基于协陪义动词的中文隐式实体关系抽取

万常选 甘丽新 江腾蛟 刘德喜 刘喜平 刘 玉

(江西财经大学信息管理学院 南昌 330013)

(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 实体关系抽取的目标在于探测实体之间的显式关系和隐式关系.现有研究大多集中在显式实体关系抽取,而忽略了隐式实体关系抽取.针对旅游和新闻领域文本经常包含许多由协陪义动词引发的隐式实体关系,本文研究了基于协陪义动词的中文隐式实体关系抽取问题.将机器学习方法与规则相结合,借助于显式实体关系对隐式实体关系进行推理.首先,利用依存句法分析,设计了协陪义候选句型分类算法以及相应的协陪义成分识别算法;其次,根据协陪义成分和协陪义动词作用范围的特点,设计了三种句内基于协陪义动词的隐式实体关系推理规则;最后,利用协陪义句中零形回指的先行词,建立不同句子中协陪义动词的主体成分与客体成分之间的联系,实现句间基于协陪义动词的隐式实体关系抽取.另外,本文还提出了趋向核心动词特征提取算法,进一步提高了动词特征对显式实体关系抽取的效果.基于真实的旅游领域和新闻领域文本数据集进行了详细的实验测试,实验结果表明了方法的有效性.

关键词 关系抽取;隐式关系;协陪义动词;显式关系;动词特征

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2019.02795

Chinese Named Entity Implicit Relation Extraction Based on Company Verbs

WAN Chang-Xuan GAN Li-Xin JIANG Teng-Jiao LIU De-Xi LIU Xi-Ping LIU Yu

(*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013)

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

Abstract The target of named entity relation extraction is to detect explicit and implicit relations between entities. Most of the existing researches focus on explicit entity relation extraction, but ignore implicit entity relation extraction. Compared with explicit relations, implicit relations have no explicit supporting evidence in text and require additional evident from a reading of the document. Therefore, implicit relations usually need to integrate semantic associations of sentence content with relevant linguistic information, specific context semantic information and related domain knowledge for indirect inference. However, because of the ambiguity of semantic relations, the complexity of sentence structures, the uncertainty of context information and the imbalance of data, the task of implicit relation extraction is more complicated and more difficult, and it cannot be implemented using a general model. Therefore, it has been a challenge to infer implicit relations. Several works related to implicit relation extraction have been performed for European languages and especially for English. As far as we know, very few studies have been done for

收稿日期:2016-06-28;在线出版日期:2017-05-21.本课题得到国家自然科学基金项目(61562032,61662027,61173146,61363039,61363010,61462037)、江西省自然科学基金项目(20152ACB20003,20161BAB202057)、江西省高等学校科技落地计划项目(KJLD12022,KJLD14035)、江西省教育厅科技研究项目(GJJ150819,GJJ160783)、江西省高校人文社会科学研究项目(JC161001)资助.**万常选**,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、信息检索. E-mail: wanchangxuan@263.net. **甘丽新**,博士研究生,副教授,主要研究方向为自然语言处理、信息检索. **江腾蛟**,博士,讲师,主要研究方向为 Web 数据管理、情感分析. **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 Web 数据管理、信息检索、自然语言处理. **刘喜平**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为信息检索、数据挖掘. **刘 玉**,硕士研究生,主要研究方向为数据挖掘.

Chinese language. In many text domains such as tourism and news domains, there exist many implicit entity relations triggered by company verbs. In this paper, we study the problem of Chinese implicit entity relation extraction based on company verbs. This paper proposes a two-stage scheme that takes into account both explicit relation extraction and implicit relation extraction. We integrate a machine learning method with rules and use explicit entity relations to infer implicit entity relations. Firstly, the company verb vocabulary is constructed by using a variety of methods and is used to select candidates from sentences containing company verbs. Secondly, the sentence pattern classification algorithm and the corresponding component recognition algorithm are designed for company candidate sentences. According to different roles of company verbs in the sentence, we employ dependency parsing to decide company candidate sentence patterns and to classify them. Due to the different roles of company verbs in different sentence patterns, methods of recognizing components from entities involved in company actions are also different. Using dependency parsing, we design corresponding component recognition algorithms for five kinds of company candidate sentence patterns. Finally, according to whether additional knowledge and the company verb are in the same sentence, we propose two kinds of inference methods for implicit relations based on company verbs: one for implicit in-sentence relations and the other for implicit between-sentences relations, where an in-sentence relation and a between-sentences relation refer to a relation that is inferred from a single sentence and multiple sentences, respectively. According to the characteristics of company semantic components and the scope of company verbs, we design three rules for implicit in-sentence entity relation extraction based on company verbs. Furthermore, by exploiting the antecedent of the zero anaphora in a company sentence, we establish the associations between subject and object components in different sentences, which are then used to extract implicit between-sentences entity relations based on company verbs. In addition, the feature extraction algorithm of directional core verbs is proposed to improve the effect of the verb feature on the explicit entity relation extraction. Comprehensive experiments are conducted on real tourism and news texts, and experimental results show that the proposed methods are effective.

Keywords relation extraction; implicit relation; company verb; explicit relation; verb feature

1 引 言

处在大数据时代的今天,随着数据规模呈指数级爆炸式的增长和数据模式的高度复杂化,使得“信息过载”和“信息泛滥”问题日益严重.因此迫切需要借助一些自动化的工具从大数据中快速准确地提取出人们所需要的有用信息,信息抽取技术由此而产生.作为信息抽取的一个重要子任务——实体关系抽取,它的目标是从文本中提取出两个命名实体之间的语义关系,如句子“李江游览泰山.”中的两个实体“李江”和“泰山”之间存在着“游历”关系.实体关系抽取已被广泛应用于自动问答系统、大规模信息处理、知识库的自动构建、搜索引擎和机器翻译等领域.

ACE 评测中的子任务 RDC(Relation Detection and Characterization)的目标在于探测实体之间的显式和隐式关系[1].显式实体关系通过文本中的显式证据表明其关系,而隐式实体关系则需要通过额外的知识帮助其关系的提取[1-2].

隐式关系由于缺少支持具体关系类型的直接证据,通常需要额外的知识进行间接推理.然而,由于额外知识难以获取,使得隐式关系抽取任务更难,而且还无法采用通用的方法实现统一建模.因此,已有关系抽取的相关研究大多集中在显式关系抽取,很少对隐式关系抽取进行研究.所以对中文隐式关系抽取的研究是关系抽取领域中最具挑战性的任务之一.

由于中文语言表达自由灵活,导致文本中蕴含大量种类繁多的隐式实体关系,如例 1 所示.

例 1.　“访华团在农业厅领导的陪同下游览泰山.”中的实体关系如图 1 所示.

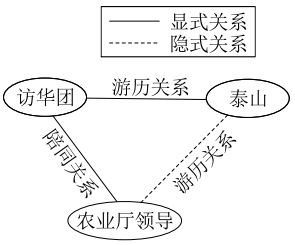

图 1　例 1 的实体关系图

在图 1 中，经典的关系抽取算法很容易探测到实体“访华团”与实体“泰山”之间存在显式的“游历”关系，因为该句中的动词“游览”能显式地表明该关系类型.而实体“农业厅领导”与“泰山”之间的“游历”关系则隐含在句中，需要提取额外的知识来辅助其关系判别.如在该句中，通过提取协陪义动词“陪同”，可知实体“访华团”和“农业厅领导”之间存在着“陪同”关系.在深层结构中，该协陪义动词表明实体“农业厅领导”不仅是实体“访华团”的陪同者，还和实体“访华团”一起实施“游览泰山”的行为，从而推理出“农业厅领导”与“泰山”之间也存在着“游历”的隐式关系.

本文主要研究旅游和新闻领域中基于协陪义动词的中文隐式实体关系抽取.旅游和新闻领域信息一般陈述了某人或某组织在某地发生过的事情.为了凸显名人的重要地位和对景点产生的名人效应，使其更容易话题化或焦点化，特别是在新闻语体中，句中经常出现协陪义动词，如以下 3 个句子中的“携、陪同、率领”均为协陪义动词.

例 2.　著名微波物理学家任之恭先生携夫人游览泰山.

例 3.　张木良陪同陈先生到北京访问.

例 4.　市政协主席率领调研组赴温岭市实地考察.

协同动词所表示的动作、状态是由两个或两个以上的个体协同作用而形成的[3].协陪义动词(又称引陪义动词)，属于协同动词的下位分类，表示的是某个个体协助、陪同、带领、护送另一个体一同进行某种活动[4].协陪义动词的客体成分和主体成分通常具有相同的语义特征，即他们共同施行某一动作行为.由此可见，在旅游和新闻领域中包含着许多由协陪义动词触发的隐式实体关系.

综上所述，本文试图以协陪义动词为核心，根据句子结构和上下文的理解构建基于协陪义动词的隐式关系抽取推理规则，致力解决中文旅游和新闻领域文本中的基于协陪义动词的隐式关系抽取问题.本文研究包括协陪义候选句的发现、协陪义句型的判断、协陪义成分的识别、基于协陪义动词的隐式关系推理以及趋向核心动词特征的提取.

首先，根据词性标注，利用协陪义动词词表发现协陪义候选句；其次，利用依存句法分析，根据协陪义动词在句中充当的成分，对协陪义候选句进行句型判断和分类；接着利用依存句法分析，针对五种不同的协陪义候选句型设计相应的协陪义成分识别算法；最后，根据现有关系抽取系统抽取到的显式关系，利用协陪义隐式关系推理规则，对基于协陪义动词的隐式实体关系进行推理.

由于协陪义中的隐式实体关系需要借助显式实体关系进行推理，则显式实体关系抽取的不准确性会带来级联错误，因此，显式实体关系抽取的性能较为关键.对于包含“到、来、去…”等趋向动词的句子，文献[5]很难提取到有效区分实体关系类型的最近句法依赖动词特征.为了解决该问题，本文在文献[5]的最近句法依赖动词特征基础上，进一步提出了趋向核心动词特征提取算法.

本文的主要创新工作包括：

(1)利用依存句法分析，设计了协陪义候选句型分类算法以及相应的协陪义成分识别算法.

(2)根据实体在协陪义句中充当的成分不同以及协陪义动词作用的范围，设计了三种句内基于协陪义动词的隐式实体关系推理规则，有效地解决了句内基于协陪义动词的隐式实体关系抽取问题.

(3)从零形回指的角度，利用零形回指的先行词建立起不同句子中主体成分与客体成分之间的桥梁，利用客体成分的显式关系，对主体成分的隐式关系进行推理，有效地解决了句间基于协陪义动词的隐式实体关系抽取问题.

(4)提出了趋向核心动词特征的提取算法，能有效提取到真正表示实体关系类型的动词，有利于进一步提高动词特征对显式关系抽取性能的影响.

(5)以旅游领域和新闻领域为数据源，通过详细的对比分析，验证了基于协陪义动词的句型和成分识别方法的有效性，同时也验证了基于协陪义动词的隐式关系推理规则以及趋向核心动词特征提取算法的有效性.

本文第 2 节介绍相关工作；第 3 节介绍基于协陪义动词的中文隐式实体关系抽取整体框架；第 4 节介绍基于协陪义动词的句型分类和成分识别方

法;第 5 节重点介绍基于协陪义动词的隐式关系推理;第 6 节提出趋向核心动词特征提取算法;第 7 节为实验及分析部分;最后部分是总结.

2 相关工作

2.1 显式实体关系抽取相关研究

现有的实体关系抽取相关研究大多集中于显式关系抽取.显式关系抽取所遵循的技术方法主要有两类:基于知识工程的方法和基于机器学习的方法.目前基于机器学习的方法已经成为关系抽取研究的主流思路.根据人工参与和对标注语料的依赖程度的不同,基于机器学习的关系抽取方法可以分为四大类:有监督的关系抽取方法、无监督的关系抽取方法、半监督的关系抽取方法和基于深度学习的关系抽取方法.

(1)基于特征向量的关系抽取方法

关系抽取通常采用有监督的机器学习方法,主要包括基于特征向量的方法和基于核函数的方法.目前,基于特征向量的关系抽取取得了较好的成效.由于特征选择对关系抽取的性能影响很大,因此基于特征向量的实体关系抽取方法的关键在于如何从自由文本及其句法结构中获取有效的特征.其中比较成熟的算法为最大熵模型和支持向量机.

现有基于特征的关系抽取相关研究多集中在显式实体关系的抽取[1,5-10].文献[1]在 ACE 关系抽取任务中首次使用基于特征的方法,采用了最大熵模型与大量特征相结合,包括词汇特征、句法特征和实体类型信息等特征;该研究表明多个层次的语言学特征能够提升关系抽取的效果.一个更大的特征集,包括词列表和更广泛的词级特征集,结合支持向量机模型,在 $F1$ 值上得到了改善[6].

传统基于特征的实体关系抽取方法只考虑句子本身的各种特征,而没有考虑句子之间的联系,文献[7]提出了概念模型,并在此基础上获取有效的空间特征,该特征不仅能获取句子本身内在的信息,而且能提取句子之间的语义信息关联;实验表明通过使用新的特征,关系抽取的精度和召回率得到了显著的提高.

文献[8]提出了将依存句法关系特征、核心谓词特征和语义角色标注特征应用于实体关系抽取,实验结果证明了该方法的有效性.

动词特征对于实体关系抽取的贡献较大,能够有效地提高关系抽取的准确率和召回率.文献[9]将祖先成分、实体间的路径、依赖动词以及实体到依赖动词的路径等特征应用于非包含关系的关系抽取任务中;实验表明依赖动词特征在较大程度上提高了实体关系抽取的性能.然而,文献[9]的依赖动词特征并非都能提取到真正表征该实体对关系类型的动词,给关系抽取带来大量噪音.

为了改善动词特征,文献[5]提出了一种基于句法语义特征的实体关系抽取方法,提取了依存句法关系组合特征和最近句法依赖动词特征,并将这两个特征应用于实体关系探测和关系抽取任务中;实验表明加入这两个新特征,实体关系探测和关系抽取的正确率和召回率都得到了较大的提升.但是,对于包含"到、来、去…"等趋向动词的句子,文献[5]很难提取到有效区分实体关系类型的最近句法依赖动词特征.

(2)基于深度学习的关系抽取方法

由于深度学习(Deep Learning)能够从大规模的训练数据中自动地学习分类特征,最近很多研究人员开始将深度学习的技术应用到关系抽取中,其目的在于减少人工参与特征的选择[11-14].

Socher 等人[11]和 Zeng 等人[12]首次提出利用深度神经网络(Deep Neural Network,DNN)解决关系分类问题.Socher 等人[11]利用递归神经网络(Recursive Neural Network,RNN)进行关系分类.该方法首先对句子进行句法解析,然后为句法树上的每个节点学习向量表示.通过 RNN,可以从句法树最低端的词向量开始,按照句子的句法结构迭代合并,最终得到该句子的向量表示,并用于关系分类.该方法能够有效地考虑句子的句法结构信息,但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息.Zeng 等人[12]采用卷积神经网络(Convolutional Neural Network,CNN)进行关系抽取.该方法采用词汇向量和词的位置向量作为卷积神经网络的输入,通过卷积层、池化层和非线性层得到句子表示.通过考虑实体的位置向量和其他相关的词汇特征,句子中的实体信息能够被较好地用到关系抽取中.

Yu 等人[13]提出了因子组合嵌入模型(Factor-based Compositional Embedding Model,FCM),利用依赖关系树和命名实体,从词向量中获取句子级和子结构级向量.Zhang 等人[14]使用 NLP 工具和词典资源获取词、位置、词性、命名实体、依存句法分析和上下文等特征,结合长短期记忆(Long Short-Term Memory,LSTM)网络进行关系分类.

上述大多数研究为了达到最优性能，在深度学习方法中结合了自然语言特点或词汇特征，如句法分析树、WordNet、词性和命名实体等.

因此，为了更少地使用外部资源和工具，Santos等人[15]提出了一种新的卷积神经网络进行关系抽取，其中采用了新的损失函数，能够有效地提高不同关系类别之间的区分性. 该方法仅仅将词向量作为特征输入即可达到最优性能. Zhou 等人[16]提出了一个基于注意力机制的双向长短期记忆(Attention-based Bidirectional Long Short-Term Memory，Att-BLSTM)网络来捕获句子中最重要的语义信息，该网络模型没有使用任何从词典资源和 NLP 系统获取的特征. Qin 等人[17]采用 CNN 自动控制从原始句子中进行特征学习，减少使用从工具和资源获得的外部特征.

上述基于深度学习的关系抽取方法都采用标准数据集 SemEval-2010 Task 8 进行实验，其抽取性能均比传统方法更好. SemEval-2010 Task 8 是为了识别一个句子中给定名词对之间的语义关系所建立的数据集. 该数据集为英文标准数据集，包含 10717 个标注样例，其中 8000 个训练实例和 2717 个测试实例；共有 9 类关系，如原因-影响、工具-代理、产品-生产商、内容-容器、实体-出生、实体-目的地、成分-整体、成员-集体、消息-主题，以及一个其他类[18]. 深度学习需要大规模的标注训练数据进行特征学习，而且深层模型难以优化，容易造成过度拟合①. 然而，目前在中文实体关系抽取方面，依然缺乏大规模的标准中文数据集，因此利用深度学习进行中文实体关系抽取的研究非常少.

综上所述，基于机器学习的关系抽取方法在标注语料资源充足的情况下能获取较好的效果，特别是将语义关系丰富的背景知识库作为训练样本后，能大大提升抽取模型的覆盖范围. 虽然基于机器学习的关系抽取方法比较前沿和主流，但其效果和应用场景都有天花板，特别是应用于工业界的具体领域时，依然需要人工干预，通过进一步结合规则的方法来提高关系抽取的性能.

2.2　隐式实体关系抽取相关研究

相关研究大多都集中于显式实体关系抽取，而很少关注隐式实体关系抽取，其原因在于[1-2,6]：

(1) 隐式关系标注结果难以达成一致. 在隐式实体关系标注过程中，由于标注者自身存在主观性，造成不同标注者的标记结果之间存在歧义性. 因此，准确标注出隐式关系对于标注者来说是个很大的挑战.

(2) 缺少隐式实体关系标注数据集. 对于已有的评测会议，在公开的评测数据集中标注的显式实体关系数量很多，而隐式实体关系数量很少.

(3) 隐式实体关系推理更难. 与显式实体关系相比，隐式实体关系由于缺少支持具体关系类型的直接证据，通常需要额外的知识或不同的系统帮助其关系推理，难度更大. 目前显式实体关系抽取方法无法很好地检测出隐式关系，因此不能直接用于解决隐式实体关系抽取问题，这在英语和汉语中是普遍存在的.

(4) 无法采用通用的方法来实现统一建模. 自然语言灵活多变，文本中蕴含的隐式实体关系类型繁多，不同关系类型的分布特点也不一样，采用额外知识和方法进行推理的方式也有所不同，因此很难采用通用的方法来实现统一建模.

文献[2]提出了结合概率抽取模型和数据挖掘方法从 Wikipedia 人物传记文本中发现隐式实体关系和模式，从而提高了实体关系抽取的性能. 文献[19]提出结合马尔可夫逻辑网和一阶逻辑方法提取维基百科中的实体关系，采用一阶逻辑范式发现隐式实体关系. 文献[20]提出了上下文依赖的知识图谱嵌入模式，该模式考虑了两种连通模式类型，即分别提取隐式实体关系的上下文连通模式(CCPS)和显式实体关系的局部连通模式(LCPS)，该方法用于链接预测和元组分类，实验结果表明该方法与经典方法对比，在链接预测和元组分类上的性能得到了显著和连续的提高. 文献[21]提出了基于规则的方法识别阿拉伯人名和组织实体之间的功能关系类型，而该方法只简单考虑了“属于类型”隐式实体关系的提取.

上述相关研究主要关注于欧洲语料库，特别是英语[2,19-20]. 由于缺乏具有影响力的大规模中文隐式关系语料库，限制了中文隐式实体关系抽取研究的发展. 并且由于中文语言表达自由灵活，主体或客体经常可以省略，导致中文隐式关系更难推理. 此外，中文句型复杂多样，中文长句经常对应英语中的多个句子. 两种语言之间的实体关系表示为隐式和显式也会有不匹配性，语义关系存在着歧义性差异. 因此，中文和英文的语言特点存在巨大的差异性，使得面向英语的隐式实体关系抽取方法不能直接应用于中文隐式实体关系抽取上.

① http://blog.sina.com.cn/s/blog_4caedc7a0102wkm9.html

综上所述，隐式实体关系抽取任务比显式实体关系抽取任务更复杂、推理难度也更大.相对于英语来说，中文隐式实体关系抽取仍然是目前实体关系抽取领域的一项既困难又富有挑战性的工作.

3 整体框架

本文提出了一种基于协陪义动词的中文隐式实体关系抽取方法，借助显式实体关系对隐式实体关系进行推理.本文研究的整体框架如图 2 所示.

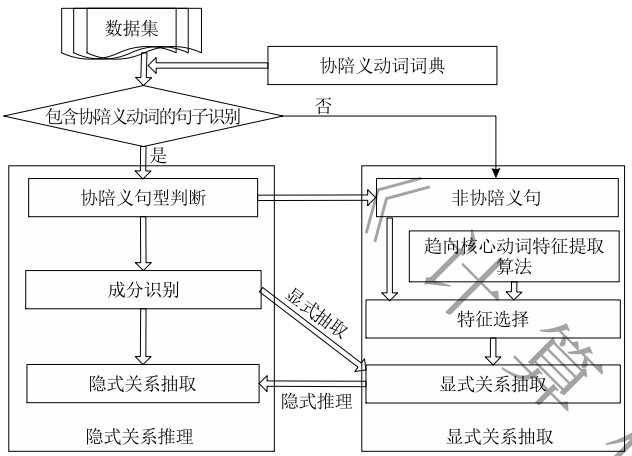

图 2 基于协陪义动词的中文隐式实体关系抽取框架

本文研究框架主要分为两部分：

(1) 显式关系抽取.在已有的最佳特征基础上，加入本文提出的趋向核心动词特征，利用 SVM 分类器进行显式关系抽取.

(2) 隐式关系推理.首先利用协陪义动词词表对句子进行筛选，得到协陪义候选句；接着根据协陪义动词在句中充当的成分不同，对协陪义候选句型进行分类；然后进一步利用依存句法分析，针对不同的协陪义候选句型，设计了相应的协陪义成分识别算法；最后设计了句内和句间基于协陪义动词的隐式关系推理规则，借助显式实体关系对隐式实体关系进行推理.本文将规则与机器学习的方法相结合，有效地解决了基于协陪义动词的隐式关系抽取问题，从而更准确地发现更多的实体关系，提高了中文实体关系抽取的性能.

4 协陪义句型分类和成分识别

协陪义指的是某一个体在另一个体的协助、陪同、带领、护送下共同进行某种活动，可以用协陪义动词和副词表示[4].协陪义动词也可以与协同副词①共现.很多协同副词一般不直接修饰协陪义动词，而是位于协陪义动词后面的动词之前，强调说明某一个体是在另一个体的协同下一起施行某一动作.协同副词的使用是一种积极的羡余现象②，目的在于突显句子的焦点和重心.协陪义句中一般都包含协陪义动词或副词，由于所有的协陪义动词都可与协同副词共现，因此本文只需考虑协陪义动词即可.

4.1 包含协陪义动词的候选句识别

协陪义动词要求其主体语义成分具有[±有意识]、[+有生命度]、[+施动性]的语义特征.由于企业、学校、部队、机关、党派、国家等组织、机关单位、社会团体都是由人组成，并且通过人来执行一系列动作的，也可看作具有这些语义特征.因此，概括为组织或团队均可充当主体成分.本文通过对数据集的分析，发现旅游和新闻领域数据集中存在很多具有协陪义的句子，且本文关注的人物/组织实体(PO)的性质正好与文献[4]中要求动作主体具有的性质相符合.

协陪义动词的两个协同参与者有主次之别，如果前者是动作的主动者，后者则是从动者，反之亦然.根据协陪义动词的主体成分和客体成分之间的关系，将协陪义动词分为两大类.

(1) I 类.表示主体成分和客体成分一同施行某种动作行为，主体成分和客体成分共同的活动方向、目的是完全一致的.主体成分是主动者，客体成分是从动者的协陪义动词.如例 5 所示.该类型动词具有“带领”类意义，具体的协陪义动词如表 1 中的 I 类所示.

例 5. 1993 年 10 月 16 日，塞舌尔卫生部副部长率卫生代表团，到泰安访问，并游览泰山.

(2) II 类.表示主体成分和客体成分一同施行某种动作行为，主体成分是从动者，客体成分是主动者的协陪义动词.根据从动者和主动者之间的关系，又可将协陪义动词细分为以下 3 个子类.

①“护送”类.强调客体成分是在主体成分的护送下共同完成动作.具体的“护送”类协陪义动词如表 1 所示.

②“协助”类.强调客体成分是在主体成分的协助下共同完成动作.具体的“协助”类协陪义动词如

① 常见的协同副词有“共、共同、齐、同、一道、一块(儿)、一路、一齐、一起、一同”等副词.

② 在现代汉语里，某一个语言组合中有多余的成分而不视为“赘疣”，称为语言的“羡余”(即“余剩”)现象.这一术语最早是由语言学界老前辈赵元任先生在《中国话文法》中提出的.他说，“虽然如此”中的“然”即“如此”，这就是“羡余”.

表 1 所示.

③“陪同”类. 强调客体成分是在主体成分的陪同下共同完成动作. 如例 2 和例 3 所示. 具体的“陪同”类协陪义动词如表 1 所示.

表 1　协陪义动词种子词表

大类型	子类型	具体动词(共 30 个)
Ⅰ类	带领类	带领、带队、带、领导、领、率、率领、引、引导、拉、牵
Ⅱ类	陪同类	伴同、伴、陪同、随同、陪伴、陪、携
	护送类	搀、搀扶、扶、护送、送
	协助类	帮、帮助、臂助、辅助、襄助、协同、协助

表 1 中的协陪义动词是由已有协陪义相关研究总结出来的[3-4]. 为了增加协陪义动词的完备性，本文将表 1 中的协陪义动词作为种子词进行扩展. 首先以表 1 中的协陪义动词作为种子词，利用哈工大信息检索研究中心同义词词林扩展版进行同义词扩展，得到候选扩展词；然后再进行人工判断和筛选，对每一个候选扩展词，利用《新华字典》判断它是否存在协陪义，若存在，则进一步利用“百度”搜索引擎，将该候选扩展词作为关键字进行网页搜索，从搜索网页结果中随机选取 30 个包含该候选扩展词的句子进行判断，若 90% 的句子都符合协陪义句型且具有协陪义，则将该候选扩展词作为协陪义动词保留下来；最终得到协陪义动词扩展词表，如表 2 所示.

表 2　协陪义动词扩展词表

大类型	子类型	具体动词(共 58 个)
Ⅰ类	带领类	引领、统率、统领、领队、提挈、带路、引路、领路、领道、向导、先导、前导、指引、导、指路、携带、随带、拖带、挈、捎、挟带、拖住、拉住、携手、拖、拖曳
Ⅱ类	陪同类	伴随、奉陪、做伴、作陪、为伴、相伴、作伴
	护送类	扶掖、牵引、挽
	协助类	扶助、赞助、援助、救助、匡助、相帮、帮忙、帮衬、支援、增援、扶持、扶掖、提携、匡扶、受助、帮带、佑助、鼎力相助、相助、帮扶、协、助

本文将协陪义动词种子词表 1 和协陪义动词扩展词表 2 进行合并，简称为协陪义动词词表(共 88 个). 为了验证协陪义动词词表的完备性，我们从百度新闻网站中随机选择了 2158 个协陪义句子进行核对，其中有 2040 个句子中的协陪义动词出现在协陪义动词词表中，即协陪义动词词表的覆盖率达到了 94.5%. 如果有必要的话，还可以采用基于“知网”词汇语义相似度计算等方法进行进一步扩展.

由于本文的目标是对具有协陪义动词的协陪义句子进行隐式实体关系抽取. 因此，首先应该利用协陪义动词词表对已分词和词性标注的句子进行协陪义候选句识别. 识别条件为：协陪义候选句中至少包含“协陪义动词词表”中的一个词，且该词的词性为动词；否则为非协陪义句.

4.2　协陪义句型判断

由于本文关注于中文旅游和新闻领域中名人或组织在某景点或地点发生的事情，而不考虑同类实体之间的关系. 因此，结合协陪义动词构成的句型以及旅游和新闻领域的数据特点，归纳出旅游和新闻领域中协陪义动词构成的基本句型主要有 4 种，分别为谓语、并列、状语和定语句型，如表 3 所示.

(1) $S+V_x+O_1+V+O_2$ 句型(简称为谓语句型). 这是协陪义动词构成的最基本句型，一般协陪义动词 V_x 在句中做谓语，充当 HED 成分.

(2) $S_1+V+O_1+S_2+V_x$ 句型(简称为并列句型). 该句型通常是至少由两个分句构成，形成并列关系. 协陪义动词 V_x 一般不出现在第一个分句里.

(3) $S+\mathrm{ADV}+V+O$ 句型(简称为状语句型). 协陪义动词基本上都可以构成“由”字句和“在…下”字句. 该句型使受事者前移至句首，充当话题，而施事者则由话题变成了对比焦点，重在表述对施事者的依赖，强调依赖于施事和取决于施事，施事者的地位相当重要，因而“由”字句和“在…下”字句后的施事者不能省略. 这类句型在旅游领域比较多，突出人物或组织实体(PO)的重要地位. 协陪义动词 V_x 主要出现在状中结构 ADV 的介宾关系 POB 中.

表 3　旅游和新闻领域协陪义动词构成的句型分类

类型	句型	符号表示	例句
谓语句型	主语＋协陪义动词＋宾语$_1$＋谓语动词＋宾语$_2$	$S+V_x+O_1+V+O_2$	张木良陪同陈先生到北京访问.
并列句型	主语＋谓语动词＋宾语$_1$＋主语＋协陪义动词	$S_1+V+O_1+S_2+V_x$	中国奶业协会赴昌平吉康考察，刘芳陪同.
状语句型	主语＋状语＋谓语动词＋宾语	$S+\mathrm{ADV}+V+O$	怀颖由董事长李继武陪同前往链景科技参观. 李金早在董事长程天富的陪同下参观了甘什黎村.
定语句型	定语＋主语＋谓语动词＋宾语	$\mathrm{ATT}+S+V+O$	由李希率领的经贸代表团访问俄罗斯.
混合句型	至少由 2 种基本句型构成	-	由张丰率领的代表团，在王浩的陪同下，考察了红旗小学.

(4) ATT+S+V+O 句型(简称为定语句型). 为了凸现各个协同参与者之间的关系及其不同的地位,协陪义动词 V_x 会出现在定中关系 ATT 中.

旅游和新闻领域文本信息的句子多为长句,一个句子中通常会包含多个人物/组织(PO)或景点/地点(TA)实体,且句式较复杂,因此可以由 4 种基本句型进行组合构成混合句型,如表 3 所示. 混合句型中通常存在多个协陪义动词,协陪义动词分别在不同句型中充当不同的成分.

根据协陪义动词在句中充当的不同成分,利用依存句法分析,对协陪义候选句型进行判断,如算法 1 所示.

算法 1. 协陪义动词构成的句型判断.

输入:一个句子中包含协陪义动词的列表 *VXList*,以及该句的依存句法分析和词性标注结果

输出:该句的句型标志 *SL*

```
Count=VXList.size() //句中包含协陪义动词的个数
AL="" //置句中协陪义动词句型组成的字符串为空
FOR (Vx∈VXList) //判断句中每个协陪义动词成分
  IF (Vx->relate 为 SBV 且 Vx 为 HED 成分)
    L='S' //该协陪义动词构成谓语句型
  ELSE IF (Vx->relate 为 SBV 且 Vx 为 COO 关系)
    L='C' //该协陪义动词构成并列句型
  ELSE IF (Vx 出现在 ADV 状中结构中)
    L='D' //该协陪义动词构成状语句型
  ELSE IF (Vx 出现在 ATT 定中关系中)
    L='T' //该协陪义动词构成定语句型
  ENDIF
  AL=AL+ L //产生协陪义动词句型的字符串
ENDFOR
//对句中所有协陪义动词构成的句型进行判断
IF (INStr(AL,'S')==Count)
    SL='S' //该句中的协陪义动词均构成谓语句型
ELSE IF(INStr(AL,'C')==Count)
    SL='C' //该句中的协陪义动词均构成并列句型
ELSE IF(INStr(AL,'D')==Count)
    SL='D' //该句中的协陪义动词均构成状语句型
ELSE IF(INStr(AL,'T')==Count)
    SL='T' //该句中的协陪义动词均构成定语句型
ELSE
    SL='M' //该为混合句型
ENDIF
```

4.3 协陪义成分识别

由协陪义动词构成的句型中必须出现主体和客体两个强制性的语义角色,缺一不可. 由于旅游和新闻领域中句子一般较长,句子结构比较复杂,导致语义角色分析准确性不高. 此外,本文关注于借助显式实体关系对协陪义句中的隐式实体关系进行推理. 因此,应首先对在句中充当主干成分的 PO 实体进行显式实体关系抽取,这样更有利于提高隐式实体关系推理的性能.

基于上述考虑,在协陪义句中,本文不考虑句中实施者和受事者的语义角色,而是从整个句子的句法结构出发,通过依存句法分析,将参与协陪义动作的 PO 实体进行成分识别. 根据 PO 实体与协陪义动词以及与 HED 成分的关系,将参与协陪义动作的 PO 实体分成 3 类.

(1) 核心实体 e_h. 该实体参与协陪义动作,在依存句法分析中与 HED 成分发生依存关系,一般在句中充当 SBV 或 FOB 等主要成分. 核心实体 e_h 与 TA 实体一般发生显式关系,可以通过已有的抽取系统进行关系抽取.

(2) 主要实体 e_f. 在一个句中,经常存在多个与核心实体 e_h 发生 COO 并列关系的 PO 实体,该类 PO 实体称为主要实体 e_f,与核心实体 e_h 具有同等地位,也参与协陪义动作的发生.

(3) 次要实体 e_s. 在协陪义动作中要求至少同时出现两个 PO 实体,除了上述核心实体 e_h 和主要实体 e_f 之外,还必须至少出现一个 PO 实体作为次要实体 e_s. 一个协陪义句中可能有多个次要实体 e_s,通常次要实体 e_s 之间也存在 COO 并列关系.

例 6. "高益槐在史妍嵋和陈中平的陪同下出席座谈会."的词性标注和依存句法分析如图 3 所示.

在图 3 中,"陪同"为协陪义动词 V_x,PO 实体"高益槐"与该句 HED 成分"出席"发生了 SBV 主谓关系,故为核心实体 e_h. 句中 PO 实体"史妍嵋"作为协陪义动作的另外一个参与实体,故为次要实体 e_s,而 PO 实体"陈中平"与"史妍嵋"发生 COO 并列关系,因此"陈中平"也为次要实体 e_s.

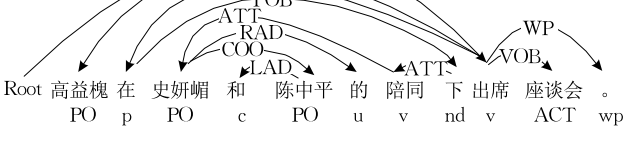

图 3 例 6 的词性标注和依存句法分析图

根据"由协陪义动词构成的句型要求必须出现核心实体和次要实体"的条件可知,如果一个句子虽然包含了协陪义动词,但无法同时找到核心实体和次要实体,则该句不具有协陪义,故划为非协陪义句. 因此,需要对协陪义候选句型进行筛选,得到真

正具有协陪义的句子，称为协陪义句.

例 7.　“徐凤翔带队赴中国药科大学考察.”的词性标注、依存句法分析如图 4 所示.

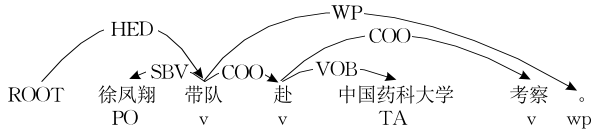

图 4　例 7 的词性标注和依存句法分析图

通过上述协陪义句型判断算法可知，例 7 为协陪义候选句中的谓语句型. 在图 4 中，“带队”为协陪义动词，PO 实体“徐凤翔”与句中 HED 成分“带队”具有 SBV 主谓关系，故“徐凤翔”为核心实体，而在该句中只有 1 个 PO 实体“徐凤翔”，因此缺少次要实体，不能满足协陪义句对 PO 实体数量和角色的要求，故该句为非协陪义句.

对协陪义候选句型进行筛选以及对协陪义句进行成分识别的具体步骤为：① 首先利用“协陪义动词词表”在句中找出所有协陪义动词 V_x；② 对每一个协陪义动词 V_x，找出与依存句法分析中 HED 成分直接或间接发生依存关系的核心实体 e_h；③ 找出与核心实体 e_h 发生 COO 并列关系的主要实体 e_f，构成主要实体集 $FENSet$；④ 找出次要实体 e_s，构成次要实体集 $SENSet$；⑤ 若核心实体 e_h 未找到或次要实体集合 $SENSet$ 为空，则该句为非协陪义句，应从协陪义候选句集合中过滤掉.

综上所述，协陪义动词在不同句型中充当的成分不同，导致对参与协陪义动作的 PO 实体进行成分识别的方法也有所不同，具体分析如下：

(1) 谓语句型 $S+V_x+O_1+V+O_2$. 核心实体 e_h 和主要实体 e_f 一般出现在协陪义动词 V_x 的左边. 核心实体是与该协陪义动词发生某种直接或间接依存关系的第一个 PO 实体，主要实体 e_f 与核心实体 e_h 是 COO 并列关系. 次要实体 e_s 一般出现在协陪义动词 V_x 的右边，与该协陪义动词也发生某种直接或间接依存关系.

例 8.　“李继武和王民星陪同怀颖参观链景科技.”的词性标注、依存句法分析如图 5 所示.

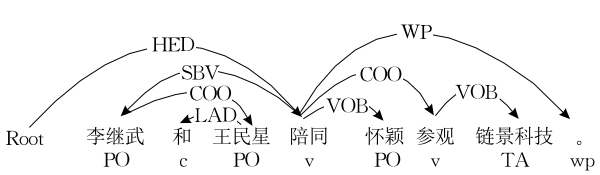

图 5　例 8 的词性标注和依存句法分析图

在图 5 中，该句是由协陪义动词“陪同”构成的谓语句型. PO 实体“李继武”和“王民星”均出现在协陪义动词“陪同”的左边，实体“李继武”与 HED 成分“陪同”构成主谓关系 SBV，因此“李继武”为核心实体；“王民星”与核心实体“李继武”为并列关系 COO，故“王民星”为主要实体；而另外一个 PO 实体“怀颖”在协陪义动词“陪同”的右边，与“陪同”构成动宾关系 VOB，充当“陪同”的宾语，因此“怀颖”被识别为次要实体.

(2) 并列句型 $S_1+V+O_1+S_2+V_x$. 该句型通常至少由两个分句构成. 动词 V 为 HED 成分，一般先于协陪义动词出现在分句中. 而由协陪义动词 V_x 构成的分句通常在由动词 V 构成的分句之后出现，且与 V 为 COO 并列关系. 核心实体 e_h 与主要实体 e_f 一般出现在包含 HED 成分 V 的分句中，且它们之间为 COO 并列关系. 核心实体 e_h 是与动词 V 发生某种直接或间接依存关系的第一个 PO 实体，充当主语成分. 次要实体出现在由协陪义动词 V_x 构成的分句中，一般位于协陪义动词 V_x 的左边，充当主语成分. 如例 9 所示.

例 9.　“中国奶业协会赴昌平吉康考察，刘芳陪同.”的词性标注、依存句法分析如图 6 所示.

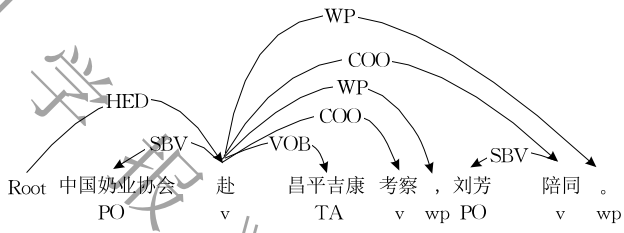

图 6　例 9 的词性标注和依存句法分析图

在图 6 中，该句由 2 个分句构成. 动词“赴”为 HED 成分，出现在第一个分句中，协陪义动词“陪同”位于第二个分句中，且与 HED 成分“赴”为 COO 并列关系. PO 实体“中国奶业协会”位于 HED 成分“赴”所在的分句中，且与“赴”具有 SBV 主谓关系，因此，PO 实体“中国奶业协会”为核心实体. 而 PO 实体“刘芳”出现在包含协陪义动词“陪同”的分句中，位于“陪同”的左边，与“陪同”具有 SBV 主谓关系，故为次要实体.

(3) 状语句型 $S+\mathrm{ADV}+V+O$. 协陪义动词 V_x 主要出现在状中结构 ADV 的“由”或“在…下”构成的介宾关系 POB 成分中. 因此，在“由”或“在…下”构成的 POB 成分中，次要实体 e_s 均出现在协陪义动词左边. 根据状语出现的位置不同可分为 2 类：

① 一般状语. 即状语出现在主语之后，HED 成分的谓语之前；

② 句首状语. 即状语是放在主语的前面,这是状语的特殊位置.

由于核心实体 e_h 和主要实体 e_f 均可在协陪义动词左右两边出现,因此无需考虑核心实体 e_h 和主要实体 e_f 的位置,只需考虑它们在句中充当的成分即可. 核心实体 e_h 与 HED 成分发生依存关系,一般具有 SBV 或 FOB 关系. 例如,例 6 是由协陪义动词"陪同"构成的状语句型,状语出现在主语"高益槐"之后,HED 成分的谓语"出席"之前,故为一般状语. 若将该句中状语调到主语前面,做句首状语,如例 10 所示.

例 10. "在史妍嵋和陈中平的陪同下,高益槐出席座谈会."的词性标注、依存句法分析如图 7 所示.

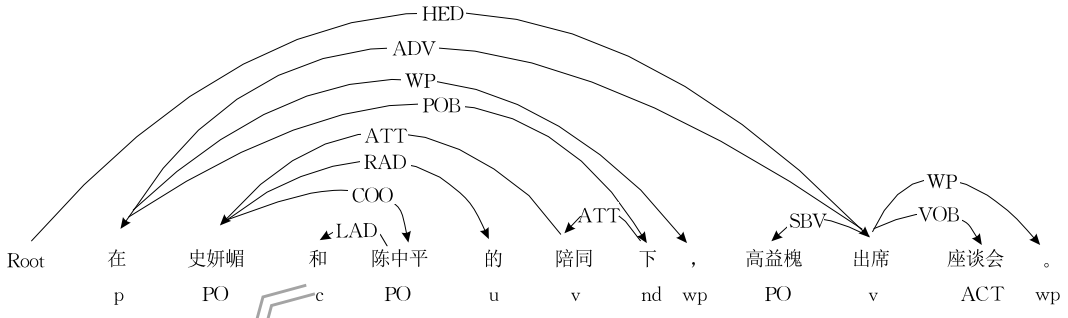

图 7 例 10 的词性标注和依存句法分析图

在图 7 中,包含协陪义动词"陪同"的状语部分出现在句首;PO 实体"史妍嵋"和"陈中平"均出现在包含协陪义动词"陪同"的介宾关系 POB 中,因此"史妍嵋"和"陈中平"皆为次要实体 e_s. 句中 PO 实体"高益槐"为 SBV 成分,他是与句中 HED 成分发生依存关系的实体,故"高益槐"识别为核心实体. 因此,将该句中包含协陪义的状语放在句首,其成分识别的结果依然与例 6 的结果一致.

(4) 定语句型 ATT+S+V+O. 协陪义动词 V_x 常出现在定中关系 ATT 中,修饰核心实体 e_h. 核心实体 e_h 与主要实体 e_f 一般位于协陪义动词 V_x 的右边,而次要实体 e_s 则位于 V_x 左边. 如例 11 所示.

例 11. "由李希率领的经贸代表团访问俄罗斯."的词性标注、依存句法分析如图 8 所示.

在图 8 中,协陪义动词"率领"修饰 PO 实体"代表团","代表团"位于"率领"的右边,且与 HED 成分"访问"具有 SBV 关系,因此"代表团"为核心实体. 而PO实体"李希"位于"率领"的左边,通过

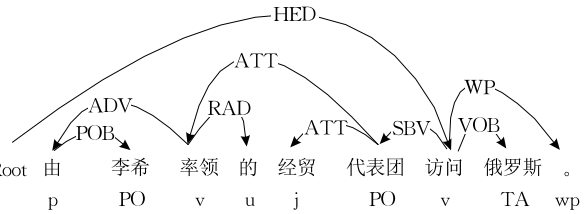

图 8 例 11 的词性标注和依存句法分析图

"由"字与"率领"发生间接的依存关系,故为次要实体.

(5) 混合句型. 该类句型至少由上述 2 种基本句型嵌套构成,因此至少包含 2 个或 2 个以上协陪义动词. 首先根据每个协陪义动词在混合句中充当的成分不同,将混合句型拆分为上述相应的基本句型,然后利用基本句型的成分识别方法进行识别. 如例 12 所示.

例 12. "由张丰率领的代表团,在王浩的陪同下,考察了红旗小学."的词性标注、依存句法分析如图 9 所示.

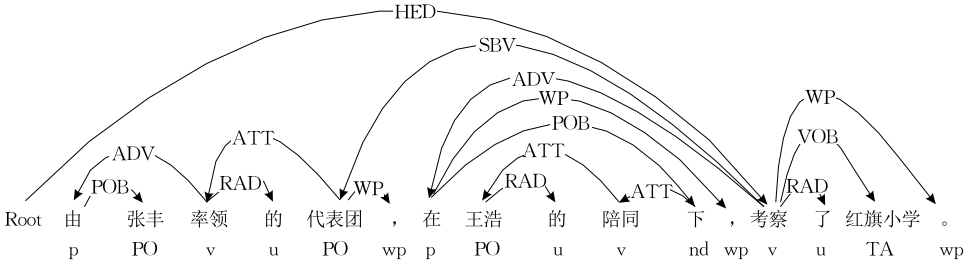

图 9 例 12 的词性标注和依存句法分析图

在图 9 中，该句包含 2 个协陪义动词“率领”和“陪同”. 第一个协陪义动词“率领”在句中是定语成分，故构成了定语句型；第二个协陪义动词“陪同”在句中充当状语成分，故构成了状语句型. 由此可见，该混合句是由定语句型和状语句型这 2 个基本句型组合构成. 因此，在成分识别时，对协陪义动词“率领”和“陪同”分别利用定语句型和状语句型的成分识别方法. 由于一个句子中只有一个 HED 成分，故与 HED 成分发生 SBV 或 FOB 依存关系的核心实体 e_h 也只有一个. 因此，混合句型中的每个协陪义动词一般共享同一个核心实体 e_h. 对于协陪义动词“率领”和“陪同”，它们的核心实体均为“代表团”. 而它们各自的次要实体则需根据其句型进行识别. 由“率领”构成的定语句型，可利用定语句型的成分识别方法，在“率领”的左边识别出 PO 实体“张丰”为次要实体，而由“陪同”构成的状语句型，在其 POB 介宾关系下可找到 PO 实体“王浩”为次要实体.

综上所述，5 种句型的协陪义成分识别的具体算法如算法 2～算法 6 所示.

算法 2.　主谓句型协陪义成分识别.

输入：一个协陪义候选句 s，包含的协陪义动词 V_x，以及该句的依存句法分析和词性标注结果

输出：该句的核心实体 e_h、主要实体集 *FENSet*、次要实体集 *SENSet* 和非协陪义句型标志 *NX*

e_h=NULL　//核心实体 e_h 初始化为空

FENSet=∅　//主要实体集 *FENSet* 初始化为空

SENSet=∅　//次要实体集 *SENSet* 初始化为空

NX=FALSE　//非协陪义句型标志初始化为 FALSE

IF ((PO实体 e_i 出现在 V_x 左边) && (e_i 与HED成分具有依存关系) && (e_i 与 V_x 具有直接或间接依存关系))　//识别核心实体

　$e_h=e_i$　//e_i 为核心实体

ELSE IF ((e_h!=NULL) && (PO 实体 e_j 出现在 V_x 左边) && (e_j 与 e_h 为 COO 并列关系))　//识别主要实体

　FENSet∪=e_j　//e_j 为主要实体，加入到 *FENSet* 中

ELSE IF ((PO 实体 e_k 出现在 V_x 右边) && (e_k 与 V_x 具有直接或间接依存句法关系))　//识别次要实体

　SENSet∪=e_k　//e_k 为次要实体，加入到 *SENSet* 中

ELSE IF ((PO 实体 e_t 出现在 V_x 右边) && (e_t 与 e_k 具有 COO 并列关系))　//识别次要实体

　SENSet∪=e_t　//e_t 为次要实体，加入到 *SENSet* 中

ENDIF

IF ((e_h==NULL) || (*SENSet*==∅))

　NX=TRUE　//协陪义候选句为非协陪义句

ENDIF

算法 3.　并列句型协陪义成分识别.

输入：一个协陪义候选句 s，包含的协陪义动词 V_x 和动词 V，以及该句的依存句法分析和词性标注结果

输出：该句的核心实体 e_h、主要实体集 *FENSet*、次要实体集 *SENSet* 和非协陪义句型标志 *NX*

e_h=NULL　//核心实体 e_h 初始化为空

FENSet=∅　//主要实体集 *FENSet* 初始化为空

SENSet=∅　//次要实体集 *SENSet* 初始化为空

NX=FALSE　//非协陪义句型标志初始化为 FALSE

IF ((动词 V 为 HED 成分) && (V 比 V_x 早出现) && (e_i 与HED成分具有依存关系) && (e_i 与 V 位于同一个分句中且出现在 V 的左边) && (e_i 与 V 具有 SBV 关系))　//识别核心实体

　$e_h=e_i$　//e_i 为核心实体

ELSE IF ((e_h!=NULL) && (PO 实体 e_j 与 HED 成分 V 位于同一个分句中且出现在 V 的左边) && (e_j 与 e_h 为 COO 关系))　//识别主要实体

　FENSet∪=e_j　//e_j 为主要实体，加入到 *FENSet* 中

ELSE IF ((PO 实体 e_k 与协陪义动词 V_x 位于同一个分句) && (e_k 位于 V_x 左边 && (e_k 与 V_x 具有 SBV 关系))　//识别次要实体

　SENSet∪=e_k　//e_k 为次要实体，加入到 *SENSet* 中

ELSE IF ((PO 实体 e_k 与协陪义动词 V_x 位于同一个分句中) && (e_t 与 e_k 具有 COO 并列关系))　//识别次要实体

　SENSet∪=e_t　//e_t 为次要实体，加入到 *SENSet* 中

ENDIF

IF ((e_h==NULL) || (*SENSet*==∅))

　NX=TRUE　//协陪义候选句为非协陪义句

ENDIF

算法 4.　状语句型的协陪义成分识别.

输入：一个协陪义候选句 s，包含的协陪义动词 V_x，以及该句的依存句法分析和词性标注结果

输出：该句的核心实体 e_h、主要实体集 *FENSet*、次要实体集 *SENSet* 和非协陪义句型标志 *NX*

e_h=NULL　//核心实体 e_h 初始化为空

FENSet=∅　//主要实体集 *FENSet* 初始化为空

SENSet=∅　//次要实体集 *SENSet* 初始化为空

NX=FALSE　//非协陪义句型标志初始化为 FALSE

IF (PO 实体 e_i 与 HED成分具有 SBV 或 FOB 依存关系)

　$e_h=e_i$　//识别核心实体

ELSE IF ((e_h!=NULL) && (PO 实体 e_j 与 e_h 为 COO 并列关系))　//识别主要实体

　FENSet∪=e_j　//e_j 为主要实体，加入到 *FENSet* 中

ELSE IF (PO 实体 e_k 与 V_x 满足依存路径 ADV-POB(e_i)-ATT(V_x))　//识别次要实体

　SENSet∪=e_k　//e_k 为次要实体，加入到 *SENSet* 中

ELSE IF (PO 实体 e_t 与 e_k 具有 COO 并列关系)

$SENSet \cup = e_t$ //e_t为次要实体,加入到 $SENSet$ 中
ENDIF
IF ((e_h==NULL) ‖ ($SENSet$==∅))
NX=TRUE //协陪义候选句为非协陪义句
ENDIF

算法 5. 定语句型的协陪义成分识别.

输入:一个协陪义候选句 s,包含的协陪义动词 V_x,以及该句的依存句法分析和词性标注结果
输出:该句的核心实体 e_h、主要实体集 $FENSet$、次要实体集 $SENSet$ 和非协陪义句型标志 NX

e_h=NULL //核心实体 e_h 初始化为空
$FENSet$=∅ //主要实体集 $FENSet$ 初始化为空
$SENSet$=∅ //次要实体集 $SENSet$ 初始化为空
NX=FALSE //非协陪义句型标志初始化为 FALSE
IF (PO 实体 e_i 是 V_x 右边的第一个 PO 实体)
$e_h=e_i$ //识别核心实体
ELSE IF ((e_h!=NULL)&&(PO 实体 e_j 与 e_h 为 COO 并列关系)) //识别主要实体
$FENSet \cup = e_j$ //e_j为主要实体,加入到 $FENSet$ 中
ELSE IF (PO 实体 e_k 出现在 V_x 左边) //识别次要实体
$SENSet \cup = e_k$ //e_k为次要实体,加入到 $SENSet$ 中
ELSE IF (PO 实体 e_t 与 e_k 具有 COO 并列关系)
$SENSet \cup = e_t$ //e_t为次要实体,加入到 $SENSet$ 中
ENDIF
IF ((e_h==NULL) ‖ ($SENSet$==∅))
NX=TRUE //协陪义候选句为非协陪义句
ENDIF

算法 6. 混合句型的协陪义成分识别.

输入:一个协陪义候选句 s,包含的协陪义动词 V_x,以及该句的依存句法分析和词性标注结果
输出:该句的核心实体 e_h、主要实体集 $FENSet$、次要实体集 $SENSet$ 和非协陪义句型标志 NX

e_h=NULL //核心实体 e_h 初始化为空
$FENSet$=∅ //主要实体集 $FENSet$ 初始化为空
$SENSet$=∅ //次要实体集 $SENSet$ 初始化为空
NX=FALSE //非协陪义句型标志初始化为 FALSE
//混合句型的协陪义成分识别
IF ((V_x->$relate$ 为 SBV)&&(V_x 为 HED 成分))
//主谓句型,调用算法 2
$V_x=ATTCRegnization(s,V_x,DPList,POList)$
ELSE IF((V_x->$relate$ 为 SBV)&&(V_x 为 COO 关系))
//并列句型,调用算法 3
$V_x=COOCRegnization(s,V_x,DPList,POList)$
ELSE IF (V_x 出现在状中结构 ADV 中)
//状语句型,调用算法 4
$V_x=ADVCRegnization(s,V_x,DPList,POList)$
ELSE IF (V_x 出现在定中关系 ATT 中)
//定语句型,调用算法 5
$V_x=ATTCRegnization(s,V_x,DPList,POList)$
ENDIF
IF ((e_h==NULL) ‖ ($SENSet$==∅))
NX=TRUE //协陪义候选句为非协陪义句
ENDIF

5 基于协陪义动词的隐式关系推理

本文以句号、问号以及感叹号作为句子结束标记.在自建的旅游和新闻领域数据集中,由于文本中的句子主要以陈述语气为主,因此绝大部分句子都是以句号为结束标记.本文以句子为单位,对由协陪义动词触发的隐式关系进行推理,试图从句内和句间两个角度进行探讨.本文关注人物/组织(PO)与景点/地点(TA)、人物/组织(PO)与活动(ACT)以及人物/组织(PO)与作品(WOR)之间的关系抽取.

5.1 句内基于协陪义动词的隐式关系推理

对一个已经完成成分识别的协陪义句进行实体关系抽取,首先将参与协陪义动作的三类 PO 实体与该句中其它非 PO 实体组成实体对,形如{〈PO,TA〉,〈TA,PO〉,〈PO,ACT〉,〈ACT,PO〉,〈PO,WOR〉,〈WOR,PO〉}.本文按照实体在句中出现的先后顺序来构建实体对,若参与协陪义动作的 PO 实体 e_x 在句中出现在其它非 PO 实体 e_i 之前,则可构建一个实体对〈e_x,e_i〉,否则为〈e_i,e_x〉.由于参与同一个协陪义动作的核心实体 e_h、主要实体 e_f 和次要实体 e_s 均为 PO 实体,因此它们均可与句中非 PO 实体 e_i 构成实体对,分别称为核心实体对〈e_h,e_i〉或〈e_i,e_h〉、主要实体对〈e_f,e_i〉或〈e_i,e_f〉、次要实体对〈e_s,e_i〉或〈e_i,e_s〉.如例 13 所示.

例 13. "旅游协会代表团和摄影家协会代表团途经泰安,在旅游局领导的陪同下游览岱庙和泰山."的词性标注和依存句法分析如图 10 所示.

在图 10 中,共有 6 个实体,具体实体名和类型为:{旅游协会代表团〈PO〉,摄影家协会代表团〈PO〉,泰安〈TA〉,旅游局领导〈PO〉,岱庙〈PO〉,泰山〈TA〉}."陪同"为协陪义动词 V_x,通过成分识别可知,实体"旅游协会代表团"为核心实体 e_h,"摄影家协会代表团"为主要实体 e_f,"旅游局领导"为次要实体 e_s.按照实体出现的顺序,这三类协陪义成分实体与句中非 PO 实体组成的实体对集分别为

EP_h(旅游协会代表团)={〈旅游协会代表团,泰安〉,〈旅游协会代表团,岱庙〉,〈旅游协会代表团,

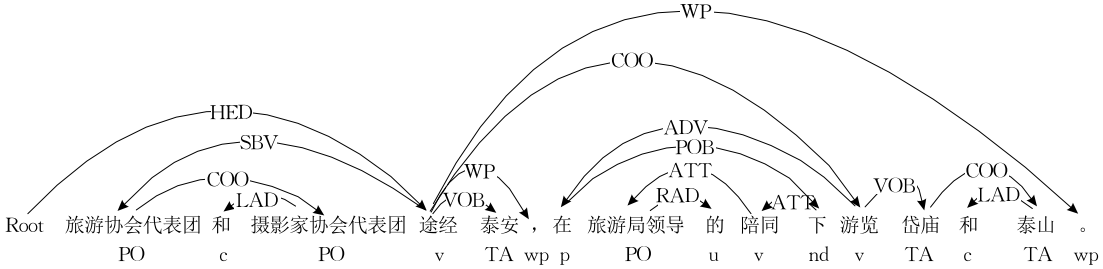

图 10　例 13 的词性标注和依存句法分析图

泰山〉}；

EP_f(摄影家协会代表团)={〈摄影家协会代表团,泰安〉,〈摄影家协会代表团,岱庙〉,〈摄影家协会代表团,泰山〉}；

EP_s(旅游局领导)={〈泰安,旅游局领导〉,〈旅游局领导,岱庙〉,〈旅游局领导,泰山〉}.

本文主要是解决句内基于协陪义动词的隐式关系抽取问题.对已经组成的实体对进行实体关系类型的判别,首先利用现有的关系抽取系统进行显式实体关系抽取,然后借助已知的显式实体关系对隐式实体关系进行推理.

根据依存句法分析,核心实体 e_h 与整个句子的 HED 成分发生 SBV 或 FOB 依存关系,因此较容易提取句中词法、句法和语义等显式特征,从而用于判定核心实体 e_h 与非 PO 实体 e_i 之间的关系类型.因此,核心实体对〈e_h,e_i〉或〈e_i,e_h〉的关系类型属于显式关系,可以通过已有的实体关系抽取系统进行抽取.如例 13 中的核心实体对集 EP_h(旅游协会代表团)={〈旅游协会代表团,泰安〉,〈旅游协会代表团,岱庙〉,〈旅游协会代表团,泰山〉}均属于显式关系,其对应的关系类型分别为:〈旅游协会代表团,泰安〉为"其他"关系(记为 OTH),〈旅游协会代表团,岱庙〉、〈旅游协会代表团,泰山〉均为"游历"关系(记为 PLA).

虽然主要实体 e_f 与句中非 PO 实体构成的主要实体对〈e_f,e_i〉或〈e_i,e_f〉也可以通过句中的显式证据抽取其关系类型.然而相对于核心实体对,主要实体对的词法、句法和语义特征没有那么直接、显式地表明其关系类型,导致对其关系抽取的性能不如对核心实体对抽取的高.由于主要实体 e_f 与核心实体 e_h 之间属于 COO 并列关系,它与核心实体 e_h 具有同等地位参与到协陪义动作中,可知主要实体对〈e_f,e_i〉或〈e_i,e_f〉与句中核心实体对〈e_h,e_i〉或〈e_i,e_h〉具有相同的关系类型.因此,本文将主要实体对的关系也作为隐式关系进行推理,通过核心实体对的关系直接推理出主要实体对的关系,具体见 CR 规则.如图 11 所示.

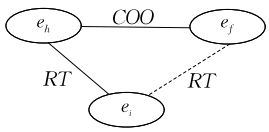

图 11　CR 规则

CR 规则.在一个协陪义句中,若已知核心实体对〈e_h,e_i〉或〈e_i,e_h〉的关系类型为 RT,则主要实体对〈e_f,e_i〉或〈e_i,e_f〉的关系类型也为 RT.

在例 13 中,如果核心实体对的关系类型已知,根据 CR 规则,则可以推理出主要实体对的关系类型,即〈摄影家协会代表团,泰安〉为"其他"关系,〈摄影家协会代表团,岱庙〉、〈摄影家协会代表团,泰山〉也均为"游历"关系.

在深层结构中,协陪义动词的次要实体与核心实体不但是协陪关系,而且通常与核心实体共同实施某一动作.因此,次要实体与主要实体也具有相同的语义特征.此外,协陪义句中存在一个隐含型空语类,该隐含型空语类(简称 P)一定会跟其他名词性词语发生照应和指代关系[4].例如,在协陪义主谓句式 $S+V_x+O_1+V+O_2$ 中,动词 V 不光受 S 支配,还受 O_1 支配,即为"多重 NP 同指".隐含型空语类补充出来即可化成式子:$S+V_x+O_1+[\mathrm{P}\ S\ O_1+V]+O_2$.显示 S 和 O_1 都参与后一个 V 结构的动作,构成了动核关系.

因此,次要实体 e_s 与句中非 PO 实体 e_i 之间能够通过协陪义动词间接或隐含地发生某种关系.它们构成的次要实体对〈e_s,e_i〉或〈e_i,e_s〉的关系,在句中没有显式证据直接支持,需要利用额外的协陪义知识帮助判定其关系,因此属于隐式关系.

为了解决句内基于协陪义动词的隐式关系抽取问题,本文利用核心实体对的显式关系,对次要实体对进行隐式关系的推理.

在不同的协陪义句型中,一个协陪义动词作用的范畴并不完全相同,具体表现为:

①在并列句型中,一个句子至少包含 2 个分

句，且协陪义动词 V_x 一般总是出现在最后分句中，该协陪义动词作用的范畴覆盖其前面分句. 因此，次要实体参与核心实体的全部活动. 如在图 6 中，次要实体“刘芳”参与了核心实体“中国奶业协会”考察“昌平吉康”的活动.

② 对于非并列句型，次要实体只有在发生协陪动作之后才能与非 PO 实体真正发生关系，而与协陪动作之前的非 PO 实体则不发生任何关系，即为“无关系”类型(记为 NON). 如在图 10 中，次要实体“旅游局领导”只是陪同核心实体“旅游协会代表团”游览了“岱庙”和“泰山”2 个景点实体，而不参与陪同核心实体“旅游协会代表团”途径“泰安”地点实体的活动.

总之，次要实体 e_s 一般都紧随协陪义动词 V_x 左右出现，因此可根据次要实体 e_s 与非 PO 实体 e_i 在句中出现的先后顺序构造次要实体对. 次要实体对 $\langle e_s, e_i\rangle$ 和 $\langle e_i, e_s\rangle$ 的关系类型可分别利用 IR 规则和 NR 规则进行推理得到，如图 12 和图 13 所示.

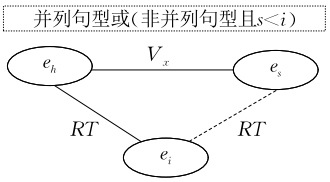

图 12 IR 规则

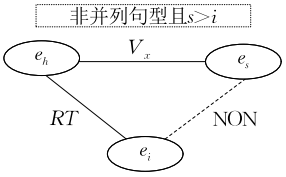

图 13 NR 规则

IR 规则. 在一个协陪义句中，已知核心实体对 $\langle e_h, e_i\rangle$ 或 $\langle e_i, e_h\rangle$ 的关系类型为 RT，次要实体为 e_s，若满足了以下条件之一，则次要实体对 $\langle e_s, e_i\rangle$ 的隐式关系类型为 RT.

① 该协陪义句为并列句型；

② 该协陪义句不为并列句型且次要实体 e_s 出现在非 PO 实体 e_i 之前.

NR 规则. 在一个协陪义句中，已知核心实体对 $\langle e_h, e_i\rangle$ 或 $\langle e_i, e_h\rangle$ 的关系类型为 RT，若该协陪义句不是并列句型且次要实体 e_s 出现在非 PO 实体 e_i 之后，则次要实体对 $\langle e_i, e_s\rangle$ 的关系类型为 NON(即“无关系”类型).

在例 13 中，该协陪义句为状语句型. 由于次要实体“旅游局领导”出现在地点实体“泰安”之后，根据 NR 规则，则可推理出次要实体对〈泰安，旅游局领导〉的关系类型为 NON. 而次要实体“旅游局领导”位于景点实体“岱庙”和“泰山”之前，根据 IR 规则，利用上述核心实体对的已知关系类型，可推理出次要实体对的隐式关系类型，即〈旅游局领导，岱庙〉、〈旅游局领导，泰山〉均为“游历”关系.

综上所述，句内基于协陪义动词的隐式关系推理的具体算法如算法 7. 在例 13 中，利用已知系统对核心实体对进行显式关系抽取，然后分别利用 CR 规则、IR 规则和 NR 规则对主要实体对以及次要实体对进行隐式关系推理，从而得到句内基于协陪义动词“陪同”的隐式关系，如图 14 所示.

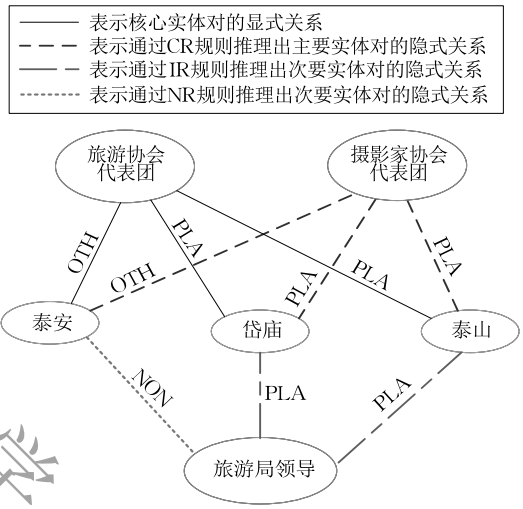

图 14 例 13 中句内基于协陪义动词的关系类型图

算法 7. 句内基于协陪义动词的隐式关系推理.

输入：一个协陪义动词 V_x 的句子 s，与该协陪义动词 V_x 相关的核心实体 e_h、主要实体 e_f 和次要实体 e_s、非 PO 实体 e_i，与该核心实体 e_h 组成的核心实体对和关系类型 $\langle e_h, e_i\rangle = RT$ 或 $\langle e_h, e_i\rangle = RT$

输出：实体对 $\langle e_i, e_f\rangle$ 或 $\langle e_f, e_i\rangle$、$\langle e_i, e_s\rangle$ 或 $\langle e_s, e_i\rangle$ 的关系类型

```
IF (e_f不为空)  //对主要实体 e_f 进行隐式关系推理
  IF (f<i)
    ⟨e_f, e_i⟩=RT  //利用 CR 规则
  ELSE
    ⟨e_i, e_f⟩=RT  //利用 CR 规则
  ENDIF
ENDIF
IF (e_s不为空)  //对次要实体 e_s 进行隐式关系推理
  IF (s 为并列句型)  //s 为并列句型
    ⟨e_i, e_s⟩=RT  //利用 IR 规则
  ELSE  //s 为非并列句型
    IF (s<i)
      ⟨e_s, e_i⟩=RT  //利用 IR 规则
```

```
        ELSE
          ⟨e_i,e_s⟩=NON  //利用 NR 规则
        ENDIF
      ENDIF
    ENDIF
```

协陪义句中存在的隐含型空语类大部分为“多重 NP 同指”，即主要实体和次要实体都参与相同的动作.但也存在少量协陪义动词，如“扶、送、帮、陪、帮助、带、拉”等，也具有[+可协同]语义特征，但由于句中动词 V 的语义特征不同，随着语境和叙述的不同，句子所表达的语法意义同样不同，使其参与意义可能相应弱化.因此导致上述隐式关系推理规则也有可能不适用.如例 14 所示.

例 14.　“朱明在李胜陪同下游览泰山，题‘一览众山小’”.

在例 14 中，动词“游览”可以是共同的行为，因此次要实体“李胜”参与了陪同主要实体“朱明”游览“泰山”.在该语境下，动词“题”理解为个人行为更恰当，因此作品“一览众山小”是由“朱明”单独完成，而“李胜”并没有参与.

因此，协陪义动词 V_x 的作用很关键，但动词 V 的作用同样不容忽视.今后研究工作将考虑把协陪义动词 V_x 与动词 V 相结合，通过对动词 V 进行某种分类，确认动词 V 与协陪义实体的同指程度，然后进一步细化和完善句内基于协陪义动词的隐式关系推理规则.

5.2　句间基于协陪义动词的隐式关系推理

已有关系抽取方法通常只考虑句子本身的特征，很少考虑句子之间的关联信息[1,8-9,23-26].上述句内基于协陪义动词的隐式关系推理方法也同样只利用了同一句子中核心实体的显式关系对主要实体和次要实体进行隐式关系推理.句内基于协陪义动词的隐式关系推理方法要求在一个句子中必须同时出现核心实体和次要实体，否则将该句作为非协陪义句直接过滤掉，而不能进入隐式关系推理阶段.如例 15～例 17 所示.

例 15.　李华陪同.

例 16.　王学力陪同调研.

例 17.　集团公司董事局主席张培良、董事长苏振佳全程陪同.

在例 15～例 17 中，每一个句子中都包含了协陪义动词“陪同”.通过协陪义成分识别可知，这 3 个句子只包含了核心实体，而缺少次要实体.由于不满足句内基于协陪义动词的隐式关系推理的条件，这 3 个句子被当作非协陪义句而直接被过滤掉，不对其进行隐式关系抽取.

事实上，虽然此类句子单从句子本身无法获取其隐式关系，但可以借助协陪义动词的语义特征，在句子之间寻找线索，帮助其隐式关系推理.如将上述 3 个句子之前的文本信息进行补充，如例 18～例 20 所示，其中 s-1 和 s-2 表示在同一个文档中的句子编号.

例 18.　〈s-1/王书坚来枣庄市调研./s-1〉〈s-2/李华陪同./s-2〉

例 19.　〈s-1/陶宏到葛店开发区调研./s-1〉〈s-2/王学力陪同调研./s-2〉

例 20.　〈s-1/天大设计院董事长洪再生到丛林考察./s-1〉〈s-2/集团公司董事局主席张培良、董事长苏振佳全程陪同./s-2〉

在例 18 的段落中，第 1 个句子 s-1 中核心实体对〈王书坚，枣庄市〉存在显式的“考察访问”关系.对于第 2 个句子 s-2，虽然也包含协陪义动词“陪同”，但由于只存在核心实体“李华”，而缺少次要实体.同理，对于例 19，在第 1 个句子 s-1 中，“陶宏”与“葛店开发区”存在显式的“考察访问”关系；而对于第 2 个句子 s-2，虽然包含协陪义动词“陪同”，但也只包含核心实体“王学力”，缺少了次要实体.对于例 20，第 1 个句子 s-1 中的“洪再生”与“丛林”存在显式的“考察访问”关系；而对于句子 s-2，虽然包含了协陪义动词“陪同”，但也只包含核心实体“张培良”和主要实体“苏振佳”，而缺乏次要实体.因此，上述 3 个句子由于协陪义动词的核心实体陪同的客体成分缺省，导致无法进行隐式关系推理.

事实上，在该语境下，可以通过文本中前面的句子将“陪同”缺省的客体成分加以补充.补充的客体成分见句子内的括号部分.如例 18 中的第 2 个句子 s-2，客体补充后为：〈s-2/李华陪同(王书坚)./s-2〉.因此，PO 实体“李华”与 PO 实体“王书坚”是“陪同”关系.又由于在 s-1 句中，〈王书坚，枣庄市〉存在显式的“考察访问”关系，因此，结合句子 s-1 和 s-2，可以推理得出 s-2 句中的 PO 实体“李华”与 s-1 句中的 TA 实体“枣庄市”也存在隐式“考察访问”关系.如图 15 所示.同理，对于例 19 和例 20 中的 s-2，将它们各自的客体成分补充后，其句子分别为：〈s-2/王学力陪同(陶宏)调研./s-2〉和〈s-2/集团公司董事局主席张培良、董事长苏振佳全程陪同(洪再生)./s-2〉.

因此，为了解决如例 18～例 20 中句子 s-2 所示

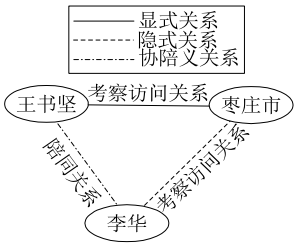

图 15 例 18 的实体关系图

的只包含协陪义主体成分句子的隐式关系抽取问题,关键在于如何借助协陪义动词的语义特征,通过上下文信息,在句子之间寻找线索,将协陪义动词的缺省客体成分补充出来.

由于通过上下文寻找到的客体成分与协陪义主体成分处于不同的句子中,因此,首先要通过已补充出来的客体成分,建立起不同句子中的主体成分与客体成分之间的桥梁,然后才能利用客体成分的显式关系对主体成分的隐式关系进行推理.

由此可见,句间基于协陪义动词的隐式关系抽取是利用协陪义动词的语义特征,通过挖掘句子之间的语义关联信息,解决只包含主体成分的协陪义句中的隐式关系抽取问题.与句内基于协陪义动词的隐式关系推理任务相比,句间基于协陪义动词的隐式关系抽取任务更加艰巨和困难.

本文拟从零形回指(也称零照应或零指代)的角度解决协陪义句中只包含主体成分的隐式实体关系抽取问题.

如例 18～例 20 中的第 2 个句子 s-2,它们均存在零形回指的现象,是省略的一种形式.陈平[27]对零形回指的定义为:"如果从意思上讲句子中有一个与上文中出现的某个事物指称相同的所指对象,但是从语法格局上看该所指对象没有实在的词语表现形式,我们便认定此处用了零形指代."

从已有汉语零形回指的研究中可知,汉语是零形回指的语言,且出现的频率很高.零形回指的使用受语篇结构、语义、语用以及主题显著等因素的影响[28].在下列情形中比较容易出现零形回指的现象:① 语篇段落中相连句子的话题涉及同一个人或事物;② 某人或某事物是整个语篇语义结构的组织核心或各句中动词、名词等成分表达的施事、动作、受事等关系能表达哪类人干哪类事,且前后句子联系明显;③ 语言使用者一般都会遵守合作原则,认为对方能确定零形回指的指称;④ 主题比较显著.

在旅游和新闻领域,为了凸显某人或某组织的重要地位或者影响,通常该人或组织比较容易成为整个篇章或段落语义结构的中心;此外,句中动词、名词等成分通常具体地描述了该人或组织在某地干了某事情,且前后句子联系较为明显,主题突出.由此可见,旅游和新闻领域的文本大多数都满足零形回指的特点.因此,在旅游和新闻领域中,文本出现零形回指的现象比较普遍.

先行词指语境中与零形回指具有相同指称内容的句法成分,也称为零形回指的所指.王厚峰等人[29]认为绝大多数的零形回指都属于动词支配的成分.因此,本文仅考虑包含协陪义动词,且只出现主体成分的句子的零形回指.

鉴于旅游和新闻领域文本具有的上述特点,包含协陪义动词的句子总是紧接在描述主题人物或组织的句子之后出现,且前后句子具有明显的联系,属于有定的零形式(Definite Null Instantiation,DNI).DNI 是指缺失的角色一定是在篇章的上下文中已经被理解,并且在上下文中能找到[30].因此,如何在句子之间寻找协陪义句中的零形回指的先行词是需要解决的关键问题之一.

先行词的寻找需要借助零形式的识别与消解的方法.零形式的识别与消解的目标在于找到未显式表达的语义成分,并在篇章中为其找到先行语[30].Tao[31]借助与所指相关动词的特殊语义要求等线索判定零形回指的先行词.殷国光等人[32]提出依据谓语核心的意义确定先行词应具有的特征,进而在前文语境中寻找具有相应语义特征的名词性词语.陈平[27]指出先行词处于主语位置时,启后性最强.受人脑短时记忆功能的限制,零形回指与先行词之间的距离不能太远,否则就很难找回[33].

本文借助协陪义动词对零形回指的先行词进行判定.协陪义句的零形回指属于句间临近零形回指,经常紧挨含有其先行词的句子出现,且属于有定的零形式,因此在上下文中能找到先行词.此外,根据协陪义动词的语义特征可知,协陪义句子的零形回指的先行词应为人物/组织 PO 实体,通常该 PO 实体在前面的句子中充当主语成分.

综上所述,利用协陪义句中零形回指的先行词,建立不同句子中协陪义动词的主体成分与客体成分之间的联系;再利用客体成分的显式关系对主体成分的隐式关系进行推理,实现句间基于协陪义动词的隐式实体关系的抽取.具体步骤如下:

① 对于一个包含协陪义动词 V_x,且只包含主体成分 e_m 的句子 s_t,依次往前查找句子,找到一个句子

$s_k(k<t)$，其主语 e_a 为人物/组织(PO)实体类型.

② 若该人物/组织实体 e_a 在本句中至少存在一个显式关系，则判定人物/组织实体 e_a 为协陪义句 s_t 的零形回指的先行词.

③ 利用先行词 e_a 的显式关系对主体成分 e_m 进行隐式关系推理，即主体成分 e_m 参与先行词 e_a 相同的活动. 首先，将与先行词 e_a 发生显式关系的每一个非 PO 实体 e_s，均与主体成分 e_m 建立联系，构成隐式关系对$\langle e_m,e_s\rangle$或$\langle e_s,e_m\rangle$；然后，利用与先行词 e_a 组成的显式关系类型$\langle e_a,e_s\rangle=RT$ 或$\langle e_s,e_a\rangle=RT$，推理出主体成分 e_m 与省略客体 e_s 之间的隐式关系类型也为 RT，即$\langle e_m,e_s\rangle=RT$ 或$\langle e_s,e_m\rangle=RT$. 具体如图 16 所示.

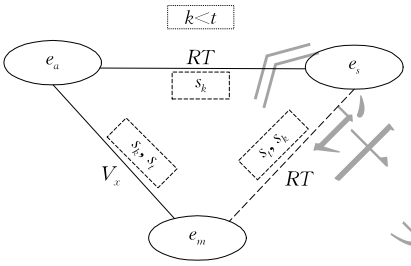

图 16　句间基于协陪义动词的隐式关系推理规则

6　趋向核心动词特征提取

趋向动词表示移动的趋向，通常为“来、去、上、下、进、出、上来、出去、开来、起来、到、到达、抵达、抵、来到、至、前往、前来”等动词[34]. 在旅游和新闻领域，趋向动词 V_q 通常表示某人/组织移动到某地干某事. 因此，由趋向动词 V_q 构成的句型通常可以表示为：PO+V_q+TA(+V+EN)，其中 PO 表示人物/组织实体，TA 表示景点/地点实体，V 一般为表示目的或结果的动词，EN 表示实体. 该句型重点在于强调某人/组织移动到某地的目的性或结果，如例 21 所示.

例 21.　“首都记者团到泰安游览.”

在例 21 中，PO 实体“首都记者团”到 TA 实体“泰安”的目的为“游览”(V). 因此，实体对〈首都记者团，泰安〉之间存在的关系为“游历”类型(PLA)，句中的动词“游览”可体现该实体之间的关系类型.

文献[5]通过提取与实体对直接或间接发生依存关系的动词作为最近句法依赖动词特征. 对于例 21 中的实体对〈首都记者团，泰安〉，文献[5]通常提取趋向动词“到”作为该实体对的最近句法依赖动词特征. 然而，该趋向动词特征却并不能表征实体对之间真正的关系类型，影响了实体关系抽取的性能. 因此，为了解决该问题，本文在文献[5]的最近句法依赖动词特征基础上，对于句型 PO+V_q+TA+V，提取动词 V 作为实体对〈PO，TA〉的动词特征，本文称为趋向核心动词特征 V_{QH}，如图 17 所示.

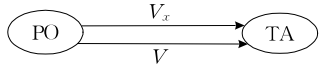

图 17　趋向核心动词特征

趋向核心动词特征提取的步骤如下，提取算法如算法 8 所示.

① 利用文献[5]的方法提取每个实体对$\langle e_i,e_j\rangle$的最近句法依赖动词特征 V_{ND}.

② 根据趋向动词词典 $QVDic$，判断该实体对的最近句法依赖动词特征 V_{ND} 是否为趋向动词；若为趋向动词，则进入到步骤③；否则 $V_{QH}=V_{ND}$.

③ 对该实体对$\langle e_i,e_j\rangle$的实体类型组合进行判断，若该实体对$\langle e_i,e_j\rangle$的实体类型组合为〈PO，TA〉，则进入到步骤④；否则 $V_{QH}=V_{ND}$.

④ 根据词性标注结果，判断该实体对所在的句子片段是否满足“PO+V_{ND}+TA+V”形式，即最近句法依赖动词特征 V_{ND} 位于实体对$\langle e_i,e_j\rangle$的中间，动词 V 位于第二个实体 e_j 的右边，动词 V 右边为标点符号 wp，且 V 与 wp 之间没有其他实体出现. 若满足“PO+V_{ND}+TA+V”形式，则取动词 V 作为该实体对的趋向核心动词特征 V_{QH}；否则 $V_{QH}=V_{ND}$.

算法 8.　趋向核心动词特征提取.

输入：一个实体对$\langle e_i,e_j\rangle$的最近句法依赖动词特征 V_{ND}，该实体对应句子的依存句法分析和词性标注结果，趋向动词词典 $QVDic$

输出：实体对$\langle e_i,e_j\rangle$的趋向核心动词特征 V_{QH}

$V_{QH}=V_{ND}$　//将 V_{ND} 作为趋向核心动词特征 V_{QH}

IF ($QVDic.contains(V_{ND})$) //V_{ND} 为趋向动词

　IF ((e_i 为 PO)&&(e_j 为 TA)) //实体类型判断

　　$V=getRightVerb(V_{ND})$ //V 为 V_{ND} 右边的动词

　　IF (V 不为空 &&(V_{ND} 位于实体 e_i 和 e_j 之间)&&(V 右边为 wp)&&(V 与该 wp 之间没有其他实体)) //满足 PO+V_{ND}+TA+V 的形式

　　$V_{QH}=V$ //将动词 V 作为趋向核心动词特征 V_{QV}

　　ENDIF

　ENDIF

ENDIF

7 实验评测

7.1 实验数据集

本文实验仍然采用我们在文献[5]中构建的3个旅游领域人文历史信息数据集,以及来源于新闻网站上的两个新闻数据集.其数据特点如下:

(1)泰山数据集.该数据集来源于"泰山文化"网站中的"泰山纪年"和"名人与泰山"版块信息.它以句子为单位,包含较多的协陪义句子和趋向动词的句子.

(2)庐山数据集.该数据集来自于"庐山之家"网站上有关"庐山历史上的今天"版块信息.它以句子为单位,包含趋向动词的句子,但协陪义句子较少.

(3)井冈山数据集。该数据集采用了"井冈山红色数字家园"网站中的"人文篇·文化传承"版块信息。它以句子为单位,包含趋向动词的句子,但协陪义句子较少.

(4)新闻数据集.该数据集来源于新闻网站中的新闻信息,分为2个子数据集:

①新闻数据集I(记为新闻I).该数据集以句子为单位,协陪义的句子占主要.

②新闻数据集II(记为新闻II).该数据集以文档为单位,由350个新闻文档组成.每个文档中至少有一个句子中包含协陪义动词,且该句中只包含主体成分.每个文档中仅保留包含协陪义动词的句子及其之前出现的句子.

本文只关心4类实体:景点/地点实体(TA)、人物/组织实体(PO)、作品实体(WOR)和活动实体(ACT).实体类型信息如表4所示.

表4 实体类型信息

实体类型	标号	范围
人物/组织	PO	个人、团队、组织等
景点/地点	TA	国家、地点、景点、别墅、宾馆等
作品	WOR	诗词、歌曲、学说、神话、著作、宣言、题词、雕塑、字画、展品、藏品、影视、御碑等
活动	ACT	会议、战争、比赛、谈判、仪式、论坛、表演等

本文依然采用文献[5]的数据预处理方法,首先采用哈尔滨工业大学LTP-Cloud[35]平台对5个实验数据集进行分词、词性标注、句法分析和实体识别,然后在此基础上加入规则进行适当修订,以便更好地符合领域特点.

本文仅对人物/组织实体(PO)与其他3类实体构成的实体对之间的关系进行抽取.关注的具体实体关系类型共10类:位于关系(LOC)、游历关系(PLA)、考察访问关系(VIS)、居住关系(LIV)、建立关系(BUI)、参与关系(PAR)、创作关系(CRE)、到达关系(ARR)、其他关系(OTH)和无关系(NON).

本文首先过滤掉无实体的句子,然后对于不同的数据集进行不同的处理:

(1)以句子为单位的泰山、井冈山、庐山和新闻I数据集,只考虑一个句子中两个实体之间的语义关系.因此,进一步过滤掉只有一类实体的句子,然后对于剩余的句子按照实体在句中出现的顺序进行两两组合构成实体对.

(2)以文档为单位的新闻II数据集,对于每一个文档,只保留两类句子:一类句子的主语为PO实体,且该PO实体有显式实体关系,标志为*TS*句;另一类句子包含协陪义动词,且该句中只包含PO主体成分,标志为*IS*句.该数据集主要用于验证句间基于协陪义动词的隐式关系抽取方法的效果.对于*TS*句,依然采用上述以句子为单位的数据集的实体对构成方法.对于*IS*句,以协陪义动词为线索,通过上下文,将其缺省的客体成分补充出来,人工标注出隐式实体关系对.

为了选择出黄金标准集,由1位博士研究生和2位硕士研究生组成标注小组,对实体关系类型和基于协陪义动词的隐式实体关系抽取的相关数据进行人工标注,当出现标注不一致的情况,则由3人讨论确定最终标注结果.5个实验数据集中的句子数以及实体对数量如表5所示.

表5 数据集预处理后的统计信息

数据集	句子数	实体对数
泰山	762	2795
庐山	1278	3605
井冈山	393	1654
新闻I	2000	5187
新闻II	738	1559

7.2 句内基于协陪义动词的隐式关系推理

句内基于协陪义动词的隐式关系推理的实验数据集需要包含较多协陪义句子,因此本文选取了泰山数据集和新闻I数据集进行实验.利用协陪义动词词表,根据包含协陪义动词候选句的识别方法,分别提取到泰山和新闻I数据集上的协陪义动词候选句为102句和1994句,说明该方法是有效的.

(1)协陪义句型判断与协陪义句子筛选

根据协陪义动词在句中充当的成分不同,利用

算法 1，对泰山和新闻 I 数据集的协陪义候选句进行句型判断，结果如表 6 所示.

表 6　协陪义句型分类和协陪义句子筛选的结果

句型	数据集	协陪义候选句	协陪义句子
谓语句型	泰山	36	34
	新闻 I	1296	954
并列句型	泰山	0	0
	新闻 I	94	93
状语句型	泰山	46	46
	新闻 I	343	343
定语句型	泰山	11	10
	新闻 I	192	192
混合句型	泰山	9	9
	新闻 I	69	69

接下来需要剔除包含协陪义动词但并不是协陪义的句子. 经过筛除掉非协陪义句子，共得到泰山和新闻 I 数据集的协陪义句子分别为 99 个和 1651 个，结果如表 6 所示.

(2) 协陪义成分识别

对于 5 种不同句型，根据算法 2～算法 6，对泰山和新闻 I 数据集上的协陪义句子进行成分识别. 在这 2 个数据集上，协陪义成分识别到的核心实体、主要实体和次要实体在 5 种句型中出现的个数、正确识别的个数、正确率 P、召回率 R 和 $F1$ 值的统计结果如表 7 所示.

表 7　协陪义成分识别结果统计

测试种类		实体成分																	
		核心实体						主要实体						次要实体					
		实际出现	规则返回	正确识别	P/%	R/%	$F1$/%	实际出现	规则返回	正确识别	P/%	R/%	$F1$/%	实际出现	规则返回	正确识别	P/%	R/%	$F1$/%
泰山	主谓句型	35	33	33	100.00	94.26	97.06	3	4	3	75.00	100.00	85.71	38	42	37	88.10	97.37	92.50
	并列句型	0	0	—	—	—	—	0	0	—	—	—	—	0	0	—	—	—	—
	状语句型	48	47	47	100.00	97.92	98.95	6	4	3	75.00	50.00	60.00	80	80	80	100.00	100.00	100.00
	定语句型	10	10	10	100.00	100.00	100.00	2	2	2	100.00	100.00	100.00	10	10	10	100.00	100.00	100.00
	混合句型	18	16	16	100.00	88.89	94.12	0	0	—	—	—	—	22	22	22	100.00	100.00	100.00
	合计	111	106	106	100.00	95.50	97.70	11	10	8	80.00	72.73	76.19	150	154	149	96.75	99.33	98.03
新闻 I	主谓句型	909	904	902	99.78	99.23	99.50	1145	1138	1137	99.82	99.30	99.56	997	992	990	100.00	98.45	99.22
	并列句型	93	89	89	100.00	95.70	97.80	94	92	92	100.00	97.87	98.92	108	106	106	100.00	98.45	99.22
	状语句型	343	342	342	100.00	99.71	99.85	418	411	403	98.05	96.41	97.23	548	548	546	100.00	99.80	99.90
	定语句型	192	193	192	99.48	100.00	99.74	202	203	202	99.51	100.00	99.75	215	216	215	100.00	100.00	100.00
	混合句型	138	138	133	96.38	96.38	96.38	156	156	152	97.44	97.44	97.44	206	206	204	100.00	98.94	99.47
	合计	1675	1666	1658	99.52	98.99	99.25	2015	2001	1986	99.25	98.56	98.90	2074	2068	2061	100.00	99.16	99.58

由表 7 可知，本文提出的根据句型进行协陪义成分识别的方法是有效的. 在旅游领域泰山数据集和新闻 I 数据集上，对于核心实体、主要实体和次要实体在 5 种句型上的识别正确率 P、召回率 R 和 $F1$ 值都非常高，充分说明了本文方法具有很好的领域适用性，同时也说明了本文方法对于数据规模具有很好的适应性.

(3) 句内基于协陪义动词的隐式关系推理

为了验证句内基于协陪义动词的隐式关系推理方法的有效性，本文将泰山和新闻 I 数据集上识别出来的核心实体、主要实体和次要实体与句中非 PO 实体进行组合，分别得到核心实体对、主要实体对和次要实体对. 本文在核心实体对关系类型已知的情况下，将核心实体对作为显式关系，主要实体对和次要实体对作为未知的隐式关系，具体标注信息如表 8 所示.

表 8　句内基于协陪义动词的实体关系标注数量

数据集	显式关系数量	隐式关系数量	合计
泰山	180	311	491
新闻 I	1876	3127	5003

从表 8 可知，在泰山和新闻 I 数据集上，协陪义句中包含大量的隐式关系，其数量远远超过显式关系数量，这正好与前文分析相一致.

根据算法 6，本文借助已知的显式实体关系进行隐式实体关系的推理. 在数据集上能够适用于 3 种规则的隐式关系个数(即实际出现个数)、应用 3 种规则分别返回的隐式关系个数、正确推理出来的隐式关系个数以及正确率 P(指关系对和关系类型都同时判断正确)、召回率 R 和 $F1$ 值的统计结果如表 9 所示.

从表 9 可知，本文提出的句内基于协陪义动词的隐式关系推理方法非常有效，在泰山和新闻 I 数据集上的总体正确率 P、召回率 R(即结果中实体对

表 9 句内基于协陪义动词的隐式关系推理结果统计

测试种类	数据集											
	泰山						新闻 *I*					
	实际出现	规则返回	正确推理	P/%	R/%	F1/%	实际出现	规则返回	正确推理	P/%	R/%	F1/%
CR 规则	18	17	14	77.78	82.35	80.00	516	497	481	93.22	96.78	94.97
IR 规则	285	285	275	96.49	96.49	96.49	2599	2543	2522	97.04	99.17	98.09
NR 规则	8	9	8	100.00	88.89	94.12	12	13	11	91.67	84.62	88.00
合计	311	311	297	95.50	95.50	95.50	3127	3053	3014	96.39	98.72	97.54

与实体关系类型均推理正确的个数/测试集中给定关系类型的总个数)和 *F*1 值均高达 95%及以上.具体分析如下:

① 协陪义成分识别的性能直接影响隐式实体关系推理的性能.因为 CR 规则、IR 规则和 NR 规则中涉及的核心实体、主要实体和次要实体均通过协陪义成分识别获取,如 CR 规则与核心实体、主要实体有关,IR 规则和 NR 规则均与核心实体、次要实体有关.从表 9 中可知,协陪义隐式关系推理在新闻 I 数据集上的整体性能高于泰山数据集,这与表 6 中的协陪义成分识别性能的高低相一致.CR 规则在泰山数据集上的性能最低,一方面由于一些未识别出的核心实体导致错误主要实体对的出现;另一方面由于泰山数据集上的主要实体识别性能最低,从而导致无法推理出正确的主要实体关系.由此可见,协陪义成分识别会给句内基于协陪义动词的隐式关系推理带来级联错误.因此,协陪义成分识别性能对于句内基于协陪义动词的隐式关系推理起着关键作用.

② 在泰山和新闻 I 数据集上,IR 规则处理的隐式关系对的数量都最多,分别高达 92%和 83%,说明协陪义动词包含的协陪意义主要体现在该类隐式关系上,因此 IR 规则在协陪义隐式关系抽取总体性能中起关键作用.IR 规则在泰山和新闻 I 数据集的性能都非常高,其正确率 *P*、召回率 *R* 和 *F*1 值均高达 96%及以上.

7.3 句间基于协陪义动词的隐式关系推理

为了验证句间基于协陪义动词的隐式关系抽取方法的有效性,需要以文档为单位,且文档内的句子具有有序编号.由于 3 个旅游领域相关数据集和新闻 I 数据集,均是以独立句子为单位放在一起,而没有保留同一文档中句子出现的顺序关系,因此无法采用这些数据集进行本实验.为此,本文选取新闻 II 数据集进行实验,将 *TS* 句作为先行词查找的句子集,其主语的显式关系类型为已知;*IS*句作为未知的隐式关系,具体标注信息如表 10 所示.

表 10 新闻 II 标注数据统计表

数据集	*TS* 句子数	*IS* 句子数	显式关系数量	隐式关系数量
新闻 II	350	388	504	1055

从表 10 可知,在新闻 II 数据集上,从 350 个新闻文档中保留了 *TS* 和 *IS* 句子共 738 个,平均每个文档中包含不到 3 个句子,其中至少包含 1 个 *IS* 句,大约 1～2 个 *TS* 句.这很好地说明了协陪义句中的零形回指属于句间临近零形回指,经常紧挨含有其先行词的句子出现,属于有定的零形式.这正好与前文分析相一致.此外,*IS* 句的数量比 *TS* 句多,说明一个文档中有可能包含了多个 *IS* 句,且这多个 *IS* 句中的零形回指的先行词很可能出现在同一个 *TS* 句中,且它们的先行词相同.

在表 10 中,将显式关系数量与隐式关系数量进行对比可知,新闻 II 数据集上的 *IS* 句中蕴含着大量的隐式关系,其数量远远超过显式关系.

本文将协陪义句的零形回指的先行词判定和句间基于协陪义动词的隐式关系推理方法结合在一起进行.首先,对于每一个 *IS* 句,其零形回指的先行词在 *TS* 句中进行查找判定,然后利用找到的先行词实体的已知显式关系,对 *IS* 句进行隐式关系推理.在新闻 II 数据集上能够适用于跨句子关系推理规则的隐式关系个数(即实际出现个数)、应用跨句子关系推理规则分别返回的隐式关系个数、正确推理出来的隐式关系个数以及正确率 *P*(指关系对和关系类型都同时判断正确)、召回率 *R* 和 *F*1 值的实验结果如表 11 所示.

表 11 句间基于协陪义动词的隐式关系推理结果

数据集	实际出现	规则返回	正确判断	*P*/%	*R*/%	*F*1/%
新闻 II	1055	989	989	100.00	93.74	96.77

从表 11 可知,本文提出的句间基于协陪义动词的隐式关系推理方法非常有效,在新闻II数据集上

的正确率 P、召回率 R 和 $F1$ 值均高达 93%及以上. 特别是正确率 P 高达 100%，说明协陪义句中零形回指的先行词判定方法和句间基于协陪义动词的隐式关系推理方法都非常准确. 召回率 R 相对低些，其原因在于有些 TS 句的句法分析存在错误，提取到非 PO 实体作为其主语，导致协陪义句中零形回指的先行词未找到，从而无法推理出句间基于协陪义动词的隐式关系. 因此，协陪义句中零形回指的先行词的判断效果对于句间基于协陪义动词的隐式关系推理起着关键作用.

7.4　基于趋向核心动词特征的关系抽取

由于包含趋向动词 V_q 构成的句型中必须包含一个 PO 实体. 因此，本文只需关注 PO 实体与其他类型实体之间的关系，即{〈PO，TA〉，〈TA，PO〉，〈PO，ACT〉，〈ACT，PO〉，〈PO，WOR〉，〈WOR，PO〉}.

鉴于泰山、庐山和井冈山数据集中包含较多趋向动词的数据，因此本文选取这 3 个数据集进行基于趋向核心动词特征的关系抽取实验.

(1) 趋向核心动词特征提取

本文首先利用文献[5]的方法提取实体对的最近句法依赖动词特征，然后选取最近句法依赖动词特征为趋向动词的实体对，根据算法 8，进一步提取趋向核心动词特征，其实验结果如表 12 所示.

表 12　趋向核心动词特征提取准确率对比表

数据集	实体对总数	趋向动词出现次数	最近句法依赖动词		趋向核心动词	
			正确提取	P/%	正确提取	P/%
泰山	2795	113	30	26.55	103	91.15
庐山	3605	200	78	39.00	187	93.50
井冈山	1654	108	70	64.81	96	88.89

在表 12 中，实体对总数表示数据集中实体对的总个数；趋向动词出现次数表示利用文献[5]提取到的最近句法依赖动词特征为趋向动词的实体对个数；正确提取表示提取到真正表示实体对之间关系类型的动词.

从表 12 中可知，本文提出的趋向核心动词特征提取方法远远优于文献[5]的最近句法依赖动词特征方法，在 3 个数据集上的正确率分别提高了 64.60、54.50 和 24.08 个百分点. 这充分说明了本文方法更有利于提取到真正表征实体对关系类型的动词.

(2) 基于趋向核心动词特征的关系抽取

本文提出的趋向核心动词特征提取方法是在文献[5]的最近句法依赖动词特征基础上进行改进. 因此，为了验证本文方法在实体关系抽取上的有效性，在使用相同实验数据情况下，在基本特征(实体类型组合特征和实体间距离特征)基础上分别加入文献[5]的最近句法依赖动词特征和本文的趋向核心动词特征，利用 LIBSVM[36] 分类器进行实体关系抽取实验. 本实验对每个数据集随机选择其中的 80%作为训练集，剩余 20%为测试集. 其关系抽取对比实验结果如表 13 所示.

表 13　趋向核心动词特征对关系抽取性能的影响

数据集	测试种类	最近依赖动词特征	趋向核心动词特征
泰山	P/%	70.63	72.50
	R/%	71.35	72.61
	$F1$/%	70.99	72.55
庐山	P/%	64.65	67.00
	R/%	65.08	67.46
	$F1$/%	64.86	67.23
井冈山	P/%	74.00	75.20
	R/%	72.48	73.09
	$F1$/%	73.23	74.13

从表 13 中可知，基于趋向核心动词特征方法在 3 个数据集上的关系抽取性能均优于基于依赖动词特征方法. 说明了基于趋向核心动词特征方法能有效提取到真正表示实体间关系类型的动词特征.

7.5　显式与隐式相结合的实体关系抽取

已有的关系抽取方法[5-6,8-9] 通常在整个数据集中进行显式实体关系抽取，并不考虑隐式实体关系的抽取. 本文构建的新闻 I 数据集全部是由协陪义句构成，显式关系全部由核心实体对组成，且隐式关系数量远远超过显式关系(见表 8). 而泰山数据集中包含了协陪义句和非协陪义句，且实体对的构成更多样化，具体可以分为两部分：① 显式关系实体对. 由非协陪义句中的实体对和协陪义句中的核心实体对组成；② 隐式关系实体对. 由协陪义句中的主要实体对和次要实体对组成. 因此，为了更好地验证句内基于协陪义动词的隐式关系推理方法的有效性，选取泰山数据集作为本实验数据. 泰山实验数据集实体关系对的具体统计信息如表 14 所示.

表 14　泰山数据集实体关系对统计表

显式关系实体对			隐式关系实体对			总计
非协陪义实体对	核心实体对	小计	主要实体对	次要实体对	小计	
2304	180	2484	18	293	311	2795

为了验证本文提出的显式与隐式相结合的实体关系抽取方法的有效性,用显式+隐式的方法与单独使用显式实体关系抽取方法进行比较.单独使用显式实体关系抽取是将泰山整个数据集上的实体关系都作为显式关系进行抽取.在单独使用显式实体关系抽取实验中,文献[5]同样采用了泰山数据进行显式实体关系抽取研究,从句法、语义角度提出了依存句法组合特征和最近句法依赖动词特征两个新特征.作为与文献[5]的同类对比方法,文献[8]新增了依存句法关系、核心谓词、语义角色标注等特征;文献[9]所提出的依赖动词特征是中文实体关系抽取领域对动词研究较为经典的方法之一.而本文的趋向动词特征是在文献[5]的最近句法依赖动词特征基础上进行改进.因此,本实验依然采用文献[8]和文献[9]作为同类对比方法.以实体关系类型组合和实体间距离为基本特征,然后分别加入文献[8]、文献[9]、本文的趋向核心动词特征与文献[5]的依存句法组合特征,得到3组实验结果分别记为郭方法(简记为Guo)[8]、董方法(简记为Dong)[9]和本文方法(简记为Ours).具体特征为:

(1)董方法特征:基本特征和依赖动词特征.

(2)郭方法特征:基本特征、依存句法关系、语义角色标注以及实体与核心谓词的距离.

(3)本文方法特征:基本特征、依存句法组合特征和最近句法依赖动词特征.

本文提出的显式与隐式相结合进行实体关系抽取方法是将关系抽取分为两个阶段进行:① 首先利用上述3种显式实体关系抽取方法,对显式关系实体对进行抽取;② 然后利用从显式实体关系抽取中得到的核心实体对关系,根据本文提出的句内基于协陪义动词的隐式实体关系推理方法,对句中的隐式关系实体对进行关系推理.结合显式实体关系抽取和隐式实体关系抽取两个阶段,从而完成对整个数据集的实体关系抽取,其关系抽取总体性能与已有方法的结果对比如表15所示.其中,"+"表示将显式关系抽取方法与句内基于协陪义动词的隐式关系推理方法(简称为句内协陪义隐式关系推理方法)相结合.

从表15可知,本文方法关系抽取的总体性能最好.具体分析如下:

(1)表15中前3种方法均进行显式实体关系抽取.本文方法在显式实体关系抽取任务中整体性能最佳.与郭方法相比,本文方法的正确率P、召回率R和$F1$分别提高12.26、9.19和10.75个百分点.特别是与董方法相比,本文方法在显式实体关系抽取任务上的性能提高最为明显,其正确率P、召回率R和$F1$的提高幅度高达4.19、28.11和18.61个百分点.这充分证明了本文提出的趋向核心动词特征和文献[5]中的依存句法组合特征在显式实体关系抽取中的有效性.同时再次充分说明了本文提出的趋向核心动词特征更能有效表征实体对之间的关系类型.

表15 显式与隐式相结合的实体关系抽取的总体性能

方法	泰山		
	P/%	R/%	$F1$/%
Guo	63.54	66.31	64.90
Dong	71.61	47.39	57.04
Ours	75.80	75.50	75.65
Guo+句内协陪义隐式关系推理方法	71.08	79.05	74.85
Dong+句内协陪义隐式关系推理方法	78.84	83.93	81.31
Ours+句内协陪义隐式关系推理方法	79.40	85.09	82.14

(2)表15中后三种方法是前3种显式实体关系抽取方法分别与本文提出的句内基于协陪义动词的隐式关系推理方法相结合得到的方法.后3种方法的实体关系抽取总体性能均得到了显著的提高,充分证明了本文提出的句内基于协陪义动词的隐式关系推理方法的有效性.董方法加入句内协陪义隐式关系推理方法后,其正确率P、召回率R和$F1$提高的幅度最大,分别高达7.23、36.54和24.27个百分点.其原因在于,对于协陪义句中的句内隐式关系实体对,董方法很难抽取到真正表征实体关系类型的动词特征,从而导致关系抽取性能低下.而本文通过将显式实体关系和隐式实体关系分开来进行抽取.董方法在显式关系抽取阶段的表现较好,然后再利用句内基于协陪义动词的隐式关系推理方法对隐式关系实体对进行推理,从而使得关系抽取的总体性能得到很大的提升.郭方法加入句内协陪义隐式关系推理方法后,其正确率P、召回率R和$F1$提高的幅度也较大,分别为7.54、12.74和9.96个百分点.这充分说明了本文提出的显式与隐式相结合进行实体关系抽取方法的有效性,综合了显式关系和隐式关系抽取两阶段的优势,从而使得实体关系抽取的整体性能达到最佳,同时也证明了本文提出的句内基于协陪义动词的隐式关系推理方法有利于实体关系抽取性能的提高.

(3)总体来看,本文方法与句内基于协陪义动词的隐式关系推理方法结合后,其关系抽取的性能最高,其正确率P、召回率R和$F1$分别达到了79.40%、85.09%和82.14%. 充分说明了本文提出的趋向核心动词特征和句内基于协陪义动词的隐式关系推理方法对关系抽取性能的提高起到了很好的作用. 与其他方法相比,本文方法与句内基于协陪义动词的隐式关系推理方法结合后对关系抽取性能提升的幅度最小,其正确率P、召回率R和$F1$分别提高了3.60%、9.59%和6.49%. 因为本文方法能够正确抽取一部分协陪义句中的句内隐式关系,从前面分析可知,本文方法在显式实体关系抽取任务中的性能已经较好,因此加入句内基于协陪义动词的隐式关系推理方法后,整体关系抽取性能提升的不是最明显. 但从关系抽取的整体性能来看,综合本文的趋向核心动词特征和句内基于协陪义动词的隐式关系推理方法(即Ours+句内协陪义隐式关系推理方法)表现最佳.

8　总结与展望

由于已有关系抽取相关研究多集中在显式实体关系抽取,从而很少关注隐式实体关系抽取. 事实上,数据集上存在着大量的隐式实体关系,特别是中文旅游和新闻领域包含许多由协陪义动词引发的隐式实体关系. 为了解决上述问题,本文提出了基于协陪义动词的中文隐式实体关系抽取方法,将实体关系抽取分为两阶段:显式关系抽取和隐式关系抽取,利用显式关系对隐式关系进行推理,充分发挥各自的优势,有利于中文实体关系抽取整体性能的提高.

进一步的研究工作主要有:

(1)将协陪义动词V_x与动词V相结合,通过对动词V进行某种分类,确认动词V与先行名词的同指程度,然后进一步细化和完善隐式关系的推理规则.

(2)进一步研究句间基于协陪义动词的隐式关系推理,考虑更多的零形式,利用更加有效的方法自动准确地识别出零形回指的先行词,有助于进一步提高句间基于协陪义动词的隐式关系抽取性能.

(3)考虑将实体识别和关系抽取统一到一个联合模型进行建模,利用实体识别和关系抽取进行相互学习和指导,进一步提升实体识别和关系抽取的性能.

致　谢　本文的研究利用了哈尔滨工业大学社会计算与信息检索研究中心免费开放的LTP平台和同义词词林扩展版、台湾大学林智仁等人开发的LIBSVM工具包;在论文投稿过程中,得到了许多评审老师的指教. 在此一并表示感谢!

参　考　文　献

[1] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations //Proceedings of the ACL Conference on Interactive Poster and Demonstration Sessions. Barcelona, Spain, 2004: 22

[2] Culotta A, McCallum A, Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text//Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York, USA, 2006: 296-303

[3] Yuan Yu-Lin. Research of quasi bidirectional verbs. Studies in Language and Linguistics, 1989, (1): 12-25(in Chinese)
(袁毓林. 准双向动词研究. 语言研究, 1989, (1): 12-25)

[4] Sheng Li-Na. Research on Company of Verbs and Adverbs in Modern Chinese [M. S. dissertation]. Nanjing Normal University, Nanjing, 2008(in Chinese)
(沈莉娜. 现代汉语动词、副词的协同义研究[硕士学位论文]. 南京师范大学, 南京, 2008)

[5] Gan Li-Xin, Wan Chang-Xuan, Liu De-Xi, et al. Chinese named entity relation extraction based on syntactic and semantic features. Journal of Computer Research and Development, 2016, 53(2): 284-302(in Chinese)
(甘丽新, 万常选, 刘德喜等. 基于句法语义特征的中文实体关系抽取. 计算机研究与发展, 2016, 53(2): 284-302)

[6] Zhou G D, Su J, Zhang J, et al. Exploring various knowledge in relation extraction//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA, 2005: 427-434

[7] Liu H, Jiang C, Hu C, et al. Efficient relation extraction method based on spatial feature using ELM. Neural Computing and Applications, 2014, 12(30): 1-11

[8] Guo Xi-Yue, He Ting-Ting, Hu Xiao-Hua, et al. Chinese named entity relation extraction based on syntactic and semantic features. Journal of Chinese Information Processing, 2014, 28(6): 183-186(in Chinese)
(郭喜跃, 何婷婷, 胡小华等. 基于句法语义特征的中文实体关系抽取. 中文信息学报, 2014, 28(6): 183-186)

[9] Dong Jing, Sun Le, Feng Yuan-Yong, et al. Chinese automatic entity relation extraction. Journal of Chinese Information Processing, 2007, 21(4): 80-85(in Chinese)

(董静，孙乐，冯元勇等. 中文实体关系抽取中的特征选择研究. 中文信息学报，2007，21(4)：80-85)

[10] Surdeanu M，Tibshirani J，Nallapati R，et al. Multi-instance multi-label learning for relation extraction//Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island，Korea，2012：455-465

[11] Socher R，Huval B，Manning C D，et al. Semantic compositionality through recursive matrix-vector spaces//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island，Korea，2012：1201-1211

[12] Zeng D，Liu K，Lai S，et al. Relation classification via convolutional deep neural network//Proceedings of the 25th International Conference on Computational Linguistics. Dublin，Ireland，2014：2335-2344

[13] Yu M，Gormley M R，Dredze M. Factor-based compositional embedding models//Proceedings of the NIPS Workshop on Learning Semantics，Montréal，Quebec，Canada，2014：1-5

[14] Zhang S，Zheng D，Hu X，et al. Bidirectional long short-term memory networks for relation classification//Proceedings of the 29th Pacific Asia Conference on Language，Shanghai，China，2015：73-78

[15] Santos C N，Xiang B，Zhou B. Classifying relations by ranking with 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. Beijing，China，2015：626-634

[16] Zhou P，Shi W，Tian J，et al. Attention-based bidirectional long short-term memory networks for relation classification//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin，Germany，2016：207-212

[17] Qin P，Xu W，Guo J. An empirical convolutional neural network approach for semantic relation classification. Neurocomputing，2016，190：1-9

[18] Hendrickx I，Kim S N，Kozareva Z，et al. Semeval-2010 task 8：Multi-way classification of semantic relations between pairs of nominals//Proceedings of the NAACL HLT Workshop on Semantic Evaluations：Recent Achievements and Future Directions. Boulder，Colorado，USA，2009：94-99

[19] Yu X，Lam W. An integrated probabilistic and logic approach to encyclopedia relation extraction with multiple features//Proceedings of the 22nd International Conference on Computational Linguistics. Manchester，England，2008：1065-1072

[20] Luo Y F，Wang Q，Wang B，et al. Context-dependent knowledge graph embedding//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon，Portugal，2015：1656-1661

[21] Hamadou A B，Piton O，Fehri H. Multilingual extraction of functional relations between Arabic named entities using NooJ platform//Proceedings of the NooJ International Conference and Workshop. Komotini，Greece，2010：192-202

[22] Hoffart J，Suchanek F M，Berberich K，et al. YAGO2：Exploring and querying world knowledge in time，space，context，and many languages//Proceedings of the 20th International Conference Companion on World Wide Web. Hyderabad，India，2011：229-232

[23] Li H，Wu X，Li Z，et al. A relation extraction method of Chinese named entities based on location and semantic features. Applied Intelligence，2013，38(1)：1-15

[24] Chen Y，Zheng Q，Zhang W. Omni-word feature and soft constraint for Chinese relation extraction//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore，Maryland，2014：572-581

[25] Jiang J，Zhai C X. A systematic exploration of the feature space for relation extraction//Proceedings of the Human Language Technologies：The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester，New York，2007：113-120

[26] Xi Bin，Qian Long-Hua，Zhou Guo-Dong，et al. The application of combined linguistic features in semantic relation extraction. Journal of Chinese Information Processing，2008，22(3)：44-50(in Chinese)
(奚斌，钱龙华，周国栋等. 语言学组合特征在语义关系抽取中的应用. 中文信息学报，2008，22(3)：44-50)

[27] Chen Ping. Discourse analysis of zero anaphora in Chinese. Studies of the Chinese Language，1987，200(5)：363-378(in Chinese)
(陈平. 汉语零形回指的话语分析. 中国语文，1987，200(5)：363-378)

[28] Xu Yu-Long. Introduction to Contrastive Linguistics. Shanghai：Shanghai Foreign Language Education Press，1992(in Chinese)
(许余龙. 对比语言学概论. 上海：上海外语教育出版社，1992)

[29] Wang Hou-Feng，He Ting-Ting. Research on Chinese pronominal anaphora resolution. Chinese Journal of Computers，2001，24(2)：136-143(in Chinese)
(王厚峰，何婷婷. 汉语中人称代词的消解研究. 计算机学报，2001，24(2)：136-143)

[30] Wu Juan，Li Ru，Wang Zhi-Qiang. Null instantiation identification and resolution in Chinese discourse. Journal of Chinese Information Processing，2016，30(3)：9-15(in Chinese)
(武娟，李茹，王智强. 汉语篇章中零形式的识别与消解. 中文信息学报，2016，30(3)：9-15)

[31] Tao L. Topic discontinuity and zero anaphora in Chinese discourse：Cognitive strategies in discourse processing//Fox B，ed. Studies in Anaphora. Amsterdam，the Netherlands：

John Benjamins, 1996: 485-511

[32] Yin Guo-Guang, Liu Wen-Xia. Research on zero anaphora in discourse of "Zuo Zhuan": A case study of "Yin Gong". Linguistic Research, 2009, (3): 6-12(in Chinese)
(殷国光，刘文霞.《左传》篇章零形回指研究——以《隐公》为例. 语文研究，2009，(3)：6-12)

[33] Hou Min, Sun Jian-Jun. Zero anaphora in Chinese and how to process it in Chinese-English MT. Journal of Chinese Information Processing, 2005, 19(1): 14-20 (in Chinese)
(侯敏，孙建军. 汉语中的零形回指及其在汉英机器翻译中的处理对策. 中文信息学报，2005，19(1)：14-20)

[34] Huang Bo-Rong, Liao Xu-Dong. Modern Chinese Revised Five Edition (Part 2). Beijing: Higher Education Press, 2011(in Chinese)
(黄伯荣，廖序东. 现代汉语增订五版(下册). 北京：高等教育出版社，2011)

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

[36] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27

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

GAN Li-Xin, Ph. D. candidate, associate professor. Her current research interests include natural language processing and information retrieval.

JIANG Teng-Jiao, Ph. D., lecturer. Her current research interests include Web data management and sentiment analysis.

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

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

LIU Yu, M. S. candidate. Her current research interests focus on data mining.

Background

As a foundation of semantic networks and ontology, named entity relation extraction has become an interesting research domain in recent years. It is very useful for many applications such as web mining, information extraction and retrieval, automatic databases filling with entities and types, questions answering task and document summarization.

The target of named entity relation extraction is to detect implicit and explicit relations between entities. Most of the existing researches focus on explicit entity relation extraction, but ignore implicit entity relation extraction. Implicit relations have no explicit supporting evidence in text and require additional evident from a reading of the document. Therefore, it has been a challenge to infer implicit relations. Several works related to implicit relations extraction have been performed for European languages and especially for English. As far as we know, very few works have been done for Chinese language.

There are a great number of implicit entity relations that are triggered by company verbs in tourism and news domain texts. Therefore, we focus on Chinese named entity implicit relation extraction which is based on company verbs. This paper proposes a two-stage scheme that takes into account both explicit relation extraction and implicit relation extraction. We integrate machine learning method with rules and use explicit entity relations to infer implicit entity relations. In order to improve the effect of verb features on the explicit entity relation extraction, we propose a feature extraction algorithm of directional core verbs based on the verb feature with nearest syntactic dependency. In the implicit relation extraction stage, company candidate sentences are selected by POS(Part of Speech) and company verb dictionary. According to dependency parsing, the sentence pattern classification algorithm and the corresponding component recognition algorithm are designed for company candidate sentences. We propose a novel method that uses explicit relations to infer implicit relations. According to characteristics of company semantic components and the scope of company verbs, we design three rules for implicit entity relation extraction in

sentences based on company verbs. In addition, by exploiting the antecedent of the zero anaphora in a company sentence, we establish the associations between the subject and object components in different sentences, which are then used to extract implicit entity relation across sentences based on company verbs. Therefore, we can effectively handle implicit entity relation extraction based on company verbs and capture more implicit entity relations. Because our work combines the advantage of explicit relation extraction and implicit relation extraction, it is beneficial to improve the overall performance of Chinese named entity relation extraction. To the best of our knowledge, our work is the first solution towards implicit relation extraction for Chinese language.

The research is partially supported by the National Natural Science Foundation of China under Grant Nos. 61562032, 61662027, 61173146, 61363039, 61363010 and 61462037, the Grand Natural Science Foundation of Jiangxi Province under Grant Nos. 20152ACB20003 and 20161BAB202057, the Ground Program on High College Science & Technology Project of Jiangxi Province under Grant Nos. KJLD12022 and KJLD14035, the Science & Technology Project of the Department of Education of Jiangxi Province under Grant Nos. GJJ150819 and GJJ160783, and the Humanities and Social Sciences Foundation of Jiangxi Provincial Universities under Grant No. JC161001.