Journal of Computer Applications
计算机应用,2012,32(8):2238 -2244
ISSN 1001-9081
CODEN JYIIDU
2012-08-01
http://www.joca.cn

文章编号:1001 -9081(2012)08 -2238 -07
doi:10.3724/SP.J.1087.2012.02238

词汇语义信息对中文实体关系抽取影响的比较

刘丹丹,彭　成,钱龙华*,周国栋
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)
(*通信作者电子邮箱 qianlonghua@suda.edu.cn)

摘　要:提出一种将《同义词词林》和《知网》的语义信息融合到基于树核函数的中文关系抽取方法,并比较和分析了两种语义信息对中文实体关系抽取的影响,同时探讨了这两种语义信息与实体类型信息之间的相互关系。实验结果表明,该方法能在一定程度上提高中文关系抽取的性能;同时,《同义词词林》能补充实体类型信息的不足,因而无论是否加入实体类型信息,其语义信息都能大幅度地提高大部分关系类型的抽取性能;而《知网》则和实体类型信息存在冲突,因此在已知实体类型信息的前提下,仅能提高个别关系类型的抽取性能。

关键词:中文实体关系抽取;树核;《同义词词林》;《知网》;语义信息
中图分类号:TP18　　**文献标志码**:A

Comparative analysis of impact of lexical semantic information on Chinese entity relation extraction

LIU Dan-dan, PENG Cheng, QIAN Long-hua*, ZHOU Guo-dong
(*School of Computer Science and Technology, Soochow University, Suzhou Jiangsu* 215006, *China*)

Abstract: A method was proposed to incorporate semantic information based on TongYiCi CiLin and HowNet into tree kernel-based Chinese relation extraction, the impact of these two kinds of semantic information on Chinese entity relation extraction was compared and analyzed, and the interrelation between lexical semantic information and entity type information was explored. The experimental results show that this method can improve the performance of Chinese relation extraction in some degree, and TongYiCi CiLin can complement the entity type information to a certain extent. Therefore, no matter whether the entity type information is involved or not, its semantic information can significantly improve the extraction performance for most of the relation types, while some conflicts exist between HowNet and the entity type information, leading to its performance improvements only for several relation types when entity types are provided.

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

0　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)是信息抽取中的重要环节,也是自然语言处理领域的热点问题之一,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,如短语"台北 大安森林公园"中的两个实体"台北"(GPE, Geo-Political Entity, 地理政治实体)和"大安森林公园"(FAC, Facility, 设施)之间存在的部分整体关系(PART-WHOLE. Geographical)。实体关系抽取作为一项应用基础性研究,对自然语言处理的许多应用如内容理解、自动问答、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。

关系抽取通常采用指导性的机器学习方法,它可以根据训练数据(即关系实例)的表达方式分为基于特征向量的方法和基于核函数的方法两类。基于特征向量的方法有文献[1-3]等,其特征包含词汇、组块、句法和语义等各种信息。在基于核函数的方法中,将关系实例表示成的离散结构有实体对所在的成分句法树[4-7]、依存树[8]或依存路径[9-10]等。以上研究都是针对英文实体关系抽取。在中文实体关系抽取中,基于特征向量的方法有文献[11-13]等;基于核函数的方法采用的离散结构有字符串[14-15]、句法树[16-17]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。文献[14]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,在人物附属关系(person-affiliation)中取得了较好的结果;文献[15]根据《知网》中的语义知识获取词汇语义相似度,在核函数中嵌入了《知网》的语义信息,对个别小类关系在不同的使用方法上作了对比分析。这表明从总体上词汇语义信息对关系抽取具有一定的积极作用,但上述文献没有对词汇语义信息对关系抽取的贡献进行深入的分析,如对哪些关系类型有效,对哪些关系类型无效,也没有全面比较《同义词词林》和《知网》两种语义信息对关系抽取影响的异同。

本文提出了一种将《同义词词林》和《知网》的语义信息加入到关系实例的结构化信息中的方法,采用基于树核函数的方法实现中文实体关系抽取,并将这两种语义对中文关系抽取的影响作了全面的对比和分析。

1　《同义词词林》和《知网》

1.1　同义词词林

《同义词词林》[18](以下简称《词林》)是一部汉语分类词

收稿日期:2012-02-24;**修回日期**:2012-04-12。
基金项目:国家自然科学基金资助项目(60873150,90920004);江苏省自然科学基金资助项目(BK2010219,11KJA520003)。
作者简介:刘丹丹(1987-),女,山东滕州人,硕士研究生,主要研究方向:信息抽取;　彭成(1987-),男,安徽六安人,硕士研究生,主要研究方向:信息抽取;　钱龙华(1966-),男,江苏苏州人,副教授,CCF 会员,主要研究方向:自然语言处理;　周国栋(1967-),男,江苏溧阳人,教授,博士生导师,CCF 高级会员,主要研究方向:自然语言处理。

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)》,是哈尔滨工业大学社会计算与信息检索研究中心在《同义词词林》的基础上研制的,最终的词表包含 77 492 条词语,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。《词林》中的语义类别体现了良好的层次关系,这种层次关系可以为自然语言处理提供不同颗粒度的语义类别信息。通过初步实验发现,加入《词林》中"词群"级别的语义信息时对关系抽取的帮助最大,所以后续的实验中除非特别说明,都是指加入《词林》中"词群"级别的语义类别信息。

表 1　《词林》词语编码表

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

《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。根据上述编码特点,本文只使用了前面的 7 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。

1.2　知网

与《词林》不同,《知网》[29-30]不是义类词典,并不是简单地将所有的"概念"归结到一个树状的概念层次体系中,而是试图用一系列的"义原"来对每一个"概念"进行描述。《知网》一共采用了 1 500 多个义原,这些义原分为以下几个大类:1) Event|事件;2) entity|实体;3) attribute|属性值;4) aValue|属性值;5) quantity|数量;6) qValue|数量值;7) SecondaryFeature|次要特征;8) syntax|语法;9) EventRole|动态角色;10) EventFeatures|动态属性。

与《词林》中的一个词语可能对应多个语义编码相似,在《知网》中每一个词语的概念定义也用多个义原来描述,即词语与义原之间是一对多关系,且第一基本义原反映了一个概念最主要的特征,因此本文仅抽取了词语的第一基本义原作为它的语义类别。这一点与仅考虑某一词语在《词林》中的第一个语义编码的做法相类似。这样统一的处理方法便于较公平地比较《词林》和《知网》两种语义资源对关系抽取的影响,尽管这种方法对《知网》的使用较为浅显。

2　融入语义信息的中文关系抽取

在分析《词林》和《知网》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,还包含其他的词汇信息,如动词、形容词和副词等。根据文献[3]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;文献[7]的研究表明,在英文关系抽取中加入动词的原形可提高关系抽取的性能。鉴于此,本文只考虑关系实例中的两个实体词汇及相关动词在《词林》和《知网》中的语义类别信息。

2.1　实体词汇语义类别与结构化信息的结合方法

由于文献[17]提出的合一句法和语义关系树在中文关系抽取中取得了较好的性能,因此本文采用该句法树作为关系实例的基本表达方式。同时,为了实现语义信息和该句法树的结合,本文将《词林》或《知网》中取得的实体词汇及动词的语义类别都挂到句法树根节点下面,从而构成了用于关系抽取的结构化信息。

例如,在关系实例"台北 大安森林公园"中(其中实体词汇的中心词用下划线表示),实体"台北"和"大安森林公园"对应的《词林》"词群"编码分别为 Cb25A 和 Bn20A,对应的《知网》第一基本义原分别为"地方"和"设施"。若同时考虑实体词汇对应的《词林》和《知网》语义信息,则将它们所对应的《词林》词群编码和《知网》第一基本义原同时挂在句法树的根节点下,如图 1 所示。其中句法树结构采用最短路径包含树(Shortest Path-enclosed Tree, SPT),而 SC1、SC2 分别表示其子节点为实体 E1 和实体 E2 的词汇所对应的《词林》语义编码,SHN1 和 SHN2 分别表示其子节点为实体 E1 和 E2 的词汇所对应的《知网》第一基本义原。"Bn20A"和"设施"为"大安森林公园"的中心词"公园"的词群编码和第一基本义原。

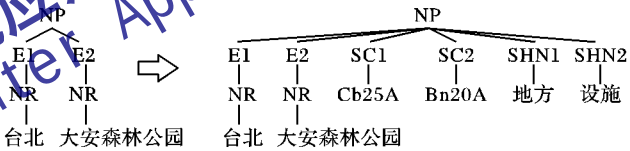

图 1　加入实体《词林》词群和《知网》第一基本义原语义后的句法树

2.2　动词语义类别与结构化信息的结合方法

关系实例中的一些动词对关系的识别起着重要作用。在 ACE2005 中文语料中,有很多动词的语义都很相似,比如"从"、"到"、"来"、"进入"、"访"、"访问"等,这些词都为趋向动词。若把这些动词在《词林》或《知网》对应的语义信息加入到结构化信息中,则能增加同一类型关系实例的结构树之间的相似性。例如:在"巴拉克 可能 于 12 月 访 美"、"代表团 于 10 月 30 日 至 11 月 1 日 访问 了 丹麦"、"领导人 第一 次 访问 安曼"、"官员 访问 华盛顿"等句子中,都存在着物理位置(PHYS)关系。这四个实例的动词为"访"、"访问"(以下划线显示),对应的《词林》词群语义编码都为 Hi02A,对应的《知网》第一基本义原语义都为"看望"。当考察这些关系实例的动词在《词林》或《知网》中的语义信息对中文关系抽取的影响时,分别把"Hi02A"或"看望"挂在标识为 SCV 和 SHNV 的父节点下面。

2.3　《词林》和《知网》语义信息的产生流程

为了将实体词汇和动词的语义信息加入到句法树中,在生成了关系实例的合一句法和语义关系树之后,需从《词林》和《知网》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、LEX2,找出 E1 和 E2 之间离 E2 较近的动词 VLEX。

2) 在《词林》和《知网》中分别查找 LEX1、LEX2 和 VLEX 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》或《知网》中查找

相应的语义类别编码。设在《词林》中得到的语义类别分别为 CODE1、CODE2 和 VCODE,在《知网》中得到的语义类别分别为 HCODE1、HCODE2 和 HVCODE。

4)考虑《词林》语义信息时,则将 CODE1、CODE2 和 VCODE 分别挂在句法树根节点下的 SC1、SC2、SCV 节点下面;考虑《知网》语义信息时,则将 HCODE1、HCODE2 和 HVCODE 分别挂在句法树根节点下的 SHN1、SHN2、SHNV 节点下面。

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》和《知网》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,而语义词典是不收录频度较少的专用名词的,不过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》和《知网》中,但其分词后的中心词"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。

2.4 基于树核函数的中文关系抽取流程

在构建了融入《词林》和《知网》语义信息的结构化信息之后,就可以采用指导性的机器学习方法来实现关系抽取,即首先在训练集上用分类器训练得到分类模型,然后在测试集上用分类模型进行关系抽取并评估抽取方法的优劣。其抽取流程为:

1)将语料库所有文件分为 5 等份,其中 4 份作为训练集,另外 1 份作为测试集;

2)对训练集和测试集中的句子采用 Charniak 句法分析器[21]进行句法分析,从而得到每个句子的句法树;

3)对每个句子中的所有实体进行两两配对得到关系实例,抽取每个关系实例的合一句法和语义关系树;

4)将合一句法和语义关系树和从《词林》或《知网》中提取的语义类别结合为相应的结构化信息(具体细节参见 2.3 节);

5)对训练集中的关系实例采用 SVM 分类器训练得到分类模型;

6)利用分类模型对测试集中的关系实例进行关系抽取,并计算相应的准确率、召回率和 $F1$ 指标。

3 实验设置与结果分析

3.1 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语料。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 38 个,微博和其他类 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隐马尔可夫模型(Hidden Markov Model, HMM)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 SVM^{Light}TK 工具包[23],由于该工具包是一个二元分类器,本文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,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($F1$),其计算公式如下:

$$P = C/T,\quad R = C/N,\quad F1 = 2PR/(P+R)$$

其中:C 为某类被正确分类的实例个数,T 为分类器预测的某类实例总数,N 为测试数据中某类实例总数,而 $F1$ 为 P 和 R 的调和平均值。

3.2 实验结果与分析

3.2.1 《词林》和《知网》语义信息对关系抽取的性能影响

表 2 和表 3 分别列出了加入《词林》语义信息(指加入实体词汇对应的词林"词群"编码)和《知网》语义信息(指加入实体词汇对应的知网"第一基本义原")后的性能及其同基准系统(指不包含实体类型信息的最短路径包含树[17])之间在各个大类和小类类别上的性能差异。其中:$F1$ 为基准系统在 5 个数据集上的平均值,ΔF 分别为在 5 个数据集上的 $F1$ 的平均变化值,S 表示该关系类别的实例数,Q 为该类别的实例数占总数的百分比,$\bar{F}$ 为 ΔF 的加权值(即 $\Delta F * Q/100$),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。其中"词林-BL"、"知网-BL"和"(词林 + 知网)-BL"分别表示在基准系统(BL)的基础上加入词林语义信息、知网语义信息和同时加入词林和知网语义信息。每一个性能指标的最大值和最小值分别用波浪线和单底划线标出。

表 2 《词林》和《知网》语义信息对关系抽取大类类别的性能影响

关系大类	S	Q	BL	词林-BL		知网-BL		(词林 + 知网)-BL	
			$F1$	ΔF	$\bar{F}$	ΔF	$\bar{F}$	ΔF	$\bar{F}$
PHYS	1 552	17.0	16.2	2.9	0.5	3.6	0.6	5.6	1.0
PART-WHOLE	2 249	24.6	64.9	4.6	1.1	5.1	1.3	6.2	1.5
PER-SOC	652	7.2	44.9	3.2	0.2	3.4	0.2	3.8	0.3
ORG-AFF	2 166	23.7	69.3	4.0	0.9	4.3	1.0	5.9	1.4
ART	623	6.8	23.4	7.0	0.5	7.8	0.5	8.4	0.6
GEN-AFF	1 905	20.8	63.9	6.5	1.4	6.6	1.4	8.2	1.7
合计	9 147	100	56.0	4.6	4.6	5.3	5.3	6.4	6.4

从表 2 可以看出,在基准系统的基础上分别加入《词林》和《知网》语义信息时,各个大类的抽取性能都有大幅度的提高,使得大类抽取的总体性能 $F1$ 值分别增加了 4.6 和 5.3;而两者同时加入时,总体性能则提高了 6.4。这说明《词林》和《知网》语义信息对关系抽取都有一定的帮助,且两者具有较大的相似性和一定的互补性,具体体现在:

1)在所有关系大类上,《知网》的性能略高于《词林》,而两者同时加入时的性能均不同程度地大于加入单一语义信息时的性能,这说明了两者之间存在着互补性;单独加入《词林》或《知网》语义时,其 $F1$ 值相差并不大,这说明了两者之间存在着相似性。

2)对于两种语义信息,ART 大类的 ΔF 值都是最大,且提高幅度相近,而影响最小的分别是 PHYS 和人物—社会关系(PER-SOC)大类。

3)通用附属关系(GEN-AFF)大类的 ΔF 值不是最大,但 $\bar{F}$ 值却是最大,这是由于该类关系实例在语料库中占有较大

的比例(约为 20%);反之,PER-SOC 大类的 ΔF 值较低,其相应的实例所占比例也较低(约为 7%),因而导致其 $\bar{F}$ 值最小。

表 3　《词林》和《知网》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	S	Q	BL	词林-BL		知网-BL		(词林+知网)-BL	
				$F1$	ΔF	$\bar{F}$	ΔF	$\bar{F}$	ΔF	$\bar{F}$
PHYS	Located	1 335	14.5	13.0	3.1	0.45	5.1	0.74	5.4	0.79
	Near	217	2.4	35.6	0.2	0.00	2.0	0.05	2.0	0.05
PART-WHOLE	Geographical	1 257	13.7	58.3	5.8	0.79	6.0	0.82	6.7	0.91
	Subsidiary	978	10.7	65.5	8.5	0.91	7.1	0.76	10.2	1.09
	Artifact	14	0.2	0.0	0.0	0.00	0.0	0.00	0.0	0.00
PER-SOC	Business	186	2.1	34.8	8.5	0.18	4.8	0.10	5.4	0.11
	Family	382	4.2	50.0	1.4	0.06	0.8	0.04	0.9	0.04
	Lasting-Personal	84	0.9	12.6	2.0	0.02	-6.2	-0.06	2.1	0.02
ORG-AFF	Employment	1 560	17.0	71.3	4.2	0.72	3.1	0.54	5.3	0.90
	Ownership	22	0.3	34.7	-4.0	-0.01	-5.3	-0.01	-5.3	-0.01
	Founder	17	0.2	34.0	-16.0	-0.04	-24.0	-0.05	-24.0	-0.05
	Student-Alum	69	0.8	15.5	1.8	0.01	-3.0	-0.02	-1.1	-0.01
	Sports-Affiliation	69	0.8	25.8	6.6	0.05	1.6	0.01	5.9	0.04
	Investor-Shareholder	85	0.9	12.2	5.7	0.05	5.7	0.05	5.5	0.05
	Membership	344	3.8	46.6	10.0	0.38	7.4	0.28	10.6	0.40
ART	UOIM	623	6.8	25.2	6.6	0.45	5.3	0.36	7.7	0.52
GEN-AFF	CRRE	732	8.0	55.9	7.9	0.63	8.8	0.71	10.4	0.83
	Org-Location	1 173	12.8	64.7	7.8	1.00	7.0	0.90	9.3	1.06
	合计	9 147	100	52.7	5.9	5.90	5.5	5.50	6.9	6.90

从表 3 可以看出,与大类抽取不同的是,单独加入两种语义信息时,并非所有小类的性能都得到了提高,而是呈现出不同的趋势,但其总体性能却都有提高,而两者同时加入的总体性能同样也有一定程度的提高。特别地,对将近一半的小类关系,包括 Located, Geographical, Subsidiary, Employment, Membership, UOIM, CRRE 和 Org-Location 等,两者单独提高性能的程度相似,而同时加入时性能又得到进一步的提高。

不过,从具体小类关系来看,两者略有差别。加入《知网》比加入《词林》的 ΔF 值增加的小类有 4 个,分别为 Located, Near, Geographical 和 CRRE 等,其中 Located 关系性能提高得最明显。而前者比后者的 ΔF 值降低 2 及以上的小类有 5 个,分别为 Business、Lasting-Personal 和 Student-Alum 等,其中 Lasting-Personal 降幅最大。经过分析,发现 Located 关系实例大多数都为“人”与“国家”之间的关系,例如“留学生 在 美国”、“官兵 在 伊拉克”、“他 前往 沙特阿拉伯”;而 PER-SOC 中三个小类的关系实例都为“人”与“人”之间的关系,例如“留学生 的 同学”、“他 最 好 的 朋友”。可以看出这两种关系中的实体都与“人”有关,“人”在《知网》中的概念较泛化,没有进一步划分,其对应的第一基本义原都为“人”;而在《词林》中的概念较细化,例如“留学生”、“官兵”、“他”,它们所对应的词群编码分别为“Ae13B”、“Ae10B”、“Di04A”。Located 关系中的人物实体不需要进一步划分,因而过于细化的《词林》语义信息不利于关系的抽取;而在 PER-SOC 关系中的人物实体需要进一步划分,因而细化的《词林》语义信息有利于关系的抽取。由此可见,不同的关系类型对语义类别信息的颗粒度要求不同,不能一概而论。

综上所述,单独加入《词林》或《知网》语义信息时,各个大类抽取的性能都有大幅度的提高,且提高幅度相差不大,并且同时加入这两种语义信息时性能最好,这说明两者具有较大的相似性和一定的互补性。而对小类关系抽取而言,《词林》和《知网》两种语义信息还在部分小类上表现出不同的趋势,这进一步说明了两者具有互补性,因此只有两者结合起来才能更好地发挥语义信息在中文实体关系抽取中的作用。

3.2.2　动词语义对关系抽取性能的影响

表 4 和表 5 分别列出了加入关系实例中动词所对应的《词林》语义信息和《知网》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 S、Q、$F1$、ΔF 和 $\bar{F}$ 等各列的含义与表 2 相同,而“SCV-BL”、“SHNV-BL”和“(SCV+SHNV)-BL”分别表示在基准系统的基础上加入关系实例中的动词对应的《词林》词群语义类别、知网第一基本义原以及同时加入两者。表 4 中 ΔF 和 $\bar{F}$ 的最大值用下划线显示。由于加入动词语义后,大部分小类关系的性能没有变化甚至有一定程度的降低,因此表 5 中仅列出了加入动词语义后性能有一定提升的小类关系,且超过 1 的 ΔF 值用下划线显示。

从表 4 可以看出,在基准系统的基础上,分别加入动词对应的《词林》或《知网》语义信息时,大类抽取的总体性能只分别提高了 0.3 和 0.2,且两者的结合也不能再提高其性能了。不过,PHYS 大类关系的 $F1$ 值却有一定幅度的提升,这是由于其中的 Locate 小类性能提高的缘故,这一点可以从表 5 中体现出来。同时,从 Membership 和 UOIM 的 ΔF 值可以看出,《词林》和《知网》语义信息的结合对 UOIM 关系有益,而对 Membership 关系却有副作用。

除此之外,还可以看出,动词的两种语义信息对其他一部分关系小类抽取性能的影响程度基本相似,例如 Lasting-Personal、Membership 和 UOIM 等。其主要原因是由于动词的《词林》语义信息和《知网》语义信息的颗粒度基本相同,其所对应的语义泛化程度大体相似,例如关系实例“七里淀村 建起 了 青年民兵之家”中,动词“建起”所对应的《词林》词群编码为“Hc05A”,其对应的语义为“建立 设立 创立 命名”,其所对应的《知网》第一基本义原为“建造”,两者基本相似。

综上所述,加入动词语义信息虽然对关系抽取的总体性能影响甚微,但是对 PHYS 大类关系却有较大影响,对某些特定的小类关系也有不同程度的影响,并且两种语义信息对一些小类的影响程度基本相似。这说明当对某些特定的关系类型进行抽取时,考虑动词语义信息确实可以起到一定的作用。

表 4 动词在《知网》和《词林》中的语义信息对大类关系抽取的性能影响

关系大类	S	Q	BL	SCV-BL		SHNV-BL		(SCV + SHNV)-BL	
			$F1$	ΔF	$\bar{F}$	ΔF	$\bar{F}$	ΔF	$\bar{F}$
PHYS	1 552	17.0	16.2	1.4	0.24	0.8	0.14	1.9	0.33
PART-WHOLE	2 249	24.6	64.9	−0.1	−0.02	−0.1	−0.03	−0.2	−0.05
PER-SOC	652	7.2	44.9	0.4	0.03	0.5	0.04	0.7	0.05
ORG-AFF	2 166	23.7	69.3	0.0	0.00	0.0	0.00	0.3	0.06
ART	623	6.8	23.4	0.0	0.00	−0.4	−0.03	0.4	0.03
GEN-AFF	1 905	20.8	63.9	0.3	0.06	0.0	0.01	0.1	0.02
合计	9 147	100	56.0	0.3	0.30	0.2	0.20	0.3	0.30

表 5 动词在《知网》和《词林》中的语义信息对小类关系抽取的性能影响

关系大类	关系小类	S	Q	BL	SCV-BL		SHNV-BL		(SCV + SHNV)-BL	
				$F1$	ΔF	$\bar{F}$	ΔF	$\bar{F}$	ΔF	$\bar{F}$
PHYS	Located	1 335	14.5	13.0	1.3	0.18	0.2	0.03	1.2	0.18
	Near	217	2.4	35.6	−0.2	0.00	0.9	0.02	0.3	0.01
PER-SOC	Lasting-Personal	84	0.9	12.6	3.9	0.04	3.9	0.04	3.9	0.04
ORG-AFF	Membership	344	3.8	46.6	1.1	0.04	1.2	0.04	0.7	0.03
ART	UOIM	623	6.8	25.2	0.3	0.02	0.5	0.03	1.1	0.07
合计		9 147	100	52.7	0.2	0.20	0.3	0.30	0.3	0.30

3.2.3 已知实体类型下语义信息的影响

前面的实验分别比较和分析了实体所在词汇和动词词汇的语义信息对中文关系抽取的影响，另外实体类型也是一种实体语义信息，且对关系抽取具有很大的作用[17]，那么实体类型信息和实体语义信息之间存在着什么样的关系呢？下面，首先从总体性能上比较《词林》、《知网》语义信息和实体类型信息对中文关系抽取的性能影响，然后从具体关系类型上进行分析。

需要说明的是，由于考虑动词语义对总体性能的影响很小，并且它只对个别小类的抽取有帮助，而对有些小类甚至还有损害作用，因此下面的实验中没有考虑动词语义信息。

1）从总体性能上比较《词林》、《知网》语义信息与实体类型信息的相互影响。

表 6 比较了在基准系统的基础上，加入不同组合的《词林》、《知网》语义信息和实体类型信息（包括实体大类和小类信息）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下划线表示。

表 6 《词林》、《知网》语义信息和实体类型信息的性能比较

实体信息	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
基准 + 词林	76.4	50.2	60.6	74.8	48.1	58.6
基准 + 知网	76.9	51.0	61.3	73.5	48.2	58.2
基准 + 词林 + 知网	78.5	51.8	62.4	75.8	49.1	59.6
基准 + 实体类型	80.4	56.5	66.4	77.1	54.3	63.7
基准 + 实体类型 + 词林	81.8	56.5	66.8	79.8	54.6	64.8
基准 + 实体类型 + 知网	80.8	55.8	66.0	77.6	53.7	63.4
基准 + 实体类型 + 词林 + 知网	82.5	55.4	66.3	80.0	52.9	63.7

从表 6 可以看出，与在没有实体类型信息的基础上单独加入或同时加入两种语义信息均能提高关系抽取性能，在实体类型信息的基础上，无论是大类抽取还是小类抽取，都是加入《词林》语义信息时 $F1$ 值性能最高，尽管提高并不明显（大类、小类的 $F1$ 值分别提高 0.4、1.1），而在加入《知网》语义信息时，总体性能稍微降低（大类、小类的 $F1$ 值分别降低了 0.4、0.3），因此同时加入两者时，总体性能也无法再提高了。这说明从总体上看，实体类型信息和《词林》语义信息可以相互补充，但是和《知网》语义信息却有冲突。

2）从具体关系类型的性能上比较《词林》、《知网》语义信息和实体类型的相互影响。

表 7 和表 8 分别列出了在已知实体类型信息的基础上，加入《词林》和《知网》语义信息后各个大类和小类类别上的性能差异，分别用“（类型 + 词林）-类型”、“（类型 + 知网）-类型”表示，其中 S、Q、$F1$、ΔF 和 $\bar{F}$ 的含义与表 2 相同，而“类型-BL”指在基准系统的基础上加入实体类型信息后的性能变化。表 7 中的 ΔF 和 $\bar{F}$ 的最大值用下划线显示，而表 8 中在实体类型的基础上加入《词林》或《知网》语义信息时超过 1 的 ΔF 值用下划线显示。

从表 7 可以看出，在基准系统的基础上，单独加入实体类型信息时，性能提高非常显著。在实体类型的基础上加入《词林》语义信息时，有 3 个大类（PER-SOC、ORG-AFF 和 GEN-AFF）的 $F1$ 值提高 1 以上；而加入《知网》语义信息时，只有组织附属关系（ORG-AFF）大类的 $F1$ 值提高，且提高幅度只有 0.5。这说明在实体类型的基础上，《词林》语义信息能在一定程度上补充实体类型信息的不足，而《知网》语义信息的加入反而削弱了实体类型信息对关系抽取的提升作用。对表 8 的结果进行进一步的分析可以发现其原因，即在实体类型的基础上加入《词林》语义信息时，ΔF 值在 1 以上的小类有 12 个（如 Near、Subsidiary、Business 等），所占比例较高；而在加入《知网》语义信息时，ΔF 值在 1 以上的小类只有 3 个，分别为 Near、Ownership、Investor-Shareholder 关系，且这 3 个小类所占比例很小，大部分小类的 ΔF 值小于 0。

另外，从该表中还可以发现《词林》语义信息和《知网》语义信息之间的相似点和不同点，具体体现在：

1）相似点。有些小类关系，如 Near、Subsidiary、Ownership、Investor-Shareholder 等，在实体类型的基础上，无论加入《词林》语义还是《知网》语义，都一定程度地提高了抽取的性能。这说明对于这些类别而言，《词林》语义和《知网》语义都能在一定程度上补充实体类型信息的不足。例如“法放公社 党委”和“国际奥委会 下属 的 体育兴奋剂和生物化学小组”关系实例中第一个实体的实体类型虽然不同，分别为“Population-Center”、“Sports”，但其词汇的中心词所对应的

《词林》词群编码都为“Di09D”;又如关系实例“院校 音乐系”、“银行 总部”等的第一个实体的实体类型分别为“Educational”、“Commercial”,但其词汇对应的第一基本义原都为“场所”。这些关系实例均为 PART-WHOLE. Subsidiary 关系,但在只加入实体类型信息时却都识别不出来,而加入《词林》和《知网》语义信息后进一步增加了关系实例之间的相似性,从而提高了关系的识别率。有些小类关系,如 Located、Geographical、UOIM 等,无论加入《词林》语义还是《知网》语义,都降低了抽取的性能,尤其是 UOIM 关系。这说明对这些小类关系而言,实体词汇的语义信息对关系抽取并不起作用。

2)不同点。有些小类关系,如 Business、Family 和 Lasting-Personal 等,加入《词林》语义信息后其 ΔF 值有显著的提升,但是加入《知网》语义信息后提升却很小甚至有一定程度的降低。说明对于这些关系类型而言,《词林》语义和实体类型信息能很好地相互补充,而《知网》语义的影响很小甚至会损害实体类型信息的作用。经过分析,发现其原因与在未知实体类型信息下加入《词林》或《知网》语义信息时某些关系小类性能差别迥异的原因类似,都是由于《词林》对某些实体的词汇划分较细致,而《知网》则过于泛化而引起的。

表 7　《词林》和《知网》语义信息在已知实体类型下的大类抽取性能比较

关系大类	S	Q	类型	类型-BL		(类型 + 词林)-类型		(类型 + 知网)-类型	
			$F1$	ΔF	$\tilde{F}$	ΔF	$\tilde{F}$	ΔF	$\tilde{F}$
PHYS	1 552	17.0	30.2	14.0	2.38	−0.9	−0.15	−1.4	−0.24
PART-WHOLE	2 249	24.6	76.2	11.4	2.78	0.0	−0.01	−0.2	−0.04
PER-SOC	652	7.2	48.9	4.0	0.29	1.0	0.07	−0.6	−0.04
ORG-AFF	2 166	23.7	77.0	7.7	1.83	1.3	0.32	0.5	0.12
ART	623	6.8	45.1	21.8	1.48	−3.3	−0.22	−2.7	−0.18
GEN-AFF	1 905	20.8	75.2	11.3	2.34	1.7	0.36	−0.3	−0.07
合计	9 147	100	66.4	10.4	10.40	0.4	0.40	−0.4	−0.40

表 8　《词林》和《知网》语义在已知实体类型的情况下的小类抽取性能比较

关系大类	关系小类	S	Q	类型	类型-BL		(类型 + 词林)-类型		(类型 + 知网)-类型	
				$F1$	ΔF	$\tilde{F}$	ΔF	$\tilde{F}$	ΔF	$\tilde{F}$
PHYS	Located	1 335	14.5	27.9	14.9	2.16	−1.8	−0.27	−0.8	−0.12
	Near	217	2.4	40.5	4.9	0.12	1.2	0.03	1.4	0.03
PART-WHOLE	Geographical	1 257	13.7	73.1	14.8	2.03	−0.5	−0.07	−0.4	−0.05
	Subsidiary	978	10.7	80.6	15.2	1.62	1.2	0.13	0.9	0.09
	Artifact	14	0.2	0.0	0.0	0.00	0.0	0.00	0.0	0.00
PER-SOC	Business	186	2.0	35.7	0.9	0.02	11.4	0.24	−1.5	−0.03
	Family	382	4.2	47.2	−2.8	−0.12	2.2	0.09	−0.5	−0.02
	Lasting-Personal	84	0.9	2.2	−10.3	−0.10	10.7	0.10	−2.2	−0.02
ORG-AFF	Employment	1 560	17.0	77.7	6.5	1.10	1.7	0.29	−0.4	−0.06
	Ownership	22	0.3	14.7	−20.0	−0.05	1.3	0.00	6.7	0.02
	Founder	17	0.2	0.0	−34.0	−0.07	0.0	0.00	0.0	0.00
	Student-Alum	69	0.8	14.1	−1.4	−0.01	4.6	0.03	0.3	0.00
	Sports-Affiliation	69	0.8	35.4	9.6	0.07	5.4	0.04	0.3	0.00
	Investor-Shareholder	85	0.9	14.4	2.2	0.02	5.4	0.05	5.4	0.05
	Membership	344	3.8	57.3	10.6	0.40	2.6	0.10	0.2	0.01
ART	UOIM	623	6.8	46.1	20.9	1.42	−2.7	−0.18	−1.3	−0.09
GEN-AFF	CRRE	732	8.0	66.9	11.0	0.88	4.2	0.34	−0.5	−0.04
	Org-Location	1 173	12.8	80.4	15.7	2.01	0.0	0.00	−0.4	−0.05
	合计	9 147	100	63.7	11.0	11.00	1.1	1.10	−0.3	−0.30

综上所述,在已知实体类型的基础上,两种语义信息对各个关系类型的抽取性能影响不同,《词林》语义信息可以进一步提高大部分关系类型的抽取性能,而《知网》语义信息只能提高个别关系类型的抽取性能。

4　结语

本文通过在句法树中加入实体词汇和相关动词的语义信息的方法来对《同义词词林》和《知网》的语义信息对中文关系抽取的影响作了系统的比较和分析。通过实验发现,一方面,动词语义信息能提高少数关系类型的抽取性能,另一方面,对于实体词汇而言,无论是否加入实体类型信息,《同义词词林》的语义信息都能大幅度地提高大部分关系类型的抽取性能;而《知网》语义信息在未知实体类型信息时,能取得与《同义词词林》相似的效果,但在已知实体类型信息的前提下,则仅能提高个别关系类型的抽取性能。这说明《同义词词林》在一定程度上能补充实体类型的不足,而《知网》则和实体类型信息存在冲突。当然需要说明的是,本文仅使用了《知网》中的部分语义信息,即词汇的第一基本义原。今后的工作将采用《知网》来计算词汇相似度[20],从而更合理地探索它对关系抽取的影响。

参考文献:

[1] ZHOU GUODONG, SU JIAN, ZHANG JIE, *et al*. Exploring various knowledge in relation extraction [C]// ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2005: 427 – 434.

[2] CHAN Y S, ROTH D. Exploiting background knowledge for relation extraction [C]// COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 152 – 160.

[3] SUN A, GRISHMAN R, SEKINE S. Semi-supervised relation extraction with large-scale word clustering [C]//HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 521 - 529.

[4] ZHANG MIN, ZHANG JIE, SU JIAN, *et al*. A composite kernel to extract relations between entities with both flat and structured features [C]// ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006: 825 - 832.

[5] ZHOU GUODONG, ZHANG MIN, JI DONGHONG, *et al*. Tree kernel-based relation extraction with context-sensitive structured parse tree information [C]// EMNLP-CoNLL 2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsbing: Association for Computational Linguistics, 2007: 728 - 736.

[6] ZHOU GUODONG, ZHANG QIAOMING. Kernel-based semantic relation detection and classification via enriched parse tree structure[J]. Journal of Computer Science and Technology, 2011, 26(1): 45 - 56.

[7] QIAN LONGHUA, ZHOU GUODONG, ZHU QIAOMING. Employing constituent dependency information for tree kernel-based semantic relation extraction between named entities[J]. ACM Transaction on Asian Language Information Processing, 2011, 10(3): No. 15.

[8] CULOTTA A, SORENSEN J. Dependency tree kernels for relation extraction [C]// ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004: 423 - 429.

[9] BUNESCU R C, MOONEY R J. A shortest path dependency kernel for relation extraction [C]// HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2005: 724 - 731.

[10] NGUYEN T-V T, MOSCHITTI A, RICCARDI G. Convolution kernels on constituent, dependency and sequential structures for relation extraction [C]// EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009, 3: 1378 - 1387.

[11] 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.

[12] 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, 91.

[13] LI WENJIE, ZHANG PENG, WEI FURU, *et al*. A novel feature-based approach to Chinese entity relation extraction [C]// HLT-Short '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Stroudsburg: Association for Computational Linguistics, 2008: 89 - 92.

[14] CHE WANXIANG, JIANG JIANMIN, SU ZHONG, *et al*. Improved-edit-distance kernel for Chinese relation extraction [C]// IJCNLP-05: The Second International Joint Conference on Natural Language Processing. Jeju, Korea: [s. n.], 2005: 134 - 139.

[15] 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.

[16] 黄瑞红, 孙乐, 冯元勇, 等. 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.

[17] 虞欢欢, 钱龙华, 周国栋, 等. 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.

[18] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林[M]. 2 版. 上海: 上海辞书出版社, 1996.

[19] 董振东, 董强. KDML——知网知识系统描述语言[EB/OL]. (2010 - 08 - 20) [2011 - 12 - 05]. http://www.keenage.com/html/e_index.html.

[20] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算[C]// 第三届汉语词汇语义学研讨会. 台北: [出版者不祥], 2002: 59 - 76.

[21] CHARNIAK E. Eugene Charniak 个人主页[EB/OL]. [2012 - 01 - 30]. http://www.cs.brown.edu/~ec/.

[22] ICTCLAS. ICTCLAS 汉语分词系统[EB/OL]. [2012 - 01 - 30]. http://ictclas.org/.

[23] MOSCHITTI A. Alessandro Moschitti 个人主页[EB/OL]. [2012 - 01 - 30]. http://ai-nlp.info.uniroma2.it/moschitti/.

(上接第 2237 页)

表 1 各方法的期望及方差

降维方法	1 维		3 维	
	正确率期望/%	方差	正确率期望/%	方差
PCA	89.92	0.0005	94.03	0.0007
LDA	92.32	0.0108	92.00	0.0110
LPP	71.95	0.0036	91.89	0.0015
SLPP	90.67	0.0307	92.21	0.0112
KSSLPP	96.00	0.0004	96.00	0.0004

5 结语

本文从半监督角度出发,提出了保持全局和局部结构的核半监督数据降维方法,该方法充分考虑了传统的有监督和无监督算法的特点,有效利用了样本的所有特性。实验表明该算法的识别性能较以前的算法有了进一步的提升。KSSLPP 算法中,不同参数的选择,以及不同核函数的选择,会对算法的性能将产生什么样的影响,这些可以作为以后研究的内容。

参考文献:

[1] van der MAATEN L J P, POSTMA E O, van den HERIK H J. Dimension reduction: A comparative review, TiCC-TR 2009-005 [R]. Tilburg: Tilburg University, 2009.

[2] DUDA R O, HART P E, STORK D G. Pattern classification [M]. 2nd ed. New York: John Wiley & Sons, 2001. 170.

[3] JOLLIFFE I T. Principal component analysis [M]. 2nd ed. New York: Springer, 1986.

[4] FISHER R A. The use of multiple measurements in taxonomic problems [J]. Annals of Eugenics, 1936, 7(2): 179 - 188.

[5] ZHANG DAOQIANG, ZHOU ZHI-HUA, CHEN SONGCAN. Semi-supervised dimensionality reduction [C]// Proceedings of the 7th SIAM International Conference on Data Mining. Cambridge: MIT Press, 2007: 629 - 634.

[6] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000.

[7] 高绪伟. 核 PCA 特征提取方法及其应用研究[D]. 南京: 南京航空航天大学, 2009.

[8] HE XIAOFEI, NIYOGI P. Locality preserving projections [C]// Proceedings of 17th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2003: 585 - 591

[9] 申中华, 潘永惠, 王士同. 有监督的局部保留投影降维算法[J]. 模式识别与人工智能, 2008, 21(2): 233 - 239.

[10] SCHÖLKOPF B, SMOLA A, MÜLLER K-R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5): 1299 - 1319.

[11] VAPNIK V N. Statistical learning theory [M]. New York: Wiley, 1998.

[12] 孙即祥, 姚伟, 腾书华. 模式识别[M]. 北京: 国防工业出版社, 2009.