计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150842
53(2):284-302,2016

基于句法语义特征的中文实体关系抽取

甘丽新　万常选　刘德喜　钟　青　江腾蛟
(江西财经大学信息管理学院　南昌　330013)
(数据与知识工程江西省高校重点实验室(江西财经大学)　南昌　330013)
(spiderganxin@163.com)

Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features

Gan Lixin, Wan Changxuan, Liu Dexi, Zhong Qing, and Jiang Tengjiao
(*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　Named entity relations are a foundation of semantic networks and ontology, and are widely used in information retrieval and machine translation, as well as automatic question and answering systems. In named entity relationships, relationship feature selection and extraction are two key issues. Characteristics of Chinese long sentences with complicated sentence patterns and many entities, as well as the data sparse problem, bring challenges for Chinese entity relationship detection and extraction tasks. To deal with above problems, a novel method based on syntactic and semantic features is proposed. The feature of dependency relation composition is obtained through the combination of their respective dependency relations between two entities. And the verb feature with the nearest syntactic dependency is captured from dependency relation and POS (part of speech). The above features are incorporated into feature-based relationship detection and extraction using SVM. Evaluation on a real text corpus in tourist domain shows above two features from syntactic and semantic aspects can effectively improve the performance of entity relationship detection and extraction, and outperform previously best-reported systems in terms of precision, recall and $F1$ value. In addition, the verb feature with nearest syntactic dependency achieves high effectiveness for relationship detection and extraction, especially obtaining the most prominent contribution to the performance improvement of data sparse entity relationships, and significantly outperforms the state-of-the-art based on the verb feature.

Key words　relationship extraction; relationship detection; syntactic feature; semantic feature; support vector machine (SVM)

摘　要　作为语义网络和本体的基础,实体关系抽取已被广泛应用于信息检索、机器翻译和自动问答系统中.实体关系抽取的核心问题在于实体关系特征的选择和提取.中文长句的句式较复杂,经常包含多

收稿日期:2015-09-22;**修回日期**:2015-12-22
基金项目:**国家自然科学基金项目**(61173146,61562032,61363039,61363010,61462037);**江西省高等学校科技落地计划项目**(KJLD12022);**江西省教育厅科技研究项目**(GJJ12733,GJJ13249)
This work was supported by the National Natural Science Foundation of China (61173146, 61562032, 61363039, 61363010, 61462037), the Ground Program on High College Science & Technology Project of Jiangxi Province (KJLD12022), and the Science & Technology Project of the Department of Education of Jiangxi Province (GJJ12733, GJJ13249).
通信作者:**万常选**(wanchangxuan@263.net)

个实体的特点以及数据稀疏问题,给中文关系探测和关系抽取任务带了挑战.为了解决上述问题,提出了一种基于句法语义特征的实体关系抽取方法.通过将2个实体各自的依存句法关系进行组合,获取依存句法关系组合特征,利用依存句法分析和词性标注选择最近句法依赖动词特征.将这2个新特征加入到基于特征的关系探测和关系抽取中,使用支持向量机(support vector machine, SVM)方法,以真实旅游领域文本作为语料进行实验.实验表明,从句法和语义上提取的2个特征能够有效地提高实体关系探测和关系抽取的性能,其准确率、召回率和$F1$值均优于已有方法.此外,最近句法依赖动词特征非常有效,尤其对数据稀疏的关系类型贡献最大,在关系探测和关系抽取上的性能均优于当前经典的基于动词特征方法.

关键词 关系抽取;关系探测;句法特征;语义特征;支持向量机

中图法分类号 TP311

处在大数据时代的今天,数据呈现出规模巨大、模态多样和高速增长等特征,使得"信息过载"问题日益严重,因此迫切需要快速、准确地获取用户真正所需信息的技术手段——信息抽取技术.实体关系抽取是信息抽取中的一个非常重要的子领域,其任务是从自然语言文本中提取出2个命名实体之间所存在的语义关系,例如,句子"邓兆祥游览庐山."中的2个实体"邓兆祥"和"庐山"之间存在着"游历"关系.作为自然语言处理的基础,实体关系抽取为海量信息处理、中文信息检索、知识库自动构建、自动问答、机器翻译和自动文摘等众多自然语言处理任务提供了重要的技术支持.

关系抽取的研究是以MUC评测会议和后来取代MUC的ACE评测会议为主线进行的,大量先进的信息抽取方法被提出来,有力地促进了关系抽取研究的完善发展.实体关系抽取所遵循的技术方法基本可以归纳为:基于模式匹配的方法、基于词典驱动的方法、基于本体的方法、基于机器学习的方法以及混合抽取方法[1].近几年的研究趋势表明,基于机器学习的方法逐渐成为关系抽取研究的主流思路.关系抽取通常采用有监督的机器学习方法,它可以根据关系实例的表示方式不同分为2类: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.由于特征的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的研究重点不在机器学习方法本身,而在于如何准确地获取各种词法、句法和语义等语言学特征,并把它们有效地集成起来,从而产生描述实体间语义关系的各种特征[2-10].

本文对旅游领域的景点人文信息进行实体关系抽取.旅游领域的景点人文信息通常是综合概括了名人或组织在某景点发生的事情.

例1."1937年6月4日,周恩来第一次登上庐山,入住仙岩旅馆,同蒋介石进行国共第二次合作谈判."

该句比较长,由4个短句构成,共包含5个实体;若只关注景点与人物/组织、景点与活动之间发生的显性关系,按照实体出现的顺序,可组成8个实体对,其中有4个实体对属于"无关系"类型,即实体对中的2个实体之间不存在关系.具体信息如表1所示:

Table 1 Information of Entities and Entity Relationships in Exp. 1

表1 例1中实体和实体关系信息

Number	Type	Information of Examples
#Entities=5	People/Organization(2)	周恩来、蒋介石
	Tourist Attraction(2)	庐山、仙岩旅馆
	Action(1)	国共第二次合作谈判
#Entity Pairs=8	Arrive Entity Relationship(1)	〈周恩来,庐山〉
	Live Entity Relationship(1)	〈周恩来,仙岩旅馆〉
	Participate Entity Relationship(2)	〈周恩来,国共第二次合作谈判〉、〈蒋介石,国共第二次合作谈判〉
	None Entity Relationship(4)	〈庐山,蒋介石〉、〈庐山,国共第二次合作谈判〉、〈仙岩旅馆,蒋介石〉、〈仙岩旅馆,国共第二次合作谈判〉

从例1可以看出,旅游领域的景点人文信息中的句子通常比较长,一个句子中经常包含多个实体信息,由此构成的实体对的数量也较多,且实体类型的数量分布不均匀.因此,旅游领域的景点人文信息的数据特点给实体关系探测和关系抽取任务带来了挑战.

1) 相对于简单句子的实体关系探测和关系抽取,长句的句式较复杂,使得实体关系探测和关系抽取的任务更加艰难.

2) 长句中经常包含多个实体信息,而且跨长距离的实体对所在的句子中通常存在多个动词,因此,如何选择能够有效地表征实体对之间有无语义关系以及具体关系类型的动词成为关系探测和关系抽取的关键.

3) 目前关系抽取的最大挑战在于训练数据不足,关系实例在各个类别上的分布极不均匀,主要集中在几个类上,如"游历"关系、"考察访问"关系和"无关系";有些类别的实例数目较少,如"建立"关系和"离开"关系.这正是目前关系抽取领域所面临的数据稀疏问题,严重影响了关系抽取的性能,给关系抽取带来了很大的难度.

由于句法结构在关系识别中起到非常重要的作用.依存语法通过分析语言单位内成分之间的依存关系揭示其句法结构,主张句子中核心动词是支配其他成分的中心成分,而它本身却不受其他任何成分的支配,所有受支配成分都以某种依存关系从属于支配者.依存句法分析可以反映出句子各成分之间的语义修饰关系,它可以获得长距离的搭配信息,并与句子成分的物理位置无关.句子中的实体必定会作为一个短语结构出现在依存结构中,将实体对应的依存句法关系进行组合能在一定程度上反映出相应实体之间的关系特征.

文献[5,7]指出特征子空间中的基本特征以及基本特征的组合能够有效地提升关系抽取性能.同时,已有研究表明,依存句法关系能有效地提高实体关系抽取的性能[3,6].因此,本文按照2个实体出现的先后顺序,将2个实体各自的依存句法关系进行组合,得到了依存句法关系组合特征.

本文提出依存句法关系组合特征的原因在于:依存句法关系组合特征具有有序性,即按照2个实体出现的先后顺序对各自对应的依存句法关系进行组合,比单独使用依存句法关系特征能更好地表示实体对在句中对应的句法结构.例如,"张学良将军离开庐山回武汉.",该句中存在的实体对为〈张学良,庐山〉,这2个实体的依存句法关系分别为主谓结构(SBV)与动宾结构(VOB),表明2个实体在句中分别充当SBV中的主语和VOB中的宾语.如果单独使用实体的依存句法关系特征,在关系分类判别过程中,虽然同时采用这2个实体的依存句法关系特征值SBV和VOB,但是由于分类过程中特征是无序的,因此可能出现的组合情况为SBV-VOB或VOB-SBV,而这2种依存句法关系的组合是有区别的.

从句法结构来看,例句中的实体对〈张学良,庐山〉表示主谓-动宾结构,用SBV-VOB表示更恰当,它反映出实体对〈张学良,庐山〉之间可能存在着语义关系.而在句子"857年(唐大中十一年),距李邕写《复东林寺碑》126年后,庐山东林寺再次大修,又请人写碑记之."中,实体对〈复东林寺碑,庐山东林寺〉之间不存在着任何关系,该实体对的依存句法关系组合为VOB-SBV.第1个实体"复东林寺碑"在子句中充当宾语成分;第2个实体"庐山东林寺"在子句中充当主语成分.而这2个实体分别在其子句中构成了完整的语义.

由于不同实体关系类型的依存句法关系组合特征的分布存在差异性,该特征具有一定的区分度,可以较好地反映出相应实体之间的关系类型特征.实验结果也验证了它在关系探测和关系抽取中的有效性.

由于动词能够很好地识别实体对之间的关系类型[6,8],很多实体关系通常可以通过动词来引发,类似于事件抽取中的事件经常由触发词而触发[11-13]的现象.动词特征的提取在整个特征提取过程中占有非常重要的位置,直接影响了关系抽取性能的好坏.由于跨长距离的实体对所在句子中通常会包含多个动词,因此,为了解决从多个动词中选择有效地表征实体对关系类型的动词选择问题,本文提出了最近句法依赖动词特征.

文献[8]中提出的依赖动词特征存在着2个问题:

1) 选择距离位置较后实体最近的动词作为依赖动词特征,并非都能提取到真正表征该实体对关系类型的动词,因此会影响实体关系类型的判别.例如,"毛岸青一行参观了庐山风景点,并参观了毛泽东在庐山居住过的美庐别墅、175号别墅以及芦林一号别墅和庐山会议会址.",该句中共存在着2个动词"参观"和"居住".对于实体对〈毛岸青一行,庐山会议会址〉,文献[8]抽取出的依赖动词特征为"居住",因此很可能将该实体对归为"居住"关系类型.

事实上,该实体对之间存在的关系类型为“游历”关系,而真正表征该关系类型的动词为“参观”.

2) 依赖动词特征并非都能有效地帮助实体之间有无关系的探测以及关系类型的区分,有时甚至会带来噪音,特别是在关系探测上该问题尤为突出.这是因为,对于关系探测中“无关系”类型的实体对,大多数情况下并不存在使实体对发生关系的依赖动词.因此,文献[8]提取的依赖动词特征给关系探测带来了大量噪音信息,不利于实体之间有无关系的区分,从而会影响关系探测的性能.例如,“蒋介石在庐山指挥东北战事.”,使用文献[8]的抽取方法,该句中的实体对〈庐山,东北战事〉的依赖动词为“指挥”,因此很可能会误判为“参与”关系;实际上该实体对之间不存在任何关系,即为“无关系”类型.

因此,为了解决上述问题,本文提出了语义特征——最近句法依赖动词特征(由于是通过依存句法分析来提取的,故称为最近句法依赖动词特征).通过对数据集的分析,我们发现以下事实.

1) 数据集中的句子基本上为陈述句和主动句.在陈述句中,不管句子怎么变,动词总是在第二位,第一位可以是主语或宾语.在主动句中,主语是谓语所表示的动作行为的发出者.根据陈述句和主动句的特点可知:如果2个实体之间存在语义关系,那么句中经常会存在这样的一个动词,通过该动词能够直接或间接地将这2个实体连接起来,并且第1个实体是该动词所表示的动作行为的发出者.

2) 如果2个实体之间不存在任何语义关系,则存在2种情况:①2个实体之间不存在动词使得它们发生语义关系;②每个实体均与不同的动词发生联系,而不同动词之间又不存在语义关联,即这2个实体无法通过一个相同的动词进行语义连接.

因此,为了减少大量噪音的引入,动词特征应该仅选择那些区分性较强的动词,以便有效地区分实体之间有无语义关系.

针对上述分析,本文的目标是通过依存句法分析和词性来提取一个句子中实体对$\langle e_i, e_j\rangle$的最近句法依赖动词特征.1)根据依存句法分析,分别提取实体e_i和e_j的依存关联节点e_i'和e_j';2)选择与第2个实体e_j的依存关联节点e_j'发生直接依存关系的动词V_j;3)获取与第1个实体e_i的依存关联节点e_i'直接发生SBV或FOB(前置宾语)关系的动词V_i;4)根据动词V_i与V_j的依存关系,确定实体对的最近句法依赖动词DV,如果V_i与V_j为同一个词或存在着并列关系,则确定V_j为最近句法依赖动词DV,否则将最近句法依赖动词DV置为空值Null.

本文提出的最近句法依赖动词特征能够有效地区分实体之间有无语义关系,特别是Null动词,具有较强的区分性,大大地减少了文献[8]依赖动词特征所带来的噪音,有利于提高关系探测性能.此外,由于最近句法依赖动词特征经常能触发实体之间的关系,能够较好地表征实体关系类型,因此有利于关系类型的识别;同时还能解决由于数据不平衡引起关系抽取性能低下的问题.实验结果表明,加入最近句法依赖动词特征能够显著地提升关系探测和关系抽取的性能,准确率、召回率和$F1$值均得到了大幅提升.

1 相关研究

1.1 SVM概述

目前,基于特征向量的关系抽取方法多采用最大熵模型[3]和支持向量机(support vector machine,SVM)[4-6].研究显示,SVM在性能上优于最大熵模型[5].SVM分类效果通常都会优于传统的算法,曾被称为“现成”的分类器,并被评为机器学习领域10大经典算法之一[14].SVM是一种基于统计学习理论驱动的有指导的机器学习方法,可用于分类和回归问题.基于统计学习理论中的结构风险最小化原则,SVM通过寻找一个最佳分类超平面将训练数据分成2类,并从训练集中挑出有效的实例作为支持矢量(即决策的依据).由于最基本的SVM是一个二元分类器,且分类过程较慢,因此,根据不同的研究与应用方向,又出现了许多基于SVM的优化算法,如SMO,C-SVM,V-SVM等方法[15-17],使SVM学习的过程更迅速,效果也得到明显提升.本文采用台湾大学林智仁等人开发的LIBSVM[16]作为SVM工具包进行实体关系抽取.

1.2 基于特征向量的实体关系抽取

基于特征向量的实体关系抽取方法的核心在于如何获取有效的特征表示.特征选取主要是从自由文本及其句法结构中抽取出各种表面特征以及结构化特征.

文献[3]综合考虑了实体单词、实体类型、实体引用方式、重叠、依存树和解析树等特征,实现了最大熵模型的关系分类器,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.

文献[4]则系统地研究了如何把包括基本词组块在内的各种特征广泛组合起来,探讨了各种语言

特征对关系抽取性能的贡献;深入研究了 WordNet 和 Name List 等语义信息对关系抽取的影响;实验结果表明,基本词组块能有效提升关系抽取性能.

文献[6]在传统方法的基础上提出一种基于句法特征、语义特征的实体关系抽取方法,融入了依存句法关系、核心谓词、语义角色标注等特征,实验结果表明该方法的 $F1$ 值有明显提升.

特征子空间中的基本特征以及基本特征的组合能够有效地提升关系抽取性能[5,7]. 文献[7]系统研究了关系抽取中的特征空间,通过合一的特征空间表达形式来研究不同特征对关系抽取性能的影响;特征空间按照序列、句法和依存关系划分为不同的子空间;实验表明特征子空间中的基本特征能有效提升关系抽取性能,而复杂特征带来的性能提升有限. 文献[5]并不是通过发掘新特征来提高语义关系抽取的性能,而是通过在各种词法、语法、语义的基本特征内部及特征之间进行有效的组合,从而产生出很多组合特征;实验证明,这些组合特征对提高语义关系抽取性能做出了很大贡献.

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[8]将实体关系划为包含实体关系与非包含实体关系,针对这 2 种关系的差异,提出新的句法特征,构建不同的特征空间;在非包含关系中使用了祖先成分、2 个实体之间的路径、依赖动词以及实体到依赖动词的路径等特征;实验表明,依赖动词较大程度地提高了实体关系抽取的性能.

文献[18]提出了一种基于位置语义特征的实体关系抽取方法,利用位置特征的可计算性和可操作性,以及语义特征的可理解性和可实现性,整合了词语位置的信息增益与基于 HowNet 的语义计算结果;实验结果表明,结合位置和语义特征的关系抽取方法优于单独使用位置或语义特征的方法.

文献[19]提出了 Omni-word 特征和软约束方法实现中文关系抽取,Omni-word 使用了句中各种潜在词作为词法特征,软约束方法能够获取局部依赖,这 2 种方法能够更好地利用句子信息,降低了中文分词和句法分析错误带来的影响;实验结果表明,该方法能有效地提高中文关系抽取的性能.

文献[20]基于概念模型获得了有效的空间特征,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;实验结果表明,该特征能有效地提升关系抽取的正确率和召回率.

本文在传统的词法和实体特征基础上,通过增加句法特征和语义特征——依存句法关系组合特征和最近句法依赖动词特征,以获取实体对之间更丰富的关系特征,提高中文实体关系探测和关系抽取的性能.

2 特征分析

主要介绍提出的句法特征和语义特征——依存句法关系组合特征和最近句法依赖动词特征.

2.1 依存句法关系组合特征

依存句法通过分析语言单位内成分之间的依存关系揭示句子中各成分之间的语义修饰关系,即指出句中词语之间在句法上的搭配关系,分析出一个句子的主、谓、宾、定、状、补结构. 在 Robinson 提出的依存句法关系公理中指出:任何一个成分都不能依存于 2 个或 2 个以上的成分. 因此,句子中的每一个实体必定会作为一个语义成分出现在依存结构中. 本文对 2 个实体的依存句法关系进行组合,提出了依存句法关系组合特征.

例 2. "邓兆祥游览庐山."的依存句法分析如图 1 所示. 其中,Root 表示根节点,HED 表示指向整个句子的核心,WP 表示指向标点符号;PO 和 TA 分别表示旅游领域中的人物/组织实体和景点实体;v 和 wp 分别表示词性标注中的动词和标点符号.

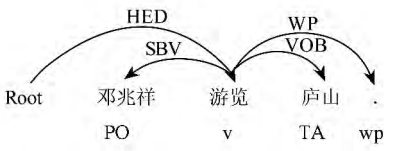

Fig. 1 Dependency parsing and POS tagging for entity relationships.

图 1 实体关系的依存句法分析和词性标注示例

图 1 中,实体对〈邓兆祥,庐山〉存在着"游历"关系,而该实体对具有 SBV-VOB 的依存句法关系组合.

由于旅游领域的景点人文信息是综合概括了名人或组织在某景点发生的事情,因此常将多个句子组合成一个长句,并没有按照语义或句式进行严格的断句. 如例 3 所示.

例 3. "张季鸾抵达庐山,蒋介石于 6 月 19 日在'美庐'会见了张季鸾."的句法分析结果如图 2 所示. 其中,COO, ADV, POB 和 RAD 分别表示的依存句法关系如表 2 所示;p, nt 和 u 分别表示词性标注中的介词、时间名词和助词.

从图 2 可以看出,实体对〈庐山,蒋介石〉之间不

存在任何关系,该实体对的依存句法关系组合为VOB-SBV.从图2的句法分析还可看出,第1个实体"庐山"所在句子的前半部分已经构成了一个语义完整的句式,"庐山"在该部分中充当宾语成分;而第2个实体"蒋介石"所在的后半个句子也已经构成了一个语义完整的句式,"蒋介石"在该部分充当主语成分.事实上,该句可以拆成2个语义完整的独立的句子.

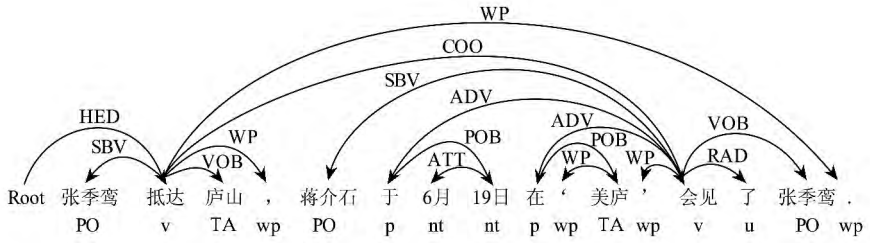

Fig. 2 Dependency parsing and POS tagging for none entity relationships.

图2 无关系的依存句法分析和词性标注示例

本文利用哈尔滨工业大学LTP-Cloud平台①对实验数据进行依存句法分析,共得到10类实体依存句法关系,具体如表2所示:

Table 2 Tag Set of Dependency Relations

表2 依存句法关系标注集

Symbol	Dependency Relation	Symbol	Dependency Relation
SBV	subject-verb	ADV	adverbial
VOB	verb-object	CMP	complement
FOB	fronting-object	COO	coordinate
IOB	indirect-object	POB	preposition-object
ATT	attribute	HED	head

本文按照实体在句中出现的先后顺序来构建实体对,若实体 e_i 在句中出现在实体 e_j 之前,则可构建一个实体对$\langle e_i, e_j\rangle$;设实体 e_i 和 e_j 的依存句法关系分别为 $e_i.dp$ 和 $e_j.dp$,则实体对$\langle e_i, e_j\rangle$的依存句法关系组合为 $e_i.dp$-$e_j.dp$.考虑实体的顺序,目的是为了使依存句法关系组合具有更好的可解释性.

例如,图1中的"邓兆祥游览庐山.",实体"邓兆祥"的依存关系为SBV,"庐山"的依存关系为VOB,按照实体出现的先后顺序构成的实体对〈邓兆祥,庐山〉的依存句法关系组合为SBV-VOB,即第1个实体"邓兆祥"在句中做主语,第2个实体"庐山"做宾语.从图1分析结果可以看出,句子的核心谓词为"游览",主语"邓兆祥"和宾语"庐山"均与动词"游览"有联系.因此,实体对〈邓兆祥,庐山〉之间存在"游历"关系.

以庐山数据集为例,通过对2个实体的依存句法关系进行组合,共得到依存句法关系组合(简称为组合类型)为64种.在庐山数据集上各种依存句法关系组合的Top-15实体对占比分布如图3所示.其中,横坐标为依存句法关系组合类型,纵坐标为占比.

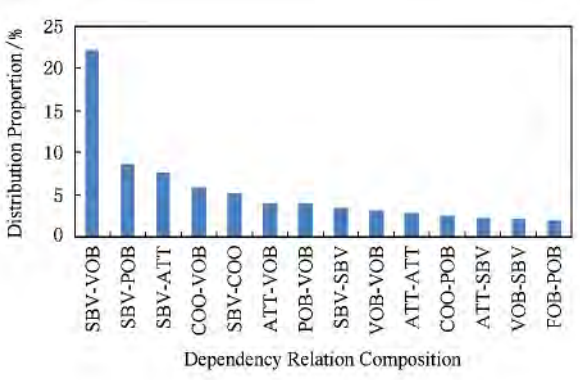

Fig. 3 Top-15 proportion distribution of entity pairs on Mount Lushan corpus.

图3 庐山数据集上依存句法关系组合的Top-15实体对占比分布

从图3可以看出,各种依存句法关系组合的实体对占比与文献[21]指出的"在关系抽取时,对实体关系有表征作用的句法结构主要有主谓宾关系、介词宾语、并列成分和修饰关系"的观点相吻合.

由于庐山数据集的句式基本上为陈述句和主动句,因此具有SBV-VOB组合的实体对的数量最多,在"考察访问"、"游历"、"参与"、"居住"等大多数实体关系类型中占主要部分.又由于具有SBV-VOB组合的实体对的数量最多,因此SBV-VOB组合不具有特别强的区分性.

然而,对于一些实体关系类型,其实体对的依存句法关系组合有比较明显的差异.例如,在庐山数据集上的"位于"关系中,SBV-POB组合占首位;在"发生"关系中,FOB-POB组合的实体对数量最多;特别

① http://www.ltp-cloud.com/demo/

是 VOB-VOB, POB-VOB 和 VOB-SBV 组合特征绝大部分只出现在“无关系”类型中. 以图 3 中的部分依存句法关系组合为例,说明一些依存句法关系组合在不同实体关系类型中的分布也不同,如表 3 所示:

Table3 Distribution of Partial Composition Types in Most Frequent Entity Relationships on Mount Lushan Corpus

表 3 庐山数据集上最频繁实体关系中部分组合类型的分布

DRC	ENO_1	E-R	ENO_2	ENO_3	C_1/%	C_2/%
FOB-POB	87	Happening	70	178	80.5	39.3
VOB-SBV	95	None	72	1642	75.8	4.4
VOB-VOB	140	None	96	1642	68.6	5.8
SBV-COO	224	Travel	80	353	35.7	22.7
SBV-POB	375	Location	109	228	29.1	47.8

在表 3 中,DRC 表示实体对的依存句法关系组合(简称为组合类型);ENO_1 表示在数据集中属于该组合类型的实体对数量;E-R 表示该组合类型的实体对中出现最多的实体关系类型(简称为最频繁实体关系类型);ENO_2 表示该组合类型的实体对中属于最频繁实体关系类型的实体对数量;ENO_3 表示数据集中包含的属于最频繁实体关系类型的实体对总数量;$C_1 = ENO_2/ENO_1 \times 100\%$,表示在该组合类型的实体对中属于最频繁实体关系类型的实体对的占比;$C_2 = ENO_2/ENO_3 \times 100\%$,表示在所有最频繁实体关系类型的实体对中属于该组合类型的实体对的占比. 从表 3 可以看出:

1) 在庐山数据集上,FOB-POB 的实体对绝大多数出现在“发生”关系中,占该关系实体对总数的 80.5%. 在“发生”关系的实体对中属于 FOB-POB 实体对占比达 39.3%,位居第一,其原因在于“发生”关系表示某个活动在某个景点发生. 如“中华世纪柏取土仪式在庐山举行.”中的实体对〈中华世纪柏取土仪式,庐山〉属于“发生”关系.

2) 同理,SBV-POB 在“位于”关系中出现最多,且该依存句法关系组合的实体对数量占“位于”关系的榜首. SBV-POB 为主谓-介宾组合,很好地反应了某人/组织在某个景点的“位于”关系.

3) 而 VOB-SBV 和 VOB-VOB 在“无关系”类型中出现最多,因为“无关系”类型反映的是 2 个实体之间不存在任何语义关系. 由于这 2 个依存句法关系组合不属于常用的句法结构,2 个实体之间一般较少发生关系. 一方面,VOB-SBV 和 VOB-VOB 的实体对均在“无关系”类型中所占比例不高,其原因在于“无关系”类型的句法结构比较杂乱,属于“无关系”类型的实体对的依存句法关系组合特征的取值数高达 52. 另一方面,从 C_1 值可以看出,VOB-SBV 和 VOB-VOB 组合对“无关系”类型还是具有一定的区分度.

总体来看,2 个实体的依存句法关系组合特征在不同实体关系类型中的分布上具有差异性,对实体关系探测和关系抽取具有一定的区分度. 基于上述分析,本文考虑将实体对的依存句法关系组合作为实体关系中的句法特征进行考量.

2.2 最近句法依赖动词特征

本文的目标是通过依存句法分析和词性来提取一个句子中 2 个实体的最近句法依赖动词特征. 根据前面对陈述句和主动句的特点分析可知,通过最近句法依赖动词可以使 2 个实体之间直接或间接地发生语义关联,且作为主语或前置宾语的第 1 个实体为该动词所表示动作行为的发出者. 存在直接语义关联和存在间接语义关联的 2 个实体之间,它们的依存路径有所不同,具体分析如下.

1) 直接语义关联实体间的最近句法依赖动词

如果实体对〈e_i, e_j〉中的 2 个实体能够通过一个最近句法依赖动词直接发生语义关联,则它们之间存在着一条满足如图 4 所示的依存句法路径. 其中,节点表示实体或动词,有向边表示从动词节点指向实体节点,边上的内容表示实体节点与动词节点之间的依存关系.

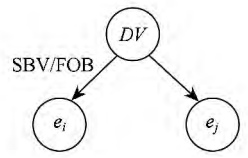

Fig. 4 Dependency paths of an entity pair with direct semantic association.

图 4 实体对直接语义关联的依存句法路径

提取使 2 个实体发生直接语义关联的最近句法依赖动词的步骤为:①找出与第 2 个实体 e_j 直接发生依存关系的动词 V_j;②找出与第 1 个实体 e_i 直接发生 SBV 或 FOB 依存关系的动词 V_i;③判断 V_j 与 V_i 是否为同一动词,若相同则实体对〈e_i, e_j〉的最近句法依赖动词 DV 为 V_j,否则置为空值 Null.

例 4. “蒋介石兴致勃勃离开庐山.”的依存句法分析和词性标注如图 5 所示. 其中,i 表示词性标注中的成语或习语.

对于图 5 中的实体对〈蒋介石,庐山〉,第 2 个实体“庐山”是动词“离开”的宾语,第 1 个实体“蒋介石”

是动词“离开”的主语,即 SBV 关系. 因此,实体对〈蒋介石,庐山〉的最近句法依赖动词为“离开”. 该动词很好地表征了实体对〈蒋介石,庐山〉之间的“离开”关系.

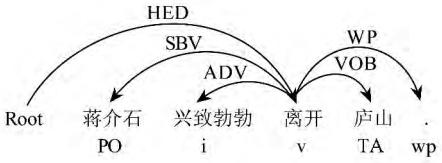

Fig. 5 Dependency parsing and POS tagging of Exp. 4.

图 5 例 4 的依存句法分析和词性标注

2) 间接语义关联实体间的最近句法依赖动词

如果实体对$\langle e_i, e_j \rangle$中的 2 个实体能够通过一个最近句法依赖动词间接发生语义关联,则它们之间存在着一条如图 6 所示的依存句法路径,其中节点表示动词、实体或非实体类型的名词. 图 6 中的依存句法路径可以分为 2 个部分:实体部分和动词部分.

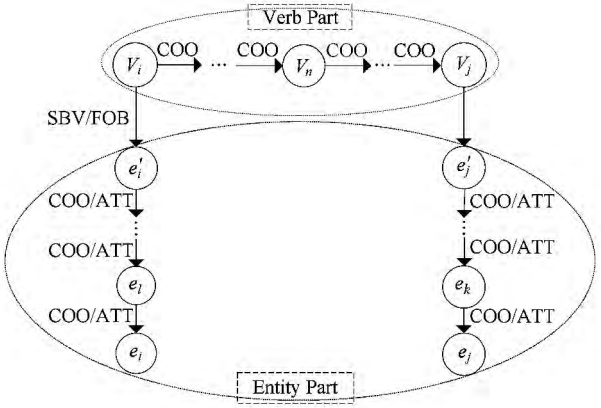

Fig. 6 Dependency paths of an entity pair with indirect semantic association.

图 6 实体对间接语义关联的依存句法路径

从图 6 可以看出,本文中的间接语义关联分为以下 2 种情况:

1) 在实体部分,实体的依存关系结构可以分为 2 类:

① 实体为并列结构. 并列结构的实体一般都是类型相同,合在一起表示一个特定的意思. 通过 COO 并列结构发现,选择与实体发生 COO 关系且依存关系距离最远的实体作为其依存关联节点.

② 实体为定中结构. 定中结构的修饰词语叫定语,被修饰词语叫中心词语,中心词语在句子中可充当主语或宾语. 通过 ATT 结构发现,选择与实体发生 ATT 关系且依存关系距离最远的非实体名词作为其依存关联节点.

2) 在动词部分,动词结构为并列结构. 如果存在多个并列结构时,通过 COO 并列结构发现,选择与第 2 个实体发生依存关系且依存距离关系最近的动词作为最近句法依赖动词.

例 5. “1927 年 1 月 28 日,蒋介石、张静江、张群、黄郛等人踏雪游览庐山风光.”的依存句法分析和词性标注如图 7 所示.

为了更清楚地显示 2 个实体对〈蒋介石,庐山〉和〈张静江,庐山〉的依存路径,本文将图 7 转换成二叉树的结构图,如图 8 所示.

对于图 8 中的实体对〈蒋介石,庐山〉:第 2 个实体“庐山”的依存关系为 ATT,是非实体名词“风光”的定语,选择名词“风光”作为“庐山”的父节点. “风光”是动词“游览”的宾语. 因此,第 2 个实体“庐山”通过“风光”间接与动词“游览”发生语义关联. 同理,第 1 个实体“蒋介石”也与非实体名词“人”发生 ATT 关系. “人”是动词“踏雪”的主语,即与名词“人”发生 SBV 依存关系的动词为“踏雪”. 动词部分为并列结构,“踏雪”和“游览”为 COO 结构,选择与第 2 个实体“庐山”最早发生依存关系的动词“游览”作为该实体对〈蒋介石,庐山〉的最近句法依赖动词特征. 最近句法依赖动词“游览”很好地辨别出实体对〈蒋介石,庐山〉之间的“游历”关系.

对于图 8 中的实体对〈张静江,庐山〉:第 2 个实体“庐山”与上述分析相同,第 1 个实体“张静江”与实体“蒋介石”为 COO 并列结构. 根据上述分析,

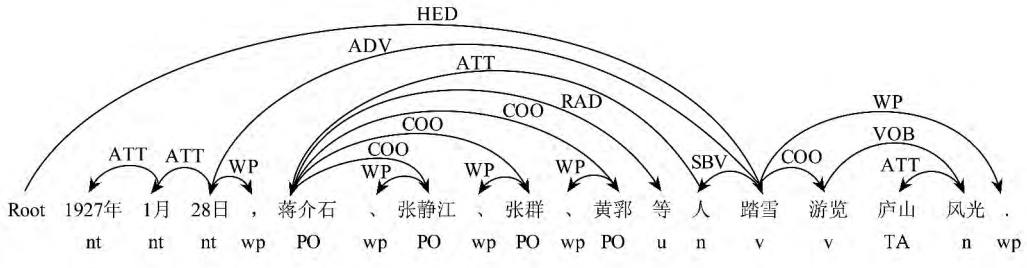

Fig. 7 Dependency parsing and POS tagging of Exp. 5.

图 7 例 5 的依存句法分析和词性标注

第 1 个实体“张静江”通过实体“蒋介石”与动词“踏雪”发生间接语义关联. 而动词部分为并列结构,与上述分析相同,因此,也选择与第 2 个实体“庐山”最早发生依存关系的动词“游览”作为该实体对〈张静江,庐山〉的最近句法依赖动词特征.

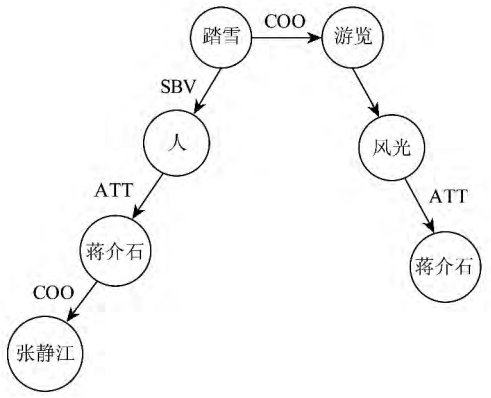

Fig. 8 Binary tree structure of dependency paths of an entity pair.

图 8 一个实体对的依存路径的二叉树结构

综上所述,根据依存句法分析和词性,实体对$\langle e_i, e_j\rangle$的最近句法依赖动词特征的提取包括 4 步(如算法 1 所示):

步骤 1. 分别提取与实体 e_i 或 e_j 存在 COO 并列结构或 ATT 定中结构关系的依存关联节点 e_i' 和 e_j',如算法 2 所示.

步骤 2. 提取与第 2 个实体 e_j 的依存关联节点 e_j' 发生依存关系的最近动词 V_j,如算法 3 所示.

步骤 3. 获取与第 1 个实体 e_i 的依存关联节点 e_i' 发生 SBV 或 FOB 关系的最近动词 V_i,如算法 4 所示.

步骤 4. 通过判断动词 V_i 与 V_j 是否为同一个动词或为 COO 并列结构关系,确定该实体对$\langle e_i, e_j\rangle$的最近句法依赖动词 DV.

利用上述算法提取实体关系类型中的高频最近句法依赖动词信息,提取结果如表 4 所示. 其中,Null 表示实体对的最近句法依赖动词为空.

算法 1. 提取最近句法依赖动词.

输入:实体对$\langle e_i, e_j\rangle$,该句的依存句法分析和词性标注结果;

输出:实体对$\langle e_i, e_j\rangle$的最近句法依赖动词 DV.

```
DV=Null;   /*初始化*/
/*提取第 1 个实体 e_i 的依存关联节点 e_i' */
IF (e_i→relate 为 COO)或(e_i→relate 为 ATT
    且 e_i→parent 为非实体名词)
  e_i'=LinkPareMethod(e_i); /*调用算法 2*/
ELSE
  e_i'=e_i;
ENDIF
/*提取第 2 个实体 e_j 的依存关联节点 e_j' */
IF (e_j→relate 为 COO)或(e_j→relate 为 ATT
    且 e_j→parent 为非实体名词)
  e_j'=LinkPareMethod(e_j); /*调用算法 2*/
ELSE
  e_j'=e_j;
ENDIF
/*提取与第 2 个实体 e_j 的依存关联节点 e_j' 发
    生依存关系的最近动词 V_j */
V_j=SENFirsVerd(e_j'); /*调用算法 3*/
/*获取与第 1 个实体 e_i 的关联节点 e_i' 发生
    SBV 或 FOB 关系的动词 V_i */
V_i=FENFirsVerd(e_i');   /*调用算法 4*/
/*通过判断动词 V_i 与 V_j 的关系,确定实体对
    的最近句法依赖动词 DV*/
IF (V_j !=Null)
  IF (V_i==V_j) /*V_i 与 V_j 为同一个动词*/
    DV=V_j;
  ELSE
    p=V_j→parent;
    V_k=V_j;
    WHILE (p→id>-1) /*当前节点不为
          根节点 Root*/
      IF (V_k→relate 为 COO) /*V_i 与 V_j 存
        在 COO 并列关系*/
        IF (V_i==V_k→parent)
          DV=V_j; /*选择动词 V_j 为最近
                直接依赖动词*/
          BREAK;
        ELSE /*遍历 p 的父节点*/
          V_k=p;
          p=p→parent; /*将 p 的父节点
              设置为当前节点*/
        ENDIF
      ENDIF
    ENDWHILE
  ENDIF
ENDIF
RETURN DV.
```

算法 2. 提取实体的依存关联节点.

输入:一个实体 e,该实体所在句子的依存句法分析和词性标注结果;

输出:实体 e 的依存关联节点.

```
p=e→parent;
WHILE (p→relate 为 COO)或 (p→relate 为
        ATT 且 p→parent 为非实体名词)
  /*如果父节点的依存关系为 COO 或 ATT,
    则继续循环*/
  p=p→parent;
ENDWHILE
RETURN p.
```

算法 3. 提取与第 2 个实体发生依存关系的距离最近动词.

输入:实体对中第 2 个实体的依存关联节点 e,该实体所在句子的依存句法分析和词性标注结果;

输出:实体对中第 2 个实体的依存关系距离最近的动词 V.

```
V=Null; /*初始化*/
p=e→parent;
                /*设置当前节点为 e 的父节点*/
WHILE (p→id>-1) /*当前节点不为根节
        点 Root*/
  IF (p 为动词节点)
    V=p→verb; /*p→verb 为与实体 e 依存
      关系距离最近的动词*/
    BREAK;
  ELSE /*如果当前节点不是动词节点,则设
        置其父节点为当前节点*/
    p=p→parent;
  ENDIF
ENDWHILE
RETURN V.
```

算法 4. 提取与第 1 个实体发生 SBV 或 FOB 关系的距离最近动词.

输入:实体对中第 1 个实体的依存关联节点 e,该实体所在句子的依存句法分析和词性标注结果;

输出:实体对中第 1 个实体的依存关系距离最近的动词 V.

```
V=Null; /*初始化*/
p=e→parent;
                /*设置当前节点为 e 的父节点*/
WHILE (p→id>-1) /*当前节点不为根节
        点 Root*/
  IF (p 为动词节点) 且 (e→relate 为 SBV 或
      FOB)
    V=p→verb; /*p→verb 为与实体 e 依存
      关系距离最近的动词*/
    BREAK;
  ELSE
    p=p→parent;
  ENDIF
ENDWHILE
RETURN V.
```

Table 4 The NSDV with High Frequency in Entity Relationships on Mount Lushan Corpus

表 4 庐山数据集上实体关系的高频最近句法依赖动词

Entity Relationship	NSDV(nearest syntactic dependency verb) with High Frequency
Travel	参观、游览、游、观赏、抵
Visit	登临、来、考察、视察、参观
Living	下榻、住、来、夜宿、避难
Build	建造、修建、兴建、建立、建
Participate	出席、参加、召开、前进、举行
Happening	举行、召开、开幕、闭幕、落下
Arriving	上、登上、来到、来、到
Leave	离开、下、出发、辞别、告辞
Creation	题写、写、题词、拍摄、书写、创作
None	Null、出席、题词、参加

从表 4 可以看出,最近句法依赖动词算法能够准确地捕获到体现 2 个实体之间关系类型的相应动词.因此,对于每个实体关系类型中出现的高频最近句法依赖动词,大都很好地表征了该实体关系类型,并且具有较强的区分度,有利于提高实体关系探测和关系抽取性能.

3 实验结果与分析

3.1 实验数据集

实验数据采用来自不同旅游网站上与旅游景点有关的人文历史信息,它综合概述了在某个景点发生的事情,包含了丰富的人物/组织与景点之间的关系,为抽取景点人文信息提供了可靠数据来源.为了验证本文方法在多样性数据上的有效性,采用了 3 个数据集进行实验:

1) 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.

2) 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.

3) 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.

本文实验的3个数据集的特点是:文档句子多为复杂长句,一个句子中经常会出现多个人物、组织或景点.

本文只关注旅游景点(TA)、人物/组织(PO)、作品(WOR)、活动(ACT)共4类实体. 本文中的实体包括具体实体和泛指实体,如“庐山花径”为具体TA实体,而“五国使者”为泛指PO实体. 考虑到人物实体中的团队名称和组织名称有时很难区分,并且本文不考虑人物与组织之间的关系,故将人物实体和组织实体统一归为PO类型实体. 实体类型信息如表5所示:

Table 5 Information of Entity Types

表5 实体类型信息

Entity Type	Symbol	Entity Scope
People/Organization	PO	个人、团队、组织等
Tourist Attraction	TA	与景点有关的名称、别墅、宾馆等
Works	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
Action	ACT	会议、战争、比赛、谈判、仪式、论坛、表演等

对3个实验数据集,在利用哈尔滨工业大学LTP-Cloud[22]平台进行分词、词性标注、句法分析和实体识别的基础上,再采用基于规则的方法进行适当修订,以便更好地符合旅游领域特点. 本文是在实体识别正确的基础上进行实体关系抽取,因此对实体识别的方法本文不加以叙述.

本文只考虑一个句子中的2个实体之间的显性关系,而不考虑跨句子的实体关系和隐性关系. 本文主要关注人物/组织(PO)与景点(TA)、人物/组织(PO)与活动(ACT)、人物/组织(PO)与作品(WOR)以及景点(TA)与活动(ACT)之间的关系探测和关系抽取,而不考虑同类实体之间的关系. 因此,如果句子中只有一个实体或者无实体,则说明此句中不存在实体关系,需要过滤此句. 对于存在2个及以上实体的句子,首先按照实体在句中出现的顺序进行两两组合,生成候选实体对,然后根据实体关系类型加入实体对类型约束条件进一步进行实体对的筛选. 实体对类型约束条件为:{〈PO,TA〉,〈TA,PO〉,〈PO,ACT〉,〈ACT,PO〉,〈PO,WOR〉,〈WOR,PO〉,〈TA,WOR〉,〈WOR,TA〉,〈TA,ACT〉,〈ACT,TA〉}.

为了选择出黄金标准集,本文选用3个人作为实体关系类型的标注者,以少数服从多数决定正确答案,当3人的答案都不一致时,则由3人讨论确定最终标注结果. 3个实验数据集中的实体关系类型数据信息具体如表6所示. 其中,“—”表示数据集中没有包含该实体关系类型的数据.

Table 6 Statistics Information of Entity Relationships of Three Corpuses in Tourism Domain

表6 旅游领域3个数据集的实体关系统计信息

Entity Relationship	Symbol	Number of Entity Pair		
		Mount Lushan Corpus	Jinggangshan Corpus	Mount Taishan Corpus
Location	LOC	228	109	109
Travel	TRA	353	100	638
Visit	VIS	387	46	272
Living	LIV	120	—	30
Build	BUI	47	—	125
Participate	PAR	361	241	292
Creation	CRE	426	194	356
Happening	HAP	178	36	101
Arriving	ARR	230	144	77
Leave	LEA	51	28	—
Other	OTH	310	107	194
None	NON	1 642	685	712
Total	12	4 333	1 690	2 906

3.2 评测指标

实验评测采用常用的评价指标:准确率P、召回率R和$F1$值,针对某一实体关系类型的抽取结果,具体评价公式为:

$$P=\frac{\text{结果中正确标注为给定关系类型的个数}}{\text{结果中标注为给定关系类型的总个数}};\quad(1)$$

$$R=\frac{\text{结果中正确标注为给定关系类型的个数}}{\text{测试集中给定关系类型的总个数}};\quad(2)$$

$$F1=\frac{2\times P\times R}{P+R}.\quad(3)$$

3.3 实验设计与结果分析

本实验共采用了3个数据集,对每个数据集进行随机选择其中的80%作为训练集,剩余20%为测试集. 本文采用台湾大学林智仁等人开发的LIBSVM[16]作为SVM工具包进行实验. 为了验证本文提出的2个特征的有效性,并与其他同类方法

进行比较,本文从关系探测和关系抽取2个角度进行实验分析.

3.3.1 关系探测

由于实验数据中存在着大量的"无关系"类型的实体对,因此从关系探测的角度分析本文提出的2个特征对系统性能的影响.关系探测的目的是为了识别一个实体对之间是否存在语义关系,属于二元分类问题.因此,本文将实验数据分成2类:将"无关系"(NON)类型单独作为一类,剩余的11类实体关系均存在语义上的关系,因此合并成另一个类,即为"有关系"(HAS)类型.3个实验数据集中的实体有无关系类型数据信息具体如表7所示:

Table 7 Entity Relationship Information for Relationship Detection in Tourism Domain

表7 旅游领域关系探测中的实体关系信息

Entity Relationship	Number of Entity Pairs		
	Mount Lushan Corpus	Jinggangshan Corpus	Mount Taishan Corpus
HAS	2691	1005	2194
NON	1642	685	712
Total	4333	1690	2906

已有许多学者对实体关系抽取进行了研究,结果已经表明实体特征、实体上下文特征、位置特征以及句法特征的有效性[3-4,16,18].因此,本文选取的实体关系基本特征包括实体类型组合、实体上下文中每个实体的左边和右边各2个词和它们的词性组合、实体间距离以及依存句法关系特征,具体如表8所示.

在上述特征中选择最佳特征组作为本文实验的基本特征,在基本特征上加入本文提出的2个新特征,分析每个新特征对关系探测所做的贡献,如表9所示.其中,"+"表示在基本特征基础上加入新的特征.

由表9可以看出,在3个数据集上,选择实体上下文信息作为特征的实体关系探测性能最差.实验中发现,只要使用实体上下文特征,实体关系探测的性能则会大大降低.这是由于每个实验数据集描述的都是某人物/组织与旅游景点发生的关系.因此,景点实体会反复出现,导致实体上下文特征不具有很好的区分性,而且该特征的数量比较多,也会导致关系探测性能低下.因此,选取实体类型组合和实体间距离作为基本特征,其关系探测性能作为基准线(baseline).

在基本特征基础上,分别加入依存句法关系特征、依存句法关系组合特征、最近句法依赖动词特征后,关系探测性能都有所提高.具体分析如下:

1) 相对于基本特征来说,分别依存句法关系特征、依存句法关系组合特征、最近句法依赖动词特征后,在3个数据集上的关系探测性能都有所提高,说明了依存句法关系特征以及本文提出的2个特征(依存句法关系组合特征、最近句法依赖动词特征)在关系探测上的有效性.

2) 与依存句法关系特征相比,在庐山和井冈山这2个数据集上,利用依存句法关系组合特征进行关系探测,准确率P、召回率R和$F1$值都有较大提高;但对泰山数据集,关系探测性能稍微有所降低.总体上说明了依存句法关系组合特征在关系探测上

Table 8 Basic Features of Entity Relationships

表8 实体关系基本特征

Feature Type	Symbol	Description
Combination of entity types	$e_1.type$-$e_2.type$	Combination of entity types for an entity pair $\langle e_1, e_2\rangle$
Entity contexts	$e_1.lw1$-$e_1.lpos1$	Combination of the first word and its POS on the left of entity e_1
	$e_1.lw2$-$e_1.lpos2$	Combination of the second word and its POS on the left of entity e_1
	$e_1.rw1$-$e_1.r\ pos1$	Combination of the first word and its POS on the right of entity e_1
	$e_1.rw2$-$e_1.r\ pos2$	Combination of the second word and its POS on the right of entity e_1
	$e_2.lw1$-$e_2.lpos1$	Combination of the first word and its POS on the left of entity e_2
	$e_2.lw2$-$e_2.lpos2$	Combination of the second word and its POS on the left of entity e_2
	$e_2.rw1$-$e_2.r\ pos1$	Combination of the first word and its POS on the right of entity e_2
	$e_2.rw2$-$e_2.r\ pos2$	Combination of the second word and its POS on the right of entity e_2
Distance between entities	$dist$	Number of words between an entity pair
Dependency relations	$e_1.dvalue$	Value of dependency relation for entity e_1
	$e_2.dvalue$	Value of dependency relation for entity e_2

Table 9 Contribution of Different Features for Relationship Detection

表 9 本文特征在关系探测中所做的贡献 %

Feature	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Entity contexts	14.35	37.88	20.81	16.43	40.53	23.38	5.99	24.28	9.61
Combination of entity types+Distance between entities (baseline)	75.31	75.64	75.47	77.50	76.92	77.21	82.35	83.10	82.72
+Dependency relations	77.58	77.83	77.70	81.60	81.66	81.62	84.34	84.48	84.41
+Dependency relation composition	78.91	79.10	79.00	83.90	82.54	83.22	83.73	84.48	84.10
+Nearest syntactic dependency verb	81.39	81.06	81.17	80.06	80.18	80.12	86.25	86.55	86.40
+Dependency relation composition+Nearest syntactic dependency verb	83.84	83.49	83.66	84.55	84.62	84.58	84.73	85.34	85.04

更有效. 其原因在于依存句法关系组合特征在“有关系”、“无关系”类型上的差异性比较大,具有更好的区分性,因此有利于提高关系探测性能.

3) 相对于其他特征来说,最近句法依赖动词特征整体上对系统性能提高的幅度最大. 加入最近句法依赖动词特征后,在庐山数据集上的准确率 P、召回率 R 和 $F1$ 值均超过 81%;在泰山数据集上的准确率 P、召回率 R 和 $F1$ 值均超过 86%. 说明了最近句法依赖动词特征能有效地提升关系探测系统的性能,其原因在于最近句法依赖动词有效地区分了实体之间有无语义关系,特别是特征值为 Null 的实体对几乎都属于“无关系”类型. 然而,对于井冈山数据集,最近句法依赖动词特征对系统性能的提高幅度不如依存句法关系特征和依存句法关系组合特征. 其原因在于该数据集上很多实体对之间不存在能够准确表征实体关系类型的动词,而通过最近句法依赖动词算法提取到特征值为 Null 的数量太多,共有 936 个,而该数据集中的“无关系”实体对个数实际上仅为 685 个. 因此导致最近句法依赖动词特征对于关系探测性能提高的效果没有上述 2 个特征明显.

4) 在基本特征的基础上,同时结合依存句法关系组合特征和最近句法依赖动词特征,在庐山和井冈山 2 个数据集上的效果表现最佳,从总体上证明了这 2 个特征在关系探测上的有效性.

为了验证本文新增特征在中文旅游领域关系探测上的有效性,在使用本文实验数据情况下,用本文提出的方法与同类方法进行比较. 郭喜跃等人在文献[6]中提出了一种基于句法语义特征的实体关系抽取方法,新增了依存句法关系、核心谓词、语义角色标注等特征,也考虑了“无关系”类型. 文献[8]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. 因此,在基本特征的基础上分别加入文献[6]、文献[8]和本文提出的特征,得到 3 组实验结果分别为郭方法(Guo[6])、董方法(Dong[8])和本文方法(Ours). 具体特征为:

1) 基本特征. 实体关系类型组合、实体间距离.

2) 董方法特征. 基本特征和依赖动词特征.

3) 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.

4) 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.

这 3 种方法在关系探测上的性能如表 10 所示:

Table 10 Comparison of Our System with Other Similar Systems for Relationship Detection

表 10 本文方法与同类方法在关系探测的实验结果对比 %

Approach	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Baseline	75.31	75.64	75.47	77.50	76.92	77.21	82.35	83.10	82.72
Guo[6]	79.71	79.68	79.04	81.59	81.66	81.62	86.55	86.70	86.63
Dong[8]	83.29	82.91	83.10	80.79	80.77	80.78	5.99	24.48	9.63
Ours	83.84	83.49	83.66	84.55	84.62	84.58	84.73	85.34	85.04

从表 10 可以看出,本文方法在中文旅游领域关系探测任务中总体上性能最好. 具体分析如下:

1) 与董方法对比,利用本文方法进行中文旅游领域关系探测在 3 个数据集上的性能都更优,特别

是在井冈山和泰山 2 个数据集上的效果更为明显.具体分析如下:

① 对于庐山数据集,本文方法的准确率 P、召回率 R 和 $F1$ 值分别提高了 0.55,0.58 和 0.56 个百分点.在井冈山数据集上,本文方法的准确率 P、召回率 R 和 $F1$ 值的提高幅度较大,分别为 3.76,3.85 和 3.80 个百分点.特别是在泰山数据集上,本文方法在关系探测上的性能远远优于董方法,其准确率 P、召回率 R 和 $F1$ 值的提高幅度高达 78.74,60.86 和 75.41 个百分点.说明了本文的 2 个特征——依存句法组合特征和最近句法依赖动词特征——对关系探测性能的提高起到了很好的作用.

② 对于泰山数据集,在基本特征集上加入董方法提取的依赖动词特征后,大大降低了关系探测的性能,其准确率 P、召回率 R 和 $F1$ 值都远远低于基准线.其原因在于:i)董方法的依赖动词特征提取算法几乎对每一个实体对都提取了动词,而该动词特征对于有无关系类型的区分度不强,几乎不能起到区分作用.ii)利用董方法提取到的动词特征数量很多,占总特征数的 86%,不具有强区分性的动词特征却带来了很多噪音干扰.这 2 个方面原因导致了董方法在关系探测时将实体对大部分都分到了"无关系"类型,因此,对实体间有无关系根本没法起到辨别的作用.然而,在基本特征集上,加入本文的 2 个特征后能有效地提高关系探测的性能(相对于基准线).说明了本文提出的最近句法依赖动词特征能够有效地表征实体之间有无语义关系,特别是 Null 动词,具有很强的区分性,大大减少了文献[8]的依赖动词特征带来的噪音,显著地提高了关系探测的性能.同时也说明了本文提出的最近句法依赖动词特征比董方法的依赖动词特征更具有鲁棒性.

2) 与郭方法相比,虽然本文方法在泰山数据集上的关系探测性能略有下降,然而在庐山和井冈山 2 个数据集上的关系探测性能效果更佳,其准确率 P、召回率 R 和 $F1$ 值的提高幅度分别为 4.13,3.81,4.62 个百分点和 2.96,2.96,2.96 个百分点.总体而言,证明了本文提出的 2 个特征在关系探测上的有效性.

3.3.2 关系抽取

为了验证本文提出的 2 个特征在中文旅游领域关系抽取的有效性,在使用本文 3 个实验数据集的情况下,用本文提出的方法与上述关系探测使用的同类方法(即郭方法[6]和董方法[8])进行比较.将这 3 种方法应用于关系抽取,其整体性能如表 11 所示:

Table 11 Comparison of Our System with Other Similar Systems for Relationship Extraction

表 11 本文方法与同类方法在关系抽取的实验结果对比 %

Approach	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
Guo[6]	54.35	59.38	56.75	70.83	69.76	70.29	62.85	64.64	63.73
Dong[8]	66.79	67.90	67.34	71.87	59.88	65.33	24.92	31.37	27.78
Ours	68.91	69.18	69.04	80.76	77.84	79.27	75.02	73.67	74.34

从表 11 可以看出,本文方法在中文旅游领域关系抽取任务中取得了最好的性能.具体分析如下:

1) 与郭方法相比,本文方法在 3 个数据上的关系抽取性能都更佳.本文方法在井冈山和泰山这 2 个数据集上的关系抽取性能提升较为明显,其准确率 P、召回率 R 和 $F1$ 值分别提高了 9.93,8.08,8.98 个百分点和 12.17,9.03,10.61 个百分点.在庐山数据集上,本文方法对系统性能的提高幅度最大,其准确率 P、召回率 R 和 $F1$ 值分别提高了 14.56,9.8 和 12.29 个百分点.其原因在于郭方法较依赖于实体关系类型的数量分布,对于数据量少的实体类型不能进行有效地抽取.例如,郭方法对于庐山数据集上的"居住"、"建立",井冈山数据集上的"考察访问"以及泰山数据集上的"居住"、"建立"等关系类型无法进行识别,其准确率 P、召回率 R 和 $F1$ 值均为 0.从表 6 中 3 个数据集的实体关系类型信息分布可知,这些无法识别的关系类型的数据量较少,导致文献[6]中提出的特征无法起作用.而本文方法对 3 个数据集上的每一个关系类型均能有效地识别,从而提高了关系抽取的整体性能.因此,本文方法在 3 个数据集上的关系抽取都获得了最佳性能,说明了本文提出的依存句法关系组合特征和最近句法依赖动词特征能够有效地提高中文旅游领域实体关系抽取的性能.

2) 与董方法对比,本文方法对于庐山数据集的关系抽取性能提高较小,其准确率 P、召回率 R 和 $F1$ 值分别提高了 2.12,1.28 和 1.70 个百分点.在

井冈山和泰山这 2 个数据集上的准确率 P、召回率 R 和 $F1$ 值都有显著的提高，在井冈山数据集上分别提高了 8.89，17.96 和 13.94 个百分点. 特别是在泰山数据集上，在基本特征集上加入董方法的依赖动词特征后，大大降低了关系抽取的性能(相对于基准线)，说明文献[8]提出的依赖动词特征带来了太多的噪音信息；而在该数据集上，本文方法却表现出绝对的优势，在准确率 P、召回率 R 和 $F1$ 值上的提高幅度分别高达 50.10，42.30 和 46.56 个百分点. 同时，也说明了本文提出的 2 个特征更具有鲁棒性. 本文方法在 3 个数据集上的关系抽取整体性能明显优于董方法，其主要原因分析如下：

① 本文提出的最近句法依赖动词特征值 Null，有利于“无关系”类型实体对的识别. 此外，对比本文方法提取的最近句法依赖动词与董方法提取的依赖动词，在庐山、井冈山和泰山 3 个数据集上不相同的数量分别占 25.4%，32.5%和 33.2%. 这部分动词特征主要是影响“有关系”类型中的具体关系类型的判别，说明了最近句法依赖动词特征能有效地表征实体关系类型.

② 依存句法组合特征对关系抽取性能的提升起到了一定的作用.

为了验证依存句法关系组合特征、最近句法依赖动词特征对关系抽取的影响，在基本特征的基础上依次加入这 2 个特征，其关系抽取的整体性能如表 12 所示. 表 13 为本文各个特征在关系抽取中具体关系类型中的表现.

从表 9 与表 12 可以看出，在 3 个数据集上，关系探测的性能高于关系抽取的性能，其原因在于：关系探测只是一个二分类问题，用于确定一个实体对之间有无语义关系；而关系抽取则是一个多分类问题，用于确定实体对之间的关系属于哪一个具体类型，因而难度更高、性能更低.

从表 12 可以看出：

1) 依存句法关系组合特征对关系抽取性能的提高贡献比较明显. 相对于基准线来说，依存句法关系组合特征在 3 个数据集上的准确率 P、召回率 R 和 $F1$ 值提高幅度最大，分别为：准确率 P 在庐山数据集上的提高幅度高达 11.42 个百分点；召回率 R 在泰山数据集上提高了 6.24 个百分点；$F1$ 值在泰山数据集上提高了 7.29 个百分点. 这说明了依存句法关系组合特征能够较好地反映出相应实体之间的关系特征，有利于关系抽取准确率的提高.

2) 最近句法依赖动词特征对关系抽取性能的提高贡献最大. 相对于基准线来说，最近句法依赖动词特征在庐山、井冈山和泰山这 3 个数据集上的准确率 P、召回率 R 和 $F1$ 值的提高幅度依次为：23.50，9.97 和 17.75 个百分点(庐山)；24.46，14.37 和 19.65 个百分点(井冈山)；24.10，13.69 和 19.24 个百分点(泰山). 这是因为实体关系大多数是由动词触发的，因此最近句法依赖动词特征能够较好地表征实体之间的关系类型，具有较好的区分度.

3) 综合本文提出的 2 个新特征后，在准确率 P、召回率 R 和 $F1$ 值方面的表现都最佳，验证了本文提出的 2 个新特征在关系抽取上的有效性.

从表 13 可以看出：

1) 在 3 个旅游领域的实验数据集中，加入依存句法关系组合特征，对于某些特定语义关系的抽取，如“位于”关系、“游历”关系、“其他”关系以及“无关系”类型等，可有效地提高关系抽取的性能. 具体分析如下：

① 在 3 个数据集上，依存句法关系组合特征对于“游历”关系、“其他”关系和“位于”关系抽取效果的提高更为明显，其原因在于它们的关系实例数量相对较多，而且依存句法关系组合特征的取值数量少，具有较好的区分性. 例如，庐山数据集中的“游历”关系的依存句法关系组合特征的取值数量为 23，且主要包含 SBV-VOB，SBV-COO 和 SBV-ATT 等组合类型.

② 对于 3 个数据集中关系实例数量较少的关系类型，加入依存句法关系组合特征后依然无法识别出这些实体关系类型，如庐山数据集中的“离开”、“建立”和“居住”，井冈山数据集中的“考察访问”和“离开”，以及泰山数据集中的“居住”和“到达”等关系类型. 虽然这些关系类型的依存句法关系组合特征的取值数量较少，但是加入依存句法关系组合特征后依然没有提高系统性能，其原因在于这些关系类型的关系实例数量太少，并且不具普遍区分性的 SBV-VOB 等组合类型占主要地位，因此数据偏倚导致依存句法关系组合特征对这些关系类型的抽取没能发挥作用.

③ 在 3 个实验数据集上对“无关系”类型抽取性能提高不明显，其原因在于，虽然该类型的关系实例数量最大，但由于它的依存句法关系组合特征的取值数量也最多，且依存句法关系组合特征的不同取值分布比较均匀. 例如，庐山数据集中的“无关系”类型的依存句法关系组合特征的取值数量高达 52，而位居第一的 SBV-VOB 组合类型包含的实体对也

Table 12 Overall Performance of Different Features in Our System for Relationship Extraction

表 12 本文各个特征在关系抽取中的整体性能 %

Feature	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Baseline	43.92	58.40	50.14	51.74	59.28	55.25	46.86	55.46	50.80
+Dependency relation composition	55.34	58.98	57.10	60.42	63.17	61.76	54.87	61.70	58.08
+Nearest syntactic dependency verb	67.42	68.37	67.89	76.20	73.65	74.90	70.96	69.15	70.04
+Dependency relation composition+Nearest syntactic dependency verb	68.91	69.18	69.04	80.76	77.84	79.27	75.02	73.67	74.34

Table 13 Performance of Different Entity Relationships with Different Features in Our System for Relationship Extraction

表 13 本文各个特征在不同关系类型抽取中的性能 %

Feature	Entity Relationship	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Baseline	LOC	0.00	0.00	0.00	28.00	33.33	30.43	0.00	0.00	0.00
	TRA	28.00	10.00	14.74	0.00	0.00	0.00	40.26	99.21	57.27
	VIS	37.61	53.25	44.09	0.00	0.00	0.00	0.00	0.00	0.00
	LIV	0.00	0.00	0.00	—	—	—	0.00	0.00	0.00
	BUI	0.00	0.00	0.00	—	—	—	0.00	0.00	0.00
	PAR	73.49	84.72	78.71	69.12	97.92	81.03	78.08	98.28	87.02
	CRE	62.50	64.71	63.58	58.00	76.32	65.91	72.97	76.06	74.48
	HAP	88.24	85.71	86.96	100.00	57.14	72.73	100.00	70.00	82.35
	ARR	24.00	13.04	16.90	34.78	57.14	43.24	0.00	0.00	0.00
	LEA	0.00	0.00	0.00	0.00	0.00	0.00	—	—	—
	OTH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NON	55.65	84.15	66.99	69.34	69.34	69.34	66.99	48.59	56.33
+Dependency relation composition	LOC	50.94	60.00	55.10	32.14	42.86	36.73	40.00	19.05	25.81
	TRA	40.51	45.71	42.95	35.29	30.00	32.43	55.20	96.06	70.11
	VIS	37.25	49.35	42.46	0.00	0.00	0.00	31.25	9.26	14.29
	LIV	0.00	0.00	0.00	—	—	—	0.00	0.00	0.00
	BUI	0.00	0.00	0.00	—	—	—	0.00	0.00	0.00
	PAR	74.36	80.56	77.33	80.00	91.67	85.44	82.61	98.28	89.76
	CRE	64.77	67.06	65.90	55.56	78.95	65.22	73.17	84.51	78.43
	HAP	82.86	82.86	82.86	100.00	71.43	83.33	88.24	75.00	81.08
	ARR	71.43	10.87	18.87	45.45	53.57	49.18	0.00	0.00	0.00
	LEA	0.00	0.00	0.00	0.00	0.00	0.00	—	—	—
	OTH	27.78	8.06	12.50	33.33	4.76	8.33	30.00	7.89	12.50
	NON	64.02	78.66	70.59	74.26	73.72	73.99	59.21	63.38	61.22
+Nearest syntactic dependency verb	LOC	65.91	64.44	65.17	71.43	71.43	71.43	50.00	33.33	40.00
	TRA	85.42	58.57	69.49	78.95	75.00	76.92	68.97	94.49	79.73
	VIS	80.33	63.64	71.01	100.00	22.22	36.36	67.31	64.81	66.04
	LIV	87.50	29.17	43.75	—	—	—	100.00	33.33	50.00
	BUI	100.00	44.44	61.54	—	—	—	83.33	40.00	54.05
	PAR	73.91	70.83	72.34	69.12	97.92	81.03	79.17	98.28	87.69
	CRE	72.00	63.53	67.50	87.10	71.05	78.26	73.33	77.46	75.34
	HAP	85.71	85.71	85.71	100.00	71.43	83.33	88.24	75.00	81.08
	ARR	64.52	43.48	51.95	90.00	64.29	75.00	44.44	26.67	33.33
	LEA	100.00	70.00	82.35	100.00	40.00	57.14	—	—	—
	OTH	40.00	22.58	28.87	80.00	19.05	30.77	45.00	47.37	46.15
	NON	63.68	86.59	73.39	68.94	81.02	74.50	69.09	53.52	60.32

Continued (Table 13) %

Feature	Entity Relationship	Mount Lushan Corpus			Jinggangshan Corpus			Mount Taishan Corpus		
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
+Dependency relation composition+Nearest syntactic dependency verb	LOC	58.00	64.44	61.05	70.00	66.67	68.29	73.33	52.38	61.11
	TRA	69.01	70.00	69.50	82.35	70.00	75.68	80.54	94.49	86.96
	VIS	74.29	67.53	70.75	100.00	22.22	36.36	68.09	59.26	63.37
	LIV	77.78	29.17	42.42	—	—	—	100.00	33.33	50.00
	BUI	80.00	44.44	57.14	—	—	—	68.18	60.00	63.83
	PAR	77.33	80.56	78.91	84.91	93.75	89.11	78.87	96.55	86.82
	CRE	73.97	63.53	68.35	90.63	76.32	82.86	80.28	80.28	80.28
	HAP	88.24	85.71	86.96	100.00	57.14	72.73	85.00	85.00	85.00
	ARR	68.75	47.83	56.41	84.00	75.00	79.25	50.00	40.00	44.44
	LEA	87.50	70.00	77.78	100.00	40.00	57.14	—	—	—
	OTH	42.86	24.19	30.93	100.00	19.05	32.00	52.78	50.00	51.35
	NON	67.33	82.32	74.07	71.43	91.24	80.13	68.18	63.38	65.69

仅占7.7%,POB-VOB,VOB-VOB,VOB-SBV等组合类型包含的实体对分别占6.5%,5.8%和4.4%,从而导致具有较高区分度的依存句法关系组合特征对"无关系"类型抽取的贡献不是很明显.说明了依存句法关系组合特征对数据分布具有一定的依赖性.

2) 加入最近句法依赖动词特征,显著地提高了实体关系抽取系统的性能,准确率 P、召回率 R 和 $F1$ 值均得到大幅度的提升,证明了该特征的有效性.具体分析如下:

① 最近句法依赖动词特征能有效地提升关系实例数量在总数据集中较少的实体类型的抽取性能.例如,庐山数据集中的"离开"、"建立"和"居住"关系,井冈山数据集中的"考察访问"和"离开"关系,以及泰山数据集中的"居住"和"到达"关系,在使用基本特征以及加入依存句法组合特征时,这些类型的关系抽取性能均为0.这是由于这些关系类型的实例数量在总数据集中最少,且各类关系数据分布很不均匀,导致基本特征以及加入依存句法组合特征都很难将它们区分.但是,在加入最近句法依赖动词特征后,这些关系类型的准确率 P、召回率 R 和 $F1$ 均有显著的提高.例如,庐山数据集中的"离开"关系的准确率 P、召回率 R 和 $F1$ 分别达到了100%,70%和82.35%;"建立"关系的准确率 P、召回率 R 和 $F1$ 值分别达到了100%,44.44%和61.54%;"居住"关系的准确率 P、召回率 R 和 $F1$ 值分别达到了87.50%,29.17%和43.75%.井冈山数据集中的"考察访问"和"离开"关系的准确率 P、召回率 R 和 $F1$ 分别提升了100.00,22.22,36.36个百分点(考察访问关系)和100.00,40.00,57.14百分点(离开关系).泰山数据集中的"居住"和"到达"关系的准确率 P、召回率 R 和 $F1$ 值分别达到了100.00%,33.33%,50.00%(居住关系)和44.44%,26.67%,33.33%(到达关系).

② 对于庐山数据集中的"参与"关系和泰山数据集中的"发生"关系,在加入最近句法依赖动词特征后,其 $F1$ 值有所降低.其原因在于这2类关系的最近句法依赖动词特征中包含较多的Null值,而包含Null的实体对主要属于"无关系"类型,且该类型数据占总数据集的榜首,远远超过其他任何关系类型,因此容易将包含Null值的这2类关系误分为"无关系"类型,导致了这2类关系抽取性能的下降.

3) 在依存句法关系组合特征基础上进一步加入最近句法依赖动词特征后,其关系抽取性能有了明显的提升,说明了最近句法依赖动词特征可以减少不具有普遍区分度的依存句法关系组合特征带来的噪音.

总体来看,在基础特征的基础上综合加入依存句法关系组合特征和最近句法依赖动词特征后,其整体性能表现最优,说明了本文提出的这2个特征能有效地提升关系抽取性能.

4 结 论

中文长句的句式较复杂,经常包含多个实体的特点以及数据稀疏问题,给中文实体关系探测和关系抽取任务带来了挑战.为了解决上述问题,本文提出了一种基于句法语义特征的实体关系抽取方法.在传统特征基础上选择最佳特征组作为基本特征,然后进行扩展,利用依存句法分析和词性标注结果获取依存句法关系组合特征和最近句法依赖动词特征,

选择 SVM 作为机器学习的实现途径,以真实旅游领域文本作为语料进行实验,验证了该方法在关系探测和关系抽取上的有效性.

本文的主要创新工作包括:

1) 提出了句法特征——依存句法关系组合特征.通过将 2 个实体各自的依存句法关系进行组合,分析了不同实体关系类型的依存句法关系组合特征的差异性,有助于提升实体关系探测和关系抽取的性能.

2) 提出了语义特征——最近句法依赖动词特征.通过依存句法分析和词性标注来选择最近句法依赖动词特征,主要贡献在于:①最近句法依赖动词特征的 Null 值能有效地区分实体有无语义关系,有利于提高关系探测的性能;②最近句法依赖动词特征能够较好地表征实体关系类型,有利于具体关系类型的识别,而且较好地解决了数据分布不均衡带来的问题,能够显著地提升关系抽取的性能.

在未来的工作中,将进一步研究跨文档中隐式关系的抽取,从而挖掘出更多的实体关系.

参 考 文 献

[1] Xu Jian, Zhang Zhixiong, Wu Zhenxin. Review on techniques of entity relaion extraction [J]. New Technology of Library and Information Service, 2008, 24(8): 18-23 (in Chinese)
(徐健, 张智雄, 吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术, 2008, 24(8): 18-23)

[2] Che Wanxiang, Liu Ting, Li Sheng. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2005, 19(2): 1-6 (in Chinese)
(车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 1-6)

[3] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]//Proc of the ACL 2004 on Interactive Poster and Demonstration Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1-4

[4] Zhou G D, Su J, Zhang J, et al. Exploring various knowledge in relation extraction [C] //Proc of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2005: 427-434

[5] Xi Bin, Qian Longhua, Zhou Guodong, et al. The application of combined linguistic features in semantic relation extraction [J]. 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[J]. 中文信息学报, 2008, 22(3): 44-50)

[6] Guo Xiyue, He Tingting, Hu Xiaohua, et al. Chinese named entity relation extraction based on syntactic and semantic features [J]. 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[J]. 中文信息学报, 2014, 28(6): 183-186)

[7] Jiang J, Zhai C X. A systematic exploration of the feature space for relation extraction [C] //Proc of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07). Stroudsburg, PA: Association for Computational Linguistics, 2007: 113-120

[8] Dong Jing, Sun Le, Feng Yuanyong, et al. Chinese automatic entity relation extraction [J]. 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[J]. 中文信息学报, 2007, 21(4): 80-85)

[9] Chan Y S, Roth D. Exploiting background knowledge for relation extraction [C] //Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 152-160

[10] Sun A, Grishman R, Sekine S. Semi-supervised relation extraction with large-scale word clustering [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2011, 1: 521-529

[11] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction [C] //Proc of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Stroudsburg, PA: Association for Computational Linguistics, 2009: 209-212

[12] Qin B, Zhao Y, Ding X, et al. Event type recognition based on trigger expansion [J]. Tsinghua Science & Technology, 2010, 15(3): 251-258

[13] Li P F, Zhu Q M, Zhou G D. Using compositional semantics and discourse consistency to improve Chinese trigger identification [J]. Information Processing & Management, 2014, 50(2): 399-415

[14] Harrington P. Machine Learning in Action [M]. Greenwich, CT: Manning, 2012

[15] John C Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines [R]. Seattle: Microsoft Research, 2003

[16] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J/OL]. ACM Trans on Intelligent Systems and Technology, 2011, 2(3): 1-27. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[17] Li X, Lord D, Zhang Y, et al. Predicting motor vehicle crashes using support vector machine models [J]. Accident Analysis & Prevention, 2008, 40(4): 1611-1618

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

[19] Chen Y, Zheng Q, Zhang W. Omni-word feature and soft constraint for Chinese relation extraction [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 572-581

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

[21] Kang Lili. Research and implementation of open Chinese entity relation extraction [D]. Shenyang: Northeastern University, 2013 (in Chinese)
(**康丽丽. 开放式中文实体关系抽取的研究与实现**[D]. **沈阳: 东北大学**, 2013)

[22] Che W, Li Z, Liu T. Ltp: A Chinese language technology platform [C] //Proc of the 23rd Int Conf on Computational Linguistics: Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 2010: 13-16

Gan Lixin, born in 1982. PhD candidate. Lecturer at Jiangxi Science and Technology Normal University. Her research interests include information retrieval, information extraction and data mining, etc.

Wan Changxuan, born in 1962. Received his PhD degree in computer science from Huazhong University of Science and Technology in 2003. Professor and PhD supervisor at Jiangxi University of Finance and Economics. Senior member of China Computer Federation. His research interests include Web data management, sentiment analysis, data mining and information retrieval, etc.

Liu Dexi, born in 1975. Received his PhD degree in computer science from Wuhan University in 2007. Professor and senior member of China Computer Federation. His research interests include information retrieval and natural language processing, etc (dexi. liu@163. com).

Zhong Qing, born in 1991. Master candidate. Her research interests include information extraction and data mining, etc (zhognqingwjj@gmail. com).

Jiang Tengjiao, born in 1976. PhD candidate. Lecturer. Her research interests include sentiment analysis, XML information retrieval and Web data management, etc (tj_jjang@163. com).