复杂中文文本的实体关系抽取研究

王 苑 徐德智 陈建二
(中南大学信息科学与工程学院 长沙 410083)

摘 要 实体关系抽取是信息抽取研究领域中的重要研究课题之一。针对已有方法在处理复杂文本上的不足,提出了复杂中文文本的实体关系抽取方法。结合中文文本的语法特征,提出了7条抽取关系特征序列的启发式规则,并采用语义序列核和KNN机器学习算法结合的方法来分类和标注关系的类型。通过对ACE评测定义下的两个子类的实体关系抽取,关系抽取的平均F值达到了76%,明显高于传统的基于特征向量和最短依存路径核的方法。

关键词 实体关系抽取,语法特征,启发式规则,语义序列核
中图法分类号 TP393,TP391 **文献标识码** A

Entity Relation Extraction for Complex Chinese Text

WANG Yuan XU De-zhi CHEN Jian-er
(College of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract Entity Relation Extraction is one of the important research fields in Information Extraction. Aiming at the problem of inefficiency of existing approaches dealing with entity relation extraction, this paper presented a novel approach. This new approach proposes seven heuristic rules to extract relation feature sequence through combining with grammar feature of Chinese text, and applies the semantic sequence kernel function with KNN learning algorithm to fulfill the entity relation extraction task. Experiments are carried out on two kinds of relation types defined in the ACE guidelines, results show that the new approach achieves an average F-score up to 76%, significantly higher than the traditional feature-based approaches and traditional shortest path for dependency kernel approaches.

Keywords Entity relation extraction, Grammar feature, Heuristic rule, Semantic sequence kernel

实体关系抽取是指自动识别包含在自然语言文本中的两个实体之间的预定义关系。所谓实体是指文本中包含的特定事实信息,如人物、组织机构、地理位置等。实体关系抽取在数据结构化、信息检索和自动应答系统等领域有着重要的研究意义。美国国家标准技术研究院(NIST)在2008年组织的自动内容抽取(ACE, Automatic Content Extraction)评测中定义了7种实体关系类型和18种子类型。

目前,针对中文语料的实体关系抽取研究方法主要有基于特征向量的方法[1,2]、基于改进的语义序列核方法[3]和基于Bootstrapping的方法[4]。这些方法大都只考虑一个句子只存在两个实体情况下的关系抽取。实际上,包含3个或3个以上实体的句子是很多的,正确地抽取这些实体的关系是个研究难点。大量针对英文语料的实体关系抽取研究从实验上证明,句子的句法信息和语义信息对于实体关系的抽取非常有效。本文的主要工作集中于研究包含多个实体的句子的实体关系抽取;结合中文语料的语法特征,对已有的基于最短依存路径核的中文实体关系抽取方法进行改进,使之更适合于包含多个实体的句子的关系抽取。

1 相关工作

在已有的中文实体关系抽取方法中,基于特征向量的方法[1,2]适合于只包含两个实体的句子的关系抽取,这是因为基于特征向量的方法主要考虑的是描述实体关系的特征词的提取。当句子中存在3个或3个以上的实体时,不仅需要提取出描述实体关系的特征词,还需要区别出该特征词的归属,即区别该特征词描述的是哪一对实体之间的关系。因此,基于特征向量的方法在处理包含多个实体的句子的关系抽取时,性能往往会很差。

基于改进的语义序列核方法[3]适合于句子比较短、比较简单的情况,更好的应用应该是和基于特征向量的方法相结合。基于Bootstrapping的方法[4],关键部分是种子的选择和迭代模式的生成,而现有的模式生成只限制于包含两个实体的关系句子,关系抽取也只限于包含两个实体的关系句子。

已有的加入语法信息的关系抽取方法采用的语法解析工具主要有两种:句法解析工具和依存文法解析工具。这两种工具解析的结果相对应地都有个专属的名称:句法树和依存

到稿日期:2008-09-09 返修日期:2008-11-26 本课题受国家自然科学基金重点项目(60433020),湖南省自然科学基金(06JJ50142),湖南省国土资源厅科技计划项目(200718)资助。

王 苑(1984-),女,硕士生,主要研究方向为信息处理等,E-mail:wangyuan_csu@yahoo.com.cn;徐德智(1963-),男,教授,主要研究方向为Web计算、语义网等;陈建二(1954-),男,教授,博士生导师,主要研究方向为计算机网络、计算机理论等。

树。基于句法树[5]和基于依存树的方法[6]由于对句法分析的准确率要求比较地高，往往不太适合中文的实体关系抽取。而基于依存图（把依存树看成有向图）中依存路径的方法[7,8]，由于考虑的只是连接两个实体的依存路径，对句法分析的准确率要求则相对要低很多，该方法的不足之处是不适合于包含多个实体的句子的关系抽取。当句子中存在多个实体时，由于依存图是个连通图，任意的两个实体都存在着最短依存路径，因此很可能会存在最短依存路径重叠或是交叉的情况。若不对任意两个实体的路径做筛选，很可能会导致错误的关系抽取。

针对已有方法不能有效解决多实体句子的关系抽取，本文结合基于最短依存核路径方法在提取特征词上的优势和基于语义序列核在计算对象相似度上的优势，提出了新的复杂文本的实体关系抽取方法。

2 复杂文本的实体关系抽取方法

本文方法主要从两个方面进行介绍：介绍关系特征序列的获取方法和介绍关系特征序列的相似度计算方法。在不引起歧义的情况下，本文中的文本特指句子。为方便描述，引入下列定义。

定义 1（复杂句子） 包含 3 个或 3 个以上实体的句子。

定义 2（简单句子） 不是复杂句子。

定义 3（关系特征序列） 句子中描述两个实体关系的特征词集合。

2.1 关系特征序列的获取

文献[7]论述了连接两个实体的最短依存路径可作为这两个实体的关系特征序列，其论点是基于简单句子的。复杂句子由于包含着多个实体且句子的依存图是连通的，因此很可能存在着最短依存路径重叠、交叉的情况，即会存在连接两个实体的最短路径可能会包括不属于描述这两个实体关系的特征词情况。以句子 S1“孙玮的这种进取性格第一次在摩根士丹利工作期间就赢得了麦晋桁的青睐。”为例，该句子的依存图如图 1 所示。

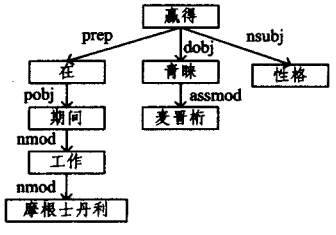

图 1 句子 S1 的依存图片段

该句子存在两个人名实体“孙玮”和“麦晋桁”、一个机构组织实体“摩根士丹利”。该句子中“孙玮”和“摩根士丹利”存在着雇佣关系，而“麦晋桁”和“摩根士丹利”不存在预定义关系。采取文献[7]中的方法提取出的“摩根士丹利”和“麦晋桁”的关系特征序列是“在摩根士丹利工作期间就赢得麦晋桁青睐”。由于包含着特征词“工作”，在利用词频信息抽取实体关系时很可能会导致实体关系抽取错误。分析句子 S1 可发现特征词“工作”应属于“孙玮”和“摩根士丹利”的关系特征序列。由此可见，复杂句子的实体关系抽取，纯粹抽取最短依存路径为关系特征序列是不足的。

仔细观察图 1，可发现“麦晋桁”处于动词“赢得”的直接宾语的分支上。而“摩根士丹利”处于“赢得”介词限定词的分支上，依据汉语语法知识[9]，谓语前的介词结构的语义指向是指向主语，因此图 1 中“摩根士丹利”和“麦晋桁”不存在语义指向关系，应认为“摩根士丹利”和“麦晋桁”不存在预定义关系。

因此，对复杂句子的关系抽取，应对实体之间的最短依存路径进行必要的限制和筛选。本文通过对大量语料的分析，结合汉语语法知识，引入了下列启发式规则。

2.1.1 获取关系特征序列启发式规则

引入的启发式规则包括下列 7 条：

1）文献[9]指出当动词带宾语时，介词短语一般不可以放在动词后，而谓语前的介词结构的语义指向是指向主语。因此对于任意一对实体 1 和实体 2，若它们的依存路径为图 2 所示的形式，则可认为这对实体不存在预定义的关系。在没有歧义的情况下，本文图中的所有虚线箭头都表示连接两个词的依存路径长度不小于 1，大写字母如 A、B、C 和 D 之类都表示任意词。

2）文献[10]指出补语成分的语义是指向谓语的，即作为补语成分的语义和主语是无关的。因此对于任意两个实体 1 和实体 3，若它们的依存路径为图 3 所示的形式，则可认为这对实体不存在预定义关系。

3）若存在某一条最短依存路径 Path1 包含另一条最短依存路径 Path2，为避免路径重叠带来的干扰作用，Path1 端连接的实体，对应关系特征序列应该取 Path1 与 Path2 没有重叠的部分。

存在一种特殊的情况。若依存图结构如图 4 所示的情况时，由于实体 1 与实体 2 是并列关系，因此实体 1 与实体 3 的关系和实体 2 与实体 3 的关系是一样的。

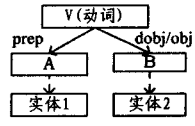

图 2 某句子依存图片段

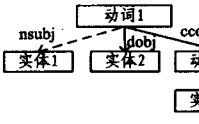

图 3 某句子的依存图片段

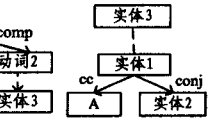

图 4 某句子的依存图片段

图 4 中的 cc 表示 coordination，conj 表示 conjunct。这种结构的依存图表示实体 1 和实体 2 是并列关系。

4）对于任意一对实体 1 和实体 2，若它们的依存路径为图 5(a)、(b)所示的形式，则提取的关系特征序列为“实体 1＋动词 i＋实体 2”。图 5 中的点划虚线箭头代表动词 1 到动词 i 的长度不小于 1，并且动词 1 到动词 i 的最短路径上不存在这样的动词，其中该动词被某词以 nsubj 的方式依存。

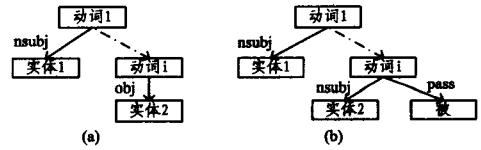

图 5 某句子的依存图片段

依存树中除唯一一个作为独立成分的动词外，其他的词必须依存于其他另一个词。由于任意一个动词都必然会有主语，所以当两个实体分别以 nsubj，obj 方式依赖于动词，且它们的依存路径上又不存在以 nsubj 方式依赖于该路径上的词时，这两个实体应该是主语、宾语的关系，并且它们的谓语应

该取最靠近宾语的那个动词。

5)若存在任意两个实体分别处于某一动词的 nsubj,obj 的分支上,则只考虑这两个实体的关系提取。如图 6 所示,只考虑实体 2 与实体 3 的关系抽取,不考虑实体 1 与实体 2 或是实体 1 与实体 3 的关系抽取。

6)当任意两个实体都处于两个不同 obj 的分支上时,不考虑这两个实体的关系抽取。由于处在 obj 上的两个实体不存在着动作的关系,不存在着限定(modifier)的关系,即它们不存在语义指向关系,因此可认为它们在句子中不存在预定义的关系。

7)在提取任意两个实体的最短依存路径时,若它们的最短依存路径如图 7 所示,则提取的特征序列应包括动词 2。

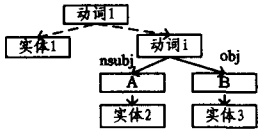

图 6 某句子的依存图片段

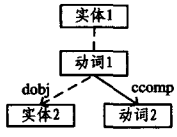

图 7 某句子的依存图片段

图 7 中,虚线表示方向任意且长度不小于 1 的依存路径。由于补语的语义要么指向被补充的动词,要么指向被补充动词的宾语,因此在提取两个实体的特征序列时,不仅包括连接两个实体的最短路径,还应包括图 7 形式中的动词 2。

2.1.2 获取关系特征序列的过程

在使用 Stanford 大学开发的中文依存文法解析器时,发现该分析器解析的文本的词数越少,往往越能取得比较好的解析效果。因此在使用依存文法分析器对句子进行解析时,需要对句子进行预处理。

算法 1 句子的预先处理算法

输入:经过分词后的句子

输出:经过初始预处理后的句子

Step1 当位于两个实体之前或之后存在着标点符号时(如分号、逗号或顿号时),删去出现该标点符号之前或之后的所有词和该标点符号;

Step2 当位于两个实体中间的部分形式为"* * 实体 1 * * , * * * * , * * 动词 * * 实体 2"时,删去两个逗号之间间隔的部分,其中符号 * 表示任意词;

Step3 由于文本中包含大括号的词往往是起着解释的作用,因此删去文本中包含在大括号内的词以及该大括号;

Step4 删去出现在句子中一些特殊符号,如破折号、分号等。

对经过预处理的句子采用算法 2 获取表示实体的关系特征序列。

算法 2 改进后特征序列提取算法

输入:经过预处理的句子

输出:关系特征序列

Step1 句子解析。采用依存文法解析器对句子进行解析,结果以依存树的形式表示;

Step2 特征序列提取。对出现在句子中的任意实体对,按照前面的 7 个启发式规则,提取初始关系特征序列;

Step3 排序。对初始关系特征序列的词,按照原句子中出现的先后顺序排列,每个词都标注相应的词性;

Step4 验证。为防止由于依存文法解析器解析句子错误,对 Step3 提取出的关系特征序列进行验证。Step3 的提取特征序列至少应包括一个出现在间隔在两个实体中的动词。若间隔在两个实体中的词没有动词,则应包括所有间隔在两个实体中的名词。若没有动词也没有名词,则应包括所有间隔在两个实体中的词;

Step5 扩展。为避免数据稀疏对后面相似度的影响,将对 Step4 提取的特征序列进行扩展。扩展的方法是对 Step4 中提取的词增加词性标注(POS),因此最终提取的序列形式为 $X=X_1X_2\cdots X_n$,其中 X_i 为二元组(p,w),p 代表了词 X_i 的词条,w 代表 X_i 的词性。

2.2 关系特征序列的相似度计算方法

本文采用语义序列核来计算关系特征序列的相似度,语义序列核的详细介绍请参见文献[3]。两个序列的相似度计算公式为

$$K(X,Y)=\frac{1}{Z_x(X,Y)}\sum_{n=1}^{k}K_n(X,Y) \tag{1}$$

其中 X,Y 为关系特征序列;$Z_x(,)$为标准化因子,定义为

$$Z_x(X,Y)=\sqrt{\sum_{n=1}^{k_1}K_n(X,X)\times\sum_{n=1}^{k_2}K_n(Y,Y)} \tag{2}$$

其中 k_1 为 X 的长度,k_2 为 Y 的长度,$K_n(X,Y)$为语义序列核函数,定义为

$$K_n(X,Y)=\sum_{u\in\Sigma^n}\sum_{i:u=X[i].p}\sum_{j:u=Y[j].p}\lambda^{l(i)+l(j)}\times\prod_{k=1}^{n}\mathrm{SIM}(X_{i_k}.w,Y_{j_k}.w) \tag{3}$$

其中 $i=[i_1,i_2,\cdots,i_n]$和 $j=[j_1,j_2,\cdots,j_m]$分别表示 X 和 Y 索引的一个子集,$n\leqslant|X|$,$m\leqslant|Y|$,$X[i]$和 $Y[j]$分别是序列 X 和 Y 的一个子序列,$l(i)$和 $l(j)$分别为 $X[i]$和 $Y[j]$在原序列中的跨度,λ 为衰减因子。这里取 $\lambda=0.5$,SIM 函数根据哈工大同义词林提供的语义知识计算两个词汇之间的语义相似度[11]。采用语义序列核来计算关系特征序列相似度的好处在于考虑到了序列的语义知识,减少了词频的影响,提高了匹配的目的性。

3 实验结果及分析

3.1 实验数据

实验选择 ACE 中定义的两个子类 Employment 和 Located 为预定义的关系类别。实验的语料来自 Web 上选择的文档。包含人名实体和机构实体的句子总共有 2500 个,存在 Employment 关系的句子有 800 个,不存在 Employment 关系的句子有 1700 个。而包含人名实体和地方实体的句子总共有 2400 个,其中存在 located 关系的句子有 800 个,不存在 located 关系的句子有 1600 个。每个关系类别分别随机抽取 1/3 为测试集、2/3 为训练语料。

实验对以下 3 种方法做了比较。方法 1 是传统的基于特征向量的方法,通过向量的内积来计算对象之间的相似度;方法 2 是基于传统的最短依存路径核的方法;方法 3 是本文介绍的方法。本文中采用的分类器是上述的语义序列核的方法和 HNN 学习算法联合构造的分类器,采用的分词器是中科院开发的 ICTCLAS 分词器,该分词器的准确率达到了 98%。

3.2 对比实验结果及分析

表 1 为 Employment 和 Located 关系的抽取结果。表中的 P 表示系统抽取的准确率;R 表示系统的召回率;F 测度综合以上两个标准,反映了系统的整体性能。

表 1 Employment 和 Located 关系抽取结果(%)

方法	Employment			Located		
	P	R	F	P	R	F
方法 1	49.6	56.5	52.9	54.5	59.4	56.8
方法 2	63.8	63.7	63.8	62	67.3	64.5
方法 3	74.4	84.1	78.9	73.2	80	76.4

从表 1 中可看出方法 3 有很大的优势，这是在使用大训练集的情况下得到的结果。为了验证新方法是否具有更好的泛化能力，进行第 2 阶段的实验。在本阶段实验中，以 Employment 的实体关系抽取为例，训练集合的规模每次递减，随机抽取 20%，40%，60%，80%，100%的实例来进行训练，实验结果如图 8 所示。

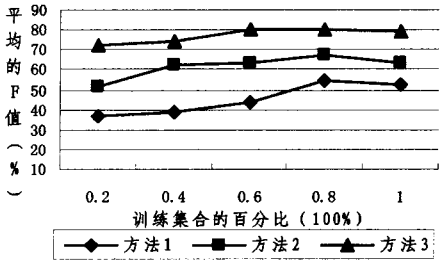

图 8　不同规模训练集合下的关系提取结果

实验证明，本文方法有更好的泛化能力。即使是在只有 20%训练语料的情况下，依然能取得比较好的效果，而其他两种方法在训练集合减少时精确率和召回率都有明显下降。图 8 显示，当训练语料的规模在 80%时，F 值比规模在 100%时高。经分析发现，由于测试集中包含没有预定义关系的测试样本比较多，在随机抽取测试集时，若测试语料规模比较小，这些测试样本的准确率比较高，则导致 F 值较高。

结束语　本文针对现有的中文实体关系抽取方法不能很好处理复杂句子的实体关系抽取，提出了复杂句子的实体关系抽取方法。充分考虑了句子的语法特征，弥补了传统方法不能处理关系特征序列交叉、重叠的缺陷。引入了语义序列核计算关系特征序列相似度，并引入了序列语义，减少了词频的影响，提高了匹配的目的性。经过实验分析，本方法相比已有的方法，其抽取正确率和召回率都有所提高。

未来的工作包括以下两个方面：第一方面，在实体关系抽取方法中加入模式匹配的处理；第二方面，实体关系推理处理，对存在多个实体的对象，根据已知的实体关系推出未知的实体关系。

参 考 文 献

[1] 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

[2] 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

[3] 刘克彬，李芳，刘磊，等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展，2007，44(8)：1406-1411

[4] Li Wei-gang，Liu Ting，Li Sheng. Bootstrapping for extracting relations from large corpora[J]. Journal of Electronics (CHINA)，2008，25(1)：89-96

[5] Zhang Min，Zhong Guo-dong，Aw Aiti. Exploring syntactic structured feature over parse trees for relation extraction using kernel methods[J]. Information Processing and Management，2008，44：687-701

[6] Culotta A，Sorensen J. Dependency tree kernels for relation extraction[C]// Proceedings of the 42nd Annual Meetings of the Association for Computational Linguistics (ACL-04). Barcelona，Spain July，2004：423-429

[7] BUnescu R C，Mooney R J. A Shortest Path Dependency Kernel for Relation Extraction[C]// Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005：724-731

[8] Huang Rui-hong，Sun Le，Feng Yuan-yong. Study of kernel-based Methods for Chinese Relation Extraction[C]// the LNCS，Springer，AIRS's 08. 2008：698-604

[9] 魏庭新，吕文华. 现代汉语介词结构位置的考察及影响其位置的句法、语义因素的分析[D]. 北京：北京语言大学，2004

[10] 李锦姬，范晓. 现代汉语补语研究[D]. 上海：复旦大学，2003

[11] Che Wang-xiang，Jiang Jian-min，Su Zhong，et al. Improved-edit-distance Kernel for Chinese Relation Extraction[C]// Proc. of the Second International Joint Conference on Natural Language Processing (IJCNLP-05). 2005：132-137

(上接第 192 页)

从表 2 中可以看出同一个服务的不同操作具有不同的 QoS，本文所实现的 QosCollectionFrame 有效采集了 Web 服务的 QoS 数据并对其进行了有效处理。

结束语　本文以 Web 服务 QoS 为研究对象，主要研究了 QoS 数据采集及 QoS 数据处理计算的方法和技术。具体包括：分析了 Web 服务的调用过程与 QoS 属性的关系，研究了几种 QoS 数据的采集方法；设计实现了以 QoS 为中心的多源 QoS 数据采集系统 QosCollectionFrame，对 QoS 数据进行存储、处理以及计算。最后本文将系统集成在北大软件资源库中，通过应用实例验证了 QosCollectionFrame 的可行性。

参 考 文 献

[1] Menascé D A. QoS Issues in Web Services [J]. IEEE Internet Computing，2002，6(6)：72-75

[2] Chen Hongan，Yu Tao，Lin Kwei-jay. QCWS：An Implementation of QoS-Capable Multimedia Web Services [A]// IEEE. Proceedings of the IEEE Fifth International Symposium on Multimedia Software Engineering [C]. USA：IEEE Computer Society Press，2003：38-45

[3] 杨胜文，史美林. 一种支持 QoS 约束的 Web 服务发现模型 [J]. 计算机学报，2005，28(4)：589-594

[4] Yu Tao，Lin Kwei-jay. A Broker-Based Framework for QoS-Aware Web Service Composition [A]// IEEE. Proceedings of the IEEE International Conference on e-Technology，e-Commerce and e-Service，2005 [C]. USA：IEEE Computer Society Press，2005：22-29

[5] Zeng Liangzhao，Benatallah B，Dumas M. Quality Driven Web Services Composition [A]// ACM. Proceedings of the 12th International Conference on World Wide Web (WWW)，Budapest，Hungary，2003 [C]. USA：ACM Press，2003：411-421

[6] 赵俊峰，谢冰，张路，等. 一种支持领域特性的 Web 服务组装方法 [J]. 计算机学报，2005，28(4)：731-738

[7] Yu Tao，Lin Kwei-Jay. Service Selection Algorithms for Web Services with End-to-end QoS Constraints [A]// IEEE. Proceedings of the IEEE International Conference on e-Commerce Technology [C]. USA：IEEE Computer Society Press，2004：129-136

[8] 邵凌霜，李田，赵俊峰，等. Web 服务 QoS 管理框架[J]. 计算机学报，2008

[9] Java API for XML-Based RPC (JAX-RPC)[OL]. http://java.sun.com/webservices/jaxrpc

[10] Java Web Services Technologies At a Glance[OL]. http://java.sun.com/webservices/technologies/index.jsp

[11] 赵俊峰. 构件库反馈管理及运行时应用支持技术的研究[D]. 北京：北京大学，2005