汉语实体关系模式的自动获取研究

邓 擘[1] 郑彦宁[1] 傅继彬[2]
(中国科学技术信息研究所 北京 100038)[1] (北京理工大学计算机系 北京 100081)[2]

摘 要 中文信息抽取系统中实体关系模式的自动获取对于整个系统具有重要意义。在bootstrap方法的基础上,根据汉语在形式表达上的多样性特点,使用统计学习技术来自动获取新模式。实验表明,该方法在人工干预很少的情况下,能够快速查找新模式,且新模式的获取不受应用领域限制。因此该方法对于提高信息抽取系统的性能是有效的。

关键词 信息抽取,实体关系,模式匹配,相似度
中图法分类号 TP312 **文献标识码** A

Study of Obtaining Chinese Entity Relation Pattern Automatically

DENG Bo[1] ZHENG Yan-ning[1] FU Ji-bin[2]
(Institute of Scientific and Technical Information of China, Beijing 100038, China)[1]
(Department of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)[2]

Abstract Obtaining Chinese entity relation pattern automatically is very important for entire information extraction system. Based on method of bootstrap and features that Chinese can express same meaning by many forms, using technology of statistical learning to obtain new pattern automatically. Experiment shows the method can find new pattern rapidly and need very small manual work, the process can't be limited by extract region. So the method is effective for promoting the function of information extraction system.

Keywords Information extraction, Entity relation, Pattern match, Similarity

1 引言

实体关系抽取的目的是发现并判断句子中出现的不同实体对之间的语义关系。实体关系的识别有多种方式,比较常见的有基于特征的方法与基于核函数的方法。其中基于核函数的关系抽取技术近年来获得了大量深入的研究,该技术用于判断实体关系的特征不再仅限于有限的数量,而是可以扩展到大量的甚至无限的特征[1],这种方法也使得关系抽取的准确率与召回率有较大程度提高[2,3]。但是该方法也有其不足之处,由于核函数的确定没有一定的规律可循,在使用该方法抽取关系前必须要预先定义好核函数[4]。由于不同的核函数在处理不同关系类型的句子时优劣各不相同,因此目前没有一个核函数在识别实体关系时具有完全的优势。

模式的方法本质上是一种基于特征匹配的方法,而且由于模式的方法一般来说简单直观,使用的特征有限,不需要做复杂计算,因此其工作效率往往较高。但由于关系抽取中模式的表现形式多样,变化较大,因此难以用人工的方法预先全部定义。所以长期以来,在基于模式的方法中,模式的自动获取技术就成为研究的关键。在Agichtein的文献[5]中,使用基于bootstrap方法的种子来逐步得到新模式,但该方法的缺陷在于模式的获取并不能完全自动化,还需要人工来判定其准确性。后来Xu则使用语义种子的方法来学习关系抽取模式,以及使用种子驱动的机器学习算法来抽取具有不同复杂度的实体关系[6,7]。文献[8]在哈瑞分布式假设的基础上,使用推理技术来得到潜在的、易于被忽视的模式。文献[9]为了能够更容易地进行系统移植,完全没有再使用预先定义模式的方法,而是以句子中出现的名词性实体对间的短语为线索来发现术语定义。文献[10]通过Wordnet技术对模式的相似度进行对比,以此发现新的模式。文献[11]使用语义方法来发现新模式并比较它们的相似性,最后根据比较结果归纳出好的模式。

本文根据bootstrap方法的种子思想,构造了汉语实体关系模式自动获取学习算法。它将模式作为种子,通过循环处理来不断产生新模式;第2节介绍了关系模式与句子的相似度计算方法;第3节介绍了实体关系模式的自动抽取方法;第4节给出了实验结果,并对其做了分析;最后是结论。

2 实体关系模式的相似度计算

2.1 词语串的相似度比较

在比较两个关系模式的相似度时,需要计算两个模式中涉及的词语的语义相似度。本文使用《同义词词林》[12]来计算。在计算词语相似度时,任意两个词语 w_1,w_2 的相似程度由其在树中的距离间接反应。设两个词语的相似度为 $S(w_1,w_2)$,距离为 $Dist(w_1,w_2)$,则使用式(1)来计算两个词语的相似度。

$$S(w_1,w_2)=\frac{1}{Dist(w_1,w_2)+1} \quad (1)$$

到稿日期:2009-03-06 返修日期:2009-06-01 本文受国家自然科学基金项目(项目编号:70803048)资助。
邓 擘(1976—),男,博士后,主要研究方向为自然语言处理,E-mail:dengbo-999@163.com。

显然，词语相似度值分布在0～1区间内。

在词语比较语义相似度的基础上，可以进一步计算两个词语串的语义相似度。词语串是指由一列词语构成的一个字符串，其中最小的语义单位是词语。例如“敬爱的/j 领袖/n”就是一个词语串。

对于两个不同的词语串，要比较它们的相似度，可采用如下方法。首先将两个词语串中的词按词性分类，如把所有名词性词语归为一类，动词归为另一类等。然后在比较时，把不同词语串中同词性的词相互进行比较。如将某一词语串中的所有名词取出，分别与另一个串中的所有名词比较，计算它们的语义相似度，以找到两个词语串中彼此最相似的名词。找到这两个词后，将这两个词分别从两个名词集中滤除，同时记录二者的相似度，然后查找下一对最相似的名词，直到某个词集中所有名词都被滤除。最后将刚才计算所得的词语相似度值相加，这样就得到了两个词语串所有名词的相似度值。同理处理动词等词类集合，最后把所有不同词类集合的相似度值相加就得到了两个词语串的综合相似度值。设两个词串分别为 $c=(c_1,c_2,\cdots,c_n)$，$d=(d_1,d_2,\cdots,d_m)$，任意两个词语的相似度计算函数为 $S(c_i,d_j)$，则两个词语串 c，d 的总相似度可以表示为：

$$\mathrm{Sim}(c,d)=\sum_{\delta_k}\sum_{i,j\in\delta_k}\mathrm{Max}(S(c_i,d_j)) \tag{2}$$

其中，δ_k 表示某个词类集合。在比较词串的相似度时，某些词性的词语的相似度比其他词语更重要，例如动词，此时则可以通过其相似度计算函数 $S(c_i,d_j)$ 乘以一个大于1的权值来增加其权重。

2.2 实体关系模式与句子的相似度计算

在大量语言实例中，可以观察到有许多从内容到结构都非常类似的句子，这些句子表达的意思也常常具有共同之处，因而这些句子实际上可能表示的是相同或类似的关系，所不同的仅仅是句子中出现的具体实体可能不一样。如果把这些句子中出现的实体去掉，可以看到这些句子在使用的词语和表达结构上具有很多相似之处。因此若能够把这些句子所共同具有的词语和表达结构抽取出来，那么这种表达结构就成为一种模式，它具有一定的普遍性。其中用于确定实体间关系的句子表示结构就称为实体关系模式。实体关系模式由一些固定词语构成，具有一定的结构，是一种表示实体间某种关系的常见表达方式。在此采用与文献[5]中类似的方式将实体关系模式定义为一个五元组，其形式为 $(L,T1,M,T2,R)$，其中 L,M,R 分别表示处于模式左、中、右位置的3个词语串，$T1,T2$ 表示两个命名实体的类型标记。一个具体的实体关系模式如：(〈经验，丰富，的〉，人名，〈被〉，组织名，〈聘任，为，工程师〉)。这个模式可以匹配这样的句子，如“有经验的张三最近被微软公司聘任为技术工程师”。模式中不同位置的词语串可以没有，但不能都没有。

实体关系模式与句子的语义相似度计算主要通过比较二者的命名实体类型，以及模式左、中、右词语串中的词语与句子相应位置的词语的语义相似度来确定。

首先模式与句子中出现的实体对类型要一致，即模式与句子在对应位置的实体对($T1,T2$)的类型要完全相同。如果一个模式的 $T1$ 类型是人名，而句子在相应位置的实体类型是地名，则认为二者的表达结构是不同的，因此它们也视为不具有比较基础。

在上述实体类型一致的前提条件下，就可以分别计算模式的左、中、右词语串与句子在相应位置的词语串的语义相似度，计算方法前文已述。值得注意的是，文献[13]中通过统计认为，判定实体关系的主要特征取决于句子中出现的中心动词。这一研究结果说明，在模式中出现的动词对判定句子所表示的实体关系具有重要的特征表示作用，因此在计算模式与句子中的动词相似度时，可以给它们赋以较大的权重。这样，通过分别计算模式与句子的词语串相似度，就可得到二者的整体相似度。模式与句子的相似度计算公式可表示为：

$$\mathrm{Sim}(P_1,P_2)=\mathrm{Sim}(L,L')+\mathrm{Sim}(M,M')+\mathrm{Sim}(R,R') \tag{3}$$

其中，L'，M'，R' 分别表示句子与模式对应位置的词语串。使用语义方式进行模式与句子的匹配，其优点是当二者在形式上不完全一致时，仍然可以保证对它们进行相似度比较。

3 实体关系模式的自动获取

在中文中相同的意思通常可以用多种不同的形式表达，这就是句子表达形式的多样性。而在互联网高度发达的今天，大量同一事件可能会被多个新闻网页以各种不同的方式描述。如果能够收集所有描述同一事件的网页，并将这些网页中对同一事件描述的相关句子集中起来，就可以发现它们虽然表达方式各有不同，但表达的意思基本一致，因此这些事件中出现的实体对之间的关系也应该是相同的。根据这一想法，如果能将这些表达相同关系的不同表达结构抽取出来，那么它们就成为表达特定关系的一些模式。而根据那些能与这些模式匹配的句子，也可以推断其中所包含的实体应该具有与模式相同的关系。

在Snowball信息抽取系统[5]中，关系模式通过实体对与关系模式的相互循环作用来得到。由于要以实体对为种子，因此在循环过程中需要对抽出实体对的正确性进行人工判断，这就限制了系统的扩展能力，使之获取实体关系模式的数量难以提高，而且也使得该方式的应用领域受到限制，即系统的移植能力不高。

基于上述考虑，本文提出了基于bootstrap方法的实体关系模式的自动获取技术，该方法主要分为以下4个步骤：

1)人工构造少量实体关系模式作为初始种子集，使用这些种子模式与训练集中的句子匹配，找出适当的多个实体对，构成一个列表，同时保存出现这些实体对的句子。

2)对上述列表中的每一对实体，通过检索技术找到含有该实体对的所有句子，构成特征句子集。

3)使用种子模式逐一与特征句子集中的句子匹配，以查找与第1)步中保存的句子具有相同实体关系、不同表达结构的句子，将这些句子构成一个实体关系句子集。

4)对实体关系句子集进行提取关系模式的处理，可以自动找到多个与种子模式有相同实体关系的新模式。

使用以上方法得到的新模式，其实体关系是相同的，但语言结构可能并不相同。如果在一次这样的学习抽取过程中产生的模式数量不够多，则还可以将新模式再加入到种子模式集，反复循环这一过程直到产生足够多的关系模式。该方法的过程如图1所示。

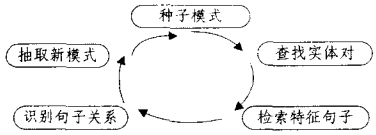

图 1　实体关系模式获取图

在上述各步骤中使用到的模式与句子匹配相似度计算方法，前文已述。

4　实验结果与分析

使用 Java 语言开发了关系抽取试验系统的代码，并选择职务变动和出生事件两个领域来抽取实体关系模式，主要处理具有雇佣关系和出生关系的句子。由于没有 ACE 的测试数据集，因此从网上手工收集了这两个领域的一些句子，在经过手工标注后用于实验。首先人工构造了种子模式集、训练集和测试集。种子模式集中有预先人工构造的 6 个关系模式，其中 4 个模式表示人名与组织机构间的雇佣关系，其余两个模式表示人名与地名间的出生关系。训练集则分别由两个领域的句子构成，属于职务变动领域的有 60 个句子，属于出生领域的有 78 个句子。而测试集则用于对新抽取的关系模式进行标注测试，其在两个领域各使用了 600 个句子，其中各领域的句子中有一半不具有雇佣关系或出生关系。

在实验中，6 个初始种子模式经过匹配处理后，共产生了 26 个新模式，属于职务变动领域的有 14 个，属于出生领域的有 12 个。表 1 列出了部分种子模式与提取出的新模式。

表 1　种子模式与新模式列表

序号	L	T1	M	T2	R
1		nr	任/v	ns	书记/n
2		nr	出生/v 于/p	ns	
3	经过/p 大会/n 选举	nr	当选/v 为/p	ns	省委/n 书记/n ./w
4	1991 年/t 9 月/t	nr	出任/v 中共/j	ns	省委/n 秘书长/n ./w
5		nr	在/p	ns	诞生/v ./w
6		nr	的/出生地 /n 是/v	ns	

表中前两个是种子模式，后面 4 个是提取出的新模式。经人工考察，新模式中共有 4 个不正确，两个领域内各有两个，剔除后用余下的模式分别对两个领域的测试集进行关系标注处理，这样得到的该次实验的准确率 P、召回率 R 以及 F 值如表 2 所列。

表 2　实验结果

领域	P	R	F
职务变动	73.03%	50.67%	59.83%
出生事件	70.3%	34%	45.83%

从新模式的内容来看，新模式与种子模式有很大相似程度，但在具体表示词语上并不完全相同。例如用于辅助判定实体关系的中心动词，在种子模式中仅有两种，而在新模式中却出现了多种与其语义相关的动词，这说明新模式大大扩展了旧模式的表达范围，并且这些新模式所表达的实体关系基本也都是正确的。但同时也应看到，抽取出的新模式并不都是正确的，在两个领域内的新模式中都出现了错误。经考察发现，造成错误的主要原因在于词语相似度计算不精确，使得在抽取模式时记分不准，超过阈值而误判为新模式。

从新模式对测试集的标注结果来看，准确率还是比较高的，但召回率尚不够。这主要有两方面原因：一个是新模式还不够多，对于测试集中出现的不同句子表达结构并不能保证完全覆盖，而仅能识别那些常见或常用的句型；另一个是新模式在内容上比较粗糙，模式中含有不少噪声词，而测试句子中也同样含有大量未在模式中出现的噪声词，这样当模式与测试句匹配时，就可能造成二者的匹配相似度较低，它低于接受阈值时就会使匹配失败，最后导致一些句子被漏选。这一问题主要是由模式提取方式造成的，当用于归纳提取的句子较少时，为了保留尽可能多的识别特征而不得不保留一些词语，这就可能引入一些噪声词语。

结束语　本文以 bootstrap 方法为基础来自动获取实体关系模式。该方法的优势在于使用人工干预较少，仅需要在初始状态时加入少量种子模式，并构造简单的训练集，此后新模式的判断与产生完全使用统计学习来完成，大大减少了人工工作，对于系统的移植有很大便利；而当模式在使用中数量不足时，也可以通过该方法随时扩充模式集，这就进一步提高了模式集的覆盖性。通过这种方法产生的模式，彼此间具有很强的相关性，但表示结构又不相同，因此对具有不同表达结构，但意思相近或相同的句子具有很好的适应性。下一步的研究是如何对含有噪声词的模式进行高度优化，使其在匹配时准确性更高，以及如何对模式的可信度进行评估，以进一步优选高质量的模式。

参 考 文 献

[1] Huang Ruihong EE, Sun Le, Feng Yuanyong. Study of Kernel-Based Methods for Chinese Relation Extraction[J]. AIRS, 2008: 598-604

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

[3] Zhou G D, Zhang M, Ji DH, et al. Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information[C]// ACL-2007. Prague, 2007: 728-736

[4] Zhao Shubin, Grishman R. Extracting Relations with Integrated Information Using Kernel Methods[C]// ACL-2005. Ann Arbor

[5] Agichtein E, Gravano L. Snowball : extracting relations from large plain-text collections[C]// ACM DL 2000. 2000: 85-94

[6] Xu , Uszkoreit F H. *Minimally supervised learning of relation extraction rules using semantic seeds*[R]. A seminar talk at the National Center for Text Mining

[7] Xu, Uszkoreit F H, Li H. A seed-driven bottom-up machine learning framework for extracting relations of various complexity[C]// ACL-2007. Prague, 2007: 584-591

[8] Lin D, Pantel P. Discovery of inference rules for question answering[J]. Natural Language Engineering, 2001, 7: 343-360

[9] Sekine S. Automatic paraphrase discovery based on context and keywords between NE pairs[C]// IWP2005. South Korea, 2005: 80-87

[10] Stevenson M, Greenwood M. Learning Information Extraction Patterns Using WordNet [C] // GWNC-2006. South Korea, 2006: 52-60

[11] Stevenson M, Greenwood M. A Semantic Approach to IE Pattern Induction[C]// ACL-2007. Prague, 2007: 379-386

[12] 梅家驹，高蕴奇，竺一鸣，等. 同义词词林[M]. 上海：上海辞书出版社，1983

[13] Zhou Guodong, Su Jian. Exploring Various Knowledge in Relation Extraction[C]// ACL-2005. Ann Arbor, 2005: 427-434