述问题，研究结合不同多词元链接建模方式的各自优势，提出分离-融合抽取范式，将跨事件信息获取和事件要素抽取的建模过程进行分离，再将获取的跨事件信息融合到目标事件的要害预测中。遵循该研究范式，进一步提出相应的多词元链接模型，通过设计不同的多词元链接构建和两阶段融合机制，以实现建模不同事件间关联依赖性的同时，保持了高效直接的要害链接推理。

主要贡献：该研究为利用多词元链接进行通用文本级别事件要素抽取构建了新范式。与现有事件建模方式存在**各自优势**不同，本文工作提出的分离-融合抽取范式，构建了跨事件信息获取和事件要素抽取过程的分离再融合，从而实现了现有不同事件建模方式优势的**高效统一**。在该范式的指导下，提出了新的多词元链接模型，有效地构建与利用了不同文本级别事件要素抽取共有的跨事件依赖信息。

1.3 论文组织结构

本文工作面向事件抽取系统存在的共性数据依赖，根据**规划研究目标** ⇒ **明确研究内容** ⇒ **介绍现有工作** ⇒ **分析共性问题** ⇒ **阐述技术路线** ⇒ **总结研究成果**的行文结构，阐明相应的研究解决方法。如图1.3所示，将全文划分为六个章节，具体组织安排

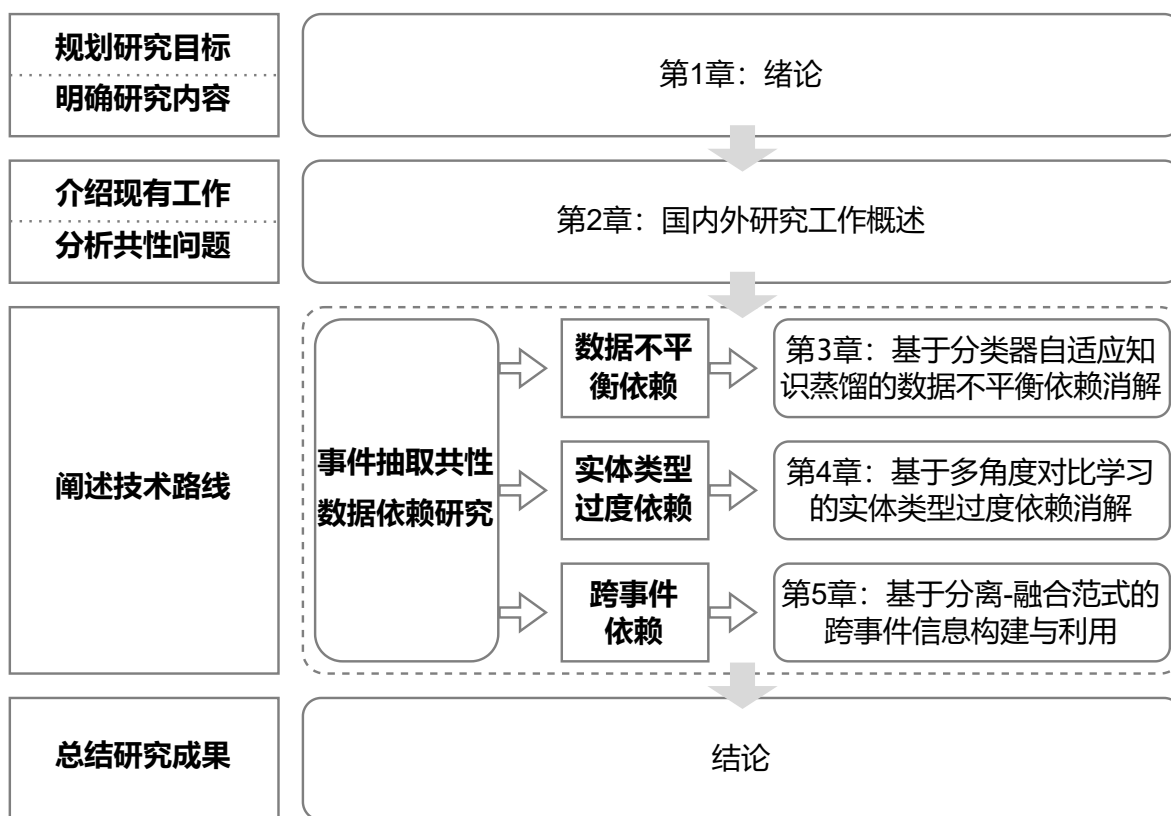


图 1.3 论文组织结构示意图

如下：

第一章为绪论，介绍了事件抽取任务的研究背景和意义，概述存在于该任务中的共性数据依赖，提炼相应的研究内容和贡献。

第二章为国内外研究工作概述，按照事件抽取系统中的不同阶段分别梳理发展脉络，分析存在的研究问题，归纳研究思路。

第三章围绕事件检测中的数据不平衡依赖消解，研究数据不平衡问题影响整体性能的原因与方式，验证句子级别识别信息在消解其性能损失方面的有效性，进而基于分类器自适应引导的知识蒸馏方法，自动从原始文本中捕捉句子级别信息，实现模型架构通用的数据不平衡自动消解框架。

第四章围绕事件要素抽取中的实体类型过度依赖消解，研究从不同角度消解事件要素抽取建模时对于实体类型信息的过度依赖，通过监督对比学习技术分别从正样本和负样本角度减少实体类型对于实例特征表示的过度影响，并利用循环训练策略优化消解过程，实现文本语义信息建模质量的提升。

第五章围绕通用文本级别事件要素抽取中的跨事件信息构建与利用，研究融合不

同多词元链接模型建模事件要素的优势，构建两种多词元链接和两阶段融合机制，以实现跨事件信息获取和事件要素抽取过程的分离再融合，在有效构建事件依赖的同时，保持单事件建模的简单链接预测，提升通用事件要素抽取系统的性能。

最后一章为结论，总结本文研究工作成果，并展望进一步的研究计划。

第 2 章 国内外研究工作概述

本章分别介绍事件检测和两种不同定义下的事件要素抽取技术的发展脉络，并梳理各自存在的共有问题，分析研究思路。

2.1 事件检测

事件检测 (Event Detection) 是从文本中抽取结构性事件知识的关键步骤^[22,23]，其旨在识别出具有特定事件类型的事件触发词。事件检测是以事件为中心的信息抽取技术的必要基础，在建模事件要素抽取、事件关系抽取和事件事实性识别等事件知识图谱构建任务时不可或缺，同时也广泛应用于社交内容分析^[24]、知识系统^[10,25]、故事生成^[26]等多种任务中。

早期的事件检测主要基于模式和特征进行建模。其中，基于模式的方法先构造事件模式，然后在抽取阶段根据事件模式匹配出对应的事件检测信息。事件模式一般基于原始文本或注释信息构建，在早期发展阶段由具备专业语言学知识的专家进行创建。进一步，一些研究利用少量事件模式种子，结合弱监督方法或自助法 (Bootstrapping) 实现更多事件模式的自动获取^[27-30]。

而基于特征的方法研究将事件检测分类的线索特征转化为特征向量，并使用机器学习方法进行建模。这些特征可以被划分为局部特征和全局特征。局部特征主要为词法特征和句法特征，包括完整的词、词性标签和依存句法特征等。例如，Chieu 等人^[31]引入一元文法和二元文法等特征信息，并利用最大熵分类器 (Maximum Entropy Classifier) 进行事件的检测。Chen 等人^[32]将包括前续字符、现有字符和后续字符的字符级别特征与词级别特征进行结合，并基于最大熵马尔可夫模型 (Maximum Entropy Markov Model) 检测中文事件触发词。而全局特征则不限于从事件触发词所在句子中得到，其可以为基于跨句子、跨文档和跨事件等获取的实体类型、事件要素信息和词义信息等^[33-38]，通常被用于局部信息的补充。例如，Ji 等人^[33]将相关文档中相同词在不同事件实例中的语义一致性作为全局信息依据，辅助局部的事件预测。Liao 等人^[35]构建无监督主题模型如 LDA (Latent Dirichlet Allocation)，以获取文档级别的全局信息，用于事件检测任务。Li 等人^[37]在建模中融合了不同事件的依赖关系，如事件触发词间的依赖路径和上下文信息。然而，基于特征的方法通常使用独热向量表示

构造的特征，易导致数据稀疏问题。更重要的是，不管基于模版还是特征的方法，都需要具备语言学知识的专业人士进行构建，对于一些低资源语言来说人力成本变得更加昂贵。因此，随着深度学习技术的快速进步，其能够自动从文本中学习语义表示的优势在事件检测任务中得以突显。接下来，本章根据使用的神经网络架构类型进行分别介绍。

2.1.1 基于卷积神经网络的方法

卷积神经网络（Convolutional Neural Networks, CNN）实质为多层感知机的变种形式，受益于其卷积核带来的权重共享和较小感受野，能够以较小的参数规模和计算复杂度实现特征信息的获取。在自然语言处理中，通常将传统二维卷积核的其中一个维度设置为词向量的维度，只改变另一维卷积操作的尺寸大小，将文本中多个词（或字符）组成的片段信息线性映射到特征空间中^[39]，实现自动捕捉这些文段的语义和句法信息。具体来说，给定 n 个词组成的句子文本 $\{w_1, w_2, \dots, w_n\}$ ，对应的向量输入为 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ，假设卷积窗口大小为 k ，则可得任意一个利用该卷积核运算的特征结果：

$$a_i = f(\mathbf{W} \otimes \mathbf{x}_{i:i+k-1} + b) \quad (2.1)$$

其中 \mathbf{W} 和 b 为待学习训练的参数， $f(\cdot)$ 为激活函数， $\mathbf{x}_{i:i+k-1}$ 表示从向量 \mathbf{x}_i 到 \mathbf{x}_{i+k-1} 的拼接， \otimes 为卷积计算。在实际应用中，为了建模特征的多样性，在每个卷积层中采用多个卷积核进行对应特征的捕捉。

在事件检测中，Nguyen 和 Grishman^[40] 为最早应用 CNN 的研究者。在其构建的方法中，首先将文本中每个词对应的词向量、实体类型向量和相对于目标预测词的位置向量进行串接，然后输入到包含不同尺寸卷积窗口的卷积层中，最后通过最大池化操作得到目标词的特征表示，从而进行最终的事件预测。为了在编码词法特征和句法特征时能够更多地关注到事件触发词信息，Chen 等人^[41] 在 DMCNN 模型中提出了一个动态池化操作去替代传统的最大池化操作，其根据目标预测词的位置将经过卷积层得到的特征表示划分为两部分，分别进行最大池化操作。如图2.1所示，根据“died”的位置将卷积层得到的向量表示划分为两个部分进行池化处理。与使用卷积层挖掘连续文段信息不同，Nguyen 等人^[42] 设定窗口长度并枚举所有符合该长度要求的非连续文段。然后，使用卷积层建模这些文段，并采用动态规划算法得到其中语义信息最显著的序列，用于最终的事件预测，从而实现非连续词间的长依赖信息的获取。

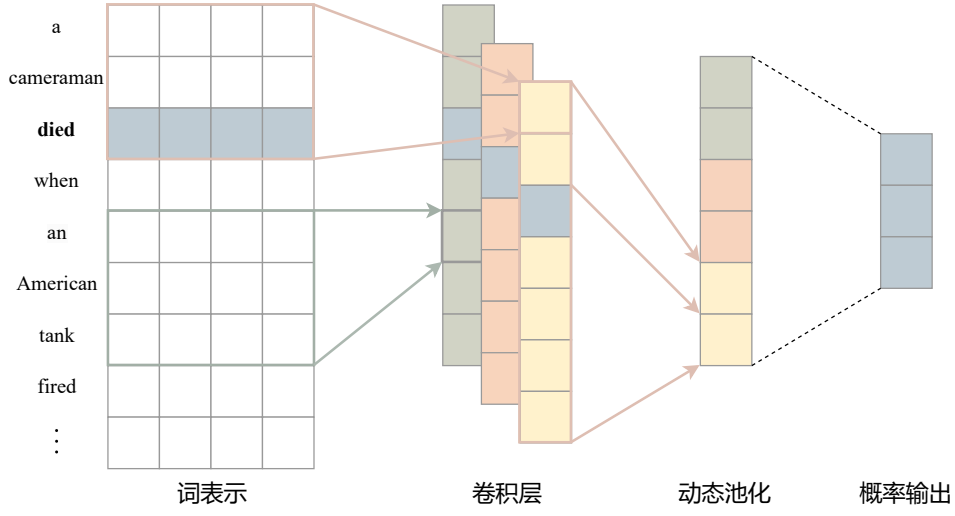


图 2.1 DMCNN 模型示意图

2.1.2 基于循环神经网络的方法

循环神经网络 (Recurrent Neural Networks, RNN)^[43] 常用于序列信息的建模, 其与传统的全连接网络不同, 每层的节点之间存在连接, 从而能够更好地捕捉序列中的顺序关联。由于自然语言文本可以视作词组成的序列, RNN 可以高效利用到单词顺序对特征编码和结果预测的影响, 因而广泛应用在自然语言理解和自然语言生成任务中。具体地, 如图2.2所示, 传统的单步 RNN 可以看作是相同的计算操作在不同时间步的重复叠加组合, 假设 n 个词组成的文本序列向量输入为 $\{x_1, x_2, \dots, x_n\}$, 则第 t 个时间步对应的输出向量 y_t 可得:

$$\begin{aligned} h_t &= f(W^h(h_{t-1}; x_t) + b^h) \\ y_t &= g(W^y h_t + b^y) \end{aligned} \quad (2.2)$$

其中 $f(\cdot)$ 和 $g(\cdot)$ 通常为非线性激活函数, h_t 为当前时间步的隐藏层状态向量, W^h 、 W^y 、 b^h 和 b^y 为不同时间步共享的训练参数。进一步, 以 LSTM^[44] 和 GRU^[45] 为代表的 RNN 变体被提出, 以缓解传统 RNN 建模长序列时产生的梯度消失和梯度爆炸现象。

在事件检测任务中, Ghaeini 等人^[46] 最先应用了 LSTM, 其根据候选触发词将文本划分为左侧上下文、候选触发词本身和右侧上下文, 分别使用 LSTM 进行编码并进行特征表示的串接, 以用于最终的多分类事件预测。Jagannatha 等人^[47] 则将医疗领域的事件检测转化为序列标注问题, 结合预定义事件类型和 BIO (Begin, Inside,

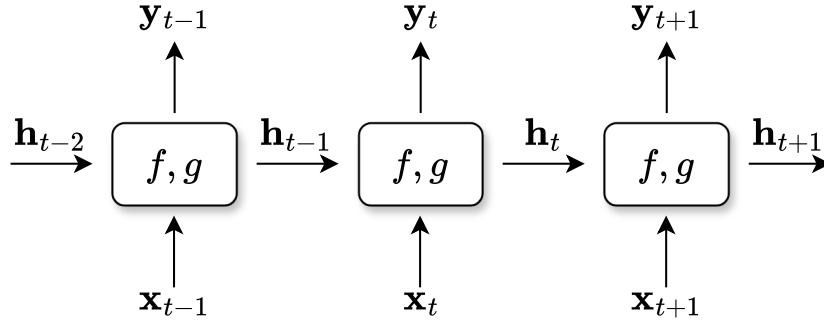


图 2.2 RNN 示意图

Outside) 标注策略, 使用双向 LSTM 或 GRU 进行建模。Orr 等人^[48] 将依存句法信息转化为有向非循环图, 并基于 GRU 提出了一个新的变种模型, 以更好地在事件检测中利用句法信息和上下文信息。为了解决 GRU 和 LSTM 等网络中门控计算依赖之前时间步的信息而导致的无法并行计算的问题, Zhang 等人^[49] 基于简单循环单元 (Simple Recurrent Unit, SRU), 构建两个 SRU 模型分别用于获取事件检测中单词级别和字符级别的语义特征, 以实现性能和计算效率的更好平衡。

虽然 RNN 相比于 CNN 能够建模更长的词间依赖关系, 但其建模的有效上下文长度仍然受限, 为了进一步提高上下文信息的利用效率, 一些研究者在此基础上引入了注意力机制。例如, Chen 等人^[50] 利用跨句信息和不同事件间的相关性增益事件检测的性能。在该模型中, 其一方面基于双向 LSTM 网络在句子级别得到每个词的特征表示, 然后使用两层注意力机制获取对应的文档级别特征表示, 并构建门控机制实现不同级别信息的融合。另一方面, 其构建了多层 LSTM 的变体网络层 TLSTM, 其在 LSTM 门控计算中增加了通过注意力机制得到的前一层预测的所有词的事件类型标签信息, 从而隐式地构建不同事件预测结果的关联性。Lou 等人^[51] 同样基于 LSTM 网络和注意力机制构建一个编码器-解码器架构, 以利用跨句范围内的语义信息和事件间的依赖关系。在编码器中, 使用双向 LSTM 网络进行编码, 并基于自注意机制融合长距离上下文语义信息。在解码器中, 分别使用融合了前一个词标签预测信息的双向 RNN 解码器, 并串接相应的解码表示, 以获取能够双向捕捉同一句子内不同事件关联依赖的事件标签向量。进一步考虑到跨句范围的事件依赖关系, 其引入了另一个 LSTM 网络编码已得到的事件标签向量, 从而聚合跨句范围内不同句子的整体事件信息。最后, 为了能够在解码中利用到聚合的事件信息, 作者们再次叠加新的双向解码层并融合前后相邻句子的事件信息。通过重复该融合过程, 更远距离的句子事件信息得以被捕获, 从而增益目标事件的检测。Liu 等人^[52] 探索不依赖事件触发词标注信息

去预测文本中存在的事件。作者们将该预测过程建模为多标签问题，使用 LSTM 编码得到词的特征表示，并使用不同的预定义事件类型向量作为查询向量，得到相应的文本整体表示，以预测对应事件类型的存在。

此外，一些研究针对 CNN 和 RNN 的特点，探索融合各自的优势进行事件检测任务的建模。例如，Feng 等人^[53]提出了一个融合 CNN 和 RNN 的混合神经网络，以用于多语言场景下的事件检测任务。其在使用双向 LSTM 网络建模上下文序列信息的同时，利用 CNN 获取局部文本特征表示。然后，串接这两种建模网络得到的信息，以进行多种语言数据集中的触发词检测。Lu 等人^[54]认为现有事件检测模型具备良好地区分不同上下文中容易混淆的高频触发词的能力，但缺乏识别罕见或未出现的触发词的泛化知识。为此，其一方面结合了 RNN 和注意力机制，建模以候选触发词为中心的上下文表示，以提供事件检测的触发词区分能力，并构建额外的二分类辅助任务增强候选触发词对其上下文表示的影响。另一方面，使用 DMCNN^[41]建模与候选触发词无关的上下文表示，以提供事件检测的触发词泛化能力。并且，利用生成对抗技术^[55]降低候选触发词对其上下文表示的影响。最后，基于门控机制实现不同上下文表示的高效融合，以增强事件检测的整体性能。

2.1.3 基于图神经网络的方法

句法表示能够提供单词与其句法相关上下文的直接连接，但无论 CNN 还是 RNN，均主要聚焦于建模自然语言文本中的序列。因此，为了更加高效直接地利用这类句法信息，近年来基于图进行建模的图神经网络 (Graph Neural Networks, GNN) 在自然语言处理的各个领域均得到广泛应用。在事件检测任务中，主要使用的是 GNN 中的图卷积神经网络 (Graph Convolutional Neural Networks, GCN)^[56]。给定图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ，其中 \mathcal{V} 为图中节点的集合， \mathcal{E} 为图中边的集合。假设第 l 层图卷积的特征表示 $\mathbf{H}^l = \{\mathbf{h}_i^l\}$ ，则第 $l+1$ 层图卷积的特征表示可计算如下：

$$\mathbf{h}_i^{l+1} = f\left(\sum_{(i,j) \in \mathcal{E}} \alpha_{ij}^l \mathbf{W}^l \mathbf{h}_j^l + \mathbf{b}^l\right) \quad (2.3)$$

其中 \mathbf{W}^l 和 \mathbf{b}^l 为待学习训练的参数， $f(\cdot)$ 为非线性激活函数， α_{ij}^l 为节点 i 指向节点 j 的边对应的权重值。图2.3展示了由四个节点组成的一个 GCN 示例。

Liu 等人^[57]和 Nguyen 等人^[58]为最早在事件检测任务中使用 GCN 的研究者。在

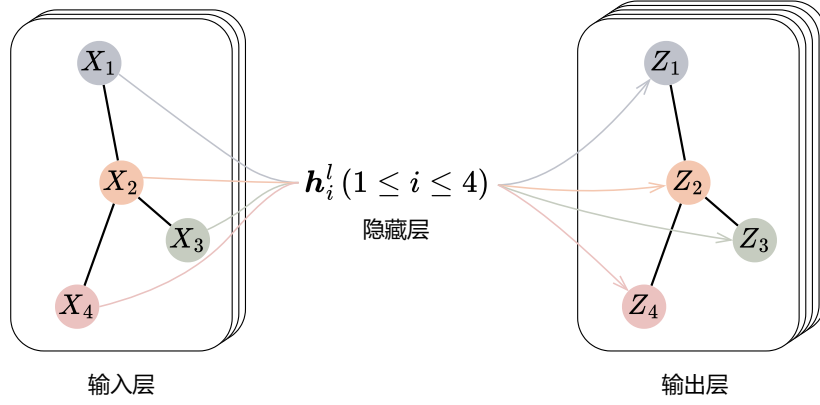


图 2.3 GCN 示意图

其研究工作中，首先使用自然语言处理工具将文本解析成结构化的依存句法树，然后设计策略将得到的依存句法树转化成可以用于建模 GCN 的图。具体地，图中的节点为句子中所有的单词，而节点间的边可以区分为三类。第一类边为直接从依存句法树中得到的正向边。假设在该树结构中，从单词 w_i 到单词 w_j 存在任意类型的依存句法关系，则在图中构建由节点 i 指向节点 j 的有向边。第二类边为所有第一类边的逆向边，以保证图中信息的逆向传播。第三类边则为每个节点指向自身的闭环边，使得构建的图能够形成闭环。基于所构建的图，进行多层 GCN 的迭代，实现依存句法树中多跳依存信息的高效获取。首先，利用双向 LSTM 网络层进行编码，以捕捉序列上下文信息。进一步，将编码得到的特征表示输入到 GCN 中，以获取构建的图中的结构化信息，其中不同类型的边使用不同的学习训练参数 \mathbf{W}^l 和 \mathbf{b}^l ，权重值 α_{ij}^l 计算如下：

$$\alpha_{ij}^l = \sigma(\mathbf{h}_j^l \mathbf{V}^l + d^l) \quad (2.4)$$

其中 \mathbf{V}^l 和 d^l 为权重计算时学习的参数，且同一种类型的边计算时进行参数共享。 $\sigma(\cdot)$ 为 sigmoid 激活函数。为了进一步在 GCN 中减少与候选触发词无关的信息干扰，Lai 等人^[59] 基于候选触发词的特征表示生成对应的门控向量，从而实现对其他单词特征表示的过滤筛选：

$$\mathbf{g}^l = \sigma(\mathbf{W}_g^l \mathbf{e}_t) \quad (2.5)$$

$$\mathbf{m}_i^l = \mathbf{g}^l \odot \mathbf{h}_i^l \quad (2.6)$$

其中 \mathbf{W}_g^l 为学习的参数， \mathbf{e}_t 为候选触发词在输入 GCN 前的编码表示， \odot 为元素依次相乘操作。基于该门控方法获取的不同单词特征表示，最终用于辅助候选触发词的事

件预测。此外，由于图网络中不同层建模上下文语义信息的侧重点有所不同，因此需要保证门控向量的多样性。为此，其设计了额外的正则化损失减少相同门控向量筛选过滤同一单词不同层特征表示后的相似程度。同时，作者们利用依存句法树中的依赖路径距离指导模型更好地感知不同单词对于候选触发词的重要程度。

此外，由于在事件检测常用的 ACE2005 数据集中，超过一半的事件触发词在其对应的依存句法树中和关联实体或者单词没有直接连接。为此，Liu 等人^[57]和 Nguyen 等人^[58]通过叠加多层 GCN 进行建模。然而，随着 GCN 层数的叠加，图中节点的特征表示将会变得愈加相似，导致过平滑问题^[60]。为了解决该问题，Yan 等人^[61]直接构建包含多跳依存句法关系的图，以替代在基于单跳依存句法关系构建的图上叠加多层 GCN。与 Liu 等人^[57]和 Nguyen 等人^[58]使用相同的策略，其首先将解析得到的依存句法树转化成包含正向边、逆向边和闭环边三种类型的有向图。不同的是，Yan 等人^[61]进一步根据各个类型得到对应的邻接矩阵，并计算任意 k 跳依存句法关系图的邻接矩阵如下：

$$\mathbf{A}_{type}^k = (\mathbf{A}_{type})^k \quad (2.7)$$

其中 \mathbf{A}_{type} 可为正向边、逆向边和闭环边分别对应的邻接矩阵。基于 \mathbf{A}_{type}^k ，为了捕捉到依存句法树中单词间 k 跳的句法关系，其使用单层图注意力网络（Graph Attention Network, GAT）^[62]替代多层 GCN 进行编码如下：

$$\mathbf{h}_i = f\left(\sum_{j=1}^n (u_{ij} a_{ij} (\mathbf{W}_a \mathbf{p}_j + \epsilon))\right) \quad (2.8)$$

$$u_{ij} = \frac{\exp(e_{ij})}{\sum_{j \in \mathcal{N}_i} \exp(e_{ij})} \quad (2.9)$$

其中 $e_{ij} = g(\mathbf{W}_1 [\mathbf{W}_2 \mathbf{p}_i; \mathbf{W}_2 \mathbf{p}_j])$ ， \mathbf{W}_1 、 \mathbf{W}_2 、 \mathbf{W}_a 和 ϵ 为待学习训练的参数， a_{ij} 表示 \mathbf{A}_{type}^k 对应的图中节点 i 指向节点 j 的边对应的权重值， \mathbf{p}_j 为图中节点 j 的输入向量， \mathcal{N}_i 表示在 \mathbf{A}_{type}^k 对应的图中节点 i 的邻接节点集合， $f(\cdot)$ 和 $g(\cdot)$ 分别表示 ELU 和 LeakyReLU 激活函数。最后，利用注意力机制聚合每个单词在不同跳句法关系图上建模获取的特征表示，以用于最终的事件类型预测。

与上述研究工作不同，最近的方法考虑进一步利用依存句法关系的类型信息。Cui 等人^[63]提出了在 GCN 中构造一个邻接张量，以利用其中的向量表示图中任意两个节点的关系。若图中节点间的依存句法关系类型相同，则对应的向量初始化也相同。基

于此，分别设计了图中节点表示和边表示的相互迭代更新方法，以实现高效建模邻接张量中的异质信息。Liu 等人^[64]在 Cui 等人^[63]研究工作的基础上，考虑了依存句法树中句法关系存在方向性的区别，初始化了两个非对称的邻接张量。然而，静态的邻接张量仍无法有效建模依存句法信息之外不同节点间潜在的依赖关联，为此其进一步构建了一个自注意力模块，以动态更新自注意力张量，并融入到节点表示和边表示的迭代计算中。

此外，Mi 等人^[65]利用依存句法树和语义角色标注构造双重关系图，在考虑句法结构的同时，引入互补的语义关系信息，以减少句法信息带来的噪音和冗余。具体地，在构建双重关系图时，保留在依存句法树中与候选触发词直接连接的单词和句法关系，将其他句法关系替换成表示与候选触发词的连接距离的虚拟关系。同时，根据语义角色标注结果引入语义关系的连接。基于构建的关系图，构建关系增强的 GAT，其基于对应关系向量和上下文信息进行注意力机制的权重计算。

2.1.4 基于预训练语言模型的方法

近年来，预先基于大规模无标签文本数据进行训练，再根据具体下游任务微调的建模方式广泛应用在各种自然语言处理任务中。在事件检测任务中，通常使用以 BERT (Bidirectional Encoder Representations from Transformers)^[66]为代表的双向语言模型作为进一步微调使用的预训练语言模型。BERT 主要由多层 Transformer^[67]编码器组成，其每一层的架构如图2.4所示。该架构由 N (一般设置为 6) 个相同的子网络层组成，每个子网络层主要包括多头注意力和前馈神经网络，并辅以加和连接和层归一化操作。在多头注意力 (MultiHead) 中，并行学习 h 组不同的线性变换层参数分别将自注意力计算输入中的键矩阵 K 、值矩阵 V 和查询矩阵 Q 映射到各自的特征空间中，然后拼接对应的自注意力计算结果，其中的 h 即为多头的数量，具体可表示如下：

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= [\text{head}_1; \dots; \text{head}_h] W^O \\ \text{head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \quad (2.10)$$

其中 W_i^Q , W_i^K , W_i^V 和 W^O 为学习训练的参数，Attention 为缩放点积注意力，计算如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.11)$$

其中 d_k 为键矩阵 K 和查询矩阵 Q 中的向量维度。

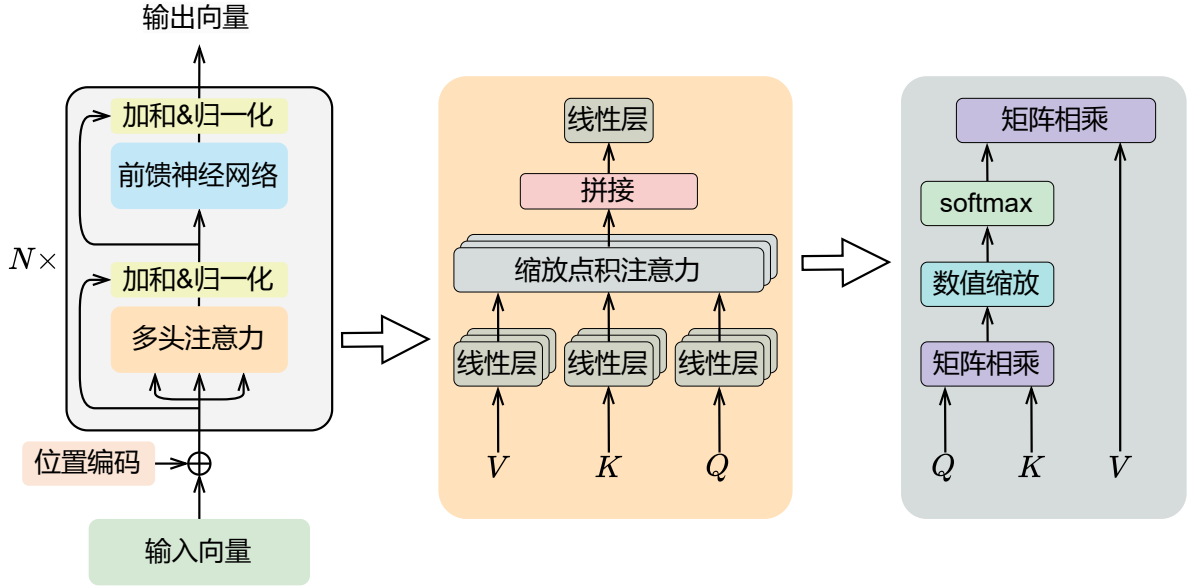


图 2.4 Transformer 编码器示意图

在预训练阶段，BERT 利用文档级别的大规模英文文本语料集，其包括英文维基百科数据和 BooksCorpus 语料库，执行如下两个预训练任务：

(1) 遮掩语言模型 (Masked Language Model, MLM)：受完形填空的启发，该任务随机选取 15% 已切分的 WordPiece^[68] 输入序列中的词元进行遮掩，并使用 BERT 预测被遮掩的词元。但该遮掩训练方式与微调阶段存在不一致。因此，基于此 15% 被选中的词元作进一步操作：保留其中 80% 词元的遮掩操作，取消其中 10% 词元的遮掩操作，并将剩下的词元随机替换成其他词元。通过提供预测词元的不确定性，促使模型更多地学习到全局上下文信息。

(2) 下一句预测 (Next Sentence Prediction, NSP)：以文本对方式构建该任务输入，表示为 “[CLS] + S_A + [SEP] + S_B + [SEP]”，其中 [CLS] 和 [SEP] 为特殊词元。该任务仅保留语料数据中一半句子的真实下一句作为 S_B ，另一半句子的 S_B 从语料集中随机选取得到，其旨在通过特殊词元 [CLS] 预测 S_B 是否被随机替换。通过此任务的训练，提升 BERT 模型在自然语言推理和问答等任务上的语言理解能力。

在下游任务微调阶段，根据任务的特点，选择单个文本或文本对输入到预训练 BERT 模型中，并叠加额外的设计模块进行训练损失微调。在事件检测任务中，Yang 等人^[69] 首先利用 BERT 模型获取各个词的特征表示并构建多分类器进行事件类型的建模。Wang 等人^[70] 将 DMCNN^[41] 模型中的 CNN 替换成 BERT，保留其动态池化

操作，构建对应的 DMBERT 进行事件预测。Liu^[71] 等人将事件检测问题转化为机器阅读理解任务，其简单地构建了一个特殊查询命令 “[EVENT]” 并和待抽取文本进行串接输入到 BERT 中，对于每个单词使用多分类的方法建模其对应的事件类型信息。类似地，Du 等人^[72] 设计了四类静态的问答模版用于事件检测，包括 “what is the trigger”、“trigger”、“action” 和 “verb”，并同样使用多分类器进行事件类型的建模预测。与 Liu 等人^[71] 和 Du 等人^[72] 不同，一些研究考虑显示地在预训练语言模型中利用事件类型的语义知识，深度建模其与待抽取上下文间的交互关系。例如，Li 等人^[73] 为了在 BERT 中利用到不同事件类型名称本身的语义信息，其将事件检测任务转化为多轮问答。具体地，首先设计固定的问题模版预测事件触发词的开始和结束位置，然后枚举所有的预定义事件类型并和已预测的触发词文本、触发词位置进行串接，根据特殊字符 [CLS] 进一步确定所属的事件类型。该方法不依赖预定义的事件类型数目，可扩展到新的事件类型预测，在零样本和少样本场景下同样表现出良好的性能潜力。Wang 等人^[74] 使用事件类型名称和高频原型触发词作为对应事件类型的描述，然后枚举所有的事件类型描述并分别与待抽取文本串接，利用二分类方式建模对应的事件触发词。进一步，在相同的框架下，Wang 等人^[75] 丰富了事件类型的描述模版，增加了事件定义、相关的事件要素类型、自动学习的软模版和人工构建的事件模版等信息。

进一步，一些工作基于 BERT 深入研究事件检测模型的鲁棒性、知识增强和预测模式等问题，以更好地挖掘预训练语言模型的潜力。例如，现有事件检测建模过于依赖词法信息，欠缺良好的上下文建模能力，从而导致在预测罕见或未出现的触发词时性能下降明显^[54]。针对该问题，Liu 等人^[76] 提出了两种上下文遮掩策略，以提升 BERT 模型在事件检测任务上的鲁棒性。第一种为句内策略，旨在提升模型学习句内上下文中的显著语义信息。具体地，固定已训练的事件检测模型的参数，其首先只遮掩候选触发词并使用遮掩向量作为查询向量计算对应的上下文注意力分布 α_u ，然后随机采样多种遮掩序列同样计算其上下文注意力分布。如果其中存在遮掩序列能够提升模型的预测结果，则以对应的上下文注意力分布作为训练目标，减少其与 α_u 的方差。第二种为句间策略，其基于单词分布假设^[77]，增加相同事件类型的触发词被遮掩后的上下文间的相似度，反之则减少对应的上下文间的相似度。类似地，Tong 等人^[78] 也观察到现有方法在预测罕见和未出现的事件触发词时性能表现不佳，并将其归因于数据集中存在的事件类型长尾分布。为了解决该问题，其基于知识蒸馏技术，将根据词义获取的开放域触发词知识用于提升对这两类事件触发词的识别。具体地，首先使用 WordNet^[79] 数据库从词义的角度实现对 New York Times^[80] 语料库的快速大

规模标注, 以获取开放域触发词知识。然后, 构建知识蒸馏框架, 促使教师网络和学生网络在监督数据集和大规模标注数据集上的预测概率均保持相似, 从而实现开放域触发词知识的迁移。其中, 在教师网络输入文本中使用两个特殊字符标注触发词的边界信息, 而在学生网络输入文本中则遮掩对应的触发词, 从而促使基于上下文信息的学生网络从教师网络中更好地学习识别触发词的能力。Liu 等人^[81]使用基于 BERT 训练的事件检测模型量化区分事件类型的潜在模式并根据模式的不同采取相应策略增强预测性能。具体地, 利用可解释领域的积分梯度算法 (Integrated Gradient)^[82], 计算文本中每个单词 w_i 对于事件预测的贡献值, 具体如下:

$$\alpha_{w_i} = (\mathbf{x}_i - \mathbf{x}'_i) \times \int_{\alpha=0}^1 \frac{\partial \mathcal{L}(t; \mathbf{X}' + \alpha \times (\mathbf{X}_s - \mathbf{X}'))}{\partial \mathbf{x}_i} d\alpha \quad (2.12)$$

其中 \mathbf{x}_i 为 w_i 由 BERT 编码得到的特征表示, \mathbf{X}_s 为文本中所有单词对应的特征表示组成的序列, \mathbf{X}' 为对应于 \mathbf{X}_s 的零向量序列, \mathbf{x}'_i 为 \mathbf{X}' 序列中第 i 个向量, t 为预测的事件类型的独热向量表示。

基于此, 进一步归一化得到的预测贡献值, 并根据每个单词的贡献值, 计算不同事件类型的触发词在整体训练集中的平均贡献。然后, 设置阈值且根据每个事件类型触发词的平均贡献将其划分为触发词依赖的事件或上下文依赖的事件。对于前者, 构建基于 BERT 的事件检测模型并在输入中使用额外的单词显著贡献向量来标记各个单词的贡献值信息, 辅助模型聚焦于具备高贡献值的单词进行建模。对于后者, 构建另一个基于 BERT 的事件检测模型并在输入中串接具有高贡献值的上下文片段, 辅助模型更好地利用相关上下文信息。在测试阶段, 集成此两个事件检测模型的预测, 得到最终的检测结果。

此外, 存在一些研究工作基于自回归预训练语言模型进行事件检测的数据增广。例如, Veyseh 等人^[83]使用 GPT-2^[84]生成训练数据。为了降低生成数据中的噪音信息, 其设计了教师-学生网络。教师网络使用高质量数据集进行训练得到, 而学生网络则同时使用高质量数据和生成数据进行训练, 并保持与教师网络的知识一致性。

近年来, 以 ChatGPT、PaLM^[85] 和 LLaMA^[86] 为代表的大语言模型 (Large Language Models, LLMs) 在自然语言处理的很多研究领域取得了令人印象深刻的性能表现。在事件检测任务上, 一些研究者基于 ChatGPT 进行了相关性能的评估。例如, Li 等人^[87]基于零样本场景评测了 ChatGPT 在事件检测任务上的性能。具体地, 其设计的输入提示模版包括对于任务的描述、预定义的事件类型集合、待抽取的文本内容和形式

化的输出格式要求。Gao 等人^[88] 丰富了输入提示模版的内容，一方面其扩充了对于各个预定义事件类型的详细描述，同时给出了正确样例和错误样例的输入输出信息。基于此，选取测试集中部分样例进行了性能评测。进一步，Han 等人^[89] 设计了与 Li 等人^[87] 相似的提示模版，并引入情景学习 (In-Context Learning, ICL)^[90] 和思维链 (Chain-Of-Thought, COT)^[91] 技术，提升了少样本场景下 ChatGPT 在事件检测任务上的性能。具体地，在原有提示模版的基础上，添加 5 个从训练集中随机采样的训练数据的文本信息和对应的正确事件检测结果作为示例，从而构建出 5-shot ICL 提示模版。对于 5-shot COT 提示模版，其关于事件检测的解释过程由人工进行构造。虽然这些研究采取了提供负样例、ICL 和 COT 等不同策略提升性能，但评估结果均表明 ChatGPT 在事件检测任务上的性能仍与现有监督模型存在明显差距。

2.2 给定实体提及的事件要素抽取

事件要素抽取 (Event Argument Extraction) 旨在根据事件检测得到的事件触发词和事件类型信息，进一步识别事件中存在的要素以及对应的要素类型。识别得到的要素作为结构化的语义信息，广泛应用于多种下游任务^[5,6,10,92]。

现有的事件要素抽取任务存在两种不同的定义方式。第一种定义方式主要基于自动内容提取项目 (Automatic Content Extraction, ACE)¹。该项目在评估设置中将实体提及的识别和事件要素抽取分离，因此在事件要素抽取中可以直接利用到实体提及的标签信息。具体地，给定待抽取文本、事件触发词和事件类型，本定义下的事件要素抽取将文本中已知存在的实体提及均视作候选事件要素，旨在识别出所有作为事件要素出现的实体提及，并给出对应的预定义事件要素类型。由于 ACE 项目聚焦于句子级别的信息标注，待抽取文本一般为单个句子。随着近年来文本信息朝着复杂化方向演变，出现了以 RAMS^[93] 和 WikiEvents^[94] 为代表的文档级别事件要素抽取数据集，其将实体提及的识别包含在事件要素抽取任务中。为了更好地研究适合不同文本级别的事件要素抽取模型，最近的一些事件要素抽取方法^[95] 给出了不同的定义方式。具体地，给定由单个句子或多个句子组成的待抽取文本、事件触发词和事件类型，该定义下的事件要素抽取旨在识别出每个事件要素的文段范围以及对应的预定义事件要素类型。与第一种定义不同的是，其要素类型的预测范围特定于给定事件类型，而非数据集中所有预定义的要素类型。

¹<http://projects.ldc.upenn.edu/ace/>

基于定义方式的不同,相应事件要素抽取研究工作的建模方式也存在较大差异。因此,本章主要介绍基于第一种定义方式的技术发展情况,其相关工作均根据 ACE 项目或类似的定义进行建模和评估。而第二种定义方式下的事件要素抽取技术将在下一章节进行梳理。

与事件检测任务相似,早期的事件要素抽取方法同样基于人工设计的事件模式或特征进行构建。其中,基于事件模式的方法结合词性标注结果和候选事件模式,匹配出对应的事件要素信息^[96]。而基于特征的方法主要使用词法特征和句法特征^[37]。词法特征一般包括单词的不同形态,如其完整形式和小写形式,以及词性标注信息等。而句法特征一般从依存句法解析结果中获取,表示为单词间依存关系形成的句法树。更具体地,给定一个单词,其主要句法特征信息一般为该单词在句法树中的深度、存在句法依存关系的词及对应的依存关系标签等。

进一步,虽然在本章梳理的事件要素抽取任务中,事件要素分布范围限定在单个句子,但存在一些研究工作基于更大范围的文本利用全局特征增益单个句子的要素抽取性能,如跨文档、跨句子和跨事件中的实体提及、事件要素和语义信息等。例如, Ji 等人^[33]利用相同实体在相关文档内的相同事件中通常以相同要素类型出现的一致性规律作为全局信息依据,辅助局部的事件要素抽取。在此基础上, Liao 等人^[97]进一步考虑相同实体在不同类型的事件中要素类型呈现的一致性。例如,在 ACE2005 数据集中, *Attack* (攻击) 事件中以 *Target* (目标) 要素类型出现的实体存在较大概率在 *Injure* (受伤) 事件和 *Die* (死亡) 事件中以 *Victim* (受害者) 要素类型出现。不同于 Ji 等人^[33]和 Liao 等人^[97]利用全局信息辅助调整局部抽取结果, Hong 等人^[36]直接使用全局的跨事件实体类型间的关系作为参与要素分类的特征。具体地,其基于 ACE2005 数据集的统计结果发现大多数具有相同实体类型的实体提及出现在相同类型的事件中,且进一步观察到这些实体提及中的大多数在事件中仅以少数几种要素类型出现。为了利用实体类型与事件类型和要素类型的一致性,其设计了事件类型、实体类型、领域内共现实体类型等特征,以用于句内要素识别的二分类预测和要素类型的多分类预测。此外, Liao 等人^[35]使用 Latent Dirichlet Allocation (LDA)^[98]主题模型得到每个文档的主题分布信息,然后将该全局信息作为特征应用到要素抽取中。

近年来,基于神经网络的事件要素抽取方法成为主流,其克服了模式构建和特征工程的困难。接下来,本章根据神经网络架构的类型,对这些方法进行总结。

2.2.1 基于卷积神经网络的方法

与图2.1所示的 DMCNN 模型类似, Chen 等人^[41] 也将其应用到事件要素抽取任务上。在卷积层部分, 其与事件检测中使用的 DMCNN 模型相同。但在动态池化部分, 考虑到相同候选要素可能在不同的事件中以不同的要素类型出现, 其在图2.1的两部分划分的基础上, 基于候选要素的位置作进一步划分, 以根据事件触发词和候选要素建模到更加关键的信息。然后, 利用划分为三段子文本的特征表示分别进行最大池化并拼接, 以用于事件要素的预测。Wang 等人^[99] 基于 DMCNN, 进一步考虑不同要素类型间存在的语义概念的相关性。例如, 要素类型 *Buyer* (买家) 和 *Seller* (卖家) 在抽象的概念层级中均与概念 *Person* (人物) 和 *Org* (机构) 存在关联, 因此这两个要素类型间的语义概念相关性要强于其他和这两个概念没有关联的要素类型。为了利用这种语义概念信息, 其在 DMCNN 的基础上进一步构建了层次模块化注意力。对于给定的要素类型和每个关联概念的向量表示, 分别计算每个单词对应的注意力得分, 并进一步得到每个单词基于所有关联概念的平均注意力得分, 以权重求和对应的特征表示, 从而获取给定要素类型的上下文表示向量。在计算每个要素的预测概率时, 将对应的上下文表示向量与 DMCNN 编码表示进行拼接, 以利用到相关的语义概念信息。

2.2.2 基于循环神经网络的方法

Nguyen 等人^[100] 最早在事件要素抽取中使用 RNN。其构建预训练词向量、实体类型向量和二元依存句法关系向量作为输入, 使用双向 GRU 网络获取文本中每个单词的特征表示。为了捕捉事件检测和事件要素抽取结果的相关性, 其根据文本中单词的顺序依次进行预测。具体地, 当预测到特定的单词时, 首先判断其是否为事件触发词。若该单词为事件触发词, 则在预测候选事件要素的类型时, 考虑利用之前已经预测得到的事件类型和要素类型信息。进一步, Sha 等人^[101] 考虑将句法信息融合到双向 LSTM 网络中。具体地, 对于不同的依存句法关系及其相应的逆关系, 设置不同的权重值, 并基于此修改 LSTM 隐藏层特征表示的计算:

$$h_t = o_t \odot \tanh(c_t) + d_t \odot \left(\frac{1}{|S_{in}|} \sum_{(i,p) \in S_{in}} a_p h_i \right) \quad (2.13)$$

其中 \mathbf{o}_t 和 \mathbf{c}_t 分别为 LSTM 的输出门控向量和记忆单元向量, S_{in} 中的每个元素 (i, p) 表示与第 t 个单词存在依存句法关系的单词顺序索引和对应的关系类型索引, a_p 为对应的关系权重值, \mathbf{d}_t 用于控制句法信息的过度影响, 其基于第 t 个单词的输入向量和第 $t-1$ 个单词的隐藏层特征表示学习得到。此外, 该方法在要素分类中考虑同时建模不同要素间的交互。为此, 其首先利用线性层得到任意候选要素间的交互向量。然后, 给定每个候选要素, 对其所有交互向量进行逐元素最大池化操作, 以捕捉最显著的交互信息。同时, 基于交互向量建模不同候选要素间的注意力权重和上下文要素表示。最后, 串接每个候选要素的最大池化交互向量、上下文要素表示、编码特征表示和事件触发词的编码特征表示, 以进行要素类型的预测。

此外, Nguyen 等人^[102] 利用 RNN 编码得到的特征表示同时建模实体提及识别、事件检测和事件要素抽取任务, 构建多任务学习损失进行训练, 实现不同任务间的知识共享和隐式依赖捕捉, 以提升事件要素抽取任务的性能。在建模过程中, 其沿用了 Nguyen 等人^[100] 的方法, 按照单词顺序依次预测事件触发词和事件要素, 并在要素预测时利用之前步骤预测的相关要素类型和事件类型信息。

2.2.3 基于预训练语言模型的方法

第一种定义方式下的事件要素抽取方法主要使用的预训练语言模型为 BERT。例如, Wang 等人^[99] 构建了事件要素抽取的基线模型 DMBERT, 其将 DMCNN 中的卷积层替换为 BERT^[103]。进一步, Wang 等人^[99] 在其提出的 HMEAE 框架中, 使用 DMBERT 替代 DMCNN 获取文本编码表示。在 DMBERT 的基础上, Xi 等人^[104] 引入 Nguyen 等人^[100] 用于存储已经预测得到的事件类型和要素类型信息而设计的二元记忆矩阵, 构建可以利用跨事件信息的 BERT(Inter) 模型。同时, Xi 等人^[104] 参照 Sha 等人^[101] 建模不同事件要素间的交互向量和上下文要素表示, 构建可以捕捉事件内不同事件要素关联性的 BERT(Intra) 模型。

然而, BERT(Intra) 隐式地构建事件要素间交互的方式无法直接使用其他要素预测结果, 且要素类型的本身语义信息也未得到充分利用。基于此, Xi 等人^[104] 将事件要素抽取转化为序列到序列的建模问题, 其中输入序列为标注了特定事件触发词信息的句子文本, 输出序列为文本中候选要素对应的要素类型序列。为了双向利用事件内不同事件要素的类型预测语义信息, 其使用编码器-解码器框架并针对性地提出了实体级别的双向循环解码架构。具体地, 在编码器的输入文本尾端融合给定事件触发词

的事件类型信息，并利用 BERT 获取每个单词的特征表示。而在解码器中，根据事件触发词和待抽取的实体提及位置信息，对编码器输出的特征表示使用 DMCNN^[41] 中的动态池化操作获取实例的整体特征表示。另一方面，使用 CNN 编码包含了每个单词在之前序列预测中获取的要素类型标签向量，从而融合已预测的事件要素信息，并同样利用最大池化操作得到其中最显著的事件要素信息的特征表示。然后，基于这两部分的特征表示，使用多分类器进行事件要素类型的解码。进一步，该模型在解码部分构建了正向和逆向两个不同的建模过程，在推断时使用双向建模得到的事件要素信息的特征表示参与最终的事件要素预测。

2.3 未给定实体提及的事件要素抽取

如上一章节所述，最近一些研究工作将事件要素抽取任务定义为旨在同时识别出事件要素的文段范围和要素类型，其抽取过程未使用实体提及识别任务的结果。因此，本章调研未给定实体提及的事件要素抽取的研究现状。随着基于 Transformer 架构的预训练语言模型技术的不断成熟，其预训练-微调模式已成为未给定实体提及的事件要素抽取方法的主流选择。其中，BERT、RoBERTa^[105] 和 BART^[106] 为本章介绍的方法中常用的预训练语言模型。

RoBERTa 使用了与 BERT 完全相同的模型结构，但在预训练设置上提出了如下改进：（1）在更大规模的英文语料库上进行预训练，相比于 BERT 使用的数据量扩增了 10 倍有余。（2）针对更大规模的训练数据，相应地将训练时的批数据大小增加到 8000，远超 BERT 设置的 256。（3）直接串联连续的多个句子作为输入，并移除 BERT 预训练中构建的 NSP 任务。（4）在 MLM 预训练任务中采用动态遮掩替代 BERT 中的静态遮掩设置。在 BERT 的静态遮掩设置中，被遮掩的 WordPiece 词元一旦确定，则在整个预训练过程中不再改变。而 RoBERTa 则将初始训练语料复制 10 份，分别随机选择不同的 WordPiece 词元进行遮掩，然后在训练的不同轮数使用经过不同遮掩选择得到的输入序列。

BART (Bidirectional and Auto-Regressive Transformers) 为序列到序列预训练语言模型，与 BERT 和 RoBERTa 不同，其使用了与 Transformer^[67] 相同的编码器-解码器架构。BART 本质上为一个将经过噪音信息破坏的文本还原成对应初始序列信息的去噪自编码器。在预训练中，BART 主要使用了如下的噪音引入方式：（1）遮掩词元：与 BERT 在 MLM 预训练任务中的设置相同，随机采样文本中的词元进行遮掩处理。（2）

删除词元：随机采样文本中的词元进行删除处理，提升模型感知缺失词元位置信息的能力。(3) 填充文本：采样不定长度的词元片段进行遮掩，该方式受 Joshi 等人在 SpanBERT^[107] 预训练语言模型中的设置启发。具体地，采样长度服从泊松分布（参数 λ 为 3）的词元片段进行遮掩处理。当长度为 0 时，则在文本中直接插入代表遮掩处理的特殊词元 “[MASK]”。该方式使得 BART 模型能够感知缺失的词元片段中的词元数量。(4) 旋转文本：随机选择文本中的一个词元，以该词元为界将整个文本划分为两部分，并旋转对调这两部分文本段，使得旋转后整个文本的开始词元为选择的词元。该设置可以提升模型对于原始文本的开头信息的识别。(5) 置换句子：随机置换文本中的句子顺序。基于上述的文本破坏方式，构建相应的重建文本损失进行 BART 模型的预训练。BART 模型广泛应用于序列生成、序列分类和词分类等任务中。

基于预训练语言模型，可进一步将现有未给定实体提及标注的事件要素抽取方法总结为如图 2.5 所示的四种类型。在该图中，“computer” 以要素类型 *Artifact*（人工制品）参与到 *Transfer – Ownership*（转让所有权）事件中。接下来，分别介绍各自类型的相关工作和方法特点。

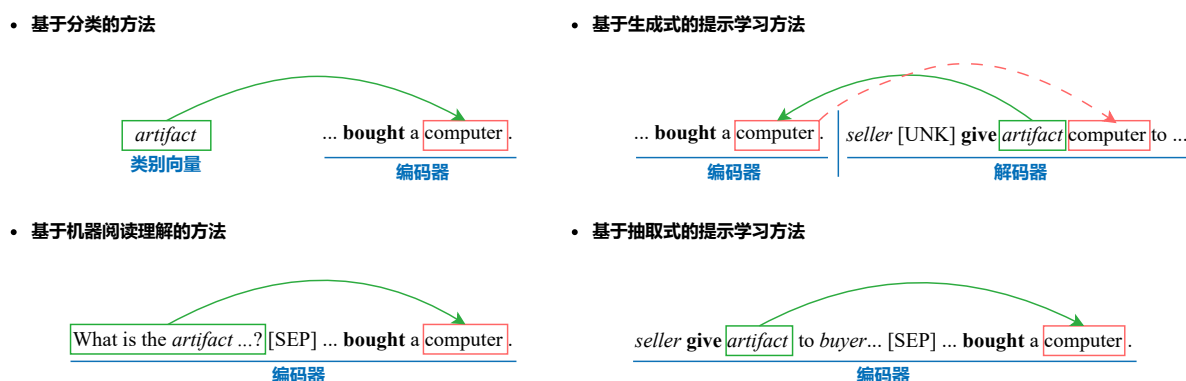


图 2.5 四类事件要素抽取方法示意图

2.3.1 基于分类的方法

基于分类的方法使用传统的分类方式，构建随机初始化的分类器进行事件要素的预测。Ma 等人^[108] 利用预训练语言模型获取能够感知触发词信息的序列上下文表示，并进一步设计了融入依存句法结构信息的 Transformer 架构，得到融合句法信息的编码表示，最后使用多分类器同时完成要素文段范围和类型的建模。具体地，其首先使用 BERT 编码待抽取文本，并利用 BERT 输入中的分隔向量（segment embedding）区分对应词是否为事件触发词。对于每个词，在其 BERT 输出的编码表示基础上，均串

接触发词和触发词对应的事件类型等信息，以增强对于事件信息的感知。然后，根据文本的依存句法解析结果，修改 Transformer 中的多头自注意力机制，将每个词注意力计算中的键和值限制为与其存在依存句法关系的词，从而使得其聚焦句法关联词间的自注意力建模。最后，结合预定义要素类型和 BIO 标注策略，利用多分类器同时完成对事件要素的文段和类型的抽取。与 Ma 等人^[108]直接使用序列标注建模的方式不同，一些研究者将事件要素抽取分解为子任务依次解决。例如，Zhang 等人^[109]将文档级别事件要素抽取拆解为要素头单词识别和基于头单词的文段扩充两个子问题。在要素头单词识别中，对于要素类型 r ，利用特定于 r 的双仿射注意力机制 (Biaffine)^[110] 参数建模候选词作为该要素类型出现的概率：

$$\Pr_r(t, c) = \frac{\exp \text{Biaffine}_r(e_t, e_c)}{\sum_{c' \in \mathcal{C} \cup \{\epsilon\}} \exp \text{Biaffine}_r(e_t, e_{c'})} \quad (2.14)$$

其中 e_t 和 e_c 为事件触发词和候选词经 BERT 编码后得到的特征表示， \mathcal{C} 为候选词集合， ϵ 表示要素类型 r 不存在对应要素。而在获取头单词后，根据设置的最大要素文段长度，枚举所有可能的候选左边界和右边界词。然后，将基于头单词的文段扩充子问题转化为寻找包含该头单词的左右边界词问题，并视每个候选边界词为一个类别，构建两个多分类任务分别建模。同样地，Xu 等人^[111]将文档级别事件要素抽取转化为要素边界识别和要素类型分类两个子任务依次进行建模。在要素边界识别任务中，使用两个不同的 BERT 编码器捕捉局部和全局上下文信息。其中，局部信息的编码通过遮掩跨句词间的注意力计算得到。进一步，为了在局部信息中融入更加丰富的语义结构知识，其基于抽象语义表示 (Abstract Meaning Representation, AMR)^[112] 的解析结果，使用 GCN 对于局部编码表示进行语义关联信息的聚合。最后，基于门控机制融合包含了语义结构知识的局部信息和文档全局信息，并使用两个二分类损失构建要素边界的识别。而在要素类型分类中，同时使用触发词、事件类型、识别出的要素以及要素与触发词间的交互等信息进行要素类别的判断。此外，Liu 等人^[113]将事件要素抽取转化为基于枚举文段的分类问题。具体地，其使用 BERT 或 RoBERTa 模型联合编码事件信息、待抽取文本和事件相关的要素类型信息，并对于给定的枚举文段，使用同时与该文段和事件触发词存在相关性的上下文和要素类型信息作为其整体表示。在此基础上，一方面根据对应的边界表示构建二分类边界损失，另一方面融合整体表示、边界表示和事件触发词表示等信息，构建事件要素类型的多分类损失。

2.3.2 基于机器阅读理解的方法

本类方法为每个要素类型构建相应的问题，将抽取对应要素转变为机器阅读理解 (Machine Reading Comprehension, MRC)^[114] 问题，其能够同时获取要素的边界和类型信息，并利用到要素类型本身的语义知识。Du 等人^[72] 探索了三种不同的要素类型问题模版，包括只使用要素类型名称、结合要素类型名称与对应语义类型的疑问词、结合 ACE 项目中关于要素类型的标注指南与对应疑问词。然后，使用 BERT 对待抽取文本和问题模版进行联合编码，并参照传统的 MRC 方法构建要素开头和要素结尾的查询向量，使用 softmax 归一化建模文本中每个词分别为要素开始词和要素结尾词的概率。由于给定要素类型可能不存在或存在多个事件要素，在要素边界推断时需人为设置预测阈值。为此，其进一步根据验证集设计启发式算法，以动态决定划分要素文段和非要素文段边界概率值的最佳阈值。同样地，Li 等人^[73] 设计了包含事件触发词、事件类型和待抽取的要素类型等信息的模版作为 MRC 中的对应问题。为了抽取可能存在的多个事件要素，其根据每个词的编码表示构建 BIO 的三分类器，以替代 Du 等人^[72] 采用预测阈值的做法。与 Du 等人^[72] 和 Li 等人^[73] 基于规则模版构建问题不同，Liu 等人^[71] 利用无监督问题生成技术，获取与待抽取文本上下文和主题相关的自然语言表达的问题，以提升 MRC 模型获取答案的能力。其无监督问题生成主要包括两个部分：问题主题生成和问题上下文生成。问题主题生成与 Du 等人^[72] 的设置相似，通过要素类型的语义类型使用不同的疑问词作为问题开端。而在问题上下文生成中，设计两个不同的编码器-解码器架构，以分别学习陈述性文本映射到问题描述和对应的相反映射。进一步，构建由大规模陈述性文本和未对齐的问题组成的数据语料，利用领域内自编码、去噪自编码和跨领域在线还原翻译任务进行训练。最后，使用将陈述性文本映射到对应问题的编码器-解码器模型生成待抽取文本对应的问题上下文。Zhou 等人^[115] 根据要素文段识别和要素类型识别分别构建两个对偶的 MRC 任务，通过参数共享和联合训练，实现相互协作，以提升建模性能。具体地，基于 BERT 分别编码待抽取文本、由要素文段和要素类型转换的问题，并利用流动注意力模块^[116] 实现问题表示和待抽取文本表示的交互。

进一步，一些研究利用 MRC 技术解决文档级别事件要素抽取带来的新问题与挑战。例如，Liu 等人^[117] 为了解决文档中存在的事件数据稀疏问题，利用 MRC 实现两种不同的数据增广。一方面，利用机器阅读理解、语义角色标注^[118] 和事件要素抽取的跨任务数据集进行 MRC 模型的预训练，实现相似任务中隐式知识的传递。另一方

面,使用预训练后的模型对无标注信息的文档进行事件要素信息的标注,以生成事件要素抽取任务中的更多数据实例。Wei 等人^[119]考虑到文档数据中事件要素可能分布在不同句子中,使得模型无法较好地捕捉事件触发词与要素间的长距离依赖以及事件要素间的隐式关联。为此,其在 MRC 模型中引入知识蒸馏技术,以提升模型在长范围文本场景下的事件内部推理和依赖知识建模能力。具体地,在其知识蒸馏框架中,对于每个要素类型,在教师网络的输入中引入事件内其他要素文段和类型信息,并使用 MRC 进行对应要素文段的建模。而在学生网络中,同样构建 MRC 抽取要素文段。基于此,从特征表示和预测概率分布两种角度将教师网络中引入的相关要素知识迁移到学生网络中。

此外,存在研究工作利用序列到序列框架直接生成对应的要素文段,以替代从文本中建模要素文段的范围。例如,Lu 等人^[120]首先构建问题生成模型,为每个要素类型获取包含了其他要素信息的问题,然后基于 BART 或 T5^[121]模型生成所有归属该要素类型的要素文段。其问题生成模型同样基于 BART 或 T5 构建,输入序列为待抽取文本,对应输出序列则融合了 ACE 项目中关于目标要素类型的标注指南和待抽取文本中存在的其他要素文段信息。

2.3.3 基于生成式的提示学习方法

本类方法构建提示模版,利用序列到序列模型生成待抽取的事件要素信息,其相比基于机器阅读理解的方法,能够更加直接地同时抽取多个具有相同要素类型的事件要素。Li 等人^[94]将文档级别事件要素抽取建模为基于提示模版进行填空的条件生成任务。其使用的提示模版为数据标注时参照的事件本体模版,并将模版中具体的要素类型名称替换为特殊占位符,即挖空操作。进一步,将经过挖空的提示模版和文本串接后输入到 BART 模型的编码器中,并在解码器中依次建模已完成要素文段填空后的提示模版序列。然后,比对填空的文段内容和对应被挖空的要素类型名称,实现要素信息的抽取。若挖空槽位的所属要素类型对应文本中多个要素文段,则在解码时使用“and”进行串联。Zeng 等人^[122]使用了和 Li 等人^[94]完全相同的模型架构和提示模版进行文档级别的事件要素抽取。不同的是,其在此基础上对于每个输入文本同时构建一个标注了相邻事件要素类型标签的信息增广文本,并通过对齐这两种输入文本中事件要素的编码表示,使得模型更好地感知跨事件要素的语义一致性。

类似地,Hsu 等人^[123]使用序列生成模型构建适用于不同级别文本的事件要素抽

取。其与 Li 等人^[94]的研究工作相比,主要存在以下不同:(1)在构造提示模版时融合了更丰富的事件信息,包括事件类型定义和与事件语义密切相关的关键词。(2)在对事件提示模版进行挖空时,使用自然语言词汇如“some attacker”、“some organization”等替代 Li 等人^[94]使用的特殊词符进行占位。该设计使得模型可以更好地学习到要素类型的语义信息和事件整体结构信息。(3)基于人工设计进行事件提示模版的构造,未直接利用数据集中的事件本体模版知识。在序列生成模型的基础上,Hsu 等人^[124]进一步利用 AMR 语义图知识。其将文本的 AMR 语义解析图转化为线性序列,并编码表示为若干向量表示,添加到模型 Transformer 架构的键向量和值向量中,从而实现结构化图知识与自然语言文本的异质融合。

进一步,Du 等人^[125]研究当文档中存在较多事件时,在序列生成模型中利用已预测的跨事件知识进行信息增益。在预测目标事件时,使用预训练的文本相似度衡量模型 S-BERT^[126]检索出最相关的已预测事件序列,并作为额外的记忆知识和文本以及事件提示模版串接输入到生成模型的编码器中,为目标事件的要素抽取提供跨事件信息。而当预测下一个事件的要素时,根据当前目标事件要素抽取结果转化的事件序列信息将同样参与到相似度检索中,从而不断丰富待检索的事件知识。进一步,Ren 等人^[127]研究在序列生成框架中利用多种检索增强策略提升事件要素抽取性能,包括上下文一致检索、模式一致检索和自适应混合检索。其中,上下文一致检索策略与 Du 等人^[125]采用的方法相似,模式一致检索策略利用事件类型标签进行检索查询,而最后一种策略则采用连续空间中的伪示例采样进一步提升序列生成模型的类比能力。

2.3.4 基于抽取式的提示学习方法

本类方法在构建提示模版的基础上,使用模版中要素类型对应的槽位表示作为查询信息,构建与机器阅读理解相似的答案获取过程,以抽取要素文段内容。该类方法结合了 MRC 和提示学习的优势,较好地挖掘了预训练语言模型中的先验知识。

Ma 等人^[95]最早提出了抽取式提示学习方法,以应用于句子级别和文档级别的事件要素抽取任务,其改变了之前提示学习方法基于挖空操作构建完形填空的建模方式。对于给定事件,其使用包含了该事件中所有候选要素类型的人工设计或自动学习的提示模版,并标记各自要素类型对应的槽位词位置。然后,利用 BART 模型建模,以文本作为编码器输入,事件提示模版作为解码器输入,得到融合了文本上下文信息的提示模版表示。进一步,基于标记的槽位词位置,分别获取不同要素类型对应的开

头与结尾查询表示向量，而后利用 MRC 技术构建要素文段范围。此外，其使用了二分图匹配问题中的匈牙利算法优化多答案情况下的训练损失计算。Zhang 等人^[128]同样将事件要素抽取任务转化为基于事件模版的槽位查询问题，使得其能够与语义角色标注任务相统一，从而利用到语义角色标注的数据集进行知识迁移，提升模型在低资源场景下的性能。与 Ma 等人^[95]不同的是，其使用 RoBERTa 模型替代 BART 模型，直接对文本和事件模版进行联合编码。Nguyen 等人^[129]基于 Ma 等人^[95]的研究工作，提出了增加软提示模版的学习，以构建每个训练实例的定制化模版输入，实现从外部文档中获取相关的上下文信息。

此外，一些研究在抽取式提示学习方法框架内考虑跨事件信息的利用。例如，Li 等人^[130]在利用 BART 模型获取目标事件中所有要素类型表示的基础上，分别构建事件内和事件外的要素类型依赖图。其中，事件外的图基于目标事件和与其最相似的事件进行构造。在此基础上，分别利用 GCN 建模各自的图结构信息，并在每层 GCN 迭代后进行跨图表示的融合。最后，利用融合了相似事件信息的不同要素类型表示和 MRC 技术，获取对应的要素文段。He 等人^[131]则进一步在整个上下文范围内利用跨事件信息，将多事件的并行抽取过程转化为表格生成任务。具体地，该表格提示模版的表头为多个事件提示模版的串接，而每一个行则对应包含了事件触发词和要素类型槽位的单个事件信息。在建模中，其拆分 RoBERTa 模型的不同网络层分别作为编码器和解码器，以替代 BART 模型，并通过设计结构化感知的自注意力机制并行获取不同事件内的要素类型表示，进而实现多事件要素抽取。

2.3.5 其他方法

最近，存在一些研究工作利用最优传输 (Optimal Transport)^[132]、图的链接预测 (Link Prediction)^[133] 和链推理 (Chain Reasoning)^[134] 等方式建模文档级别的事件要素抽取任务。此外，与事件检测任务相同，一些研究基于大语言模型评估了事件要素抽取的性能。例如，Li 等人^[87]在零样本场景下评测了 ChatGPT 的句子级别事件要素抽取性能，其输入由任务描述、给定事件类型对应的预定义事件要素类型集合、待抽取文本和输出格式要求等。Han 等人^[89]构建了和 Li 等人^[87]相似的输入模版，并引入 ICL 和 COT 技术，提升 ChatGPT 在句子级别事件要素抽取上的性能表现。然而，现有方法缺乏对更具挑战性的文档级别事件要素抽取的性能评估，本文将在章节5中给出 ChatGPT 在不同文本级别的事件要素抽取数据集上的详细实验结果。

2.4 现有研究存在的问题

前文分别基于神经网络架构的使用类型对事件检测和给定实体提及的事件要素抽取、以及基于建模方式对未给定实体提及的事件要素抽取进行了详细的梳理。然而，即使采用了不同的神经网络架构或者建模方式，这些方法均存在各自的共性问题如下：

(1) 数据不平衡依赖带来的共性识别事件能力的下降。现有事件检测数据集中存在事件的实例比例过小，其导致的数据不平衡依赖限制了现有基于不同网络架构的事件检测模型识别事件的能力，进而降低了整体检测的性能。虽然已有方法采用偏差损失函数^[50,61,63]或构建识别正实例/负实例^[135]的辅助任务进行消解，但均需要额外的超参数，以适应不同模型架构和不同程度的数据不平衡，缺乏通用性和扩展性。

(2) 实体提及带来的共性类型信息过度依赖。现有给定实体提及的事件要素抽取方法，尽管使用的网络架构不同，但均主要研究更好地建模实体提及与事件触发词间的上下文交互^[41,99,104]，从而利用实体提及带来的信息增益。然而，现有方法均忽略了实体提及的类型信息与要素类型呈现出的强关联数据特性，缺乏对于这种强关联性带来的潜在实体类型信息过度依赖问题的研究和统一解决方法。

(3) 不同级别文本中存在的共性跨事件依赖关系构建限制。现有适用于不同级别文本的事件要素抽取主要基于未给定实体提及的定义方式进行构建，其主要类型划分如章节2.3所总结。大多数方法均忽略了不同事件间的依赖关系，仅有少数方法^[122,125,130,131]在建模时利用了这种关联依赖进行性能增益。然而，这些方法受限于构建的提示模版质量或长范围文本等因素。因此，缺乏同时解决建模方式和跨事件信息构建与利用的通用事件要素抽取方法。

根据这些共性数据依赖挑战，本文研究通用性的应对方法。具体地，研究提升事件识别能力的通用技术，在第3章中提出分类器自适应知识蒸馏的事件检测方法，以消解数据不平衡依赖；研究减少建模过程产生的实体类型信息依赖的统一思路，在第4章中提出多角度实体类型过度依赖消解的事件要素抽取方法，以消解对实体类型的过度依赖；研究同时解决通用文本级别事件要素建模的内在局限和构建事件间的关联依赖，在第5章中提出基于分离-融合范式的多词元链接事件要素抽取模型，以实现跨事件信息的高效构建与利用。

第 3 章 基于分类器自适应知识蒸馏的数据不平衡依赖消解

事件检测作为事件抽取的基础子任务，其数据集中存在严重的数据不平衡现象，限制了不同事件检测模型系统的事件识别能力。针对该问题，现有方法主要利用人工设计的损失函数减少不平衡数据产生的负面影响。然而，这些方法特定于具体的模型架构和任务设计，依赖额外超参数，阻碍了其在不同数据不平衡情况下的扩展应用。为此，本章提出了一种基于分类器自适应知识蒸馏方法，以消解数据不平衡依赖问题。实验结果证明，该方法可以通用有效地提升不同事件检测模型的性能。此外，经本章实验验证，该方法可有效迁移到其他存在数据不平衡依赖的信息抽取任务中，如关系抽取等，证明了其良好的任务普适性。

3.1 引言

随着神经网络技术的快速发展，现有事件检测方法的性能取得了显著提升。然而，事件检测数据集中不存在事件的句子过多，导致这些方法普遍遭受数据不平衡依赖的困扰。例如，在 ACE2005 数据集的训练集部分，超过 76% 的句子文本不存在事件，由此产生的过量负实例（非事件触发词）限制了现有方法高效捕捉正实例（事件触发词）的特性，使得这些方法在识别事件时性能表现有待提升。为了消解此种依赖，一些研究工作致力于增强对于正实例特性的建模或减少过高的负实例比例产生的负面影响。例如，Dos 等人^[136]在训练中不使用负实例数据，并利用成对排序损失捕捉正实例的通用特性。Chen 等人^[50]在事件检测的训练损失中引入偏置参数，使得正实例和负实例以不同的权重进行训练。Ye 等人^[135]提出了一个多任务框架，通过增加一个额外的加权损失函数进行正负实例识别能力的学习，进而更好地捕捉正实例的特性。然而，这些方法依赖于人工设计的损失函数，其需要额外的超参数以适应不同数据集的不平衡程度差异，缺乏可扩展性。此外，这些方法只针对特定的模型架构和任务设计，缺乏通用性。因此，本章研究通用的事件检测框架，在不依赖额外超参数的情况下，自动消解数据不平衡问题。

与现有工作聚焦于调整正实例或负实例的训练权重不同，本章基于数据不平衡依赖问题影响事件检测性能的方式，研究减轻其性能影响的通用方法。根据 Ye 等人^[135]的研究工作，可知数据不平衡依赖导致现有模型易将正实例预测为负实例，反之亦

然。因此，数据不平衡依赖限制了现有模型识别正负实例的能力，进而降低了整体事件检测性能。

为了提升正负实例的识别性能，本章首先定义了句子级别识别信息，其表示给定实例的所在句子中是否存在事件。例如，给定文本“Mary died on Thursday in Memphis.”和实例“died”，已知“died”触发了 *Die* 事件，则其句子级别识别信息为在该句子中存在事件。若给定文本“Mary lives in Memphis.”和实例“lives”，已知其文本中不存在任何事件触发词，则其句子级别识别信息为在该句子中不存在事件。基于此，在事件检测中引入了该信息，并通过预实验评估了其在消解数据不平衡依赖导致的性能下降方面的优势。表3.1具体展示了在不同事件检测基线模型¹中引入句子级别识别信息后的性能变化，其中“TI”和“TC”分别表示事件触发词（正例）识别²和事件检测的F1性能指标，“TI+”和“TC+”分别表示在模型输入中引入句子级别识别信息后对应的“TI”和“TC”。可以观察到，该信息的引入使得不同事件检测基线模型的事件触发词识别性能均显著提升，且提升的识别性能进一步增强其在事件检测任务的整体性能。然而，句子级别识别信息需要根据标注的事件检测标签转化得到，因此无法在推理测试阶段进行利用。

表 3.1 ACE2005 数据集上引入句子级别识别信息在 F1 (%) 指标上的结果对比

模型	TI	TI+	TC	TC+
Bi-LSTM ^[44]	70.1	78.6	67.8	71.7
JMEE ^[57]	75.2	79.4	72.8	75.4
MOGANED ^[61]	75.9	79.8	73.4	76.6
DMBERT ^[70]	79.4	84.4	74.6	80.0
DMRoBERTa	80.1	85.5	75.5	81.9

为了解决该挑战，本章构建两种网络，分别采用两类不同的输入，其主要区别为是否引入了句子级别识别信息，并实现这两种网络间的知识引导，从而弥补句子级别识别信息在不同阶段的获取差异。因此，本章进一步考虑利用知识蒸馏技术解决。传统的知识蒸馏^[137]将教师网络输出的软标签分布作为学生网络的监督信息，以实现将知识从教师网络转移到学生网络。然而，事件检测数据集中负实例的比例过高，导致教师网络的软标签分布中包含的正实例信息较少，知识利用效率较低。

为此，本章提出了基于分类器自适应知识蒸馏的数据不平衡依赖消解方法（Clas-

¹DMRoBERTa 为本章结合 DMBERT 和 RoBERTa^[105] 而提出的一个变体模型。

²事件触发词识别在表述上等同于事件识别。

sifier-Adaptation Knowledge Distillation, CAKD), 有效提升不同事件检测基线模型的性能。首先, 根据训练实例的标签转化得到对应的句子级别识别信息, 并作为事件识别增强网络输入的一部分参与训练, 并在训练完成后固定其对应的分类器参数。然后, 移除输入中的句子级别识别信息以训练事件检测网络, 并共享固定的事件识别增强网络的分类器参数。通过设置额外的训练任务, 使得该分类器参数引导事件检测网络从原始文本中自动学习句子级别识别信息, 从而提高触发词识别任务性能并进一步增强事件检测的性能。本章的主要贡献如下:

1. 研究数据不平衡依赖问题影响事件检测性能的方式, 引入了句子级别识别信息消解其导致的性能下降。
2. 提出了一种分类器自适应知识蒸馏的事件检测通用方法, 自动捕捉句子级别识别信息, 以消解数据不平衡依赖。
3. 基于多种事件检测基线模型, 在 ACE2005 数据集上进行了充分的实验。实验结果表明, 本章提出的方法能够通用有效地提升不同模型的事件检测性能。
4. 将本章提出的方法应用到同样存在数据不平衡依赖问题的关系抽取任务上, 基于 TACRED 数据集的实验证明了其在信息抽取任务上良好的扩展性。

3.2 方法设计

遵循最近的研究工作, 给定 n 个词组成的句子实例 $\{w_1, w_2, \dots, w_n\}$, 事件检测任务旨在为每个词 w_i ($1 \leq i \leq n$) 分配预定义的事件类型标签 (包括特殊事件类型 *None*, 其表示对应的词不为事件触发词)。

本章提出的基于分类器自适应知识蒸馏的数据不平衡依赖消解方法主要由事件识别增强网络和事件检测网络组成, 可应用于不同的事件检测基线模型。图3.1展示了基于 DMBERT 基线模型的 CAKD 方法架构。接下来, 本章将作具体介绍。

3.2.1 事件识别增强网络

事件识别增强网络旨在在训练中引入句子级别识别信息, 该信息可根据训练集的事件标签直接获取, 不依赖于任何外部知识或资源。具体地, 首先根据句子中是否存在事件触发词, 将该句子标记为 “*Positive*” 或 “*Negative*”。如果句子中不存在任何事

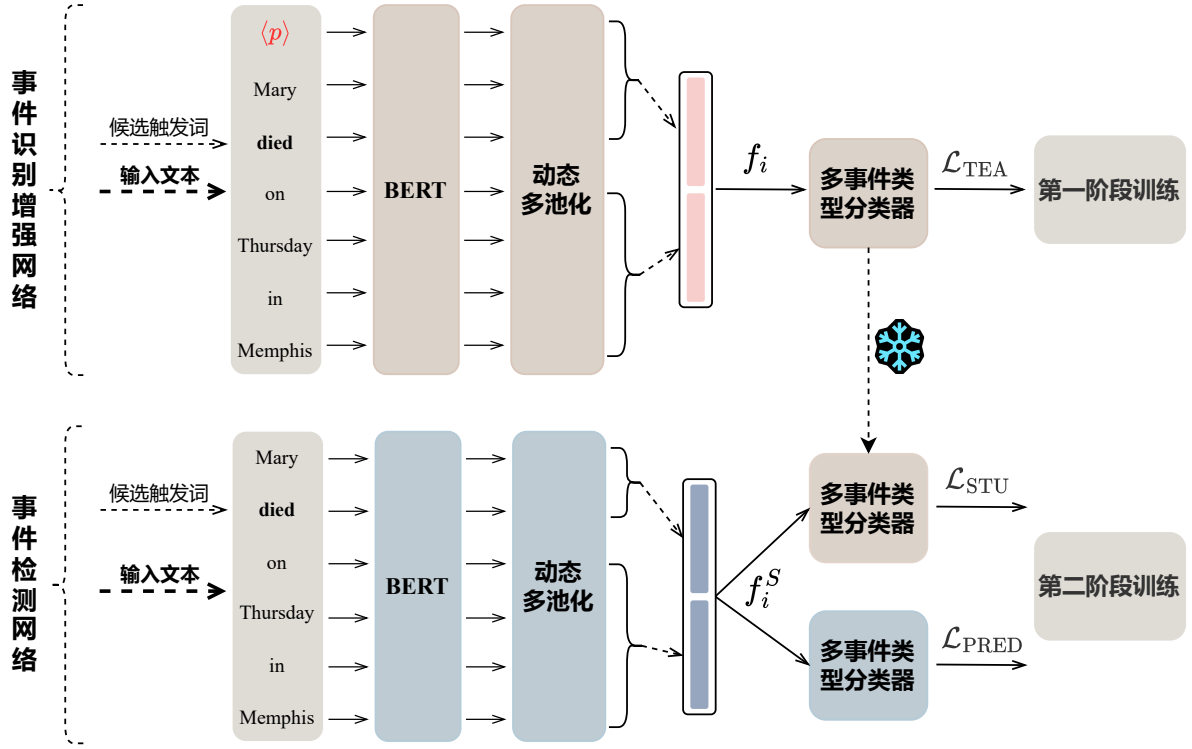


图 3.1 基于 DMBERT 的 CAKD 方法架构图

件触发词，则该句子被标记为 “*Negative*”。如果至少存在一个事件触发词，则该句子被标记为 “*Positive*”。然后，基于在事件识别增强网络中使用的事件检测基线模型的类型，采用两种不同的方式在输入部分引入句子级别识别信息并进行相应编码，分别为：

基于预训练词向量的事件检测模型 对于句子中任意一个词 w_i ($1 \leq i \leq n$)，首先拼接以下两类向量，表示为 \mathbf{x}_i ：

- **原始输入向量** 采用其事件检测基线模型的原始输入设置。不同基线模型的输入向量有所差别，通常由预训练词向量、实体类型向量、词性标注向量和位置向量中的全部或部分串接得到。
- **句子级别识别向量** 通过随机初始化得到两个向量 \mathbf{s}_P 和 \mathbf{s}_N ，用于提供句子级别识别信息。若词 w_i 所在句子被标记为 “*Positive*”，则 w_i 对应的句子级别识别向量为 \mathbf{s}_P ，否则其对应的句子级别识别向量为 \mathbf{s}_N 。

然后，使用事件检测基线模型对输入 $\mathbf{X}_{\text{teacher}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ 进行编码。为了描述方

便，本章选取 Bi-LSTM^[44] 作为基线模型，得到编码表示 $\mathbf{f}_i \in \mathbb{R}^{d_{out}}$ 如下：

$$\mathbf{f}_i = \left[\overrightarrow{LSTM}(\mathbf{x}_i); \overleftarrow{LSTM}(\mathbf{x}_i) \right] \quad (3.1)$$

基于预训练语言模型的事件检测模型 对于句子中任意一个词 $w_i (1 \leq i \leq n)$ ，首先根据句子级别识别信息在所在句子中插入可训练学习的特殊词元。具体地，若词 w_i 所在句子被标记为 “Positive”，则在句首插入特殊词元 $\langle p \rangle$ ，否则在句首插入特殊词元 $\langle n \rangle$ 。假设句首插入的特殊词元为 $\langle p \rangle$ ，且选取 DMBERT^[70] 作为事件检测基线模型，则首先使用 BERT^[103] 进行编码，得到特征表示如下：

$$\{\mathbf{s}_P, \mathbf{h}_1, \dots, \mathbf{h}_j, \dots, \mathbf{h}_n\} = \text{BERT}(\langle p \rangle, w_1, \dots, w_j, \dots, w_n) \quad (3.2)$$

其中 \mathbf{s}_P 为 $\langle p \rangle$ 对应的特征表示， $\mathbf{h}_j (1 \leq j \leq n)$ 为 $w_j (1 \leq j \leq n)$ 对应的特征表示。在此之后，使用一个动态多池化操作分段聚合编码后的特征表示：

$$[\mathbf{h}_{1,i}]_k = \max \{[\mathbf{s}_P]_k, [\mathbf{h}_1]_k \dots, [\mathbf{h}_i]_k\} \quad (3.3)$$

$$[\mathbf{h}_{i+1,n}]_k = \max \{[\mathbf{h}_{i+1}]_k \dots, [\mathbf{h}_n]_k\} \quad (3.4)$$

其中 $[\cdot]_k$ 表示一个向量的第 k 个元素值。最后，串接上述表示得到 w_i 的特征表示 $\mathbf{f}_i \in \mathbb{R}^{d_{out}}$ ：

$$\mathbf{f}_i = [\mathbf{h}_{1,i}; \mathbf{h}_{i+1,n}] \quad (3.5)$$

然后，对于不同类型的事件检测基线模型，将特征表示 \mathbf{f}_i 输入到多事件类别分类器，以计算概率分布 $p(w_i)$ 如下：

$$p(w_i) = \text{softmax}(\mathbf{W}_S \mathbf{f}_i + \mathbf{b}_S) \quad (3.6)$$

其中 $\mathbf{W}_S \in \mathbb{R}^{C \times d_{out}}$ 和 $\mathbf{b}_S \in \mathbb{R}^C$ 表示分类器中需要学习训练的参数，其中 C 为预定义的事件类型数目。假设给定的批数据 \mathcal{B} 中存在 K 个词，其交叉熵损失 \mathcal{L}_{TEA} 计算如下：

$$\mathcal{L}_{\text{TEA}} = - \sum^K \log p(e|w_i) \quad (3.7)$$

其中 $p(e|w_i)$ 表示事件识别增强网络将词 w_i 预测为正确的事件类型 e 的概率值。

3.2.2 事件检测网络

事件检测网络旨在建模事件检测任务，其使用原始的文本信息作为输入。然后，一方面基于固定的事件识别增强网络分类器参数设计辅助训练任务，引导事件检测网络自动从原始文本中捕获句子级别识别信息。同时，学习训练另一个分类器，以用于事件类型预测。

在事件检测网络输入部分，若使用基于预训练词向量的事件检测模型如 Bi-LSTM 进行编码，则移除 $\mathbf{X}_{\text{teacher}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ 中拼接的句子级别识别向量 \mathbf{s}_P 或 \mathbf{s}_N ，记作 $\mathbf{X}_{\text{student}} = (\mathbf{x}_1^S, \dots, \mathbf{x}_n^S)$ 。然后，利用另一个 Bi-LSTM^[44] 得到编码表示 $\mathbf{f}_i^S \in \mathbb{R}^{d_{out}}$ 如下：

$$\mathbf{f}_i^S = [\overrightarrow{LSTM}(\mathbf{x}_i^S); \overleftarrow{LSTM}(\mathbf{x}_i^S)] \quad (3.8)$$

若使用基于预训练语言模型的事件检测模型如 DMBERT，则首先利用另一个 BERT 编码不包含特殊词元 $\langle p \rangle$ 或 $\langle n \rangle$ 的原始文本：

$$\{\mathbf{h}_1^S, \dots, \mathbf{h}_j^S, \dots, \mathbf{h}_n^S\} = \text{BERT}(w_1, \dots, w_j, \dots, w_n) \quad (3.9)$$

在此之后，同样使用一个动态多池化操作分段聚合编码后的特征表示，并进一步串接得到 $\mathbf{f}_i^S \in \mathbb{R}^{d_{out}}$ ：

$$[\mathbf{h}_{1,i}^S]_k = \max \{ [\mathbf{h}_1^S]_k \dots, [\mathbf{h}_i^S]_k \} \quad (3.10)$$

$$[\mathbf{h}_{i+1,n}^S]_k = \max \{ [\mathbf{h}_{i+1}^S]_k \dots, [\mathbf{h}_n^S]_k \} \quad (3.11)$$

$$\mathbf{f}_i^S = [\mathbf{h}_{1,i}^S; \mathbf{h}_{i+1,n}^S] \quad (3.12)$$

然后，对于不同类型的事件检测基线模型，将特征表示 \mathbf{f}_i^S 输入到事件识别增强网络分类器中，计算概率分布 $p^S(w_i)$ 如下：

$$p^S(w_i) = \text{softmax}(\mathbf{W}_S \mathbf{f}_i^S + \mathbf{b}_S) \quad (3.13)$$

给定包含 K 个词的批数据 \mathcal{B} ，其交叉熵损失 \mathcal{L}_{STU} 计算如下：

$$\mathcal{L}_{\text{STU}} = - \sum^K \log p^S(e|w_i) \quad (3.14)$$

其中 $p^S(e|w_i)$ 表示将词 w_i 预测为正确的事件类型 e 的概率值。同时，将特征表示 \mathbf{f}_i^S 输入到另一个多事件类别分类器中，得到概率分布 $p^E(w_i)$ 如下：

$$p^E(w_i) = \text{softmax}(\mathbf{W}_E \mathbf{f}_i^S + \mathbf{b}_E) \quad (3.15)$$

其中 $\mathbf{W}_E \in \mathbb{R}^{C \times d_{\text{out}}}$ 和 $\mathbf{b}_E \in \mathbb{R}^C$ 表示该分类器中需要训练的参数。给定包含 K 个词的批数据 \mathcal{B} ，则其交叉熵损失 $\mathcal{L}_{\text{PRED}}$ 计算如下：

$$\mathcal{L}_{\text{PRED}} = - \sum^K \log p^E(e|w_i) \quad (3.16)$$

其中 $p^E(e|w_i)$ 表示事件检测网络将词 w_i 预测为正确的事件类型 e 的概率值。

3.2.3 训练和推理

在训练阶段，首先基于损失 \mathcal{L}_{TEA} 训练事件识别增强网络。在完成事件识别增强网络训练后，固定其分类器参数 \mathbf{W}_S 和 \mathbf{b}_S ，并基于如下损失训练事件检测网络：

$$\mathcal{L}_{\text{OVERALL}} = \mathcal{L}_{\text{STU}} + \mathcal{L}_{\text{PRED}} \quad (3.17)$$

其中最小化 \mathcal{L}_{STU} 能够在训练过程中通过共享固定的事件识别增强网络参数 \mathbf{W}_S 和 \mathbf{b}_S ，引导事件检测网络自动从文本中捕获句子级别识别信息。同时，最小化 $\mathcal{L}_{\text{PRED}}$ 使得参数 \mathbf{W}_E 和 \mathbf{b}_E 聚焦于事件类型的分类建模。在推理阶段，使用 $p^E(w_i)$ 获取 w_i 对应的事件类型。

3.3 实验评估

3.3.1 实验设置

1. 数据集 对于事件检测任务，本章使用了包含 599 个文档和 33 种事件类型的 ACE2005 语料库^[138]，该语料库数据来源为报纸、新闻专线数据和广播新闻，由自动内容提取项目（Automatic Content Extraction, ACE）完成标注。其存在英文、中文和阿拉伯文版本，遵循最近的研究工作^[41,54]，本章使用英文版本（以下简称 ACE2005 数据集），并分别使用了 529、40 和 30 个文档作为训练集、验证集和测试集。此外，在训练集中，仅 3325 个句子存在至少一种事件，而剩余的 10594 个句子中不存在任何事件，因此其数据不平衡问题显著。

2. 评价指标 本章实验遵循事件检测任务的标准评估^[37,41]。如果事件触发词和对应的事件类型均与数据集标注的信息完全相同，则认定该事件触发词被正确检测，本章模型采用包括特殊事件类型 *None*（空）的多分类方式同时识别事件触发词和预测对应的事件类型。基于此，本章使用不区分事件类别的微平均方式计算准确率（P）、召回率（R）和 F1 值（F1）作为评估指标，具体如下：

$$P = \frac{TP}{TP + FP} \quad (3.18)$$

$$R = \frac{TP}{TP + FN} \quad (3.19)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3.20)$$

其中 TP 为检测出的正确事件触发词数目， FP 为检测出的错误事件触发词数目，而 FN 则表示未被检测出的事件触发词数目。

3. 基线方法 本节选择以下方法参与实验比较，这些方法均可基于不同的事件检测基线模型进行性能评估，具体包括：

- **基线模型**：仅利用 $\mathcal{L}_{\text{PRED}}$ 训练事件检测网络，其退化为对应的事件检测模型。
- **基线模型 + Sim-CAKD**：本章提出的 CAKD 方法的简化版本。具体地，事件检测网络不额外学习训练新的多事件类别分类器，而直接利用事件识别增强网络的分类器进行事件类型的建模。因此，公式 3.17 中 $\mathcal{L}_{\text{OVERALL}} = \mathcal{L}_{\text{STU}}$ 。在推理阶段，同样利用事件识别增强网络的分类器获取事件检测任务的结果。

- **基线模型 +MTL**: 遵循 Ye 等人^[135] 消解关系抽取中数据不平衡依赖的思路, 通过共享网络编码, 增加额外的事件触发词识别任务, 采用多任务学习方式进行建模。下文将介绍关键细节。

对于基线模型 +MTL 方法, 假设给定句子中任意一个词 w_i ($1 \leq i \leq n$), 根据选定的事件检测基线模型编码得到特征表示为 \mathbf{f}_i^M 。在事件触发词识别任务中, 计算对应的二分类损失 \mathcal{L}_{ID} 如下:

$$p^I(w_i) = \sigma(\mathbf{W}_I \mathbf{f}_i^M + \mathbf{b}_I) \quad (3.21)$$

$$\mathcal{L}_{ID} = - \sum^K (L_i \log p^I(w_i) + (1 - L_i) \log(1 - p^I(w_i))) \quad (3.22)$$

其中 $\mathbf{W}_I \in \mathbb{R}^{1 \times d_{out}}$ 和 $\mathbf{b}_I \in \mathbb{R}^1$ 分别表示该二分类器的训练参数, $\sigma(\cdot)$ 为 sigmoid 激活函数。当词 w_i 为事件触发词, 则 L_i 设置为 1, 否则设置为 0。而在事件检测任务中, 将特征表示 \mathbf{f}_i^M 输入到多事件类别分类器中, 得到对应的事件检测任务损失 \mathcal{L}_{CL} :

$$p^C(w_i) = \text{softmax}(\mathbf{W}_C \mathbf{f}_i^M + \mathbf{b}_C) \quad (3.23)$$

$$\mathcal{L}_{CL} = - \sum^K \log p^C(e|w_i) \quad (3.24)$$

其中 $\mathbf{W}_C \in \mathbb{R}^{C \times d_{out}}$ 和 $\mathbf{b}_C \in \mathbb{R}^C$ 表示该分类器中需要训练的参数。基于损失 \mathcal{L}_{ID} 和 \mathcal{L}_{CL} , 则可得该多任务学习方法的总训练损失:

$$\mathcal{L}_M = \alpha \mathcal{L}_{ID} + (1 - \alpha) \mathcal{L}_{CL} \quad (3.25)$$

其中 α 为权重超参数。

在本章所提的 CAKD 方法和上述基线方法中, 不仅使用 Bi-LSTM^[44], 还利用以下事件检测基线模型参与性能验证: (1) Liu 等人提出的 JMEE^[57], 其融合了自注意力机制和 GCN, 以编码句法结构信息和捕捉不同事件间的关联。(2) Yan 等人提出的 MOGANED^[61], 其使用 GAT 网络聚合依存句法树中的多跳句法信息, 以有效建模非邻接依存关系。(3) Wang 等人提出的 DMBERT^[70], 其将 DMCNN^[41] 模型中的 CNN 替换成 BERT, 并保留动态多池化操作。(4) 本章提出的 DMRoBERTa, 其将 DMBERT 中的 BERT 替换成对应的改进版预训练语言模型 RoBERTa^[105], 其他部分保持不变。

4. 实验配置 对于基于预训练词向量的事件检测模型，各自原始输入向量保持不变，一般由预训练词向量、随机初始化的实体类型标签向量、位置向量和词性标注向量等组成，而其句子级别识别向量的维度均设置为 300。此类事件检测模型的实验服务器配置均为：CPU 型号 Intel(R) Xeon(R) Gold 5218，主频 2.30GHz；内存 128G；GPU 计算卡型号 NVIDIA GeForce RTX 1080Ti，显存 11G。对于基于预训练语言模型的事件检测模型，分别加载 HuggingFace 上发布的“google-bert/bert-base-uncased”文件³和“FacebookAI/roberta-base”文件⁴作为编码使用的 BERT 和 RoBERTa 预训练模型参数。学习率均设置为 $5e-5$ ，并使用 AdamW 进行训练优化。此类事件检测模型的实验服务器配置均为：CPU 型号 Intel(R) Xeon(R) Gold 5320，主频 2.20GHz；内存 256G；GPU 计算卡型号 NVIDIA GeForce RTX 3090，显存 24G。此外，对于所有事件检测模型，其第一阶段训练均采用早停策略决定训练的轮次。当事件识别增强网络连续 3 轮在验证集上的 F1 指标没有出现提升，则结束该阶段的训练。在第二阶段训练中，若使用的基线模型为 Bi-LSTM，则事件检测网络的轮次设置为 30。对于其他事件检测基线模型，其事件检测网络采用对应事件检测模型的训练轮次或策略⁵。本章实验均基于 Pytorch^[139] 框架完成。

3.3.2 性能结果

表3.2展示了在 ACE2005 数据集上的事件检测实验结果，其中 * 表示重新运行其发布的源代码得到的 F1 指标值。可以看出，本章所提的 CAKD 方法可以显著增强所有事件检测基线模型的 F1 指标，其涵盖了基于预训练词向量和预训练语言模型两种不同的类型。其中，CAKD 方法能够大幅提升基于预训练语言模型的事件检测基线模型，分别在 DMBERT 和 DMRoBERTa 的 F1 指标上取得了 4.0% 和 3.8% 的性能提升。进一步，可以观察到对于所有的事件检测基线模型，CAKD 方法在 F1 性能指标上都超过了相应的简化方法 Sim-CAKD，其证明了将预测事件类型和学习句子级别识别信息的训练参数进行分隔，能提升对应基线模型的性能。对于 MTL 方法，其能够提升 Bi-LSTM、DMBERT 和 DMRoBERTa 等基线模型的 F1 性能表现，但其性能提升幅度均不如 CAKD 方法。而且，在使用了 GNN 网络架构的 JMEE 和 MOGANED 模型上，MTL 方法降低了其对应的 F1 指标结果。因此，相比于 MTL 方法，本章所提的 CAKD

³<https://huggingface.co/google-bert/bert-base-uncased>

⁴<https://huggingface.co/FacebookAI/roberta-base>

⁵DMRoBERTa 参照 DMBERT 的训练轮次设置。

方法表现出在不同类型基线模型上的性能优势和通用性。此外，可以发现 CAKD 方法在不同基线模型上均主要通过显著增强预测的召回率提高整体的 F1 指标。该发现可以归因于本章提出的 CAKD 方法能够提升基线模型自动从文本中学习句子级别识别信息的能力，以更好地获取给定实例的正负实例情况，从而在事件检测中能以更高的置信水平预测出比例较小的正例数据。与之相反，基线模型本身在缺乏获取句子级别识别信息的能力时，倾向于避免将更多的给定实例预测为正例，导致较低的召回率。

表 3.2 ACE2005 数据集上的事件检测性能结果

模型	P(%)	R(%)	F1(%)
Bi-LSTM	68.6	67.0	67.8
Bi-LSTM+MTL	66.4	70.6	68.4
Bi-LSTM+Sim-CAKD	67.2	71.5	69.3
Bi-LSTM+CAKD	68.2	71.7	69.6
JMEE	75.3	70.5	72.8*
JMEE+MTL	71.9	65.3	68.5
JMEE+Sim-CAKD	74.0	73.1	73.6
JMEE+CAKD	75.0	72.4	73.7
MOGANED	81.0	67.0	73.4*
MOGANED+MTL	76.5	66.1	70.9
MOGANED+Sim-CAKD	79.4	69.1	73.9
MOGANED+CAKD	78.7	72.3	75.4
DMBERT	77.6	71.8	74.6
DMBERT+MTL	75.3	77.7	76.5
DMBERT+Sim-CAKD	76.0	81.2	78.5
DMBERT+CAKD	75.6	81.9	78.6
DMRoBERTa	71.1	80.5	75.5
DMRoBERTa+MTL	76.4	76.0	76.2
DMRoBERTa+Sim-CAKD	71.6	85.7	78.0
DMRoBERTa+CAKD	73.7	85.9	79.3

3.3.3 与基于预训练语言模型的基线性能比较

结合本章研究工作、DMBERT^[70] 和 RoBERTa^[105] 的发表时间，本章选取基于预训练语言模型且性能表现优秀的事件检测模型与 CAKD 方法作性能对比，具体包括：(1) Lai 等人提出的 GatedGCN^[59]，其基于 BERT 进行编码，并利用依存句法树在 GCN 中构建候选触发词特征表示对其他单词信息的过滤筛选。(2) Tong 等人提出的 EKD^[78]，其基于 BERT 进行编码，并构建知识蒸馏框架，利用从 WordNet 获取的

开放域触发词知识提升对训练集中稀疏和未出现事件触发词的识别。(3) Veyseh 等人提出的 GPTEDOT^[83], 其基于 BERT 进行编码, 并利用 GPT-2 为事件检测生成额外的训练数据。此外, 考虑到 ChatGPT 在多种自然语言处理任务上表现优异, Han 等人^[89] 利用情境学习 (In-Context Learning, ICL)^[90] 构建 5-shot ICL 提示模版作为 ChatGPT 的输入, 以生成 ACE2005 数据集上所有测试样本的事件检测结果。本章展示其 F1 指标作为性能参考。

表3.3展示的结果表明, DMRoBERTa+CAKD 在 F1 指标上获得了最优性能, 而 DMBERT+CAKD 也在 F1 指标上超越了 GatedGCN 和 EKD 模型。此结果表明, 虽然只采用了简单的预训练语言模型编码和池化操作, 但在应用了本章提出的 CAKD 方法后, 与上述引入了外部知识资源如依存句法信息、WordNet 和额外生成数据的事件检测模型相比, 仍表现出具有竞争力的实验结果, 进一步验证了 CAKD 方法的性能效率。另外, 可以观察到 ChatGPT 的输入受限于少样本场景, 在事件检测任务上性能表现不佳。

表 3.3 ACE2005 数据集上基于预训练语言模型的事件检测性能结果对比

模型	P(%)	R(%)	F1(%)
ChatGPT	-	-	27.3
GatedGCN	78.8	76.3	77.6
EKD	79.1	78.0	78.6
GPTEDOT	82.3	76.3	79.2
DMBERT+CAKD	75.6	81.9	78.6
DMRoBERTa+CAKD	73.7	85.9	79.3

3.3.4 事件触发词识别能力分析

本节比较 CAKD 方法和相应的基线模型在事件触发词识别任务上的 F1 性能指标。如表3.4所示, 在不同基线模型上引入 CAKD 方法, 其识别事件触发词的性能均显著提升, 证明其能够通用地消解数据不平衡依赖问题带来的正负例识别能力的下降。

3.3.5 数据不平衡程度分析

本节分析在不同程度数据不平衡场景下, CAKD 方法、MTL 方法以及相应的基线模型方法在事件检测任务上的性能变化。具体地, 在分析实验中采用 DMRoBERTa

表 3.4 ACE2005 数据集上的事件触发词识别任务 F1 (%) 指标结果

模型	基线模型	基线模型 +CAKD
Bi-LSTM	70.1	71.6
JMEE	75.2	77.2
MOGANED	75.9	78.7
DMBERT	79.4	82.8
DMRoBERTa	80.1	84.3

作为基线模型，并依次从 ACE2005 数据集训练集中随机删减等间隔数量存在事件的句子数据，以构建存在不同数据不平衡程度的训练数据。如图3.2所示，随着数据不平衡程度的增加，不同方法的 F1 指标均呈现下降趋势，但 DMRoBERTa+CAKD 的下降最为平缓，在不同场景下的 F1 指标均显著优于同样应用了数据不平衡依赖消解的 DMRoBERTa+MTL，验证了本章提出的 CAKD 方法在消解不同数据不平衡依赖的优势。进一步，可以观察到 DMRoBERTa 相比于应用了数据不平衡依赖消解的 DMRoBERTa+CAKD 和 DMRoBERTa+MTL，在更加极端的不平衡数据场景下性能劣势显著增加，表明了数据不平衡依赖消解的必要性。

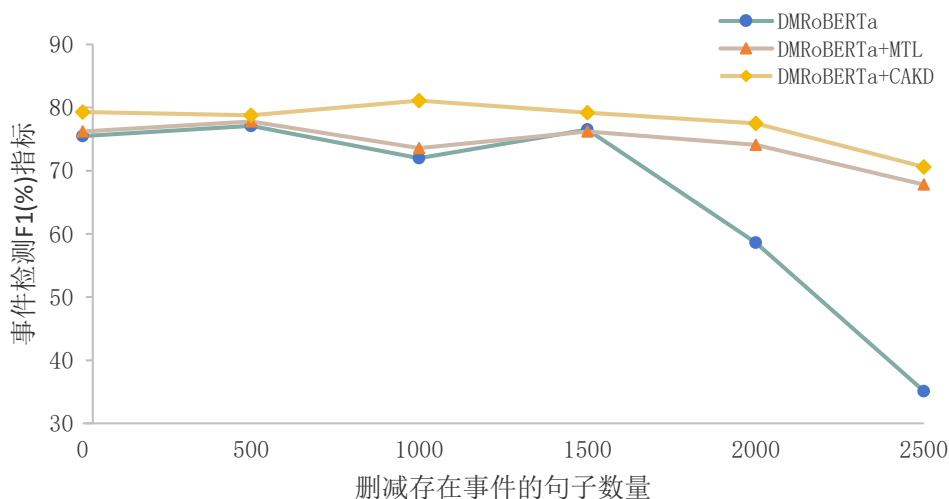


图 3.2 不同程度数据不平衡下不同方法在 F1 (%) 指标上的结果对比

3.3.6 案例研究

本节基于 ACE2005 数据集中的具体实例进行案例分析。表3.5展示了本章提出的 CAKD 方法和对应的基线模型在具体实例上的事件检测结果，其中粗体单词表示事件触发词，“频率”为对应事件类型的触发词占总事件触发词数目的比例。可以看到，

在实例 (1) 中, 虽然 “acquitted” 较为明显地触发了 *Acquit* (宣判无罪) 事件类型, 基线模型仍由于该事件类型的稀疏性而将 “acquitted” 预测为非事件触发词。类似地, 在实例 (2) 中, 由于 “rally” 为罕见的事件触发词, 基线模型无法识别出 “rally” 触发了 *Demonstrate* (认定) 事件类型。与之相比, CAKD 方法能够基于文本内容建模句子级别识别信息, 具备了更好地预测句子中是否存在事件的能力, 相应提升了对于罕见事件类型和罕见事件触发词的识别, 从而给出了正确的检测结果。

表 3.5 ACE2005 数据集上的案例分析

实例	正确事件标签	基线模型	基线模型 +CAKD	频率
(1) The Pakistani supreme court last year acquitted Ayub Masih.	Acquit	None	Acquit	0.01%
(2) Judge Shahid Rafiq..., found Ranjha Masih guilty of defiling Koranic verses during a protest rally by the minority Christian community in 1998.	Demonstrate	None	Demonstrate	1.47%

3.3.7 其他信息抽取任务上的扩展实验

为了进一步验证 CAKD 方法在其他任务上的良好扩展性, 本章将 CAKD 方法迁移到信息抽取的另一关键任务——关系抽取上。在该任务中, 数据不平衡问题同样严重存在。例如, 在广泛使用的关系抽取数据集 TACRED^[140] 上, 约 80% 的实体对中存在关系。因此, 本章在 TACRED 数据集上比较 CAKD 方法和章节 3.3.1 中介绍的基线方法的性能结果。同样地, 在这些方法中, 使用不同的基线模型, 包括 Bi-LSTM、Bi-GRU、PA-LSTM^[140]、C-GCN^[141]、C-AGGCN^[142] 和 GDPNet^[143]。其中, 在 CAKD 和 Sim-CAKD 方法中, 采用和事件检测任务相似的策略, 构建两个网络, 在其中一个网络的输入中引入句子级别识别信息, 并分别利用这些基线模型得到两个网络对应的编码表示, 以替代公式 3.1 和公式 3.8 中的 f_i 和 f_i^S , 而方法中的其他部分则保持不变。

表 3.6 展示了在 TACRED 数据集上的关系抽取实验结果, 其中 * 表示重新运行其发布的源代码得到的 F1 指标值。可以看出, 将本章提出的 CAKD 方法迁移到不同的基线模型上, 均提升了在关系抽取任务上的 F1 性能结果, 有效地证明了本章所提方法的良好任务扩展性。同样地, 其对应的简化方法 Sim-CAKD 也提升了不同关系抽取模型的整体性能, 但其平均性能表现与 CAKD 方法存在一定差距, 验证了与章节 3.3.2 类似的结论, 即分隔预测关系类型和学习句子级别识别信息的训练参数, 能提

升对应基线模型的性能。对于 MTL 方法，其能够提升 Bi-LSTM 和 Bi-GRU 模型的 F1 指标，但在其他关系抽取基线模型上表现不佳，进一步证明了该方法只适用于部分类型的网络架构模型。

表 3.6 TACRED 数据上的关系抽取性能结果

模型	P(%)	R(%)	F1(%)
Bi-LSTM	65.7	58.9	62.1
Bi-LSTM+MTL	63.4	62.8	63.1
Bi-LSTM+Sim-CAKD	63.8	62.3	63.0
Bi-LSTM+CAKD	66.2	62.0	64.1
Bi-GRU	70.1	59.1	64.2
Bi-GRU+MTL	64.0	65.8	64.9
Bi-GRU+Sim-CAKD	66.7	64.5	65.6
Bi-GRU+CAKD	67.4	64.1	65.7
PA-LSTM	65.7	64.5	65.1
PA-LSTM+MTL	66.2	63.4	64.8
PA-LSTM+Sim-CAKD	69.1	61.9	65.3
PA-LSTM+CAKD	67.1	66.4	66.7
GCN	69.8	59.0	64.0
GCN+MTL	67.4	59.9	63.4
GCN+Sim-CAKD	68.1	62.5	65.2
GCN+CAKD	68.5	61.7	64.9
C-GCN	69.9	63.3	66.4
C-GCN+MTL	69.8	62.2	65.8
C-GCN+Sim-CAKD	70.4	63.5	66.8
C-GCN+CAKD	69.7	65.0	67.3
C-AGGCN	71.8	66.4	67.7*
C-AGGCN+MTL	69.2	64.1	66.6
C-AGGCN+Sim-CAKD	71.6	64.7	68.0
C-AGGCN+CAKD	70.7	65.5	68.0
GDPNet	72.0	69.0	70.5
GDPNet+MTL	69.9	66.9	68.4
GDPNet+Sim-CAKD	71.0	70.2	70.6
GDPNet+CAKD	71.3	70.6	70.9

3.4 本章小结

本章首先研究了数据不平衡依赖对于事件检测性能的影响方式，并以此构建了句子级别识别信息进行辅助消解。进一步，本章提出了一种新的分类器自适应知识蒸馏

的事件检测通用方法，其在事件识别增强网络输入部分通过学习的向量或特殊词元引入句子级别识别信息，并共享分类器参数引导事件检测网络自动捕捉该句子级别识别信息，实现事件检测的性能增强。本章在 ACE2005 数据集上使用不同事件检测基线模型进行了实验评估，评估结果表明本章提出的方法均能提升相应的性能结果。此外，本章通过实验验证了提出的方法能够自动适应不同数据不平衡程度和迁移到其他信息抽取任务上的能力，有效验证了提出的方法的多维度通用性。

第 4 章 基于多角度对比学习的实体类型过度依赖消解

事件要素抽取任务是事件抽取的关键组成部分，成为了限制事件抽取整体性能的瓶颈。正如章节2.2所述，现有事件要素抽取的建模存在两种不同的定义，其不同定义的一个显著区别为是否将实体提及识别与事件抽取进行分隔，本章节研究在已经给定实体提及的情况下进行的事件要素抽取。

基于给定的实体提及，大多数事件要素模型在建模过程中利用了其正面信息增益，而忽视了实体类型和要素类型间的强关联对性能产生的负面影响。为了更好地研究该负面影响，本章节首先定义实体类型过度依赖并通过实验验证其降低了多种事件要素抽取模型的整体性能。进一步，本章提出了一种基于多角度对比学习的实体类型过度依赖消解方法解决该问题。实验评估表明，提出的方法可以通用有效地消解不同事件要素抽取模型对于实体类型信息的过度依赖。

4.1 引言

作为基于实体的抽取任务，事件要素抽取^[41,99,104]主要研究在给定上下文语义信息的情况下如何更好地建模实体提及和事件触发词间的复杂交互。进一步，结合现有的事件要素抽取方法和以 ACE2005 为代表的事件要素抽取数据集，可以观察到以下两个数据特征：（1）实体类型为事件要素抽取的实体提及提供了主要信息来源。具体地，按照 Peng 等人^[144]在关系抽取任务中的类似设置，将 ACE2005 数据集中的实体提及替换成对应的实体类型，并以此训练一个事件要素抽取任务的代表性模型 DMBERT^[70]。实验结果表明 DMBERT 取得了和替换前几乎相同的性能结果。（2）大多数实体类型在事件中仅以少数特定要素类型参与^[36]。例如，ACE2005 数据集中出现了 43 种实体类型，但有超过 74% 的实体类型以少于 5 种的要素类型参与到事件中。

基于上述两个数据特征，可以得知实体类型信息对于事件要素抽取任务是至关重要的，且与要素类型密切相关。现有事件要素抽取方法基于实体类型信息的重要性，更多地关注其正面信息增益，而忽略了实体类型和要素类型间的相关性带来的潜在负面影响。如表4.1所示，给定目标抽取文本“A rocket holding the first of two Mars rovers blasted off Tuesday on a seven-month voyage to the planet.”和 *Transport*（运输）事件的事件触发词“voyage”，一个性能优良的事件要素抽取模型需要识别出实体类型为

Landing 的实体提及 “two Mars rovers” 没有以任何要素类型参与到该事件中，即对应的事件要素类型为空。然而，由于其他实体类型为 *Landing* 的实体提及在训练数据中较高频率地以要素类型 *Vehicle* 出现在事件中，正如表4.1中所示的句1和句2，使得现有事件要素抽取模型容易受到实体类型 *Landing* 的误导，将实体提及 “two Mars rovers” 的要素类型错误地预测为 *Vehicle*。

表 4.1 实体类型过度依赖示例

实例	实体类型	要素类型标签	要素类型预测
例: A rocket holding the first of two Mars rovers blasted off Tuesday on a seven-month voyage to the planet.	Landing	None	Vehicle ×
句 1: The rovers' landing sites, on opposite sides of the planet, were chosen for their likelihood of holding evidence of water.	Landing	Vehicle	Vehicle ✓
句 2: The bus was ripped to shreds while traveling between a residential area and Haifa university.	Landing	Vehicle	Vehicle ✓

为此，本章首先定义了实体类型过度依赖并通过预实验评估证明了不同事件要素抽取基线模型均受到实体类型过度依赖问题的影响，使得整体性能出现下降。由于实体类型过度依赖是由建模时对实体类型信息的严重依赖导致的，消解该问题可以考虑提升模型学习除实体类型外的更多语义信息的能力。

广泛应用于计算机视觉^[145,146]和自然语言处理^[147,148]的对比学习技术是解决这一问题的有效方法，尤其是监督对比学习技术^[149–151]，其可将属于同一类型实例（称之为正样本）的特征表示距离拉近，属于不同类型实例（称之为负样本）的特征表示距离推远。然而，在事件要素抽取任务中，直接应用传统的监督对比学习技术只能根据要素类型信息进行特征学习，而无法同时考虑到实体类型信息。

为了解决上述提到的挑战，本章提出基于多角度对比学习的实体类型过度依赖消解方法。在不同的事件要素抽取模型的基础上，提出了通用的多角度实体类型过度依赖消解（Multi-view Entity Type Overdependency Reduction, METOR）模型，其中多角度指正样本角度和负样本角度。基于正样本角度，本章提出了一种加权-选择的对比学习方法，以提升具有相同要素类型但具有不同实体类型的实例特征表示间的相似性。基于负样本角度，本章提出了一种伪正对比学习方法来降低具有不同要素类型但具有相同实体类型的实例特征表示间的相似性。此外，为了提升上述两种对比学习方

法间的协作能力，在模型训练过程中提出了一种高效的循环训练策略。本章的主要贡献概括如下：

- 定义了实体类型过度依赖问题并通过预实验验证了不同事件要素抽取基线模型均受到该问题的影响而导致整体性能下降。
- 针对实体类型过度依赖问题，本章提出了基于多角度对比学习的实体类型过度依赖消解方法。该方法可以应用于不同基线模型中，由两种对比学习方法和一种循环训练策略组成，其中循环训练策略能够提升不同对比学习方法间的训练协作。
- 在广泛使用的 ACE2005 数据集上进行了大量实验。实验结果表明，本章提出的方法能够通用有效地消解不同基线模型的实体类型过度依赖问题，并基于预训练架构取得当前最优性能。

4.2 实体类型过度依赖定义与验证

如上一章节所述，实体类型为事件要素抽取任务提供了关键信息且和要素类型信息强关联。为了更好地研究实体类型和要素类型之间的关联性，本节首先形式化地定义实体类型过度依赖，并通过预实验评估其对于事件要素抽取性能的影响。

4.2.1 定义

一般来说，事件要素抽取模型由两个主要部分组成：编码器和分类器。其中编码器将实例编码成特征表示，而分类器识别实例的特征表示和不同要素类型的表示间的相似性。由于实体类型构成了实体提及的主要信息，则以实体提及为重要组成部分的特征编码表示依赖于实体类型。假设 $G(r_k)$ 是由要素类型为 r_k 的实例构成的群组， $C(e_i, r_k)$ 是由实体类型为 e_i 和要素类型为 r_k 的实例构成的簇，而 $\mathbf{h}_{r_k}^M$ 和 \mathbf{h}_{e_i, r_k}^M 分别是 $G(r_k)$ 和 $C(e_i, r_k)$ 由给定事件要素抽取模型 M 的编码器编码得到的表示，则簇级别的实体类型依赖定义如下：

定义 1 (实体类型依赖). 给定 $C(e_i, r_k)$ 和由 M 编码得到的对应编码表示 \mathbf{h}_{e_i, r_k}^M ，如果存在 $C(e_i, r_l)$ 和 $C(e_j, r_k)$ ($i \neq j, l \neq k$) 满足如下条件，则称 $C(e_i, r_k)$ 依赖于实体类型 e_i ：

$$\mathbf{h}_{e_i, r_k}^M \cdot \mathbf{h}_{e_i, r_l}^M > \mathbf{h}_{e_i, r_k}^M \cdot \mathbf{h}_{e_j, r_k}^M \quad (4.1)$$

在定义1中, 如果 $C(e_i, r_k)$ 和 $C(e_i, r_l)$ 之间的相似性比 $C(e_i, r_k)$ 和 $C(e_j, r_k)$ 之间的相似性高, 则认定 $C(e_i, r_k)$ 满足实体类型依赖定义。该定义阐述了由于实体类型 e_i 和要素类型 r_k 的关联性使得 $C(e_i, r_k)$ 与群组 G_{r_l} 中的 $C(e_i, r_l)$ 的特征表示间的相似性反而比其与同在一个群组中的 $C(e_j, r_k)$ 的特征表示间的相似性更高的现象。

为了进一步评估这种实体类型依赖是否影响事件要素抽取模型的性能, 需要定义簇之间的语义不一致性。Ohashi 等人^[152] 研究发现如果实例特征表示的相似性与其类别标签不一致, 则相应的分类器更加容易出现预测错误。因此, 簇级别的语义不一致性可以定义如下:

定义 2 (语义不一致). 给定 $C(e_i, r_k)$ 和由 M 编码得到的对应编码表示 \mathbf{h}_{e_i, r_k}^M , 如果存在 $G(r_l)$ ($l \neq k$) 满足如下条件, 则称 $C(e_i, r_k)$ 存在语义不一致:

$$\mathbf{h}_{e_i, r_k}^M \cdot \mathbf{h}_{r_l}^M > \mathbf{h}_{e_i, r_k}^M \cdot \mathbf{h}_{r_k}^M \quad (4.2)$$

在定义2中, 虽然 $C(e_i, r_k)$ 中实例的要素类型是 r_k 而不是 r_l , 但 $C(e_i, r_k)$ 与群组 $G(r_k)$ 的特征表示间的相似性比其与群组 $G(r_l)$ 的特征表示间的相似性低。然而, 理想情况下 $C(e_i, r_k)$ 应该与自身所属的群组相似性更高, 否则将出现语义不一致现象, 从而使得编码器 M 对应的分类器更容易预测出错。因此, 语义不一致性可用于确定存在实体类型依赖的簇是否影响事件要素抽取模型的性能。基于此, 定义簇级别的实体类型过度依赖如下:

定义 3 (实体类型过度依赖). 给定 $C(e_i, r_k)$ 和由 M 编码得到的对应编码表示 \mathbf{h}_{e_i, r_k}^M , 如果依次满足如下条件, 则称 $C(e_i, r_k)$ 存在实体类型过度依赖:

- (a) 存在 $C(e_i, r_l)$ 和 $C(e_j, r_k)$ ($i \neq j, l \neq k$) 使得 $C(e_i, r_k)$ 满足定义1。
- (b) 存在 $G(r_l)$ ($l \neq k$) 使得 $C(e_i, r_k)$ 满足定义2。

在定义3中, 如果簇 $C(e_i, r_k)$ 满足实体类型依赖的定义且该依赖导致语义不一致, 进而导致分类器容易预测错误^[152], 则认为该簇过度依赖实体类型。因此, 实体类型过度依赖本质上为实体类型依赖导致的语义不一致。以图4.1为例, 若存在 $C(e_1, r_2)$ 和 $C(e_5, r_1)$ 使得 $C(e_1, r_1)$ 满足定义1, 且存在 $G(r_2)$ 使得 $C(e_1, r_1)$ 满足定义2, 则可以判定 $C(e_1, r_1)$ 对实体类型 e_1 过度依赖。

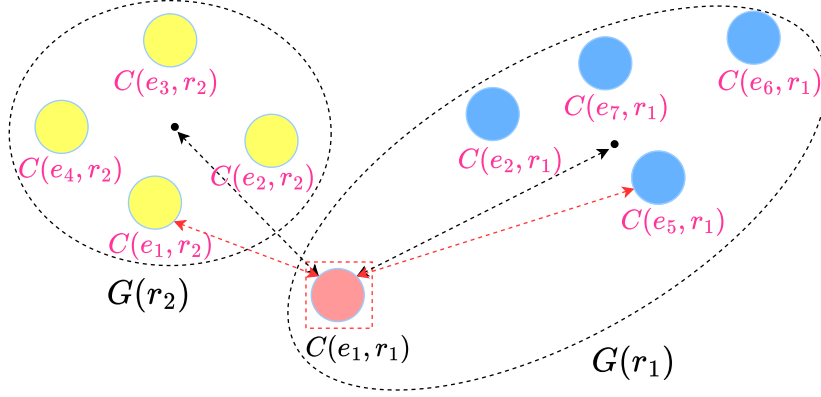


图 4.1 实体类型过度依赖示例图

4.2.2 预实验验证

基于实体类型过度依赖的定义，本章节通过预实验验证其对于事件要素抽取性能的影响。

数据集和模型 在数据集方面，预实验评估选择了广泛使用的事件检测语料库 ACE2005 数据集，其具体细节将在主实验章节进行阐述。在模型方面，由于 RNN、CNN 和 BERT 是事件要素抽取最常用的三类编码器，预实验基于所选数据集训练了三种不同的事件要素抽取模型，包括 AttRNN、DMCNN^[41] 和 DMBERT^[70]。其中，基于 RNN 的事件要素抽取模型如 JRNN^[100] 和 dbRNN^[101] 未公布源代码，因此本文受 Lu 等人^[54] 研究工作的启发，利用增加了注意力机制的双向 LSTM 作为使用 RNN 架构的基线模型，记作 AttRNN。此外，由于训练集和测试集之间的数据分布相似，实验中将根据实体类型依赖、语义不一致和实体类型过度依赖的定义，分别评估测试集中相应簇的性能。由于事件要素抽取的测试集依赖于事件检测的预测结果，本文遵循 Wang 等人和 Xi 等人^[99,104] 的实验设置，分别使用训练好的 AttRNN、DMCNN 和 DMBERT 来检测事件类型和相应的触发词。考虑到测试的实例数量取决于不同事件检测模型正确预测的事件数量，本实验中三种基线模型对应的测试集实例数分别为 4013、3630 和 4081。

实验步骤 首先，对于给定模型，本章使用其编码器获取测试集中每个实例的特征编码表示。然后，基于要素类型，测试集中的所有实例被划分为不同的群组，使得每个群组中实例的要素类型是相同的。进一步地，根据实体类型，将每个群组中的实例划分为不同的簇。因此，同一簇中实例的实体类型和要素类型是相同的。对于 AttRNN、DMCNN 和 DMBERT 各自对应的编码器，根据上述划分步骤可分别得到簇

的数目为 156、148 和 161。基于此，本章通过简单地计算各个群组/簇中所有实例特征表示的平均值，以用作其整体表示，其避免了引入额外计算参数和降低了计算复杂度。然后，本实验根据实体类型依赖、语义不一致和实体类型过度依赖的定义对每个簇进行判定，并分别计算符合不同定义的测试子集的性能（F1 值）。

评估验证 预实验结果如表4.2所示，其展示了根据不同定义类型划分的测试子集的簇数目及相应性能结果。从表中可以观察到：（1）对于不同的事件要素抽取模型，符合实体类型过度依赖的实例簇的性能都较低。其与其他类型的实例簇之间的巨大性能差异表明，现有事件要素抽取基线模型不擅长处理实体类型过度依赖问题。（2）符合实体类型过度依赖的实例簇和语义不一致的实例簇之间的性能差异表明，与语义不一致相比，由实体类型和要素类型之间的相关性导致的实体类型过度依赖进一步降低了模型的性能。（3）未对实体类型信息过度依赖的簇和整个测试集的性能差异表明，如果事件要素抽取模型缓解了实体类型过度依赖问题，则整体性能可以得到提升。

表 4.2 测试集中不同类型实例簇的数目及性能结果

模型	实体类型依赖 (数目/F1 (%))	语义不一致 (数目/F1 (%))	实体类型过度依赖 (数目/F1 (%))	非实体类型过度依赖 (数目/F1 (%))	所有簇 (数目/F1 (%))
AttRNN	100 / 40.6	56 / 38.9	48 / 18.1	108 / 63.6	156 / 50.9
DMCNN	112 / 46.0	76 / 32.6	52 / 15.6	96 / 71.4	148 / 53.5
DMBERT	95 / 38.6	48 / 40.7	31 / 15.3	130 / 68.9	161 / 57.2

由于不同簇中的实例数量差异较大，本章节进一步统计了符合实体类型过度依赖的簇中实例的数目占据所有测试实例的比例，记作“实体类型过度依赖比例”。表4.3展示了不同事件要素抽取模型的比例结果。本文将在章节4.4中评估下一节所提的实体类型过度依赖消解方法是否可以降低“实体类型过度依赖比例”，进而提高整体性能。

表 4.3 符合实体类型过度依赖的簇中实例统计信息

模型	实体类型过度依赖的簇中实例数	总实例数	实体类型过度依赖比例 (%)
AttRNN	2845	4013	70.9
DMCNN	2799	3630	77.1
DMBERT	1281	4081	31.4

从预实验中可以观察到，实体类型过度依赖问题降低了事件要素抽取模型的整体性能。为了缓解该问题，本文考虑降低“实体类型过度依赖比例”。进一步，其是由实体类型依赖导致，则可以利用以下两种方式处理：（1）增加具有相同要素类型但实体类型不同的实例特征表示之间的相似性；（2）降低具有不同要素类型但实体类型相

同的实例特征表示之间的相似性。基于此，下一节将具体介绍将上述两种方式整合的所提方法，从而提高事件要素抽取的性能。

4.3 方法设计

本章将详细介绍基于多角度对比学习的实体类型过度依赖消解方法，其可以应用于不同的事件要素抽取基线模型。为了描述方便，本章选取 DMBERT^[70]，以其为基础介绍相应的多角度实体类型过度依赖消解的事件要素抽取模型 METOR。图4.2展示了所提模型的总体架构，其主要由四个模块组成：编码器、正样本处理、负样本处理和循环训练策略。

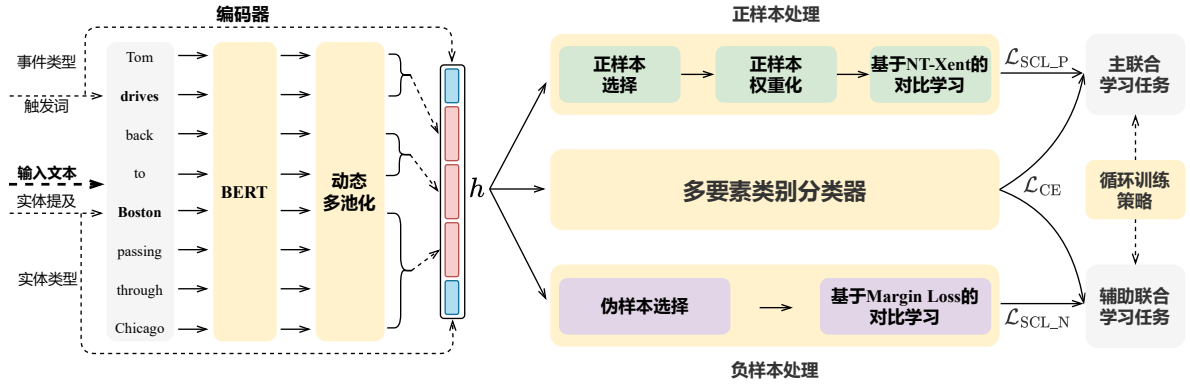


图 4.2 METOR 模型架构图

4.3.1 编码器

编码器模块旨在将每个实例编码成对应的特征表示。由于事件要素抽取为事件检测的后续任务，本章节遵循 Wang 等人^[99]和 Xi 等人^[104]的研究工作，先利用 Wang 等人^[99]提出的 DMBERT 事件检测模型进行事件触发词和相应事件类型的预测，并将所有实体提及设置为待抽取的事件要素候选。由于句子中可能存在多个事件要素候选，本章将这些候选分成多个实例处理。假设给定的事件要素候选为 arg ，其实体类型为 e ，所参与预测的事件触发词为 tri ，对应的事件类型为 t ，首先利用 BERT^[103] 对所在句子文本 $x = \{w_1, \dots, tri, \dots, arg, \dots, w_n\}$ 进行编码得到特征表示：

$$\{\hat{h}_1, \dots, \hat{h}_{p_{tri}}, \dots, \hat{h}_{p_{arg}}, \dots, \hat{h}_n\} = \text{BERT}(w_1, \dots, tri, \dots, arg, \dots, w_n) \quad (4.3)$$

其中 p_{tri} 和 p_{arg} 分别表示 tri 和 arg 在输入句子中的位置。

在此之后，使用一个动态多池化操作分段聚合编码后的特征表示：

$$[h_{1,p_{tri}}]_k = \max \left\{ [\hat{h}_1]_k \cdots, [\hat{h}_{p_{tri}}]_k \right\} \quad (4.4)$$

$$[h_{p_{tri}+1,p_{arg}}]_k = \max \left\{ [\hat{h}_{p_{tri}+1}]_k \cdots, [\hat{h}_{p_{arg}}]_k \right\} \quad (4.5)$$

$$[h_{p_{arg}+1,n}]_k = \max \left\{ [\hat{h}_{p_{arg}+1}]_k \cdots, [\hat{h}_n]_k \right\} \quad (4.6)$$

其中 $[\cdot]_k$ 表示对应向量的第 k 个值。然后，本章随机初始化所有的事件类型标签和实体类型标签，分别表示为 $\mathbf{W}_T \in \mathbb{R}^{n_t \times d_s}$ 和 $\mathbf{W}_E \in \mathbb{R}^{n_e \times d_s}$ 。其中， d_s 、 n_t 和 n_e 分别表示事件类型（实体类型）向量的维度、预定义的事件类型数目和预定义的实体类型数目。此外， $\mathbf{W}_T(t)$ 表示使用事件类型 t 从向量矩阵 \mathbf{W}_T 中查找出的事件类型向量。类似地， $\mathbf{W}_E(e)$ 表示根据实体类型 e 查找出的实体类型向量。最后，串接上述表示，得到实例的特征表示 $\mathbf{h} \in \mathbb{R}^{d_{out}}$ 如下：

$$\mathbf{h} = [h_{1,p_{tri}}; h_{p_{tri}+1,p_{arg}}; h_{p_{arg}+1,n}; \mathbf{W}_T(t); \mathbf{W}_E(e)] \quad (4.7)$$

4.3.2 正样本处理

本模块提出了一种选择-加权的对比学习方法，该方法包括三个步骤：正样本选择、正样本加权和基于归一化温度标度交叉熵（NT-Xent）的对比学习。首先，该方法基于给定实例的实体类型选择正样本，使得模型能够聚焦于处理与给定实例的实体类型无关的正样本。然后，计算全局共现信息和语义相关性信息，以对所选正样本的相关性进一步权重化处理。基于此，模型可以提升特定正样本的对比学习贡献，其包含了更多与给定实例要素类型相关的关键语义信息，从而学习到实体类型信息之外的更多有效语义知识。最后，计算基于 NT-Xent 的选择-加权对比学习方法的训练损失。

正样本选择 给定一批包含 K 个实例的批数据 \mathcal{B} ，对于其中的每个实例 x_i ，首先选择与其要素类型相同的实例作为正样本集合 $\mathcal{P}(x_i)$ ，选择与其要素类型不同的实例作为负样本集合 $\mathcal{N}(x_i)$ 。基于 $\mathcal{P}(x_i)$ ，本节采用一个简单的选择策略去获取新的正样本集合：

$$\mathcal{P}_s(x_i) = \{x_p : x_p \in \mathcal{B}, (e_p \neq e_i) \wedge (r_p = r_i) \wedge (p \neq i)\} \quad (4.8)$$

其中 r_i 和 e_i 分别表示实例 x_i 的要素类型和实体类型。本选择策略只用于包含了不止

一个样本的 $\mathcal{P}(x_i)$ ，以避免造成对比数据的稀疏。

正样本加权 在完成正样本选择后，本章根据要素类型 r_i ，对 $\mathcal{P}_s(x_i)$ 中的每个正样本分配一个权重。传统的对比学习方法在对比学习损失中同等地对待所有正样本，无法区分具有不同实体类型的正样本的重要程度。为此，本章基于要素类型 r_i ，从全局共现信息和语义相关性信息两个方面，计算不同正样本对于 x_i 的重要性。

全局共现信息可表示为记录了整个训练集中不同实体类型和要素类型共现数量的共现矩阵 $I_{e,r}$ 。假设 \mathcal{E} 和 \mathcal{R} 分别为训练集中的实体类型和要素类型集合，批数据 \mathcal{B} 中第 p 个实例（即 x_p ）存在于集合 $\mathcal{P}_s(x_i)$ 中，其实体类型为 e_p 。受 TF-IDF^[153] 启发，首先定义实体类型频率（ETF），用于表示 e_p 与 r_i 同时出现的频率：

$$\text{ETF}(e_p, r_i) = \frac{I_{e_p, r_i}}{\sum_{m \in \mathcal{E}} I_{m, r_i}} \quad (4.9)$$

此外，本章使用与 e_p 同时出现的要素类型的数量来衡量 e_p 的重要性。要素类型数量越少， e_p 则越重要。因此，反向要素类型频率（IRTF）可根据如下定义得到：

$$\text{IRTF}(e_p) = \log \frac{|\mathcal{R}|}{|\{r \in \mathcal{R} : I_{e_p, r} > 0\}| + 1} \quad (4.10)$$

基于 ETF 和 IRTF，可得实例 x_p 对于实例 x_i 在全局共现信息方面的重要性得分：

$$w_1(x_p, x_i) = \text{ETF}(e_p, r_i) \times \text{IRTF}(e_p) \quad (4.11)$$

然而，在事件类型不同的实例中，具有相同实体类型的实体提及可能以不同的要素类型参与在事件中。因此，对于不同的事件类型，即使实例的实体类型相同，对给定实例的重要性也应有所区分。因此，考虑 x_p 对 x_i 的重要性时仅使用全局共现信息不足以建模不同事件类型的影响。基于此，本节进一步利用语义相关性信息。首先，将所有要素类型标签随机初始化成向量矩阵 $\mathbf{W}_R \in \mathbb{R}^{n_r \times d_s}$ ，其中 n_r 为预定义的要素类型的数目（包括特殊要素类型“空”）， d_s 是要素类型向量的维度。然后，对于给定的实例 x_i ，得到融合事件类型信息的实体类型和要素类型表示：

$$\mathbf{e}_i^t = \mathbf{W}_T(t_i) \odot \mathbf{W}_E(e_i) \quad (4.12)$$

$$\mathbf{r}_i^t = \mathbf{W}_T(t_i) \odot \mathbf{W}_R(r_i) \quad (4.13)$$

其中 \odot 表示元素依次相乘, r_i 、 e_i 和 t_i 分别表示实例 x_i 的要素类型、实体类型和事件类型。 $\mathbf{W}_T(t_i)$ 表示使用事件类型 t_i 从向量矩阵 \mathbf{W}_T 中查找出的事件类型向量。类似地, 可以分别得到实体类型向量 $\mathbf{W}_E(e_i)$ 和要素类型向量 $\mathbf{W}_R(r_i)$ 的含义。由于 \mathbf{W}_T 和 \mathbf{W}_E 同时也是公式 4.7 中串接向量的一部分, 事件类型和实体类型的语义信息实现了在不同模块间的共享。同样, 可以得到任意正样本 x_p 的融合事件类型信息的实体类型表示 \mathbf{e}_p^t 。然后, 本章基于语义相关性信息, 利用注意力机制计算 x_p 对于 x_i 的重要性得分, 并根据 softmax 函数得到归一化结果 $w_2(x_p, x_i)$:

$$s(x_p, x_i) = \mathbf{v}^\top \tanh(\mathbf{U} [\mathbf{e}_i^t; \mathbf{e}_p^t; \mathbf{r}_i^t]) \quad (4.14)$$

$$w_2(x_p, x_i) = \frac{\exp(s(x_p, x_i))}{\sum_{e_j \in \mathcal{E}} \exp(s(x_j, x_i))} \quad (4.15)$$

其中 $\mathbf{v} \in \mathbb{R}^{3d_s}$ 和 $\mathbf{U} \in \mathbb{R}^{3d_s \times 3d_s}$ 分别表示可训练学习的向量和矩阵参数。然后, 可得 x_p 对于 x_i 的最终重要性得分:

$$\mathbf{w}(x_p, x_i) = \alpha w_1(x_p, x_i) + (1 - \alpha) w_2(x_p, x_i) \quad (4.16)$$

其中 α ($0 < \alpha < 1$) 为权重参数。

基于 NT-Xent 的对比学习 遵循 Chen 等人^[145] 和 Khosla 等人^[149] 的研究, 本章首先将实例 x_i 的特征表示 \mathbf{h}_i 映射到一个新空间上, 以提高对比学习的质量:

$$\mathbf{z}_i = \mathbf{W}_1 f(\mathbf{W}_2 \mathbf{h}_i) \quad (4.17)$$

$\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_2}$ 和 $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_{out}}$ 为可训练学习的矩阵参数, $f(\cdot)$ 为 ReLU 激活函数, 其中 d_1 和 d_2 为维度参数。在选择出正样本并对其权重化后, 本章基于 NT-Xent^[145] 计算相应的对比学习损失 \mathcal{L}_{SCL_P} 如下:

$$\mathcal{L}_{SCL_P} = \sum_{i=1}^K \frac{-1}{|\mathcal{P}_s(x_i)|} \sum_{p \in \mathcal{P}_s(x_i)} \mathbf{w}(x_p, x_i) \cdot \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p) / \tau)}{\sum_{n \in K \setminus i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_n) / \tau)} \quad (4.18)$$

其中 $\text{sim}(\mathbf{z}_i, \mathbf{z}_p) = \mathbf{z}_i^\top \mathbf{z}_p / \|\mathbf{z}_i\| \|\mathbf{z}_p\|$, τ 表示温度参数。

4.3.3 负样本处理

本模块提出了一种伪正对比学习方法，该方法包括两个步骤：伪样本选择和基于边际损失（Margin Loss）的对比学习。首先，基于实体类型从负样本中选择出伪正样本。然后，利用这些选择出的样本作为对比参考，降低给定实例和与其实体类型相同的负样本之间的特征表示相似性。

伪样本选择 给定一批包含 K 个实例的批数据 \mathcal{B} ，对于其中的每个实例 x_i ，首先选择对应的伪正样本和伪负样本，选择过程如下：

$$\mathcal{P}_{pseudo}(x_i) = \{x_p : x_p \in \mathcal{B}, (e_p \neq e_i) \wedge (r_p \neq r_i) \wedge (p \neq i)\} \quad (4.19)$$

$$\mathcal{N}_{pseudo}(x_i) = \{x_n : x_n \in \mathcal{B}, (e_n = e_i) \wedge (r_n \neq r_i) \wedge (n \neq i)\} \quad (4.20)$$

其中伪正样本集合和伪负样本集合分别表示为 $\mathcal{P}_{pseudo}(x_i)$ 和 $\mathcal{N}_{pseudo}(x_i)$ 。

基于边际损失的对比学习 首先，本章计算 $\mathcal{P}_{pseudo}(x_i)$ 和 $\mathcal{N}_{pseudo}(x_i)$ 中实例特征表示的平均值，分别得到 \mathbf{z}_i^P 和 \mathbf{z}_i^N ，以表示这两个集合的整体信息。然后，利用 $\mathcal{P}_{pseudo}(x_i)$ 作为对比参考进行对比学习，以降低实例 x_i 与属于 $\mathcal{N}_{pseudo}(x_i)$ 的样本之间的相似性。由于 $\mathcal{P}_{pseudo}(x_i)$ 中的样本本质上仍是 x_i 的负样本，使用实例级的对比学习损失 NT-Xent 以增加伪正样本与给定实例之间的相似性是不合适的。因此，本章使用边际损失来对伪正样本集合和伪负样本集合的整体表示进行对比学习：

$$\mathcal{L}_{SCL_N} = \sum_{i=1}^K \max \{ \mathbf{z}_i^N \cdot \mathbf{z}_i - \mathbf{z}_i^P \cdot \mathbf{z}_i + \gamma, 0 \} \quad (4.21)$$

其中 γ 为边际超参数。

4.3.4 循环训练策略

本模块首先计算事件要素抽取任务的损失。给定实例 x ，将其从编码器模块得到的特征表示 \mathbf{h} 输入到多要素类别分类器，以计算概率分布 $p(x)$ 如下：

$$p(x) = \text{softmax}(\mathbf{W}^c \mathbf{h} + \mathbf{b}^c) \quad (4.22)$$

其中 $\mathbf{W}^c \in \mathbb{R}^{n_r \times d_{out}}$ 和 $\mathbf{b}^c \in \mathbb{R}^{n_r}$ 分别表示分类器中要训练的参数。给定包含 K 个实例的批数据 \mathcal{B} ，交叉熵损失 \mathcal{L}_{CE} 计算如下：

$$\mathcal{L}_{CE} = - \sum^K \log p(r|x) \quad (4.23)$$

其中 $p(r|x)$ 表示将实例 x 预测为正确的要素类型 r 的概率值。在测试阶段， $p(x)$ 用于获取事件要素抽取的预测要素类型。

进一步，本章注意到优化 \mathcal{L}_{SCL_N} 将不可避免地提升实例 x_i 和 $\mathcal{P}_{pseudo}(x_i)$ 中的样本（本质上为负样本）间的特征表示相似度。然而，根据公式4.18，优化 \mathcal{L}_{SCL_P} 会降低实例 x_i 和所有负样本间的特征表示相似度。因此， \mathcal{L}_{SCL_N} 和 \mathcal{L}_{SCL_P} 的训练目标是不一致的。为此，本章提出了一个循环训练策略，该策略依次使用两个不同的联合训练目标来协调 \mathcal{L}_{SCL_P} 和 \mathcal{L}_{SCL_N} 的不一致。具体来说，将联合优化 \mathcal{L}_{SCL_P} 和 \mathcal{L}_{CE} 视为主联合学习任务，先训练 f 轮，其中 f 是超参数。然后，将联合优化 \mathcal{L}_{SCL_N} 和 \mathcal{L}_{CE} 视作辅助联合学习任务，再训练一轮。重复上述过程直到训练完成。通过这种方式，可以逐步调整处理 \mathcal{L}_{SCL_P} 和 \mathcal{L}_{SCL_N} 训练目标的不一致。本章提出的训练策略表示如下：

$$\mathcal{L}_{JOINT} = \begin{cases} \mathcal{L}_{SCL_P} + \mathcal{L}_{CE} & epoch \bmod f \neq 0 \\ \mathcal{L}_{SCL_N} + \mathcal{L}_{CE} & epoch \bmod f = 0 \end{cases} \quad (4.24)$$

其中 $epoch$ 表示完成训练的轮次。

4.4 实验评估

在实验评估中，本章所提的方法不仅应用于基线模型 DMBERT，也应用于 AttRNN 和 DMCNN 基线模型，分别记作 METOR (RNN) 和 METOR (CNN)。在 METOR (RNN) 和 METOR (CNN) 中，将编码器模块中的 DMBERT 分别替换为 AttRNN 和 DMCNN，其他模块保持不变。由于事件要素抽取任务依赖于事件检测的结果，本章遵循 Wang 等人^[99] 和 Xi 等人^[104] 的研究设定，使用经过训练的 AttRNN、DMCNN 和 DMBERT 分别检测事件触发词和相应的事件类型，作为 METOR (RNN)、METOR (CNN) 和 METOR 模型的输入。接下来，本章将介绍实验设置、性能结果、消融研究、实体类型信息深入分析、超参数敏感性分析、计算复杂性分析和案例研究。

4.4.1 实验设置

1. 数据集 遵循最近主流的给定实体提及的事件要素抽取研究工作^[57,100,104]，本章实验评估在 ACE2005 数据集上进行。与章节3.3.1相同，其为 ACE2005 语料库^[138] 中的英文版本数据集。根据 Wang 等人^[99] 的预处理设置¹，该数据集包含 599 个文档、33 种事件类型、46 种实体类型和 35 种事件要素类型。此外，引入 *None*（空）作为特殊要素类型，以表示相应的候选要素在给定实例中为非事件要素。进一步，该数据集被划分为 529、30 和 40 个文档，分别作为训练集、验证集和测试集。

2. 评价指标 本章实验遵循事件要素抽取任务的标准评估。如果事件类型、偏移量和要素类型与数据集的标注完全相同，则认定候选要素被正确抽取，其中偏移量指候选要素文段在句子中的开始和结束位置索引。基于此，本文使用准确率（P）、召回率（R）和 F1 值（F1）作为评估指标，其计算方式参照章节3.3.1。

3. 实验配置 在编码器模块中，对于 METOR（RNN）本章使用预训练 Glove 词向量、随机初始化事件类型向量、随机初始化实体类型向量和随机初始化位置向量，维度分别为 100、5、5 和 5，并在模型中使用双向 LSTM 和自注意力机制。对于 METOR（CNN），使用与 DMCNN 相同的输入向量和超参数。类似地，本章将 5 维随机初始化的实体类型向量添加到 METOR（CNN）的输入中。对于 METOR，编码使用的 BERT 预训练模型为 HuggingFace 发布的“google-bert/bert-base-uncased”版本²，实体类型向量维度 d_s 设置为 50。在正样本处理模块中，本章将 d_1 设置为 512， d_2 设置为 512，温度参数 τ 设置为 0.1，权重 α 设置为 0.7。在负样本处理模块中，将 γ 设置为 0.1。在循环训练策略中，METOR（RNN）、METOR（CNN）和 METOR 的循环频率 f 分别设置为 4、3 和 3。此外，其批数据大小均设置为 80，训练轮次设置为 10，优化器使用 AdamW，对应的学习率分别设置为 1×10^{-3} 、 1×10^{-3} 和 5×10^{-5} 。对于不同的模型，训练均在相同实验服务器上进行，具体配置为：CPU 型号 Intel(R) Xeon(R) Gold 5320，主频 2.20GHz；内存 256G；GPU 计算卡型号 NVIDIA GeForce RTX 3090，显存 24G。上述模型均基于 Pytorch^[139] 框架进行构建。

4. 基线模型 本节选择以下模型作为实验比较的基线：

- DMCNN^[41] - 该模型使用了 CNN 和动态多池化操作提取实例的特征表示。

¹<https://github.com/thunlp/HMEAE>

²<https://huggingface.co/google-bert/bert-base-uncased>

- JRNN^[100] - 该模型提出了包含离散二元矩阵的 RNN，其利用了事件类型和要素类型、要素类型间的相互依赖性。
- dbRNN^[101] - 该模型在 RNN 中引入了依存句法信息，并构建事件内候选要素间的交互向量以捕捉其关联性。
- DMBERT^[70] - 该模型使用了 BERT 和动态多池化操作来提取实例的特征表示。
- PLMETOR^[69] - 该模型基于 BERT 来预测事件要素类型并生成额外的训练实例，以进一步提高性能。
- HMEAE^[99] - 该模型提出了一种利用事件要素类型语义相关性的层次模块化注意力网络。
- BERT (Inter)^[104] - 该模型基于 BERT 并利用了跨事件的相互依赖关系^[100]。
- BERT (Intra)^[104] - 该模型基于 BERT 并利用了事件内部的要素依赖关联^[101]。
- BERD^[104] - 该模型提出了一种能够利用事件内其他要素的类型预测语义信息的编码器-解码器框架。

本章注意到 RCEE_ER^[71] 在 F1 指标上取了优秀的性能表现，但其受益于额外的数据增广和无监督数据。因此，本章遵循 Xi 等人^[104] 在 BERD 模型中的基线模型设置，未将 RCEE_ER 模型作为实验性能比较的基线。此外，本章将 ChatGPT 在 ACE2005 数据集上的性能评估保留到章节 5.3 进行展示与分析。

4.4.2 性能结果

表 4.4 展示了本章所提方法与基线模型在 ACE2005 数据集上的性能结果，其中 † 表示将本章提出的方法应用在 DMBERT 基线模型时，其性能在双侧 t 检验下显著 ($p < 0.01$) 优于之前性能最优的事件要素抽取模型 BERD。根据结果可进一步观察到，与 AttRNN 相比，METOR (RNN) 在 F1 指标上实现了 3.9% 的性能提升。即使与利用了事件类型和要素类型、要素类型间的相互依赖性的 JRNN 相比，METOR (RNN) 也取得了相差无几的 F1 指标结果。此外，dbRNN 的 F1 指标显著优于 METOR (RNN)，本文将其归因于两个原因：(1) dbRNN 利用了额外的句法信息。(2) dbRNN 同时优化了事件检测和事件要素抽取任务的性能，而 METOR (RNN) 没有受益于这种联合

优化。当本章提出的方法应用在 DMCNN 时, METOR (CNN) 相比该基线在 F1 指标上实现了 5% 的提升, 比同样使用了 DMCNN 作为编码器的 HMEAE (CNN) 提升了 2.8%。并且, METOR (CNN) 的 F1 指标优于大多数没有使用预训练模型的事件要素抽取模型, 并与 dbRNN 相比时展现了相当的性能结果。同时, 可以观察到 METOR (CNN) 的性能优于使用了 BERT 预训练模型作为编码器的 DMBERT、PLMETOR 和 BERT (Inter)。

当在编码器模块使用 DMBERT 时, METOR 取得了当前最优的 F1 性能, 且相比于之前性能最优的 BERD 模型提升了 2.1%。此外, METOR 的 F1 指标比相应的 DMBERT 基线高出了 5.2%。总的来说, 由于本章所提的 METOR 模型可以学习到除实体类型之外的更多语义信息, 从而消解了实体类型过度依赖问题, 实现了在事件要素抽取任务上对所有基线模型的性能超越。

表 4.4 ACE2005 数据集上的事件要素抽取性能结果

模型	P(%)	R(%)	F1(%)
AttRNN	50.6	51.1	50.9
DMCNN	62.2	46.9	53.5
JRNN	54.2	56.7	55.4
dbRNN	66.2	52.8	58.7
HMEAE (CNN)	57.3	54.2	55.7
METOR (RNN)	51.9	58.0	54.8
METOR (CNN)	56.2	61.1	58.5
+ BERT (<i>base</i>)			
DMBERT	58.8	55.8	57.2
PLMETOR	62.3	54.2	58.0
BERT (Inter)	58.4	57.1	57.8
BERT (Intra)	56.4	61.2	58.7
HMEAE (BERT)	62.2	56.6	59.3
BERD	59.1	61.5	60.3
METOR	60.3	64.7	62.4[†]

4.4.3 消融研究

本章通过消融研究分析所提方法中每个模块的有效性, 包括正样本处理 (Positive Sample Processing)、负样本处理 (Negative Sample Processing) 和循环训练策略 (Cyclic Training Strategy), 其中 $\Delta F1$ 表示与完整方法在 F1 指标上的性能差距。如表 4.5 所

示，将本章所提的方法应用到不同的基线模型上，三个模块都能提升相应基线模型的性能。本节选择 METOR 模型进行着重讨论。首先，如果移除正样本处理和负样本处理模块，METOR 模型在 F1 指标上的性能将分别下降 2.0% 和 1.3%，这表明从正样本和负样本角度消解实体类型过度依赖都可以有效地提升事件要素抽取性能。其次，如果将不同模块的损失直接求和并进行联合训练，以替代循环训练策略，则 METOR 模型在 F1 上的性能将下降 2.9%。这表明循环训练策略可以有效地调和正样本处理模块和负样本处理模块中的不同对比学习目标。此外，可以观察到即使移除掉正样本处理模块或负样本处理模块，METOR 模型仍优于之前性能最优的 BERD 模型。本文将其归因于这两个模块都推动了模型学习除实体类型之外的更多语义信息，从而消解了实体类型过度依赖问题，提升了整体性能。

表 4.5 ACE2005 数据集上的消融研究结果

模型	P(%)	R(%)	F1(%)	Δ F1
METOR (RNN)	51.9	58.0	54.8	-
w/o Positive Sample Processing	48.7	55.7	52.0	-2.8
w/o Negative Sample Processing	51.1	57.2	54.0	-0.8
w/o Cyclic Training Strategy	50.8	52.7	51.7	-3.1
METOR (CNN)	56.2	61.1	58.5	-
w/o Positive Sample Processing	52.9	59.2	55.9	-2.6
w/o Negative Sample Processing	57.4	56.3	56.8	-1.7
w/o Cyclic Training Strategy	54.0	57.4	55.7	-2.8
METOR	60.3	64.7	62.4	-
w/o Positive Sample Processing	59.7	61.2	60.4	-2.0
w/o Negative Sample Processing	59.3	62.9	61.1	-1.3
w/o Cyclic Training Strategy	57.1	62.1	59.5	-2.9

4.4.4 实体类型信息分析

作为基于实体的任务，实体类型信息对于事件要素抽取任务的影响是多方面的。一方面，实体类型是事件要素抽取任务中实体提及的主要信息来源。另一方面，其导致事件要素抽取模型产生了实体类型过度依赖，进而降低了模型的整体性能。因此，本节将对实体类型信息进行深入全面的分析。

不同变体模型的实体类型信息利用效率分析 本节设计两种变体模型作为额外的比较基线，即“Baseline+type”和“Baseline+type+CL”，其中“Baseline”可以为 AttRNN、

DMCNN 或 DMBERT。具体地,“Baseline”模型不使用实体类型信息,而“Baseline+type”使用实体类型信息作为输入的一部分。此外,对于“Baseline+type+CL”,其除了使用实体类型信息作为输入外,还引入监督对比学习方法^[149]损失,与公式4.23中的交叉熵损失直接求和后并进行联合训练。

表4.6展示了本章提出的方法与不同变体模型的性能比较。可以观察到,当“Baseline”为 AttRNN、DMCNN 和 DMBERT 时,“Baseline+type”在 F1 指标上的性能与其相比分别提升了 0.8%、1.2% 和 1.2%。这表明在事件要素抽取基线模型的输入中利用实体类型信息,其性能能够得到一定的提升。此外,当“Baseline”为 DMCNN 和 DMBERT 时,“Baseline+type+CL”可以进一步提高“Baseline+type”在 F1 指标上的性能,分别为 0.2% 和 0.6%,其表明传统的监督对比学习方法可以进一步提高基线模型利用实体类型信息的效率。进一步,由于 METOR 模型使用了实体类型信息并针对该信息设计了多角度对比学习方法,使得其能在利用实体类型信息的同时,从文本中学习除实体类型信息之外的更多语义信息。因此,与 AttRNN、DMCNN 和 DMBERT 相比, METOR 模型在 F1 上的性能分别提升了 3.9%、5.0% 和 5.2%。

表 4.6 ACE2005 数据集上不同变体模型的性能结果对比

模型	P(%)	R(%)	F1(%)	Δ F1
AttRNN	50.6	51.1	50.9	-
AttRNN+type	51.8	51.6	51.7	+0.8
AttRNN+type+CL	50.2	53.4	51.7	+0.8
METOR (RNN)	51.9	58.0	54.8	+3.9
DMCNN	62.2	46.9	53.5	-
DMCNN+type	51.5	58.3	54.7	+1.2
DMCNN+type+CL	53.8	56.1	54.9	+1.4
METOR (CNN)	56.2	61.1	58.5	+5.0
DMBERT	58.8	55.8	57.2	-
DMBERT+type	55.7	61.3	58.4	+1.2
DMBERT+type+CL	55.9	62.4	59.0	+1.8
METOR	60.3	64.7	62.4	+5.2

实体类型过度依赖程度分析 在章节4.2中,可以得知不同的事件要素抽取基线模型均存在实体类型过度依赖问题。如果反映实体类型过度依赖程度的“实体类型过度依赖比例”指标降低,则对应的整体性能也会提升。为了直观验证本章提出的方法在解决实体类型过度依赖问题上的有效性,本节将基于“实体类型过度依赖比例”指标进行具体分析。如图4.3所示,与相应的“Baseline”模型相比,“Baseline+type+CL”的

“实体类型过度依赖比例”均略有下降，表明普通监督对比学习可以在一定程度上消解实体类型过度依赖问题。进一步，可以观察到将本章提出的方法应用到不同的基线模型时，与相应的“Baseline”模型相比，其“实体类型过度依赖比例”均显著下降。因此，本章所提方法可以有效降低对实体类型的过度依赖程度。此外，可以观察到“Baseline+type+CL”在处理实体类型过度依赖问题的有效性不如本章所提方法，其原因为普通的监督对比学习主要解决定义2中的语义不一致问题，而没有聚焦于消解实体类型过度依赖问题。

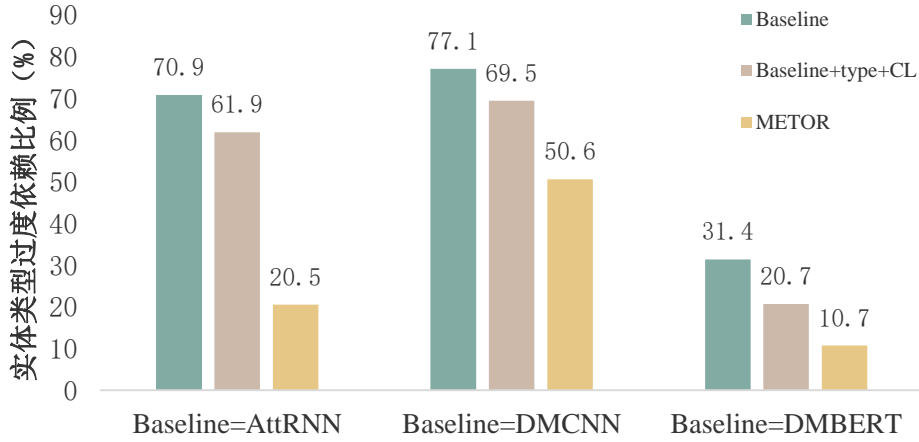


图 4.3 不同模型的实体类型过度依赖程度比较

4.4.5 超参数敏感性分析

本章分析不同超参数对 METOR 模型性能的影响，包括公式4.18中的温度参数 τ 、公式4.21中的边际参数 γ 和公式4.24中的循环训练频率 f 。

如图4.4(a)所示，METOR 模型在温度参数 τ 为 0.1 时性能最佳。当其在 0.02 到 0.5 的范围内变化时，METOR 模型的 F1 性能变化幅度不超过 1.1%，这证明了 METOR 模型在不同 τ 的设置下的鲁棒性。

从图4.4(b)中可以观察到，当边际参数 γ 设置为 0.1 时，METOR 模型的 F1 指标性能达到最佳。当边际参数 γ 设置为 0.05 时，从负样本角度消解实体类型过度依赖问题的程度存在不足，METOR 模型的 F1 的性能略有下降。而当边际参数 γ 大于 0.1 时，其性能也会下降。这是由于给定实例和伪正样本集合中的负样本之间的特征表示过度相似，导致正样本处理模块和负样本处理模块间的训练目标差距较大。然而，由于 METOR 模型中的循环训练策略能够极大缓解由较大边际参数 γ 引起的训练目标不一致性问题，其性能的下降较为有限。

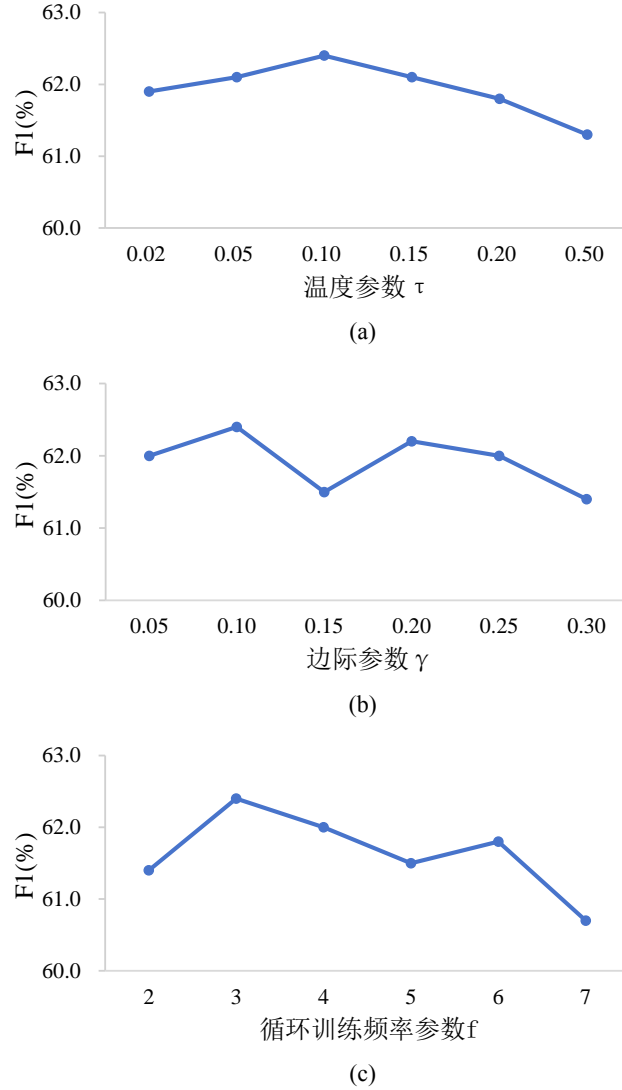


图 4.4 METOR 模型参数敏感性分析

从图4.4(c)中可以观察到,当循环训练频率 f 设置为 3 时, METOR 模型实现了最佳性能。当循环频率 f 设置为 2 时,循环训练策略中主任务和辅助任务的训练频率没有差异,导致 METOR 模型在两个不一致的训练目标间频繁切换,从而性能出现下降。此外,当循环频率 f 大于 3 时,导致循环训练策略中负样本处理模块的训练频率较低,其将接近于退化为不包含负样本处理模块的 METOR 模型。

4.4.6 计算复杂度分析

虽然本章所提方法在实验中应用到了不同的事件要素抽取模型,但计算复杂度的分析是相似的。因此,本节着重讨论基于 DMBERT 的 METOR 模型的情况,具体计算

如下：在编码器模块中，BERT 的计算复杂度主要来源于多头注意力^[67]，为 $O(N^2d + Nd^2)$ ，其中 N 是输入序列的最大长度， d 是 BERT 隐藏层输出向量的维度。在正样本处理模块中，语义相关性信息的计算复杂度为 $O((d_s^2 + d_s)n_e)$ ，其中 d_s 是事件类型/实体类型/要素类型向量的维度， n_e 是给定事件要素抽取数据集中预定义实体类型的数目。对于基于 NT-Xent 的对比学习来说，其计算复杂度为 $O((d + d_s + d_2)d_1 + Kd_2 + K^2)$ ，其中 d_1 和 d_2 是维度参数， K 是批数据大小。在负样本处理模块中，计算复杂度为 $O(d_2)$ 。在循环训练策略模块中，多类分类器的计算复杂度为 $O(n_r(d + d_s))$ ，其中 n_r 是给定事件要素抽取数据集中预定义要素类型的数目。

因此，METOR 模型的总计算复杂度为 $O(Nd^2 + (N^2 + d_1 + n_r)d + (d_1 + K)d_2 + n_e d_s^2 + (d_1 + n_r + n_e)d_s + K^2)$ 。如果忽略较小的参数 n_e ， n_r 和 d_s ，则其计算复杂度可以简化为 $O(Nd^2 + (N^2 + d_1)d + (d_1 + K)d_2 + K^2)$ 。进一步，与计算复杂度为 $O(Nd^2 + N^2d)$ 的基线模型 DMBERT 相比，所提模型的计算复杂度增长了 $O(d_1d + (d_1 + K)d_2 + K^2)$ 。由于在本章的超参数设置中，存在 $d_1 + K < d$ ， $d_2 < d$ 和 $K < N$ ，使得增长的计算复杂度 $O(d_1d + (d_1 + K)d_2 + K^2)$ 约等于 $O(d^2 + N^2)$ 。因此，可以得知 METOR 模型的计算复杂度略高于 DMBERT，但其 F1 性能显著提高了 5.2%。

4.4.7 案例研究

本章节选取 ACE2005 数据集测试集中的六个实例进行案例研究。表 4.7 展示了使用 DMBERT、DMBERT (CL) 和本章所提的 METOR 模型对这些实例的预测结果，其中 DMBERT (CL) 为表 4.6 中的变体模型 DMBERT+type+CL。

在实例 (1) 中，“rally”触发了 *Demonstrate* 事件，且事件候选要素为 “Manhattan”，其实体类型和要素类型分别为 *County – or – District* 和 *Place*。由于实体类型 *County – or – District* 通常以要素类型 *Place* 参与到事件中，为该要素类型预测提供了重要信息。因此，DMBERT、DMBERT (CL) 和本章所提的 METOR 模型通过实体类型 *County – or – District* 提供的语义信息给出了正确的要素类型预测。

在实例 (2) 中，候选事件要素 “the court” 的实体类型和要素类型分别为 *Government* 和 *Adjudicator*。虽然实体类型 *Government* 无法提供明显的语义信息来预测要素类型，但当 “verdict” 触发的 *Convict* 事件给出时，其可以为预测提供信息提示。换言之，*Government* 实体经常在 *Convict* 事件中以要素类型 *Adjudicator* 出现，所以三个模型均能正确预测要素类型。因此，前两个实例证明了实体类型可以为预测其频

表 4.7 ACE2005 数据集上的案例分析

实例	实体类型	DMBERT	DMBERT (CL)	METOR
(1) Liana Owen drove 10 hours from Pennsylvania to attend the rally in Manhattan with her parents.	County-or-District	Place ✓	Place ✓	Place ✓
(2) He asked the court to invalidate the verdict and throw out the criminal case against Pasko.	Government	Adjudicator ✓	Adjudicator ✓	Adjudicator ✓
(3) Separately, former WorldCom CEO Bernard Ebbers failed on April 29 to make a first repayment of 25 million dollars on a 400-million-dollar loan from MCI, the Journal said, citing SEC documents.	Individual	Giver ×	Recipient ✓	Recipient ✓
(4) U.S. Defense Secretary Donald H. Rumsfeld discussed the resolution with Prime Minister Tony Blair and Defense Secretary Geoff Hoon on Friday as Rumsfeld returned from a tour of Iraq, Afghanistan and the Persian Gulf region.	Nation	Destination ×	Origin ✓	Origin ✓
(5) It was not clear how many people were in the cafe at the time of the blast.	Indeterminate	None ×	None ×	Target ✓
(6) The space agency has been under intense scrutiny since February when the space shuttle Columbia disintegrated over Texas, killing all seven crew members.	Air	Instrument ×	Instrument ×	None ✓

繁参与的要素类型提供有用的预测线索。

在实例（3）中，在由“loan”触发的 *Transfer – Money* 事件中，候选事件要素“former WorldCom CEO Bernard Ebbers”的实体类型和要素类型分别为 *Individual* 和 *Recipient*。由于要素类型 *Giver* 和 *Recipient* 均与实体类型 *Individual* 在 *Transfer – Money* 事件中频繁关联，DMBERT 无法在该实例中正确区分 *Giver* 和 *Recipient*。然而，DMBERT (CL) 和本章所提的 METOR 模型均受益于对比学习，避免被同一事件中具有相同实体类型的要素类型所混淆，给出了正确的识别结果。实例（4）展示了另一个类似的实例。

在实例（5）中，候选事件要素“how many people”的实体类型和要素类型分别为 *Indeterminate* 和 *Target*。从语义上来看，如果候选要素为一个 *Indeterminate* 实体，则该实体很有可能在事件中不作为要素出现。这种强语义关联导致了 DMBERT 对要素类型的预测错误。进一步，由于多数实体类型为 *Indeterminate* 的候选要素在训练集中不为事件要素出现，DMBERT (CL) 使用传统的对比学习来避免对于实体类型 *Indeterminate* 过度依赖的效果不佳。然而，本章所提的 METOR 模型消解了这种实

体类型过度依赖问题，避免了该错误。

在实例 (6) 中，在由 “killing” 触发的 *Die* 事件中，候选事件要素 “the space shuttle Columbia” 的实体类型和要素类型分别是 *Air* 和 *None*。和实例 (5) 相似，如果候选事件要素为一个 *Air* 实体，则该实体很有可能在事件中以要素类型 *Instrument* 出现，这阻碍了 DMBERT 和 DMBERT (CL) 从文本中理解语义信息，导致它们做出错误的预测。与之相比，METOR 模型正确预测了要素类型。因此，METOR 可以同时减少假负 (false negative) 预测（如实例 (5) 所示）和假正 (false positive) 预测（如实例 (6) 所示），从而取得优秀的事件要素抽取性能表现。

4.5 本章小结

本章首先研究了事件要素抽取的实体类型依赖性，并利用预实验评估其对不同事件要素抽取基线模型整体性能的影响。预实验表明了不同基线模型均面临不同程度的实体类型过度依赖问题，从而降低了整体性能。为此，本章提出了基于多角度对比学习的实体类型过度依赖消解方法。该方法分别从正样本和负样本角度消解模型在训练过程中对于实体类型的过度依赖，并设计了一种循环训练策略使这两种对比学习方法能够高效协作。本章基于 ACE2005 数据集进行了充分的实验，结果表明所提方法能够有效地对不同事件要素抽取基线模型进行实体类型过度依赖问题的消解，且当应用在 DMBERT 基线时，性能显著优于当前最佳的事件要素抽取模型。

第 5 章 基于分离-融合范式的跨事件信息构建与利用

以 ACE2005 数据集为代表的句子级别数据集被主流的事件抽取研究广泛采用,但随着真实世界文本内容的复杂性不断增加,事件要素信息越来越多地跨越出现在一个文档中的多个句子中。因此,研究同时适用于句子级别和文档级别事件要素抽取的通用模型成为新的研究热点和挑战。与上一章节不同,该研究遵循章节2.2所述的关于事件要素抽取的第二种定义,即不使用实体提及的标注信息,而在预测要素类型的同时识别出其文段范围。

现有通用文本级别事件要素抽取模型存在各自的局限性,而基于多词元链接的建模方法能够有效避免这些模型的内在局限,并展现出良好的性能潜力。因此,本章研究利用多词元链接方法进行通用文本级别的事件要素抽取。然而该方法中的单一事件建模方式缺乏对跨事件依赖关系的利用,多事件建模方式需要更长的要素类型序列编码和额外的链接操作,增加了预测推理的难度。为此,本章提出了一种基于分离-融合范式的跨事件信息构建与利用方法,并相应提出了一种新的多词元链接模型,避免了当前方法的各自局限,且实现了从不同级别文本中高效利用跨事件依赖信息。实验结果证明了本章提出的模型在所有句子级和文档级数据集上都优于目前性能最优的事件要素抽取模型。

5.1 引言

最近,基于抽取式^[95,129-131]和生成式^[123,125,154]的提示学习方法在同时适用于句子级别和文本级别的事件要素抽取任务上取得了显著的性能提升。然而,前者在抽取具有相同要素类型的多个事件要素时,受到提示模版中预先设定的重复要素槽位数量的限制,后者在抽取长距离事件要素的性能表现不佳。此外,这两类方法的性能表现均较为依赖于提示模版的设计质量。

因此,不同于基于提示学习的方法,最近的一些研究^[74,155,156]将输入文本和所有涉及的要素类型串接成自然语言序列,然后联合编码,并进一步实施多词元¹链接操作,从而实现事件要素抽取或通用信息抽取的建模。其中,所有事件要素文段和要素类型的并行链接操作能够直接高效地抽取具有相同要素类型的多个事件要素,并提升

¹ “词元”对应于英文表达中的“token”,这里特指经过预训练语言模型分词器分词后得到的语义单元。

长文本中分散的事件要素间的交互。此外，该类方法不需要精心设计的提示模版。根据每次同时抽取的事件数量，可以进一步将该类方法划分为两种建模方式：单事件抽取和多事件抽取。

第一种方式的研究^[74,156]将待抽取文本和只在目标事件中出现的要素类型串接后作为输入，然后实施两种链接操作去建模目标事件。如图5.1所示，其中被红色标记的为事件触发词，被黄色标记的为事件要素类型，不同颜色的箭头表示不同种类的链接操作，事件要素“James McDade”在 *Die* 事件和 *Attack* 事件中扮演了不同的要素类型，其抽取建模过程是分离的。虽然这种单事件抽取的方式便于进行简单的链接预测，但其忽略了建模不同事件间的显著关联依赖^[122,130,131]。

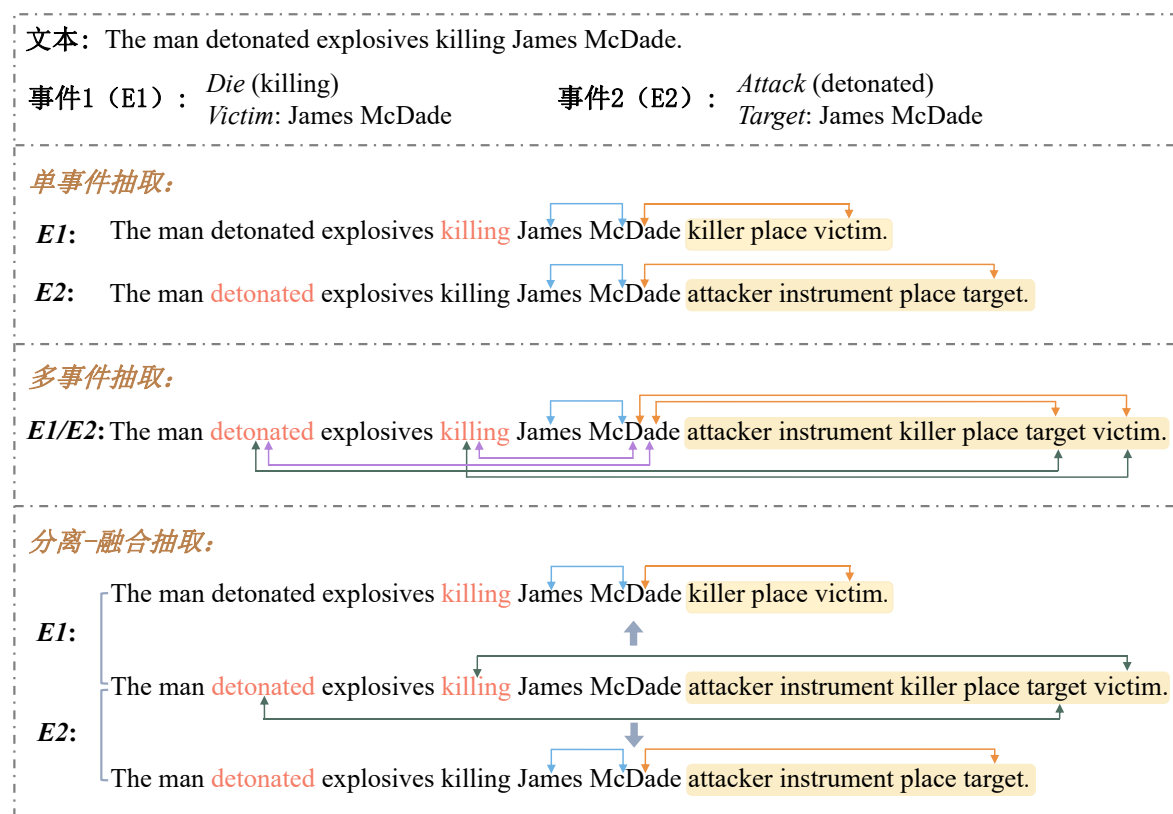


图 5.1 基于不同建模方式的多词元链接事件要素抽取示意图

因此，第二种方式的研究^[155]一次性抽取多个事件的事件要素。然而，两个新的问题随之而来。首先，与输入文本串接的不再是只在单个事件中出现的要素类型，而是在所有事件中出现的要素类型，这使得通过链接预测选择出正确的要素类型变得更加困难。其次，与单事件抽取方式相比，其需要两种额外的链接操作来确定抽取的事件要素文段和对应的要素类型归属的事件触发词，导致预测阶段积累更多的错误。图5.1展示了同时从 *Die* 事件和 *Attack* 事件中抽取事件要素“James McDade”时，其

额外增加的触发词-要素文段和触发词-要素类型链接操作。

为了利用上述两种抽取方式的优点，本章设计了一种基于分离-融合范式的跨事件信息构建与利用方法，其先分离跨事件信息获取和事件要素抽取的过程，然后将获取的跨事件信息融合到最终的事件要素抽取中。通过这种分离-融合范式，最终的事件要素抽取可以同时保留单事件抽取方式的简单链接预测和多事件抽取方式的跨事件依赖关系建模能力。图5.1说明了这一范式，其中中间部分从事件要素抽取过程中分离出来，通过触发词-要素类型链接获取跨事件信息，向上和向下的箭头表示跨事件信息的融合。

遵循设计的范式，本章提出了一种新颖的多词元链接事件要素抽取模型去获取分离的跨事件信息并进一步两阶段地融合到最终的事件要素预测 (**Separate acquisition of cross-event information and Two-fold Fusion, Sep2F**)。为了分离跨事件信息获取和事件要素抽取过程，本章设计了两个多词元链接模块。具体来说，针对多事件中的每个事件，引入一个链接模块来连接事件触发词和其中出现的要素类型。因此，不同事件触发词的特征表示并行聚合了在各自事件中出现的要素类型，提供了关键的跨事件信息。同时，引入另一个链接模块抽取目标事件的事件要素。其利用两种链接操作获取事件要素的文段范围和对应的要素类型。在此基础上，本章提出了一个两阶段融合模块将获取的跨事件信息融合到目标事件的事件要素抽取中。具体地，其首先动态融合来自上述两个链接模块的文本特征表示。然后，利用融合的文本特征表示获取跨模块词元链接分数。该链接分数被进一步融合到最终的预测中。这两阶段的融合过程相互影响，提供了显著的性能贡献。本章的主要贡献可以概括如下：

- 提出了一种新的基于分离-融合范式的跨事件信息构建与利用方法，其能够同时建模跨事件依赖信息和保留单事件抽取方式的优点。
- 基于分离-融合范式，提出了一种新的多词元链接事件要素抽取模型，其避免了当前通用事件要素抽取方法的各自局限，且实现了从不同级别文本中通用高效地利用跨事件信息。具体地，设计了两个链接模块，以分别获取跨事件信息和抽取目标事件的事件要素。此外，引入了一个两阶段融合模块，以确保在要素抽取过程中充分利用到获取的跨事件信息。
- 在四个不同文本级别的数据集上进行了充分的实验，实验结果表明本章所提的模型均显著超越目前性能最优的事件要素抽取基线方法。

5.2 方法设计

本章使用 (X, T, C) 表示每个数据实例，其中 X 是输入文本， T 是目标事件， C 表示在文本 X 中除目标事件 T 以外的其他事件。具体地， T 进一步表示为 (e, t, \mathcal{R}^e) ，其中 e 是事件类型， t 是事件触发词， \mathcal{R}^e 为只在事件 e 中出现的要素类型集合。类似地， C 可表示为 $\{(\tilde{e}_i, \tilde{t}_i, \mathcal{R}^{\tilde{e}_i}) \mid i \leq |C|\}$ ，其为目标事件 T 提供了跨事件信息。基于此，本章所提模型旨在抽取目标事件 T 的要素集合 \mathcal{A} ，其集合 \mathcal{A} 中的每个事件要素 $a^{(r)}$ 表示要素类型为 $r \in \mathcal{R}^e$ 的 X 中文本片段。

如图5.2所示，本章提出的模型由三个模块组成：多事件链接构造、目标事件链接构造和两阶段融合，其遵循了本章设计的分离-融合范式。具体地，前两个模块将跨事件信息的获取和目标事件的要素抽取过程分离，最后一个模块实现了获取的跨事件信息的融合。接下来，本章将具体介绍各个模块。

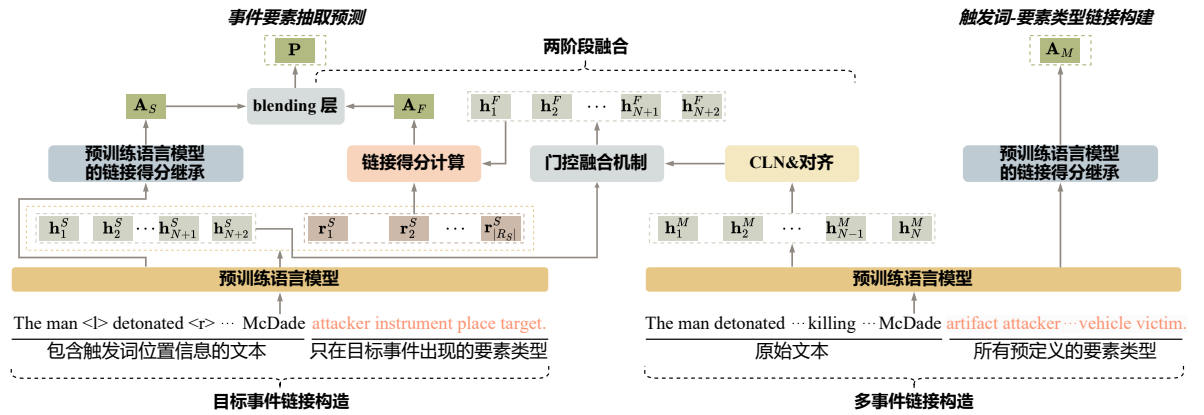


图 5.2 Sep2F 模型架构图

5.2.1 多事件链接构造

本模块通过聚合多事件的要素类型信息来获取跨事件信息，其中多事件包含目标事件 T 和所有其他事件 C 。为了实现这种信息聚合，本模块并行地建立不同的事件触发词和它们对应的事件要素类型之间的链接。首先，联合编码待抽取文本和给定数据集中预定义的所有要素类型。然后，根据事件中的触发词-要素类型链接引入标签矩阵和得分矩阵。最后，给出相应的训练损失。

编码 首先，本章利用对应的事件要素名称描述每个事件要素类型，即单个自然语言描述的词。对于名称中包含多个词的少数要素类型，本模型在训练过程中学习额外

的特殊词元进行表示。然后，拼接给定数据集中预定义的所有要素类型的名称描述，记作 R_M 。在此之后，使用预训练语言模型（PLM）对序列 R_M 和输入文本 X 进行联合编码如下：

$$\mathbf{E}^M = (\mathbf{h}_1^M, \dots, \mathbf{h}_N^M, \mathbf{r}_1^M, \dots, \mathbf{r}_{|R_M|}^M) = \text{PLM}(X \oplus R_M) \quad (5.1)$$

其中 \mathbf{h}_n^M ($1 \leq n \leq N$) 和 \mathbf{r}_n^M ($1 \leq n \leq |R_M|$) 分别为输入文本 X 和事件要素序列 R_M 中的词元向量。

标签矩阵 为了学习到多事件各自的要素类型信息，本模块设计了标签矩阵 $\mathbf{L}_M \in \mathbb{B}^{(N+|R_M|) \times (N+|R_M|)}$ 。对于 T 或 C 中的每个事件，假设其触发词的开始和结束词元向量分别是 \mathbf{h}_i^M 和 \mathbf{h}_j^M 。进一步，对于该事件中出现的每个要素类型，假设其对应的词元向量是 \mathbf{r}_k^M 。然后，实施链接操作来连接触发词和该要素类型。具体来说，分别构造链接词元对 $(i, k + N)$ 和 $(k + N, j)$ ，并相应地设置 $L_M[i][k + N]$ 和 $L_M[k + N][j]$ 为“True”。对于没有链接的词元对，本模块将 \mathbf{L}_M 中的相应值标记为“False”。图5.3展示了 \mathbf{L}_M 的一个示例，其中被红色标记的为事件触发词词元，被黄色标记的为所有预定义的事件要素类型词元，“-”表示“False”。根据图5.3可知，“detonated”触发的事件中存在要素类型 *Target*，而“killing”触发的事件中存在要素类型 *Victim*。

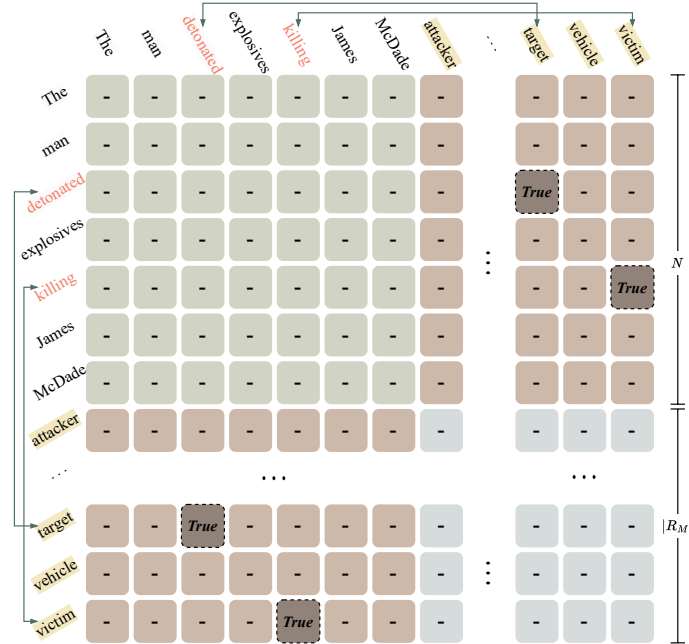


图 5.3 标签矩阵 \mathbf{L}_M 示例图

得分矩阵 受 Tang 等人^[157]研究工作的启发，本模块继承基于 Transformer 架构的

预训练语言模型的多头自注意力结果作为不同词元对之间的链接分数。具体来说，根据选定的预训练语言模型的最后一层，获取其未经过 softmax 归一化的多头自注意力权重，并进行均值计算：

$$\mathbf{A}_M = \frac{1}{P} \sum_p \frac{\mathbf{Q}_p \mathbf{K}_p^\top}{\sqrt{d_h}} \quad (5.2)$$

其中 P 为多头的数目， d_h 为多头自注意机制中查询向量和键向量的维度， \mathbf{Q}_p 和 \mathbf{K}_p 分别为查询矩阵和键矩阵， $\mathbf{A}_M \in \mathbb{R}^{(N+|R_M|) \times (N+|R_M|)}$ 表示多事件的触发词-要素类型链接得分。

训练损失 本模块使用训练损失 \mathcal{L}_{TR} 去指导模型学习不同事件触发词及其相应出现的要素类型之间的链接，如下所示：

$$\begin{aligned} \mathcal{L}_{\text{TR}} = & - \frac{1}{(N + |R_M|)^2} \sum_i \sum_j (L_{i,j}^M \log \sigma(A_M[i][j]) + \\ & (1 - L_{i,j}^M) \log (1 - \sigma(A_M[i][j]))) \end{aligned} \quad (5.3)$$

其中 $\sigma(\cdot)$ 为 sigmoid 激活函数。当 $L_M[i][j]$ 为 “True” 时，设置 $L_{i,j}^M$ 为 1，否则设置 $L_{i,j}^M$ 为 0。

5.2.2 目标事件链接构造

为了抽取给定目标事件 (e, t, \mathcal{R}^e) 的事件要素，本模块对输入文本和只在事件 e 中出现的要素类型进行联合编码，然后定义对应的多词元链接标签矩阵和得分矩阵，并给出相应的训练损失计算。

编码 与多事件链接构造模块不同，本模块拼接所有 \mathcal{R}^e 集合中的事件要素类型的名称描述，记作 R_S 。然后，遵循 Ma 等人^[95]的研究工作在文本 X 中插入两个特殊词元 $\langle \ell \rangle$ 和 $\langle r \rangle$ ，以便标记触发词 t 的位置：

$$X_S = (x_1, \dots, \langle \ell \rangle, t, \langle r \rangle, \dots, x_{|X|}) \quad (5.4)$$

在此之后，将包含触发词位置信息的文本 X_S 和要素类型序列 R_S 进行串接，并利用

另一个预训练语言模型对其进行编码：

$$\mathbf{E}^S = (\mathbf{h}_1^S, \dots, \mathbf{h}_{N+2}^S, \mathbf{r}_1^S, \dots, \mathbf{r}_{|R_S|}^S) = \text{PLM}(X_S \oplus R_S) \quad (5.5)$$

其中 \mathbf{h}_n^S ($1 \leq n \leq N+2$) 和 \mathbf{r}_n^S ($1 \leq n \leq |R_S|$) 分别为 X_S 和 R_S 中的词元向量。

标签矩阵 为了标记给定目标事件中的事件要素文段范围和要素类型，本模块定义了标签矩阵 $\mathbf{L}_S \in \mathbb{B}^{(N+|R_S|+2) \times (N+|R_S|+2)}$ 。对于目标事件中的每个事件要素，假设其文段的开始和结束词元向量分别是 \mathbf{h}_i^S 和 \mathbf{h}_j^S ，其要素类型对应的词元向量是 \mathbf{r}_k^S 。基于此，首先构造链接词元对 (i, j) 和 (j, i) 并设置 $L_S[i][j]$ 和 $L_S[j][i]$ 为 “True”，以用来标记事件要素的文段。同时，通过要素文段-要素类型链接操作来标记事件要素对应的要素类型信息。相应地，构造两个链接对 $(i, k+N+2)$ 和 $(k+N+2, j)$ ，并设置 $L_S[i][k+N+2]$ 和 $L_S[k+N+2][j]$ 为 “True”。对于没有链接的词元对，将 \mathbf{L}_S 中的相应值设置为 “False”。图5.4展示了 \mathbf{L}_S 的一个示例，其中被红色标记的为事件触发词词元，被黄色标记的为只在该事件触发词对应的事件类型中出现的事件要素类型词元，“-” 表示 “False”。根据图5.4可知，在 “detonated” 触发的事件中，存在事件要素 “James McDade”，其对应的要素类型为 *Target*。

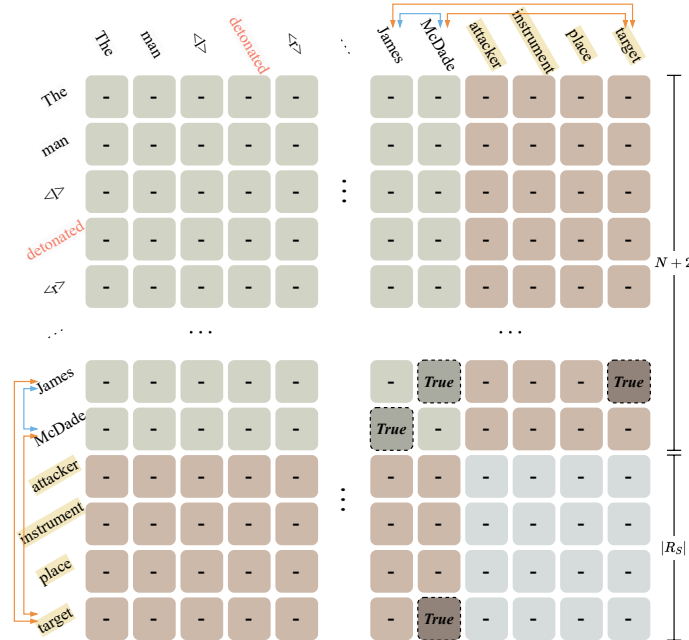


图 5.4 标签矩阵 \mathbf{L}_S 示例图

得分矩阵 与多事件链接构造模块中的公式5.2相同，本模块同样根据所使用的预训练语言模型的最后一层，利用其未通过 softmax 归一化的平均多头自注意权重作为

多词元间的链接分数，表示为 $\mathbf{A}_S \in \mathbb{R}^{(N+|R_S|+2) \times (N+|R_S|+2)}$ 。

训练损失 为了利用到获取的跨事件信息，首先采用两阶段融合模块（其将在下一模块中作详细介绍）来得到最终的词元链接预测矩阵：

$$\mathbf{P} = \text{TFF}(\mathbf{E}^M, \mathbf{E}^S, \mathbf{A}_S) \quad (5.6)$$

其中 TFF 指两阶段融合模块。然后，可以得到训练损失 \mathcal{L}_{EAE} 如下：

$$\mathcal{L}_{\text{EAE}} = - \frac{1}{(N + |R_S| + 2)^2} \sum_i \sum_j (L_{i,j}^S \log P[i][j] + (1 - L_{i,j}^S) \log (1 - P[i][j])) \quad (5.7)$$

其中当 $L_S[i][j]$ 为 “True” 时，设置 $L_{i,j}^S$ 为 1，否则设置 $L_{i,j}^S$ 为 0。

5.2.3 两阶段融合

本模块利用聚合了跨事件要素类型信息的文本表示来增强目标事件的要素抽取。具体地，首先动态融合来自上述两个链接模块的文本表示，即第一阶段融合。然后，使用融合的文本表示来获得目标事件的另一个词元链接得分矩阵，其不同于 \mathbf{A}_S 。这两个词元链接得分矩阵进一步被融合为最终的链接预测，即第二阶段融合。

第一阶段融合 由于 \mathbf{E}^M 中的文本向量涉及多事件信息，本模块首先利用条件层归一化（Conditional Layer Normalization, CLN）^[158,159] 来获取与目标事件（触发词）相关的上下文向量。对于目标事件触发词 t ，本模块简单地将其输入到一个参数冻结的预训练语言模型中，并使用得到的第一个词元向量来表示该触发词，记作 \mathbf{t} 。然后，对于 \mathbf{E}^M 中的每个词元向量 \mathbf{h}_n^M ($1 \leq n \leq N$)，可得到相应的向量 $\hat{\mathbf{h}}_n^M$ ($1 \leq n \leq N$) 如下：

$$\boldsymbol{\alpha}_t = \mathbf{t} \mathbf{W}_\alpha + \mathbf{b}_\alpha \quad (5.8)$$

$$\boldsymbol{\beta}_t = \mathbf{t} \mathbf{W}_\beta + \mathbf{b}_\beta \quad (5.9)$$

$$\hat{\mathbf{h}}_n^M = \text{CLN}(\mathbf{h}_n^M, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \boldsymbol{\alpha}_t \odot \left(\frac{\mathbf{h}_n^M - \mu}{\sigma} \right) + \boldsymbol{\beta}_t \quad (5.10)$$

其中 $\mathbf{W}_\alpha \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{W}_\beta \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{b}_\alpha \in \mathbb{R}^{d_1}$ 和 $\mathbf{b}_\beta \in \mathbb{R}^{d_1}$ 为可训练的参数， μ 和 σ 分别为根据 \mathbf{h}_n^M 中的向量元素计算得到的平均值和标准偏差， d_1 为向量 \mathbf{t} 的维度。然后，

在 $(\hat{\mathbf{h}}_1^M, \dots, \hat{\mathbf{h}}_N^M)$ 中插入零向量 $\mathbf{0}$ ，以便与文本向量 \mathbf{E}^S 进行对齐：

$$(\bar{\mathbf{h}}_1^M, \dots, \bar{\mathbf{h}}_{N+2}^M) = (\hat{\mathbf{h}}_1^M, \dots, \mathbf{0}, \hat{\mathbf{h}}_i^M, \dots, \hat{\mathbf{h}}_j^M, \mathbf{0}, \dots, \hat{\mathbf{h}}_N^M) \quad (5.11)$$

其中 i 和 j 对应于触发词 t 的词元向量在 $(\mathbf{h}_1^S, \dots, \mathbf{h}_{N+2}^S)$ 中的开始和结束位置。最后，动态融合不同的文本向量。给定两个词元向量 \mathbf{h}_n^S 和 $\bar{\mathbf{h}}_n^M$ ($1 \leq n \leq N+2$)，本模块使用门控融合机制以获取词元向量 \mathbf{h}_n^F ：

$$\mathbf{g}_n = \sigma([\mathbf{h}_n^S; \bar{\mathbf{h}}_n^M] \mathbf{W}_G) \quad (5.12)$$

$$\mathbf{h}_n^F = \mathbf{g}_n \odot \mathbf{h}_n^S + (1 - \mathbf{g}_n) \odot \bar{\mathbf{h}}_n^M \quad (5.13)$$

其中 $\sigma(\cdot)$ 为 sigmoid 函数， $\mathbf{W}_G \in \mathbb{R}^{2d_1 \times d_1}$ 为可训练学习的矩阵， \odot 表示元素依次相乘操作。

第二阶段融合 首先，串接融合的文本向量 $(\mathbf{h}_1^F, \dots, \mathbf{h}_{N+2}^F)$ 和要素类型对应的词元向量 $(\mathbf{r}_1^S, \dots, \mathbf{r}_{|R_S|}^S)$ ，表示为 \mathbf{F} 。然后，本模块计算另一个词元链接得分矩阵 $\mathbf{A}_F \in \mathbb{R}^{(N+|R_S|+2) \times (N+|R_S|+2)}$ 如下：

$$\mathbf{F}_Q = \mathbf{F} \mathbf{W}_Q, \mathbf{F}_K = \mathbf{F} \mathbf{W}_K \quad (5.14)$$

$$\mathbf{A}_F = \mathbf{F}_Q \mathbf{F}_K^\top \quad (5.15)$$

其中 $\mathbf{W}_Q \in \mathbb{R}^{d_1 \times d_2}$ 和 $\mathbf{W}_K \in \mathbb{R}^{d_1 \times d_2}$ 为可训练学习的矩阵。最后，根据 \mathbf{A}_F 和章节 5.2.2 中得到的得分矩阵 \mathbf{A}_S ，使用 blending 层^[160] 融合如下：

$$\mathbf{P} = \sigma(\mathbf{A}_S + \mathbf{A}_F - \tau) \quad (5.16)$$

其中 τ 为可训练的参数。

5.2.4 训练和推理

训练 本章提出模型的整体训练损失可根据如下得到：

$$\mathcal{L} = \alpha \mathcal{L}_{\text{EAE}} + (1 - \alpha) \mathcal{L}_{\text{TR}} \quad (5.17)$$

Algorithm 1 目标事件的要素推理**Input:** 预测的链接对集合 $\mathcal{C}((s_k, e_k)), k \in (1, \dots, |\mathcal{C}|)$.**Output:** 事件要素集合 \mathcal{A} .

```

1:  $cand\_span = set()$ .
2:  $start2r\_dict = \{\}, end2r\_dict = \{\}$ .
3: for  $(s_k, e_k) \in \mathcal{C}$  do
4:   if  $s_k \leq N + 2$  and  $e_k \leq N + 2$  then
5:     将元组  $(s_k, e_k)$  添加到集合  $cand\_span$ 
6:   else if  $s_k \leq N + 2$  and  $e_k > N + 2$  then
7:     将键值对  $(s_k, e_k - N - 2)$  添加到字典  $start2r\_dict$ 
8:   else if  $s_k > N + 2$  and  $e_k \leq N + 2$  then
9:     将键值对  $(e_k, s_k - N - 2)$  添加到字典  $end2r\_dict$ 
10:  end if
11: end for
12: for  $(s, e) \in cand\_span$  do
13:   for  $r \in (set(start2r\_dict[s]) \cap set(end2r\_dict[e]))$  do
14:     将元组  $(s, e, r)$  添加到集合  $\mathcal{A}$ 
15:   end for
16:   for  $r \in (set(start2r\_dict[e]) \cap set(end2r\_dict[s]))$  do
17:     将元组  $(e, s, r)$  添加到集合  $\mathcal{A}$ 
18:   end for
19: end for

```

其中 α ($0 < \alpha < 1$) 为权重参数。

推理 基于词元链接预测矩阵 \mathbf{P} ，对于其中的每一个链接对 (i, j) ，如果对应预测值 $P[i][j]$ ($1 \leq i, j \leq N + |R_S| + 2$) 超过设定的阈值参数 δ ，则将该链接对添加到链接对集合 \mathcal{C} 中。然后，使用 \mathcal{C} 作为输入，运行如算法1所示的推理算法，得到目标事件的要素集合 \mathcal{A} 。对于 \mathcal{A} 中的每个元素 $(start, end, role)$ ， $start$ 和 end 分别表示要素的开始和结束位置， $role$ 为对应的要素类型在 R_S 中的索引。

5.3 实验评估

5.3.1 实验设置

1. 数据集 为了评估所提模型，本章将在以下数据集上进行实验：

- ACE2005^[138] - 该数据集详细介绍见章节3.3.1。本章实验使用其事件标注数据进行句子级别事件要素抽取的评估。在数据预处理上，与章节4.4.1不同，本章遵循 PAIE^[95] 和 TabEAE^[131] 等通用事件要素抽取模型的设置，使用 Du 等人^[72] 的预处理脚本。该脚本遵循 DyGIE++^[161] 模型的设置。

- RAMS^[93] - 该文档级别事件要素抽取数据集来源于新闻文章。与原始数据集将同一上下文中的多个事件视为不同的实例不同，本章实验遵循 TabEAE 的预处理过程，聚合每个实例的同一上下文中出现的其他事件的标注信息。
- WikiEvents^[94] - 该文档级别数据集来源英文维基百科。本章实验遵循 PAIE 的预处理程序，采用以每个目标事件触发词为中心的预定义窗口去截断过长的上下文内容，避免超出预训练语言模型的输入长度限制。
- MLEE^[162] - 该文档级别事件要素抽取数据集基于生物医学领域的英文出版物摘要标注得到。在预处理方面，遵循 TabEAE 参照 Trieu 等人^[163]的预处理设置。此外，由于 MLEE 数据集中不存在验证集，本章实验遵循 TabEAE 使用训练集调整模型的超参数。

对于所有文档级别数据集，遵循 He 等人^[131]设置预定义的窗口长度为 250，将每个文档拆分为多个上下文片段。本章实验所用数据集的具体统计数据如表 5.1 所示。

表 5.1 数据集详细统计信息

数据集	ACE2005	RAMS	WikiEvents	MLEE
# 事件				
训练集	4,202	7,329	3,241	4,442
验证集	450	924	345	-
测试集	403	871	365	2,200
# 事件要素				
训练集	4,859	17,026	4,552	5,786
验证集	605	2,188	428	-
测试集	576	2,023	566	2,764
# 事件类型	33	139	50	23
# 要素类型	22	65	59	8

2. 评价指标 遵循之前通用文本级别事件要素抽取研究工作^[95,131]，本章采用两个指标来衡量模型性能。(1) 要素识别 F1 得分 (Argument Identification F1 score, Arg-I): 当其事件要素文段的边界与事件中任意一个标注的要素相匹配，则认为该事件要素被正确识别。基于该识别准则，本章进一步计算对应的 F1 值评估要素识别性能，其 F1 值计算参照章节 3.3.1。(2) 要素分类 F1 得分 (Argument Classification F1 score, Arg-C): 当其事件要素文段的边界和要素类型都与事件中任意一个标注的要素相匹配，则认为该事件要素被正确分类。基于该分类准则，本章进一步计算对应的 F1 值评估要素

分类性能，其 F1 值计算参照章节3.3.1。在实验中，使用三个不同的随机种子训练本章提出的 Sep2F 模型，并通过平均值计算得到相应的性能结果。

3. 实验配置 为了便于与最近的研究工作进行公平比较，本章实验在模型中使用 RoBERTa-base 和 RoBERTa-large^[105] 作为预训练语言模型，其分别基于 Huggingface 上发布的“FacebookAI/roberta-base”文件²和“FacebookAI/roberta-large”文件³进行参数初始化。基于此两种模型，均采用配备了线性学习率调度器的 Adam 优化器和比率为 0.1 的 warmup 策略。 δ 为等式5.7中二分类预测值 $P[i][j]$ 的推断阈值，因此实验中遵循常规的二分类设置，在所有数据集上均设置 δ 为 0.5。在 RAMS 数据集上设置训练轮次为 50，在其他数据集上设置训练轮次为 100。在 ACE2005 数据集上设置批数据大小为 16，在其他数据集上设置批数据大小为 8。对于 ACE2005， d_2 设置为 32，而在其他数据集上 d_2 设置为 64。此外，基于每个数据集的验证集，使用网格搜索调整各自模型的学习率和训练权重 α 。具体地，在区间 $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ 内调整权重 α ，在区间 $\{3, 4, 5\} \times 10^{-5}$ 和 $\{1, 2, 3\} \times 10^{-5}$ 内分别搜索基于 RoBERTa-base 和 RoBERTa-large 的模型的最佳学习率。表5.2列出了在不同数据集上最终选择的训练权重和学习率，其中表格上半部分为基于 RoBERTa-base 的模型参数详情，表格下半部分为基于 RoBERTa-large 的模型参数详情。对于基于 RoBERTa-base 的模型，不同数据集上的训练均在相同实验服务器上完成，具体配置为：CPU 型号 Intel(R) Xeon(R) Gold 5320，主频 2.20GHz；内存 256G；GPU 计算卡型号 NVIDIA GeForce RTX 3090，显存 24G。对于基于 RoBERTa-large 的模型，其具体配置为：CPU 型号 Intel(R) Xeon(R) Silver 4314，主频 2.40GHz；内存 128G；GPU 计算卡型号 NVIDIA A100 PCIE，显存 40G。本章实验均基于 Pytorch^[139] 框架完成。

表 5.2 超参数设置

超参数	ACE2005	RAMS	WikiEvents	MLEE
学习率	3e-5	5e-5	4e-5	4e-5
权重 α	0.7	0.7	0.8	0.6
学习率	2e-5	3e-5	2e-5	1e-5
权重 α	0.8	0.5	0.7	0.8

4. 基线模型 本章选择以下模型作为与所提的 Sep2F 模型进行实验比较的基线，其均在句子级别和文档级别数据集上评估了事件要素抽取任务的性能：

²<https://huggingface.co/FacebookAI/roberta-base>

³<https://huggingface.co/FacebookAI/roberta-large>

- EEQA^[72] - 该模型将事件要素抽取任务转化为机器阅读理解问题进行解决。
- BART-Gen^[94] - 该模型将事件要素抽取建模为基于提示模版的条件生成任务。
- PAIE^[95] - 该模型提出了抽取式的提示学习方法，其利用要素类型对应的槽位表示查询对应的要素文段。
- APE(Single)^[154] - 该模型通过将提示模版中的要素类型映射至对应的实体类型，实现了跨事件共享知识的学习。
- PGAD^[164] - 该模型引入了扩散模型，以获取多样化的上下文感知提示表示。
- EDGE^[130] - 该模型同时设计了事件内和跨事件的图，以丰富要素类型语义信息。
- TabEAE^[131] - 该模型将多事件要素抽取转换为表格生成，以捕捉事件间的关联。

最近，基于情境学习 (In-Context Learning, ICL)^[90] 的 ChatGPT 在各种自然语言处理任务中表现出色。但是，目前缺乏对不同事件要素抽取数据集的全面评估，尤其是文档级别数据集。因此，本章按照 Han 等人^[89] 的设置，构建 5-shot ICL 提示模版作为每个测试样本的输入，并使用 OpenAI 接口⁴获取事件要素抽取的结果。具体地，本章使用两个版本的 ChatGPT 评估事件要素抽取的性能：*gpt-3.5-turbo* 和 *gpt-4*，其分别由 *gpt-3.5* 和 *gpt-4* 模型构建。

此外，UnifiedEAE^[165] 和 APE^[154] 等研究工作同时利用了多个数据集提升目标数据集的性能，以受益于额外的数据资源。因此，本章未将其作为主实验结果的比较基线，而是另作单独的分析。

5.3.2 性能结果

表5.3展示了在四个数据集上的事件要素抽取实验结果，其中粗体表示在各自数据集上的最优性能，下划线表示在各自数据集上的次优性能结果，预训练模型所在列中的“b”和“l”分别表示对应的为 base 版本和 large 版本。从表5.3中可以观察到，本章提出的基于 RoBERTa-large 的模型在所有数据集上都取得了当前最优的性能。具体地，与之前性能最优的 TabEAE 模型相比，其在 ACE2005、RAMS、WikiEvent 和 MLEE 数据集上，分别在 Arg-I/Arg-C F1 指标上取得了 1.6%/2.0%、1.4%/1.0%、2.6%/2.8%

⁴<https://platform.openai.com/>

和 2.4%/2.5% 的性能提升。而相比于使用了 base 版本预训练模型的最优基线，本章提出的基于 RoBERTa-base 的模型在 ACE2005、RAMS 和 WikiEvent 数据集上分别在 Arg-I/Arg-C F1 指标上取得了 0.9%/2.9%、1.4%/2.4% 和 3.7%/4.1% 的性能提升。此外，本章所提的 base 模型在 WikiEvents 和 MLEE 数据集上超越了所有使用 large 版本预训练模型的基线，且在 ACE2005 和 RAMS 数据集上相比 large 基线模型也展现了具有竞争力的性能。因此，可以得知本章提出的 Sep2F 模型在处理不同文本级别的事件要素抽取任务时均表现出色。

表 5.3 四种数据集上的事件要素抽取性能结果

模型	预训练模型	ACE2005		RAMS		WikiEvents		MLEE	
		Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C	Arg-I	Arg-C
ChatGPT	GPT-3.5	35.6	30.0	25.1	19.3	13.6	11.6	15.7	11.8
ChatGPT	GPT-4	37.5	33.2	23.8	20.9	16.7	15.3	16.9	14.9
BART-Gen	BART-b	59.6	55.0	50.9	44.9	47.5	41.7	-	-
PAIE	BART-b	73.6	69.8	54.7	49.5	68.9	63.4	-	-
APE(Single)	BART-b	74.1	70.1	54.8	49.6	66.2	62.1	-	-
PGAD	BART-b	74.1	70.5	53.7	45.9	69.6	64.1	-	-
EDGE	BART-b	75.3	70.6	55.2	49.7	68.2	62.8	-	-
Sep2F	RoBERTa-b	76.2	73.5	56.6	52.1	<u>73.3</u>	<u>68.2</u>	<u>76.8</u>	<u>75.6</u>
EEQA	RoBERTa-l	72.1	70.4	51.9	47.5	60.4	57.2	70.3	68.7
BART-Gen	BART-l	69.9	66.7	51.2	47.1	66.8	62.4	71.0	69.8
PAIE	BART-l	75.7	72.7	56.8	52.2	70.5	65.3	72.1	70.8
PAIE	RoBERTa-l	76.1	73.0	57.1	52.3	70.9	65.5	72.5	71.4
APE(Single)	BART-l	75.3	72.9	56.3	51.7	70.6	65.8	-	-
TabEAE	RoBERTa-l	<u>77.2</u>	<u>75.0</u>	<u>57.3</u>	<u>52.7</u>	71.4	66.5	75.1	74.2
Sep2F	RoBERTa-l	78.8	77.0	58.7	53.7	74.0	69.3	77.5	76.7

另外，可以观察到不同版本的 ChatGPT 在事件要素抽取任务上的性能都明显落后于当前有监督模型。尤其在处理更具挑战性的文档级别事件要素抽取任务时，性能差距通常会进一步扩大。因此，提升 ChatGPT 等大语言模型在事件要素抽取任务上的性能表现仍有待深入研究。

5.3.3 不同多词元链接模型性能比较

如章节 5.1 中所述，基于多词元链接操作的现有事件要素抽取模型主要存在两种方式：单事件抽取和多事件抽取。然而，这些方法将事件要素抽取任务作为端到端的通用信息抽取或事件抽取中的一个子任务进行建模。因此，其实验设置和训练数据集与单独处理事件要素抽取任务存在差异。为了更加公平地比较不同的多词元链接模型

的性能，本章根据这些多词元链接模型的建模方式设计以下变体模型：(1) **SingleE** 在目标事件链接构造模块中移除两阶段融合，以抽取单个事件的要素信息，对应于单事件抽取方式。(2) **SingleE-L** 在模型输入中串接给定数据集的所有预定义要素类型名称与待抽取文本，其他设置与 **SingleE** 相同。(3) **MultiE** 同时抽取文本中多个事件的事件要素，对应多事件抽取方式。具体地，将事件要素抽取转化为多个 < 事件触发词, 要素类型, 要素文段 > 三元组抽取问题，并参照 Tang 等人在 UniRel^[157] 模型中基于多词元组成的实体进行关系三元组抽取的代码实现⁵。由于本章实验在事件要素抽取中使用正确标注的事件触发词作为输入，MultiE 将在推理过程中根据标注信息纠正事件触发词的文段范围预测错误。此外，与 MultiE 输入文本串接的为只在待抽取事件中出现的要素类型名称。(4) **MultiE-L** 将所有预定义的要素类型名称与待抽取文本串接，其他设置与 MultiE 相同。

表 5.4 四种数据集上不同多词元链接模型在 Arg-C F1 (%) 指标上的结果对比

模型	ACE2005 [1.35]	RAMS [1.25]	WikiEvents [1.78]	MLEE [3.32]
SingleE	75.4	53.0	66.5	74.4
SingleE-L	72.9	49.1	55.8	73.0
MultiE	69.9	49.3	47.9	57.2
MultiE-L	68.1	40.5	26.5	8.4
Sep2F	77.0	53.7	69.3	76.7

表5.4为上述的多词元链接模型在四个数据集上关于 Arg-C F1 (%) 指标的实验结果，其均使用 RoBERTa-large 进行编码，且 “[.]” 表示每个对应数据集中单个数据实例存在的平均事件数目。首先，从表5.4中可以总结出本章提出的模型在四个数据集上的性能均超越了所有变体模型。接下来，给出详细比较如下：

- **SingleE vs Sep2F**: 与本章提出的 Sep2F 模型相比，未利用跨事件信息的 SingleE 的 Arg-C F1 指标在 ACE2005、RAMS、WikiEvents 和 MLEE 数据集上分别下降了 1.6%、0.7%、2.8% 和 2.3%。
- **MultiE vs Sep2F**: MultiE 在事件要素抽取任务上未表现出有竞争力的性能，尤其是在 WikiEvents 和 MLEE 数据集上，MultiE 和本章所提的 Sep2F 的性能存在显著差距。

⁵<https://github.com/wtangdev/UniRel>

- **SingleE vs SingleE-L**: 在不同数据集上 SingleE-L 的 Arg-C F1 指标均低于 SingleE, 其验证了在文本和要素类型名称间进行链接预测时, 串接更多的要素类型会降低预测的性能。此外, 图5.5进一步展示了这两个模型在不同数据集上要素类型串接长度和 Arg-C F1 指标的关联性, 其中“增加的要素类型长度”表示 SingleE-L 输入中拼接的要素类型序列相比于 SingleE 输入中拼接的最长要素类型序列增加了的长度, “Arg-C F1 下降值”表示 SingleE 和 SingleE-L 的性能差距。根据图5.5可以观察到, 由于 SingleE-L 与 SingleE 在 RAMS 和 WikiEvents 数据集上的要素类型串联比其他数据集更长, 导致其在这两个数据集上的 Arg-C F1 指标差异也会相应地更加显著。
- **MultiE vs MultiE-L**: 与 SingleE 和 SingleE-L 的比较结果相比, MultiE 和 MultiE-L 的性能差距进一步增加, 其增加程度和不同数据集上单个数据实例的平均事件数目密切相关。例如, MLEE 数据集上单个数据实例的平均事件数超过三个, 导致 MultiE-L 性能的大幅下降。
- **SingleE-L vs MultiE-L**: 尽管 MultiE-L 模型同时抽取多个事件的事件要素, 以便考虑不同事件间的关联性。但是, 当 MultiE-L 和 SingleE-L 保持相同的要素类型串接时, 其 Arg-C F1 指标在所有数据集上均低于 SingleE-L 模型。此结果表明, MultiE-L 无法高效地在单个数据实例中处理更多的词元链接预测。因此, 其跨事件信息带来的性能提升未能弥补更多累积错误导致的性能下降。此外, WikiEvents 和 MLEE 数据集上单个实例的平均事件数量相比其他数据集较多, 导致在这两个数据集上 SingleE-L 和 MultiE-L 的性能差异更大。

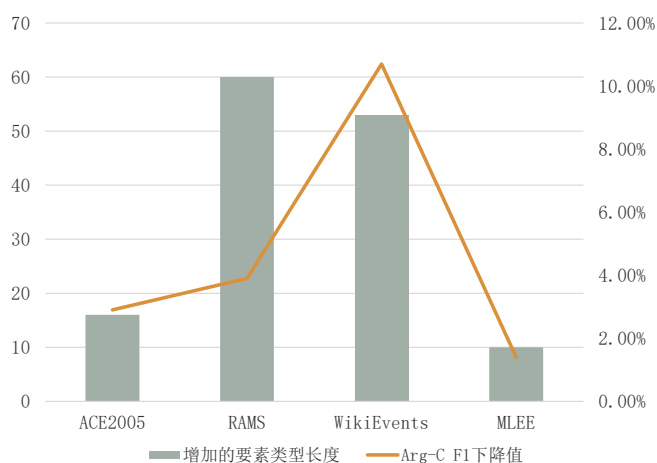


图 5.5 要素类型串接长度与 Arg-C F1 指标的关联性

5.3.4 单事件/多事件实例性能分析

遵循 He 等人^[131] 的实验设置，本节根据目标抽取事件所在上下文中存在的其他事件的数量（即 $|C|$ ），将每个数据集的测试实例分为两组。当 $|C| = 0$ 时，实例中只存在目标抽取事件。如果 $|C| > 0$ ，则实例中存在多个事件。然后，基于分组得到各自的统计和性能信息如表 5.5 所示，其中 $[\cdot]$ 中的值表示对应分组中测试实例的数量，Sep2F-b 和 Sep2F-l 分别表示基于 RoBERTa-base 和 RoBERTa-large 的 Sep2F 模型。值得注意的是，表中其他的基线模型均基于 RoBERTa-large 进行构建。从表 5.5 中可以观察到：（1）在两个分组中，基于 RoBERTa-large 的 Sep2F 模型在 Arg-C F1 指标上均超越了之前性能最优的基线。这种性能的提升可以归因于本章所提的分离-融合范式，其同时保留了单事件抽取方式和多事件抽取方式的优点。（2）基于 RoBERTa-base 的 Sep2F 模型在 WikiEvents 和 MLEE 数据集包含多个事件的实例分组上表现出了令人印象深刻的性能结果，其超越了两个最具性能竞争力的基于 RoBERTa-large 的 SingleE 和 TabEAE。这验证了本章提出的模型在处理具有多个事件的实例时的良好表现。（3）尽管 SingleE 和 PAIE 模型均未利用到跨事件信息，但与 PAIE 相比，SingleE 在所有数据集上均表现出更好的性能结果，其验证了基于多词元链接的模型解决事件要素抽取任务的潜力和优势。

表 5.5 四种数据集上单事件/多事件实例对应的 Arg-C F1 (%) 指标结果

模型	ACE2005		RAMS		WikiEvents		MLEE	
	$ C = 0$ [185]	$ C > 0$ [218]	$ C = 0$ [587]	$ C > 0$ [284]	$ C = 0$ [114]	$ C > 0$ [251]	$ C = 0$ [175]	$ C > 0$ [2025]
SingleE	74.8	76.0	52.9	53.3	68.4	65.5	83.1	73.8
MultiE	70.8	69.1	50.0	47.5	45.5	49.1	53.7	57.5
PAIE	71.0	73.9	52.7	52.1	65.3	65.4	78.9	70.1
TabEAE	73.4	76.1	52.9	52.5	67.3	66.2	81.1	73.6
Sep2F-b	72.8	74.0	52.1	52.0	66.8	69.0	85.5	74.9
Sep2F-l	75.4	78.3	53.2	54.6	68.8	69.5	84.3	76.1

5.3.5 消融研究

本章通过消融研究分析基于 RoBERTa-large 的 Sep2F 模型中不同组件的有效性，包括第一阶段融合（First-fold Fusion）、第二阶段融合（Second-fold Fusion）、Blending 层（Blending Layer）、文本中存在的其他事件信息（Surrounding Events C ）和训练损

失 \mathcal{L}_{TR} (Training Loss \mathcal{L}_{TR})。根据表5.6所示的在 Arg-C F1 (%) 指标上的消融结果, 所有组件均能提升本章提出的 Sep2F 模型的性能。具体地, 可以观察到: (1) 如果移除第一阶段融合或第二阶段融合, Sep2F 模型在各个数据集上的性能均将下降。这表明两个阶段的融合均对性能提升贡献显著。(2) 移除两个阶段中任意一阶段的融合, Sep2F 模型在各个数据集上, 尤其是 WikiEvents 和 MLEE 数据集, 取得与之前性能表现最优的事件要素抽取基线模型相当的 Arg-C F1 指标。这展现了本章提出的模型即使利用简单的融合方法, 仍然能在事件要素抽取任务上表现良好, 进一步验证了分离-融合范式的优势。(3) 当移除 Blending 层并直接对两个词元链接得分矩阵进行加和, Sep2F 模型的 Arg-C F1 指标在所有数据集上均下降了 1.0% 以上, 这表明 Blending 层可以较好地集成不同的词元链接预测分数。(4) 如果移除给定目标事件上下文中的其他事件 C , 并在多事件链接构建模块中保留目标事件触发词和在该事件中出现的要素类型之间的链接, 则可以发现 Sep2F 模型在所有数据集上的性能均出现下降, 从而明确地证明了跨事件依赖信息对于模型的性能增益。此外, 可以发现 Sep2F 模型在 ACE2005 和 RAMS 数据集上性能略有下降, 而在 WikiEvents 和 MLEE 数据集上则性能下降的更加显著。该观察结果可以归因于表5.4中展示的各个数据集中单个实例的平均事件数目的差异。与 ACE2005 和 RAMS 数据集相比, 本章提出的模型可以利用到 WikiEvents 和 MLEE 数据集上更多事件的跨事件信息。因此, 与其他两个数据集相比, 移除目标事件周围的其他事件信息会导致 Sep2F 模型在 WikiEvents 和 MLEE 数据集上更显著的性能下降。(5) 如果移除多事件链接构造模块中的训练损失 \mathcal{L}_{TR} (即设置公式5.17中的训练权重参数 α 为 1), 则意味着 Sep2F 模型无法通过调整 \mathcal{L}_{TR} 而利用到上下文中其他事件的要素类型信息。从消融结果可以看出, 由此丢失的跨事件信息使得模型的 Arg-C F1 指标在 ACE2005、RAMS、WikiEvents 和 MLEE 数据集上分别下降了 1.2%、1.8%、2.3% 和 1.9%。

表 5.6 四种数据集上关于 Arg-C F1 (%) 指标的消融研究结果

模型	ACE2005	RAMS	WikiEvents	MLEE
Sep2F	77.0	53.7	69.3	76.7
- First-fold Fusion	74.8	51.2	66.3	75.6
- Second-fold Fusion	72.5	50.2	67.8	74.2
- Blending Layer	75.9	52.2	66.6	75.9
- Surrounding Events C	76.5	52.6	66.2	74.3
- Training Loss \mathcal{L}_{TR}	75.8	51.9	67.0	74.8

5.3.6 与资源增强模型的性能比较

本章将 Sep2F 模型与聚焦于利用跨数据集知识的 UnifiedEAE 和 APE 模型作比较，其中 APE 还额外人工构建了学习重叠知识的提示模版。图5.6展示了性能比较结果，其中上半部分的图展示的为基于 base 版本预训练语言模型的比较，而下半部分的图展示的为基于 large 版本预训练语言模型的结果。在使用 base 版本的预训练语言模型时，本章提出的 Sep2F 模型在所有三个数据集上均优于 UnifiedEAE 和 APE 模型。此外，当使用 large 版本预训练模型时，Sep2F 模型在 ACE2005 和 WikiEvents 数据集上的性能均优于 APE，并在 RAMS 数据集上也取得与 APE 相当的 Arg-C F1 指标。因此，可以得知尽管参与比较的基线受益于额外的训练资源和人为设计的经验知识，本章提出的模型在使用不同版本预训练语言模型时仍能在绝大部分数据集上实现最佳性能，进一步证明了其在事件要素抽取任务上的优越性。

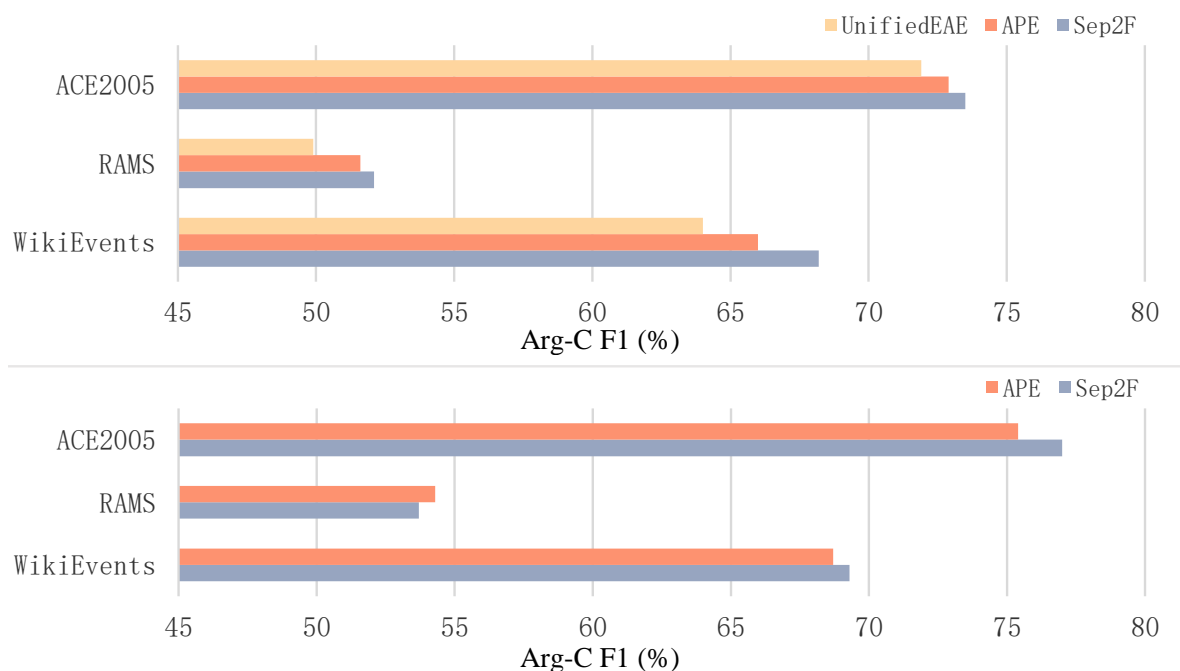


图 5.6 三种数据集上 Sep2F 与资源增强模型的性能结果对比

5.3.7 案例研究

本章节基于 WikiEvents 数据集中的一个具体实例进行定性的案例分析。图5.7展示了本章提出的 Sep2F 模型和 SingleE 模型在该实例上的事件要素抽取结果，其不同事件的触发词被不同颜色标记，且这两个模型均基于 RoBERTa-large 进行构建和训练。

可以观察到, SingleE 模型未能预测出由“booming”触发的 *Attack* 事件中的两个事件要素, 但 Sep2F 模型给出了正确的预测结果。本章推测其原因为 Sep2F 模型可以利用到由“killed”触发的事件中的要素类型信息。同样, 由于跨事件要素类型 *Destroyer* 提供了重要的依赖信息提示, Sep2F 模型避免了事件触发词“death”的干扰, 将“He”的要素类型识别为 *Killer* 而不是 *Victim*, 但 SingleE 模型给出了错误的抽取结果。

<p>The diagram illustrates semantic relations between entities and events in the text. Arrows indicate relations: 'Victim' from 'person' to 'killed'; 'Target' from 'killed' to 'bombing'; 'Killer' from 'Tsarnaev' to 'bombing'; 'Attacker' from 'Tsarnaev' to 'bombing'; 'Destroyer' from 'He' to 'destruction'; and 'Killer' from 'He' to 'death'.</p>		
事件类型: <i>Attack</i>	事件触发词: booming	
标注的Attacker: <i>Tsarnaev</i>	SingleE: <i>N/A</i> ×	Sep2F: <i>Tsarnaev</i>
标注的Target: <i>person</i>	SingleE: <i>N/A</i> ×	Sep2F: <i>person</i>
事件类型: <i>Die</i>	事件触发词: death	
标注的Killer: <i>He</i>	SingleE: <i>N/A</i> ×	Sep2F: <i>He</i>
标注的Victim: <i>N/A</i>	SingleE: <i>He</i> ×	Sep2F: <i>N/A</i>

图 5.7 WikiEvents 数据集上的案例分析

5.3.8 训练权重超参数分析

本章研究公式 5.17 中训练权重 α 对于 Sep2F 模型性能的影响。表 5.7 展示了不同 α 取值下 Sep2F 模型在各个数据集上的 Arg-C F1 性能指标, 可以观察到本章提出的 Sep2F 模型在不同数据集上均性能稳定, 证明了其良好的鲁棒性。

表 5.7 四种数据集上不同 α 取值对应的 Arg-C F1 (%) 指标结果

训练权重 α	ACE2005	RAMS	WikiEvents	MLEE
0.5	76.1	53.7	68.6	76.6
0.6	76.8	53.0	68.7	76.4
0.7	76.7	53.1	69.3	76.6
0.8	77.0	52.7	69.3	76.7
0.9	76.4	52.8	67.9	76.4

5.4 本章小结

本章提出了一种新颖的多词元链接事件要素抽取模型。具体地，提出了两个链接模块来分离跨事件信息的获取和目标事件的要素抽取过程。此外，提出了一种新的两阶段融合模块，以确保获取的跨事件信息能够有效地增益目标事件的要素抽取。因此，本章所提模型利用了跨事件依赖线索并保留了单事件抽取方式的优点。本章在四个不同级别的基准数据集上进行了充分实验，结果表明提出的模型在所有数据集上均取得了事件要素抽取的最优性能。

结论

1. 本文工作总结

作为信息抽取的关键技术，事件抽取任务在智能问答、事实核查、推荐系统等多样化自然语言场景扮演着至关重要的支撑作用，且以事件抽取技术为基础构建的事件知识图谱同样提供了广阔的应用前景。因此，事件抽取是提升文本结构化知识构建与应用的自动化和智能化水平的必然要求。目前，得益于深度学习技术的蓬勃发展，现有方法利用特定神经网络架构研究事件抽取的不同阶段任务，展现了显著的性能进步。但是，面临**数据不平衡依赖**、**实体类型过度依赖**和**跨事件依赖**等已有模型架构存在的共性数据挑战，这些方法缺乏通用的解决方案。为此，本文围绕事件抽取中的共性数据依赖进行了深入研究。接下来，将总结本文的研究成果和相应创新点如下：

(1) 针对事件检测中存在的数据不平衡依赖，首先研究了其影响事件检测性能的方式，相应引入能够有效消解其导致的性能下降的句子级别识别信息。在此基础上，提出了基于分类器自适应知识蒸馏的事件检测方法。该方法在事件识别增强网络中构建句子级别识别输入信息，通过分类器自适应的知识蒸馏技术增强事件检测网络自动捕获句子级别识别信息的能力，实现事件检测性能的提升。实验评估表明，提出的方法可以有效消解数据不平衡依赖问题对于不同事件检测模型的性能影响，并展现了自动适应不同程度数据不平衡和跨任务扩展的能力。

创新点：本文首次**主动**消解了不同事件检测模型系统面临的共性数据不平衡依赖，基于其对性能影响的方式引入了句子级别识别信息进行应对；提出了分类器自适应知识蒸馏方法，实现句子级别识别信息的自动获取，通用有效地完成了不同模型系统和数据不平衡程度的依赖消解。

(2) 针对事件要素抽取中存在的实体类型过度依赖，首先定义了该问题，并验证了多种不同的事件要素抽取架构均存在对于实体类型的过度依赖，进而影响事件要素类型的抽取表现。在此基础上，提出了基于多角度对比学习的实体类型过度依赖消解方法。该方法采用两种改进的监督对比学习技术，分别从正样本和负样本两个角度实现对实体类型过度依赖的消解，且通过循环训练策略进一步提升多角度消解的效率。实验评估表明，提出的方法可以有效消解不同事件要素抽取架构的实体类型过度依赖，并在基于预训练语言模型的架构上表现出当前最优性能。

创新点：本文首次探究了实体提及的类型信息对要素类型建模的**负面影响**，根据

实体类型和要素类型的数据特点验证了该负面影响在不同要素抽取架构中存在的普遍性；提出了两种改进的监督对比学习模块和相辅的循环训练策略，实现了从不同角度降低建模过程对于实体类型的依赖程度，通用有效地消解了实体类型过度依赖对不同要素抽取架构的不利影响。

(3) 针对不同文本级别事件要素抽取中均存在的跨事件依赖，提出了分离-融合抽取范式，将跨事件信息获取和事件要素抽取先分开建模再进行融合，从而利用到不同抽取范式的优势。在此基础上，提出了一个基于分离-融合抽取范式的多词元链接模型，通过采用两个分隔的多词元链接构建和进一步的两阶段融合，实现了建模跨事件信息和保持简单链接推理的高效兼容。实验评估表明，本文提出的模型性能在句子级别和文档级别数据集上均超过了当前最优模型，且在包含多事件信息的文本中表现出更加显著的性能提升。

创新点：本文首次提出了能够**高效统一**地利用不同事件要素抽取方式各自优势的分离-融合新范式；提出了相应的多词元链接模型，在有效避免现有通用文本级别事件要素抽取方法各自局限性的同时，实现了跨事件依赖关系的高效构建与利用。

2. 下一步研究展望

本文深入研究了事件抽取任务中不同模型架构和文本级别存在的数据不平衡依赖、实体类型过度依赖和跨事件依赖等共有数据特性，提供了有效的解决方法，为文本信息的智能化应用发展提供了有利支撑。然而，结合本文研究内容和当下人工智能的发展潮流，依然存在以下需要进一步探索的研究问题：

(1) 研究基于多词元链接的端到端事件抽取。当前主流的端到端事件抽取方法主要基于序列生成模型，在训练和推理阶段依赖枚举部分或全部事件类型提示模版进行事件信息的获取，存在计算效率不高的局限。而基于多词元链接的方法能够实现文本、事件类型和要素类型的并行化编码和链接构建，在保证计算效率的同时，提供了良好的端到端事件抽取研究潜力。然而，相比于事件要素抽取，事件抽取任务的建模提升了链接推理的复杂性，并进一步增加了跨事件信息利用的难度。本文将在下一步工作中研究如何实现端到端事件抽取任务的有效拆解和相适应的融合机制。

(2) 研究基于大语言模型增强的事件抽取。现有监督模型较为依赖对于特定词或词对信息的映射学习，缺乏事件的复杂推理能力和抽象知识建模能力。最近，大语言模型技术通过指令微调和学习反馈，具备了对未见或罕见事件模式的良好泛化推理能力。因此，本文后续将研究如何利用大语言模型构建抽象层级更高的事件概念知识，

辅助增强传统事件抽取模型性能的进一步提升。

(3) 研究基于部分标注场景的事件抽取。由于事件抽取数据的标注需要专业的语言学知识和经验,无法保证每个事件数据的完整信息均被有效识别。此外,其标注信息的细粒度特点决定了相应的训练语料构建成本相当高昂。因此,无论是从客观存在的事件信息不完整还是语料标注的成本考虑,研究基于部分标注场景的事件抽取方法具备了广阔的应用前景和价值意义。本文将在接下来的工作中研究如何基于对比学习、伪标签生成和样本置信度评估等技术,构建适应部分标注场景的事件抽取方法。

参考文献

- [1] Peng T, Li Z, Zhang L, et al. FSUIE: A Novel Fuzzy Span Mechanism for Universal Information Extraction [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 16318–16333.
- [2] Jia Z, Yan Z, Han W, et al. Modeling Instance Interactions for Joint Information Extraction with Neural High-Order Conditional Random Field [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 13695–13710.
- [3] 黄河燕, 袁长森, 冯冲. 融合实体和上下文信息的篇章关系抽取研究 [J]. 自动化学报, 2023, 49: 1–10.
- [4] 蔡宇翔, 骆姐, 甘洋镭, et al. 基于跨度边界感知的嵌套命名实体识别 [J]. 软件学报, 2024: 1–14.
- [5] Wu X, Huang K-H, Fung Y, et al. Cross-document misinformation detection based on event graph reasoning [C]. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 543–558.
- [6] Liu F, Yacoob Y, Shrivastava A. COVID-VTS: Fact Extraction and Verification on Short Video Platforms [C]. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023: 178–188.
- [7] Kuhnle A, Aroca-Ouellette M, Basu A, et al. Reinforcement learning for information retrieval [C]. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 2669–2672.
- [8] Zhang W, Zhao X, Zhao L, et al. Drl4ir: 2nd workshop on deep reinforcement learning for information retrieval [C]. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 2681–2684.
- [9] Boyd-Graber J, Börschinger B. What Question Answering can Learn from Trivia Nerds [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7422–7435.
- [10] Wen H, Lin Y, Lai T, et al. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system [C]. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, 2021: 133–143.
- [11] Liu C-Y, Zhou C, Wu J, et al. CPMF: A collective pairwise matrix factorization model for upcoming event recommendation [C]. In 2017 International Joint Conference on Neural Networks (IJCNN),

- 2017: 1532–1539.
- [12] Abdollahi S, Gottschalk S, Demidova E. LaSER: Language-specific event recommendation [J]. *Journal of Web Semantics*, 2023, 75: 100759.
 - [13] Lu J, Ng V. Span-based event coreference resolution [C]. In *Proceedings of the AAAI conference on artificial intelligence*, 2021: 13489–13497.
 - [14] Lu Y, Lin H, Tang J, et al. End-to-end neural event coreference resolution [J]. *Artificial Intelligence*, 2022, 303: 103632.
 - [15] Liu K, Chen Y, Liu J, et al. Extracting events and their relations from texts: A survey on recent research progress and challenges [J]. *AI Open*, 2020, 1: 22–39.
 - [16] Zhuang L, Fei H, Hu P. Syntax-based dynamic latent graph for event relation extraction [J]. *Information Processing & Management*, 2023, 60 (5): 103469.
 - [17] Guan S, Cheng X, Bai L, et al. What is Event Knowledge Graph: A Survey [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35 (7): 7569–7589.
 - [18] Cheng D, Yang F, Wang X, et al. Knowledge graph-based event embedding framework for financial quantitative investments [C]. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 2221–2230.
 - [19] Souza Costa T, Gottschalk S, Demidova E. Event-QA: A dataset for event-centric question answering over knowledge graphs [C]. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020: 3157–3164.
 - [20] Wu T, He S, Liu J, et al. A brief overview of ChatGPT: The history, status quo and potential future development [J]. *IEEE/CAA Journal of Automatica Sinica*, 2023, 10 (5): 1122–1136.
 - [21] Pan J Z, Razniewski S, Kalo J-C, et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges [J]. *TRANSACTIONS ON GRAPH DATA AND KNOWLEDGE*, 2023: 1–38.
 - [22] Grishman R. Information extraction: Techniques and challenges [C]. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14–18, 1997*, 1997: 10–27.
 - [23] Ahn D. The stages of event extraction [C]. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 2006: 1–8.
 - [24] Zhou D, Zhang X, He Y. Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings [C]. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017: 808–817.

- [25] Li M, Zareian A, Lin Y, et al. Gaia: A fine-grained multimedia knowledge extraction system [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020: 77–86.
- [26] Li Q, Li P, Bi W, et al. Event Transition Planning for Open-ended Text Generation [C]. In Findings of the Association for Computational Linguistics: ACL 2022, 2022: 3412–3426.
- [27] Xu F, Uszkoreit H, Li H. Automatic event and relation detection with seeds of varying complexity [C]. In Proceedings of the AAAI workshop event extraction and synthesis, 2006: 12–17.
- [28] Liao S, Grishman R. Filtered ranking for bootstrapping in event extraction [C]. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010: 680–688.
- [29] Kilicoglu H, Bergler S. Effective bio-event extraction using trigger words and syntactic dependencies [J]. Computational Intelligence, 2011, 27 (4): 583–609.
- [30] Cao K, Li X, Fan M, et al. Improving event detection with active learning [C]. In Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015: 72–77.
- [31] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text [C]. In Eighteenth national conference on Artificial intelligence, 2002: 786–791.
- [32] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction [C]. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 2009: 209–212.
- [33] Ji H, Grishman R. Refining event extraction through cross-document inference [C]. In Proceedings of ACL-08: Hlt, 2008: 254–262.
- [34] Gupta P, Ji H. Predicting unknown time arguments based on cross-event propagation [C]. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009: 369–372.
- [35] Liao S, Grishman R. Acquiring topic features to improve event extraction: in pre-selected and balanced collections [C]. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, 2011: 9–16.
- [36] Hong Y, Zhang J, Ma B, et al. Using cross-entity inference to improve event extraction [C]. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011: 1127–1136.
- [37] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features [C]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013: 73–82.
- [38] Li X, Nguyen T H, Cao K, et al. Improving event detection with abstract meaning representation [C].

- In Proceedings of the first workshop on computing news storylines, 2015: 11–15.
- [39] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1510.03820, 2015.
 - [40] Nguyen T H, Grishman R. Event detection and domain adaptation with convolutional neural networks [C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015: 365–371.
 - [41] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 167–176.
 - [42] Nguyen T H, Grishman R. Modeling skip-grams for event detection with convolutional neural networks [C]. In Proceedings of the 2016 conference on empirical methods in natural language processing, 2016: 886–891.
 - [43] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. nature, 1986, 323 (6088): 533–536.
 - [44] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735–1780.
 - [45] Cho K, Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [C]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1724–1734.
 - [46] Ghaeini R, Fern X, Huang L, et al. Event Nugget Detection with Forward-Backward Recurrent Neural Networks [C]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016: 369–373.
 - [47] Jagannatha A N, Yu H. Bidirectional RNN for medical event detection in electronic health records [C]. In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, 2016: 473.
 - [48] Orr J W, Tadepalli P, Fern X. Event Detection with Neural Networks: A Rigorous Empirical Evaluation [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 999–1004.
 - [49] Zhang Y, Xu G, Wang Y, et al. Empower event detection with bi-directional neural language model [J]. Knowledge-Based Systems, 2019, 167: 87–97.
 - [50] Chen Y, Yang H, Liu K, et al. Collective event detection via a hierarchical and bias tagging net-

- works with gated multi-level attention mechanisms [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 1267–1276.
- [51] Lou D, Liao Z, Deng S, et al. MLBiNet: A Cross-Sentence Collective Event Detection Network [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 4829–4839.
- [52] Liu S, Li Y, Zhang F, et al. Event detection without triggers [C]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 735–744.
- [53] Feng X, Qin B, Liu T. A language-independent neural network for event detection [J]. Science China Information Sciences, 2018, 61: 1–12.
- [54] Lu Y, Lin H, Han X, et al. Distilling discrimination and generalization knowledge for event detection via delta-representation learning [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4366–4376.
- [55] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview [J]. IEEE signal processing magazine, 2018, 35 (1): 53–65.
- [56] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks [C]. In International Conference on Learning Representations, 2016.
- [57] Liu X, Luo Z, Huang H-Y. Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 1247–1256.
- [58] Nguyen T, Grishman R. Graph convolutional networks with argument-aware pooling for event detection [C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 5900–5907.
- [59] Lai V D, Nguyen T N, Nguyen T H. Event Detection: Gate Diversity and Syntactic Importance Scores for Graph Convolution Neural Networks [C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 5405–5411.
- [60] Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications [J]. AI open, 2020, 1: 57–81.
- [61] Yan H, Jin X, Meng X, et al. Event Detection with Multi-Order Graph Convolution and Aggregated Attention [C]. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 5770–5774.
- [62] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks [J]. arXiv preprint

- arXiv:1710.10903, 2017.
- [63] Cui S, Yu B, Liu T, et al. Edge-Enhanced Graph Convolution Networks for Event Detection with Syntactic Relation [C]. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 2329–2339.
 - [64] Liu A, Xu N, Liu H. Self-attention graph residual convolutional networks for event detection with dependency relations [C]. In Findings of the Association for Computational Linguistics: EMNLP 2021, 2021: 302–311.
 - [65] Mi J, Hu P, Li P. Event Detection with Dual Relational Graph Attention Networks [C]. In Proceedings of the 29th International Conference on Computational Linguistics, 2022: 1979–1989.
 - [66] Kenton J D M-W C, Toutanova L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]. In Proceedings of NAACL-HLT, 2019: 4171–4186.
 - [67] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30: 6000–6010.
 - [68] Wu Y, Schuster M, Chen Z, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation [J]. arXiv preprint arXiv:1609.08144, 2016.
 - [69] Yang S, Feng D, Qiao L, et al. Exploring pre-trained language models for event extraction and generation [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5284–5294.
 - [70] Wang X, Han X, Liu Z, et al. Adversarial training for weakly supervised event detection [C]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 998–1008.
 - [71] Liu J, Chen Y, Liu K, et al. Event extraction as machine reading comprehension [C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 1641–1651.
 - [72] Du X, Cardie C. Event Extraction by Answering (Almost) Natural Questions [C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 671–683.
 - [73] Li F, Peng W, Chen Y, et al. Event extraction as multi-turn question answering [C]. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 829–838.
 - [74] Wang S, Yu M, Chang S, et al. Query and Extract: Refining Event Extraction as Type-oriented Binary Decoding [C]. In Findings of the Association for Computational Linguistics: ACL 2022, 2022: 169–182.

- [75] Wang S, Yu M, Huang L. The art of prompting: Event detection based on type specific prompts [J]. arXiv preprint arXiv:2204.07241, 2022.
- [76] Liu J, Chen Y, Liu K, et al. How does context matter? On the robustness of event detection with context-selective mask generalization [C]. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 2523–2532.
- [77] Harris Z S, Harris Z S. Distributional Structure [J]. Papers in Structural and Transformational Linguistics, 1970: 775–794.
- [78] TONG M, XU B, WANG S, et al. Improving event detection via open-domain event trigger knowledge.(2020) [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual Conference, 2020: 5–10.
- [79] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to WordNet: An on-line lexical database [J]. International journal of lexicography, 1990, 3 (4): 235–244.
- [80] Sandhaus E. The new york times annotated corpus [J]. Linguistic Data Consortium, Philadelphia, 2008, 6 (12): e26752.
- [81] Liu J, Chen Y, Xu J. Saliency as evidence: Event detection with trigger saliency attribution [C]. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022: 4573–4585.
- [82] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks [C]. In International conference on machine learning, 2017: 3319–3328.
- [83] Veyseh A P B, Lai V, Démoncourt F, et al. Unleash GPT-2 power for event detection [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 6271–6282.
- [84] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1 (8): 9.
- [85] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways [J]. Journal of Machine Learning Research, 2023, 24 (240): 1–113.
- [86] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models [J]. arXiv preprint arXiv:2302.13971, 2023.
- [87] Li B, Fang G, Yang Y, et al. Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness [J]. arXiv preprint arXiv:2304.11633, 2023.
- [88] Gao J, Zhao H, Yu C, et al. Exploring the feasibility of chatgpt for event extraction [J]. arXiv preprint

- arXiv:2303.03836, 2023.
- [89] Han R, Peng T, Yang C, et al. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors [J]. arXiv preprint arXiv:2305.14450, 2023.
- [90] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877–1901.
- [91] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in Neural Information Processing Systems, 2022, 35: 24824–24837.
- [92] Fung Y, Wang H, Wang T, et al. DeepMaven: Deep question answering on long-distance movie/TV show videos with multimedia knowledge extraction and synthesis [C]. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023: 3033–3043.
- [93] Ebner S, Xia P, Culkin R, et al. Multi-Sentence Argument Linking [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 8057–8077.
- [94] Li S, Ji H, Han J. Document-Level Event Argument Extraction by Conditional Generation [C]. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 894–908.
- [95] Ma Y, Wang Z, Cao Y, et al. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction [C]. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022: 6759–6774.
- [96] Cao K, Li X, Ma W, et al. Including new patterns to improve event extraction systems [C]. In The Thirty-First International Flairs Conference, 2018: 487–492.
- [97] Liao S, Grishman R. Using document level cross-event inference to improve event extraction [C]. In Proceedings of the 48th annual meeting of the association for computational linguistics, 2010: 789–797.
- [98] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine Learning research, 2003, 3 (Jan): 993–1022.
- [99] Wang X, Wang Z, Han X, et al. HMEAE: Hierarchical modular event argument extraction [C]. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 5777–5783.
- [100] Nguyen T H, Cho K, Grishman R. Joint event extraction via recurrent neural networks [C]. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 300–309.

- [101] Sha L, Qian F, Chang B, et al. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction [C]. In Proceedings of the AAAI conference on artificial intelligence, 2018: 5916–5923.
- [102] Nguyen T M, Nguyen T H. One for all: Neural joint modeling of entities and events [C]. In Proceedings of the AAAI conference on artificial intelligence, 2019: 6851–6858.
- [103] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 4171–4186.
- [104] Xiangyu X, Ye W, Zhang S, et al. Capturing Event Argument Interaction via A Bi-Directional Entity-Level Recurrent Decoder [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 210–219.
- [105] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv preprint arXiv:1907.11692, 2019.
- [106] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871–7880.
- [107] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans [J]. Transactions of the association for computational linguistics, 2020, 8: 64–77.
- [108] Ma J, Wang S, Anubhai R, et al. Resource-Enhanced Neural Model for Event Argument Extraction [C]. In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 3554–3559.
- [109] Zhang Z, Kong X, Liu Z, et al. A two-step approach for implicit event argument detection [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7479–7485.
- [110] Dozat T, Manning C D. Simpler but More Accurate Semantic Dependency Parsing [C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018: 484–490.
- [111] Xu R, Wang P, Liu T, et al. A Two-Stream AMR-enhanced Model for Document-level Event Argument Extraction [C]. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5025–5036.
- [112] Banarescu L, Bonial C, Cai S, et al. Abstract meaning representation for sembanking [C]. In Pro-

- pceedings of the 7th linguistic annotation workshop and interoperability with discourse, 2013: 178–186.
- [113] Liu W, Cheng S, Zeng D, et al. Enhancing Document-level Event Argument Extraction with Contextual Clues and Role Relevance [C]. In Findings of the Association for Computational Linguistics: ACL 2023, 2023: 12908–12922.
 - [114] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text [C]. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2383–2392.
 - [115] Zhou Y, Chen Y, Zhao J, et al. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering [C]. In Proceedings of the AAAI conference on artificial intelligence, 2021: 14638–14646.
 - [116] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension [C]. In International Conference on Learning Representations, 2016.
 - [117] Liu J, Chen Y, Xu J. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction [C]. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 2716–2725.
 - [118] Atkins S, Rundell M, Sato H. The contribution of FrameNet to practical lexicography [J]. International Journal of Lexicography, 2003, 16 (3): 333–357.
 - [119] Wei K, Sun X, Zhang Z, et al. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 4672–4682.
 - [120] Lu D, Ran S, Tetreault J, et al. Event Extraction as Question Generation and Answering [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023: 1666–1688.
 - [121] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of machine learning research, 2020, 21 (140): 1–67.
 - [122] Zeng Q, Zhan Q, Ji H. EA2E: Improving Consistency with Event Awareness for Document-Level Argument Extraction [C]. In Findings of the Association for Computational Linguistics: NAACL 2022, 2022: 2649–2655.
 - [123] Hsu I-H, Huang K-H, Boschee E, et al. DEGREE: A Data-Efficient Generation-Based Event Extraction Model [C]. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 1890–1908.

- [124] Hsu I-H, Xie Z, Huang K-H, et al. AMPERE: AMR-Aware Prefix for Generation-Based Event Argument Extraction Model [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 10976–10993.
- [125] Du X, Li S, Ji H. Dynamic Global Memory for Document-level Argument Extraction [C]. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022: 5264–5275.
- [126] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [C]. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 3982–3992.
- [127] Ren Y, Cao Y, Guo P, et al. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 293–306.
- [128] Zhang Z, Strubell E, Hovy E. Transfer learning from semantic role labeling to event argument extraction with template-based slot querying [C]. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 2627–2647.
- [129] Nguyen C, Mãn H, Nguyen T. Contextualized soft prompts for extraction of event arguments [C]. In Findings of the Association for Computational Linguistics: ACL 2023, 2023: 4352–4361.
- [130] Li H, Cao Y, Ren Y, et al. Intra-Event and Inter-Event Dependency-Aware Graph Network for Event Argument Extraction [C]. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 6362–6372.
- [131] He Y, Hu J, Tang B. Revisiting Event Argument Extraction: Can EAE Models Learn Better When Being Aware of Event Co-occurrences? [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 12542–12556.
- [132] Veyseh A P B, Van Nguyen M, DERNONCOURT F, et al. Document-level event argument extraction via optimal transport [C]. In Findings of the Association for Computational Linguistics: ACL 2022, 2022: 1648–1658.
- [133] Yang Y, Guo Q, Hu X, et al. An AMR-based Link Prediction Approach for Document-level Event Argument Extraction [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 12876–12889.
- [134] Liu J, Liang C, Xu J, et al. Document-level event argument extraction with a chain reasoning paradigm [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 9570–9583.

- [135] Ye W, Li B, Xie R, et al. Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1351–1360.
- [136] dos Santos C, Xiang B, Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks [C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 626–634.
- [137] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. arXiv preprint arXiv:1503.02531, 2015.
- [138] Doddington G R, Mitchell A, Przybocki M, et al. The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation [C]. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’ 04), 2004: 837–840.
- [139] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch [C]. In NIPS 2017 Autodiff Workshop, 2017.
- [140] Zhang Y, Zhong V, Chen D, et al. Position-aware Attention and Supervised Data Improve Slot Filling [C]. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 35–45.
- [141] Zhang Y, Qi P, Manning C D. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2205–2215.
- [142] Guo Z, Zhang Y, Lu W. Attention Guided Graph Convolutional Networks for Relation Extraction [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 241–251.
- [143] Xue F, Sun A, Zhang H, et al. Gdpnet: Refining latent multi-view graph for relation extraction [C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14194–14202.
- [144] Peng H, Gao T, Han X, et al. Learning from Context or Names? An Empirical Study on Neural Relation Extraction [C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 3661–3672.
- [145] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C]. In International conference on machine learning, 2020: 1597–1607.
- [146] He K, Fan H, Wu Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning [C]. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 9726–9735.

- [147] Fang H, Wang S, Zhou M, et al. Cert: Contrastive self-supervised learning for language understanding [J]. arXiv preprint arXiv:2005.12766, 2020.
- [148] Wang D, Ding N, Li P, et al. CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 2332–2342.
- [149] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning [J]. Advances in neural information processing systems, 2020, 33: 18661–18673.
- [150] Liu H, Zhang F, Zhang X, et al. An Explicit-Joint and Supervised-Contrastive Learning Framework for Few-Shot Intent Classification and Slot Filling [C]. In Findings of the Association for Computational Linguistics: EMNLP 2021, 2021: 1945–1955.
- [151] Yang C, Zou J, Wu J, et al. Supervised contrastive learning for recommendation [J]. Knowledge-Based Systems, 2022, 258: 109973.
- [152] Ohashi S, Takayama J, Kajiwara T, et al. Text classification with negative supervision [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 351–357.
- [153] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information processing & management, 1988, 24 (5): 513–523.
- [154] Zhang K, Shuang K, Yang X, et al. What Is Overlap Knowledge in Event Argument Extraction? APE: A Cross-datasets Transfer Learning Model for EAE [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 393–409.
- [155] Lou J, Lu Y, Dai D, et al. Universal Information Extraction as Unified Semantic Matching [C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 13318–13326.
- [156] Liu C, Zhao F, Kang Y, et al. RexUIE: A Recursive Method with Explicit Schema Instructor for Universal Information Extraction [C]. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 15342–15359.
- [157] Tang W, Xu B, Zhao Y, et al. UniRel: Unified Representation and Interaction for Joint Relational Triple Extraction [C]. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 7087–7099.
- [158] Yu B, Zhang Z, Sheng J, et al. Semi-open information extraction [C]. In Proceedings of the Web Conference 2021, 2021: 1661–1672.
- [159] Xu J, Xu W, Sun M, et al. Extracting Trigger-sharing Events via an Event Matrix [C]. In Findings

- of the Association for Computational Linguistics: EMNLP 2022, 2022: 1189–1201.
- [160] Wolpert D H. Stacked generalization [J]. *Neural networks*, 1992, 5 (2): 241–259.
- [161] Wadden D, Wennberg U, Luan Y, et al. Entity, Relation, and Event Extraction with Contextualized Span Representations [C]. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019: 5784–5789.
- [162] Pyysalo S, Ohta T, Miwa M, et al. Event extraction across multiple levels of biological organization [J]. *Bioinformatics*, 2012, 28 (18): i575–i581.
- [163] Trieu H-L, Tran T T, Duong K N, et al. DeepEventMine: end-to-end neural nested event extraction from biomedical texts [J]. *Bioinformatics*, 2020, 36 (19): 4910–4917.
- [164] Luo L, Xu Y. Context-Aware Prompt for Generation-based Event Argument Extraction with Diffusion Models [C]. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023: 1717–1725.
- [165] Zhou J, Zhang Q, Chen Q, et al. A Multi-Format Transfer Learning Model for Event Argument Extraction via Variational Information Bottleneck [C]. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2022: 1990–2000.

攻读学位期间发表论文与研究成果清单

- [1] 第一作者. Multi-view entity type overdependency reduction for event argument extraction[J]. Knowledge-Based Systems 2023. (SCI 一区)
- [2] 第二作者 (导师一作) . Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data[J]. Information Sciences 2021. (SCI 一区)
- [3] 第一作者. Dialogue State Distillation Network with Inter-slot Contrastive Learning for Dialogue State Tracking[C]. AAAI 2023. (CCF-A 类国际会议)
- [4] 第一作者. Separation and Fusion: A Novel Multiple Token Linking Model for Event Argument Extraction[C]. NAACL 2024. (CCF-B 类国际会议, 已接收)
- [5] 第三作者. A Multi-turn Machine Reading Comprehension Framework with Rethink Mechanism for Emotion-Cause Pair Extraction[C]. COLING 2022. (CCF-B 类国际会议)
- [6] 第六作者. Effective Named Entity Recognition with Boundary-aware Bidirectional Neural Networks[C]. WWW 2021. (CCF-A 类国际会议)
- [7] 第六作者. Modularized Interaction Network for Named Entity Recognition[C]. ACL 2021. (CCF-A 类国际会议)
- [8] 第六作者. A collaborative learning framework for knowledge graph embedding and reasoning[J]. Knowledge-Based Systems 2024. (SCI 一区)