文章编号：1003-0077(2014)02-0091-09

《同义词词林》在中文实体关系抽取中的作用

刘丹丹，彭 成，钱龙华，周国栋

（苏州大学 自然语言处理实验室，江苏 苏州 215006；
苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：语义信息在命名实体间语义关系抽取中具有重要的作用。该文以《同义词词林》为例，系统全面地研究了词汇语义信息对基于树核函数的中文语义关系抽取的有效性，深入探讨了不同级别的语义信息和一词多义等现象对关系抽取的影响，详细分析了词汇语义信息和实体类型信息之间的冗余性。在 ACE2005 中文语料库上的关系抽取实验表明，在未知实体类型的前提下，语义信息能显著提高抽取性能；而在已知实体类型的情况下，语义信息也能明显提高某些关系类型的抽取性能，这说明《词林》语义信息和实体类型信息在中文语义关系抽取中具有一定的互补性。

关键词：中文实体关系抽取；树核函数；同义词词林；语义信息

中图分类号：TP391 **文献标识码**：A

The Effect of TongYiCi CiLin in Chinese Entity Relation Extraction

LIU Dandan, PENG Cheng, QIAN Longhua, ZHOU Guodong

(Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu 215006, China;
School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Semantic information plays an important role in the semantic relation extraction between named entities. Taking "TongYiCi CiLin" as an example, this paper systematically investigates the effectiveness of lexical semantic information on tree kernel-based Chinese semantic relation extraction, particularly the influence of different levels of semantic information and polysemy phenomenon, as well as details about the redundancy between lexical semantic information and entity type information. The experiments of relation extraction on the ACE2005 Chinese corpus shows that semantic information can significantly improve the extraction performance without entity types, while in the case of known entity types, semantic information can also noticeably enhance the extraction performance for some relation types. This implies a certain degree of complementarity between "CiLin" semantic information and entity type information in Chinese semantic relation extraction.

Key words: Chinese entity relation extraction; tree kernel; TongYiCi CiLin; semantic information

1 引言

命名实体间语义关系抽取（简称实体关系抽取，或关系抽取）是信息抽取中的一个重要研究内容，其任务是从自然语言文本中提取出两个命名实体之间所存在的语义关系，例如，短语"美国总统*克林顿*的*平壤*之行"中的两个实体"克林顿"（PER）和"平壤"（GPE）之间存在的物理位置关系（PHYS. Located）。作为一项应用基础性研究，实体关系抽取对自然语言处理的许多应用如内容理解、问题回答、自动文摘、机器翻译、文本分类以及信息过滤等都具有重要的意义。

无论是采用指导性的机器学习方法，还是采用无指导的聚类方法，关系抽取研究的关键问题都是如何有效的表达关系实例并计算关系实例之间的相似度。基于特征向量的方法[1-5]将关系实例表示成高维特征空间中的一个向量，通过计算向量之间的

收稿日期：2012-01-06 **定稿日期**：2012-08-21

基金项目：国家自然科学基金（60873150，90920004）、江苏省自然科学基金（BK2010219，11KJA520003）。

相似度来表示实例之间的相似度，其特征包含词汇、组块、句法和语义等各种信息。基于核函数的方法则将关系实例表示成离散结构，如实体对所在的成分句法树[6-10]、依存树[11]或依存路径[12-13]等，它通过计算离散结构之间的相似度来表示实例之间的相似度。由于它能探索高维空间中的隐含结构化特征，因此在关系抽取及自然语言处理的其它任务中获得了广泛的应用。在中文实体关系抽取中，基于特征向量的方法有文献[14-16]等。基于核函数的方法采用的离散结构有字符串[17-18]、句法树[19-20]等。

众所周知，语义信息对实体间语义关系的抽取具有重要的作用。目前关系抽取中使用到的语义信息主要分为以下三类：实体类型语义信息、实体词汇的聚类信息和实体词汇的语义信息。实体类型语义信息包括实体大类和实体小类信息，无论是从语义关系的定义，还是实验结果来看，这类信息对关系抽取的性能具有很大的提升作用，因而几乎所有的关系抽取系统都使用实体类型信息。不过，目前使用的实体类型信息都是基于手工标注的结果，实际识别出的实体类型，特别是小类信息，肯定含有噪音，从而使得其作用受到一定的影响。文献[4-5]先采用聚类的方法得到实体词汇的语义编码，然后在基于特征向量的关系抽取中使用该语义编码，实验结果表明其对关系抽取的性能提高具有一定的促进作用。但由于特征匹配的限制，语义编码必须截断后才能使用。在中文关系抽取中，文献[17]采用编辑距离核函数来计算关系实例的字符串之间的相似度，并考虑了词汇之间在《同义词词林》中的语义相似度，在 person-affiliation 关系中取得了较好的结果。不过，他们没有单独比较词汇语义相似度的贡献，也没有考虑对其它类型的关系抽取的影响。文献[18]采用字符串核的方法进行 ACE 语料库上的三个大类的中文关系抽取，并在子串比较的时候考虑其词汇在《知网》中的词义相似度，实验表明语义相似度能提高大部分关系类型的抽取性能。

综上所述，语义信息确实能够提高关系抽取的性能，但目前还没有一个系统全面的研究来分析语义信息对中文关系抽取的有效性，如对哪些关系类型有效，有效程度如何，以及词汇语义信息和实体类型信息之间的冗余度等。针对这些问题，本文以《同义词词林》为例，采用基于树核函数的方法来研究语义信息在中文实体语义关系抽取中的作用，旨在发现语义信息对哪些关系类型影响最大。

本文第 2 节介绍了《同义词词林》及其编码方式；第 3 节讨论《词林》语义类别信息与结构化信息的结合；第 4 节给出了实验设置及结果分析；最后第 5 节是总结部分。

2 同义词词林

《同义词词林》[21]（以下简称《词林》）是一部汉语分类词典，其中每一条词语都用一个编码来表示其语义类别。本文所用的《词林》为《词林（扩展版）》，是哈尔滨工业大学信息检索研究室在《同义词词林》的基础上研制的。最终的词表包含 77 492 条词语，其中一词多义的词语为 8 860 个，共分为 12 个大类，94 个中类，1 428 个小类，小类下再以同义原则划分词群，最细的级别为原子词群，这样词典中的词语之间就体现了良好的层次关系。不同级别的分类结果可以为自然语言处理提供不同颗粒度的语义类别信息。

《词林》的 12 个大类分别用一位大写英文字母 A 到 L 来表示，中类编号在大写字母后面加一位小写英文字母表示，小类编号再加两位十进制整数表示，词群编号再加一位大写英文字母表示，原子词群编号再加两位十进制整数表示，最后一位的标记有 3 种，其中“=”代表“相等”、“同义”；“#”代表“不等”、“同类”，属于相关词语；“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。根据编码特点，本文没有使用第八位编码。具体的标记如表 1 所示。如词语“公园”的语义编码为“Bn20A01=”，大类（B）表示“物”，中类（Bn）表示“建筑物”，小类（Bn20）表示“园林”，原子词群（Bn20A01）表示“园林 公园 花园 庄园 园 苑”，词群（Bn20A）并没有赋予专门的名称。

表 1 《词林》词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	B	n	2	0	A	0	1	=/#/@
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

3　《词林》语义信息与结构化信息的结合

在分析《词林》语义信息对基于树核函数的中文关系抽取的影响之前，首先需要考虑两个问题：一是应该加入哪些词汇的语义信息；二是词汇的语义信息如何与句法树中的结构化信息相结合。

在表示关系实例结构化信息的句法树中，除两个实体名称外，还包含其它的词汇信息，如动词、形容词和副词等。根据文献[5]的研究，加入实体名称的聚类语义信息有利于提高关系抽取的性能，而其他词汇的语义信息则没有效果。鉴于此，本文只考虑关系实例中的两个实体词汇在《词林》中的语义类别信息。

3.1　实体词汇的《词林》语义类别与结构化信息的结合

对实体而言，其语义信息和句法树中的结构化信息相结合的方法有两种：一是直接将语义类别信息加入到句法树中；二是通过复合核函数的方法将基于结构化信息的树核函数和基于语义类别信息的核函数结合起来。在 ACE RDC 2004 英文语料库上的实验表明[9]，由于后者能调整两种核函数的贡献，因此性能比前者略有提高。但本文的重点在于探索语义信息对关系抽取的作用，为避免复合系数的调整问题，我们采用与文献[20]相似的方法，将语义信息挂在句法树的根结点下面，从而构成合一句法和语义关系树。

例如，在关系实例“*台北 大安森林公园*”中，实体“台北”对应的《词林》“原子词群”编码为 Cb25A11，“词群”编码为 Cb25A，“小类”编码为 Cb25，“中类”编码为 Cb，“大类”编码为 C。如果考虑《词林》“词群”级别的语义信息，就将其对应的语义类别编码“Cb25A”挂在句法树的根结点下，如图 1 所示。其中句法树结构采用最短路径包含树（SPT，Shortest Path-enclosed Tree），而 SC1、SC2 分别表示其子结点为实体 E1 和实体 E2 的词汇所对应的语义编码，“Bn20A”为“大安森林公园”的中心词“公园”的词群编码。

3.2　实体词汇的一词多义信息与结构化信息的结合

一词多义是自然语言中的普遍现象，它对自然语言处理的很多任务都有影响。在 ACE 2005 中文

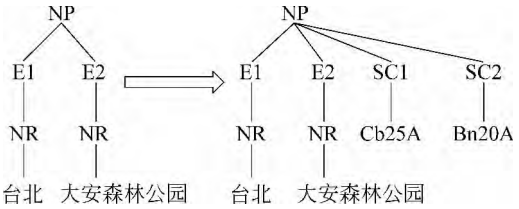

图 1　加入实体《词林》词群语义类别后的句法树

语料库上的统计表明，在《词林》中具有一词多义的实体词汇占其总数的 1/5 还多，因而实体词汇的“一词多义”现象对关系抽取具有一定的影响。

在关系实例中，不同的“一词多义”的实体词汇（简称为多义实体）所具有的词义数是不同的，统计表明词义数为 2 和 3 的多义实体占所有多义实体的 80％左右，而词义数 7 以上的多义实体则非常之少。因此，在考察“一词多义”对关系抽取影响的实验时，我们仅考虑词义数为 2－6 的《词林》语义信息。例如，当词义数为 2 时，图 1 中的实体 E1“台北”在《词林》中具有 2 个词义，其词群编码分别“Cb25A”、“Di03B”。把这两个编码都挂在具有相同标识（即 SC1）的父节点下面，即表示实体 1 的词汇具有两个含义，这样在计算两棵树的相似度时，只要其中任何一个语义编码匹配，相似度就能得到提高。

3.3　实体词汇的《词林》语义信息的获取

为了将实体词汇的语义信息加入到句法树中，在生成了关系实例的 SPT 树之后，需从《词林》中抽取出语义类别信息，并将它插入到句法树中，其处理流程如下：

① 从句法树中找出实体 E1 和 E2 所对应的词汇 LEX1 和 LEX2；

② 在《词林》中查找 LEX1 和 LEX2 的语义类别编码；

③ 如果某一词汇的语义类别编码不存在，则将该词汇进行分词，取分词后最右边的词汇，再在《词林》中查找相应的语义类别编码。设得到的语义类别分别为 CODE1，CODE2；

④ 按照《词林》的不同语义级别对 CODE1，CODE2 进行截段，得到最终的编码分别为 C1，C2；

⑤ 将 C1，C2 分别挂在句法树根结点下的 SC1，SC2 结点下面。

需要说明的是，第 3 步中的分词非常必要，因为很多实体词汇无法在《词林》中找到相应的语义编码。据统计，这一类实体词汇的数量超过实体总数的 1/4。其主要原因是，很多实体的名称都是较少

出现的专用名词，而语义辞典是不收录频度较少的专用名词的，但其中心词则是普通名词，通常可以找到其语义类别。例如，在图1的实例中，“大安森林公园”没有收录在《词林》中，但分词后的中心词“公园”却可以找到语义编码。另外，在分词时，对于人名则不作处理，因为人名虽然不能在《词林》中找到语义编码，但对其进行分词却也没有意义。

最后，当要处理多义实体的一词多义时，则需要在执行第2步时从《词林》中同时找出多个含义所对应的语义编码，同时加入到句法树中。

4 实验设置与结果分析

本节首先给出实验设置，包括所使用的语料库、分词工具和分类器及性能评估指标，然后给出实验结果，并对其进行分析。

4.1 实验设置

本文采用ACE 2005中文语料库作为中文语义关系抽取的实验数据。该语料库定义了中文实体之间的6个关系大类，18个关系小类，它包含633个文件，其中广播新闻类298个，新闻专线类238个，微博和其它97个。采用句法分析器进行句法分析，在去除个别句法分析器不能正确处理的句子后，最终得到关系正例9 147个，关系负例97 540个。

本文的分词工具采用中国科学院计算技术研究所研制的基于多层HMM模型的汉语词法分析系统ICTCLAS[22]。分类器采用支持卷积树核函数的SVMLight TK[23]工具包，由于该工具包是一个二元分类器，我们采用一对多的方法将它转换为多元分类器。特别地，相似度计算采用SST(SubSet Tree)核，衰减系数为0.4。为了充分利用语料库资源，减少语料库变化对实验结论的影响，本文实验采用五倍交叉验证策略，最后取5次平均值作为最终的性能。评估标准采用常用的准确率(P)，召回率(R)和F1指标(F1)。

4.2 实验结果与分析

(1)《词林》不同级别的语义信息对中文关系抽取的影响

图2比较了《词林》的不同级别(即“大类”、“中类”、“小类”、“词群”、“原子词群”)的语义信息对大类和小类关系抽取性能(即F1值)的影响，其中基准系统是指不加入任何语义信息时SPT树所取得的性能，每一次实验分别加入一个级别的语义类别信息，横坐标表示《词林》语义信息的不同级别，并且从左到右粒度不断变细，纵坐标则为关系的抽取性能，性能最高的F1值用粗体显示。

从图2可以看出，分别加入《词林》的“小类”/“词群”级别的语义信息后大类/小类关系抽取的性能最佳，分别比基准系统的F1值提高了4.8/5.9个百分点，这说明《词林》语义信息能显著提高中文关系抽取的性能。

该图同时也表明，无论是大类关系抽取，还是小类关系抽取，随着加入《词林》的语义信息的粒度的细化，F1值都是先升高后降低，且在“小类”/“词群”级别时，性能达到最大值，这说明过于细化或泛化的语义信息都对关系抽取不利。

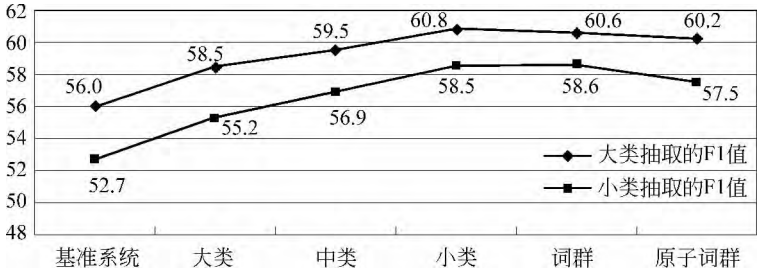

图2 《词林》不同级别的语义信息对中文关系抽取的性能影响

由于加入《词林》的“小类”或“词群”语义信息，对大类和小类的F1值差别都不大(相差0.2或0.1)，因此在后续实验中选取“小类”或“词群”级别的语义原则上都可以。除非特别说明，本文的后续实验都选择“词群”级别的语义信息加入到句法树中。

(2)《词林》语义信息对中文关系抽取具体类别的影响

由前面的实验可以知道，在基准系统的基础上，加入“小类”或“词群”语义信息，关系抽取的性能最高。表2和表3分别列出了加入“词群”语义信息后的性能及其同基准系统之间在各个大类和小类类别

上的性能差异，其中 P/R/F1 为在 5 个数据集上的平均值，△P/△R/△F 分别为在 5 个数据集上的 P/R/F1 的平均变化值，# 表示该关系类别的实例数，% 为该类别的实例数占总数的百分比，～F 为 F1 值的加权平均（即△F＊%/100），它表明了某个类别上 F1 值的变化对总体性能变化的贡献度。每一个性能指标的最大值和最小值分别用加粗的双底划线和单底划线标出。

从表 3 中可以看出，与大类抽取不同的是，加入“词群”语义信息后，并非所有小类的性能都提高，而是呈现出不同的趋势，从△F 值来看：

• F1 值增加幅度在 3 点以上的小类有 10 个，如 Membership(10.0)，Business/Subsidiary(8.5)和 CRRE(7.9)等。这是由于这些关系中的专用名词或其中心词在《词林》中具有相同的词群编码，因此语义信息的加入增加了树结构的相似性。例如，在“*共产党 领袖*”、“*塞尔维亚民主党* 提名 的 *候选* 人”等短语中都存在着 Membership 关系，由于词汇的稀疏性问题，在基准系统中都被误识别为 Employment 关系，而加入实体 E1 的词汇语义编码(Di07A)后，相似度得到提高；

• Near 小类几乎没有增加，Artifact 小类没有变化，而 Founder 和 Ownership 小类则显著降低。这是由于某些词汇的分词错误导致了错误的语义编码，造成了关系的误识别。例如，关系实例“雅虎 创办人”为 Founder 关系，但实体“雅虎”分词后的中心词“虎”明显改变了实体的语义类别，从而导致该关系实例被错误识别。

表 2　“词群”语义信息对关系抽取大类类别的性能影响

关系大类	#	%	P	R	F1	△P	△R	△F	～F
PHYS	1 552	17.0	66.5	11.1	19.1	**2.2**	**1.9**	**2.9**	0.49
PART-WHOLE	2 249	24.6	76.0	64.0	69.5	**6.5**	3.2	4.6	1.13
PER-SOC	652	7.2	78.9	34.7	48.1	4.9	2.3	3.2	**0.23**
ORG-AFF	2 166	23.7	83.2	65.5	73.3	3.3	4.2	4.0	0.94
ART	623	6.8	67.9	19.6	30.3	5.7	5.1	7.0	0.47
GEN-AFF	1 905	20.8	79.0	63.6	70.4	3.2	**8.3**	6.5	**1.36**
合计	9 147	100	76.4	50.2	60.6	4.2	4.4	4.6	4.6

表 3　“词群”语义信息对关系抽取小类类别的性能影响

关系大类	关系小类	#	%	F1	△P	△R	△F	～F
PHYS	Located	1 335	14.5	16.1	5.2	1.9	3.1	0.45
	Near	217	2.4	35.8	5.6	−0.5	0.2	0
PART-WHOLE	Geographical	1 257	13.7	64.0	7.6	4.5	5.8	0.80
	Subsidiary	978	10.7	73.9	7.7	9.0	8.5	0.91
	Artifact	14	0.2	0.0	0.0	0.0	0.0	0
PER-SOC	Business	186	2.1	43.3	8.1	7.0	8.5	0.18
	Family	382	4.2	51.4	4.9	0.3	1.4	0.06
	Lasting-Personal	84	0.9	14.6	−1.7	1.3	2.0	0.02
ORG-AFF	Employment	1 560	17.0	75.5	4.1	4.3	4.2	0.71
	Ownership	22	0.3	30.7	10.0	−5.0	−4.0	−0.01
	Founder	17	0.2	18.0	**−20.0**	**−13.3**	**−16.0**	**−0.04**
	Student-Alum	69	0.8	17.3	−5.3	1.4	1.8	0.01
	Sports-Affiliation	69	0.8	32.3	13.3	4.2	6.6	0.05
	Investor-Shareholder	85	0.9	18.0	**25.0**	3.5	5.7	0.05
	Membership	344	3.8	56.6	10.3	**9.6**	**10.0**	0.38

续表

关系大类	关系小类	#	%	F1	△P	△R	△F	~F
ART	UOIM	623	6.8	31.8	5.2	5.0	6.6	0.45
GEN-AFF	CRRE	732	8.0	63.8	3.1	9.8	7.9	0.63
	Org-Location	1 173	12.8	72.5	5.1	9.4	7.8	**1.00**
合计		9 147	100	58.6	5.7	5.4	5.9	5.9

将表2和表3综合起来考虑，可以发现：

• 由于GEN-AFF大类中的两个小类均有大幅度提高，且所占比例较高（约20%），因而导致该大类的性能贡献度~F最大；

• ORG-AFF大类中的各个小类表现差别迥异，因而虽然该大类所占比例较高，但总体性能贡献值却小于PART-WHOLE和GEN-AFF两大类。

综上所述，《词林》语义信息对所有大类关系抽取的F1值都有不同程度的提高，尤其对ART和GEN-AFF两大类的影响最大；而对大部分小类关系抽取的性能也有不同程度的提高，如Membership，Subsidiary，Business和CRRE等提高幅度较大，而对Founder和Ownership等部分小类则明显降低。

（3）《词林》中实体词汇的一词多义现象对关系抽取性能的影响

图3比较了《词林》中实体词汇的一词多义对大类和小类关系抽取性能（即F1值）的影响，每一次实验都是在前面实验的基础上再加入一个额外的语义信息，横坐标表示词义数从1变化到6，纵坐标则表示抽取性能的F1值。同样，最高性能用粗体表示。

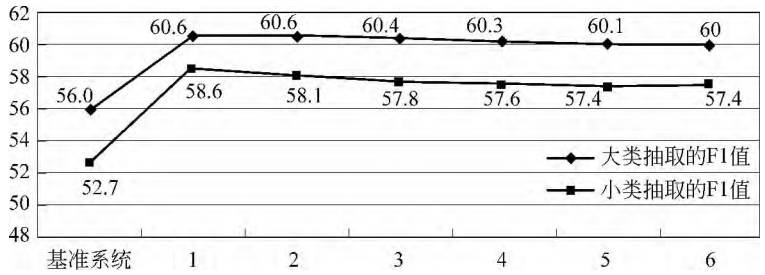

图3 《词林》中的实体词汇的一词多义对中文关系抽取的性能影响

由图3可以看出，加入一词多义信息并不能改善关系抽取的性能，反而随着多义词词义数的不断增加，F1值逐渐下降。通过分析，发现其原因是由于关系实例中的实体词汇在ACE新闻类语料库中的语义通常都是较为常见的一种，考虑一词多义（即加入该实体不常用的语义）后，反而增加了噪音信息，并且树的结构更为庞大，从而降低了关系抽取的性能。

（4）《词林》语义信息与实体类型信息的冗余度

实体本身也有大类和小类等类别信息，它们和实体词汇的语义信息之间是否存在冗余呢？我们首先从总体性能上分析了《词林》语义信息和实体类型信息的性能影响，然后从具体关系类别上进行比较。

1. 从总体性能上比较《词林》语义信息与实体类型信息的影响

表4比较了在基准系统的基础上，加入不同组合的《词林》词群语义信息和实体类型信息（实体大类和小类）后中文关系抽取的总体性能，其中大类和小类关系抽取的最高性能用粗体表示。

表4 《词林》语义信息和实体类型信息的性能比较

实体类型/《词林》语义信息	大类关系抽取			小类关系抽取		
	P/%	R/%	F1	P/%	R/%	F1
基准系统	72.2	45.8	56.0	69.1	42.7	52.7
实体大类	79.2	54.8	64.8	76.0	52.6	62.2
实体小类	80.2	55.3	65.4	77.7	53.5	63.4
词林词群	76.4	50.2	60.6	74.8	48.1	58.6
实体大类+词林词群	80.8	55.9	66.1	78.6	54.0	64.0
实体小类+词林词群	81.4	55.6	66.0	79.4	53.8	64.1
实体大类+实体小类	80.4	56.5	66.4	77.1	54.3	63.7
实体大类+实体小类+词林词群	**81.8**	**56.5**	**66.8**	**79.8**	**54.6**	**64.8**

从表4可以看出，同基准系统相比，加入实体大类、实体小类和《词林》语义等所有信息后，无论是大

类抽取，还是小类抽取都取得了最好的性能，F1 值分别为 66.8/64.8，且 P 值和 R 值同时显著提高，这说明这些语义信息对中文关系抽取都有一定的作用。此外，该表还表示：

• 单独加入实体大类、实体小类或词林词群等信息之一，实体小类取得了最好的性能提高。这说明实体小类信息能更准确地刻画实体的本质，更好地区分关系的类型，而《词林》词群语义信息尽管类别更细，但它是针对通用领域的，不一定最适合新闻领域的关系抽取；

• “实体小类＋词林词群”的大类 F1 值比“词林词群”的大类 F1 值高出 5.4 点，而比“实体小类”的大类 F1 值只高出 0.6 点，这说明就关系抽取而言，实体小类覆盖了词林词群中的大部分语义信息，反之则不然。同理，实体大类也覆盖了词林词群中的大部分语义信息，因为“实体大类＋词林词群”的大类 F1 值比“词林词群”的大类 F1 值高出 5.5 点，而比“实体大类”的大类 F1 值只高出 1.3 点。

• 最后很重要的一点是，在“基准系统”的基础上加入“词林词群”，大类抽取的 F1 值提高了 4.6 点，小类抽取的 F1 值提高了 5.9 点，而在“实体大类＋实体小类”的基础上，再加入“词林词群”，大类抽取的 F1 值只提高了 0.4 点，小类抽取的 F1 值也只提高了 1.1 点。可以看出实体类型的加入严重削弱了语义信息对抽取性能的提高幅度，那么这是否意味着语义信息对关系抽取来说意义就不大了呢？答案是否定的。其一，我们现在加入实体类型时，假设它是完全正确的。在实际的命名实体识别系统中，总会有错误产生，尤其是对于实体小类，因而实际应用中的实体类型是有噪音的，它对性能的提高不可能有预期的那么大，而《词林》语义信息则是从现存的语义辞典《同义词词林》中提取的，它不存在这个问题。其二，语义信息对不同关系类型的抽取性能表现出多样性，这就是下面的分析所要说明的问题。

2. 从具体关系类型的性能上比较“词群”语义和实体类型的影响

为了比较《词林》语义信息和实体类型信息的冗余性对具体关系类型抽取的影响，表 5 列出了各个小类关系的 F1 值、△F 值。其中“词林词群-BL”和“实体类型-BL”分别表示在基准系统的基础上加入词林词群或实体类型（实体大类＋实体小类）后的 F1 值和△F 值，“（类型＋词群）-类型”表示在实体类型的基础上加入词林词群后的 F1 值和△F 值，小类关系按此△F 值降序排列。从表 5 中可以看出：

表 5　实体类型信息与词林语义在小类关系上的 F1 值及其变化

关系小类	词林词群-BL		实体类型-BL		（类型＋词群）-类型	
	F1	△F	F1	△F	F1	△F
Business	43.3	8.5	35.7	0.9	47.1	11.4
Lasting-Personal	14.6	2.0	2.2	−10.3	12.9	10.7
Sports-Affiliation	32.3	6.6	35.4	9.6	40.9	5.5
Investor-Shareholder	18.0	5.7	14.4	2.2	19.9	5.4
Student-Alum	17.3	1.8	14.1	−1.4	18.6	4.6
CRRE	63.8	7.9	66.9	11.0	71.1	4.2
Membership	56.6	10.0	57.3	10.6	59.8	2.6
Family	51.4	1.4	47.2	−2.8	49.4	2.2
Employment	75.5	4.2	77.7	6.5	79.5	1.7
Ownership	30.7	−4.0	14.7	−20.0	16.0	1.3
Subsidiary	73.9	8.5	80.6	15.2	81.8	1.2
Near	35.8	0.2	40.5	4.9	41.7	1.2
Org-Location	72.5	7.8	80.4	15.7	80.4	0.0
Founder	18.0	−16.0	0.0	−34.0	0.0	0.0
Artifact	0.0	0.0	0.0	0.0	0.0	0.0

续表

关系小类	词林词群-BL		实体类型-BL		(类型+词群)-类型	
	F1	△F	F1	△F	F1	△F
Geographical	64.0	5.8	73.1	14.8	72.6	−0.5
Located	16.1	3.1	27.9	14.9	26.0	−1.8
UOIM	31.8	6.6	46.1	20.9	43.4	−2.7

• 在表格中双划线以上的小类关系，如 Business，Lasting-Personal 和 Sports-Affiliation 等，在实体类型的基础上再加入《词林》语义信息时，其性能提高幅度(△F 值)都在 1 点以上。尤其是三个小类关系(用底划线表示)，Business、Lasting-Personal 和 Student-Alum，单独加入实体类型并不能明显提高性能(0.9/−10.3/−1.4)，甚至降低，但在加入实体类型后，《词林》语义信息显示了它更强劲的性能提升作用。这说明对于这些小类关系而言，实体类型信息和《词林》语义信息可以相互补充，并且只有这样才能更好地抽取这些小类关系；

• 在表格中双划线以下的小类关系(除占比例较少的 Founder 和 Artifact 小类关系之外)，如 Org-Location，Geographical，Located 等，实体类型的加入，严重削弱了《词林》语义信息对抽取性能的提升作用。即单独加入实体类型就已经取得了非常显著的性能提升，再加入《词林》语义信息不会明显提高其性能，特别是对 Geographical/Located/UOIM 等小类，《词林》语义信息的加入反而损害了它们的抽取性能，这说明对这些小类关系而言，实体类型信息已包含了大部分的《词林》语义信息内涵，两者冗余度较高。

综上所述，虽然从总体性能上看，在已知实体类型的前提下，加入《词林》语义信息的效果不明显，但是，如果是对某些特定语义关系的抽取，如 Business，Lasting-Personal 和 Student-Alum 以及 Sports-Affiliation，Investor-Shareholder 和 CRRE 等，加入《词林》语义信息还是非常有用的。

5 总结与展望

本文利用了现有的中文语义资源《同义词词林》，探讨了《词林》语义对中文关系抽取的影响，通过实验我们发现，《词林》词群级别的语义信息能显著提高中文关系抽取的性能，但考虑一词多义却不能提高抽取性能。另外，《词林》词群语义和实体类型信息存在着一定程度的冗余，因此在已知实体类型的前提下加入《词林》词群语义时关系抽取总体性能提高较少，但是对某些特定语义关系的抽取，如 Business，Lasting-Personal 等，性能却有明显的提升，这说明只有《词林》语义信息和实体类型信息相互补充，相辅相成，才能更好地提升中文语义关系抽取的性能。

下一步的研究工作我们将从以下几个方面展开，一是通过将词汇语义相似度嵌入到树核函数中的方法来考虑语义信息对关系抽取的影响，并和本文的方法进行比较；二是考虑实体信息自动标注的情况下，实体类型和词汇语义信息对关系抽取的影响；三是将中文抽取方面的研究工作推广到英文关系抽取中，考察 WordNet 对关系抽取的影响。

参考文献

[1] Nanda Kambhatla. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations [C]//Proceedings of the ACL. Morristown, NJ, USA, 2004: 178-181.

[2] Zhou GuoDong, Su Jian, Zhang Jie, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the ACL, 2005:427-434.

[3] Zhou G D, Qian L H, Fan J X. Tree kernel-based semantic relation extraction with rich syntactic and semantic information[C]//Proceedings of the Information Sciences, 2010:1313-1325.

[4] Chan Y S, Roth D. Exploiting Background Knowledge for Relation Extraction[C]//Proceedings of the COLING, 2010:152-160.

[5] Sun A, Grishman R, Sekine S. Semi-supervised Relation Extraction with Large-scale Word Clustering [C]//Proceedings of the ACL, 2011:521-529.

[6] Zhang M, Zhang J, Su J, et al. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features[C]//Proceedings of the COLING-ACL. Sydney, Australia, 2006:825-832.

[7] Zhou G D, Zhang M, Ji D H, et al. Tree Kernel-based

Relation Extraction with Context-Sensitive Structured Parse Tree Information[C]//Proceedings of the EMNLP/CoNLL. Prague,Czech, 2007:728-736.

[8] Zhou G D, Zhu Q M. Kernel-based semantic relation detection and classification via enriched parse tree structure[J]. Journal of Computer Science and Technology. 2011. 26(1):45-56.

[9] Qian L H, Zhou G D, Kong F, et al. Exploiting constituent dependencies for tree kernel-based semantic relation extraction[C]//Proceedings of the COLING. Manchester, 2008:697-704.

[10] Qian L H, Zhou G D, Zhu Q M. Employing Constituent Dependency Information for Tree Kernel-based Semantic Relation Extraction between Named Entities [C]//Proceedings of the ACM Transaction on Asian Language Information Processing. 2011. 10(3): Article 15(24pages).

[11] Culotta A, Sorensen J. Dependency tree kernels for relation extraction[C]//Proceedings of the ACL. Barcelona, Spain, 2004:423-429.

[12] Bunescu R C, Raymond J M. A Shortest Path Dependency Kernel for Relation Extraction[C]//Proceedings of the EMNLP. Vancover, B.C, 2005:724-731.

[13] Nguyen T T, Moschitti A, Riccardi G. Convolution Kernels on Constituent, Dependency and Sequential Structures for Relation Extraction[C]//Proceedings of the EMNLP, 2009: 1378-1387.

[14] **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.

[15] **董静，孙乐，冯元勇，黄瑞红. 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, 91.

[16] Li W J, Zhang P, Wei F R, et al. A Novel Feature-based Approach to Chinese Entity Relation Extraction [C]//Proceedings of the ACL. Columbus, Ohio, USA, 2008: 89-92.

[17] Che W X, Jiang J M, Su Z, et al. Improved-Edit-Distance Kernel for Chinese Relation Extraction[C]//Proceedings of the IJCNLP. 2005: 132-137.

[18] **刘克彬，李芳，刘磊，韩颖. 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.

[19] **黄瑞红，孙乐，冯元勇，黄云平. 基于核方法的中文实体关系抽取研究**[J]. **中文信息学报**, 2008, 22(5): 102-108.

[20] **虞欢欢，钱龙华，周国栋，朱巧明. 基于合一句法和实体语义树的中文语义关系抽取**[J]. **中文信息学报**, 2010,24(5): 17-23.

[21] **梅家驹，竺一鸣，高蕴琦，殷鸿翔. 同义词词林(第二版)**[M]. **上海:上海辞书出版社**, 1996.

[22] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese Lexical Analyzer ICTCLAS[C]//Proceedings of the 2nd SIGHAN workshop affiliated with 41th ACL. Sapporo Japan, 2003:184-187.

[23] Moschitti A. A Study on Convolution Kernels for Shallow Semantic Parsing[C]//Proceedings of the ACL. Barcelona, Spain, 2004:335.

刘丹丹(1987—)，**硕士研究生，主要研究领域为信息抽取**。
E-mail: liudandan219@163.com

钱龙华(1966—)，**副教授，硕士生导师，主要研究领域为自然语言处理**。
E-mail: qianlonghua@suda.edu.cn

彭成(1987—)，**硕士研究生，主要研究领域为信息抽取**。
E-mail: 719864778@qq.com