·人工智能及识别技术· 文章编号:1000-3428(2016)06-0201-07 文献标志码:A 中图分类号:TP311

基于依存分析的开放式中文实体关系抽取方法

李明耀[1,2],杨 静[1,2]

(1.上海市多维度信息处理重点实验室,上海 200241; 2.华东师范大学 计算机科学技术系,上海 200241)

摘 要: 实体关系抽取是信息抽取的组成部分,其目标是确定实体之间是否存在某种语义关系。由于中文语法错综复杂、表达方式灵活、语义多样等固有性质的限制,导致在中文中以动词作为关系表述容易引起实体间的关系含糊不清。为此,利用依存分析,提出一种开放式中文实体关系抽取方法。对输入的单句进行依存分析,通过依存分析输出的依存弧判断单句是否为动词谓语句,如果是动词谓语句则结合中文语法启发式规则抽取关系表述。根据距离确定论元位置,对三元组进行评估,输出符合条件的三元组。在SogouCA和SogouCS语料库上的实验结果表明,提出的方法适用于大规模语料库,具有较好的性能与可移植性。与基于卷积树核的无监督层次聚类方法相比,F值提高了16.68%。

关键词: 开放式信息抽取;中文实体关系抽取;依存分析;无监督;启发式规则

中文引用格式: 李明耀,杨 静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程,2016,42(6):201-207.

英文引用格式: Li Mingyao, Yang Jing. Open Chinese Entity Relation Extraction Method Based on Dependency Parsing[J]. Computer Engineering, 2016, 42(6): 201-207.

Open Chinese Entity Relation Extraction Method Based on Dependency Parsing

LI Mingyao[1,2], YANG Jing[1,2]

(1. Shanghai Key Laboratory of Multidimensional Information Processing, Shanghai 200241, China;
2. Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China)

【Abstract】 Entity relation extraction is a part of the Information Extraction(IE). Its objective refers to determining whether there is a kind of semantic relationship between entities. To break the limitations of complex Chinese grammar, flexible expression and various semantic, which results in the vague relationship between entities simply using verbs as relational expressions in Chinese, this paper presents an open Chinese entity relation extraction method using dependency parsing. This method first does dependency parsing to the input sentence. Whether it is verb predicate sentence can be judged through the dependency arc by dependency parsing. If it is verb predicate sentence, relationship expression can be extracted combined with Chinese grammar heuristic rule. The location of the argument is determind according to the distance, evaluating the triples and outputting these qualified triples. Experimental results on SogouCA and SogouCS corpus show that the proposed method is suitable for large-scale corpus, and has good performance and portability. Contrast with unsupervised clustering method based on kernel tree, F-measure is increased by 16.68%.

【Key words】 Open Information Extraction(OIE); Chinese entity relation extraction; dependency parsing; unsupervised; heuristic rule

DOI: 10.3969/j.issn.1000-3428.2016.06.036

1 概述

近年来,随着互联网技术的发展,万维网逐渐成为一个取之不尽用之不竭的信息来源,如何迅速得到人们感兴趣的信息成为研究者关注的热点。信息抽取(Information Extraction,IE)技术在这种情况下产生,信息抽取的主要目标是从自由文本中抽取指定的命名实体(Entity)、语义关系(Relationship)、事实事件(Event)等信息,把没有结构的自由文本转化成有结构的信息。实体关系抽取(Relation Extraction,RE)是指确定实体之间是否存在某种语义关系,是信息抽取的组成部分,包括

基金项目:上海市科委基金资助项目(14511107000)。

作者简介:李明耀(1989-),男,硕士研究生,主研方向为数据挖掘、信息抽取;杨 静,副教授、博士。

收稿日期:2015-05-25 修回日期:2015-07-19 E-mail:myli@ica.stc.sh.cn

文本挖掘、机器学习和自然语言处理等技术，在自动问答系统、搜索引擎、知识图谱构建等有着广泛的应用。

传统的信息抽取是面向限定领域文本的、限定类别实体、关系和事件等的抽取，面对日益增多不规范的和开放的海量数据，传统的基于人工标注语料库的机器学习方法[1]遇到了严峻的挑战。开放式信息抽取(Open Information Extraction，OIE)在这种背景下产生，目的是从大规模、含有噪音、结构不一、不规则和重复的 Web 页面中，抽取出不限定类别的命名实体、语义关系、事实事件等，并形成有结构的信息输出[2-3]。

本文基于 ReVerb 系统，提出一个自动的中文实体关系抽取方法，即利用依存分析的关系抽取(Relation Extraction with Dependency Parsing，REDP)。该方法利用依存句法分析，简称依存分析(Dependency Parsing，DP)实现开放式中文实体关系抽取。首先对一个句子进行依存分析，再结合中文语法启发式规则和依存分析的结果抽取关系表述，并根据距离确定论元位置，最后进行三元组输出。

2 相关工作

目前，开放式实体关系抽取的研究主要针对英文。文献[2]设计并实现了 TextRunner 系统，将动词作为关系表述。文献[4]利用维基百科的 Infobox 中的信息实现 WOE 系统。文献[5]在句法和词汇上做出限制并开发了 ReVerb 系统，选取以动词表示命名的实体关系，使得超过 30% 的三元组准确率达到了 80% 甚至更高。文献[6]发现 ReVerb 系统中有 65% 的错误是由正确的关系表述与错误的论元(Argument)组成的三元组，在论元识别上改进了以往的简单启发式方法，并实现了 R2RA 系统，改进了论元识别的准确率。目前，开放式中文实体关系抽取的成果也比较少。理论上，上述英文开放式实体关系抽取方法可以直接用于中文。但是由于中文语法错综复杂，表达方式灵活，语义多样等固有性质的限制，将英文开放式实体关系抽取的方法直接用在中文上是非常困难的，总结如下：

(1) 英文不需要分词，词与词之间有空格作为分割，而中文需要借助分词技术。

(2) 英文与中文的句子组成成分不相同，次序也不一样。

(3) 英文中每个句子只有一个谓语动词，而中文没有唯一的谓语动词。

(4) 英文中可以认为论元在关系表述的两侧，但在中文中，语法的复杂性使得这一规律不一定成立。

另外，在中文中以动词作为关系表述，容易导致实体间的关系含糊不清，例如“上海市公安局和上海海关缉私局成立联合专案组，迅速开展案件侦查”。如果按照动词表示关系表述，那么可以得到 Triple1-(上海市公安局，上海海关缉私局，成立)、Triple2-(上海市公安局，上海海关缉私局，联合)、Triple3-(上海市公安局，上海海关缉私局，开展)、Triple4-(上海市公安局，上海海关缉私局，侦查)，但上述这些关系都是含糊不清的，不能很明确地表示两者之间的语义关系。

开放式实体关系抽取可以分为半监督、远程监督和无监督 3 种方法。其中，半监督的抽取方法需要少量的人工种子数据，远程监督的方法需要一个大规模的知识库，无监督的抽取方法不需要任何人工标注数据。

基于 BootStrapping 算法实现的中文实体关系抽取是半监督方法在中文环境下的一个成功例子[7]。BootStrapping 算法通过种子模板抽取特征词，利用最近邻原则自动生成更多的抽取模板，是一个领域无关、自动化的方法，但是在扩展的过程中会加入很多不正确的抽取模板，并且使得错误不断积累，影响最终的效果。针对上述缺陷，标签传播(Label Propagation)算法[8]和协同学习(Co-training)算法[9]先后被提出。但是，这些方法并没有用到句法特征，更多的还是依赖于一系列初始种子，同时也存在初始种子选择困难的问题，文献[10]的实验结果表明初始种子的质量能够直接影响程序输出结果。

无监督的方法主要解决有监督方法中需要大量标注语料的局限性以及半监督方法中种子选择的问题[11]。假设具有一样语义关系的实体组合具有相近的语境，同时语境中的高频词汇可以被认为是实体关系。在这个基础上，首先标记出语料中的命名实体以及上下文，然后对得到共现的实体对以及它们的上下文进行聚类，最后标记每一个类簇，以核心词汇作为关系表述。后续很多学者在此基础上进行改进探索，他们的方法集中于改进聚类方法或者聚类特征来提高准确率。文献[12]对实体上下文的特征词赋予一定的权重构造特征向量，并采用改进的聚类算法。文献[13]改变了传统方法中的字符串匹配，在核函数的计算过程中嵌入了上下文的词汇相似度。文献[14]以句法解析器为基础获得实体对的上下文特征词，采用基于 k 均值的联合聚类算法。文献[15]以卷积树核函数作为相似度计算方法，并采用分层聚类算法。无监督的聚类方法在很大程度上减少了人工的介入，消除了预定义关系类别、不依

赖标注的语料以及人工指定的规则。但这类方法仍存在一些不足,例如特征获取不准确、聚类结果不合理、准确率较低以及聚类数目、聚类中心难以确定等。

文献 [2] 提出了无监督的另一种方法,用命名实体之间的动词作为关系短语,并实现了 TextRunner 信息抽取系统。随后很多方法被提出用于提高 TextRunner 的性能。文献 [5] 采用先识别关系描述词再识别实体,这与 TextRunner 系统先识别实体再识别关系描述词不同,由于词汇信息和句法信息对抽取关系表述进行限制,Reverb 系统获得了优于 TextRunner 的性能。句法信息主要是指关系描述词必须满足如下规则:

V | VP | VW* P

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

词汇信息是指,利用句法信息找出来的关系描述词数目必须大于某一阈值。识别出关系描述词以后,选择关系描述词左右两边最近的两个名词短语构成三元组。理论上,上述这些实体关系抽取的方法可以用于开放式中文实体关系抽取。但是,由于中文语法错综复杂、表达方式灵活、语义多样等固有性质的限制很难直接用在中文实体关系抽取上。文献 [16] 基于中文语法特征,利用句法结构分析针对状态谓语、前修饰和模块化提出了不同的抽取方法,然后根据上述 3 种关系类型以及实体数目确定论元位置,最后按照关系表述类型对其进行裁剪。

文献 [17] 针对有监督和半监督方法的不足,首先提出了远程监督的方法,利用已有的结构化知识库(FreeBase) 实现实体间的关系匹配,解决了有监督学习中需要大量标记语料库的不足以及半监督学习中的低准确率和语义漂移的问题,取得了不错的效果。文献 [18] 通过半人工化方式形成的互动百科数据,构建人物关系知识库并将人物对知识库中的关系实例互相匹配,得到未标记出关系的人物对和标记关系的人物对集合。最后,使用标签传播算法完成未标记人物对的关系匹配。但是,这种方法的前提是存在较大规模的知识库,需要尽可能多地含有关系类别以及对应的关系实例。知识库中关系类别的数量会直接影响到能够抽取出来的关系类别,每种关系类别中的关系实例数量会直接影响到特征的数量,最终影响抽取关系抽取的准确率、召回率和 F 值。此外,在中文上,也很难找到一个大规模的可用的关系知识库。

3　利用依存分析的开放式中文实体关系抽取

本节将会详细描述 REDP 的理论和算法过程。该方法以大规模的自由文本作为关系抽取的目标文本,并借助哈工大的语言云平台对分句后的自由文本进行分词、词性标注、命名实体识别和依存分析的预处理; 然后,进行三元组抽取,这是整个 REDP 的核心部分,主要包括利用中文语法启发式规则抽取关系表述和根据距离确定论元位置。最后,算法将输出表示实体关系的三元组的集合。

在本文中关系表述和论元参照文献 [16] 的定义:

定义 1　关系表述是指能够体现语义关系的短语片段。

定义 2　论元是指参与关系表述的语义角色,即命名实体,本文中主要包括机构名、人名和地名。

3.1　依存分析

依存分析的目的是通过分析句子中各个成分之间的依赖关系,从而揭示句子的句法结构。依存分析认为句子中的支配者是核心动词,而其他任何成分支配核心动词,所有被支配者都以某种形式依赖于支配者[19]。直观地讲,依存分析识别句子中的“主谓宾”、“定状补”这些语法成分与这些成分的位置无关,分析各成分之间的语义修饰关系,获得远距离的搭配信息。

文献 [20] 给出了依存语法的 4 条公理:

(1) 每个句子中只有一个要素是独立的。

(2) 其他要素都依赖于某一要素。

(3) 任何一个要素只能依赖于一个要素。

(4) 若要素 A 直接依赖于要素 B,同时要素 C 在句中位于 A 和 B 之间,那么 C 直接依赖于 A 或者 B 或者 A 和 B 之间的某一要素。

在中文文本处理的实践中,文献 [20] 给出了依存关系的第 5 条公理:

(5) 中心要素左右两边的其他要素相互不发生关系。

句子成分间相互依存和被依存的现象普遍存在于中文能够独立运用的语言单位之中,具有普遍性。依存句法分析的结果是一种结构化数据,利用 LTP 对“上海市公安局和上海海关缉私局成立联合专案组,迅速开展案件侦查。”进行依存分析,将会得到如图 1 所示的结果。如表 1 所示,结合表述文法,总结在 REDP 中常用的依存句法分析标注关系。

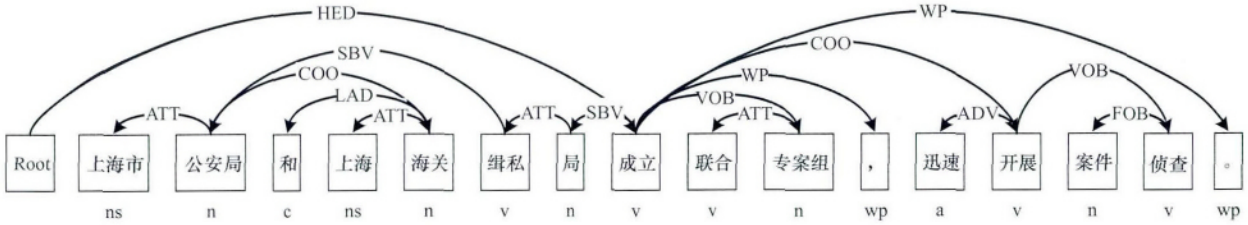

图 1 依存分析例子

表 1 依存分析标注关系

关系类型	依存弧	例子
动宾关系	VOB	我送她一束花(送→花)
间宾关系	IOB	我送她一束花(送→她)
定中关系	ATT	红苹果(红←苹果)
状中结构	ADV	非常美丽(非常←美丽)
动补结构	CMP	做完了作业(做→完)

3.2 结合中文语法的关系表述抽取

在英文中可以将实体之间的谓语动词作为关系表述，而在中文中以谓语动词作为关系表述，容易导致实体间的关系含糊不清。同时，动词谓语句的谓语是动词或者动词短语，它在日常用语中占了很大的比重，是汉语中常见的句型[19]。在 REDP 中，可以通过表 1 中的依存弧 VOB 和依存弧 IOB 确定动词谓语句。

动词谓语句主要分为 3 大类：动词作谓语，动词短语作谓语，复杂的动词短语作谓语。

(1) 动词作谓语

动词作谓语的情况比较简单，仅仅只有动词作为谓语，关系表述可以由下面的文法产生：

关系表述 1→动词

例如 “P1 生于浙江绍兴。”，在该句中“生于”作为谓语动词，“P2 和 P3 是夫妻。”，在该句中“是”作为谓语动词，但是没有充分表达出 P2 和 P3 是夫妻关系。

(2) 动词短语作谓语

动词短语作谓语是以动词为主体，主要分为 3 大类：偏正短语，后补短语和动宾短语。其中，偏正短语可以由修饰词和中心词构成，分为两大类：定语中心语，状语中心语[21]。结合表 1 中的依存弧如表 2 所示，可以通过 VOB 确定动宾关系、CMP 确定后补短语以及 ATT 或者 ADV 确定偏正短语。动词短语做谓语的关系表述可以由以下的文法产生：

关系表述 2→动词 + 宾语

关系表述 3→动词 + 补语

关系表述 4→状语 + 动词

表 2 动词短语作谓语的实例

类别	组成	例子	依存弧
动宾短语	动 + 宾	保护同学	VOB
后补短语	动 + 补	保护好	CMP
偏正短语	状 + 动	好好学习	ATT，ADV

(3) 复杂的动词短语作谓语

复杂的动词短语作谓语是一个动词同时带有状、宾、补语或其中的两个[21]。例如 “状 + 动 + 补 + 宾”，“状 + 动 + 补”，“状 + 动 + 宾”，“动 + 补 + 宾”。复杂的动词短语可以利用表 1 中的依存弧，再按照一定的顺序组合表示，可以由以下的文法产生：

关系表述 5→状语 + 动词 + 补语 + 宾语

关系表述 6→状语 + 动词 + 补语

关系表述 7→状语 + 动词 + 宾语

关系表述 8→动词 + 补语 + 宾语

根据本节中文动词语法规则分析可以看出，动词谓语句的常见形式，动词出现在状语的后面以及出现在宾语和补语的前面，也可能是这几个成分同时出现。也就是说，谓语动词依赖于它前后的成分。通过关系表述 1 ~ 关系表述 8 的比较和分析，总结了关系表述可以由下面的文法产生：

关系表述→状语* 动词 + 补语? 宾语?

其中，* 表示出现 0 次或者任意多次；+ 表示出现 1 次或者任意多次；? 表示出现 0 次或者 1 次。图 2 展示了状语、动词、补语和宾语之间的依存关系。

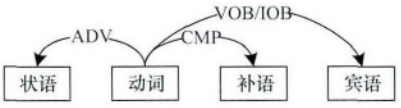

图 2 关系表述文法结构

根据表 1 和关系表述文法，进一步分析图 1 中的例子，首先可以通过依存弧 VOB 确定动宾关系：“成立专案组和开展侦查”。然后按照关系表述文法对动宾关系进一步完善，在“成立专案组”中，依存弧 ATT 表示定中关系，“联合”修饰“专案组”，最后可以得到关系表述“成立联合专案组”。在“开展侦查”中，依存弧 ADV 表示状中结构，“迅速”修饰“开展”，依存弧 FOB 表示前置宾语，“案件”修饰“侦查”，最后可以得到关系表述“迅速开展案件侦查”。

3.3　论元的确定

在 ReVerb 系统[5]中，抽取关系表述左右两边最近的两个名词短语作为论元，这种方法并不能直接适用到中文上。因为在中文表达中，关系表述可由状语、动词、补语和宾语组成情况更为复杂。

随机选取 2 000 个含有关系表述的句子，以人工的方式统计了论元位置分布情况，如表 3 所示。总结得出，中文在论元的抽取上与英文最大的区别在于: 关系表述中可能会含有论元以及论元都在关系表述的左侧。

表 3　论元位置分布

类别	例子	三元组	所占比重/%
关系表述中含有论元	P4 生于山西。 山西临汾仍然被认为是中国污染最严重的城市之一。	Triple(P4，山西，生于山西) Triple(山西临汾，中国，被认为是中国污染最严重的城市之一)	73.24
关系表述中不含有论元	P2 和 P3 是夫妻。 南安市安监局已督促南安市交通局组织人员赶赴现场进行调查。	Triple(P2，P3，是夫妻) Triple(南安市安监局，南安市交通局，组织人员赶赴现场进行调查)	26.27

基于论元和论元、论元和关系表述之间的距离，可以为每两个论元组合计算一个置信度，当式(1) 中 $Confidence(L_i, L_j)$ 达到最大值时，选择 L_i 作为论元 1，L_j 作为论元 2，假定论元 1 总是位于论元 2 的左边，具体的计算公式为:

$$Confidence(L_i, L_j) = \frac{1}{L_i - L_j} + \frac{1}{L_i - R} + \frac{1}{L_j - R + 1}(L_i > L_j) \quad (1)$$

句子经过分词，识别出关系表述和论元后，从右往左依次将词、关系表述或者论元的位置标记为 0，1，…，N。在式(1) 中，L 表示论元的位置，R 表示关系表述的位置。在第 1 个分式中，$L_i - L_j$ 表示论元 1 和论元 2 的距离; 在第 2 个分式中，$L_i - R$ 表示论元 1 和关系表述的距离; 在第 3 个分式中，$L_j - R + 1$ 表示论元 2 和关系表述的距离，分母中加 1 的目的是为了防止除数为 0，因为论元 2 有可能出现在关系表述中，距离越大表示论元和论元之间、论元和关系表述之间存在语义关系的可能性越小，置信度也会越低。

在得到关系表述和论元后就可以进行三元组输出。在本文中对于论元 1 为空、论元 2 为空或者论元的长度为 1 的情况进行简单的过滤处理。

4　实验与结果分析

4.1　实验设置

在本文实验中，选取搜狗实验室的全网新闻数据(SogouCA) 2012 完整版、搜狐新闻数据(SogouCS) 2012 完整版作为语料库，用于验证本文的方法适用于不同的大规模语料库，同时还具有很好的移植性。同时为了比较本文提出方法的 F 值，选取了 ACE RDC 2005 中文标注语料库，在该语料库上将本文方法与基于卷积树核的无监督层次聚类方法进行对比。在本地安装了哈工大的语言云 LTP，模型版本选择了 3.2.0。通过正文提取和分句后，借助 LTP 对每个句子进行分词、词性标注、命名实体识别和依存分析。

在本文实验中，采用准确率 P、召回率 R 和 F 值进行评价。实验结果将分别从准确率 P、召回率 R 和 F 值进行分析，具体的计算公式为:

$$P = \frac{\text{抽取出的正确关系表述数量}}{\text{抽取出的关系表述数量}} \quad (2)$$

$$R = \frac{\text{抽取出的正确关系表述数量}}{\text{文档集合中实际含有的关系表述数量}} \quad (3)$$

$$F = \frac{P \times R \times 2}{P + R} \quad (4)$$

目前，一般对命名实体间的关系都是取其上下文中的一个词，而不是一个短语片段，因此，无法找到公认的关系表述客观判断方法，在本文中采用手工评价的方式对实验结果进行评价[16]。

4.2　关系表述比例的确定

为了分析关系表述在句子中所占的不同比例 r，r 的计算公式为:

$$r = \frac{\text{关系表述长度}}{\text{句子长度}} \quad (5)$$

选择近似最优的比例 r 作为阈值。同时为了方便合理地评测实验结果，随机选取 SogouCA 语料库中的 5 000 个句子进行对比实验。针对不用的 r 值，实验结果如图 3 所示。

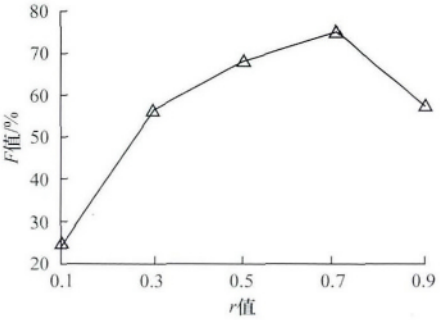

图 3　不同 r 值时对应的 F 值比较

从实验结果可以看出，F 值随着 r 增大先上升后下降，当 r 值为 0.7 时达到了最优值，可见 r 值的选

择对开放式中文实体关系抽取有重要作用。产生这种结果的主要原因是: 当 r 值为较小值时,抽取到的关系表述较短,基本上都是由 1 个词组成,在很大程度上限制了召回率和准确率。当 r 值逐渐增大时,抽取到的关系表述逐步向短语片段靠近,准确率和召回率得到了提升,从而使 F 值得到了提升。当 r 值继续增大时,关系表述越来越长,引入了很多噪声,在这个阶段准确率降低,召回率也随之降低,进而导致 F 值降低。下面举例说明:

(1) "P4 生于山西。" REDP 从中可以抽取出 Triple-{P4,山西,生于山西},此时 $r=0.5$,如果 r 值设置得比较低,将会把这种正确的三元组过滤掉。

(2) "联合国数据显示,叙利亚冲突中死亡的人数已超过 1.2×10^4 人,约 2.3×10^5 人成为难民,百万人需要人道主义援助。" REDP 从中可以抽取出 Triple-{联合国,叙利亚,"显示叙利亚冲突中死亡的人数已超过 1.2×10^4 人,约 2.3×10^5 人成为难民,百万人需要人道主义援助"},此时 $r=0.86$,如果 r 值设置得比较低,将会把这种错误的三元组包含进来。

由上述实验结果和分析可知,在开放式中文实体关系抽取中,r 值的选择对 F 值有很大的影响。由于人工评测结果工作量比较大,不再对 r 值进行更为细致的修正,最后确定 r 值为 0.7 时,本文方法能够达到最好的实验结果。

4.3 对比实验设置

4.3.1 不同语料库的对比实验

为了分析本文提出的方法在不同语料库下的实验结果,在 SogouCA 新闻语料库和 SougoCS 新闻语料库上进行开放式中文实体关系抽取对比实验,并且所有文本的预处理过程、参数设置均相同,抽样句子数都为 5 000。

SogouCA 新闻语料库和 SougoCS 新闻语料库的规模、准确率、召回率和 F 值如表 4 所示。从文本数以及句子数可以看出,SogouCA 新闻语料库中的文本长度比 SougoCS 新闻语料库中的短。总体来说,最后的准确率、召回率和 F 值差别不大,这是由于新闻的质量不同造成的。不同质量的文本,依存分析后的正确性也是不同的。

表 4 不同语料库的实验结果

语料库	规模/GB	文本数/10^6	句子数/10^6	P/%	R/%	F/%
SogouCA	1.8	1.3	13.18	77.58	75.00	76.27
SogouCS	1.8	1.4	11.44	75.00	74.29	74.64

通过不同语料库上的对比实验结果可知,开放式中文实体关系抽取方法能够充分利用依存分析以及中文语法规则,并且适用于不同的大规模语料库,同时还具有很好的移植性。

4.3.2 不同抽取方法的对比实验

在无监督的关系抽取中,也可以采用不同的方法进行关系抽取。为了比较不同关系抽取方法的实验结果,用本文提出的利用依存分析的中文实体关系抽取方法在 ACE RDC 2005 中文标注语料库进行实验,并与基于卷积树核的无监督层次聚类方法[15]进行比较分析,在本实验中 r 值仍为 0.7。基于卷积树核的无监督层次聚类方法是以无监督的分层聚类算法为核心,采用了卷积树核函数用来计算相似度,并以最短路径的子树作为结构化表示形式[15]。

从表 5 的实验结果看,本文方法在 F 值上比基于卷积树核的无监督层次聚类方法提高了 16.68%,在准确率和召回率上都有比较大的提高。这主要是由于基于卷积树核的无监督层次聚类方法采用了最小完全树中两个实体之间最短路径的子树作为关系实例的结构化信息表示,导致了特征获取不准确。同时不同实体和不同关系会存在相同最短路径的子树,这就使得两棵子树相似度较高,在聚类阶段被聚到了同一个簇中,导致聚类结果不合理。

表 5 本文方法与其他无监督方法的对比 %

方法	P	R	F
本文方法	76.89	76.68	76.78
基于卷积树核的无监督层次聚类方法	62.90	57.60	60.10

由表 5 可知,本文方法通过利用依存分析以及中文语法启发式规则能够很好地抽取出多种类型的关系实例,也能保证抽取的准确率、召回率和 F 值,避免了聚类方法特征获取不准确、聚类结果不合理、准确率较低以及聚类数目、聚类中心难以确定的缺点。

5 结束语

本文提出了一种利用依存分析的开放式中文实体关系抽取方法,它利用依存分析并结合中文语法启发式规则抽取关系表述,再根据距离确定论元位置,最后进行三元组输出。实验结果说明,本文方法能够从大规模的自由文本中抽取出实体关系,解决了半监督方法初始种子选择困难的问题; 克服了在中文上远程监督方法很难找到一个大规模可用的知识库; 突破了无监督聚类方法特征获取不准确以及聚类数目难以确定的瓶颈; 避免了中文语法错综复杂、表达方式灵活、语义多样等固有性质的限制,保证了关系抽取的准确率、召回率和 F 值。

下一步将继续利用依存分析的优势,挖掘实体的属性和值。同时,结合文本挖掘出的实体关系实现中文知识图谱的全网自动挖掘和建立。

参考文献

[1] Bach N, Badaskar S. A Review of Relation Extraction [D]. Pittsburgh, USA: Carnegie Mellon School, 2007.

[2] Banko M, Cafarella M J, Soderland S, et al. Open Information Extraction from the Web [C] //Proceedings of the 20th International Joint Conference on Artifical Intelligence. New York, USA: ACM Press, 2007: 2670-2676.

[3] 赵　军,刘　康,周光有,等. 开放式文本信息抽取 [J]. 中文信息学报,2011,25(6):98-110.

[4] Wu Fei, Weld D S. Open Information Extraction Using Wikipedia [C] //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. New York, USA: ACM Press, 2010: 118-127.

[5] Fader A, Soderland S, Etzioni O. Identifying Relations for Open Information Extraction [C] //Proceedings of Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM Press, 2011: 1535-1545.

[6] Etzioni O, Fader A, Christensen J, et al. Open Information Extraction: The Second Generation [C] //Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Berlin, Germany: Springer, 2011: 3-10.

[7] 张素香,李　蕾,秦　颖,等. 基于 Boot Strapping 的中文实体关系自动生成 [J]. 微电子学与计算机,2006,23(12):15-18.

[8] Chen Jinxiu, Dong Hong. Relation Extraction Using Label Propagation Based Semi-supervised Learning [C] //Proceedings of the 21st International Conference on Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2006: 129-136.

[9] Cvitas A. Relation Extraction from Text Document [C] //Proceedings of the 34th International Convention on Manufactured Imports Promotion Organization. Washington D. C., USA: IEEE Press, 2011: 23-27.

[10] Kozareva Z, Hovy E. Not all Seeds are Equal: Measuring the Quality of Text Mining Seeds [C] //Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA: Association for Computational Linguistics, 2010: 618-626.

[11] Hasegawa T, Sekine S, Grishman R. Discovering Relations Among Named Entities from Large Corpora [C] //Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 415.

[12] 张志田. 无监督关系抽取方法研究 [D]. 哈尔滨: 哈尔滨工业大学,2007.

[13] 刘建舟,邵雄凯. 一种改进的中文实体关系抽取方法 [J]. 软件导刊,2011,10(4):27-29.

[14] 王　晶. 无监督的中文实体关系抽取研究 [D]. 上海: 华东师范大学,2012.

[15] 黄　晨,钱龙华,周国栋,等. 基于卷积树核的无指导中文实体关系抽取研究 [J]. 中文信息学报,2010,24(4):11-17.

[16] 郑珊珊. 基于中文语法特征的开放领域实体关系抽取 [D]. 上海: 华东师范大学,2013.

[17] Mintz M, Bills S, Snow R, et al. Distant Supervision for Relation Extraction Without Labeled Data [C] //Proceedings of the 47th Annual Meeting of the ACL. Stroudsburg, USA: Association for Computational Linguistics, 2009: 1003-1011.

[18] 潘　云,布　勒,布丽汗,等. 利用中文在线资源的远程监督人物关系抽取 [J]. 小型微型计算机系统,2015,36(4):701-706.

[19] 黄伯荣,廖序东. 现代汉语 [M]. 3 版. 北京: 高等教育出版社,2002.

[20] 黄昌宁,苑春法,潘诗梅. 语料库,知识获取和句法分析 [J]. 中文信息学报,1992(3):1-6.

[21] 郑贵友. 现代汉语语法讲义 [EB/OL]. (2010-04-18). http://www.blcu.edu.cn/jwch/newjwc/jpkc/chengjuan/jiaoan_2.doc.

编辑　顾逸斐

(上接第 200 页)

[11] Mueen A, Keogh E, Young N. Logical-shapelets: An Expressive Primitive for Time Series Classification [C] //Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2011: 1154-1162.

[12] Hills J, Lines J, Baranauskas E, et al. Classification of Time Series by Shapelet Transformation [J]. Data Mining and Knowledge Discovery, 2014, 28(4): 851-881.

[13] Lines J, Bagnall A. Alternative Quality Measures for Time Series Shapelets [M] //Yin Hujun, Costa J A F, Barreto G. Intelligent Data Engineering and Automated Learning-IDEAL 2012. Berlin, Germany: Springer-Verlag, 2012: 475-483.

[14] Mueen A, Keogh E, Young N. Logical-shapelets: An Expressive Primitive for Time Series Classification [C] //Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2011: 1154-1162.

[15] Rakthanmanon T, Keogh E. Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets [C] //Proceedings of the 13th SIAM International Conference on Data Mining. Philadelphia, USA: SIAM, 2013: 668-676.

[16] Chang K W, Deka B, Hwu W M W, et al. Efficient Pattern-Based Time Series Classification on GPU [C] //Proceedings of 2012 IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2012: 131-140.

编辑　金胡考