基于句法语义依存分析的中文金融事件抽取

万齐智[1),3)] 万常选[1),3)] 胡 蓉[2),3)] 刘德喜[1),3)]

[1)](江西财经大学信息管理学院 南昌 330032)

[2)](江西财经大学软件与物联网工程学院 南昌 330032)

[3)](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 事件抽取在自然语言处理应用中扮演着重要的角色,如股票市场趋势预测.传统事件抽取较为关注触发词和论元所属类型的正确性,较少地结合应用需求去分析研究事件抽取效果及使用价值.在财经领域,事件作用对象及动作是关注的重点.因此,本文聚焦于金融事件,抽取三元组事件 $ET(Sub,Pred,Obj)$.在中文财经新闻中,存在大量事件嵌套和成分共享等现象,致使易出现事件漏抽和事件成分缺失的情况.为了解决这些问题,本文建立一个句法和语义依存分析相结合的中文事件抽取框架,归纳了 4 种常见缺省结构,并设计相应的补全规则.首先,基于句法依存树,分析动词词法和句法结构,建立核心动词链,使得每个核心动词对应一个事件,解决事件漏抽问题.然后,在句法依存树的基础上添加语义依存关系,建立事件间语义关联,得到句法语义依存分析(Syntactic Semantic Dependency Parsing,SSDP)树.第三,调整 SSDP 树,优化句法结构,形成 SSDP 图,使得同等句法结构的词结点处于相同层级,为后续事件抽取提供途径.第四,归纳 4 种常见缺省结构,设计相应补全规则,解决事件成分缺失问题.最后,在中文财经新闻标题和 CoNLL2009 中文语料上进行详细的实验测试,实验结果表明该方法是有效的.

关键词 中文事件抽取;核心动词链;句法语义依存分析图;事件语义关联;缺省补全

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2021.00508

Chinese Financial Event Extraction Base on Syntactic and Semantic Dependency Parsing

WAN Qi-Zhi[1),3)] WAN Chang-Xuan[1),3)] HU Rong[2),3)] LIU De-Xi[1),3)]

[1)](*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330032)

[2)](*School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang* 330032)

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

Abstract As a sub-task of information extraction, event extraction plays an important role in nature language process applications, such as stock market trend forecast, which can provide strong clues for events users, e.g. investors, managers and government, to analyze the market and make decisions. At present, most of the studies about event extraction pay more attention to the type correctness of triggers and arguments, and not consider the effect and value of event extraction based on application requirements. We call this type of event extraction traditional event extraction. The event types and standards in traditional event extraction are derived from ACE2005 containing 8 categories and 33 sub-categories, KBP2015 and ERE, et al. However, there are some limitations in application of them to event extraction in specific financial domain. For example, there is not the *overweight* event type in ACE2005, which is a special behavior in the financial

收稿日期:2019-09-10;在线发布日期:2020-03-01.本课题得到国家自然科学基金项目(61972184,61562032,61762042)、江西省教育厅科学技术研究项目(GJJ180198,GJJ180252)资助.**万齐智**,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究方向为信息抽取、自然语言处理、数据挖掘.E-mail:wanqizhi1006@163.com.**万常选**(通信作者),博士,教授,博士生导师,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、信息检索.E-mail:wanchangxuan@263.net.**胡 蓉**,硕士,助理研究员,主要研究方向为信息抽取、自然语言处理、大数据分析.**刘德喜**,博士,教授,博士生导师,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.E-mail:dexi.liu@163.com.

domain. In this paper, we focus on the financial news and extract open events without types. In the field of finance and economics, most event users are more concerned with the objects and actions that events affect. Therefore, combined with the application requirement, we propose to extract the financial event $ET(Sub, Pred, Obj)$, where Sub, $Pred$ and Obj represent subject, predicate and object respectively. However, Chinese financial news generally suffers from the event nesting and component default problem, which result in event omission and key element missing of events. To tackle this issue, with the expression habits and characteristics of Chinese linguistics, we build a Chinese event extraction framework based on syntactic and semantic dependency parsing. Then summarize four common default structures and design corresponding completion rules. In particular, at the beginning of this paper, we summarize four prominent phenomena in the extraction of events from the headlines of financial news, and explore the cause of these problems, no in-depth analyzing the relevance of syntactic and semantic structure or lack of it. After that, we employ the syntactic dependency parsing tree and lexical structure, and propose the core verb chains, which make sure that each core verb corresponds to an event solving event leakage problem. Thirdly, we add semantic dependency relation between events on the basis of syntactic dependency tree, which is called Syntactic Semantic Dependency Parsing (SSDP) tree. In order to better separate the detected events and their properties, we adjust and optimize SSDP tree to form the SSDP graph, where the word nodes of the same syntactic structure are at the same level, providing a way for subsequent event extraction. Fourthly, with the division of default structure in linguistic, we summarize four common default structures and propose ten corresponding completion rules to solve the problem of component default. Meanwhile, the whole Chinese event extraction algorithm based SSDP graph is shown at the end of the section. Finally, this paper depicts a detailed experimental situation. The experimental dataset, labeling standard and evaluation index are given. Subsequently, the method in this paper is verified on two datasets, financial news titles and common field news titles. At the end, we conduct comprehensive benchmarks on Chinese financial news titles and CoNLL2009 Chinese Corpus. The experimental results show that the proposed methods are effective.

Keywords　Chinese event extraction; core verb chain; syntactic semantic dependency parsing graph; event semantics relevance; default complement

1　引　言

事件抽取作为信息抽取的子任务，在自然语言处理应用中扮演着较为重要的角色，如股票市场趋势预测[1-3]. 投资者、上市公司以及政府对股票市场趋势都比较感兴趣，趋势预测可为其分析市场、做出决策提供有力参考. 相关工作[1-4]利用自然语言处理技术分析了网络文本对股市趋势预测的影响，发现金融新闻报道的事件是股市趋势预测的重要依据[1]. 因此，事件抽取的内容及其质量至关重要，将直接影响股市趋势预测效果.

目前大部分事件抽取都是基于 ACE2005①(定义了事件的 8 种大类、33 种小类)、KBP2015② 和 ERE 标准[5]，这些标准及数据集应用于宏观经济预测等特定领域的事件抽取存在一定的局限性，如在标准中并未定义股票“增持”事件类型. 文献[6-7]虽针对公司新闻和中文财经领域制定了适合自身的事件类型，但都局限于较小范围内的某些特定事件. 目前对于哪些事件会影响股价走势尚未有定论，致使自定义类型的事件可能对预测作用不大，且还要求研究人员具备丰富的财经知识和经验，一定程度上增加了研究难度. 所以，本文聚焦于财经新闻，采取开放模式进行事件抽取.

财经领域较为关注事件作用对象及动作. 本文

① http://projects.ldc.upenn.edu/ace/
② https://tac.nist.gov//2015/KBP/

结合应用需求，确定抽取三元组事件 $ET(Sub, Pred, Obj)$. 其中 Sub 为主语，$Pred$ 表示谓语(事件的核心，触发整个事件发生，一般动词居多[8]，后续称为核心动词)，Obj 代表宾语，上述 3 个要素均可称为事件的属性或成分. 文献[1]虽然也研究了上述三元组事件抽取，但做了较多限制，如谓语短语需以动词开始、介词结束，主语和宾语需为处于谓语左右两侧的名词等. 这会导致较多有价值的事件因不满足条件而被舍弃，如语句 S_1“港股恒指跌 0.14%”. 其中，动词“跌”作为谓语触发事件，并未以介词结束；同时，该文献未考虑复合句中因共享成分而导致的事件成分缺失问题，使得抽取的事件不完整，一定程度上降低了事件使用价值.

中文作为话题驱动语言，为了表达的连贯性和简洁性，常省略某些语言成分，即句子存在缺省[8]. 根据中心理论[9]，主语、谓语和宾语作为句子的主要成分. 但是，主语是最有可能缺省的，其次是宾语，最后为其他位置上的词语[10-11]. 从句法结构和语义方面划分，可分为直接省略和间接省略. 如语句 S_2“英首相让步，考虑爱尔兰担保协议”为直接省略，后半句缺省主语“英首相”；语句 S_3“京东营收增速首次跌破 30%，年内市值蒸发逾 400 亿美元”属于间接省略，后半句已存在主语“市值”，但语义并不完整，缺少前半句的“京东”作为修饰. 对于直接省略，根据是否由介词引起，又可分为介词引发和直接结构省略. 如语句 S_4“中国动力飙近 21%，与中国能源达战略性合作框架”后半句因介词“与”引导，缺少部分主语“中国动力”. 中文语句表达十分灵活，缺省结构较为复杂多样化. 因此，如何抽取完整的事件是本文致力解决的一个关键问题.

新闻标题一般需要简明扼要地概括新闻内容. 财经新闻标题偏好采用动作行为的表达形式，致使语句中出现大量动词，且较多连续动词. 如“3 位创投股东拟清仓减持套现超 20 亿，博天环境一字跌停”. 其中，“清仓”、“减持”、“套现”、“超”等一系列动词描绘整个过程，可认为标识一个事件，而动词“跌停”单独触发另一个事件. 如何识别哪些动词触发事件，哪些动词作为简单的成分，即确定语句中蕴含的事件数和谓语，是本文致力解决的另一个关键问题.

针对上述两个关键问题，本文归纳了在财经新闻标题中抽取事件时较为凸显的 4 种现象：

(1) 事件漏抽. 一条新闻标题常包含多个事件，只抽取了其中部分事件.

(2) 事件成分缺失. 抽取的事件成分不全，主要由主语或宾语省略所致.

(3) 事件成分抽取错误. 抽取的事件成分信息在语义上与文本语义存在出入.

(4) 事件语义放大. 缺少限定范围，使得抽取事件语义大于原文语义或语义不明，主要因修饰语省略引起. 如语句 S_3，事件 ET_1(市值，蒸发，400 亿美元)虽已抽取了 Sub 属性，但缺乏修饰定语“京东”，使得事件 ET_1 语义放大，指向不明，缺乏使用价值.

出现上述 4 种现象，主要是因为没有深入分析句法和语义结构上的关联或是缺少关联. 其中，前两种现象属于句法结构，应探寻事件间和共用成分间的关联规则；后两种现象则侧重于语义，需要从语义角度分析其存在的关联. 因此，本文采用句法和语义依存分析相结合的方法，建立句法语义依存分析(Syntactic Semantic Dependency Parsing，SSDP)图. 同时，基于 SSDP 图，归纳常见的缺省结构，制定缺省补全规则. 首先，根据句法依存结构，设计规则，建立核心动词链. 其次，添加语义依存关系，建立 SSDP 树. 再次，基于核心动词链和语义结构，优化 SSDP 树，形成 SSDP 图. 最后，基于 SSDP 图，分析扩展事件间的语义关系，提出 4 种缺省结构，并设计相关补全规则，解决抽取事件的成分缺失问题.

本文的主要贡献包括：

(1) 建立核心动词链. 基于句法依存结构，分析动词词法及句法依存结构，提出核心动词链建立规则.

(2) 建立句法和语义依存分析相结合的 SSDP 图. 借助句法依存树，添加语义依存关系，建立包含事件间语义关联的 SSDP 树；基于核心动词和语义结构，将 SSDP 树调整为 SSDP 图，使得核心动词和同等结构成分的结点尽量处于同一层级.

(3) 归纳 4 种常见缺省结构，提出相关补全规则. 根据中文使用习惯和语料数据，归纳了 4 种常见缺省结构，并设计有效的查询补全规则.

本文第 2 节介绍相关工作，分析目前相关研究的进展及优缺点；第 3 节分析核心动词词性及句法结构，归纳核心动词链的建立规则，为探测事件提供依据；第 4 节首先探讨基于缺省补全的中文事件抽取面临的挑战，然后描述句法和语义依存分析相结合的 SSDP 图的构建方法，为补全事件缺失成分搭建查询桥梁；在第 5 节中，讨论 4 种常见缺省结构，并分析其补全规则，解决抽取事件的成分缺失问题；第 6 节介绍本文实验数据集、实验方法和实验结果，

验证本文方法的有效性；最后，第 7 节对全文进行总结，并就未来工作提出展望，为即将开展的后续研究指明方向.

2　相关工作

事件抽取作为信息抽取的子任务，在知识挖掘领域起着非常重要的作用. 近几年，事件抽取的主要研究重点是，如何利用不同的线索信息提高事件触发词或论元所属类型的正确率，较少地结合应用需求去分析研究事件抽取效果及使用价值. 我们将前者称为传统事件抽取，后者称为应用需求驱动的事件抽取.

(1) 传统事件抽取的研究进展

传统事件抽取一般分为 4 个子任务，触发词识别/分类和论元识别/分类，前者称为事件探测. 目前，无论是事件探测还是完整的事件抽取，涉及识别或抽取语句中包含事件数的研究非常少.

在事件抽取方面，文献[12]为解决新事件类型在标准数据集上识别效果不佳的问题，选择新领域数据训练模型，但因缺乏标注数据，提出一种可快速收集新事件类型训练数据的方法，并通过已有标准数据集，训练一个可在新类型上识别 Actor、Place 和 Time 等论元的模型，一定程度上解决了新类型事件的抽取问题. 文献[13]针对基于 CRF 的事件抽取联合模型的缺陷进行扩展，旨在解决事件多标签问题，但需对事件进行分类训练. 另外，借助同一大类事件下，不同子类事件间元素存在高关联性，采取多任务学习方法解决由分类训练带来的数据稀疏问题.

文献[14]提出一种动态多池化的卷积神经网络以保持多事件信息，实现语句中多事件抽取；同时可自动抽取词法级和语句级特征，缓和严重依赖 NLP 工具的现象. 文献[15]利用双向循环神经网络和人工设计的特征联合抽取事件触发词和论元. 文献[16]研究论元与论元间的句法依赖关系，为其建立依赖桥，结合双向循环神经网络方法，提高了同一事件的论元被完整抽取的概率. 文献[17]利用同一语句包含的多个事件触发词之间存在高关联性，通过引入句法依存树和基于注意力的图卷积网络，借助其他事件触发词类型信息进一步确定当前事件触发词所属类型，从而提高事件抽取效果.

在事件探测方面，文献[18]研究的问题类似于文献[17]，也是借助事件间的关联来提升事件分类的效率. 但不同的是，文献[18]指出，较多可利用的、有关联的事件位于不同语句中，只考虑单个语句中的事件，存在一定局限性. 因此，设计一个门控多级注意力机制，自动提取并动态融合句子级和文档级信息. 文献[19]分析以往研究主要针对特定领域或特定事件类型存在的局限性，提出在开放领域中探测无类型约束的事件. 随后提及由此带来的 2 个问题：① 事件无统一定义；② 无足够训练数据. 为了克服问题①，选择识别所有可能的事件. 但通过公布的语料可知，基本限于一条语句只包含一个事件，即只考虑一条语句中包含一个事件的情况.

文献[20]针对以往工作只利用一次上下文信息的情况，提出利用动态记忆网络多次使用上下文信息，提高事件触发词分类效果. 文献[21]通过生成对抗方法，解决由语义信息映射的高维特征空间中存在虚假特征干扰的问题，提高了事件探测效果.

通过对事件抽取相关研究的梳理发现，事件抽取主要集中于利用寻找的线索提高事件识别或抽取的效果，与本文研究问题还是存在一定的差别. 但是，本文获取的事件内容与语义角色标注在形式上存在一定的相似性. 语义角色标注主要标注论元与谓词之间的角色关系，属于浅层语义分析.

(2) 语义角色标注的研究进展

语义角色标注(Semantic Role Labeling，SRL)包含 4 个子任务，分别是谓词识别/消歧和论元识别/分类. 针对 SRL 的研究，基本上都是基于 CoNLL 提供的标注语料库，这些语料库大部分已标注了谓词[22-23]，所以很多研究的重点主要聚焦于论元与谓词之间的角色关系. 近些年，深度神经网络方法在 SRL 上已经取得了较好的效果[22-29]，尤其是 LSTM.

深度学习方法较少考虑句法特征，但直观上句法结构利于 SRL，为了验证这个假设，文献[24-25，30]均采用基于现有的模型，如 BiLSTM，设计嵌入句法结构的模式，使得深度学习模型可利用输入的句法结构实现 SRL. 研究表明，深度学习模型嵌入句法结构可提高 SRL 效果. 虽然句法结构能够提供一定信息，但因其对语言类型和领域外数据的鲁棒性不高，所以也存在较多的研究未利用句法结构[24-26]. 文献[31]则采取折衷方案，利用超级标签获取部分句法结构信息，提高了 SRL 效果.

除此之外，也有研究针对 SRL 基于跨度和基于依赖的 2 种标注形式进行了分析. 文献[23]指出，由于 2 种标注形式的存在，使得很多下游应用不知采取何种形式更为有利，从而提出一种统一 2 种标注

形式的端到端 SRL 模型.文献[26]分析了基于 BIO 标签的神经网络需要已标注谓词作为输入的一部分、且无法包含跨层级特征等缺点,提出一种端到端模型,用于联合预测所有谓词和论元跨度,以及它们之间的关系.

文献[22]受助于人类在处理未见过事情时借鉴相似问题处理方法的启发,提出一种不依赖句法结构的方法(BiLSTM+AMN),该方法利用训练集中语句及其标签关联记忆线索,帮助论元角色标注.

上述研究较好地推动了 SRL 研究的进展,但针对本文提出的研究问题,发现仍存在以下不足:

① 不能结合应用需求识别以事件为单位的谓词.SRL 多以动词为单位进行识别,而在语料中,语句通常包含较多具有动词词性的非谓词,导致识别的事件数远多于实际的事件数.

② 绝大多数研究未考虑论元补全,少量研究只实现了简单论元补全,即同一个论元与不同谓词间的角色关系.

③ 由于 CoNLL 提供了谓词标注,因此部分研究只考虑了识别论元与谓词之间的角色关系,并没有研究谓词识别问题.

④ 大部分研究基于英文语料,由于中文需要分词,因此借助句法结构信息的模型不能较好地适用于中文语料.

⑤ 特定领域的标注数据不足,尤其是中文标注数据,无法满足需大量标注数据的深度学习方法,使得 SRL 效果不佳.

(3)财经领域的事件抽取

对于应用需求驱动的事件抽取,以需求为导向,有针对性地抽取所需事件.

文献[7]聚焦于财经领域中事件信息分散于多个语句的现象,自定义财经领域事件类型,并提出抽取文档级事件的方案.同时,采用远程监督实现自动标注财经领域训练数据,克服特定领域标注数据集不足的问题.

文献[6]以了解公司大体情况为需求,针对公司新闻文本,分别采用 SVM 和 RNN-LSTM 方法探测自定义的 10 种不同经济事件.

文献[32]以证券和金融市场决策者需了解事件各方面的综合信息为出发点,分析了基于单个文档抽取事件的局限性,利用不同机构可能报道同一事件以及事件存在冗余信息的线索,提出在开放域新闻集群中抽取事件的无约束类型,并归纳通用的事件模式.

文献[1]首次提出采用结构化信息表示事件,将抽取的事件用于预测股价波动.文中事件定义为 4 元组 $E=(O_1,P,O_2,T)$,其中 O_1 为行动者,P 代表谓语,O_2 是目标者,T 为时间戳(主要用于对齐股票时间).该文利用开放信息抽取技术[12-13],无需事先定义事件类型和人工标注训练语料.但在抽取谓语和论元时添加了句法和词汇限制[13].

该文献存在的不足:

(1)谓语抽取的约束条件过于严苛.在新闻语料中,存在较多谓语不符合约束条件.

(2)论元识别存在一定的局限性.首先,充当论元的词不一定为名词短语,且也不一定为距离谓语最近的名词短语.

(3)没有考虑成分缺省情况.财经新闻语料存在大量的成分缺省,不完善缺失成分将会大大降低抽取事件的使用价值.

3 建立核心动词链

本节首先分析基于句法依存的核心动词句法结构,然后总结建立核心动词链的规则,最后给出建立核心动词链的算法.

3.1 核心动词词法及句法分析

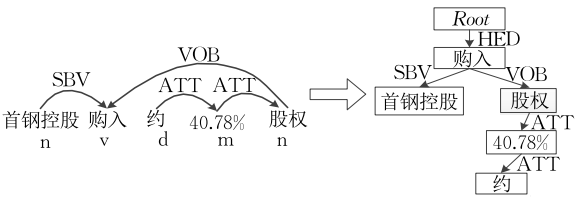
3.1.1 依存句法分析树

依存句法分析(Dependency Parsing,DP)是自然语言处理中的关键技术之一,其基本任务是确定句子的句法结构或者句子中词汇之间的依存关系[33].主要包括两方面的内容,一是确定语言的语法体系,即对语言中合法句子的语法结构给予形式化定义;二是依存句法分析技术,即根据给定的语法体系,自动推导出句子的句法结构,分析句子所包含的句法单位以及这些句法单位之间的依存关系.依存句法分析树(称为 DP 树)则将句法单位之间的依存关系以树的形式表示.本文的依存句法分析采用哈尔滨工业大学语言技术平台(Language Technology Platform,LTP)[34],LTP 共定义了 14 种依存关系,如表 1 所示.

表 1 LTP 中依存关系名及含义

标记	解释	标记	解释
SBV	主谓关系	FOB	前置宾语
VOB	动宾关系	ADV	状中结构
IOB	间宾关系	CMP	动补结构
POB	介宾关系	IS	独立结构
ATT	定中关系	DBL	兼语
COO	并列关系	LAD	左附加关系
HED	核心关系	RAD	右附加关系

语句 S_5“首钢控股购入约 40.78%股权”，其依存句法分析结果如图 1(a)所示，DP 树如图 1(b)所示. 其中，图 1(a)中的 n，v，d，m 分别代表名词、动词、副词和数词；在图 1(b)中，“购入”与父结点 *Root* 关系为 HED，是本语句核心词，结点之间的边代表句法依存关系. 由于核心词的词性通常为动词，所以也将核心词称之为核心动词.



(a) 依存句法分析　　(b) DP树

图 1　语句 S_7 的依存句法分析和 DP 树

3.1.2　核心动词句法分析

中文语句表达常采用并列句和复合句，在财经新闻标题中更为突出. 财经新闻标题一般采取正、副标题形式，副标题对正标题起补充说明，更为详细地阐述正标题的内容. 下面以一个简单的例子来说明核心动词的句法结构.

例 1.　“果源价格分化严重 苹果期货增仓上涨”. 副标题“苹果期货增仓上涨”对正标题“果源价格分化严重”中的具体果源(苹果)价格情况进行描述. 该语句的 DP 树如图 2 所示(为了简化，树中省略了标点符号依存关系，后续的 DP 树中也全部省略).

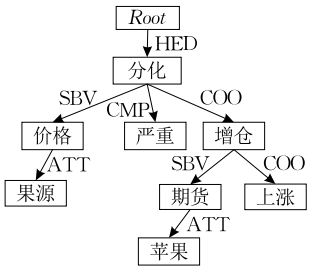


图 2　例 1 的 DP 树

对于例 1，共包含 3 个事件 ET_2(果源价格，分化，)、ET_3(苹果期货，增仓，)和 ET_4(苹果期货，上涨，). 在图 2 的 DP 树中，只存在一个语句核心动词“分化”. 如果每个核心动词触发一个事件，则导致 ET_3 和 ET_4 事件漏抽.“增仓”作为 ET_3 的谓词，是 ET_2 谓词“分化”的孩子结点，且依存关系为 COO，而 ET_4 的谓词“上涨”又作为 ET_3 的谓词“增仓”的孩子结点.

从图 2 中分析可知，LTP 针对一条语句，只会给出一个核心动词，但我们可以根据其依存关系和词性，参照每个核心动词对应一个事件的标准，划分出多个核心动词，从而形成一条核心动词链. 如何建立核心动词链将在下一小节介绍.

3.2　核心动词链建立

通过对大量的语料和上节的 DP 树进行分析，发现 3 条线索：①事件的谓词一般由动词充当；②一个语句中事件间的谓词在 DP 树中为父子结点，且保持连续，如“分化”→“增仓”→“上涨”；③一个语句中事件谓词之间父子结点的边为 COO. 另外，在语言学中，并列的词语在句法结构上应该拥有相同地位或性质，即它们之间应采用并列符号进行关联，如 LTP 采用的 COO. 通过对一个语句中动词并列符号的识别，可较好地分离语句中包含的若干事件.

因此，根据上述线索，提出一个事件分离方法(核心动词链的建立规则)，具体规则如下：

规则 1.　如果 LTP 给出的语句核心词是动词，则默认属于核心动词链中；否则考虑其满足 COO 关系的孩子结点，直到找到动词为止.

规则 2.　加入的结点是与核心动词链中结点构成 COO 关系的动词结点，且确保添加的动词从语句核心词开始一直保持 COO 关系的连续性，一旦中断则不再考虑后续动词.

规则 3.　如果 LTP 给出的语句核心词是非动词，且其孩子中没有满足 COO 关系的动词结点，则该语句不生成核心动词链.

上述规则彼此间具有一定的逻辑依赖性. 规则 1 是构建核心动词链的起点；规则 2 是对扩充核心动词链的新结点进行词性、连续性和句法依存关系判断，其中原始的连续性来源于规则 1；而规则 3 是不满足规则 1 的情况(即核心动词链为空).

添加至核心动词链中的每个动词结点需满足以上所有规则，本文以下部分所说的核心动词均指处于核心动词链中的结点，所以链中结点数即为语句蕴含的事件数. 因此，利用核心动词链方法，可解决本文提出的如何确定语句中蕴含事件数的关键问题.

针对图 2 中“增仓”和“上涨”结点，按照规则应全部添加至核心动词链中，但它们反映同一事件(即主语和宾语相同)的不同情况. 为了避免将一个事件拆分成多个事件而降低事件信息的连贯性和完整性，本文做了如下优化：对于语句中位置连续的核心动词(如果核心动词之间只包含副词，也认为连续)，则将所有核心动词合并为一个整体，表示一系列连贯动作，如例 1 中事件 ET_3 与 ET_4 合并为事件 ET_5(苹果期货，[增仓，上涨]，).

综合核心动词链的建立及优化规则，可得到核

心动词链的建立算法，如算法 1 所示.

算法 1. *CoreVerbChain*(*CVC*,*curNode*,*DPtree*).

输入：核心动词链 *CVC*，当前核心动词结点 *curNode*，语句 DP 树 *DPtree*

输出：添加了新发现核心动词的核心动词链 *CVC*

```
FOR(cnode∈CNS) //CNS 为 curNode 的孩子结点集合
  IF (cnode.postag 为动词且 cnode.dp 关系为 COO)
    //postag 为结点词性，dp 为结点的依存句法关系
    IF (cnode 与 curNode 在原句中相邻或中间只包含副词) //有连续核心动词
      IF (CVC 为空) //处理初始 curNode 为非动词，
        //且 CVC 为空无法合并连续核心动词的情况
        将 cnode 加入 CVC；
      ELSE
        将 cnode 添加至 CVC 中的 curNode 列表中；
        //合并连续核心动词
      END IF
    ELSE  //无连续核心动词
      将 cnode 加入 CVC；
      //将满足规则的核心动词添加至核心动词链
    END IF
    CoreVerbChain(CVC,cnode,DPtree)； //递归查找
  END IF
END FOR
```

4 SSDP 图

本节讨论 SSDP 图构建. 首先分析解决本文所提问题面临的挑战，然后介绍 SSDP 树的建立过程，最后描述 SSDP 树转变为 SSDP 图的过程.

4.1 基于缺省补全的中文事件抽取的挑战

目前，随着机器学习和深度学习相关技术的飞速发展，大量的方法用于解决事件抽取问题，且取得了较好的效果，但是这类方法需要大量人工标注数据作为训练集. 对于中文财经新闻领域，人工标注的数据十分匮乏，较大地影响了上述方法的抽取效果. 而且，本文采取开放模式抽取开放性事件（无具体类型的事件），无任何标准可用于触发词和论元标注，一定程度上又增大了人工标注的难度. 因此，针对本文提出的问题，建议选取规则匹配方法.

依存句法结构蕴含着丰富的信息，无论是深度学习还是规则匹配方法，均将其作为一条重要线索. 针对本文的研究问题，句法依存关系可以用于识别结构上的成分缺省，从而启动成分补全.

然而，仅仅采用依存句法分析方法，无法完全解决上述缺省补全问题. 一方面，一条语句可采用不同的表达形式，致使句法结构多样化，增加了补全复杂度；另一方面，语句的表达存在时序性（即事件之间具有先后顺序）和一定的语义关系（如因果关系、转折关系等），而且事件缺省的成分常包含于该事件之前的其它事件中，故补全缺省成分的前提是需要建立事件间的语义关联.

4.2 基于句法语义依存分析的 SSDP 树构建

针对汉语言中的缺省，研究成果并不多，且定义及范围没有统一的标准[11]. 黎锦熙[35]认为经常出现的省略包括对话省、自述省和承前省；吕叔湘[36]将缺省分为当前省、承上省和概括省；王力[37]则分为承说省和习惯省. 随着汉语法学中“三个平面”理论（语法，语义，语用）的提出，语法学者对缺省从认知角度有了如下三种基本认识[38].

（1）句法结构上界定. 指结构中必不可少的成分没有出现的句法结构省略.

（2）语义结构上界定. 指应该说出的意思没有说出来的语义省略.

（3）语用交际界定. 指因语言环境需要的语用省略. 其中，语言环境涉及较为广泛，可以是社会文化背景、语言上下文或交际的现场情景.

新闻标题较为短小、独立，语言上下文中的语用省略偏少，因此本文依据上述缺省结构的界定，提出一种句法与语义分析相结合的事件抽取方法，称之为句法语义依存分析（SSDP）方法.

语义依存分析（Semantic Dependency Parsing，SDP），用于刻画词汇间语义依存关系，与语义角色标注存在一定的关联. SRL 只关注句子谓词与其主要论元之间的关系，而 SDP 不仅关注谓词与论元，还关注谓词与谓词、论元与论元、论元内部的语义关系，对句子语义信息的刻画更加完整全面. SDP 属于深层语义分析，不仅可为我们调整 DP 树中部分结点结构提供语义分析，还可为我们建立事件间关联提供途径.

例 1 的 SDP 树如图 3 所示. 其中，Exp、Host、eCoo、Cons、Feat 和 Mann 分别表示当事关系、宿主角色、并列关系、结局角色、描写角色和方式角色.

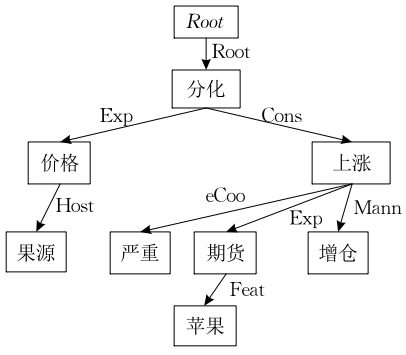


图 3 例 1 的 SDP 树

考虑到目前可用 SDP 工具的正确性一般(图 3 中“严重”结点错误地依存于“上涨”结点)，且与 DP 在结构上有时会存在冲突(作用对象不一致)，因此，本文只利用 SDP 建立核心动词间关联. 为进一步降低冲突的可能性，建立过程需按如下方式进行.

首先，对 DP 树进行剪枝，只保留主语、核心动词和宾语等主干成分，减少 DP 树中的结点数量；其次，对剪枝后的 DP 树进行语义依存分析，获取核心动词间语义关联；最后，将获取的语义关联添加至原始 DP 树中.

另外，核心动词一定程度上代表事件，事件之间的语义依存关系采用 *eXX*(如 eCoo，eSucc 和 ePurp)表示，因此针对核心动词间非 *eXX* 关系的情况，在依赖的孩子结点中查询获取，并作为核心动词间语义关联. 例如，图 3 中“上涨”与“分化”结点之间的关系为 Cons，需在其孩子结点中获取 eCoo 关系.

本文针对 DP 树中结点关系，设计事件关系二元组 $ERT(dp, sdp)$. 其中，dp 为句法依存关系，sdp 表示语义依存关系. 将添加了语义依存关系的 DP 树称为 SSDP 树，其构建算法如算法 2 所示.

算法 2.　*SSDPtreeBuild*(*DPtree*, *CVC*).

输入：语句 DP 树 *DPtree*，核心动词链 *CVC*

输出：句法语义依存分析树 *SSDPtree*

FOR(*coreVerb*∈*CVC*)

　获取 *coreVerb* 在 *DPtree* 中对应结点 *coreVerbNode*；

　获取 *coreVerbNode* 的主谓宾主干结构 *cvnMain*；

　将 *cvnMain* 按原词顺序组合形成主干语句 *senMain*；

　通过 SDP 工具获取 *senMain* 的语义依存关系 *senSdp*；

　coreVerbNode.sdp=*senSdp*[*coreVerb*]；

　//修改 *DPtree* 中核心动词结点语义依存关系

END FOR

SSDPtree=*DPtree*；

RETURN *SSDPtree*；

基于图 2 所示的 DP 树，将图 3 中核心动词“分化”、“上涨”间的语义依存关系 eCoo 添加至 DP 树中，构建的 SSDP 树如图 4 所示，同时进行了核心动词合并. 其中，“分化”和“[增仓，上涨]”之间通过 eCoo 连接，表示事件间存在并列关联，但在句法结构上仍为父子关系. 下节将介绍如何将 SSDP 树调整为 SSDP 图.

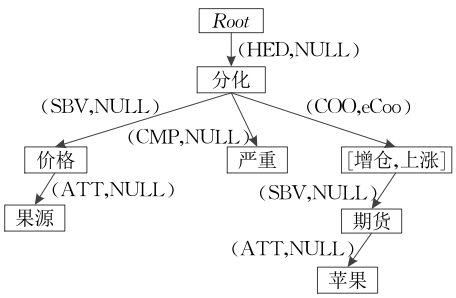


图 4　图 2 添加 *sdp* 且合并核心动词后的 SSDP 树

4.3　基于核心动词和语义结构的 SSDP 树调整

同一条语句中，每个事件的发生虽然存在前后顺序，但它们在句法结构上(包括每个事件的核心动词、主语及宾语等)应处于相同地位，这样不仅使得句子句法结构一目了然，还有利于事件的确定和 *ET* 元组中成分的抽取. 因此，本节针对 SSDP 树做了一定的优化和调整，剪除无效路径，降低树的高度，使得调整后的 SSDP 树更趋于扁平化，缩短搜索路径. 因调整后的 SSDP 树已不符合树的定义，故将其称之为 SSDP 图. 具体调整方法如下：

(1) 核心动词调整. 提升处于核心动词链中的每个核心动词结点层级，使得调整后的 SSDP 图中所有核心动词结点与核心根结点(*Root* 结点的直接孩子结点，未调整前只有语句核心词称为核心根结点)具有相同层级，即调整为 *Root* 结点的直接孩子结点. 如图 5 所示，将“[增仓，上涨]”结点调整为 *Root* 结点的直接孩子，使得与“分化”结点处于同级，但其原始关系仍保留，并采用有向虚线进行连接，方向代表事件的时序性.

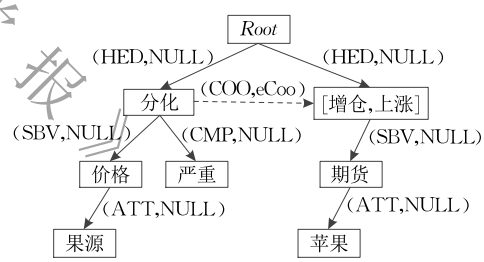
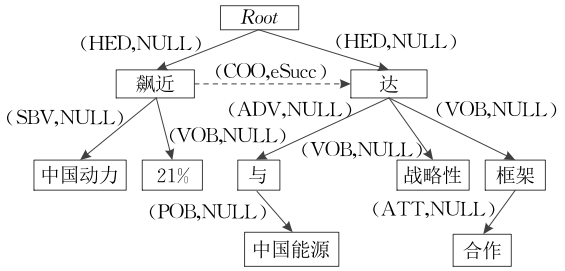


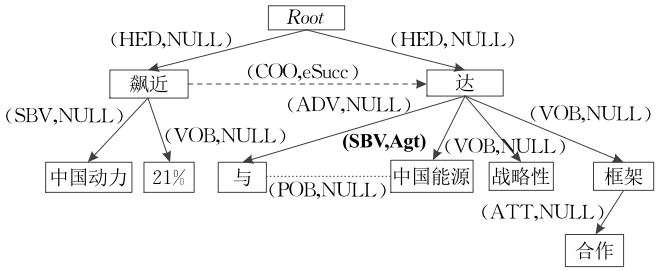
图 5　图 4 经核心动词调整后的 SSDP 图

核心动词调整是将 SSDP 树调整为 SSDP 图的关键，不仅有利于事件的划分，而且一定程度上间接促使了同等成分的结点也处于相同层级. 如图 5 中具有 SBV 关系的“期货”和“价格”结点，ATT 关系的“果源”和“苹果”结点均处于相同层级.

(2) 介词结构调整. 提升介词引导的充当主语或宾语的结点层级，使其作为对应核心根结点的直接孩子结点. 图 6(a)展示了语句 S_4 的 SSDP 树经核心动词调整后得到的 SSDP 图，其中，eSucc 表示顺承关系.“中国能源”语义上为“达”的主语，但结构上是“与”的直接孩子；图 6(b)在图 6(a)的基础上，对介词结构进行调整，将“中国能源”调整为“达”的直接孩子，句法依存关系从 POB 调整为 SBV，同时保留原始依存关系(POB，NULL)，采用无向虚线连



(a) 经核心动词调整后的SSDP图

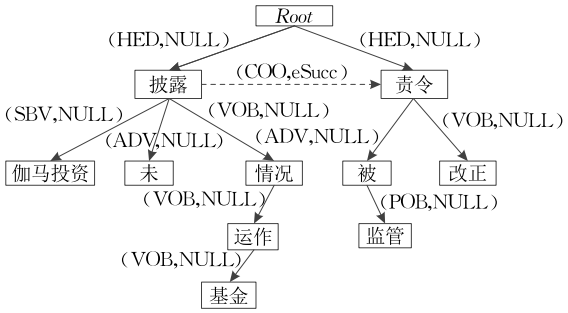


(b) 经介词结构调整后的SSDP图

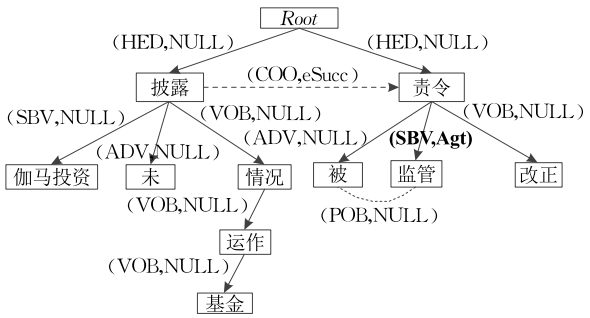
图 6 语句 S_4 的 SSDP 树经调整后的 SSDP 图

接. 其中,Agt 表示施事关系.

(3) 被动语态调整. 提升被动语句对应结点层级,修改句法依存关系. 如语句 S_6"伽马投资未披露基金运作情况,被监管责令改正"的 SSDP 树经核心动词调整和被动语态调整后的 SSDP 图,分别如图 7(a)和图 7(b)所示.



(a) 经核心动词调整后的SSDP图



(b) 经被动语态调整后的SSDP图

图 7 语句 S_6 的 SSDP 树经调整后的 SSDP 图

其中,"被"结点的直接孩子结点"监管"调整为"责令"的直接孩子结点,且添加其对应依存关系(SBV,Agt).

此处,"被"字属于特殊介词,虽然调整过程与介词结构相似,但其缺省结构补全规则存在区别. 被动语态需对调主语和宾语,且"被"字在调整后的图结构中无意义. 而图 6(b)中介词"与"起并列连接作用,在成分补全时应与左右成分一同纳入.

基于上述调整规则,可得到将 SSDP 树调整为 SSDP 图的算法,如算法 3 所示.

算法 3. *SSDPtreeAdjust*(*SSDPtree*,*CVC*).

输入:语句 SSDP 树 *SSDPtree*,核心动词链 *CVC*

输出:语句 SSDP 图 *SSDPgraph*

FOR(*coreVerb*∈*CVC*) //*CVC* 中每个核心动词 *coreVerb*
  获取 *coreVerb* 在 *SSDPtree* 中对应结点 *coreVerbNode*;
  建立由 *Root* 指向 *coreVerbNode* 之间的关联;
  //提升 *coreVerbNode* 结点的层级,但保留核心动词
  //结点之间的原始关系
  FOR(*node*∈*CCVNS*)
    //*CCVNS* 为 *coreVerbNode* 的孩子结点集合
    IF (*node* 为被动语态词) //被动语态调整
      获取语句的语义依存关系 *senSdp*;
      FOR(*cnode*∈*CNS*) //*CNS* 为 *node* 的孩子结点集合
        IF (*cnode* 为右孩子结点) //存在主语
          建立由 *coreVerbNode* 指向 *cnode* 的边;
          cnode.dp=SBV;
          cnode.sdp=*senSdp*[*cnode*];
        ELSE IF (*cnode* 为左孩子结点) //存在宾语
          建立由 *coreVerbNode* 指向 *cnode* 的边;
          cnode.dp=VOB;
          cnode.sdp=*senSdp*[*cnode*];
        END IF
      END FOR
    ELSE IF (*node* 为介词)
      获取语句的语义依存关系 *senSdp*;
      FOR(*cnode*∈*CNS*) //*CNS* 为 *node* 的孩子结点集合
        IF (*cnode* 与 *coreVerbNode* 的语义关系为主谓关系) //介词引发的主谓关系调整
          建立由 *coreVerbNode* 指向 *cnode* 的边;
          cnode.dp=SBV;
          cnode.sdp=*senSdp*[*cnode*];
        ELSE IF (*cnode* 与 *coreVerbNode* 的语义关系为动宾关系) //介词引发的动宾关系调整
          建立由 *coreVerbNode* 指向 *cnode* 的边;
          cnode.dp=VOB;
          cnode.sdp=*senSdp*[*cnode*];
        END IF
      END FOR
    END IF

END FOR
END FOR
SSDPgraph=*SSDPtree*;
RETURN *SSDPgraph*;

综上所述，SSDP 图的构建过程主要包含 3 步，如算法 4 所示. 第 1 步，核心动词链的建立，如算法 1 所示；第 2 步，SSDP 树的生成，如算法 2 所示；第 3 步，SSDP 树的调整，如算法 3 所示.

算法 4.　SSDP 图构建.

输入：语句 *sen*
输出：语句 SSDP 图 *SSDPgraph*
CVC=∅；//将核心动词链 *CVC* 置为空
利用 LTP 工具获取 *sen* 的依存句法分析结果 *DPresult*；
根据 *DPresult* 生成 *sen* 的 *DPtree*；
Root=*DPtree* 的根结点；
HEDnode=*Root* 的孩子结点；//只有一个孩子结点
IF(*HEDnode* 词性为动词)
　将 *HEDnode* 加入 *CVC*；
EDN IF
CoreVerbChain(*CVC*,*HEDnode*,*DPtree*)；
//*CVC* 返回满足核心动词链建立规则的核心动词
IF(*CVC* 不为空)
　SSDPtree=*SSDPtreeBuild*(*DPtree*,*CVC*)；
　SSDPgraph=*SSDPtreeAdjust*(*SSDPtree*,*CVC*)；
END IF

5　缺省结构及成分补全

本节先介绍 4 种常见缺省结构以及缺省补全规则，然后描述基于 SSDP 图的中文事件抽取算法.

5.1　缺省结构

目前关于缺省分类的划分未有统一标准，较多文献基于中文宾州树库(Chinese TreeBank，CTB)[39]和 Ontonotes 3.0 等语料库划分的缺省类别进行研究，主要包含 6 类缺省，如表 2 所示. 其中，NONE-*T*、NONE-*PRO*和 NONE-*pro*占比最大[11].

表 2　CTB 及 Ontonotes 3.0 中缺省分类

类别	描述
NONE-*T*	缺省为主题或从句实施者
NONE-*	缺省在“把”字句、“被”字句
NONE-*PRO*	从句中缺省明显主语
NONE-*pro*	缺省的为主语或宾语
NONE-*RNR*	发生预指的缺省形式
NONE-*?*	其他类型

根据上述分类规则并结合新闻语料分析，本文将事件成分缺省主要分成以下 4 种结构.

(1) 直接成分缺省. 根据缺省成分的复杂性，可分为简单缺省和组合缺省.

① 简单缺省. 缺省成分结构简单，可单独作为其它事件的某个成分(如主语). 语句 S_2 中简单缺省结构的 SSDP 图如图 8 所示. 其中，ePurp 代表目的关系，每个事件用虚线框标识，事件 ET_6(英首相，让步，)中简单主语成分“英首相”作为 ET_7(，考虑，爱尔兰担保协议)事件的主语，采用点横相间的有向虚线连接，表示其层级关系，并添加依存关系(SBV，Agt).

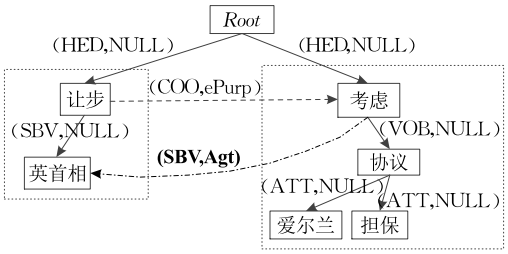


图 8　语句 S_2 中简单缺省结构的 SSDP 图

② 组合缺省. 某个组合整体作为其它事件的某个成分. 语句 S_7“油价再遭痛击，拖累期市”中组合缺省结构的 SSDP 图如图 9 所示. 其中，事件 ET_8(油价，遭，痛击)整体作为事件 ET_9(，拖累，期市)中“拖累”缺失的主语，添加 ET_8 与“拖累”结点的依存关系(SBV，Agt).

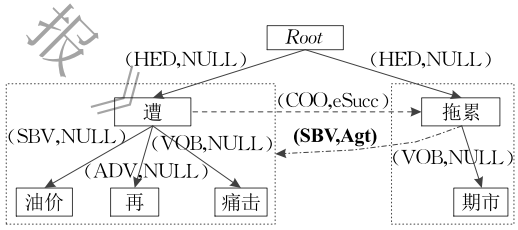


图 9　语句 S_7 中组合缺省结构的 SSDP 图

(2) 介词引发缺省. 由介词引发的部分成分缺失. 语句 S_4 中介词引发缺省结构的 SSDP 图如图 10 所示. 其中，介词“与”引导关联“中国动力”和“中国能源”. 因“中国能源”结点为 SBV 关系，故添加“中国动力”与“达”结点间的依存关系(SBV，Agt).

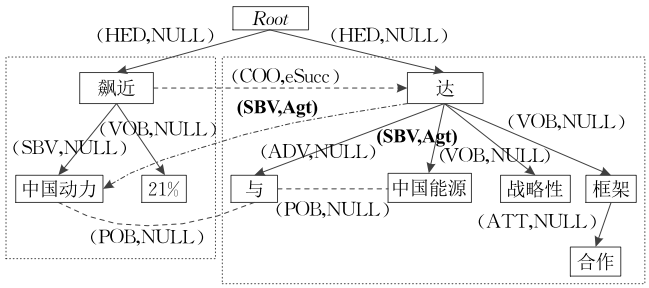


图 10　语句 S_4 中介词引发缺省结构的 SSDP 图

(3)被动语态缺省. 由“被”字等介词引发的被动语态的成分缺省.“被”字属于特殊介词,首先按照介词引发的缺省过程构建依存图,然后建立共享成分与缺省事件的宾语关系. 语句 S_6 中被动语态缺省结构的 SSDP 图如图 11 所示. 其中,Pat 为受事关系,“被”结点只起引导连接作用,既然引导的成分关系已修改,则其相关边可直接剪除. 剪枝后的依存图如图 12 所示.

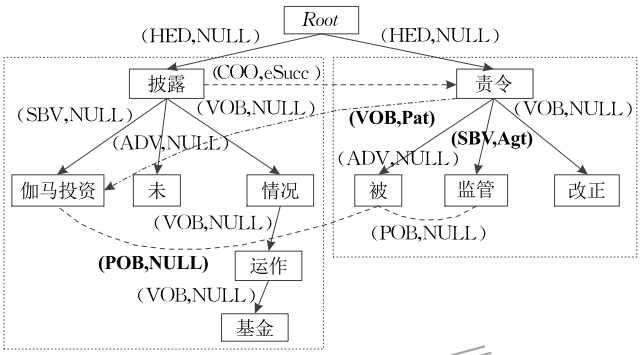


图 11 语句 S_6 中介词“被”引发缺省结构的 SSDP 图

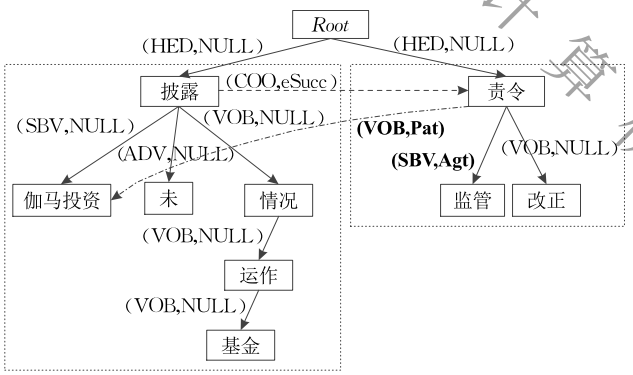


图 12 图 11 剪枝“被”后的 SSDP 图

(4)间接修饰缺省. 语义上存在修饰关系的缺省结构. 间接修饰缺省主要是反映事件间论元之间关系,充当修饰作用的一般为关联事件的主语或其主语的定语. 如语句 S_3,其定语“京东”作为“市值”的修饰;语句 S_8“深圳成立私募基金,规模为 170 亿元”,其宾语“私募基金”修饰后半句的主语“规模”. 语句 S_3 中间接修饰缺省结构的 SSDP 图如图 13 所示.

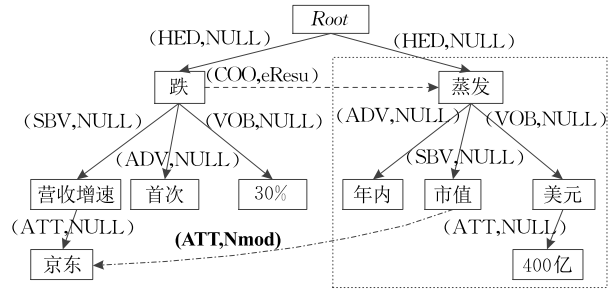


图 13 语句 S_3 中间接修饰缺省结构的 SSDP 图

示,添加了“京东”与“市值”之间的定语关系(ATT,Nmod). 其中,eResu 表示因果关系,Nmod 表示名字修饰角色.

5.2 补全规则

通过上节缺省结构分析可知,补全缺省成分可在与本事件时间最近的早期事件中查找,但并非所有缺省都需进行补全,存在语句本身无主语的情况,如“识别减值风险,严防商誉雷”. 因此,何时启动缺省补全机制、如何获取补全内容,是缺省补全的两大难点,尤其是间接缺省,无法从句法结构上进行判断,必须借助语义分析.

不同的缺省类型,其补全启动时机和规则存在差异,下面分别对主语和宾语缺省进行分析.

在语法结构中,动词分为及物和不及物两类. 宾语缺省补全需结合语句核心动词类型共同分析. 如果核心动词为不及物,其缺省属正常情况,无需启动补全机制. 当核心动词为及物动词,语句一般会跟随宾语对象,或以指代词形式给出. 真正的宾语缺省大多由介词或被动语态引发,本文前述已对这些结构做了调整. 由介词引发的缺省,可根据介词的识别,启动补全操作. 而被动语态前期已作成分关系调整,可直接识别抽取,也无需补全.

由语料分析发现,共享主语的事件间的语义依存关系集中于因果(eResu)、顺承(eSucc)和目的(ePurp)关系. 对于简单的并列句,其句子成分相对完善,通常不会共享主语,即使存在成分缺省,一般默认为事件实际无主语,不启动补全操作. 本文主语补全时机和规则主要围绕上述 3 种语义关系,我们称这些关系为引发关系. 下面针对具体情况分别讨论.

5.2.1 直接成分缺省补全

直接成分缺省是基于依存句法结构进行判断,当 SSDP 图中的核心根结点不存在 dp 为 SBV 孩子结点时,说明只是句法结构上存在主语缺失,由于存在部分实际无主语情况,所以是否需要补全缺省,还需再结合语义依存关系进行分析,从而提出了规则 4～规则 6.

规则 4. 如果由核心根结点触发的事件不存在具有语义依存关系的较早事件,则不必补全.

规则 5. 如果存在直接成分缺失,且 ERT 中 sdp 为非引发关系,若最近关联事件只存在一个主语,则在最近的关联事件中查询获取关联事件的主语,补全缺省主语,即简单缺省补全.

规则 6.　如果存在直接成分缺失，且 *ERT* 中 *sdp* 为引发关系，若最近关联事件存在多个主语，则取最近关联事件中距离当前事件最远的主语（关联事件第一个主语），补全缺省主语.

规则 4 要求，补全操作的前提必须是共享主语的句子在当前事件之前发生，且存在语义依存关系. 这符合语句表达逻辑. 因此，规则 4 是其它缺省规则执行的前提；规则 5 和规则 6 分别讨论不同 *sdp* 关系下的缺省补全情况. 规则 5 和规则 6 均是依照人们使用语言的习惯，取关联事件中位于语句最前面的主语作为缺省补全.

图 8 展示了规则 5 补全过程. 事件 ET_7 存在主语缺失，因此在最近关联事件 ET_6 中查询核心根结点的直接孩子结点，且 *dp* 为 SBV. 针对规则 6，我们通过一个示例进行说明.

例 2.　“上航飞东京一航班因机械故障返航，已另调配飞机”的 SSDP 图如图 14 所示. 其中，“调配”事件缺少主语，其关联事件“返航”存在多个主语，因此取最远主语“上航”作为“调配”事件主语.

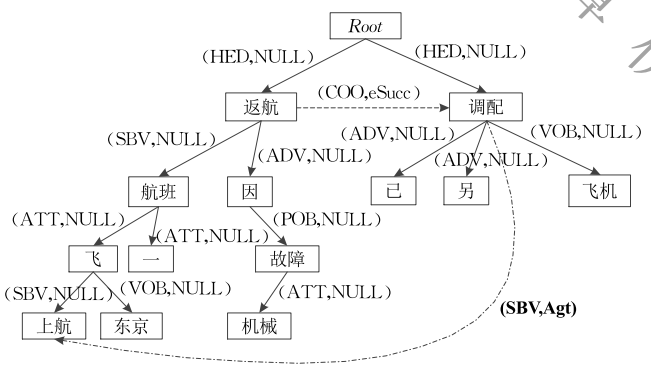


图 14　例 2 的 SSDP 图

在财经领域中，作为共享主语，一般以细分的名词居多，如公司、股票的简称、机构团体等专有名词等. 因此，当存在直接成分缺失，*ERT* 中 *sdp* 为引发关系，且最近关联事件只存在一个主语时，则还需借助该主语的词性细分缺省补全情况.

记 *POL*＝{ni，nz，nh，j} 为词性集，其中 ni、nz、nh 和 j 分别表示机构团体、专有名词、人名和简称. *POL* 为简单的词性集合，无领域特性，在规则判断过程中无需复杂的计算，直接词性对比即可.

规则 7.　当最近关联事件的主语词性不属于 *POL*，且主语存在定语时，则取主语第一个定语补全缺省主语.

规则 8.　当最近关联事件的主语词性不属于 *POL*，且主语不存在定语，则取关联事件整体补全缺省主语，即组合缺失补全.

规则 9.　当最近关联事件的主语词性为名词或属于 *POL*，则直接取关联事件主语补全缺省主语.

规则 7～规则 9，对外是规则 6 的互补形式，讨论 *sdp* 为引发关系但主语唯一的情景；对内则分析关联事件主语词性. 财经领域标题常描述同一个主体的不同方面情况，当关联事件主语词性属于 *POL*，则该主语作为缺省补全成分的概率较大，如例 4；若不属于 *POL* 且存在定语时，定语常为专有词汇，从而共享此定语，即缺省定语，如例 3 所示.

例 3.　“自主品牌车市寒冬如何活下去，不少沦为车展背景板”包含事件 ET_{10}（，沦为，车展背景板）. 其中，主语“寒冬”不属于 *POL*，但其存在 ATT “自主品牌”. 当事件 ET_{10} 补全主语时，根据规则 7 可获取“自主品牌”.

语句 S_7 的 SSDP 图如图 9 所示. 主语“油价”作为普通名词，且不存在定语，满足规则 8，所以事件 ET_8 整体作为事件 ET_9 的主语.

例 4.　“离岸人民币贬值，跌破 6.93 关口”包含事件 ET_{11}（，跌，6.93 关口）. 其中“离岸人民币”作为一个专有名词，符合规则 9，事件 ET_{11} 的主语补全为“离岸人民币”，其 SSDP 图如图 15 所示.

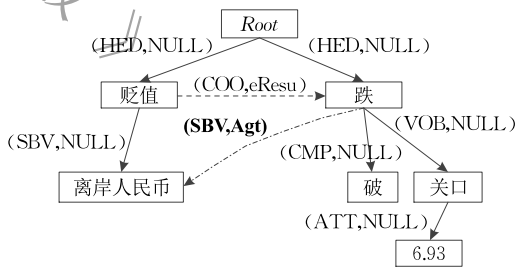


图 15　例 4 的 SSDP 图

5.2.2　介词及被动语态缺省补全

中文的介词经常连接多个名词性词语，针对前面已出现的名词，在后续搭配介词的描述中一般将其省略. 简单地从句法上分析，事件已经存在相关成分，但逻辑上深入分析发现，相关成分并不完备. 同时，对于特殊介词“被”字，既兼顾了介词特点，又包含了主语和宾语语义的反转，也需要特殊处理，以便于缺省补全. 因此，专门提出规则 10 和规则 11 进行处理.

规则 10.　如果为介词引发的成分缺省，且 *sdp* 为引发关系，则在最近关联事件中查找主语补全缺

省的主语或宾语.

规则 11. 如果由被动语态引起的成分缺省,且 sdp 为引发关系,则取最近关联事件的主语作为缺省事件的宾语成分.

图 10 和图 12 分别展示了规则 10 和规则 11 的补全情况.在图 10 中,介词"与"触发启动补全,在关联事件 ET_{12}(中国动力,飙近,21%)中查找其主语"中国动力",并将其与"与中国能源"合并作为 ET_{13} 的主语.在图 11 中,由"被"字引起的被动语态,根据规则 11,获取关联事件 ET_{14}(伽马投资,披露,基金运作情况)中主语"伽马投资"作为 ET_{15} 缺省宾语.

5.2.3 间接缺省补全

除了句法结构上直观的缺省,还存在语义上的间接缺省.间接缺省主要缺省修饰语,常由公司或机构等充当.如果缺省事件本身已经存在词性属于 POL 的名词作为主语,说明已限定范围,缺少修饰成分的可能性较小,此时无需补全;当事件存在主语,主语词性不属于 POL,且 sdp 关系为引发关系时,才进行间接缺省补全,其规则如下.

规则 12. 如果关联事件主语的词性属于 POL,且主语存在定语,同时定语的词性也属于 POL,则在最近关联事件中取距离本事件最远的定语(关联事件第一个定语),补全主语的缺省修饰部分.

规则 13. 如果关联事件主语的词性属于 POL,且主语不存在定语,则取最近关联事件中距离本事件最远的主语(关联事件第一个主语),补全主语的缺省修饰部分.

规则 12 和规则 13 一定程度上属于规则 9 的细化,且同时兼顾了规则 7 存在定语的情况.不同的是,规则 9 为句法结构不存在主语时的缺省补全,而规则 12 和规则 13 是解决存在主语的修饰缺省.另外,较多词性属于 POL 的公司词语位于描述本公司各项指标的定语中或直接代表默认指标(即充当主语),因此补全修饰缺省可主要考虑这些情景.如图 13 满足规则 12 补全条件,故获取 ATT"京东"作为事件 ET_1 主语"市值"的修饰补充.

尽管上述规则可涵盖绝大部分缺省,但还是存在遗漏情况,如共享成分为事件宾语.因难于判断需补全的成分在关联事件中扮演的角色,且该情况在语料中占比较小,因此本文暂未考虑此情形,在后期研究中将进一步分析此情况.

5.3 基于 SSDP 图的中文事件抽取算法

综上所述,本文研究的事件抽取主要包括 3 步.第 1 步,依次扫描 SSDP 图中核心根结点及其孩子结点;第 2 步,抽取事件主语、谓语和宾语,并判断是否启动补全;第 3 步,基于 5.2 节补全规则获取补全内容.过程如算法 5 所示.

算法 5. 基于 SSDP 图的中文事件抽取.

```
输入:语句的 SSDPgraph
输出:事件列表 ETL
FOR(coreRootNode∈CNS)
  //CNS 为 Root 的孩子结点集合,即所有核心根结点
  ET=∅;
  ET.pred=coreRootNode.tag;
  //tag 为 coreRootNode 结点的词标签
  FOR(cnode∈CCRNS)
    //CCRNS 为 coreRootNode 的孩子结点集合
    IF(cnode.dp 为 SBV)
      ET.sub=cnode.tag;
    END IF
    IF (cnode.dp 为 VOB 或 FOB)
      ET.obj=cnode.tag;
    END IF
    IF(cnode.tag 为"被"字且 coreRootNode.sdp 属于
        引发关系) //规则 11
      将最近关联事件的主语添加至 ET.obj;
    ELSE IF (cnode.prep 不为空且 coreRootNode.sdp
          属于引发关系) //规则 10
          //prep 为结点的介词关联标识
      将最近关联事件的主语添加至 ET.sub;
    ELSE IF(cnode.dp 为 SBV 且 coreRootNode.sdp 不
          属于引发关系) //规则 5
      将最近关联事件的主语添加至 ET.sub;
    ELSE IF (cnode.dp 为 SBV 且 coreRootNode.sdp
          属于引发关系)
      IF (最近关联事件中存在多个主语) //规则 6
        将第一个主语的 subNode.tag 添加至 ET.sub;
      ELSE IF (最近关联事件中只存在一个主语)
        IF (aubNode.postag 不属于 POL 且 subNode 存
            在定语) //规则 7
          将 subNode 的第一个定语添加至 ET.sub;
        ELSE IF (subNode.postag 不属于 POL 且 subNode
            不存在定语) //规则 8
          将最近关联事件中所有结点的 tag 组合添加
          至 ET.sub;
        ELSE IF (subNode.postag 属于 POL) //规则 9
          将最近关联事件 subNode.tag 添加至 ET.sub;
        END IF
      END IF
```

ELSE IF（*cnode.dp* 为 SBV 且 *cnode.postag* 不属于 *POL* 且 *coreRootNode.sdp* 为引发关系）
IF（关联事件主语 *subNode.postag* 属于 *POL* 且 *subNode* 存在定语且 *subNode* 的定语的 *postag* 属于 *POL*）//规则 12
将最近关联事件中第一个定语添加至 *ET.sub*；
ELSE IF（关联事件主语 *subNode.postag* 属于 *POL* 且 *subNode* 不存在定语）//规则 13
将最近关联事件中第一个主语添加至 *ET.sub*；
END IF
END IF
END FOR
将 *ET* 添加至 *ELT* 列表；
END FOR

6　实验测评

在实验中，依存句法分析、语义角色标注均采用哈尔滨工业大学语言技术平台 LTP[①]，语义依存分析使用哈工大联合科大讯飞公司共同推出的“哈工大-讯飞语言云”平台[②].

6.1　实验数据集

本实验定位于财经新闻标题，数据采自新浪财经网[③]滚动新闻，同时为了确保数据来源多元化，还选取了东方财富网[④]数据，用于验证抽取方法针对不同数据集的鲁棒性.

（1）数据集

本文选取新浪财经网（简称新浪网）2018 年 1 月至 12 月财经新闻标题，共计 492 336 条；东方财富网（简称东方网）2019 年 5 月至 6 月部分财经新闻数据，共计 978 条. 数据集中抽取事件及相关指标的统计结果如表 3 所示.

表 3　数据集中抽取事件及相关指标的统计结果

数据集	新闻数	事件数	无主语数	无宾语数
新浪网	492336	724294	229346	198991
东方网	978	1537	503	384
合计	493314	725831	229849	199375

其中，事件数为采用本文方法由计算机抽取得到的结果，非人工标注结果；无主语数、无宾语数分别指计算机直接抽取（即没有进行主语、宾语补全）时没有抽取到主语、宾语的事件数量. 新浪网中平均每条新闻标题中有 1.47 个事件，东方网中平均每条新闻标题中有 1.57 个事件.

为了验证本文所提规则的覆盖性，我们从新浪网数据集中随机选取 5000 条财经新闻标题，由计算机计算得到的规则覆盖情况如表 4 所示. 其中，事件数为 7575，事件间存在语义依存关系且后面的事件在句法结构上没有主语的事件对总数为 1401（可能存在直接成分缺省），事件间存在语义依存关系且后面的事件在句法结构上存在主语的事件对总数为 1460（可能存在间接缺省），占比的单位为%.

表 4　数据集中本文规则覆盖情况

规则	数量	总数	占比	规则	数量	总数	占比
规则 1	4842	5000	96.84	规则 8	154	1401	10.99
规则 2	2393	5000	47.86	规则 9	376	1401	26.84
规则 3	158	5000	3.16	规则 10	37	1460	2.53
规则 4	4098	7575	54.10	规则 11	64	1460	4.38
规则 5	685	1401	48.89	规则 12	31	1460	2.12
规则 6	49	1401	3.50	规则 13	73	1460	5.00
规则 7	27	1401	1.93	—	—	—	—

由表 4 可知，利用 CVC 比原始 DP 树多识别的事件占比增加 56.44%（包含事件的语句数为 4842，CVC 能够识别的事件数为 7575）. 另外，语料中直接成分缺省的补全规则占比较大，间接缺省仅占小部分，规则覆盖的总体情况与中文实际表达习惯较为吻合.

（2）文本预处理

新闻标题数据集中存在两个导致句法分析产生错误的问题：① 新闻标题常包含专家、企业的意见或评述，如标题“银保监会：信托公司监管评级将增设支持民企评分细项”中“银保监会：”，这些成分对于所需信息的抽取帮助甚微，且干扰 LTP 句法依存分析；② 新闻正、副标题之间一般以空格隔开，LTP 对于空格并不认为隔开两个子句，句法依存的分析效果不佳. 因此，在预处理阶段，首先去除正标题之前的内容（一般为“：”之前），然后以中文逗号代替正、副标题之间的空格.

（3）标注数据集

原始新闻标题数据集巨大，人工很难完成所有数据的标注，因此随机选择部分数据进行人工标注（新浪网 1200 条，东方网 500 条），并以此验证事件抽取效果. 标注数据集中标注事件及相关指标的统计结果如表 5 所示.

① http://ltp.ai/docs/index.html
② https://www.xfyun.cn/services/semanticDependence
③ http://finance.sina.com.cn/roll/#pageid=384&lid=2519&k=&num=50&page=1
④ http://finance.eastmoney.com/news/cywjh.html

表 5 标注数据集中标注事件及相关指标的统计结果

数据集	新闻数	事件数	无事件数	主语数	宾语数	补全主语数	补全宾语数
新浪网	1200	1898	68	1569	1350	523	475
东方网	500	718	15	571	572	213	149
合计	1700	2616	83	2140	1922	737	624

其中,事件数是指人工标注得到的事件数量;无事件数是指人工标注没有发现事件的新闻标题数量;主语数是指有主语(含补全主语)的事件数量;补全主语数是指从新闻标题直接标注得到的事件中缺省主语或主语不完整,但人工可以从相关联的事件中发现并补全主语的事件数量;宾语数、补全宾语数的概念分别类似于主语数和补全主语数.在合计标注的数据集中,补全主语数、补全宾语数分别占到了主语数和宾语数的 34.44%、32.47%,进一步验证了中文语句中缺省情况的普遍性以及补全的重要性.

本文共选用具备较强财经知识的 3 位教师作为事件标注者.标注标准:① 如果语句不是由动词触发,则标注为无事件;② 如果语句有核心动词,则认为存在事件,且每一个核心动词触发一个事件,但将同一个语句中相邻核心动词的多个事件合并为一个事件(核心动词间只包含副词也看成是相邻);③ 如果事件应该存在主语或宾语,则无论是否缺省,均标注为存在主语或宾语;④ 如果事件存在主语或宾语缺省,则标注为补全主语或补全宾语,并给出补全后的主语或宾语.当出现标注结果不一致的情形,则由 3 人讨论确认最终标注结果.标注一致性评测结果是 3 位标注者标注结果完全相同的数量占标注总数量的比例(单位:%),如表 6 所示.

表 6 人工标注数据集一致性评测

数据集	核心动词	事件	主语	宾语	补全主语	补全宾语
新浪网	98.63	97.79	97.71	99.11	94.46	98.32
东方网	98.33	97.35	97.20	98.78	94.37	96.64
合计	98.55	97.67	97.57	99.01	94.30	97.92

由表 6 可知,核心动词易判断,宾语及补全宾语结构简单,其标注一致性均比较高.补全主语因缺省结构复杂,标注一致性最低,分歧主要集中于简单缺省和组合缺省的判断,二者作为补全成分有时均成立,难以明确地区分.

6.2 评测指标

为了更好地理解测评指标,先对标注数据集中的相关统计指标进行说明,具体如表 7 所示.为了简化,可将正确抽取数、不完整抽取数分别简称为正确数、不完整数.

表 7 标注数据集中的相关统计指标

指标符号	指标含义	指标说明
LQ	标注数(Labeled Quantity)	人工标注得到的数量
EQ	抽取数 (Extracted Quantity)	基于本文方法由计算机抽取得到的数量
CEQ	正确抽取数 (Correct Extracted Quantity)	在抽取结果中抽取正确(即抽取结果也是人工标注结果)的数量
WEQ	错抽数 (Wrong Extracted Quantity)	在抽取结果中抽取错误(即抽取结果不是人工标注结果)的数量
MQ	漏抽数 (Missed Quantity)	人工标注结果中没有被计算机抽取到的数量
WQ	错误数 (Wrong Quantity)	在抽取过程中没有正确抽取的数量,包含错抽数和漏抽数
IEQ	不完整抽取数 (Incomplete Extracted Quantity)	核心动词抽取正确但其他属性抽取错误的事件数量

注:(1)对于事件抽取,正确数 *CEQ* 是指所有事件属性均抽取正确的事件数量,错抽数 *WEQ* 是指核心动词未抽取正确的事件数量;(2)对于事件抽取,抽取数 *EQ* 等于正确数 *CEQ*、错抽数 *WEQ* 和不完整数 *IEQ* 之和;(3)对于事件属性(核心动词、主语或宾语)抽取,抽取数 *EQ* 等于正确数 *CEQ* 和错抽数 *WEQ* 之和;(4)对于人工标注没有发现事件的新闻标题,如果计算机也没有抽取到事件,这表明计算机抽取正确,但由于标注数 *LQ* 中无法反映无事件数,因此在计算事件抽取评测指标(准确率、召回率和 *F*1 值)的时候均不考虑无事件数指标.

根据以上概念,错抽数(对于事件抽取还包括不完整抽取数)影响抽取准确率,错误数(包括错抽数和漏抽数,对于事件抽取还包括不完整抽取数)影响抽取召回率.基于表 7 的相关统计指标,可得到准确率(*Precision*,*P*)、召回率(*Recall*,*R*)和 *F*1 值 3 种评测指标的计算公式如下:

$$P = CEQ / EQ,$$

$$R = CEQ / LQ,$$

$$F1 = 2 \times P \times R / (P + R).$$

6.3 实验结果

针对事件及事件属性(核心动词、主语或宾语)抽取,首先给出表 7 所列各指标在标注数据集中的统计结果,再分别就准确率 *P*、召回率 *R* 和 *F*1 值 3 种指标进行评测,以评价本文方法的抽取效果.

6.3.1 实验统计数据

(1)核心动词抽取.核心动词抽取是事件抽取的关键,不仅反映了事件探测效果,还直接决定其他属性抽取的意义.标注数据集中核心动词抽取的各指标统计结果如表 8 所示.

表 8 核心动词抽取的统计结果

数据集	抽取数	正确数	错抽数	漏抽数
新浪网	1899	1763	136	67
东方网	715	691	24	16
合计	2614	2454	160	83

(2)主语抽取. 主语作为事件的实施者,其重要性不言而喻. 标注数据集中全部主语抽取和补全主语抽取的各指标统计结果如表 9 所示. 其中,全部主语抽取为所有主语抽取情况,包含补全主语抽取. 由表 9 可知,对于合计数据集而言,补全主语错抽数占全部主语错抽数的比例高达 85.66%,说明在本文语料中,全部主语错抽的影响主要源于补全主语错抽.

表 9　主语抽取的统计结果

数据集	全部主语				补全主语			
	抽取数	正确数	错抽数	漏抽数	抽取数	正确数	错抽数	漏抽数
新浪网	1579	1376	203	48	594	408	186	11
东方网	573	518	55	22	218	183	35	8
合计	2152	1894	258	70	812	591	221	19

(3)宾语抽取. 标注数据集中全部宾语抽取和补全宾语抽取的各指标统计结果如表 10 所示.

表 10　宾语抽取的统计结果

数据集	全部宾语抽取				补全宾语抽取			
	抽取数	正确数	错抽数	漏抽数	抽取数	正确数	错抽数	漏抽数
新浪网	1337	1211	126	35	537	436	101	3
东方网	563	520	43	19	177	135	42	3
合计	1900	1731	169	54	714	571	143	6

由表 10 可知,对于合计数据集而言,补全宾语错抽数占全部宾语错抽数的比例高达 84.62%,说明在本文语料中,全部宾语错抽的影响主要源于补全宾语错抽.

(4)事件抽取. 事件抽取包含事件所有属性的抽取,因此正确抽取是指事件全部属性均抽取正确. 标注数据集中事件抽取的各指标统计结果如表 11 所示.

表 11　事件抽取的统计结果

数据集	抽取数	正确数	不完整数	错抽数	漏抽数
新浪网	1899	1579	184	136	67
东方网	715	627	64	24	16
合计	2614	2206	248	160	83

6.3.2　缺省结构覆盖率

为了体现本文所提规则的覆盖情况,按照第 5 节描述的缺省结构对人工标注语料进行了统计,具体如表 12 所示. 表中数据为语料中缺省结构出现次数在缺省总数的占比(单位:%). 其中,“宾作主”表示事件宾语充当或修饰其他事件主语的情况.

表 12　缺省结构覆盖率

数据集	简单	组合	间接	介词	被动	宾作主
新浪网	69.22	11.09	12.81	2.68	3.06	1.15
东方网	74.65	8.45	9.86	2.82	4.23	0.00
合计	70.79	10.33	11.96	2.72	3.40	0.82

由表 12 可知,表中 6 种缺省结构涵盖了语料所有缺省情况,本文考虑了前 5 种,其覆盖率合计值达 99.18%. “宾作主”情况仅在新浪网中出现,覆盖率为 1.15%,说明本文提出的规则涵盖了绝大部分的缺省情况,覆盖率可以保证.

6.3.3　实验评测

(1)核心动词及事件抽取

评测结果如表 13 所示.

表 13　核心动词及事件抽取的效果

数据集	核心动词抽取/%			事件抽取/%		
	准确率	召回率	$F1$ 值	准确率	召回率	$F1$ 值
新浪网	92.84	92.89	92.86	83.15	83.19	83.17
东方网	96.64	96.24	96.44	87.69	87.33	87.51
合计	93.88	93.81	93.84	84.39	84.33	84.36

由表 13 可知,合计的核心动词抽取的 $F1$ 值达 93.84%,验证了按照核心动词链建立规则识别确认事件的有效性. 以上结果主要受益于事件绝大部分由动词触发,而每个事件是独立的,在语言学句法结构上均采用并列关系进行事件关联. 本文的核心动词抽取方法遵循了这一特点,将 SSDP 树中核心动词进行拆分,形成 SSDP 图,图中 *Root* 结点的每个孩子均为核心动词. 然而,核心动词在抽取过程中还存在一些不足,如多词性问题,词性的准确性一方面影响依存句法结构,另一方面影响核心动词的识别,在一定程度上降低了核心动词抽取的效果,后期工作可考虑结合动词搭配的论元来确定多词性词语的词性. 另外,新浪网的 $F1$ 值为 92.86%,较东方网的 96.44%低了 3.58 个百分点,主要是由两个原因导致的:① 新浪网的新闻标题更偏好于采用大量动词表达,一定程度上降低了识别新闻标题中核心动词的准确率;② 新浪网的新闻标题对正文的概括更为精简,词汇之间的关联降低,不利于句法分析.

对于事件抽取,通过 $F1$ 值可以发现,其抽取效果也不错,合计的 $F1$ 值为 84.36%,验证了本文方法对事件抽取的有效性. 但较于核心动词抽取,因同时添加了正确抽取主语和宾语的要求,$F1$ 值由核心动词抽取的 93.84%降至事件抽取的 84.36%,降低了 9.48 个百分点. 核心动词抽取正确时事件抽取错误(即事件不完整抽取的情况)的统计结果如表 14 所示,其中主语错抽数、宾语错抽数中都包含了“主宾语均错抽数”. 针对合计数据集,核心动词抽取正确时的主语错抽数、宾语错抽数与不完整数的占比分别为 68.95%和 46.37%,说明主语比宾

语被错抽的可能性更大.这是因为主语省略较宾语省略更为普遍,且形式多样化,规则难以全面覆盖、完全适用.

表 14 核心动词抽取正确时主语和宾语错抽的统计结果

数据集	不完整数	主语错抽数	宾语错抽数	主宾语均错抽数
新浪网	184	126	81	23
东方网	64	45	34	15
合计	248	171	115	38

(2)主语抽取

评测结果如表 15 所示.其中,合计的全部主语抽取的 $F1$ 值达 88.26%,验证了本文方法对于主语抽取的有效性.

表 15 全部主语及补全主语抽取的效果

数据集	全部主语抽取/%			补全主语抽取/%		
	准确率	召回率	$F1$ 值	准确率	召回率	$F1$ 值
新浪网	87.14	87.70	87.42	68.69	78.01	73.05
东方网	90.40	90.72	90.56	83.94	85.92	84.92
合计	88.01	88.50	88.26	72.78	80.30	76.36

同时,主语抽取的效果严重依赖于核心动词抽取的效果.分析如下:① 如果核心动词抽取正确,则主语被正确抽取的可能性较大.例如,针对合计数据集,基于表 14 可计算得到核心动词抽取正确时的主语错抽数与表 8 中的核心动词正确数(2454)的占比为 6.97%,即核心动词抽取正确时主语错抽率仅为 6.97%;② 如果核心动词抽取错误,则主语被错抽的概率就要大得多.例如,由于核心动词错抽导致主语和宾语错抽的统计结果如表 16 所示.

表 16 核心动词错抽导致主语和宾语错抽的统计结果

数据集	核心动词错抽数	主语错抽数	宾语错抽数	主宾语均错抽数
新浪网	136	77	45	34
东方网	24	10	9	8
合计	160	87	54	42

在表 16 中,主语错抽数、宾语错抽数中都包含了“主宾语均错抽数”.针对合计数据集,主语错抽数占核心动词错抽数的比例为 54.38%,即核心动词抽取错误导致的主语错抽率高达 54.38%.

对于补全主语抽取,其合计的 $F1$ 值为 76.36%,验证了本文提出的主语缺省补全规则的有效性.但相较于其他属性的抽取效果,补全主语的 $F1$ 值最低.主要源自于如下几个方面:

① 未考虑利用关联事件的宾语补全缺省主语的情况.存在利用关联事件中的宾语补全缺省主语的情况,如语句 S_8,宾语“私募基金”作定语补全“规模”.

② 主语省略形式多样化.缺省事件需要补全的主语以多样化的形式处于关联事件中,给出的规则难以适用于所有情况.

③ 语义依存关系存在错误.本文借助了结点间语义依存关系,但对于核心动词间的语义依存关系,SDP 工具存在语义依存关系结果分析错误的情况,使得不满足相关规则,导致抽取错误.

另外,对比两个标注数据集,补全主语抽取的效果存在差异,主要是由于新浪网的新闻标题过于精简并采用多动词所致,凸显了词语多词性问题带来的影响.

(3)宾语抽取

评测结果如表 17 所示.由表 17 可知,合计的全部宾语和补全宾语抽取的 $F1$ 值分别为 90.58%和 85.35%,验证了本文方法对于宾语及补全宾语抽取的有效性.

表 17 全部宾语及补全宾语抽取效果

数据集	全部宾语抽取/%			补全宾语抽取/%		
	准确率	召回率	$F1$ 值	准确率	召回率	$F1$ 值
新浪网	90.58	89.70	90.14	81.19	91.79	86.17
东方网	92.36	90.91	91.63	76.27	90.60	82.82
合计	91.11	90.06	90.58	79.97	91.51	85.35

同样,宾语抽取的效果也是严重依赖于核心动词抽取的效果.例如,针对合计数据集,基于表 14 可计算得到核心动词抽取正确时的宾语错抽数与表 8 中的核心动词正确数(2454)的占比为 4.69%,即核心动词抽取正确时宾语错抽率仅为 4.69%;基于表 16 可计算得到宾语错抽数占核心动词错抽数的比例为 33.75%,即核心动词抽取错误导致的宾语错抽率高达 33.75%.

相对于主语抽取,无论是全部还是补全,宾语抽取的效果均要更好.由前文叙述可知,宾语缺省形式较为常规化,主要由介词和被动语态引起,其余大部分为含有宾语和无宾语情况(不及物动词作为核心动词),规则容易总结,且适用性较好,使得其效果好于主语抽取,但对标宾语抽取规则本身,还存在 2 点不足:① 被动语态只考虑了“被”字结构,中文中还存在一些其他表示被动的词语,如“遭”字结构等;② 语义依存关系用于判别介词引发的宾语缺省存在一定的局限性,其准确率还有待进一步提高.

另外,对于主语或宾语抽取,其效果除了受核心动词抽取的影响以及与补全主语或补全宾语抽取的效果有关之外,还会受 LTP 分词、依存句法分析结果的影响.标注数据集中包含了部分分词及依存

句法分析错误的情况，对于分词及依存句法分析正确的数据对象，本文方法在事件及各属性上的抽取效果均会有一定程度的提高.但中文分词本身就是一个很具挑战性的开放难题，需要考虑语言学特点和词语语义等情况，有待提出更好的解决方法.

针对两个人工标注数据集，事件及各属性抽取效果的直观对比分析结果分别如图 16 和图 17 所示，其中，横坐标为事件及各属性下的 3 种指标，纵坐标为各指标值(单位：%).从图 16 和图 17 可以看出，东方网的抽取效果整体略好于新浪网，这主要受核心动词抽取影响.

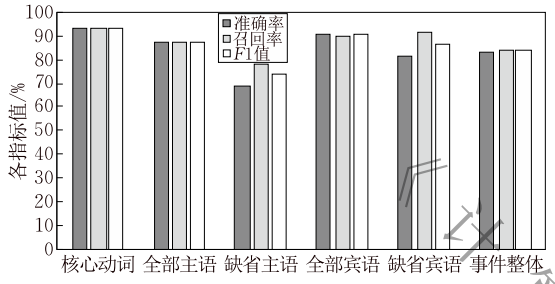


图 16　新浪网的事件及各属性抽取效果对比

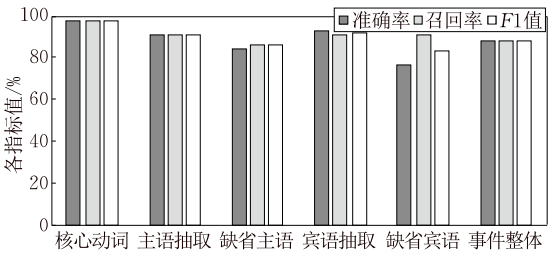


图 17　东方网的事件及各属性抽取效果对比

6.3.4　其他领域抽取效果

为了体现本文方法具有较好的扩展性，选择对开放域新闻进行事件抽取实验.本文随机选取新浪网 2019 年新闻标题 600 条.其中，人工标注的事件数、全部主语数和补全主语数分别为 957、804 和 125，无宾语缺省情况；实验抽取的各项指标的统计数据如表 18 所示.

表 18　新浪新闻中主语补全及事件抽取的统计结果

事项	抽取数	正确数	不完整数	错抽数	漏抽数
主语补全	143	100	0	43	11
事件抽取	956	853	69	34	18

由标注数据可知，补全主语数在全部主语数中的占比为 15.55%，远低于财经新闻领域，且通过标注发现，缺省结构基本上集中于简单主语缺省.这是由于领域的特性所致.

财经领域新闻标题大多描述某个公司或企业不同方面的相关信息，在中文表达中，为了简洁，同一个主语在后续相邻语句的表达中，无论充当相同成分(简单缺省)还是作定语修饰成分(间接缺省)，常省略.而开放域新闻常为一个事件的发生如何影响其他事件，单个事件的成分比较健全，所以缺省数较少.另外，财经领域包含较多数值词，即描述事件的具体情况，这使得事件之间存在较多因果关系，而因果关系中的结果，有较大一部分是由一个语句整体所致，所以，财经领域存在一定的组合缺省.由此可知，财经领域较其他领域，不仅存在较多的缺省情况，且缺省的形式较为丰富，进一步佐证了本文研究财经金融领域的事件抽取及缺省成分补全具有较大的现实意义.

在抽取效果上，测评结果如表 19 所示.

表 19　新浪新闻中事件抽取及主语补全的效果

事件抽取/%			补全主语抽取/%		
准确率	召回率	F1 值	准确率	召回率	F1 值
89.23	89.13	89.18	69.93	80.00	74.63

从表 19 与表 13、表 15 对比可知，无论是事件抽取还是主语补全，在开放域的效果均与金融领域相当，且略有提升，说明本文方法在领域扩展上具有较好的适应性，鲁棒性较强.对于事件抽取，F1 值较表 13 中提高了 6.01 个百分点，主要是因为领域专业词汇较少，分词及句法结构分析的结果较好.在补全主语抽取方面，因开放域缺省结构简单，其 F1 值比表 15 提升了 1.58 个百分点.

6.4　与其他方法的实验对比

(1) 基线实验选择

本文从两个方面选择对比方法.一方面选择 DP[34] 和 SDP 抽取方法，验证 SSDP 组合的有效性；另一方面选择 SRL[LTP][40]、SRL[Mate]①等 SRL 方法和 DMCNN[14]、JRNN[15] 和 JMEE[17] 等事件抽取方法作为对比方法，验证基于 SSDP 的事件抽取及基于所提规则的缺省补全方法的优势.其中，2 种 SRL 方法直接给出语句包含的主语、谓语和宾语，3 种事件抽取方法对语句包含的词语进行触发词、论元和论元角色分类.

需要说明的是，DMCNN、JRNN 和 JMEE 方法均聚焦于设计先进方法进行传统事件抽取，不属于针对事件抽取中的某个特定问题而设计的方法，如

① https://code.google.com/archive/p/mate-tools/

训练数据不足或篇章级事件等,因此本文选择这些方法作为基线实验.这些方法在原文中做了很多论元角色的判断,但在我们实现相关方法时,实验标注的数据集中只包含主语、谓语、宾语和其他角色的分类,仅考察这些方法对事件的主语、谓语和宾语的正确分类效果.

另外,由于语言类别和语料的不同,本文对相关方法做了以下几点修改:① ACE2005 语料包含事件类型、实体类型等信息,在本文实验中将此信息输入为空;② 触发词和论元及角色按照 ACE2005 划分的全部类别进行分类,但本文只对主语、谓语和宾语抽取的效果进行对比.

(2)基线实验数据集及参数设置

为了避免 DMCNN、JRNN 和 JMEE 等方法因训练数据不足难以发挥其抽取效果,同时为了进一步验证本文方法在非标题(长句)、开放领域数据下的抽取能力,我们不仅在 6.1 节描述的标注数据集上而且在 CoNLL2009 中文语料上进行事件抽取实验.其中,CoNLL2009 中文语料分训练集、验证集和测试集 3 部分,包含的语句数分别为 22277 条、1762 条和 22 条.

在本文标注数据集上,我们随机选择 30%作为测试集,剩余的作为训练集,并从训练集中随机选择 10%作为验证集.对于 CoNLL2009,因测试集太小,我们随机从训练集中不放回地抽取 313 条语句增加至测试集,即最后确定训练集 21964 条、验证集 1762 条、测试集 335 条.实验涉及的词向量由 Word2Vec① 工具在本文 2 个数据集上训练得到,词向量维度分别与文献[14]、文献[15]和文献[17]保持一致,Word2Vec 其余参数的设定标准依据词汇语义相似度.每条语句分词个数最大设定为 100.对于基线实验模型所需的超参数取值,采取网格搜索函数 GridSearchCV 选择最优值,基于 CoNLL2009 的模型超参数最终取值情况如表 20 所示.

表 20 基于 CoNLL2009 的超参数取值情况

	超参数				
	batch_size	epochs	dropout	activation	learn_rate
DMCNN	64	10	0.2	relu	0.010
JRNN	32	5	0.2	tanh	0.001
JMEE	32	8	0.5	tanh	0.001

在实验测试方面,CoNLL2009 未全部标注本文提出的缺省补全信息,如间接缺省和组合缺省等.因此,我们按照前述的标注标准对测试集进行了补充标注.在 CoNLL2009 上的测评均基于该补充标注的测试集.另外,本文规定只有触发词、论元及论元角色全部正确分类(仅指事件的主语、谓语和宾语全部正确,包括主语和宾语的缺省内容补全),才认定为本文的事件抽取正确.所以,测试过程分为三步,首先进行触发词抽取,然后判断触发词抽取情况,仅当触发词抽取正确时才启动论元抽取,最后依据二者的抽取情况进行综合计算,得出最终抽取结果.

(3)对比分析结果

DP 和 SDP 方法分别分析语句的句法和语义依存情况.为了使实验具有可比性和说服力,我们假设:① DP 和 SDP 方法均按照核心动词链的建立和调整规则进行扩展;② 两种方法都采取缺省补全规则进行属性查询补充;③ 因为介词和被动语态引起的缺省,需要结合语义分析结果进行 SSDP 树结构调整,所以 DP 方法对上述两种缺省不做调整;④ DP 方法不建立事件间语义关联,对于任何缺省均按补全规则查询.由于 SRL 方法给出了论元角色标注结果,因此可直接通过标注的角色获取事件 *ET* 各属性.

在新浪网标注数据集上,5 种方法抽取的统计数据如表 21 所示.由于 DMCNN、JRNN 和 JMEE 方法都是通过语句中词的分类直接判断抽取效果,因此在表 21 中未给出这 3 种方法的统计结果.

表 21 新浪网上事件抽取的统计结果

抽取方法	抽取数	正确数	不完整数	错抽数	漏抽数
SSDP	1899	1579	242	78	64
DP	1892	1441	371	80	86
SDP	1759	655	869	235	374
SRL[LTP]	1949	898	696	355	304
SRL[Mate]	2533	696	828	1009	139

在新浪网标注数据集上,8 种方法抽取的准确率、召回率和 *F*1 值如表 22 和图 18 所示.其中,图 18 中横坐标为 3 种评测指标下的事件抽取的 8 种方法,纵坐标为各指标值(单位:%).

表 22 新浪网上事件抽取的效果对比

抽取方法	准确率	召回率	*F*1 值
SSDP	83.15	83.19	83.17
DP	76.16	75.92	76.04
SDP	37.24	34.51	35.82
SRL[LTP]	46.07	47.31	46.69
SRL[Mate]	27.48	37.06	31.56
DMCNN	56.14	39.75	46.55
JRNN	43.55	52.94	47.79
JMEE	53.64	52.87	53.26

① https://code.google.com/p/word2vec/

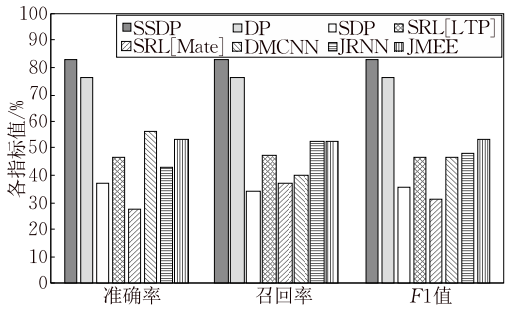


图 18　新浪网上 8 种方法的事件抽取效果对比

由表 22 和图 18 可知，本文方法明显优于其他方法. 这是因为 SSDP 结合了句法和语义双重结构特征. 因此，添加语义关联、调整优化 DP 树，可提高事件的识别能力，进而提升属性抽取效果.

在 $F1$ 值上，DP、SDP 比 SSDP 分别低了 7.5、47.35 个百分点. DP 主要是因为事件间缺乏语义关联，没有根据其关联类型进行补全机制判断，即未考虑事件间语义，而直接采用句法结构进行缺省补全，导致效率有所下降. 另外，DP 的 $F1$ 值较高，说明了 DP 句法结构分析效果不错，也间接验证了该工具被普遍采用的原因. 而 SDP 则主要是由于语义依存结构分析效果不佳所导致的，同时不支持自定义词典的添加，也进一步增加了词汇间语义依存的错误率，使得抽取错误数急剧增加.

SRL[LTP]和 SRL[Mate]的抽取效果较差，$F1$ 值分别为 46.69%和 31.56%. 这主要是受核心动词识别效果的影响，因为中文存在大量多词性的动词，SRL 基本将所有动词均识别为语句谓语，导致事件错抽数较高，从而大幅降低了抽取效果.

关于 DMCNN、JRNN 和 JMEE 方法的抽取效果，可从两个方面探讨. 一方面，对比 SSDP，其抽取效果不太理想，$F1$ 值减少了 29.91～36.62 个百分点；另一方面，对比原文献[14-15，17]，事件抽取效果在总体上与原文中给出的结果相当，但均有所降低. 主要原因包括：① 本文语料(新闻标题)较短，语句含有的上下文信息有限，深度学习难以提取较多有用信息；② 本文语料无事件类型和实体类型等信息，输入特征减少；③ 原文未考虑缺省补全情况；④ 中文需要分词，而分词存在一定错误，同时也会降低依存句法分析效果；⑤ 供模型训练的可用语料偏少.

在 CoNLL2009 中文语料上，4 种方法的事件抽取效果如表 23 和图 19 所示.

表 23　CoNLL2009 上事件抽取的效果对比

抽取方法	准确率	召回率	$F1$ 值
SSDP	76.08	65.16	70.20
DMCNN	61.74	43.67	51.16
JRNN	57.83	48.48	52.75
JMEE	64.91	48.05	55.22

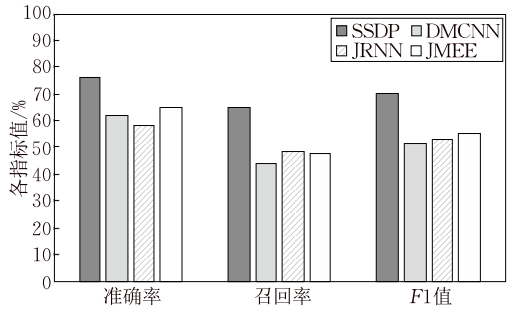


图 19　CoNLL2009 上 4 种方法的事件抽取效果对比

由表 22、表 23 可知，针对 CoNLL2009 中文语料，相较于财经新闻标题语料，虽然 SSDP 方法的事件抽取效果出现了较大幅度的降低，DMCNN、JRNN 和 JMEE 方法的事件抽取效果均有一定幅度的提高. 但是，SSDP 方法的 $F1$ 值仍高于 DMCNN、JRNN 和 JMEE 方法 14.98～19.04 个百分点.

通过分析，SSDP 方法效果降低的主要原因包括：① CoNLL2009 中文语料中的长句不利于 LTP 结构分析，关键的核心动词结构错误将导致其包含的事件无法识别；② 长句不利于缺省补全，长句覆盖的结构复杂且词语多，增大了补全的难度；③ 出现部分长句整体做为宾语的结构，因宾语在语言学结构上不存在 COO 并列关系扩展，本文未考虑此情况，致使长句宾语包含的大量事件无法识别. 同时，DMCNN、JRNN 和 JMEE 方法效果提高的原因可能是增加了训练数据所致.

实验结果充分说明，在金融领域的中文事件抽取及事件成分缺失补全方面，本文方法具有明显的优势和较强的适应性.

7　总结与展望

事件抽取对宏观经济趋势预测具有重要意义，目前事件抽取侧重于抽取分类的正确性，未结合应用需求进行分析，难以较好地应用于特定领域.

本文针对金融领域财经新闻标题数据，归纳了事件漏抽、事件成分缺失、事件成分抽取错误及事件语义放大等 4 种现象，提出了句法和语义依存分析相结合的事件抽取框架——SSDP 图. 首先，利用

LTP 工具获得语句的依存句法分析结果，并将其转换为 DP 树；其次，归纳核心动词链的建立规则，解决事件漏抽问题；第三，引入事件间语义依存关系，构建 SSDP 树；第四，根据核心动词链、介词结构和被动语态结构调整 SSDP 树，形成 SSDP 图；最后，基于 SSDP 图，建立事件成分缺失补全规则，同时抽取中文金融事件.

下一步的研究工作主要包含：

(1)通过 LTP 进行依存句法分析时发现，多词性词语的句法结构分析效果较差，如何利用论元信息进一步确定多词性词语在具体语句中的词性，是有待克服的一个障碍.

(2)本文抽取的事件结构，只考虑了 ET 三元组，抽取哪些信息将对股市趋势预测等应用有价值，是我们感兴趣的工作.

(3)由于 SSDP 方法中用到的事件间语义依存分析较为简单、粒度较粗，如何制定针对财经领域金融事件间语义关联，将是未来的工作之一.

致 谢 本文的研究工作利用了哈尔滨工业大学社会计算信息检索研究中心免费开放的 LTP 平台、哈尔滨工业大学联合科大讯飞公司共同推出的讯飞开放平台，在此一并表示感谢！最后，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] Ding X, Zhang Y, Liu T, et al. Using structured events to predict stock price movement: An empirical investigation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1415-1425

[2] Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction//Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina, 2015: 2327-2333

[3] Ding X, Zhang Y, Liu T, et al. Knowledge-driven event embedding for stock prediction//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING). Osaka, Japan, 2016: 2133-2142

[4] Xie B, Passonneau R, Wu L, et al. Semantic frames to predict stock price movement//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria, 2013: 873-883

[5] Aguilar J, Beller C, McNamee P, et al. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards//Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Baltimore, Maryland, 2014: 45-53

[6] Jacobs G, Lefever E, Hoste V. Economic event detection in company-specific news text//Proceedings of the 1st Workshop on Economics and Natural Language Processing (ACL). Melbourne, Australia, 2018: 1-10

[7] Yang H, Chen Y, Liu K, et al. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations (ACL). Melbourne, Australia, 2018: 1-6

[8] Li Peng-Feng, Zhou Guo-Dong, Zhu Qiao-Ming. Semantics-based joint model of Chinese event trigger extraction. Journal of Software, 2016, 27(2): 280-294(in Chinese)
(李培峰，周国栋，朱巧明. 基于语义的中文事件触发词抽取联合模型. 软件学报，2016，27(2)：280-294)

[9] Yeh C L, Chen Y C. Zero anaphora resolution in Chinese with shallow parsing. Journal of Chinese Language and Computing, 2007, 17(1): 41-56

[10] Li P, Zhu Q, Zhou G. Argument inference from relevant event mentions in Chinese argument extraction//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria, 2013: 1477-1487

[11] Tang Wen-Wu, Guo Yi, Xu Yong-Bin, et al. The default common object identification based on condition random fields. Journal of Chinese Information Processing, 2016, 30(6): 208-214(in Chinese)
(唐文武，过戈，徐永斌等. 基于条件随机场的评价对象缺省识别. 中文信息学报，2016，30(6)：208-214)

[12] Chan Y S, Fasching J, Qiu H, et al. Rapid customization for event extraction//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL). Florence, Italy, 2019: 31-36

[13] He Rui-Fang, Duan Shao-Yang. Joint Chinese event extraction based multi-task learning. Journal of Software, 2019, 30(4): 1015-1030(in Chinese)
(贺瑞芳，段绍杨. 基于多任务学习的中文事件抽取联合模型. 软件学报，2019，30(4)：1015-1030)

[14] 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 (ACL). Beijing, China, 2015: 167-176

[15] 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, California, 2016: 300-309

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

[17] Liu X, Luo Z, Huang H. Jointly multiple events 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

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

[19] Araki J, Mitamura T. Open-domain event detection using distant supervision//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, USA, 2018: 878-891

[20] Liu S, Cheng R, Yu X M, et al. Exploiting contextual information via dynamic memory network for event detection//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 1030-1035

[21] Hong Y, Zhou W, Zhang J, et al. Self-regulation: Employing a generative adversarial network to improve event detection//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia, 2018: 515-526

[22] Guan C, Cheng Y, Zhao H. Semantic role labeling with associated memory network//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis, Minnesota, 2019: 3361-3371

[23] Li Z, He S, Zhao H, et al. Dependency or span, end-to-end uniform semantic role labeling//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, 2019: 6730-6737

[24] Xia Q, Li Z, Zhang M, et al. Syntax-aware neural semantic role labeling//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, 2019: 7305-7313

[25] He S, Li Z, Zhao H, et al. Syntax for semantic role labeling, to be, or not to be//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia, 2018: 2061-2071

[26] He L, Lee K, Levy O, et al. Jointly predicting predicates and arguments in neural semantic role labeling//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne, Australia, 2018: 364-369

[27] Mehat S V, Lee J Y, Carbonell J. Towards semi-supervised learning for deep semantic role labeling//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 4958-4963

[28] Tan Z, Wang M, Xie J, et al. Deep semantic role labeling with self-attention//Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 4929-4936

[29] He L, Lee K, Lewis M, et al. Deep semantic role labeling: What works and what's next//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver, Canada, 2017: 473-483

[30] Li Z, He S, Cai J, et al. A unified syntax-aware framework for semantic role labeling//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium, 2018: 2401-2411

[31] Kasai J, Friedman D, Frank R. Syntax-aware neural semantic role labeling with supertags//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis, Minnesota, 2019: 701-709

[32] Liu X, Huang H, Zhang Y. Open domain event extraction using neural latent variable models//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy, 2019: 2860-2871

[33] Zong Cheng-Qing. Statistical Natural Language Processing. 2nd Edition. Beijing: Tsinghua University Press, 2013 (in Chinese)
(宗成庆. 统计自然语言处理. 第 2 版. 北京：清华大学出版社，2013)

[34] Che W, Li Z, Liu T. LTP: A Chinese language technology platform//Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China, 2010: 13-16

[35] Li Jin-Xi. The New Chinese Grammar. 1955 Edition. Beijing: The Commercial Press, 1955(in Chinese)
(黎锦熙. 新著国语文法. 1955 年版. 北京：商务印书馆，1955)

[36] Lv Shu-Xiang. Essentials of Chinese Grammar. 1982 Edition. Beijing: The Commercial Press, 1982(in Chinese)
(吕叔湘. 中国文法要略. 1982 年版. 北京：商务印书馆，1982)

[37] Wang Li. Modern Chinese Grammar. 1985 Edition. Beijing: The Commercial Press, 1985(in Chinese)
(王力. 中国现代语法. 1985 年版. 北京：商务印书馆，1985)

[38] Qian Shi-Feng. Summary of omission definition. Journal of Language and Literature Studies, 2007, (1): 119-122 (in Chinese)
(钱世风. 省略界定综述. 语文学刊：高数版，2007，(1)：119-122)

[39] Xue N, Xia F, Huang S, et al. The bracketing guidelines for the Penn Chinese TreeBank (3.0). IRCS Technical Report Series, 2000: 39

[40] Guo J, Che W, Wang H, et al. A unified architecture for semantic role labeling and relation classification//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING). Osaka, Japan, 2016: 1264-1274

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

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

HU Rong, M. S., assistant researcher. Her current research interests include information extraction, natural language processing and big data analysis.

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

Background

As a sub-task of information extraction, event extraction plays an important role in various NLP applications including stock prediction and information retrieval. Event nesting and element defaults are common in Chinese. In this paper, we address two problems, determining the number of events contained in a Chinese sentence and extracting the structured event, which is a triple containing a subject, a predicate, and an object.

The main research for structured event extraction focuses on extracting all the properties of the triple, but don't obtain the default component of event. In addition to this, most of other existing research efforts have been put on the event extraction, but they pay more attention on the type correctness of triggers and arguments, which not consider to the completeness of events including the number of events in a sentence and the property in an event. In financial news headlines, there are a large number of verbs and component defaults, which cause the event to leak and the properties of extracted event to be incomplete. Furthermore, the event types are only for standard event types, such as ACE, which is not exact suitable for finance and economics.

Our work not only extracts all the events in a sentence, but also completes the default components, which can improve their usage value, such as for stock market trend forecasts.

In consideration of the characteristics of Chinese financial news headlines, we capture the syntactic relationships between words and summarize the rules of core verb chain formation, which can solve the problem of event leak. In addition, we add the semantic associations between events to form the SSDP tree and adjust SSDP structure to build the SSDP graph. At last, we present four default structures, and propose corresponding completion rules. To the best of our knowledge, our work is the first solution towards this problem.

The research is partially supported by the National Natural Science Foundation of China under Grant Nos. 61972184, 61562032 and 61762042, the Science & Technology Project of the Department of Education of Jiangxi Province under Grant Nos. GJJ180198 and GJJ180252.