基于深度学习的篇章级事件抽取综述

胡 蓉[1),2)] 万常选[1),2)] 万齐智[1),2)] 刘德喜[1),2)] 刘喜平[1),2)]

[1)](江西财经大学计算机与人工智能学院 南昌 330032)

[2)](江西财经大学数据与知识工程江西省高校重点实验室 南昌 330013)

摘 要 篇章级事件抽取是自然语言处理的重要任务且富有挑战,当前涌现了很多优秀的研究成果。尽管国内外存在少量篇章级事件抽取综述,但存在一些局限:(1)按文献采用的具体技术或任务实现步骤对现有研究成果进行分类,未深入分析现有研究成果间的关联与区别,未深刻理解现有研究成果分别致力于解决哪些问题;(2)简单介绍现有数据集,未能正确认识每个数据集的特点及带来的任务挑战。由于每个数据集侧重点不同,研究者们致力于解决不同的问题,因此现有梳理方式未能清晰地展示不同数据集下不同研究问题的研究进展。为此,本文重新梳理篇章级事件抽取的2个(子)任务的研究成果。首先,针对2个任务,分别明确任务目标,分析解决任务的基本思路,总结现有研究进展(基于哪些数据集解决了哪些问题)。然后,总结对应数据集的特点,归纳任务面临的挑战,再深入分析具体研究方法,并图示化展示推进情况。最后,结合有待继续攻破的问题,讨论篇章级事件抽取未来发展趋势。

关键词 篇章级事件抽取;信息抽取;事件抽取数据集;事件论元抽取;深度学习

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.00381

Document-Level Event Extraction Based on Deep Learning: A Survey

HU Rong[1),2)] WAN Chang-Xuan[1),2)] WAN Qi-Zhi[1),2)] LIU De-Xi[1),2)] LIU Xi-Ping[1),2)]

[1)](School of Computer and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330032)

[2)](Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang 330013)

Abstract Document-level Event extraction (DEE) is an important and challenging task in natural language processing, and numerous outstanding research achievements have emerged. DEE mainly focuses on two tasks, namely Document-level Event Identification and Argument Extraction (DocEI & AE), Document-level Event Argument Extraction (DocEAE). DocEI & AE indicates the complete document-level event extraction, that is, judging what types of events exist in a given document, identifying all events under each event type, and extracting arguments of corresponding roles. DocEAE refers to the event argument extraction, that is, given the event types and event triggers contained in each document, extracting event arguments of the corresponding roles triggered by each trigger. The goals of two tasks are different, and the task steps are also not exactly the same. Furthermore, the corresponding datasets also have different characteristics and focus on causing different research problems. Although there are a few surveys on document-level event extraction, they share the following two limitations. (1) The

收稿日期:2024-02-16;在线发布日期:2024-09-23。本课题得到国家自然科学基金项目(62272205, 62272206, 62076112, 62462034)、江西省教育厅科学技术研究项目(GJJ210531, GJJ2400411)、江西省研究生创新专项资金项目(YC2023-B188)、江西省自然科学基金(20242BAB25119, 20212ACB202002, 20232ACB202008)、江西省主要学科学术和技术带头人培养计划领军人才项目(20213BCJL22041)资助。**胡 蓉**,博士研究生,助理研究员,主要研究领域为信息抽取、自然语言处理、大数据分析。E-mail: hurong2014@126.com。**万常选**(通信作者),博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为Web数据管理、情感分析、数据挖掘。E-mail: wanchangxuan@263.net。**万齐智**,博士,讲师,中国计算机学会(CCF)会员,主要研究领域为信息抽取、自然语言处理、数据挖掘。**刘德喜**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、信息检索、Web数据管理。**刘喜平**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为信息检索、数据挖掘。

classification of existing researches is usually conducted based on specific techniques or task steps adopted in the literatures, without an in-depth analysis of the correlations and differences between the methods or a profound understanding of the issues each of them aims to address. (2) The description of datasets for DEE is simple and fails to understand the characteristics and task challenges of each dataset. Due to the different concerns of each dataset, the issues that researchers strive to solving vary. Therefore, current reviewing methods fail in clearly demonstrating the research progress on different issues under diverse datasets. This paper reorganizes the results of DocEI & AE and DocEAE tasks in document-level event extraction. Firstly, the task objectives and the common approach to solving the tasks are clarified and analyzed, then the current research progress (solving which issues based on the datasets) is summarized. Specifically, there are two ways to implement DocEI & AE task. One is pipeline pattern, that is, sub-tasks such as entity extraction, event type judgment and multi-event extraction are independently executed step by step. The difficulty lies in multi-event recognition, i. e., how to identify all events contained in a document. The other is the joint pattern, which integrates the above sub-tasks into one task to complete the entire extraction task, avoiding error propagation. The difficulty is designing a data structure/graph structure that can not only represent and decode all the events contained in the document, but also directly reveal the arguments playing the roles in the events under the specific event types, so as to integrate the three subtasks into one task to complete the entire event extraction task. In DocEI & AE task, researches mainly focus on ChFinAnn and DuEE-Fin corpora, which are mainly divided into path expansion, bipartite matching learning, sentence community, joint extraction strategies. The ideas of the DocEAE task can be divided into the following strategies: traditional classification, text generation, machine reading comprehension, and span selection. Secondly, the characteristics of the two types of datasets are explored, followed by the challenges of two tasks. Meanwhile, the specific research methods of the two tasks are analyzed in depth with visualizations to showcase the progress. Finally, combining with the issues that need to be further tackled, the future development trends of document-level event extraction are discussed.

Keywords document-level event extraction; information extraction; event extraction datasets; event argument extraction; deep learning

1 引 言

篇章级事件抽取(Document-level Event Extraction, DEE)是在给定的事件模板下,从给定的篇章中检测出符合模板要求的事件,并抽取相关角色的论元。根据是否需要识别事件,目前DEE研究主要围绕2个(子)任务,分别是篇章级事件识别及其论元抽取(Document-level Event Identification and Argument Extraction, DocEI & AE)和篇章级事件论元抽取(Document-level Event Argument Extraction, DocEAE)。具体地,DocEI & AE[1-4]是从给定的篇章中判断存在哪些类型的事件、识别每一种事件类型下的所有事件(注:语料中未标注事件触发词)并抽取相应角色的论元;DocEAE[5-8]是在给定事件触发词及对应的事件类型下,从篇章中抽取其触发事件的相应角色的论元。

DocEI & AE和DocEAE的目标与任务各不相同,研究框架及策略也不尽相同。研究者们基于2个任务的公开数据集,致力于解决各任务特有的研究问题,有了不同的研究进展。如DocEI & AE任务中,有人致力于解决篇章中多事件识别问题[1, 9-10],也有人设计全新的数据结构致力于集成实现事件识别及其论元抽取的各个任务步骤[11-12];DocEAE任务中,重点致力于解决篇章中复杂论元(即带有限定性描述内容的词语/短语/实体构成的文本片段所描述的复杂对象充当事件论元)抽取问题,如有人解决候选论元空间过大问题[5, 13],有人转

换为其他任务避免产生候选论元[6, 14]，也有人基于预训练语言模型采用跨度选择策略，即训练模型直接预测每个角色的论元跨度。为了解决2个任务的不同挑战，研究者们采用了不同的策略。本文将分别对这2个任务的现有研究进行梳理和总结。

目前国内外存在少量篇章级事件抽取综述，如按使用的神经网络技术或任务实现步骤整理现有研究[15-16]，或根据任务的实现方式将现有研究分为流水线模式、联合模式和端到端模式[17]。然而，现有综述存在一些局限：(1) 按文献采用的深度学习技术或任务实现步骤等对现有DEE的研究进行分类，未深入分析2个任务的现有研究分别致力于解决哪些问题及各研究间的关联与区别，使得读者无法全面把握篇章级事件抽取研究及其推进情况；(2) 简单介绍现有2个任务的数据集，未能正确认识各任务数据集的特点及其引发的研究问题，各任务遇到的挑战。

为了更好地梳理篇章级事件抽取的研究现状，使读者清晰地把握篇章级事件抽取存在的研究问题及研究推进情况。本文的策略是：首先，针对2个任务分别明确任务目标，阐述解决问题的基本思路，归纳现有研究进展。然后，分析常用数据集的特点，分析在对应数据集上各任务面临的挑战。最后，深入分析针对相应的挑战而提出的具体方法，并以图示化方式展示各方法的区别及推进情况。

本文的主要贡献如下：

(1) 理清了目前篇章级事件抽取研究围绕的2个任务的研究目标和解决问题的策略，并统筹归纳了2个任务的研究进展。其中，融入了较多个人理解的归纳图，可为相关研究人员了解篇章级事件抽取提供帮助。

(2) 深入分析了2个任务常用数据集的特点，并归纳总结了各类数据集带来的任务挑战，这是目前没有人考虑和完成的。

(3) 针对每个任务，分类梳理了解决以上挑战的具体研究方法，基于对原文充分理解的基础上，总结并绘制了各类策略框架和进展图，图示化地展示了推进过程。

2　任务研究目标和进展

2.1　任务研究目标

事件抽取[18-21]旨在从文本中抽取指定的各种类型的事件信息。早期研究聚焦于语句级事件抽取[22-30]，由于无法处理事件论元跨句情况(事件的论元散落在多个语句中)，篇章级事件抽取被关注。篇章级事件抽取主要围绕2个任务：DocEI & AE和DocEAE。

(1) DocEI & AE目标及子任务

针对DocEI & AE任务，目标是从篇章中抽取指定的各种类型的所有事件(注：语料中未标注事件触发词)，事件类型和角色限定了抽取的事件范围，论元用于填充相应角色的内容。图1展示了DocEI & AE示例，其中，上部分为篇章，下部分的事件表格为需要从篇章中抽取的事件记录(黄金标注)。在这个抽取的过程中，通常包括3个子任务：①实体抽取，即识别候选论元，如图1上部分篇章中颜色突出部分；②事件类型判断，即识别篇章中包含哪些事件类型，该篇章包含EU和EO 2种事件类型，如图1下部分事件表上方的内容；③多事件抽取，即给定事件类型下多事件识别及其论元抽取，如图1下部分每张事件表中的内容为该事件各角色对应的论元(EU类型包含1个事件$event_1$、EO类型包含2个事件$event_2$和$event_3$)。

DocEI & AE任务描述如下。假定篇章D，任务目标是：首先识别D中包含的所有事件类型$\{t\}$；然后识别每一种事件类型t下包含的所有事件$\{event\}$，并完成每个事件$event$在每个角色下的论元抽取(即找出该事件的所有<角色，论元>对)。任务可形式化描述为，$event=\{(Role_{t,k}, \{arg_{t,k,j}\})\}$，其中，$Role_{t,k}$是事件类型$t$的第$k$个角色，$arg_{t,k,j}$是篇章$D$中的一个实体，它在一个事件中充当角色$Role_{t,k}$的论元，$j$表示一个事件在角色$Role_{t,k}$上允许有多个论元。

(2) DocEAE目标及子任务

针对DocEAE任务，已知每个篇章中包含的事件类型和事件触发词(语料中已标注)，目标是将篇章中每一个触发词所触发事件的角色与论元链接起来(无需识别事件)，即找出指定事件的所有<角色，论元>对。图2展示了DocEAE的示例，其中，黑框中的文本为篇章，<t></t>标记目标触发词，事件类型由触发词确定，从而限定了角色范围；带下划线的文本片段为需要抽取的目标事件各角色的论元(黄金标注)，弧线上标识了该论元充当的角色。该过程一般包括2个子任务：①识别候选论元，即从篇章中识别所有可能充当论元的文本片段(如图2中下划线所示)，可能为复杂对象，如图2中蓝色箭头指向部分；②论元抽取，针对该事件类型下的每个角色，对每个候选论元进行分类，明确是否充当该角色的论元，如图2中带箭头的线段所示。

图1 ChFinAnn语料上DocEI & AE示例

注：图上部分为篇章，[S_i]表示第i条语句，相同颜色表示同一事件的论元(如绿色、蓝色、橙色)，红色加粗表示被相同/不同事件中多个角色共享的论元。图下部分为DocEI & AE目标，即从给定输入篇章中判断存在哪些类型的事件、识别每一种事件类型下的所有事件并抽取相应角色的论元，每张表表示一个事件，表的上方为事件类型，表的第1列为该事件类型对应的角色，第2列表示相应角色的论元(从篇章中提取)。

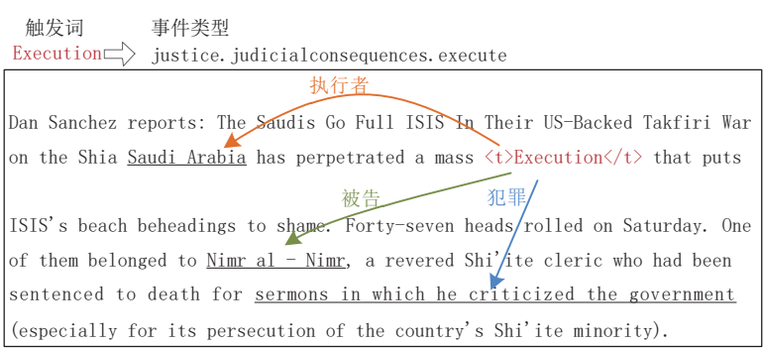


图2 RAMS语料上DocEAE示例

注：图黑框中的文本为篇章，<t></t>标记目标触发词，各带箭头的线段指向的带下画线的文本片段为DocEAE的抽取目标，即已知篇章中目标事件触发词和事件类型(无需识别事件)，抽取目标触发词所触发事件的相应角色的论元。

DocEAE任务描述如下。假定篇章D中包含多个事件，每个事件由一个触发词t标记(同时t也表示所触发事件的类型)。针对每个目标触发词t触发的事件$event_t$，任务目标是将事件角色$Role_{t,k}$与论元链接起来，即找到每个事件$event_t$的所有角色-论元对$\{(Role_{t,k}, \{arg_{t,k,j}\})\}$，其中，$Role_{t,k}$是事件类型$t$的第$k$个角色，$arg_{t,k,j}$是篇章$D$中的一个文本片段，它在事件$event_t$中充当角色$Role_{t,k}$的论元，$j$表示该事件在角色$Role_{t,k}$上允许有多个论元。

2.2 任务研究进展

2.2.1 DocEI & AE任务研究进展

DocEI & AE任务的实现模式有2种，一种是流水线模式，即分步骤独立执行实体抽取、事件类型判断和多事件抽取等子任务，难点在于多事件识别[1-3, 9-10, 31]，即如何识别篇章中包含的所有事件(注：语料中未标注事件触发词)；另一种是联合模式，即将上述子任务整合为一个任务完成整个抽取任务，避免了错误传播。联合模式的难点在于设计一个数

据结构/图结构[11-12]，既能表示和解码篇章中包含的所有事件，又能直接揭示实体在何种事件类型下的哪些事件中扮演了何种角色的论元，才能将3个子任务集成为1个任务完成整个抽取任务。

下面基于上述2种实现模式，先分别介绍它们的研究框架及策略，再归纳研究进展。

（1）研究框架及策略

图3为DocEI &. AE任务的流水线模式框架。首先抽取实体作为候选论元（图3第1部分），然后通过Transformer等神经网络捕获实体和语句表示（图3第2部分），接下来识别篇章中包含的事件类型（图3第3部分），最后完成给定事件类型下的多事件抽取（图3第4部分，包括多事件识别以及对每个候选论元进行分类，明确其是否在某个事件中充当论元、充当何种角色的论元）。不同方法中3个子任务的顺序可能有小的变动。

图4为DocEI &. AE任务的联合模式抽取策略示例[12]。首先，构建词语-词语双向事件完全图（图4左部分），节点为词语w_i，边的类型为事件类型-论元角色-论元角色。在一个事件的任意2个角色的论元（由实体充当）之间建立双向连边，并转化为在分属于2个实体中的任意2个词语之间建立双向连边。然后，模型以该图的邻接矩阵为目标进行训练。最后，针对预测的邻接矩阵进行解码（包括多事件解码和边类型解码），可得到篇章中包含的所有事件及其论元信息（图4右部分）。

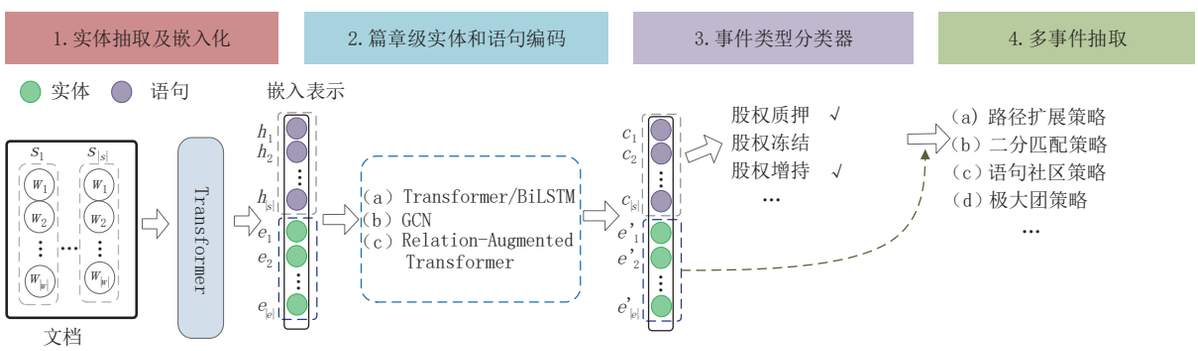


图3　DocEI &. AE任务的流水线模式框架

注：w_i和s_j分别为词语初始表示和语句初始表示，h_j和e_k分别表示Transformer编码后的语句表示和实体表示，c_j和e'_k分别表示增强上下文语义后的语句表示和实体表示。

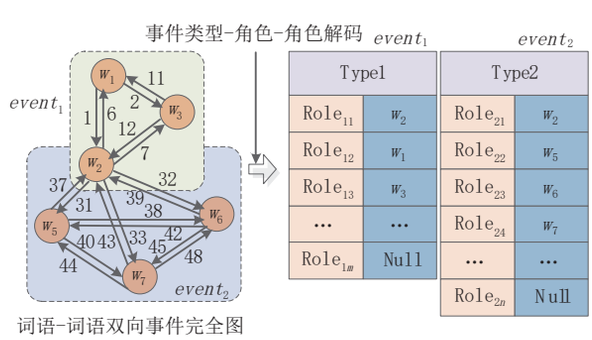


图4　DocEI &. AE任务的联合模式抽取策略示例

注：难点在于设计一个图结构，既能表示和解码DocEI & AE任务的所有事件信息[12]，又能直接揭示实体（词语）在何种事件类型下的哪些事件中扮演了何种角色的论元，从而可以将该任务转化为词语-词语邻接矩阵（由该图结构转换得到）对应的预测和解码任务，实现DocEI & AE任务的联合抽取。

（2）研究进展

现有DocEI &. AE任务的研究主要集中于ChFinAnn、DuEE-Fin语料，内容涉及中文金融公告。图5首先从策略、模型、解决的问题等方面对ChFinAnn和DuEE语料上的DocEI &. AE研究进行了总结，然后按不同策略对各模型进行了分类。主要分为如下5类。

① 路径扩展策略。在实体抽取和事件类型判断后，针对每种事件类型，文献[1]事先设定每一种事件类型下的论元角色顺序，采用路径扩展策略，使得每条路径唯一标识该事件类型下的一个事件，路径上的实体节点即为该事件的论元。文献[9-10]采用与文献[1]相同的路径扩展策略，就实体语义编码和路径扩展机制进行了探讨。

② 二分匹配学习策略。为了解决路径扩展策略依赖事先定义的论元角色顺序问题，文献[2]提出并行网络，每个事件用一张实体-论元角色二维表表示；文献[4]提出代理节点集群网络，每个代理节点表示一个事件，代理节点间建立连接来捕获事件全局信息；文献[5]迭代生成事件，并利用已生成的事件信息。以上研究都是先通过设定超参数（文档包含的事件数）来预测事件集合，再求解预测事件集合和黄金标注事件集合的最佳二分匹配。

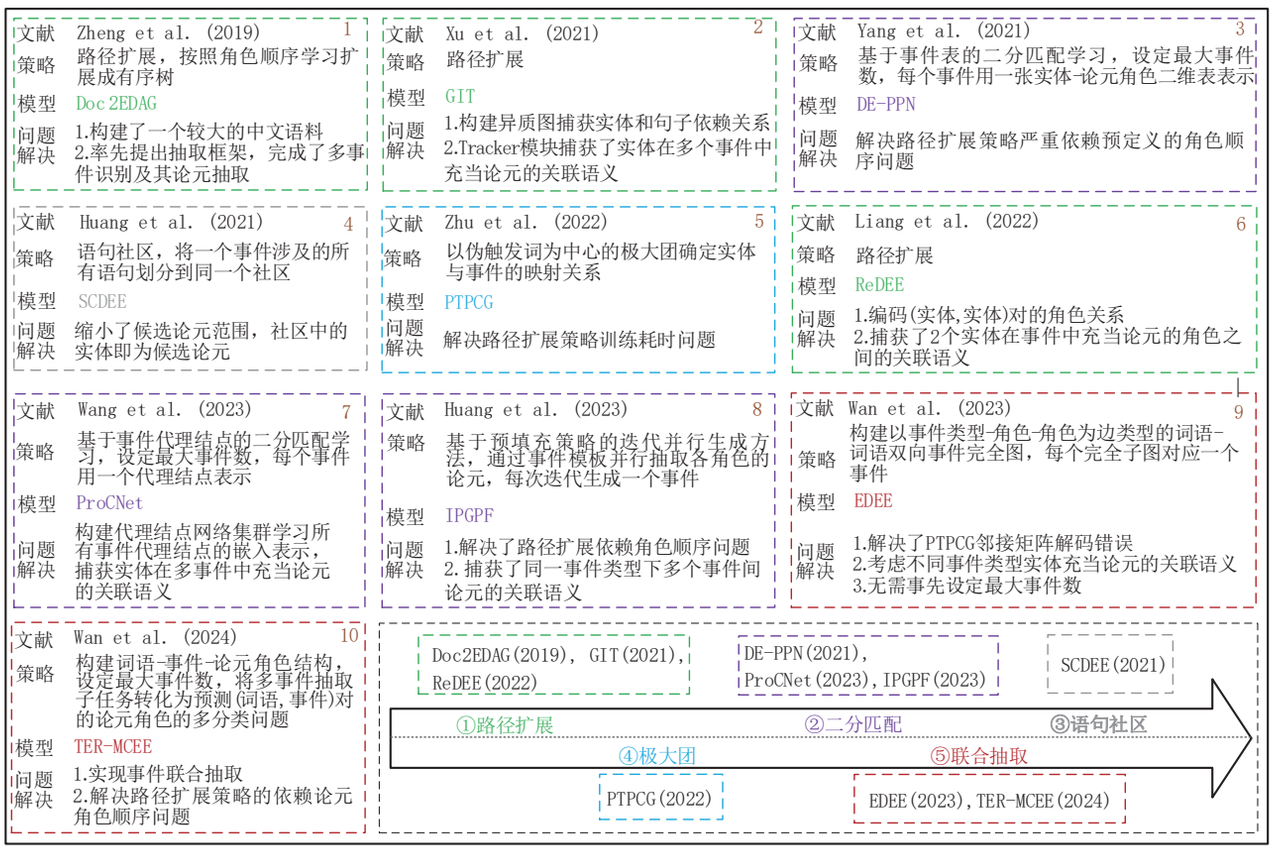


图5 ChFinAnn/DuEE-fin语料上DocEI & AE研究进展

注：图中首先对各模型进行总结；然后对各模型进行分类(右下部分所示)，其中不同虚框表示不同的策略，红色虚框表示联合抽取模式，其他都为流水线模式。

③ 语句社区和极大团策略。为了缩小事件的候选论元范围，抽取实体后，文献[3]和文献[31]首先分别通过构建语句社区和由伪触发词组成的极大团进行多事件识别(每个语句社区或极大团对应一个事件)，再将每个语句社区或基于极大团扩充得到的团中的实体作为该事件的候选论元，然后对每个事件进行事件类型判断和抽取各角色的论元。

④ 联合抽取策略。以上策略都属于流水线模式，为了克服流水线模式导致的错误传播问题，文献[11-12]分别通过构建词语-事件-论元角色结构和以事件类型-论元角色-论元角色关系为边类型的词语-词语双向事件完全图，将流水线(pipeline)模式涉及的事件抽取的所有子任务进行了集成。

2.2.2 DocEAE任务研究进展

现有DocEAE的研究主要集中于RAMS/WIKIEVENTS语料，研究思路主要分为：基于传统分类、机器阅读理解、文本生成、跨度选择等4种策略。由于各种策略实现方式不同，下面针对每种策略，分别阐述DocEAE任务的研究思路和研究进展(如图6所示)。

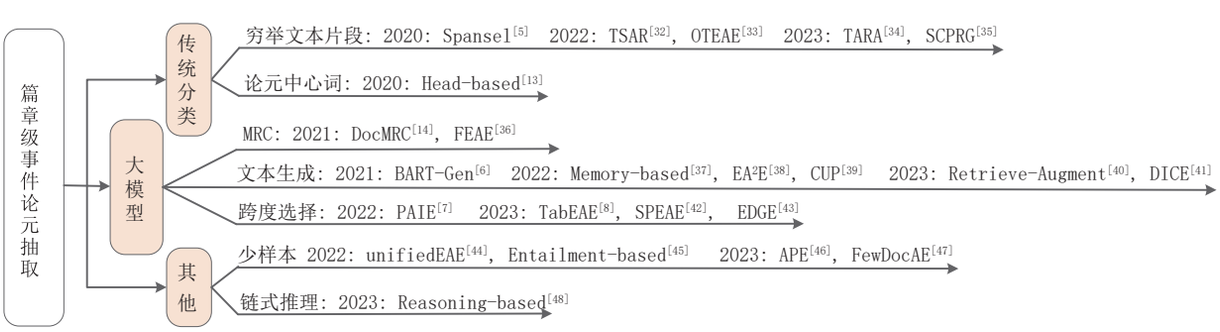


图6 RAMS/WIKIEVENTS语料上DocEAE研究进展

（1）基于传统分类。先产生候选论元，再对每个角色，分类判断每个候选论元是否充当该角色的论元。

针对复杂论元情况，文献[5]枚举一定长度范围内的所有文本片段作为候选论元，再针对每个角色，分类判断每个候选论元是否充当其论元。在其基础上，文献[32]和文献[34]采用抽象语义表示图（AMR）对篇章结构语义进行编码；文献[13]采用确定论元中心词的两阶段方法，减少候选论元的选择范围；文献[33]基于剪枝的句法依存树学习词语的表示；文献[35]考虑了非论元上下文线索和同一事件类型角色之间的关联。

（2）机器阅读理解。为每个角色生成一个问题，将问题和文本拼接成序列一起送到预训练语言模型的编码器，基于编码结果计算每个问题对应的答案（论元开始和结束位置）。

文献[14]和文献[36]将任务转换为机器阅读理解（MRC），通过确定论元开始和结束位置来避免产生候选论元。此外，文献[14]采用先在MRC数据集上进行预训练，再到DocEAE数据集上进行微调的方式，以缓解少样本带来的训练数据不足的问题。文献[36]利用相关论元（同一事件的其他论元）及其角色作为线索指导模型推理，捕获了同一事件中论元和论元之间的语义关系。

（3）文本生成。基于事件本体，为每种事件类型创建一个包含所有角色对应论元的事件模板（用占位符表示论元），采用序列到序列的编码器-解码器模型填充模板，即用具体论元代替论元占位符。

基于预训练语言模型（PLMs），文献[6]将DocEAE任务转化为基于事件模板的文本生成任务。文献[37-38]扩展生成模型，捕获了多事件间的关联语义；文献[40]将检索增强技术融入生成模型；文献[39]捕获了论元和触发词间的依赖关系；文献[41]将文本生成策略应用于临床医学领域的事件抽取。

（4）跨度选择器。基于角色的表示训练跨度选择器，分别确定论元在文本中的开始和结束位置。

文献[7]为每种事件类型设计了一个提示模板，基于PLMs返回的角色表示为每个角色生成2个跨度选择器（分别确定论元开始和结束位置），由于一个事件类型下的所有角色的跨度选择器是联合训练的，因此捕获了同一事件类型下各角色的论元之间的关联语义。在文献[7]的基础上，文献[42]基于软提示进行了扩展；文献[8]在文献[7]框架上致力于同时抽取多个事件的论元，捕获了多事件间的关联语义；文献[43]给每个事件建立事件内和事件间依赖感知图网络来捕获同一事件中的角色依赖关系。

以上4种策略中，后3种都是基于预训练语言模型，致力于充分发挥预训练语言模型在语言理解和生成方面的优势。

此外，文献[44-46]分别将迁移学习、文本蕴涵（Textual Entailment）应用于DocEAE任务，以缓解训练数据不足的问题；文献[48]采用链式推理（Chain Reasoning）解决长距离依赖；文献[47]构建了少样本数据集FewDocAE。

3　数据集

现有的篇章级事件抽取相关的数据集主要有：ChFinAnn、DuEE-fin、RAMS、WIKIEVENTS、MUC-4和DocEE。前2个为中文金融公告类语料，研究者们主要用于DocEI & AE任务研究；后4个为英文新闻类文章语料，主要用于DocEAE任务研究。

表1为6个语料基本概况，每列指标依次表示语料名称、标注方式、篇章数、事件类型数、角色数、论元数、事件数以及平均每个篇章包含的词语数和语句数、平均每个事件涉及的语句数。表2统计了常用的前4个语料在篇章多事件、跨句论元、多值论元、共享论元和嵌套论元等方面的情况。其中，“篇章多事件概况”各列依次表示多事件篇章占比、多事件篇章中事件平均数及存在跨事件类型的篇章占比；“跨句论元比例”、“多值论元比例”列分别表示跨句论元（即中心句中没有包含的论元，中心句是指触发词所在语句或包含一个事件的论元数最多的语句）、多值论元（即一个事件在同一角色上可能有多个论元）占总论元比例；“共享论元概况”各列依次表示共享论元（即同一个实体或文本片段同时充当多个角色的论元）、事件内共享论元和事件间共享论元占总论元比例；“嵌套论元概况”各列分别表示嵌套论元（即一个论元的部分片段也充当另一个角色的论元）占总论元比例、嵌套论元总数。

3.1　DocEI & AE任务语料

3.1.1　ChFinAnn

ChFinAnn是一个由事件知识库远程监督构建的、未标注事件触发词的大规模篇章级事件抽取数据集。涉及中文金融领域与股权相关的活动，包括以下5种事件类型：股权冻结（EF）、股权回购（ER）、

表1 语料概况

语料	标注方式	篇章	事件类型	角色	论元	事件	篇章词语数	篇章语句数	事件语句数
ChFinAnn	自动	32 040	5	35	289 871	47 824	912	20	6.2
DuEE-fin	人工	11 699	13	92	81 632	15 850	499	-	-
RAMS	人工	9124	139	65	21 237	9124	134	5	1.3
WIKIEVENTS	人工	246	50	59	5536	3951	771	25	1.0
MUC-4	-	1700	6	5	2641	-	291	13	-
DocEE	人工	27 485	59	356	180 528	27 000	592	27	-

表2 多事件及复杂论元统计

语料	篇章多事件概况			跨句论元	多值论元	共享论元概况			嵌套论元概况	
	比例	事件数	跨类型	比例	比例	比例	事件内	事件间	比例	论元数
ChFinAnn	29%	3	3%	43.1%	0.0%	42.9%	15.6%	28.8%	0.0%	81
DuEE-fin	29.2%	3	42%	-	9.3%	28.8%	1.0%	27.9%	11.9%	6402
RAMS	0%	0	0%	17.9%	0.0%	16.8%	16.8%	0.0%	1.6%	330
WIKIEVENTS	93%	17	88%	0.5%	17.9%	46.3%	0.7%	45.8%	29.4%	1,630

股权减持(EU)、股权增持(EO)和股权质押(EP),35种角色。由表1可知,ChFinAnn语料共32 040个篇章(文档),含47 824个事件,其中299个篇章中包含跨事件类型的事件,大部分篇章只包含同种事件类型的事件。29%的篇章包含多个事件,整个语料不存在多值论元。一个篇章平均包含912个字符、20条语句、1.5个事件和9.1个论元。官方分别将25 632个、3204个和3204个篇章用于训练、验证和测试。相较于人工标注语料,规模较大。

表3为ChFinAnn语料各种事件类型的分布情况。由表1~表3可知,ChFinAnn有以下特点:(1) 语料规模较大,且多事件大都属于同种事件类型。29.0%篇章包含多个事件,一个多事件篇章平均含3个事件,其中跨事件类型的篇章比率仅占3.0%,多发生在EU和EO事件类型。(2) 论元跨句普遍存在,且语句跨度大。98.0%的篇章存在论元跨句,平均一个事件的所有论元涉及6条语句。(3) 论元共享,且发生在事件内和事件间。42.9%的篇章存在论元共享,事件内和事件间的共享论元比例分别为15.6%和28.8%。(4) 事件类型分布不均匀。含EF事件类型的篇章非常少,仅占3.7%;而含EP事件类型的篇章非常高,达49.0%。

表3 ChFinAnn各种事件类型分布

事件类型	角色	篇章	事件	多事件比例	最大事件数
EF	8	1196	2050	32.0%	20
ER	6	3677	4557	16.1%	7
EU	6	5847	8035	24.3%	15
EO	6	6017	9046	28.0%	24
EP	9	15 602	25 035	35.4%	34
All	35	32 040	47 824	29.0%	34

3.1.2 DuEE-Fin

DuEE-Fin是百度发布的、由人工标注的篇章级事件抽取数据集,包括13种事件类型和92种角色,文本内容来源于中文金融领域的新闻和公告,涉及真实场景中的诸多问题,比如多值论元、嵌套论元等。每个事件记录都有一个标注的触发词,但未标注触发词和论元的在篇章中的位置。测试集是在线评估不公开的,且包括55 881个混淆样本(不包含任何事件记录的篇章),使得抽取工作更加复杂。由表1可知,DuEE-Fin语料共有11 699个篇章(文档),含15 850个事件。29.2%的篇章包含多个事件,同时9.3%的论元为多值论元。一个篇章平均含499个词语、1.4个事件和6.9个论元,最长篇章达3000个词语。官方分别将7015个、1171个和3513个篇章用于训练、验证和测试。

DuEE-Fin有以下特点:(1) 跨事件类型更普遍。29.2%篇章包含多个事件,其中跨事件类型的篇章比率高达42.0%。(2) 存在多值论元。对比ChFinAnn,存在9.3%的论元为多值论元。(3) 共享论元多发生于事件间。28.8%的论元为共享论元,其中大部分都发生于事件之间。

3.1.3 任务面临的挑战

目前在以上2个语料上,都是针对DocEI & AE任务进行研究,且重点是解决多事件识别问题。研究遇到如下挑战。

(1) 多事件:识别篇章中包含的各种类型的所有事件,如图1中EU类型的事件$event_1$、EO类型的

事件$event_2$和$event_3$。

（2）论元跨句：事件的论元散落在多个语句中。如事件$event_1$大部分论元在语句S_5中，但角色“股票保有量”的论元散落于语句S_8中。

（3）论元共享：事件中多个角色的论元共享同一实体，包括同一事件中的共享和多事件间的共享。如事件$event_1$的角色“交易开始日”和“交易结束日”共享同一实体“2009年1月7日”，同时事件$event_1$、$event_2$和$event_3$的角色“交易开始日”也共享同一实体“2009年1月7日”。

（4）多值论元：同一事件中同一角色有多个论元。如DuEE-Fin数据集中由触发词“融资”触发的事件类型“企业融资”中的某一事件中角色“投资方”同时有“招银国际”、“高盛”、“管理层”共3个论元，但在数据集ChFinAnn上没有该问题。

面对以上挑战，抽取策略和模型需要尽量捕获如下语义关联信息。

（1）实体与事件的映射关系。如图1中篇章包含多个事件和多个实体，如何确定实体与事件的映射关系对多事件识别至关重要。

（2）跨句论元的上下文语义。事件的论元散落在多个语句中，如图1中事件$event_1$大部分论元在语句S_5中，但角色“股票保有量”的论元散落于语句S_8中，因此，实体和语句的篇章级上下文语义直接影响后续事件类型判断和多事件抽取子任务。

（3）论元和角色的约束语义。事件的角色用于描述同一类型事件的某一方面信息，有一定的语义，扮演同一角色的论元性质相对固定。如日期角色“StartDate”的论元为日期或空值，施事者角色“Pledger”的论元大都为人名、组织机构等。

（4）角色之间的关联语义。相同或不同事件类型的不同角色之间有一定的关联，如图1中股权增持事件$event_1$和股权增持事件$event_2$的施事者角色“股票持有者”肯定不会是同一个实体充当论元，因为语义上应该是事件$event_1$的施事者将股份转给事件$event_2$的施事者，两个事件的施事者不会是同一个。

（5）事件论元间的关联语义。相同或不同实体在同一事件内或不同事件中充当相同或不同角色的论元之间有一定的关联语义。割裂事件内或事件间的事件论元间的关联语义，将影响事件论元抽取效果。

注意：DocEI & AE任务目前主要是针对ChFinAnn数据集的特点进行研究，DuEE-Fin数据集多用于跨数据集验证。

3.2　DocEAE任务语料

3.2.1　RAMS

RAMS是由人工标注的含触发词的篇章级事件抽取数据集。文本内容取自英文新闻，语言描述呈现多样化；标注时将论元的范围限制在围绕触发词的5条语句的窗口内。RAMS包括139种事件类型和65种角色。由表1可知，该数据集共3993个文档，被切分为9124个篇章，一个篇章仅含一个事件，共9124个事件。每个篇章约5条语句，平均含134个词语和2.3个论元。官方分别将7329个、924个和871个篇章用于训练、验证和测试。

RAMS有以下特点：（1）篇章较短且仅含一个事件。每个篇章平均含5条语句，仅含一个事件。（2）事件类型多样性。数据涵盖了139种事件类型和65种角色，事件类型多样化。（3）复杂论元。与财经公告相比，新闻语料存在较高比例的论元为带有限定性描述内容的文本片段，同时描述形式也更多样化，导致更难抽取。（4）事件类型分布不均匀且少样本。139种事件类型，每种类型涉及的事件数都比较少，甚至有多种事件类型的样本数仅为1。（5）论元跨句不明显。平均每个事件涉及1.3条语句，多数论元都在中心句。

3.2.2　WIKIEVENTS

WIKIEVENTS是另一个由人工标注的含触发词的篇章级事件抽取数据集。文本来源于维基百科，包括50种事件类型和59种角色。由表1可知，该数据集共246个篇章（文档），有3951个事件，93.0%的篇章包含多事件，17.9%的论元为多值论元。一个篇章平均包含771个词语、25条语句、16个事件和22.5个论元。官方分别将206个、20个、20个篇章用于训练、验证和测试。

WIKIEVENTS有以下特点：（1）语料规模较小且多事件非常普遍。整个语料只有246个篇章（文档），多事件占比93%，多事件篇章平均含高达17个事件。（2）实体多。一个篇章平均含135个实体，与论元无关的信息过多，加大了论元抽取的难度。（3）跨事件类型比例高。多事件篇章中88.0%篇章含不同事件类型的事件，给多事件抽取带来了挑战。（4）多值、共享、嵌套论元的比例高。分别存在17.9%、46.3%和29.4%的多值论元、共享论元和嵌套论元。

3.2.3　MUC-4和DocEE

MUC-4[49-50]数据集是以发生在拉丁美洲的恐

怖袭击为主题的英文新闻文章组成，包括绑架、袭击、爆炸、抢劫、纵火、强迫停工共6种事件类型，5种角色：犯罪者、组织、目标、受害者和武器[51]。该数据集包含1700个篇章（文档），规模较小且47.4%篇章未标注事件论元[52]。官方分别将1300个、100个篇章用于验证和测试[53]。

MUC-4具有如下特点：(1) 不同事件类型间的角色相似[53]。整个语料共5种角色被所有事件类型共享。(2) 数据集很小。1700个篇章中只包含2641个论元。

DocEE[52]是从维基百科中提取的新闻热点事件，包括社会紧急事件，如地震、交通事故和武装冲突，以及与人类生活有关的有趣事件，如名人事迹、体育赛事和其它以娱乐为中心的报道，共定义了59种事件类型。

DocEE有以下特点：(1) 语料规模较大且角色粒度细。356种角色使得对事件的描述非常详尽，但对模型的角色语义消歧能力提出了更高的挑战。(2) 多值论元和跨句论元现象严重。新闻报道对事件的部分角色（如事故损失、调查结果、事故原因等）进行了多方面描述，导致一个角色存在多个论元；同时角色粒度细，导致一个完整事件存在较多的论元，且这些论元分散在多个语句中。21.4%的论元属于多值论元（即该论元所对应的角色有多个论元）；一个事件平均涉及10.2条语句，跨句论元多，对模型处理长文本能力提出了挑战。(3) 复杂论元。8.4%的论元是带有限定性描述内容的文本片段。

3.2.4 任务面临的挑战

在以上4个英文语料上，目前都是针对DocEAE任务进行研究，且重点是解决复杂论元的抽取问题。研究遇到如下挑战。

(1) 语言描述形式多样化。4个英文语料来源于新闻或维基百科，相较于固定格式的财经公告类语料，语句表达形式呈现多样化，描述相同角色的论元时可能采用多种表述方式，不易于模型理解语义。

(2) 复杂论元和嵌套论元。DocEI & AE语料的事件论元都由命名实体（组织机构、人名、日期）等描述的简单对象充当，模型可抽取实体作为候选论元，而RAMS中有一定比例的论元为复杂论元（如图2中角色“Crime”的论元），导致候选论元太多，影响模型抽取效果。另外，嵌套论元也是DocEAE面临的一大挑战。

(3) 小样本。RAMS和WIKIEVENTS数据规模不大、事件类型多样且分布不均匀，导致多种事件类型的样本量较少，影响模型训练。

面对以上挑战，论元抽取策略和模型需要尽量捕获如下语义关联信息。

(1) 中心句和篇章语义理解。目标事件触发词所在的语句称为中心句，由表1可知，前2个英文语料的事件论元大都集中在中心句，因此对中心句语义的充分理解将有助于候选论元的产生和事件论元的抽取。另外，由于新闻语料语言表述多样化，充分理解整个篇章语义对论元角色分类有重要影响。

(2) 捕获角色和词语语义。小样本且事件类型丰富的情况下，每种事件类型涉及的样本量较少，导致基于深度学习模型的训练往往不足，充分捕获角色和词语的语义将减少小样本对模型分类效果的影响。

注意：DocEAE任务目前主要是针对RAMS数据集的特点进行研究，WIKIEVENTS数据集多用于跨数据集验证。针对MUC-4和DocEE数据集的研究较少[51-52, 54-55]。

4 主要研究方法

由于DocEI & AE任务和DocEAE任务的难点不同、侧重的研究问题也不同，因此，本文针对这2个任务分别分析对应的研究方法及推进情况。考虑到ChFinAnn、DuEE-Fin、RAMS和WIKIEVENTS语料最为常用，本节主要探讨基于这4个语料的研究方法。

4.1 DocEI & AE任务的研究方法

目前，DocEI & AE任务主要基于ChFinAnn语料和DuEE-Fin语料开展，由第3.1.3节可知，2个语料上存在多事件、论元跨句和论元共享等挑战，关键是多事件识别及其论元抽取。研究者们针对不同挑战提出了相应解决问题的方法，主要分为路径扩展、二分匹配学习、语句社区、极大团和联合抽取等策略。

4.1.1 路径扩展策略

路径扩展策略主要包括Doc2EDAG、GIT和RAAT等相关成果。下面首先介绍每个成果及它们的关注点；然后对比3个模型，归纳它们之间的联系和区别；最后，介绍路径扩展策略的优缺点。

(1) Doc2EDAG[1]：基于实体有向无环图的篇章级事件抽取。为了研究多事件识别和共享论元抽取问题，文献[1]构建了一个较大规模的篇章级事件

抽取语料ChFinAnn，并提出了路径扩展框架Doc2EDAG。

在多事件识别和论元抽取子任务中，针对每种事件类型，以该事件类型为根节点，按照预定义的角色顺序，针对当前树的每个叶节点开展路径扩展。每步路径扩展，实际上是对每个实体进行二分类，判断是否充当该角色的论元，以决定是否进行扩展。对每个要扩展的实体，创建实体节点并连接到当前叶节点来扩展路径。在上述扩展过程中，当有多个实体充当该角色的论元时，路径则会分出新支。对每个角色不断扩展路径分支，直至最后一个角色。显然，从根节点到叶节点的每一条路径可以唯一标识一个事件，且路径上的第k个实体节点即为扮演该事件第k个角色的论元。通过路径扩展，同时解决了多事件识别及其论元抽取问题。

图7为路径扩展示例，图中右上部分是事件记录（黄金标注），下部分为路径扩展策略生成的事件树，左上部分是角色“股票持有者”的路径扩展示例，下部分分支出的3条路径与右上部分的事件记录相对应。

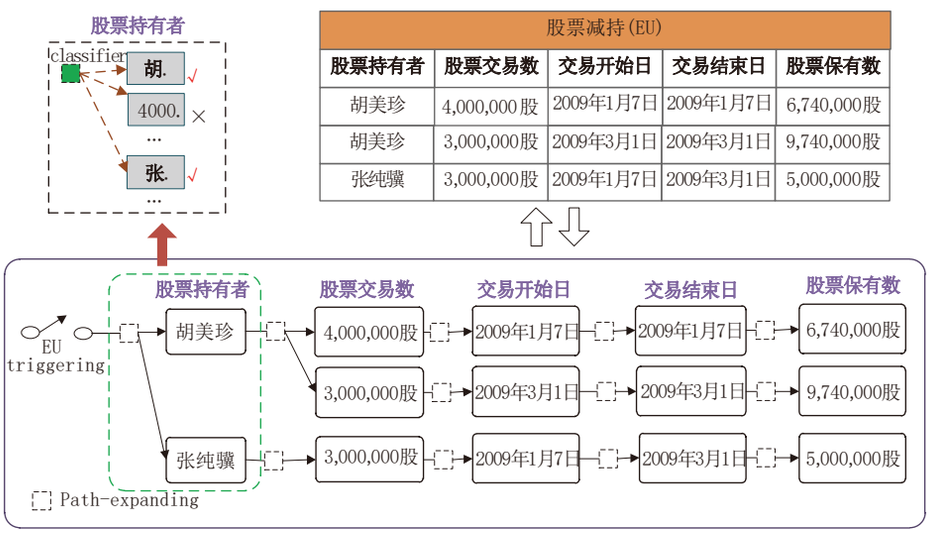


图7　路径扩展策略示例

路径扩展子任务中，为了利用已扩展的实体节点信息，每条路径独立拥有一个记忆张量$\boldsymbol{M}$来存储该路径已扩展的实体序列。图8展示了基于路径扩展策略的Doc2EDAG模型框架。左下部分蓝虚框是扩展路径时针对叶子节点开展的路径扩展子任务，需编码实体序列$\boldsymbol{E}$、路径序列$\boldsymbol{M}$和当前角色$Role_i$信息。随着路径的扩展，每条路径上的记忆张量$\boldsymbol{M}$不断更新，且只存储该条路径上已扩展的实体序列，如图8中上、下2条路径的记忆张量$\boldsymbol{M}$各不同。

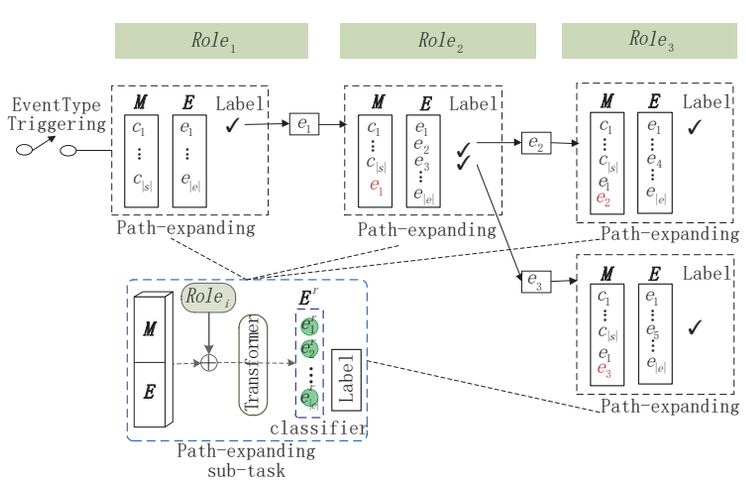


图8　基于路径扩展策略的Doc2EDAG模型框架

（2）GIT[9]：基于异质交互图的篇章级事件抽取。为了捕获跨句论元的上下文语义，文献[9]构建了异质交互图（GIT），以篇章中所有实体和语句为节点，构建以下4种边类型：任意2个语句连边、实体与所在语句连边、同属一个语句的2个实体连边、同一实体的2个提及连边。此外，为了捕获实体在多个事件中充当/不充当论元的关联语义，在路径扩展子任务中设计了一个全局记忆张量存储所有已扩展路径的实体序列，并用于每一个路径扩展子任务。

图9展示了基于路径扩展策略的GIT模型框架。左下部分蓝虚框是GIT扩展路径时针对每个叶子节点开展的路径扩展子任务，除实体外，还需编码该条路径实体序列$\boldsymbol{U}_i$、全局记忆张量$\boldsymbol{G}$和当前角色$Role_j$信息。$\boldsymbol{U}_i$用于存储第i条路径已抽取的实体序列，$\boldsymbol{G}$用于存储所有路径已抽取的实体序列信息，Tracker负责将全局记忆张量$\boldsymbol{G}$共享给所有路径扩展子任务，并实时将所有路径的抽取结果更新到$\boldsymbol{G}$。为了图片清晰，只标出了路径$\boldsymbol{U}_j$抽取角色$Role_3$时的$\boldsymbol{G}$信息。

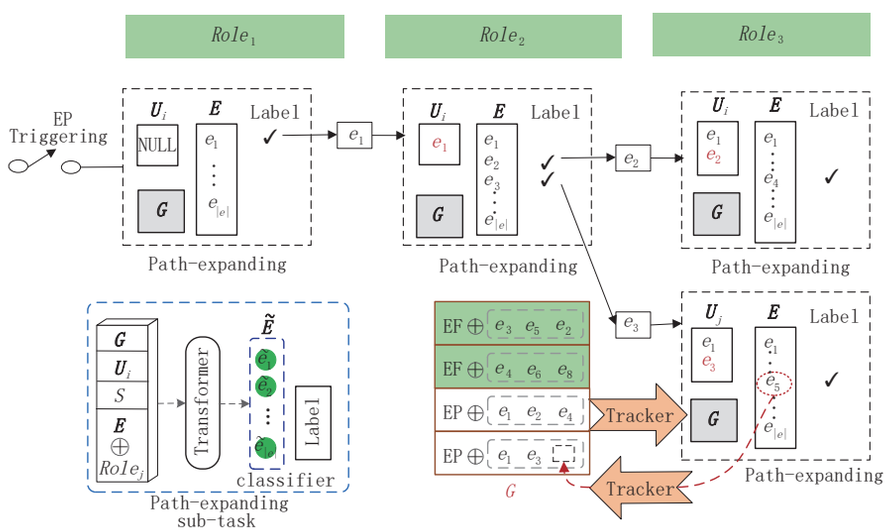


图9 基于路径扩展策略的GIT模型框架

注：假设EF事件类型已抽取完毕，现抽取EP事件类型。$\boldsymbol{U}_i$用于存储第i条路径已抽取的实体序列，$\boldsymbol{G}$在所有扩展子任务中共享，并通过Tracker模块实时更新。

（3）ReDEE[10]：基于关系增强注意力Transformer的篇章级事件抽取。为了捕获实体在同一事件中充当论元的角色之间的关联语义，文献[10]基于Doc2EDAG框架，提出了一个关系增强注意力Transformer模型（RAAT）。针对每一种事件类型，按照预定义的角色顺序，将任意2个满足$m<n$的角色$Role_m$和$Role_n$组成角色关系，其中m、n表示角色序号。每一种角色关系用一张实体-实体二维表表示2个实体在相同事件中充当论元的角色之间的对应关系。此外，用一张（实体+语句）-（实体+语句）二维表表示GIT异质交互图中节点间的关系。将以上各种关系合并为矩阵$\boldsymbol{T}$，并集成到Transformer的关系增强注意力计算中，形成具有独立注意力计算模块的RAAT。

（4）3个模型对比。GIT和RAAT基于Doc2EDAG路径扩展框架，就实体语义编码和扩展机制2个方面进行了探索。图10从这2个方面展示了上述3个模型的区别及推进情况。

综上所述，Doc2EDAG的编码较简单，且每条路径独立进行预测，割裂了实体在多事件中充当论元的关联语义。为此，GIT通过构建异质交互图丰富了实体的语义信息，同时引入Tracker模块跟踪已抽取事件，并用于后续路径扩展，进一步提升了抽取效果。ReDEE旨在捕获同一事件的论元的角色关系，设计了一种关系增强的注意力Transformer。

对于路径扩展策略，篇章中多事件的识别以及其论元抽取融入所有路径中，使得每条路径对应一个事件，巧妙地解决多事件识别问题。同时，根据给定的论元角色顺序，可以明确路径中每个节点充当的角色。然而，正是由于路径扩展的特点，使得论元角色顺序需要事先确定。由于在扩展路径中的后续节点时必须基于已构建的路径，导致抽取效果严重依赖指定的论元角色顺序，且训练十分耗时。

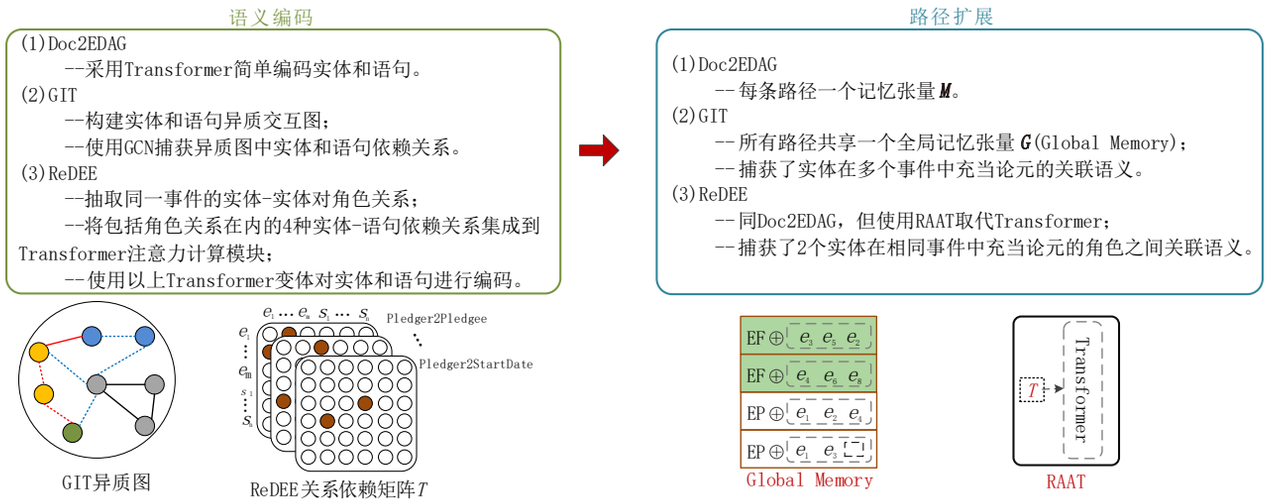


图10　3个路径扩展模型区别及推进情况

4.1.2　二分匹配策略

为了解决路径扩展策略依赖指定角色顺序的问题，文献[2]和文献[4]通过设定篇章中包含的最大事件数(避免多事件识别)，分别将每个事件建模为一张实体-论元角色二维表或一个代理节点。文献[56]则通过设定迭代次数，基于人工设计的事件模板迭代产生事件，从而明确每个事件中实体与角色的对应关系(即论元识别)。模型只需预测实体在该事件的每个角色中是否充当论元。由于预先设定事件数或迭代次数，预测的事件与黄金事件之间没有明确的对应关系，所以该策略通常还需解决如何将预测的m个事件分配给黄金标注的k个事件，即二分匹配策略。

图11右部分为二分匹配学习示例。图中橙色表示黄金标注的事件集合，蓝色表示预测的事件集合，通过定义2个事件的损失函数，以总和损失最小化为训练目标，利用匈牙利算法求解预测事件和标注事件的对应关系。

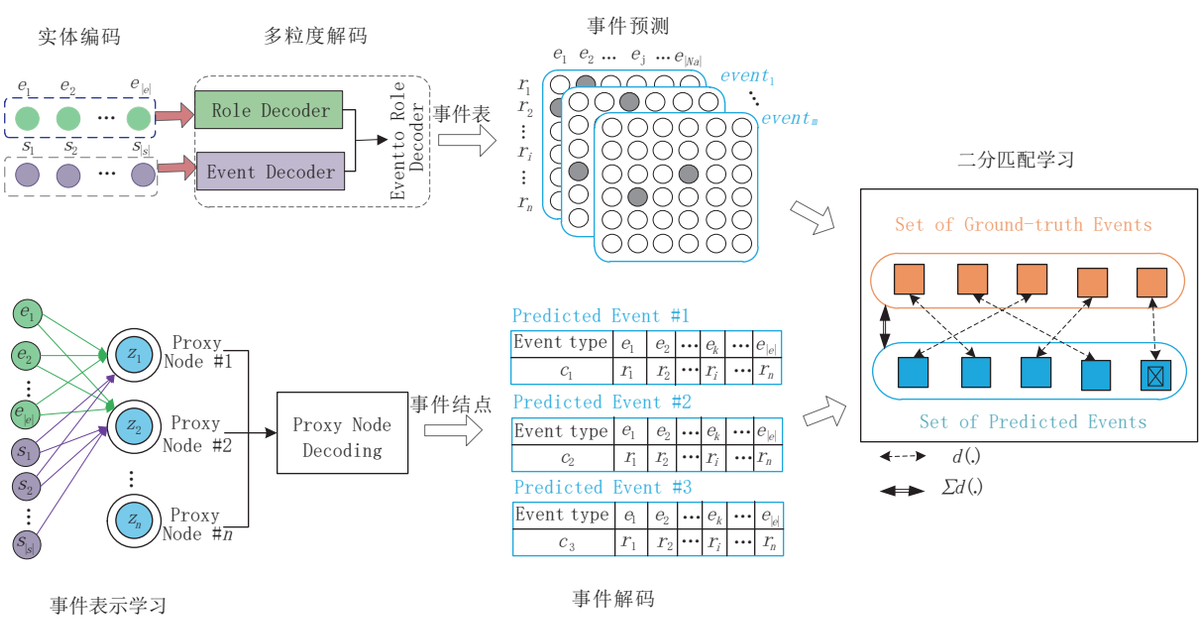


图11　基于二分匹配策略的模型框架

注：图上部分为基于并行预测网络的DE-PPN模型，图下部分为基于事件代理节点和豪斯多夫距离的ProCNet模型，其中e_i、s_j、z_l和r_k分别表示实体、语句、事件和角色。

下面我们介绍现有成果中采用二分匹配策略的3种方法，分别为DE-PPN、ProCNet和IPGPF。

(1) DE-PPN[2]。基于并行预测网络的篇章级事件抽取。文献[2]将每个事件建模为一张实体-论

元角色的二维表，并提出基于事件表的并行预测网络（DE-PPN）。DE-PPN是一个编码器-解码器模型，编码器用于获取篇章语义表示，解码器包含角色解码器和事件解码器。其中，角色解码器主要将同一个事件的论元聚集在一起，解决事件论元识别问题；事件解码器支持产生多个事件，解决多事件识别问题。

图11上部分展示了DE-PPN抽取框架，其中行r_i表示论元角色，列e_j表示实体。每个事件采用一张实体-论元角色二维表揭示在该事件中实体是否充当指定角色的论元，并以并行的方式从篇章中抽取指定事件类型下的所有事件。此方法存在2个不足：①需要事先设定篇章包含的事件数；②每个事件基于一张实体-论元角色二维表独立进行二分类，割裂了实体在多个事件中充当论元的关联语义。

（2）ProCNet[4]。基于事件代理节点和豪斯多夫距离的篇章级事件抽取。由于DE-PPN割裂了实体在多个事件中充当论元的关联语义，为了缓解该局限，文献[4]首先引入事件代理节点，将所有实体、语句分别与代理节点进行关联，构建三者之间的交互图，丰富事件代理节点的语义。然后，针对每个事件代理节点的表示解码得到对应事件，包括事件类型判断和论元抽取步骤。

图11下部分展示了ProCNet抽取框架。基于交互图学习代理节点的表示，再解码为一个事件，通过总和最小化进行全局调优，捕获了实体在多事件中充当论元的关联语义。该方法存在以下局限：①不能解决同一实体在相同事件中充当不同角色的论元；②模型依赖最大事件数的设定，设定过大会有过多代理节点映射到空事件，导致计算冗余；③未对实体和语句建模，使得实体表示缺乏上下文语义，导致跨句论元分类不够准确。

（3）IPGPF[56]。基于预填充策略的迭代生成篇章级事件抽取。同样为了解决并行预测网络DE-PPN事件间语义割裂的局限，IPGPF提出基于预填充策略的并行迭代生成方法。针对每种事件类型，人工设计一个模板（包含该类型下的所有角色），并行计算模板中每个角色的候选论元得分，从而确定各角色对应的论元，完成事件论元抽取。设定最大迭代次数，基于预填充策略（预先填充部分角色的论元）每次迭代生成一个事件，完成多事件识别。其中，历史生成的事件用于迭代生成下一个事件，捕获了同一事件类型下多个事件各角色的论元间的关联语义。

该方法存在以下不足：①不能解决不同事件类型下多事件间各角色的论元关联语义丢失问题；②模型需要设定迭代次数，设置过大将过滤过多空事件，过小则会丢失事件。

综上所述，二分匹配策略旨在破除路径扩展策略依赖指定角色顺序问题，通过设定事件数或迭代次数等超参数，避免多事件识别。针对每个事件，模型只需预测实体在该事件的哪些角色下充当论元，简化了模型的学习难度。然而，由于需要设定事件数或迭代次数，不同的事件数或超参数对抽取效果带来一些影响。此外，由于指定事件数，通常容易导致实体在事件间的关联语义丢失（如DE-PPN），尽管ProCNet和IPGPF模型在其基础上做了改进，但依旧丢失了同一实体在相同事件中充当不同角色的关联语义和在不同事件类型下多事件间各角色的论元的关联语义。

4.1.3 语句社区策略

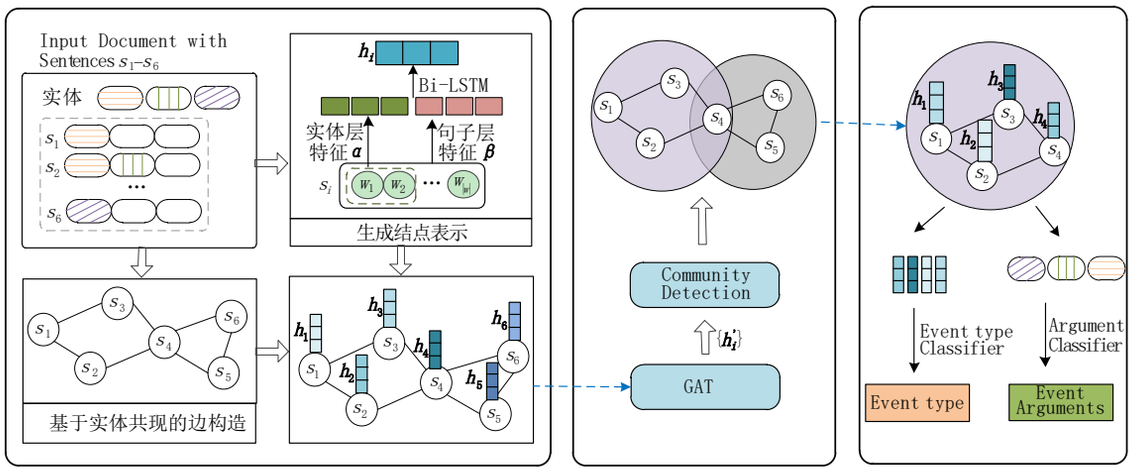
由于二分匹配策略将篇章中所有实体作为每个事件的候选论元，为此文献[31]提出语句社区概念，将一个事件涉及的语句划分到同一个社区（每个社区对应一个事件，避免多事件识别），社区中的实体即为事件的候选论元，从而缩小了候选论元的范围。然而，该策略增加了建立语句与社区映射关系的子任务。

图12为基于语句社区的SCDEE模型框架。篇章中的语句S_i为节点，将实体共现的语句连边（图13左下部分）形成篇章图；采用图注意力网络（GAT）编码图结构信息，并计算语句S_i隶属社区c的概率p（图12中间部分）。针对每个社区，进行事件类型判断及论元抽取。

由于语句社区粒度较大，不同事件对应的社区时常相同（即社区包含的语句相同），且社区包含了较多噪声实体（其他事件的实体），导致模型效果不佳。

4.1.4 极大团策略

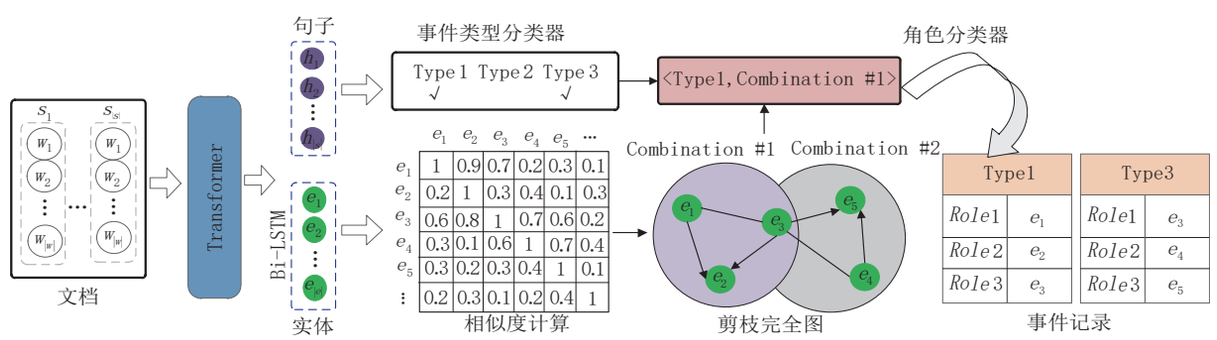
路径扩展策略属于自回归方式，需大量的计算资源，导致长篇章中的训练和推理效率不佳。二分匹配和语句社区策略需要指定事件数，给抽取效果带来影响。为了同时解决上述问题，文献[3]利用极大团的性质确定篇章中包含的事件数，解决多事件识别问题。即提出伪触发词概念，利用以伪触发词为中心的极大团确定实体与事件的映射关系，并用非自回归解码算法从剪枝完全图中解码事件论元组合，是一个快速轻量级的模型。剪枝完全图中，针对



（a）构造文档图　　（b）检测语句社区　　（c）事件类型及论元分类

图12　基于语句社区策略的SCDEE模型框架

注：s_i表示语句，$\boldsymbol{h}_i$为s_i的节点向量。首先以语句为节点，基于实体共现连边构建篇章图；再基于实体和语句表示计算语句节点的向量表示；然后检测语句社区，一个社区表示一个事件，社区中的实体作为事件候选论元。



（1）事件类型检测及实体抽取　　（2）抽取团　　（3）事件记录生成

图13　基于极大团策略的PTPCG模型框架

每种事件类型，根据角色论元的存在性和区分性选择一组伪触发词（提前设定伪触发词个数），将伪触发词连向其他非伪触发词，伪触发词之间互连，每个事件形成以伪触发词为核心的极大团，多个事件形成目标的剪枝完全图。

图13为基于极大团策略的PTPCG模型框架。模型以目标剪枝完全图对应的邻接矩阵为目标训练，针对预测得到的邻接矩阵对应的剪枝完全图，将每个极大团解码为一个事件，即$<Type_i, \{Combination_j\}>$对，再进行论元角色判断，最终生成所有事件记录。

然而，由于不同事件存在共享伪触发词问题，使得基于剪枝完全图解码事件时存在错误，即训练目标本身就存在错误。

4.1.5　联合抽取策略

第4.1.1节—第4.1.4节的方法均采用流水线模式，即将篇章级事件抽取任务分解为实体抽取、篇章包含的事件类型判断、给定事件类型下多事件识别及其论元抽取（即多事件抽取）等子任务分步骤独立完成。然而，这种策略存在错误传播，降低了抽取效果。为此，文献[11-12]分别设计了一种篇章级事件联合抽取策略，并揭示了联合抽取方法的优势。

篇章级事件联合抽取策略的关键在于设计数据结构使得能够揭示实体在何种事件类型下的哪个事件中充当何种角色的论元，所以现有方法主要致力于设计满足上述要求的数据结构。

（1）TER-MCEE[11]：基于词语-事件-论元角色的多通道事件抽取。文献[11]通过构建词语-事件-论元角色结构，揭示了词语-事件的对应关系和<词语，事件>对于论元角色的匹配关系，实现了事件的联合抽取。可将实体抽取、给定事件类型下的多事件抽取子任务集成，转化为预测<词语，事件>对的论元角色的多分类问题。为了实现事件抽取的完全联合执行，避免篇章包含事件类型判断的子任务，策略1借助多通道技术，针对每种事件类型，设

计一个通道执行上述多分类训练。汇总所有通道的损失作为整个模型的总损失，指导模型训练。策略2将所有通道融合为一个张量，捕获同一个词语在不同通道间的语义关联。

图14为TER-MCEE的模型框架[11]。该框架主要包含5个模块：句法解析模块、词语-事件-论元角色构造模块、特征整合模块、Bi-LSTM层模块、多通道论元角色预测模块。其中，词语-事件-论元角色构造模块生成<词语，事件>对的论元角色标签；多通道论元角色预测模块，每个通道处理一种事件类型，针对该事件类型训练模型预测<词语，事件>对的论元角色标签。然而，该方法需要事先设定语料包含的事件数，策略1中不同事件类型独立进行训练预测，割裂了每个任务中模型之间的交互和词语在不同事件类型的多个事件中充当论元的关联语义；策略2忽略了不同词语之间的关联语义。

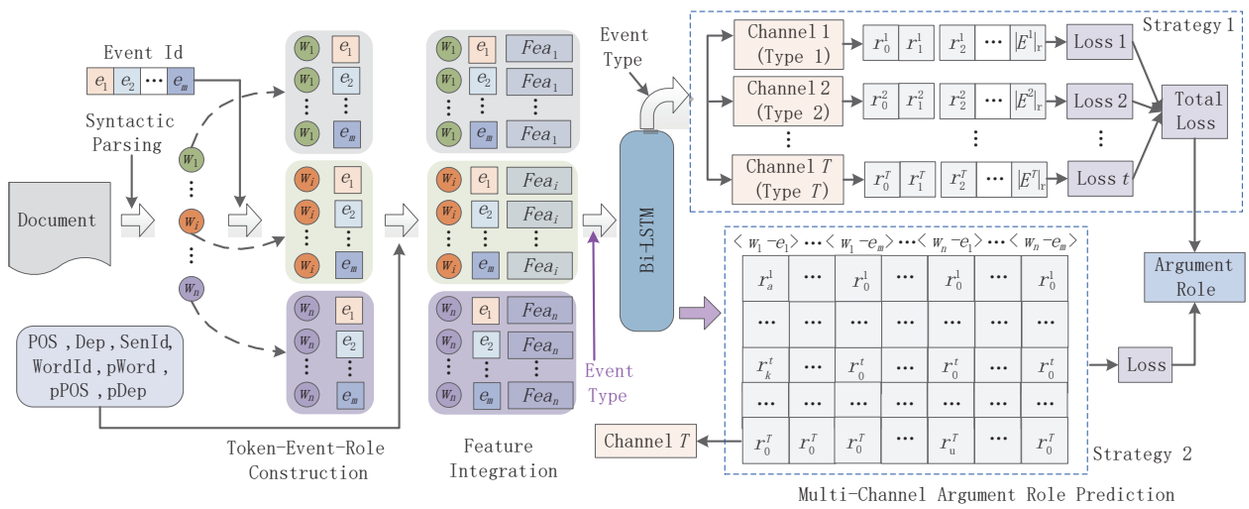


图14 基于联合抽取策略的TER-MCEE模型框架

(2) EDEE[12]：边增强的篇章级事件联合抽取。由于极大团策略中不同团的伪触发词可能相同(即共享伪触发词)，导致从基于语料构成的实体-实体邻接矩阵中解码实体-事件对应关系时本身就存在错误，即不能无损还原语料中标注的事件信息。为此，文献[12]舍弃以伪触发词为中心的策略，设计了一个以事件类型-论元角色-论元角色关系为边类型的词语-词语双向事件完全图，并借助图结构的邻接矩阵，开发了一个边增强的篇章级事件联合抽取模型，将传统流水线模式涉及的事件抽取的所有子任务进行了集成。

基于联合抽取策略的EDEE模型框架如图15所示[12]。EDEE模型对每个<词语，词语>对进行分类，判断其关系标签，得到图15中间部分所示的预测的词语-词语邻接矩阵。然后，根据预测的词语-词语邻接矩阵，解码其蕴含的所有事件和事件记录信息，包括图结构解码(确定完全子图数，解决多事件识别问题)和边类型解码(由关系三元组确定词语在哪种事件类型的某个事件中充当何种论元角色的论元，解决事件类型识别和论元抽取问题)，如图15右部所示。

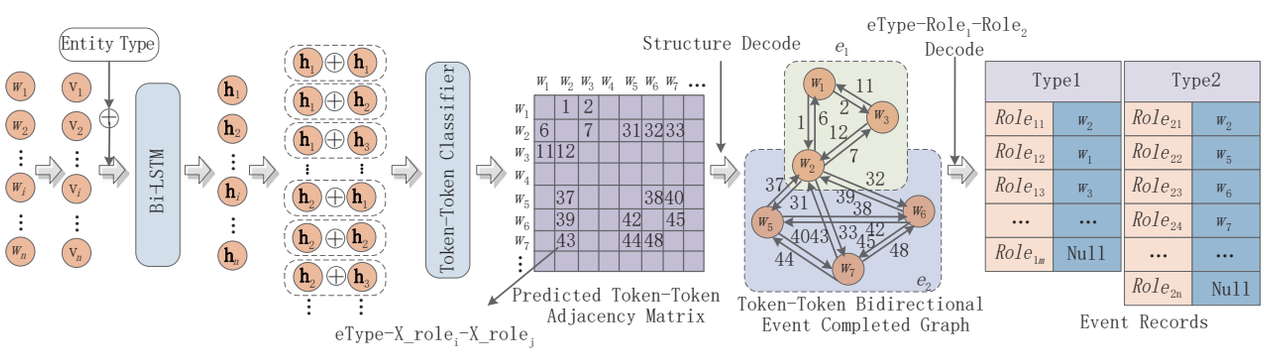


图15 基于联合抽取策略的EDEE模型框架

4.1.6　小结

图16对上述各类模型的推进情况进行了总结和展示。

路径扩展策略(Doc2EDAG、GIT、RAAT)针对每种事件类型生成一棵实体有序树，使得树中每条路径对应一个事件，路径上的实体节点为事件论元，同时解决了多事件识别和论元抽取问题。然而，该方法依赖预定义的论元角色顺序，且采用自回归策略，训练耗时严重。为了解决依赖预定义的论元角色顺序问题，本文提出了并列预测网络DE-PPN，每个事件以一张实体-论元角色的二维表表示，但割裂了实体在多个事件中充当论元的关联语义。为此，ProCNet引入事件代理节点，构建代理节点、实体和语句的交互图，丰富事件的表示；IPGPF提出事件迭代生成策略，将历史生成的事件记录用于迭代生成下一个事件，捕获了多事件间的关联语义，然后基于分配问题来求解预测事件集合与黄金标注事件集合的最佳二分匹配。

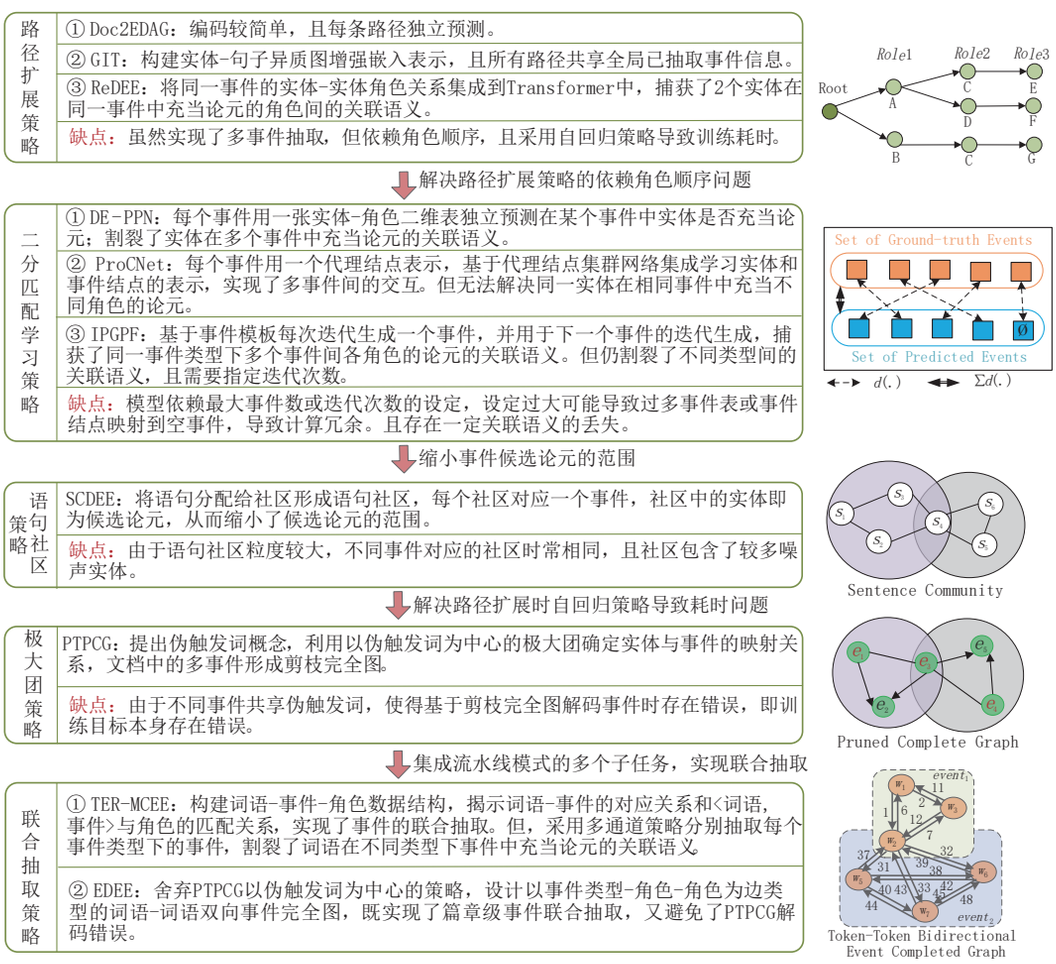


图16　DocEI & AE任务的各模型推进情况

为了缩小每个事件的候选论元，SCDEE构建语句社区，使得每个社区对应一个事件，社区中的实体即为事件的候选论元。由于语句社区粒度较大，不同事件对应的社区时常相同，且社区包含了较多噪声实体，导致模型效果不佳。为此，PTPCG提出伪触发词概念，并基于团确定实体与事件的映射关系。然而，由于不同事件存在共享伪触发词问题，使得基于团解码事件时存在错误，即训练目标本身就存在错误。

以上模型都是基于流水线模式，存在一定的错误传播。为此，文献[11-12]探讨了篇章级事件联合抽取策略。该策略的关键在于设计数据结构使得能够揭示实体在何种事件类型下的哪个事件中充当何种角色的论元。基于该目标，研究者分别提出了篇章级事件联合抽取模型TER-MCEE和EDEE。

4.2　DocEAE任务的研究方法

由第3.2节可知，RAMS和WIKIEVENTS语料存在语言表述形式多样、复杂论元和嵌套论元、小

样本等挑战。现有事件论元抽取主要基于这2个语料，研究者们针对不同挑战提出了相应解决问题的方法，主要分为传统分类、机器阅读理解、文本生成、跨度选择器等策略。对于少量不属于上述类别的成果，我们将其列入“其他”类别。

4.2.1 传统分类策略

传统分类方法的核心是候选论元的生成，主要包括如下2种策略。

(1) 穷举文本片段

针对复杂论元的情况，文献[5]首先穷举指定长度内的所有文本片段作为候选论元；然后根据触发词计算候选论元的得分，将前k个得分的文本片段作为该事件的候选论元；最后针对每种角色，引入候选论元、触发词和角色的链接得分函数，得分最高的为该角色对应的论元。后续有不少研究在文献[5]的穷举文本片段策略上，致力于捕获上下文中词语的关联和事件之间的关联。

为了分析篇章结构，丰富词语语义，文献[33]提出利用最优传输理论将单个句法依存树转换为篇章结构，同时提出一种新的正则化技术来约束不相关上下文词语在最终分类预测中的贡献。为了捕获更多的语义信息，文献[32]从局部(句内)和全局(整个篇章)两个角度对篇章进行编码，丰富了词语的表示，降低跨句上下文对抽取效果的影响；同时引入边界损失加强边界信息。

为了显式地捕获论元和事件之间的依赖关系，文献[34]将DocEAE任务转换为基于抽象语义表示(AMR)图的链接预测模型。模型首先压缩了原始AMR图中信息量较少的子图和边的类型；然后挑选穷举的前k个得分的文本片段作为候选论元，整合到AMR图中，并在图中标记同一篇章中的其他事件触发词，捕获其他事件的结构信息。

为了考虑非论元上下文线索和同一事件类型角色之间的关联，文献[35]提出一个基于文本跨度-触发词的上下文池化(STCP)和潜在角色交互模型。其中，STCP可根据预训练模型中指定的＜论元，触发词＞对的注意力权重，自适应地融合非论元线索词的语义，增强候选论元与相关上下文的表示；潜在角色交互模块，可以引导角色间进行交互，捕获角色间语义相关性，并与STCP模块的候选论元嵌入表示拼接作为论元角色分类的依据。

(2) 基于论元中心词的扩展

为了降低候选论元的选择空间，文献[13]提出两阶段候选论元抽取方法。

第一阶段：检测候选论元中心词。该阶段不直接确定候选论元，而是通过计算每个词语作为触发事件指定角色的候选论元中心词的概率，发现候选论元的中心词(通常是离论元句法依存树根节点最近的词语)。

第二阶段：围绕中心词扩展候选论元。从中心词的左右2个方向按一定长度范围进行扩展，产生所有可能的候选论元。

对于两阶段抽取方法，候选论元的选择空间从穷举文本片段转换为中心词的检测空间加上围绕中心词的候选论元的扩展空间，大大减少了候选论元的数量。

4.2.2 机器阅读理解策略

为了避免抽取候选论元和缓解少样本问题，文献[14]和文献[36]将DocEAE任务转换为机器阅读理解(MRC)问题。首先，模型根据触发词确定触发事件的事件类型，为该类型下的每个角色生成一个问题；然后，将问题和文本拼接成序列一起送到预训练语言模型中进行编码；最后，基于编码的结果计算每个问题所对应的答案在文本中的开始和结束位置，完成该问题对应角色的论元抽取。下面介绍该策略下的2种方法。

(1) DocMRC[14]：基于隐式和显式数据增强的MRC模型。隐式数据增强是利用MRC和其他任务的语料，如语义角色标注、ACE事件抽取语料库进行预训练，并将这些任务构建为MRC问题，建立统一的MRC训练框架；再使用本任务语料对模型进行微调。隐式数据增强将其他任务中学到的知识隐式地转移到篇章级事件论元抽取任务，有效提升少样本下的事件论元抽取性能。显式数据增强使用预训练的MRC模型标注新的数据样本，从而增加新的样本数量。

图17为基于机器阅读理解策略的DocMRC模型框架。模型首先基于MRC和相关任务的语料学习统一框架的MRC模型；然后基于篇章级事件抽取语料微调；最后基于MRC模型标注新的样本，显式地增加样本数据。

(2) FEAE[36]：基于课程知识蒸馏的训练策略。文献[14]通过建立论元和事件触发词之间的直接关联，捕获了同一事件论元之间的关联语义。FEAE基于MRC，在抽取指定角色的论元时，利用相关论元(同一事件的其他论元)及其角色作为线索指导模型推理，旨在捕获同一事件中论元之间的语义关系。由于推理阶段不能利用黄金标注的论元信息，

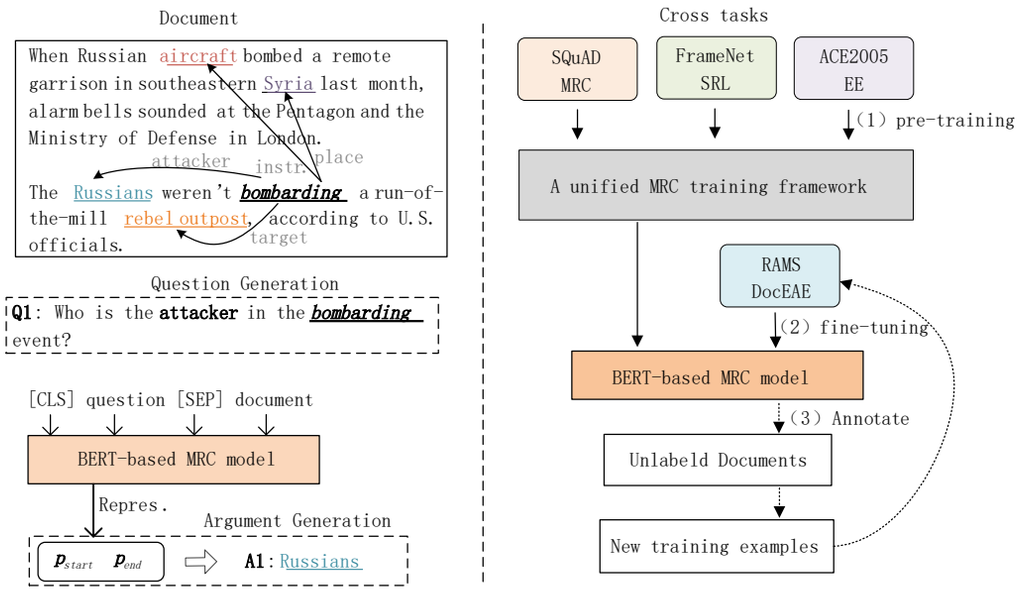


图17　基于机器阅读理解的DocMRC模型框架

注：左侧为一般的DocMRC模型框架，其中上部分为篇章，中间为针对角色“attacker”生成的问题，下面为基于预训练语言模型抽取论元的示例，**Q1**表示抽取角色“attacker”的论元的问题，“Repres.”表示基于预训练语言模型编码后的问题和篇章表示，p_{start}和p_{end}分别为基于该表示得到角色“attacker”的论元在篇章中的开始和结束位置，**A1**为对应的论元。右侧为数据增强的DocMRC模型框架。

从而引入知识蒸馏策略[57]，即先将相关论元及其角色作为线索训练教师模型；然后训练学生模型来模仿教师模型的推理（蒸馏阶段），期间以教师模型输出的监督信息（知识）来监督训练；最后以最终无额外相关论元信息的学生模型来抽取指定角色的论元。

另外，在知识蒸馏中引入课程学习[58]，通过减少指定论元的比例来增加蒸馏过程的学习复杂性，从而促进神经网络的训练。即在蒸馏开始时，将所有相关论元输入学生模型训练，然后逐渐减少论元比例，最终过渡到完全不使用论元信息。

4.2.3　文本生成策略

为了避免抽取候选论元，基于预训练语言模型，文献[6]将DocEAE任务转化为基于事件模板的生成任务，并创建了WIKIEVENTS语料。基于事件本体，为每种事件类型创建一个包含所有角色对应论元的事件模板（用占位符表示论元），采用序列到序列的编码器-解码器模型计算生成概率，最终生成填充的模板，即用具体论元代替论元占位符。

图18为基于文本生成策略的事件论元抽取模型框架。图上半部分为篇章和构建的事件模板，篇章中用<*tgr*>标记目标事件触发词（篇章中绿色部分），模板中用占位符<*arg*>表示论元。模型将篇章和模板作为输入，解码生成填充论元的模板。最后根据生成的模板，输出事件记录。

为了捕获多个事件的论元间的关联语义以及事件间论元结构之间的一致性（如事件中扮演警察角色的论元不太可能在其他事件中扮演攻击者角色的

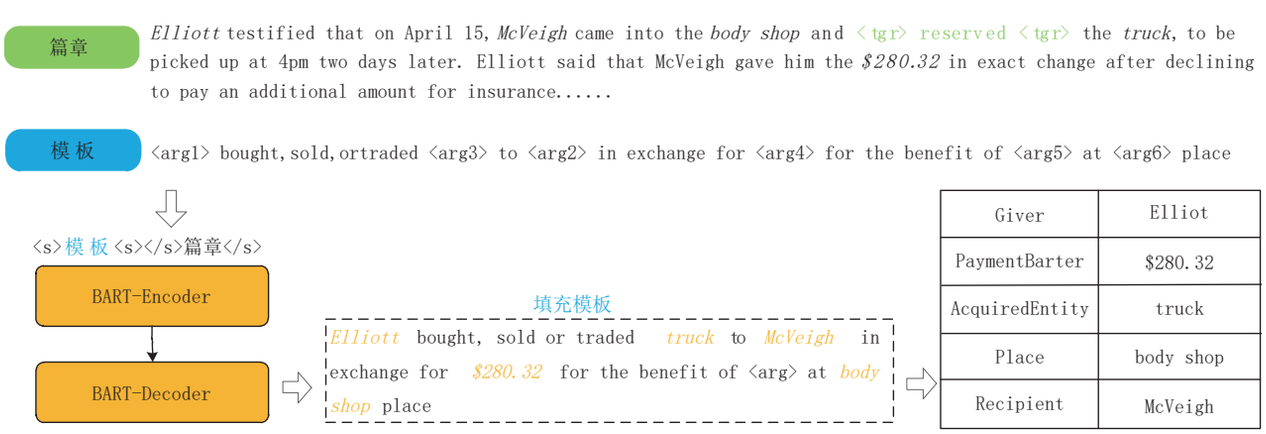


图18　基于文本生成策略的模型框架

论元），文献[37]提出一个基于全局记忆增强的编码-解码事件论元生成框架。通过构造一个全局存储器来记录已抽取的事件论元信息，并将其输入解码器训练基于文本生成的事件论元抽取模型，以捕获事件间的关联语义。在解码过程中，基于全局存储器和基于知识角色对的限制（如：已解码的实体，不再会解码成与已解码角色不兼容的角色的论元），完成论元抽取。

考虑到引入事件-事件关系有利于事件间论元抽取的一致性，文献[38]提出事件感知的论元抽取模型，通过标记篇章中其他事件的论元来增强上下文，隐式地对事件间的关系建模。训练过程中，模型通过在常规上下文（未标记任何事件论元）下不断逼近目标（黄金标注），丰富事件论元的表示。推理时，将已抽取论元加入上下文用于下一论元抽取，显式地增强同一事件中论元间的关联语义。

为了捕获论元和触发词间的依赖关系，文献[39]提出一种基于课程学习的提示调优文本生成模型，利用抽象语义表示图捕获论元和指定触发词间的依赖关系；利用课程学习框架，分4个阶段由易到难地训练论元抽取模型。每个阶段基于编码器-解码器模型，填充提示模板（用论元填充占位符），完成事件论元抽取。其中，提示模板根据不同阶段的学习难度进行调整，前2个阶段使用语料标注的模板，后2个阶段使用基于抽象语义表示图构造的模板，增强模型对不同论元和触发词的推理能力。

文献[40]将检索增强生成技术应用于事件论元抽取，即检索外部知识，并将其作为文本生成的线索。文献[41]将文本生成策略应用于临床医学领域的事件抽取。

4.2.4 跨度选择策略

跨度选择策略，即根据角色的表示训练得到2个跨度选择器：跨度开始和结束位置选择器，从而确定论元。该策略无需抽取候选论元，直接根据以上2个选择器确定论元的跨度。

（1）PAIE[7]：基于Prompt的跨度选择策略。由于基于问答的方式是每次针对一个角色进行论元抽取，没有利用同一事件中各角色论元之间的关联语义。为了捕获同一事件中各角色论元之间的关联语义，同时避免产生候选论元，文献[7]基于预训练语言模型，提出角色交互模型PAIE，其架构如图19所示。

该策略的核心思想：首先为每种事件类型设计合适的Prompt，每个角色用一个槽表示（对于多值论元，在Prompt中为该角色设计多个槽），如图19左下部下划线所示。然后，将Prompt输入预训练语言模型，生成该事件类型下每个角色的表示，如图19中间下部所示。最后，针对每个角色表示，联合训练预测开始位置、结束位置的2个跨度选择器，确定该角色的跨度（文本片段），如图19中最下面带箭头的弧线所示。针对多值论元，引入二部匹配损失处理同一角色多个跨度的匹配问题，如图19右上部所示。

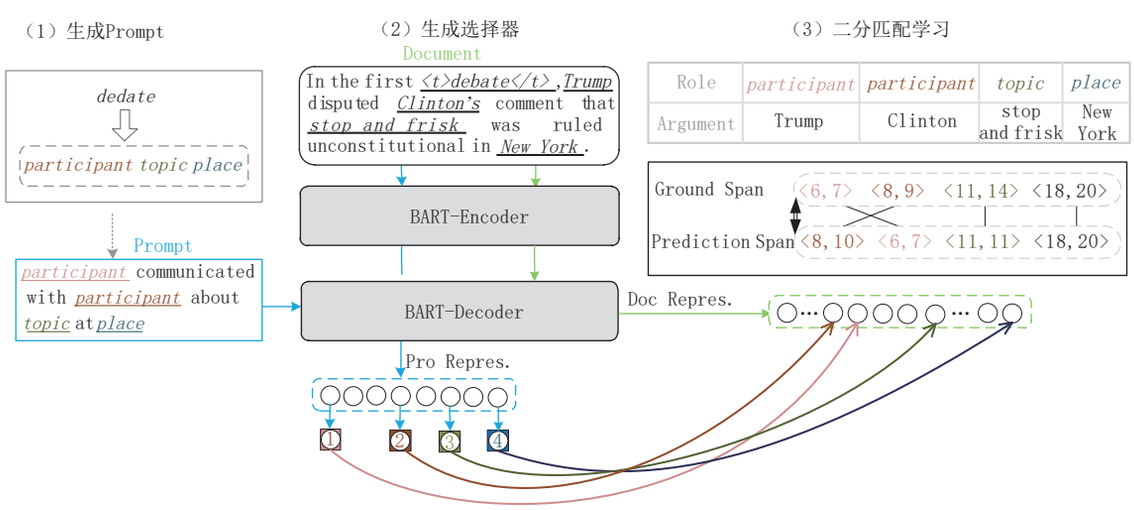


图19 基于跨度选择策略的PAIE模型框架

注：左部分为根据触发词“debate”生成Prompt；中间部分为基于给定篇章和Prompt，模型基于PLMs生成Prompt和篇章表示，前者用于生成特定事件类型的每个角色的跨度选择器（蓝色框中的每一个小圆圈代表Prompt中每一个词语的表示，①②③④分别对应Prompt中4个角色的跨度选择器）；右部分为求解预测跨度和黄金跨度的最佳二分匹配。

对于PAIE模型，由于一个事件类型下多个角色的跨度选择器是基于语料联合训练得到的，因此模型隐式地捕获了同一事件中多个角色的论元之间的关联语义。

文献[8]在PAIE基础上进行扩展，提出了基于表生成的TabEAE模型，捕获了不同事件类型多个事件论元之间的关联语义。文献[7]和文献[8]的提示方法主要依赖于人工设计的提示模板，不能根据特定篇章的上下文来更新角色的表示。为此，文献[42]提出了软提示抽取模型，通过构造事件类型-篇章图来丰富事件类型的表示，然后将其与对应触发词的嵌入表示拼接，作为角色表示的增强信息。

(2) EDGE[43]：基于事件内和事件间角色交互的依存感知图网络。文献[43]首先给每个事件建立事件内部图(事件类型-角色连通图)来捕获同一事件中的角色依赖关系；然后，设置检索模块来存储和检索当前事件的类似事件，利用检索到的事件内部图帮助构建事件间图(事件之间角色关系图)，以捕获事件之间的依赖关系。

4.2.5　其 他

为了模型能够适用于不同格式的数据集，文献[44]提出一种基于变分信息瓶颈(一种实现数据压缩和表示学习的方法)的多格式迁移学习框架。模型利用现有的多个数据集中的公共知识帮助事件论元抽取任务在新数据集上训练，即提出一个提示框架，从不同格式的数据集中学习共享格式和特定格式的知识。

与文献[44]目标一致，文献[46]提出了跨数据集的迁移学习模型。首先，通过定义跨数据集的重叠知识，将数据集上的知识划分为跨数据集的重叠知识和目标数据集的特定知识。然后，分别在2个阶段，利用文本生成方式学习跨数据集的这2部分知识，并构建了实体类型提示，激发重叠知识在事件论元抽取任务中的作用。同样为了解决多源数据集及少样本问题，文献[45]将事件论元抽取任务建模为文本蕴涵(根据一个文本片段判断另一个文本片段的真实性[59])任务，在多个蕴涵数据集上预训练蕴涵模型，再应用于事件论元抽取任务。

为了捕获跨句论元间的依赖关系，文献[48]引入链推理范式，通过局部推理形成的链式结构来描述事件论元抽取过程。文献[47]对现有的DocEAE数据集重新采样，构建了少样本数据集FewDocAE，并探索了少样本情况下的篇章级事件论元抽取。

4.2.6　小 结

图20展示了上述各类模型的推进情况。根据它们的动机和采用的策略，主要分为4大类。其中，机器阅读理解策略、文本生成策略、跨度选择策略是基于大模型得到的。在不同研究策略中，大模型发挥的作用不尽相同，如基于大模型将事件论元抽取任务转化为问答、文本生成等任务，或基于大模型获取角色的表示。下面针对这4类策略分别进行归纳。

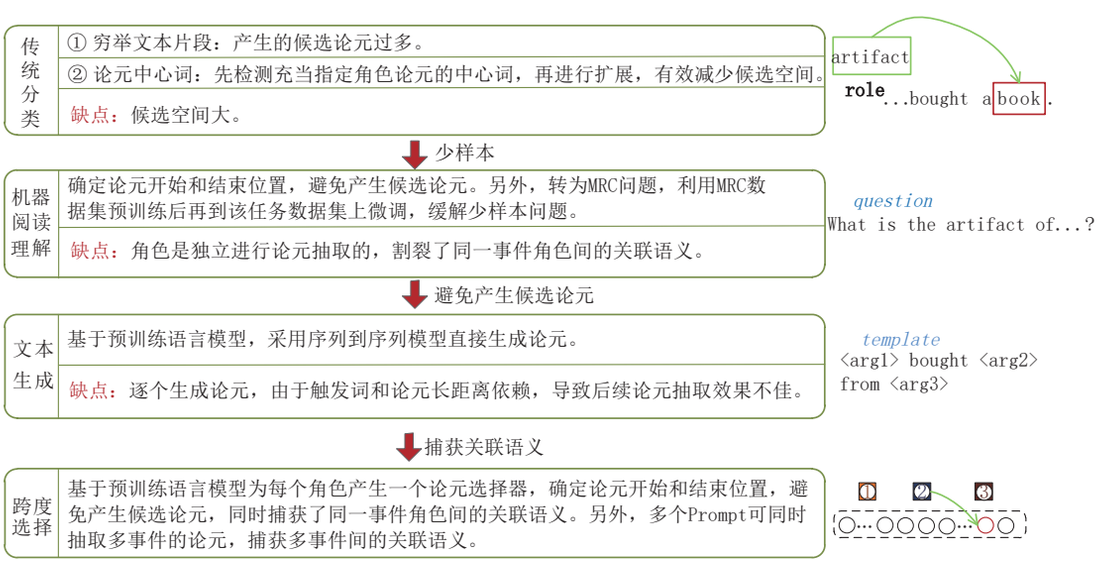


图20　基于RAMS/WIKIEVENTS语料各类模型推进情况

(1) 传统分类策略。该类方法主要采用候选跨度分类策略，关键在于候选论元的确定。早期采用穷举方法确定候选论元，这种方法导致候选论元数量巨大。随后相关研究探讨如何减少候选论元数

量，采取2阶段的策略：先确定某个中心词语代表候选论元，再从中心词扩展到整个候选论元。传统分类策略的优点在于能够充分利用篇章信息（如篇章中的各种关联信息），使得词语的语义较为丰富，但候选论元的确定存在挑战；2阶段的方法在扩展过程中存在一定难度。

（2）机器阅读理解策略。该类方法主要根据标注的触发词为对应事件类型的每个角色生成一个问题，基于预训练语言模型以问答形式逐个角色完成论元抽取，该类策略的关键在于理解问题和上下文。现有研究主要在DocEAE数据集上微调基于相关任务构建的MRC模型，有效缓解少样本问题。该类方法的优点包括：①避免了抽取候选论元；②通过构建每个角色的问题，捕获了论元和触发词之间的直接关联。缺点是每个角色的论元以及每个事件都是独立抽取，丢失了角色之间和事件之间的关联语义。

（3）文本生成策略。该类方法采用序列到序列的编码器-解码器框架填充事件模板，其关键在于增强上下文语义。早期研究直接将篇章和事件模板拼接作为上下文，后续相关研究主要探讨如何用论元与论元、论元与触发词间的关联语义来增强上下文。例如，将已抽取的最相似事件作为额外的上下文、标记篇章中其他事件的论元来增强上下文、利用AMR图捕获论元和触发词间的依赖关系、检索外部知识作为生成的线索等。该类策略既可以避免候选论元抽取难题，又能捕获同一事件各角色论元之间的关联语义，为模型提供了更多语义信息，但每个事件是独立的，忽略了事件之间的关联语义。

（4）跨度选择策略。该类方法主要基于角色表示训练跨度开始和结束位置的选择器，关键在于如何丰富角色表示的语义。早期研究者通过设计事件提示模板（Prompt），利用预训练语言模型获取角色的表示；近些年相关研究均旨在丰富角色语义，包括：①利用篇章实例增强事件类型的表示，间接丰富角色的语义；②建立事件类型-角色图，直接更新角色的表示；③采用多个事件并行抽取方式，捕获不同事件论元之间的关联语义。该类策略的好处同样是避免了候选论元抽取，且通过提示模板捕获了同一事件类型角色之间的关联语义。然而，只基于模式层面（事件本体结构）丰富角色的语义存在瓶颈，增加的语义信息十分有限，可以考虑如何在该策略框架下充分利用篇章实例信息或者根据不同的篇章差异化丰富角色语义。

5 未来趋势分析

通过梳理篇章级事件抽取的研究问题及其进展，本文对篇章级事件抽取未来发展趋势的分析如下。

（1）获取更多语义信息/缩小候选跨度的抽取策略。现有方法没有完全挖掘蕴含的语义信息，如角色的语义信息、事件类型的语义信息，角色、事件类型、触发词、篇章之间的关联语义信息，事件的上下文信息等，这些信息的利用都能提升篇章级事件抽取的效果。此外，巨大的候选跨度范围也是导致抽取效果不佳的一个重要因素。因此，如何有效编码这些信息或者减少候选跨度都可能成为篇章级事件论元抽取（DocEAE）的热点。

（2）基于少样本的篇章级事件论元抽取。目前支持篇章级事件抽取的语料有限，虽然中文数据集的规模还可以，但主要为金融公告；英文数据集数据量较小，无法有效支撑大多数深度学习模型训练。此外，数据集中的事件类型分布不均匀导致很多事件类型的样本非常少。因此，研制基于少样本的篇章级事件论元抽取（DocEAE）策略可能是未来的一个研究热点，如利用零样本和元学习等技术，减轻事件抽取任务对大规模标记数据的依赖。

（3）面向跨度论元的篇章级事件抽取。面向复杂论元的篇章级事件抽取语料主要有RAMS和WIKIEVENTS，但基于这2个数据集的研究任务目前都为篇章级事件论元抽取，即给定事件触发词信息，仅解决事件论元抽取问题，降低了抽取难度，同时也缩小了研究成果的应用范围。此外，篇章通常包含多个事件，同时解决事件的识别与跨度论元的抽取，虽然增加了抽取难度，但研究成果具有更大的应用推广价值。因此，设计面向跨度的篇章级事件识别及其论元抽取（DocEI & AE）模型也可能是未来的一个研究趋势。

（4）基于新数据集的篇章级开放事件抽取。现有中文篇章级事件抽取数据集主要包含金融公告，篇章形式和描述比较固定，且均为实体论元；英文数据集尽管包含部分跨度论元，但规模偏小。此外，这些数据集都指定了事件类型，限制了抽取的对象，不能较好地适用于新事件类型。财经评论中的语言描述复杂，涉及很多角色的事件成分，如施事者、受事者、时间和地点等，触发行为发生/状态改变的背景、范围、条件和使用的工具等，行为发生/状态改变的

结果、强度和情感等，而事件的这些成分对于完整地理解财经评论中事件的行为发生/状态改变的语义是非常重要的，同时存在大量的复杂论元和论元共享（包括部分共享和完整共享，甚至是整个事件作为其他事件的论元）现象。此外，财经评论逻辑性较强，存在较丰富的事件关系。因此，如何从财经评论中抽取篇章级开放事件具有重要的应用价值和较大的研究挑战。本文作者及其团队已标注了面向财经评论的篇章级开放事件抽取的语料，后期公布后可能会吸引一定数量的研究者关注，使得篇章级开放事件抽取或者事件关系抽取成为一个研究热点。

（5）基于大模型的篇章级事件论元抽取。由于大模型在许多任务上表现优异，研究者将会考虑如何利用大模型帮助提升事件抽取的效果，从而引发基于大模型的篇章级事件抽取的研究热潮。例如，利用大模型的零样本提示、上下文情境学习、思维链等技术，设计更好的事件论元抽取的提示Prompt，获取更多的语义信息；或者设计其它新的篇章级事件论元抽取的执行框架。

6　总　结

本文对篇章级事件抽取的研究现状进行了较为深入的回顾和分析，主要针对两类研究（子）任务DocEI & AE和DocEAE，分别从研究目标与任务、常用数据集、研究进展和主要研究方法等方面展开讨论。首先，明确任务目标，阐述解决问题的基本思路，归纳现有研究进展。然后，总结常用数据集的特点，归纳2类数据集分别针对DocEI & AE和DocEAE任务所面临的挑战。紧接着，重点深入分析不同数据集下的主要研究策略，并以图示化方式展示各方法的区别及推进情况。最后，结合篇章级事件抽取仍需攻破的问题，讨论篇章级事件抽取未来发展趋势。

致　谢　由衷地感谢论文评审专家和编辑对本文所提出的修改建议！

参　考　文　献

[1] Zheng S, Cao W, Xu W, et al. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 337-346

[2] Yang H, Sui D, Chen Y, et al. Document-level event extraction via parallel prediction networks//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). Bangkok, Thailand, 2021: 6298-6308

[3] Zhu T, Qu X, Chen W, et al. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph// Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI). Vienna, Austria, 2022: 4552-4558

[4] Wang X, Gui L, He Y. Document-level multi-event extraction with event proxy nodes and Hausdorff distance minimization// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 10118 - 10133

[5] Ebner S, Xia P, Culkin R, et al. Multi-sentence argument linking//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, USA, 2020: 8057-8077

[6] Li S, Ji H, Han J. Document-level event argument extraction by conditional generation//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, USA, 2021: 894-908

[7] Ma Y, Wang Z, Cao Y, et al. Prompt for extraction? PAIE: prompting argument interaction for event argument extraction// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, 2022: 6759-6774

[8] He Yu, Hu J, Tang B. Revisiting event argument extraction: Can EAE models learn better when being aware of event cooccurrences? //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 12542-12556

[9] Xu R, Liu T, Li L, et al. Document-level event extraction via heterogeneous graph-based interaction model with a tracker// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). Bangkok, Thailand, 2021: 3533-3546

[10] Liang Y, Jiang Z, Yin D, et al. RAAT: Relation-augmented attention transformer for relation modeling in document-level event extraction //Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Seattle, USA, 2022: 4985-4997

[11] Wan Q Z, Wan C X, Xiao K L, et al. Token-Event-Role structure-based multi-channel document-level event extraction. ACM Transactions on Information Systems (TOIS), 2024, 42: 1-27

[12] Wan Q Z, Wan C X, Xiao K L, et al. Joint document-level

event extraction via token-token bidirectional event completed Graph//Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 10481-10492

[13] Zhang Z, Kong X, Liu Z, et al. A two-step approach for implicit event argument detection//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, USA, 2020: 7479-7485

[14] Liu J, Chen Y, Xu J. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction //Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021: 2716-2725

[15] Hu Rui-Juan, Zhou Hui-Juan, Liu Hai-Yan, et al. Survey on document-level event extraction based on deep learning. Computer Engineering and Applications, 2022, 58: 47-60 (in Chinese)
(胡瑞娟, 周会娟,刘海砚等. 基于深度学习的篇章级事件抽取研究综述. 计算机工程与应用, 2022, 58: 47-60)

[16] Wan Hao-Chang, Zhou Chen-Lian, Marius Gabriel Petrescu. Survey on event extraction based on deep learning. Journal of Software, 2023, 34: 3905-3923 (in Chinese)
(王浩畅, 周郴莲, Marius Gabriel Petrescu 基于深度学习的事件抽取研究综述. 软件学报, 2023, 34: 3905-3923)

[17] Liu Q, Luan Z, Wang K L, et al. Document-level event extraction: A survey of methods and applications. Journal of Physics: Conference Series, 2023. 2504: 1-7

[18] Wan Qi-zhi, Wan Chang-xuan, Hu Rong, et al. Chinese financial event extraction base on syntactic and semantic dependency parsing. Chinese Journal of Computers, 2021, 44: 508-530 (in Chinese)
(万齐智, 万常选,胡蓉等. 基于句法语义依存分析的中文金融事件抽取. 计算机学报, 2021, 44: 508-530)

[19] Wan Q Z, Wan C X, Xiao K L, et al. A multi-channel hierarchical graph attention for open event extraction. ACM Transactions on Information Systems (TOIS), 2023, 41(1): 1-27

[20] Wan Q Z, Wan C X, Xiao K L, et al. CFERE: Multi-type Chinese financial event relation extraction. Information Sciences, 2023, 630: 119-134

[21] Wan Q Z, Wan C X, Xiao K L, et al. Construction of a Chinese corpus for multi-type economic event relation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2022, 21(6): 1-2

[22] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP). Beijing, China, 2015: 167-176

[23] Nguyen T.H, Cho K, Grishman R. Joint event extraction via recurrent neural networks//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). San Diego, USA, 2016: 300-309

[24] Sha L, Qian F, Chang B, et al. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensorbased argument interaction//Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 5916-5923

[25] Liu X, Luo Z, Huang H. Jointly multiple event extraction via attention-based graph information aggregation//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 1247-1256

[26] Chen Y, Yang H, Liu K, et al. Collective event detection via a hierarchical and bias tagging networks with gated multilevel attention mechanisms//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 1267-1276

[27] Yang S, Feng D, Qiao L, et al. Exploring pre-trained language models for event extraction and generation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy, 2019: 5284-5294

[28] Liu J, Chen Y, Liu K. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection //Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Hawaii, USA, 2019: 6754-6761

[29] Tong M, Xu B, Wang S, et al. Improving event detection via open-domain trigger knowledge//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, USA, 2020: 5887-5897

[30] Du X, Cardie C. Event extraction by answering (almost) natural questions//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2020: 829-838

[31] Huang Y, Jia W. Exploring sentence community for document-level event extraction// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-findings). Punta Cana, Dominican Republic, 2021: 340-351

[32] Xu R, Wang P, Liu T, et al. A two-stream amr-enhanced model for document-level event argument extraction// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Seattle, USA, 2022: 5025-5036

[33] Veyseh A P B, Nguyen M V, Dernoncourt F, et al. Document-level event argument extraction via optimal transport// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL-findings). Dublin, Ireland, 2022: 1648-1658

[34] Wang X, Gui L, He Y. An AMR-based link prediction approach for document-level event argument extraction// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 12876-12889

[35] Liu W, Cheng S, Zeng D, Qu H. Enhancing document-level event argument extraction with contextual clues and role

relevance//Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL-findings). Toronto, Canada, 2023: 12908-12922

[36] Wei K, Xian S, Zhang Z, et al. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2021: 4672-4682

[37] Du X, Li S, Ji H. Dynamic global memory for document-level argument extraction//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, 2022: 5264-5275

[38] Zeng Q, Zhan Q, Ji H. EA^{2}E: Improving consistency with event awareness for document-level argument extraction// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-findings). Seattle, USA, 2022: 2649-2655

[39] Lin J, Chen Q, Zhou J, et al. CUP: Curriculum learning based prompt tuning for implicit event argument extraction// Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). Vienna, Austria, 2022: 4245-4251

[40] Ren Y, Cao Y, Guo P, et al. Retrieve-and-Sample: Document-level event argument extraction via hybrid retrieval augmentation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 293-306

[41] Ma M D, Taylor A K, Wang W, et al. DICE: Data-efficient clinical event extraction with generative models//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 15898-15917

[42] VanNguyen C, Man H, Nguyen T H, et al. Contextualized soft prompts for extraction of event arguments// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-findings). Toronto, Canada, 2023: 4352-4361

[43] Li H, Cao Y N, Ren Y B, et al. Intra-Event and Inter-Event dependency-aware graph network for event argument extraction// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-findings). Singapore, 2023: 6362-6372

[44] Zhou J, Zhang Q, Chen Q, et al. A multi-format transfer learning model for event argument extraction via variational information bottleneck //Proceedings of the 29th International Conference on Computational Linguistics (COLING). Gyeongju, Republic of Korea, 2022: 1990-2000

[45] Sainz O, Gonzalez-Dios I, Lacalle O L, et al. Textual entailment for event argument extraction: Zero-and few-shot with multi-source learning//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-findings). Seattle,USA, 2022: 2439-2455

[46] 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//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 393-409

[47] Yang X J, Lu Y J, Petzold L. Few-shot document-level event argument extraction//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 8029-8046

[48] Liu J, Liang C, Xu J. Document-level event argument extraction with a chain reasoning paradigm//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 9570-9583

[49] Virginia M. 1992. Fourth message understanding conference (muc-4)

[50] Sundheim B M. Overview of the fourth message understanding evaluation and conference//Proceedings of 4th Message Understanding Conference (MUC-4). McLean, Virgini, 1992

[51] Zhou H Z, Mao K Z, et al. Document-level event argument extraction by leveraging redundant information and closed boundary loss//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Seattle, USA, 2022: 3041-3052

[52] Tong M H, Bin X, Wang S, et al. DocEE: A large-scale and fine-grained benchmark for document-level event extraction// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Seattle, USA, 2022: 3970-3982

[53] Nguyen K H, Tannier X, et al. A dataset for open event extraction in English//Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia, 2016: 1939-1943

[54] Wang B, Du X Y, Cardie C. Probing representations for document-level event extraction// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-findings). Singapore, 2023: 12675-12683

[55] Du X Y, Cardie C. Document-level event role filler extraction using multi-granularity contextualized encoding//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, USA, 2020: 8010-8020

[56] Huang G H, Xu R X, Zeng Y, et al. An iteratively parallel generation method with the pre-filling strategy for document-level event extraction//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP. Singapore, 2023: 10834-10852

[57] Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015

[58] Bengio Y, Louradour J, et al. Curriculum learning//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 41-48

[59] Sainz O, Lacalle O L, Labaka G, et al. Label verbalization and

entailment for effective zero and fewshot relation extraction// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic, 2021:1199-1212

HU Rong, Ph. D. candidate, research assistant. Her current research interests include information extraction, natural language processing and big data analysis.

WAN Chang-Xuan, Ph. D., professor. His current research interests include Web data management, sentiment analysis, data mining and information retrieval.

WAN Qi-Zhi, P. D., lecturer. His current research interests include information extraction, natural language processing, and data mining.

LIU De-Xi, Ph. D., professor. His current research interests include natural language processing, information retrieval, and Web data management.

LIU Xi-Ping, Ph. D., professor. His current research interests include information retrieval and data mining.

Background

Document-level Event extraction is an important and challenging task in natural language processing. The existing document-level event extraction is mainly divided into two directions: one direction is the complete document level event extraction(DEE), that is, judging what types of events exist in a given document, identifying all events under each event type, and extracting arguments of corresponding roles; Another direction is document-level event argument extraction (DocEAE); that is, given the event types and event triggers contained in each document, extracting event arguments of the corresponding roles triggered by each triggers. The goals and tasks of the two directions are different, and the task steps are not close to the same. The corresponding two types of data sets also have different characteristics and focus on causing different research problems. Based on the open data sets of two directions, the researchers are committed to solving the research problems unique to each direction, and have made different research progress.

Existing little surveys on document-level event extraction share the following limitations: (1) insufficient understanding of the research issues, (2) classifying existing researches based on specific techniques or task steps, without in-depth analysis of the relationship and difference between these researches, the issues these researches trying to solve, and (3) a brief introduction to common datasets which fails to properly understand their characteristics and the task challenges.

Regarding event extraction, our previous results, including (1) Chinese Financial Event Extraction Base on Syntactic and Semantic Dependency Parsing, (2) A Multi-Channel Hierarchical Graph Attention Network for Open Event Extraction, (3) Token-Event-Role Structure-based Multi-Channel Document-Level Event Extraction, (4) CFERE: Multi-type Chinese financial event relation extraction, (5) Joint Document-Level Event Extraction via Token-Token Bidirectional Event Completed Graph, and (6) Dependency Structure-Enhanced Graph Attention Networks for Event Detection published in *Chinese Journal of Computers*, *ACM Transactions on Information Systems* (TOIS), *ACM Transactions on Information Systems* (TOIS), *Information Sciences*, *The 61stAnnual Meeting of the Association for Computational Linguistics* (ACL), and *The 38th Annual AAAI Conference on Artificial Intelligence* (AAAI), which are the leading journals.

The research is partially supported by National Natural Science Foundation of China (Nos. 62272205, 62272206, 62076112, 62462034), the Science & Technology Project of the Department of Education of Jiangxi Province (No. GJJ210531, GJJ2400411), the Jiangxi Graduate Innovation Special Fund Project (YC2023-B188), the Natural Science Foundation of Jiangxi Province (Nos. 20242BAB25119, 20212ACB202002, 20232ACB202008), and the Funding Program for Academic and Technology Leaders in Major Disciplines of Jiangxi Province (No. 20213BCJL22041).

In this paper, we discussed the research tasks, task steps and developments first. Then, we summarize the characteristics, task challenges of each dataset. Subsequently, we analyze the main research methods under these datasets and display the progress graphically. Finally, we conclude the future direction of document-level event extraction task.