文章编号：

基于句法语义特征的中文实体关系抽取*

郭喜跃[1,3]，何婷婷[2]，胡小华[2]，陈前军[1,4]

（1.华中师范大学国家数字化学习工程技术研究中心，湖北省武汉市 430079；

2.华中师范大学计算机学院，湖北省武汉市 430079；

3.兴义民族师范学院信息技术学院，贵州省兴义市 562400；

4. 湖北大学信息与网络中心 湖北省武汉市 430062）

摘要：实体关系抽取的核心问题是实体关系特征的选择。以往的研究通常都以词法特征、实体原始特征等来刻画实体关系，其抽取效果已难再提高。在传统方法的基础上，本文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，选择 SVM 作为机器学习的实现途径，以真实新闻文本作为语料进行实验。实验结果表明该方法的 F1 值有明显提升。

关键词：句法特征；语义特征；实体关系抽取；SVM

中图分类号：TP391　　**文献标识码：**A

Chinese Named Entity Relation Extraction Based-on the Syntactic and Semantic Features

Xiyue GUO[1,3], Tingting HE[2] ,Xiaohua HU[2], Qianjun CHEN[1,4]

(1. National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, Hubei, 430079, China ;

2. School of Computer, Central China Normal University, Wuhan, Hubei, 430079, China ;

3.School of Information Technology, Xingyi Normal University for Nationalities, Xingyi, Guizhou, 562400, China;

4.Network Center of Hubei University, Wuhan, Hubei, 430062, China)

Abstract: Identifying the relation features between named entities is the key aspect in named entity relation extraction. Traditional methods usually chose the lexical features and other simple features as the features of named entity relations, and their capacity can hardly be improved now. On the basisof traditional methods, this paper proposed a new kind of Chinese named entity relation extraction method, adding syntactic and semantic features to the relation feature sets, such as dependency parsing, core predicate verband semantic role labeling etc. Chose the SVM as the machine learning tool, and took the true news text as the experiment corpus. The result of experiment indicated that this method could improve the F1 value obviously.

Key words: syntactic features; semantic features; named entity relation extraction; SVM

1 引言

实体关系抽取是指从自然语言描述的语料中获取命名实体之间存在的关系，比如人名与组织机构之间可能存在雇佣关系等。实体关系抽取是基于命名实体识别的一种更深层次的研究，能够为事件抽取、自动问答、机器翻译以及自然语言处理相关领域的研究提供前提保障

* **收稿日期：**　　**定稿日期：**

基金项目：本文受国家社会科学基金重大项目（No.12&2D223）;国家“十二五”科技支撑计划课题(2012BAK24B01);国家自然科学基金(61300144);国家语委“十二五”重点项目（ZDI125-1）；教育部/国家外国专家局高等学校学科创新引智计划项目（B07042）；湖北省自然科学基金重点项目（No.2011CDA034）；华中师范大学中央高校基本科研业务费项目(No. CCNU13A05014, No. CCNU13C01001, CCNU13F010) 资助。

[1]。在大数据时代下，人们需要处理的数据量越来越庞大，处于承上启下地位的实体关系抽取，其地位自然越发重要。CNKI 统计显示，自 2007 年以来，实体关系抽取研究的关注度一直呈上升趋势①，这说明实体关系抽取得到越来越多的重视。

关系抽取研究最初由 MUC 会议提出，许多学者以此为平台提出了大量的关系抽取方法，使得关系抽取在理论与实践上初步成型。继 MUC 之后，连续举办 9 届的 ACE 会议也将关系抽取作为评测内容之一，吸引了许多优秀的研究成果纷纷通过此会议发表或展示，有力地促进了关系抽取研究的完善与发展[2]。实体关系抽取的研究思路主要有基于语言规则模板的方法、基于词典驱动的方法、基于 Ontology 的方法和基于机器学习的方法等，近几年的研究趋势表明，以机器学习为主、融合多种方法的思路成为主流。

2011 年德国洪堡大学的 Philippe Thomas 等人，为研究生物医学文献进展，提出一种利用集成学习（Ensemble Learning）抽取药物之间的相互作用的方法，该方法基于不同语言特征空间，构建多种机器学习方法对比机制，选出效果最好的方法[3]；斯坦福大学的 Mihai Surdeanu 等人在 2012 年将多实例多标记学习引入到关系抽取中，形成一种新的方法，它利用带有潜在变量的图模型，并将文本中实体对和其标记融合在一起，这一方法从一定程度上克服了远距离监督学习的缺陷，而且实验表明它在两类不同领域的文本中性能表现不俗[4]；2013 年 Haiguang Li 、GongqingWu 等在 Applied Intelligence 发表了一种基于位置语义特征的命名实体关系抽取方法，该利用位置特征的可计算性和可操作性，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，最终实现了关系抽取效果的明显提升[5]。

中文实体关系抽取的研究也取得了丰硕成果。何婷婷等人于 2006 年提出一种基于种子自扩展的命名实体关系抽取方法，首先手工选取少量具有抽取关系的命名实体对，把它们作为初始关系的种子集合，通过自学习，关系种子集合不断扩展，通过计算命名实体对和关系种子之间的上下文相似度来得到所要抽取的命名实体对，这种方法有效地提高了抽取的准确率[6]；陈火旺等人在 2007 年提出一种基于 SVM 的中文实体关系抽取，它将关系抽取看作分类问题，利用 HowNet 提供的概念信息，以词特征集、词性特征集、实体属性特征集、实体出现特征集和 HowNet 概念特征集等形成训练语料，并利用 SVM 模型进行学习，从而实现实体关系抽取，实验结果表明，F1 值最高可达 76.58[7]；2013 年，陈鹏、郭剑毅等人在研究支持核函数的机器学习算法的基础上，提出了基于凸组合核函数的中文领域实体关系抽取方法，该方法的核心思想是以径向基核函数、Sigmoid 核函数及多项式核函数组成的不同组合比例的凸组合核函数将特征矩阵映射成为不同的高维矩阵，利用支持向量机训练这些高维矩阵构建不同分类模型，针对大量的旅游语料进行实验，平均 F 值为 62.9[8]。

相对而言，由于中文语言结构的独特性和语义的复杂性，中文实体关系抽取研究整体上与国外的研究还存在一定差距，常用的基于浅层语法分析获取特征的方法已经达到瓶颈。本文也将采用 SVM 模型训练语料，但于以往不同的是，该方法扩展了实体关系特征的选择范围，除了传统的词法特征、实体原始特征外，又选择了句法特征、语义特征等作为实体关系特征，主要包括语义角色标注、依存句法关系、核心谓词特征等，并依据中文的语法特点对这些特征进行有机整合，得到二元实体对之间的丰富关系特征，最后交由 SVM 进行训练和测试。

2 相关研究

2．1 句法分析

句法结构描述了句子中的短语结构、依存结构以及功能，通常用句法结构树来进行表示；句法分析研究如何通过计算机算法得到自然语言句子的句法结构。句法分析在信息检索、信息抽取、自动问答、机器翻译等领域都起到重要作用。依存结构是句法分析的一个重要方面，

① http://trend.cnki.net/TrendSearch/trendshow.htm?searchword=%u5173%u7CFB%u62BD%u53D6，访问时间: 2014-5-21

它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，主张句子中核心谓词是支配其它成分的中心成分，而它本身却不受其它任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[9]。由于句子中的命名实体必定会作为一个短语结构出现在依存结构中，那么这种依存关系也必然会反映出相应实体之间的关系特征，例如：“中国国家主席习近平在人民大会堂举行仪式，欢迎来华访问的美国总统奥巴马。”，其命名实体识别和句法分析结果如图 1：

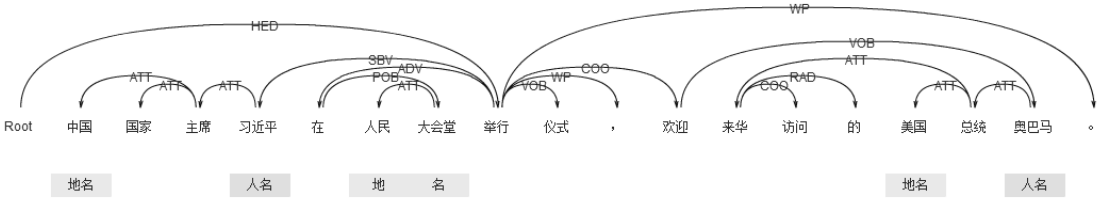


图 1 句法分析示例

从中可以看出，本句中共含有 3 个地名实体、两个人名实体，其中（“中国”，“习近平”）、(“美国”，“奥巴马”）这两组实体对之间存在明显的定中关系，而“人民大会堂”作为一个完整的地名实体，与其前面的“在”形成介宾关系。

另外，对句法分析的结果进行大量实验后发现，在所有谓词中，核心谓词对获取实体边界、承接实体关系起着关键作用，句子中命名实体与核心谓词的平均距离和命名实体与普通谓词的平均距离有明显差异，在海量语料中后者大约是前者的 1.5 倍，所以实体与核心谓词的距离也是实体之间的一种隐含的关系特征。仍以图 1 中的句子为例，句法分析结果指出句子的核心谓词是“举行”，非核心谓语有“欢迎”“来华”“访问”，经计算，各实体与核心谓词的平均距离是 5，实体与非核心谓词的平均距离是 7。

基于上述发现，本文考虑将句法分析结果的依存句法关系和核心谓词作为实体关系中的句法特征进行考量。

2．2 语义角色标注

语义角色标注以句中的谓词为支点，分析并标注各短语结构在句子中的语义成分，如施事、受事、附加语等，它是语义分析中的一种形式，在信息抽取、自动问答、自动摘要等方面已经发挥了重要的支撑作用[10]。中文语义角色标注经过几年的研究，已经逐渐成熟，目前的标注方法其 F 值最高可达 91%[11]。与在句法分析中的成分类似，命名实体在语义角色标注结果中也隐含着一些特征。例如“俄罗斯外长拉夫罗夫与美国国务卿克里在法国巴黎举行会谈。”其语义角色标注结果如图 2：

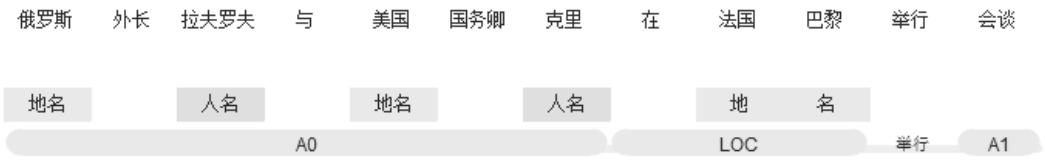


图 2 语义角色标注示例

图 2 中间的矩形框内为命名实体识别结果，下方的圆角矩形框内为语义角色标注结果。本句共含两大类 5 个命名实体；围绕核心谓词“举行”，其语义角色标注结果有 3 部分组成：A0 为施事部分，A1 为受事部分，LOC 为表示地点的附加部分。从此图可以直观看到，(“俄罗斯”，“拉夫罗夫”）和（“美国”，“克里”）这两个实体对在语义上是主语，实体“法国巴黎”是状语，作为受事部分的 A1 中无命名实体。总结来看，命名实体的类型和实体对的组合往往隐性地指示了它在语义角色标注结果中的位置和作用，而这些位置与作用也从一定程度上反映了实体之间的内存语义关系。

2．3 SVM 概述

SVM 是一种可用于分类和回归问题的、较为复杂的机器学习算法模型，曾被评为机器学习领域 10 大经典算法之一[12]。虽然其实现过程十分复杂，但可以用简单的线性分类过程来描述其基本原理。

设给定的训练数据集为：$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 x_i 为特征向量，y_i 为所属类别的标签，（x_i，y_i）为一个样本点。SVM 的目标是在特征空间中找到一个分离超平面，能够将各样本特征划分到不同的类别中。对于简单的线性可分问题，其分离超平面函数可假设为 wx+b=0（w 为法向量，b 为截距），可用（w，b）来表示，这就是支持向量。可以想象，符合这样条件的超平面可能会有多个，也即存在多组（w，b）数据，需要继续从中找出能够使间隔最大的一组（w，b）作为最终结果，这时分离超平面就可以唯一确定下来。对于特征十分复杂、线性不可分的问题，则需要在此基础上引入核函数的概念来确定分离超曲面，该核函数应该能够将高维的特征数据映射到低维空间，从而降低计算的复杂度，常用的核函数有多项式核函数、高斯核函数、神经网络核函数、RBF 核函数等[13]。

SVM 训练过程本质上为凸优化问题，可以利用已知的有效算法发现目标函数的全局最小值，这与基于贪心算法来获取局部最小解的方法有本质不同，所以 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[12]。但是最基本的 SVM 算法只能用于二类分类问题，而且分类过程较慢，所以自从 SVM 算法被提出以后，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[14-16]，这些方法使 SVM 学习的过程更迅速，效果也有明显提升。

3 本文的方法

3．1 基本流程

本方法利用哈工大 LTP-Cloud 平台对语料进行初步处理。LTP-Cloud 是由哈工大社会计算与信息检索研究中心研发的云端自然语言处理服务平台。后端依托于历时十年形成的语言技术平台，语言云为用户提供了包括分词、词性标注、依存句法分析、命名实体识别、语义角色标注在内的丰富高效的自然语言处理服务[17]。

本方法以 LTP-Cloud 对语料的词法、句法分析结果为基础，生成二元实体对，并采集所有实体对的既定特征数据从而生成训练文本，并交由 SVM 进行训练。具体的过程如图 3 所示。

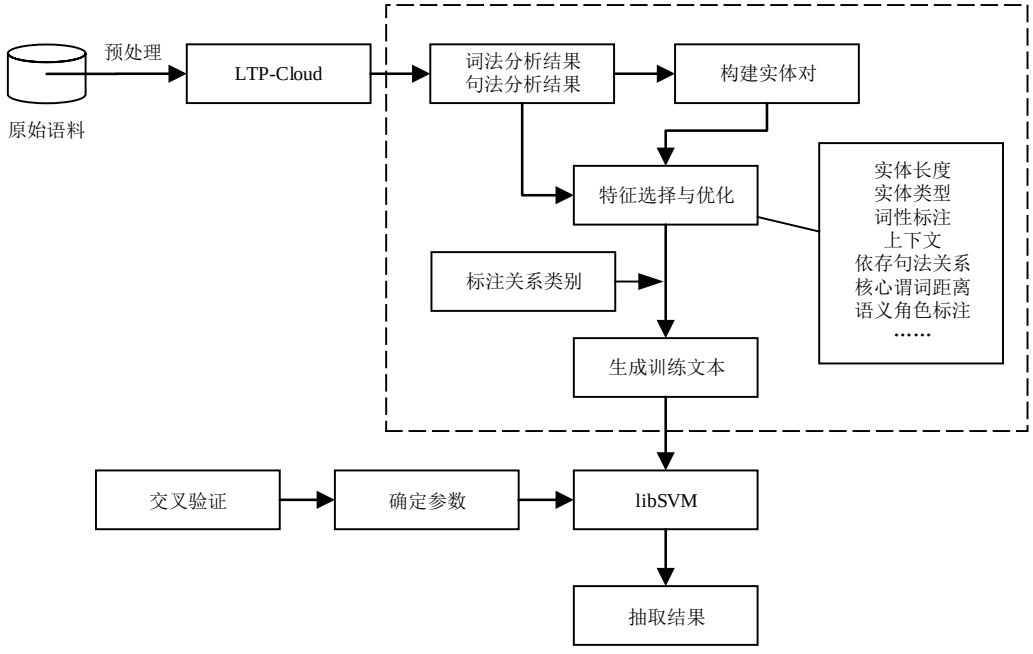


图 3 基于句法语义特征与 SVM 的实体关系抽取方法过程

图中虚线框内的模块为本方法的核心内容，也体现了本文的创新点。由于 LTP-Cloud 以单个句子为分析对象，处理结果只是带有多种标注信息的特定格式的数据，所以需要在此基础上获取句子中包含的所有命名实体，为了能够让机器较为全面地学习到实体间各种可能的关系，这里将句子中所有实体两两组合，生成实体对；如果句子中只有一个实体，或无实体，则说明此句中不存在实体关系，需要忽略此句；接着，根据已经组成的实体对，选择各种实体关系的特征。

3．2 基本特征

常规的实体关系特征主要从词法分析结果来获取，以往的研究已经表明了这些特征的有效性[18]。面向句子中所有实体组成的二元实体对，本文选择的基本实体关系特征有：

1、实体长度。根据命名实体结果的标识信息中，获取多词实体的边界，并根据其首尾词的位置来计算实体长度。

2、实体种类。目前 LTP-Cloud 能够识别的实体种类有人名、地名、组织机构名。

3、实体内容。这里采用词袋机制将实体内容由字符转换为数字。

4、实体中各词的词性标注。

5、实体的上下文环境。包括实体前后两个词的内容以及词性标注信息。

3．3 句法语义特征

本方法对 LTP-Cloud 的处理结果进行再加工，可以得到更多的句法语义特征。

1、句法依存关系。将获取实体对中每一个实体在原句中所属的句法依存关系值。

2、实体与核心谓词的距离。根据实体首词在句中的位置和核心谓词的位置，计算出每一个实体与核心谓词的距离。

3、语义角色标注。LTP-Cloud 的初步结果中包含了针对所有谓词的语义角色标注结果，但是只有基于核心谓词的语义角色标注的覆盖度是最广的，所以这里也仅选择基于核心谓词的语义角色标注结果作为这一特征来源，获取实体对中每一个实体所属的语义角色成分，将其作为实体关系的一种特征。

每组实体对的实际特征个数会随着实体长度的不同而不同;这些特征之间的相对位置并不是任意的，需要根据一定的规律合理安排，如：将实体的全局特征放到前面，实体局部特征放到后面，这样做便于机器的学习，提高精确度。

4 实验与分析

4．1 实验结果评价指标

预设了 4 种实体关系种类：人名实体与组织机构实体之间的雇佣关系（employ）、组织机构实体与地名实体之间的位于关系（locate）、属于同一种实体类型的同类关系（sameType）和无关系（no-Relation）。由于本文亦将实体关系抽取过程看作是分类的过程，所以这里的评价方式也采用常规的准确率、召回率和 F1 值。针对某一具体关系类型的抽取结果，其评价公式为：

$$\text{准确率：Precision} = \frac{\text{结果中正确标注为当前关系类型的个数}}{\text{结果中标注为当前关系类型的总个数}} \times 100\%$$

$$\text{召回率：Recall} = \frac{\text{结果中正确标注为当前关系类型的个数}}{\text{测试语料中当前关系类型总个数}} \times 100\%$$

$$\text{F1 值：Recall} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

因为分类标注问题不同于信息检索问题，所以应计算所有实体关系种类的准确率和召回

率的平均值，以此作为整体抽取结果的准确率和召回率，并由此得出整体 F1 值。

4．2 实验设计

本方法用 1998 年 1 月份的《人民日报》所有版面内容作为语料，全文共含有 4 万多个中文句子，。由于 LTP-Cloud 需要以句子为基本处理对象，所以还需采用基于规则的方法将语料内容进行分句。将上述语料通过 LTP-Cloud 处理后，可得到含有约 8.5 万个唯一实体的处理结果，由此可得到约 3.6 亿个二元实体对，将其中的 80%作为训练语料，20%作为测试语料，进一步分析出实体对中句法语义特征数据，并人工添加实体关系分类标注，最终形成训练语料。采用 libSVM 作为辅助工具，在 SVM 的训练过程中，选择 RBF 作为核函数，采用交叉验证法，得到最优参数 c=2.0，g=0.5，CV rate=73.1905。实验程序采用 Python 语言编写实现。获取实体关系特征步骤的核心算法如下：

```
for eachSentence in sentenceList
    call LTP-Cloud to process eachSentence using GET method
    get the length of each namd-entity, and add this property to the first word of each entity
    create an empty tempNeList , waiting to save all named-entities in eachSentence by order
    create an empty tempV , waiting to save the key verb’s properties of eachSentence
    for each node in eachSentence
        fill the Bags for the properties of words, ne, pos, SRL, relative etc.
        save the correct node into tempNeList
        save the node of key verb into tempV
    for eachNode in tempNeList
        find the boundary of named-entities and save a whole named-entities into neList as
a node
    if the length of neList is less than 2
        ignore eachSentence
    else
        for firstIndex in range(0,len(neList)-1)
            for secondIndex in range(0+1,len(neList))
                create named-entitied pair
                form the features of neList[first]within current named-entitied pair
                form the features of nelist[second]within current named-entitied pair
                combine the 2 feature sets to form the features of eachSentence
save the features of eachSentence into a certain file
```

4．3 实验结果

为了分析本方法所选特征集与以往方法所选特征集在关系抽取中的效果差异，这里根据不同的特征集进行了对比实验。第一组实验选取 2.2 节所述的基本特征，第二组实验在第一组所选特征的基础上加入了 2.3 节所述的句法语义特征。两组实验的具体结果统计如表 2 所示。

表 1 对比实验结果统计

特征类型	统计对象	准确率	召回率	F1 值
基本特征	雇佣关系	61.50	49.79	55.03
	位于关系	68.60	64.09	66.27
	同类关系	78.74	72.89	75.70

	无关系	71.09	75.80	73.37
	整体	73.82	72.04	72.92
基本特征+ 句法语义特征	雇佣关系	63.16	53.98	58.21
	位于关系	74.68	77.87	76.24
	同类关系	76.55	84.88	80.50
	无关系	77.25	79.45	78.33
	整体	76.03	79.85	77.89

从上述结果可以看出，第二组实验在加入更多的句法语义特征后，实体关系抽取的效果明显优于第一组实验。其中位于关系、同类关系和无关系这三种关系抽取效果的提高更为明显，这说明本方法新增特征是有效的。

另外，文献 7 所提出的中文实体关系抽取方法是中文实体关系抽取领域较为经典的方法之一，将本文第二组实验结果与文献 7 的研究结果进行了比较，可以看出，除准确率下降了 0.19%外，本文所述方法的召回率和 F1 值分别提升了 2.93%和 1.31%，具体见图 4。

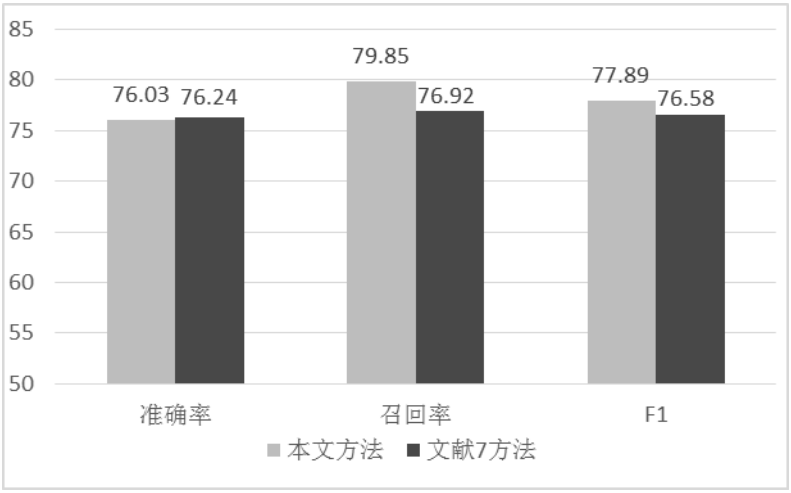


图 4 实验结果对比图

4．4 结果分析

分析上述结果不难发现，本文所述方法与以往方法相比，由于选择了句法语义特征，其实验整体表现有一定的优势，这说明上述所选特征是有效的。但是其中也存在一个明显的问题，从局部来看，部分实体关系抽取的效果相对较差，比如人名实体与组织机构实体之间的雇佣关系。在实体对中，并不是只要存在一个人名实体与一个组织机构实体，就应认定他们之间存在雇佣关系，比如“新华社记者张宿堂报道：国务院总理李鹏……”句中的实体对（新华社，张宿堂）、（国务院，李鹏）之间应该都属于雇佣关系，但对于实体对（新华社，李鹏）、（国务院，张宿堂），很明显不应属于雇佣关系，但是后两个实体对与前两个实体对的特征基本类似，只是在句中的位置、具体的词不同，所以这就容易导致分类错误。

5 总结

本文提出了一种基于句法语义特征的实体关系抽取方法，与以往的实体关系抽取方法相比，本文新增了句法分析结果和语义分析结果作为为实体关系的特征，实验结果表明此方法效果明显。本方法仅考虑了人名、地名、组织机构名这三类实体，实体间的关系也仅选择了雇佣关系、位于关系、同类关系和无关系四种，有待进一步扩展；另外，本方法以句子为处理单位，缺少篇章处理的视野，未考虑实体的指代消解问题，未来将在上述方面继续做深入研究。

参考文献

[1] Kushmerick, N., Weld, D., and Doorenbos, R. Wrapper induction for information extraction. IJCAI-97, 1997.

[2] D Zelenko, C Aone, A Richardella. Kernel methods for relation extraction[J]. The Journal of Machine Learning Research, 2003(3):1083-1106.

[3] Philippe Thomas, Mariana Neves, Illés Solt. Relation Extraction for Drug-Drug Interactions using Ensemble Learning [C]. DDI Extraction 2011.

[4] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, Christopher D. Manning. Multi-instance Multi-label Learning for Relation Extraction [C]. ACL, 2012.

[5] Haiguang Li, GongqingWu etc. A relation extraction method of Chinese named entities based on location and semantic features[J]. Applied Intelligence, 2013, Volume 38, Issue 1, pp 1-15

[6] 何婷婷,徐超等. 基于种子自扩展的命名实体关系抽取方法[J]. 计算机工程, 2006(21):183-184, 193

[7] 徐芬, 王挺, 陈火旺. 基于 SVM 方法的中文实体关系抽取[C]. 第九届全国计算语言学学术会议, 2007.

[8] 陈鹏,余正涛等. 基于凸组合核函数的中文领域实体关系抽取[J]. 中文信息学院, 2013(5):144-148.

[9] 胡宝顺, 于戈等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报, 2008(4):662-676.

[10] 李业刚, 孙福振, 李鉴柏等. 语义角色标注研究综述[J]. 山东理工大学学报（自然科学版）, 2011(6):19-14.

[11]刘怀军, 车万翔, 刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007(1):79-84.

[12] Peter Harrington. Machine Learning in Action[M]. Connecticut: Manning Publications Co., 2012.

[13] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2012.

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

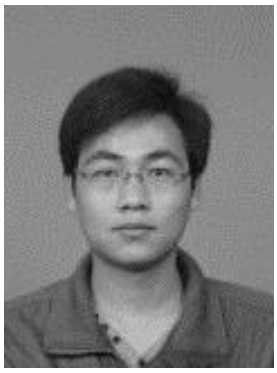
[15] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011,3(2)

[16] Xiugang Li, Dominique Lord etc. Predicting motor vehicle crashes using Support Vector Machine models[J]. Accident Analysis and Prevention, 2008(40): 1611-1618.

[17] 刘一佳. 语言云简介[DB/OL]. http://www.ltp-cloud.com/intro/, 2014-5-29.

[18] 刘丹丹,彭成, 周国栋等. 词汇语义信息对中文实体关系抽取影响的比较[J]. 计算机应用, 2012(8):2238-2244.

作者简介：郭喜跃（1983—），男，博士研究生，主要研究领域为信息抽取。Email: ihnlaoyao@126.com，**通讯作者**；何婷婷（1964—），女，博士，教授，博士生导师，主要研究领域为自然语言处理、数据库与数据挖掘。 Email: tthe@mail.ccnu.edu.cn；胡小华（1965—），男，博士，教授，博士生导师，主要研究领域为人工智能、数据挖掘、自然语言处理.。Email: xh29@drexel.edu；陈前军（1981—），男，博士研究生，主要研究方向：自然语言处理、算法及软件工程。Email: 69643573@qq.com。



(郭喜跃)

（何婷婷）

（胡小华）