中文开放式多元实体关系抽取

李 颖 郝晓燕 王 勇
(太原理工大学计算机科学与技术学院 晋中 030600)

摘 要 传统信息抽取针对特定的领域。当转换到新领域时,需要人工编写新的抽取规则和人工标记新的训练样本。开放信息抽取突破了传统信息抽取的局限性。现有的开放式信息抽取系统大多针对英文,然而,目前对于中文的研究相对较少,并主要以抽取三元组为主,没有针对中文抽取多元组的方法。因此提出了一种基于依存分析的中文开放式多元实体关系抽取方法。首先,对文本集进行预处理和依存关系分析;然后将动词视为候选关系词,将与此动词有满足条件的有效依存路径的基本名词短语视为实体词,关联两个及两个以上的实体词的关系词可与实体词组成候选多元实体关系组;最后,使用经过训练的逻辑回归分类器对多元实体关系组进行过滤。对百度百科数据集的抽取结果显示,所提方法在抽取大量实体关系多元组时准确性可达到 81%。

关键词 中文开放式信息抽取,依存分析,实体关系抽取,机器学习,OIE,word2vec
中图法分类号 TP391.1 **文献标识码** A

N-ary Chinese Open Entity-relation Extraction

LI Ying HAO Xiao-yan WANG Yong
(College of Computer Science and Technology, Taiyuan University of Technology, Jinzhong 030600, China)

Abstract Traditionally, information extraction (IE) has focused on satisfying precise, narrow, pre-specified requests from small homogeneous corpora. Shifting to a new domain requires the user to name the target relations and to manually create new extraction rules or hand-tag new training examples. Open information extraction (OIE) overcomes the limitations of traditional IE techniques, which trains individual extractors for every single relation type. Present studies have attracted much attention on English OIE. However, few studies have been reported on OIE for Chinese. This paper presented a N-ary Chinese OIE system (N-COIE). N-COIE preprocesses the sentences using the nature language processing tools, and then extracts entity-relation groups from the preprocessed sentences. Finally, N-COIE filters entity-relation groups using the trained logistic regression classifier. Empirical results show the effectiveness of the proposed system.

Keywords Chinese open information extraction, Dependency parsing, Entity-relation extraction, Machine learning, OIE, Word2vec

开放式信息抽取是一种不限领域的抽取方式。传统信息抽取依赖人工编写的规则和人工标注的训练实例,并且用户需要明白所需要抽取信息的句子类型[1]。然而,对于大量的网络信息,使用传统的信息抽取方式要实现对每一种类型关系进行抽取是不现实的。为了针对大量的异构网络信息,信息抽取方法需要转变结构,需要使用一种可一次抽取文本中所有关系的方法。

本文提出了中文开放式多元实体关系抽取方法(N-ary Chinese Open Entity-Relation Extraction, N-COIE)。N-COIE 将中文文本作为输入集,将多元实体关系集作为输出集。首先,将文本进行分句、分词、词性标注、依存句法标注等预处理。然后,根据本文编写的抽取规则从已预处理的句子中抽取出候选实体关系对。最后,通过训练 500 个样本集,得到逻辑回归训练器,通过训练好的逻辑回归分类器对已抽取出的错误多元实体关系集进行过滤。

由于对于包含并列结构或否定结构的句子,容易抽取出不完整或错误的实体关系组,因此本方法在进行抽取之前针对包含并列结构的句子或包含否定结构的句子做特殊处理。

1 相关工作

第一个开放式信息抽取系统是 TextRunner[2-3],它采用朴素贝叶斯模型,特征为词性标注和基本名词短语,训练集由宾州树库产生。随后的研究发现,使用线性条件随机场和马科夫逻辑网模型可以得到一个更好的抽取结果。

WOE 系统[4]使用维基百科数据作为训练集,在效果上得到了很好的提升;并且验证了加入依存关系特征可以在效果上得到显著提升,其代价是抽取速度减慢。

前两种方法都采用先识别实体再识别关系的方法。ReVerb[5]介绍了一种先识别关系再识别实体的方法,即首先找出满足语法约束和语义约束的动词性关系,然后通过关系词与两实体之间的位置约束找到两个实体。这种方法只需要对句子进行词性标注和匹配即可找到实体关系对。R2A2[6]融

本文受基于框架语义标注的中文篇章指代消解策略研究(2012011011-2)资助。
李 颖(1993-),女,硕士,主要研究方向为计算语言学、自然语言处理,E-mail:liying4590@126.com;郝晓燕(1970-),女,博士,副教授,主要研究方向为计算语言学、自然语言处理,E-mail:haoxiaoyan@tyut.edu.cn(通信作者);王 勇(1992-),男,硕士,主要研究方向为计算语言学、自然语言处理。

入了实体学习组件 ArgLearner，以更好地判别实体的边界。

面对大规模网络信息，以上提及的开放式信息抽取系统已经可以进行关系实体提取，但他们都针对于以动词为核心的关系，这样会遗漏以名词、形容词等为核心的关系。再者，以上系统仅对句子的局部进行分析并抽取局部实体关系组，忽略了上下文信息，使得抽取结果表现为局部性，抽取出的关系组可能并非事实。为此，新一代开放式信息抽取系统 OLLIE[7] 弥补了以往开放式信息抽取系统的不足。

针对以名词为核心的关系，文献[8]提出了一种新的方法，即对名词的属性进行了抽取，使得信息量增多，抽取的准确性更高。

对于多元开放式实体关系抽取，文献[9]在 Wanderlust[10] 的基础上提出了一种可对任意实体类型进行多元信息抽取的方法 KRAKEN[11]。

Gamallo[12] 针对英语、西班牙语、葡萄牙语、加利西亚语，使用基于规则的依存分析抽取实体关系组。

然而对于中文，目前研究较少，目前存在 COIE[13]，ZORE[14] 和 UnCORE[15] 3 个开放式信息抽取系统。COIE 使用 CKIP 解析器，首先通过“head-driven”法则识别给定语句中的中心关系词，再依次寻找中心实体词。ZORE 通过依存解析树识别实体关系组，并通过语义模型迭代抽取实体关系组。UnCORE 首先使用实体之间的距离限制和关系词之间的位置限制来获取候选二元实体关系组，然后采用全局排序和类型排序的方法获得关系指示词，最后使用关系指示词和句式规则对二元实体关系组进行过滤。

本文提出的 N-COIE 系统可抽取出多元实体关系，并且准确性达到 81%。

2　中文多元实体关系抽取

N-COIE 的输入为中文文本集，输出为抽取好的实体关系集。本系统包括 3 个主要模块：预处理模块（包括词性标注、依存关系标注）、抽取模块（包括基本名词短语识别和实体关系组抽取）和过滤模块。预处理模块用于对中文文本做基本自然语言处理和标记。抽取模块可在处理后的文本基础上根据规则进行抽取。过滤模块对错误的实体关系组进行过滤，在准确性和召回率之间做折中。

系统框图如图 1 所示。

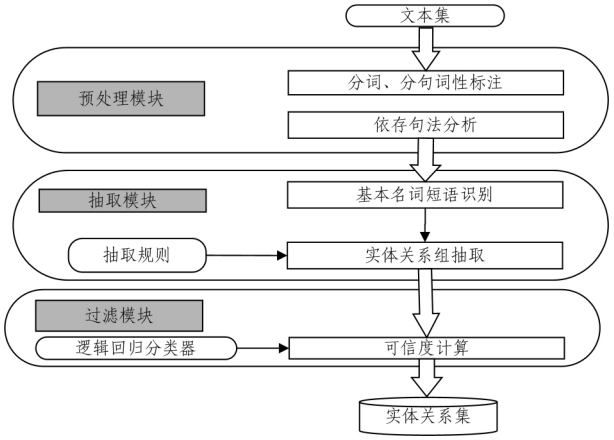

图 1　系统框图

2.1　预处理模块

N-COIE 通过使用 LTP 平台[16] 提供的自然语言处理工具对输入文本集进行处理，包括分词、词性标注和依存句法分析。

依存句法分析可通过分析出句子各个语言单位内成分之间的依存关系揭示其句法结构。直观来讲，依存句法分析识别句子的“主谓宾”、“定状补”等语法成分，并分析各成分之间的语法关系。

依存分析结果如图 2 所示。

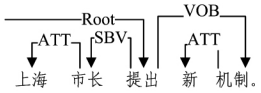

图 2　依存分析结果图

依存句法标注含义如表 1 所列。

表 1　关系类型与标注

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

2.2　基本名词短语识别算法

基本名词短语指的是最简单的、非嵌套的名词短语，不含有其他的子短语。显然基本名词短语不能再包含其他名词短语，也不能被其他名词短语所包含。例如，“苹果 CEO 库克”是一个基本名词短语。

基本名词短语可作为一个实体单元，关于实体的识别以基本名词短语为单位，避免了对于“苹果 CEO 库克”只抽取出实体“苹果”、“CEO”或“库克”等局部实体而非整体实体的情况。

基本名词短语识别算法规定，由本文所定义的修辞词和主名词组合中必须包括一个主名词，可包含 0 个或多个修饰词。修饰词与主名词定义如表 2 所列。

表 2　修饰词与主名词

修饰词	主名词
b：other noun-modifier	j：abbreviation
h：prefix	m：number
m：number	n：general noun
n：general noun	nh：person name
nh：person name	nd：direction name
nz：other proper noun	ni：organization name
ns：geographical name	nl：location name
nt：temporal noun	ns：geographical name
q：quantity	nt：temporal noun
r：pronoun	nz：other proper noun
	q：quantity
	r：pronoun

2.3　抽取算法

N-COIE 将一个句子中的动词视为候选关系词。对于之前已识别出的基本名词短语，若存在此基本名词短语到动词的可达依存关系路径，并且此路径中至少包含 SBV（主谓关系）、VOB（动宾关系）、IOB（间宾关系）、FOB（前置宾语）、DBL（兼语）、ADV（状中关系）6 种依存关系中的一种，则此基本名词短语是该动词的一个实体词。通过分析 80 个例句，验证了上述 6 种依存关系可包括动词的主体成分、客体成分和介词成分。

依次找到此动词的所有实体词，将包含两个及两个以上实体的实体关系组加入到候选实体关系集中。若包含两个实体词，则为二元关系；若包含两个以上实体词，则为多元关系。按照此方法依次找到语句中的其它实体关系对。

1)对并列结构的特殊处理

并列结构会共享一些语法成分，比如："天文学家摆放设备，观测星空"。动词"摆放"和"观测"共享了主语"天文学家"，则可分解为两个句子"天文学家摆放设备。"和"天文学家观测星空。"进行分析。这样可避免对于关系词"观测"无法抽取出实体词"天文学家"的现象。

2)对否定结构的特殊处理

对于含有否定词的句子，比如："高中生没有参加研讨会"，抽取出参加〈高中生，研讨会〉是错误的。应对否定词进行考虑，则抽取出没有参加(高中生，研讨会)。

本系统通过一个封闭的否定词集识别否定词，对于识别出的否定词，将其加入到与之关联的关系词中。否定词集为：不、无、没、莫、非、没有、防止、禁止、难以、忘记、忽视、放弃、拒绝、杜绝、无时无刻、不明不白、差点儿、差点儿没。

下面举例来说明整个抽取算法。

输入：1991 年，莫言在北京师范大学鲁迅文学院获得文艺学硕士学位。

通过表 2 规则识别基本名词短语：1991 年、莫言、北京师范大学鲁迅文学院、文艺学硕士学位。

对于动词"获得"，存在符合条件的依存路径有：ADV(获得，1991 年)；SBV(获得，莫言)；VOB(获得，文艺学硕士学位)；ADV(获得，在北京师范大学鲁迅文学院)。则可得到多元实体关系组，获得〈莫言，文艺学硕士学位，1991 年，在北京师范大学鲁迅文学院〉。

N-COIE 分析实例如图 3 所示。

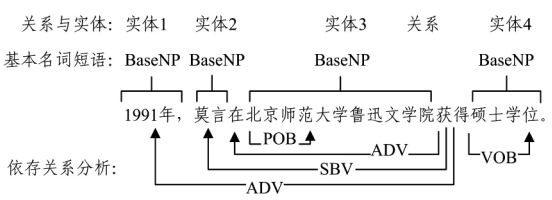

图 3 N-COIE 分析实例

2.4 分类器过滤

通过前面的抽取算法得到的实体关系集中含有一部分错误的关系组，需要过滤掉一些错误的实体关系对，使本系统在准确率和召回率之间得到很好的折中。依照先前的开放式信息抽取系统[6]，可以通过为句子分配信任度来过滤掉一部分实体关系对。逻辑回归分类器可以给每个关系分配一个信任度。

所选特征和经过 500 个句子训练后的权重如表 3 所列。

表 3 所选特征集与权重

特征	权重
关系实体对包含句子的所有信息	0.96
关系词长度小于 3	0.35
关系词长度在 3 到 5 之间	0.11
关系词长度大于 5	−1.04
有两个实体	0.13
有 3 个实体	0.36
有 4 个实体	−0.62
有多于 4 个实体	−0.45

2.5 关系词标准化

通过前面的抽取算法可以得到大量的多元实体关系组，同义关系词进行聚类的过程有助于之后多元组的应用。例如：关系词"起源，源于，发源，来自"均表示相同的含义，可用"起源"来标准化同义关系词。在实际应用中，标准化的多元组可以使问题简化。例如：在问答系统中，对于问题"中医源自哪里？"或"中医来自哪里？"由标准实体关系多元组"起源(中医，中国)"可得出答案为"中国"。

本文使用 word2vec[17-18] 对关系词进行聚类。

3 实验与分析

3.1 实验设置

本系统将中文版百度百科作为测试集，选取 30000 条有效句子，没有结束符号的句子视为无效。

3.2 对抽取器的评估

首先，通过比较所提方法与基准线来说明过滤步骤的有效性。基准线是不进行逻辑回归分类器过滤步骤的抽取结果，与 ReVerb 系统类似。

本文评估抽取出的实体关系组的准确性和召回率。当抽取出实体关系组的关系词和所有实体词都完整与正确时，视为正确关系组。对数据集进行 5 次交叉测试，并将平均结果视为最终准确率和召回率。

图 4 显示了本系统与基准线的比较结果。

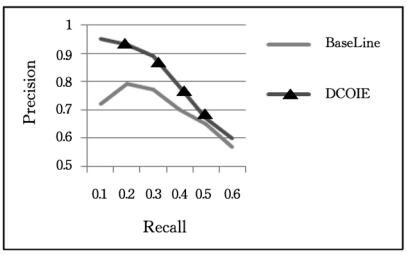

图 4 N-COIE 系统与基准线的比较

从图 4 中可以看出，本系统在各种召回率情况下准确性都高于基准线，在召回率为 0.2 时，准确率为 0.92，高于基准线 13%。在召回率为 0.4 时，准确率为 0.80，高于基准线 10%。这说明通过逻辑回归分类器可以达到很好的效果。

然后，通过 N-COIE 系统对 30000 条句子进行抽取，得到了 38512 组多元实体关系对，达到了 81%的准确性，从而验证了本系统有良好的表现，并且可以实现多元关系的抽取。表 4 为 N-COIE 系统抽取出的实体关系组样例。

表 4 网络文本中抽取的实体关系组样例

句子	关系
莫言是第一个获得诺贝尔文学奖的中国籍作家。	是〈莫言，中国籍作家〉；获得〈莫言，诺贝尔奖〉
2000 年，莫言的《红高粱》入选《亚洲周刊》评选的"20 世纪中文小说 100 强"。	入选〈《红高粱》，"20 世纪中文小说 100 强"，2000 年〉
1991 年，莫言在北京师范大学鲁迅文学院创作研究生班获得文艺学硕士。	获得〈莫言，文艺学硕士学位，在研究生班，1991 年〉

3.3 错误分析

我们分析了错误的关系组和遗漏的正确关系组。准确率分析结果如表 5 所列，召回率分析结果如表 6 所列。

表 5　准确率分析

比率/%	原因
31	关系词识别错误
27	依存分析错误
19	并列结构产生的错误
10	否定结构错误
8	实体标识错误
5	分词及词性标注错误

表 6　召回率分析

比率/%	原因
46	逻辑回归分类器过滤掉
39	依存分析错误
11	抽取规则产生的错误
4	其他

从表 5 可以看到，对于不正确的关系组，其中 31% 是由关系词识别错误产生的。这是由于中文中很多正确的关系词是由间隔的两个或两个以上动词联合组成的，例如："这项活动邀请蔡正华和陈冠华等歌手表演"，其中关系词为"邀请表演"更恰当，而非"邀请"和"表演"两个关系词。其中 27% 源于依存分析错误。其中 19% 是因为对并列结构处理时产生的错误。10% 是源于否定结构产生的错误，否定词集未能包含所有含有否定意思的词。8% 是实体识别错误，例如由于基本名词短语识别错误产生的实体识别错误。5% 是分词和词性标注错误，例如："这本书的出版标志着我们思想界的进一步解放。"，在这句话中"出版"一词是动词还是名词存在争议，LTP 平台将此词标注为动词，而对于本系统，将其标注为名词效果更好。

对于遗漏的正确关系组，46% 是因为逻辑回归分类器过滤掉，逻辑回归分类器一方面提高了准确性，而另一方面也提高了系统的召回率。39% 是源于依存分析错误。11% 源于抽取规则不完善。例如："莫言是第一个获得诺贝尔文学奖的中国籍作家。"，根据抽取规则，"获得"为关系词，而根据语义"第一个获得"更加准确。

结束语　本文提出了一种新的中文开放式信息抽取方法，可实现多元关系的抽取，并提供了中文多元信息抽取系统(N-COIE)。N-COIE 在依存分析的基础上基于规则从大规模文本中抽取出实体关系集，并且针对包含并列结构和否定结构的句子进行了特殊的处理，使准确性得到了新一步的提高。通过训练好的逻辑回归分类器对实体关系组进行过滤，进一步提高了抽取结果的准确性。实验结果表明，N-COIE 可抽取大量的实体关系组，并达到了 81% 的准确性。下一步工作将针对抽取结果中的错误实体关系组对系统进行改进，对关系词为非动词的情况进行研究，并对 N-COIE 的应用展开研究，如在知识图谱和问答系统中的应用，例如："中医源自哪里？'由标准实体关系对起源(中医，中国)可得出答案为"中国"。

参 考 文 献

[1] CHINCHOR N，MARSH E. MUC-7 Information Extraction Task Definition[C]// Proc of MUC-7. 1998.

[2] BANKO M，CAFARELL M J，SODERLAND S. Open information extraction from the Web[C]// Proc of IJCAI. 2007.

[3] BANKO M，ETZIONI O. The tradeoff between open and traditional relation extraction[C]// Proc of Annual Meeting of the Association for Computational Lingustics. 2008：28-36.

[4] WU F，WELD D S. Open information extraction using Wikipedia [C]// Proc of Annual Meeting of the Association for Computational Lingustics. 2010：118-127.

[5] FADER A，SODERLAND S，ETZIONI O. Identifying relations for open information extarctions[C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2011：1535-1545.

[6] ETZIONI O，FADER A，CHRISTENSEN J. Open information extraction：the second generation[C]// Proc of International Joint Conference on Artificial Intelligence. 2011：3-10.

[7] MAUSAM，SCHMITZ M，BART R，et al. Open Language Learning from Information Extraction[C]// Proc of Conference on Empirical Methods in Natural Language Processing and Computer Language Learning(EMNLP). 2012：523-534.

[8] XAVIER C C，DE LIMAV L S. Boosting Open Information Extraction with Noun-Based Relations[C]// LREC. 2014.

[9] AKBIK A，LOSER A. KRAKEN：N-ary Facts in Open Information Extraction[C]// Proc of AKBC-WEKEX at NAACL. 2012：199-202.

[10] AKBIK A，BROSS J. Wanderlust：Extracting semantic relations from natural language text using dependency grammar patterns [C]// Proc of the 1st Workshop on Semantic Search at 18th WWW Conference. 2009.

[11] 杨博，蔡东风，杨华. 开放式信息抽取研究进展[J]. 中文信息学报，2014，28(4)：1-11.

[12] GAMALLO P，GARCIA M，FERNADEZ-LANZA S. Dependency-based open information extraction [C]// Proc of ROBUSUN-SUP. 2012.

[13] TSENG Y H，LEE L H，LIN S Y，et al. Chinese open information extraction for knowledge acquisition[C]// EACL2014. 2014：12-16.

[14] QIU L K，ZHANG Y. ZORE：A syntax-based system for Chinese open information extraction[C]// EMNLP. 2014：1870-1880.

[15] 秦兵，刘安安，刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展，2015，52(5)：1029-1035.

[16] CHE W X，LI Z H，LIU T. LTP：A Chinese Language Technology Platform [C]// ACL. 2010：13-16.

[17] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[C]// CoRR. 2013.

[18] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. 2013：3111-3119.