第33卷 第10期 计 算 机 工 程 2007年5月
Vol.33 No.10 Computer Engineering May 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)10—0212—03 文献标识码：A 中图分类号：TP391.4

用语义模式提取实体关系的方法

邓 擘，樊孝忠，杨立公
(北京理工大学计算机科学与技术学院，北京 100081)

摘 要：研究了信息抽取中的汉语实体关系提取技术，在使用模式匹配技术的基础上引入了词汇语义匹配技术对汉语实体关系进行提取。比较了一般模式匹配技术和词汇语义模式匹配技术在汉语实体关系提取任务中的性能。实验结果表明，一般模式匹配技术在处理中文时效果较差，而词汇语义模式匹配技术更适合于处理汉语实体关系提取任务。

关键词：信息抽取；实体关系；模式匹配；词汇语义

Entity Relation Extraction Method Using Semantic Pattern

DENG Bo, FAN Xiaozhong, YANG Ligong
(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081)

【Abstract】 The paper studies entity relation extraction of Chinese in information extraction. To extract Chinese entity relation, it imports technology of word semantic match into pattern match technology. Performance between general pattern match technology and semantic pattern match technology is compared. Experiment shows that general pattern match technology can't get acceptable result in task of Chinese, but new method presented by this paper is more adapted to deal with task of Chinese entity relation extraction.

【Key words】 Information extraction; Entity relation; Pattern match; Word semantics

关系提取是信息抽取的子任务，主要目的是提取句子中的实体关系。目前有多种刻画句子或实体特征的方法[1]，这些特征被广泛应用于各种信息提取系统中。其中比较著名的有DIPRE、Snowball[2]、FASTUS 和SIFT等信息提取系统。进行关系提取经常用到模式匹配技术，上述3个系统均使用了模式匹配技术。尽管这些系统构建模式的方式各有差异，但构建模式的一般过程都是根据实体类型及上下文等一些句子特征来形成半固定的、具有一定结构的数据元组。

国外也有学者做过使用语义的方法进行关系与事件抽取的尝试。Chinatsu 使用本体的概念构造了一个关系与事件抽取系统 REES。Birte 研究了法语模式下的关系提取问题，发现了关系提取中的模式与本体状态间的依赖关系，并且认为仅仅依靠单个的本体来使用提取模式是不够的。

在基于模式匹配方式进行关系提取的信息抽取系统中，一般模式总是与每个给定句子的实体前后的词进行精确的词语及句法格式的匹配，有的相应地对匹配结果给出评估分数。从已有的研究成果来看，使用模式匹配方式工作的信息抽取系统对英文的效果比较好。但是汉语在构词、语法、语义及时态等诸多方面与英语有很大的区别。鉴于这种区别，本文在模式匹配技术中引入汉语的词汇语义的概念来对汉语句子的实体关系进行提取。作为对比，本文也使用了一般模式匹配技术来提取汉语句子的实体关系。

1 汉语实体关系提取

1.1 词汇语义的相似度计算

本文主要使用《同义词词林》[3]这部语义词典来作为计算词汇语义相似度的工具。《同义词词林》是一部比较详尽的同义词语义知识词典，它按照树状的层次结构把所有收录的词条组织在一起，把词汇分成了大、中、小3类。每个小类里都有很多词，这些词又根据词义的远近和相关性分成了若干个词群，每个词群中的词语又进一步分成了若干个行，同一行的词语或者词义相同，或者词义有很强的相关性。整个词典的层次结构如图1所示。

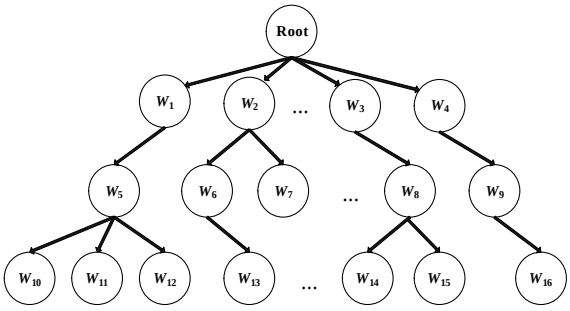

图1 词典的层次结构

整个层次结构共分为5层。根据这种树状结构，词汇语义相似度被转换为对词语间语义距离的度量。两个词语的距离越大，其相似度越低；反之，两个词语距离越小，其相似程度越大。定义词语间距离为0时，其相似度为1；词语间距离为无穷大时，其相似度为0，相似度即成为词语距离的单调递减函数。

如果把两个词语 w_1 和 w_2 的相似度记为 $Sim_{lexical}(w_1,w_2)$，两词语间距离记为 $Dis(w_1,w_2)$，那么使用《同义词词林》语义知识词典，两词语间的距离可以通过计算词汇在树状结构中相应节点间的连通距离来得到，并由此可以进一步计算两词语的语义相似度。本文使用文献[3]中提出的词汇语义相似度计算公式如下。

作者简介：邓 擘(1976－)，男，博士生，主研方向：信息检索，自然语言理解；樊孝忠，教授；杨立公，博士生

收稿日期：2006-05-23 **E-mail：**dengbo_999@bit.edu.cn

$$Sim_{lexical}(w_1,w_2)=\frac{\alpha\times(l_1+l_2)}{(Dis(w_1,w_2)+\alpha)\times Max(|l_1-l_2|,1)} \quad (1)$$

其中 l_1、l_2 分别是词语 w_1、w_2 所处的层次，α是相似度为 0.5 时 w_1、 w_2 之间的距离，α是一个可调节的参数，一般α> 0。这样就得到了计算任意两个词语的语义相似度的办法。

1. 2 关系抽取模式的产生

Snowball 系统采用了种子学习的方法来半人工地产生提取模式，本文在此也使用该方法来获取关系提取模式。对于用模式提取的种子句，主要标明如下 3 类信息，首先把种子句分词并标明各词词性；其次标明句子中所出现的两类实体的类型及其位置。因为本文主要处理人与组织机构之间的关系提取任务，所以在句子中只标明这两种实体类型。最后在句子末尾标明句中出现的实体的关系类型。以下是一个句子标注例子：

[王明/Per son] 在/p [微软公司/Org] 辛辛苦苦/z 工作/v 了/u 五/m 年/q 。[R: General-Staff]

本文使用的关系模式提取方法虽然也是从种子句中出现在实体前后的上下文来学习到关系模式，但是在从种子句中学习提取模式时，并不是机械地把出现在种子句中实体前、实体中、实体后的所有词语均作为所学模式的特征词保留下来，而只是保留了一些主要的词。因为通过对汉语句子的观察，发现汉语句子中两个实体的上下文中常会出现大量的修饰性词语，如形容词、副词、语气词等。如果把句中所有出现的词语均作为提取模式的特征词而保留下来，一方面极大增加了模式的长度，会在以后的模式匹配工作中占据较多的机器时间，另一方面长模式的匹配准确率也会有相应的下降。所以仅保留对关系识别最有判断价值的核心词语，如名词、动词等，而把一般修饰性词语，即对判断实体关系无太大价值的词滤去。这样就获得了从种子中提取模式的一个简单办法。模式的一般结构如下：

$$<Left, ET_1, Middle, ET_2, Right>$$

其中 *Left* 、 *Middle* 、 *Right* 分别是出现在种子句中实体前、实体中和实体后的所有重要词语，即不对模式匹配窗口的大小作限制。而 ET_1 和 ET_2 是种子句中两个实体的实体类型。该模式的一般向量形式为

$$t_p=(a_1,a_2,...,a_r,ET_1,b_1,b_2,...,b_s,ET_2,c_1,c_2,...,c_t)$$

其中 a 、 b 、 c 分别表示实体前、实体中、实体后的词语。若这些位置没有词语，则可以为空。以上句为例，则可产生提取模式如下：

t_p^1 = (Per，在，Org，辛辛苦苦，工作，年)

1. 3 实体关系的确定

1.3.1 模式匹配过程

在对测试集句子进行模式匹配前，除了对句子分词、标注词性以及标明句子中出现的各实体类型外，也需要过滤掉句中出现的修饰性词语。具体的模式匹配过程如下：

(1)对测试集句子过滤掉修饰性词语。

(2)提取测试句子的实体前、实体中和实体后的重要词语以及实体类型，组成待比较向量。

$$t_s=(a_1',a_2',...,a_u',ET_1',b_1',b_2',...,b_v',ET_2',c_1',c_2',...,c_w')$$

(3)比较模式中实体类型与待比较向量中实体类型是否一致，若类型一致则进行下一步。

(4)比较模式向量中词语与待比较向量中的每个词语的语义相似度，获得一张模式词语与待比较向量词语的词汇语义相似度表。具体比较过程见 1.3.2 节。

(5)在得到模式与待比较向量的词汇语义相似度表的基础上，计算二者的匹配相似度。具体计算过程见 1.3.3 节。

(6)将待比较向量与所有模式匹配的相似度按大小排序，选择相似度最大的模式的实体关系类型作为测试句子的实体关系类型。

对所有测试句子按上述 6 个步骤处理后，即可判定测试句的实体关系类型。

1.3.2 词汇语义相似度表的计算。

按式(1)可分别计算模式中每个词语与待比较向量中每个词语的语义相似度，并从中挑出相似度最大的待比较向量词语与当前模式词语构成一个三元记录，即 $(w_i,w_{i1},Sim_{\max})$，其中 w_i 是模式中的词语，w_{i1} 是待比较向量中的词语，$Sim_{\max}$ 为相似度最大值。所有模式中词语的三元记录构成一张模式词语与待比较向量中词语的语义相似度表。这样做可以避免不同位置的近义词由于没有得到比较机会而造成的匹配不上的问题。假设有如下的测试句子：

小李现在在淀粉厂上班。

经分词和实体标注后得到待比较向量如下：

t_s^1 = (Per，现在，在，Org，上班)

则按照式(1)可得上述模式 t_p^1 与此待比较向量 t_s^1 的词汇语义相似度，如表 1 所示。

表 1 模式与比较向量的词汇语义相似度

t_p^1	在	辛辛苦苦	工作	年
t_s^1	在		上班	现在
$Sim_{\max}$	1	0	$\frac{8\times\alpha}{2+\alpha}$	$\frac{4\times\alpha}{2+\alpha}$

因为“辛辛苦苦”一词与待比较向量中其它任一词均不在同一类，所以它与其它词语的距离为无穷大，即相似度为 0。

1.3.3 模式匹配相似度的计算

根据前述得到的模式与待比较向量的词汇语义相似度表，在计算模式与待比较向量的匹配相似度时主要考虑 3 点：首先是两匹配词的语义相似度；其次是两个比较词语各自在模式和待比较向量中的相对位置。若两词语均在各自相对应的位置上，比如在实体之间，则可以适当提高匹配程度。最后是两个比较词语的词性。若两词词性相同，如均为动词，则也可考虑适当提高匹配程度。这样就增大了在相同位置的词和同性词的匹配相似性。根据上述 3 个因素定义模式与待比较向量的匹配相似度计算公式如下：

$$Match(t_p,t_s)=\begin{cases}\sum\alpha(w_i,w_{ij})\beta(w_i,w_{ij})Sim(w_i,w_{ij}) & ET_1=ET_1',ET_2=ET_2'\\ 0 & ET_1\neq ET_1',ET_2\neq ET_2'\end{cases} \quad (2)$$

其中 w_i 、 w_{ij} 是在词汇语义相似度表中同一个三元记录中的两个词， w_i 是模式中的词语， w_{ij} 是待比较向量中的词语。

$\alpha(w_i,w_{ij})$ 是两词语相对位置度量函数，形式定义如下：

$$\alpha(w_i,w_{ij})=\begin{cases}1 & w_i.position=w_{ij}.position\\ 0.8 & w_i.position\neq w_{ij}.position\end{cases} \quad (3)$$

$\beta(w_i,w_{ij})$ 是两词语词性度量函数，形式如下：

$$\beta(w_i,w_{ij})=\begin{cases}1 & w_i.POS=w_{ij}.POS\\ 0.8 & w_i.POS\neq w_{ij}.POS\end{cases} \quad (4)$$

由式(2)可知，当被比较的两个词语分别在模式和待比较向量的同一位置，是相同的词，并具有相同词性时，则词汇语义匹配方式就转化为词语的精确匹配方式，即等价于一般的模式匹配方式。且此公式的模式匹配相似度是可以大于 1 的。由前述得到的词汇语义相似度表可以计算模式 t_p^1 与待比

较向量 t_S^1 的匹配相似度为

$$Match(t_P^1,t_S^1)=1+0+\frac{8\times\alpha}{2+\alpha}+0.8\times0.8\times\frac{4\times\alpha}{2+\alpha}=1+\frac{10.56\times\alpha}{2+\alpha}$$

2 试验与分析

试验中选择了人名与组织机构的关系作为提取对象，使用了人民日报的已分好词的语料库(该语料可以从网址 www.icl.pku.edu 获得)。从中选择含有人名与组织机构名的句子以及自己收集的一些句子共 600 个句子组成了试验集合。在对所有的句子进行分词、实体标注与实体关系人工识别并分类后，选取其中有代表性的 100 个句子作为关系模式提取的种子集，余下的句子作为测试集。在词汇语义相似度计算方面，使用了《同义词词林》作为语义词典。对本文中提出的模式匹配方法与待比较向量进行了比较并确定了它的关系类别。为了比较这一方法在汉语实体关系提取任务中的效果好坏，使用 Snowball 系统中的模式匹配方法作为一般模式匹配方式，在相同的试验集合上做了对比试验，并用准确率、召回率和 F 值来作为比较两种方法效果的好坏。

在 ACE 的 RDC 任务中把实体关系分为若干个层次的类别，在试验中主要处理了角色这个大类别下的 5 个子类别的识别工作。将测试集中的 500 个句子人工分为 5 个子类别，各子类别测试集的数量见表 2。

表 2 子关系类的句子数量

子类别	Owner	Client	Member	General-staff	Management
数量	87	106	102	113	92

2.1 试验结果

表 3 是使用 Snowball 系统中的模式匹配方式与本文提出的使用词汇语义模式匹配方式在含有 500 个汉语测试句子集上的试验结果。

表 3 实体关系识别结果

模式匹配方式	P_{avg}	R_{avg}	F_{avg}
一般模式匹配	0.276	0.325	0.298
词汇语义模式匹配	0.582	0.614	0.597

2.2 试验分析

从试验结果可以看出，用英文中一般常用的模式匹配方式去匹配中文句子时，匹配的准确率和召回率是很低的。经分析匹配过程，发现很多语句由于对应位置的词语与匹配模式中相应位置的词语并不相同而导致只有很低的相似度，甚至即使在模式与测试句子中都有相同的词语出现时，由于词语出现在二者的不同位置而匹配不上，这些是造成一般模式匹配方式在汉语句子中效果不好的主要原因；此外，由于汉语句子在结构上比英文灵活，表意词汇更为丰富，使得这种机械性的模式匹配方式在应用于中文时困难很大。而使用词汇语义的模式匹配方式时，则由于模式中的词语可以与测试句中出现在任意位置的任一词语的语义进行相似度比较，而不是对词语的形式进行比较，因此这种方法对句子的表达结构、词语的表现形式以及词语在测试句子中的相对位置并不敏感，从而在模式与测试句的匹配过程中常因为二者出现了相同或语义相近的词语而获得较高的相似度，也因此才有较高的准确率与召回率。但该方法也存在一些缺陷，比如当属于两个不同的子关系类别中的测试句出现较多相同或近义词时，常会造成该方法的误判，这一现象在 Member 和 General-staff、Owner 和 Management 这两对子类别的判别中出现得较多。

3 结论

普通模式匹配技术提取实体关系时，均是用模式与句子及其词语进行格式或形式上的匹配。这种匹配方式一般要求句子的行文、结构比较规范。然而对于中文，由于其结构一般比较灵活，表意方式多样，因此这种匹配技术对中文效率并不高。为提高效率，本文提出了使用词汇语义的模式匹配方法来提取中文句子中的实体关系。从试验结果来看，这一方法能够较好地符合中文的实际情况，一定程度上提高了中文实体关系的提取效果。下一步的工作是通过逐步建立并完善汉语概念本体库，用汉语词汇本体来代替同义词词典，以及改进语义模式匹配等方法来进一步提高汉语句子中实体关系的判别准确性。

参考文献

1 Zhou Guodong, Su Jian, Zhang Jie. Exploring Various Knowledge in Relation Extraction[C]//Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005: 427-434.

2 Agichtein E, Gravano L. Snowball: Extracting Relations from Large Plain-text Collections[C]//Proceedings of the 5th ACM International Conference on Digital Libraries. 2000: 85-94.

3 梅家驹, 高蕴奇, 笠一鸣, 等. 同义词词林[M]. 上海: 上海辞书出版社, 1983.

4 吴　健, 吴朝晖, 李　莹, 等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报, 2005, 28(4): 595-602.

(上接第 211 页)

5 结论

图 5 面部图片 3

图 6 面部图片 4

图 7 面部图片 4

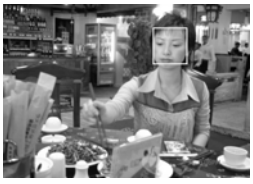

图 8 面部图片 5

通过对 35 幅图像的实验可知，只要图片光照适中，面部是正面或半侧面的图像均能够做出正确定位。利用本文算法定位面部的部分图片见图 5~图 8。本实验算法全部在 Matlab 6.5 环境下编译实现。

参考文献

1 张兆礼, 赵春晖, 梅晓丹. 现代图像处理技术及 Matlab 实现[M]. 北京: 人民邮电出版社, 2001.

2 张宏林. Visual C++数字图像模式识别技术及工程实践[M]. 北京: 人民邮电出版社, 2003.

3 崔　屹. 数字图像处理技术与应用[M]. 北京: 电子工业出版社, 1997.

4 周　杰. 人脸自动识别方法综述[J]. 电子学报, 2000, 28(4):102-106.