高技术通讯 2010 年 第 20 卷 第 2 期:127 ~ 132

doi:10.3772/j.issn.1002-0470.2010.02.004

基于加权子序列核函数的次范畴论元分析①

朱聪慧② 赵铁军 韩习武* 郑德权

(哈尔滨工业大学教育部-微软语言语音重点实验室 哈尔滨 150001)

(*黑龙江大学计算机科学与技术学院 哈尔滨 150080)

摘 要 为提高汉语动词次范畴化框架(SCFs)的分析性能,提出了一种新的次范畴论元分析方法。该方法引入了基于间隙加权子序列的核函数,以传统规则的右部作为分类类别,将规则左部作为问题输入空间,将原本规则推导的问题转化为机器学习问题。由于间隙加权子序列核函数可以考虑跨距离的词之间的依赖关系,加之机器学习方法的引入,使得论元识别精度从 55.16% 提到了 93.43%,并且极大提高了次范畴整句获取精度。

关键词 汉语动词次范畴(SCF),论元分析,主动学习,间隔加权子序列

0 引 言

近年来词汇知识库的重要性无论在计算语言学界还是在理论语言学研究中都日渐增长,动词的次范畴化(subcategorization)信息是国内外公认的相关知识库不可缺少的组成部分,动词次范畴的正确分析,可以大幅提高搭配抽取、名实体识别、句法分析以及机器翻译等其他自然语言应用的性能。次范畴化,又称次语类化,指某一句法范畴的进一步划分[1]。动词次范畴化是通过动词作谓词时所表现出的不同句法特征的分布对动词大语类的一种细化,这些句法特征也因此被称为次范畴化框架(subcategorization frame,SCF)[2],动词次范畴化信息的分析也就体现为 SCFs 及其分布的获取。传统的分析方法大都是以手写的语言学论元(argument)规则作为启发式信息,经过规则的匹配和推导,在含有标准句法信息的数据上进行的。汉语动词 SCFs 是一个跨越多个句法层次的结构,它既包含了最低层次的部分词法信息,也包含了含有结果特征的句法信息。本文对汉语动词次范畴划分方法进行了研究,提出了基于间隙加权子序列核函数的汉语动词次范畴论元分析方法,获得了高精度的论元识别性能。

1 相关研究

自从 20 世纪 90 年代初文献[3]首次提出英语动词 SCF 信息的自动获取问题以来,世界上包括英语在内的很多语种在大规模真实语料上的次范畴化自动获取研究中都取得了很大程度的进展,同时取得了有一定应用价值的结果,例如德语、捷克语、西班牙语、葡萄牙语、希腊语[4,5]。特别是在英语大规模获取方面,文献[6]利用扩展规则的方式取得了非常优异的性能。而在汉语动词 SCFs 获取方面,韩习武做了大量工作,首先他给出了汉语 SCF 的形式化定义[7],并在基于假设生成、假设检验模式上做了大量的汉语 SCF 分析工作[8-11],取得了很好的结果,并且建立了相关的汉语 SCF 语料库。大多数次范畴化自动分析方法的目标都是在给定的语料中获取谓语动词的 SCF 类型和数目。处理过程有两个典型步骤:首先,根据启发式信息无指导地生成 SCF 假设,然后,进行统计过滤,选择可靠假设。

随着研究方法的深入,动词 SCFs 的分析性能也有了大幅提升,但是目前的方法还存在几个不足。首先,这些方法必须以手写确定的论元推导规则作为启发式信息,在含有完全正确句法信息的语料上才能进行 SCF 分析。其次,因为手写规则的不完全覆盖性和不一致性,导致输入的句子只有一部分可以应用这些规则,得到最终结果。同时约有 20% 的输入句子因为没有与之符合的手写规则或者规则之间存在冲突致使其无法被分析而直接被丢弃了,造成了 SCF 分析的低召回率[11]。

为了弥补以上不足,必须解决两个关键性问题:

① 国家自然科学基金(60773069,60973169)资助项目。

② 男,1979 年生,博士生;研究方向:机器翻译、机器学习;联系人,E-mail: Conghui@hit.edu.cn

(收稿日期:2009-02-03)

一是怎样减少分析过程中对语言学知识的依赖,避免手工书写规则,二是怎样提高论元规则集的覆盖率和一致性。这里主要是为了解决规则集推导结果的不一致问题。本文不再使用传统方法中采取的简单的规则匹配和推导(满足规则左部,就应用相应的推导规则),而是将原本规则推导的问题转化为机器学习问题。应用间隙加权子序列核函数,以拥有相同右部的规则(同时指向同一论元类别的规则)的左部作为训练集,设计一个论元类别的多元分类器,在特征空间内,给出当前输入对应的每一种论元类别的概率。

实验结果表明,本文提出的论元分析算法可以在几乎不需要任何先验语言学知识的情况下自动抽取出不同论元对应的推导规则,而且配合基于统计方法的间隙加权子序列核函数方法,可以大幅提高这些自动抽取的规则的一致性和覆盖性,获得高精度的论元识别性能(与传统基于规则推导的方法相比,论元识别精度从55.16%提到了93.43%),以此结果为基础的,在含有句法噪音的数据上,整句SCFs分析结果的召回率也从83.83%提高到99.49%。

2 汉语SCF分析问题的定义和转化

韩习武首先给出了汉语SCF的形式化定义[10],汉语动词SCF的形式化定义是汉语SCF分析的基础,在他的定义中,共有137个SCF类别,12个论元类别。每一个汉语句子都只对应唯一的一个SCF类别。一个输入句子相应的SCF是由该句的核心动词和一系列论元的序列组成,例如表1中的输入句子对应的SCF就是"看:NP BAP v BP",其中动词是"v",此外还有"NP"、"BAP"和"BP"三个论元。表1的下面几行就是相应的论元和对应的句子成分,这里的"NP"、"BAP"和"BP"就是汉语SCF中的变量型论元,也可以简称为论元,类似的论元共有12个,详情见表2。

汉语SCF分析以含有句法信息的汉语句子作为输入,分析结果是每个输入句子相应的SCF。那么整句的SCF分析可以看成一个"多对一"的映射,$f:X\rightarrow Y$,其中输入集X是带有句法分析结果的句子,而结果Y是汉语SCF的集合,并且$\|Y\|=137$。也就是说也可以把问题看成对一个输入句子的多元分类问题,只不过需要判断的分类数是137个。

表1 一个汉语句子及其SCF

类别	相应成分
输入句子	保姆/nc VBA[把/p 小孩/nc VC [看/vg BVP[丢/vg 了/ut]]]。/wj
相应SCF	看:NP BAP v BP
论元NP	保姆/nc
论元BAP	VBA[把/p 小孩/nc
论元V	看/vg
论元BP	丢/vg 了/ut

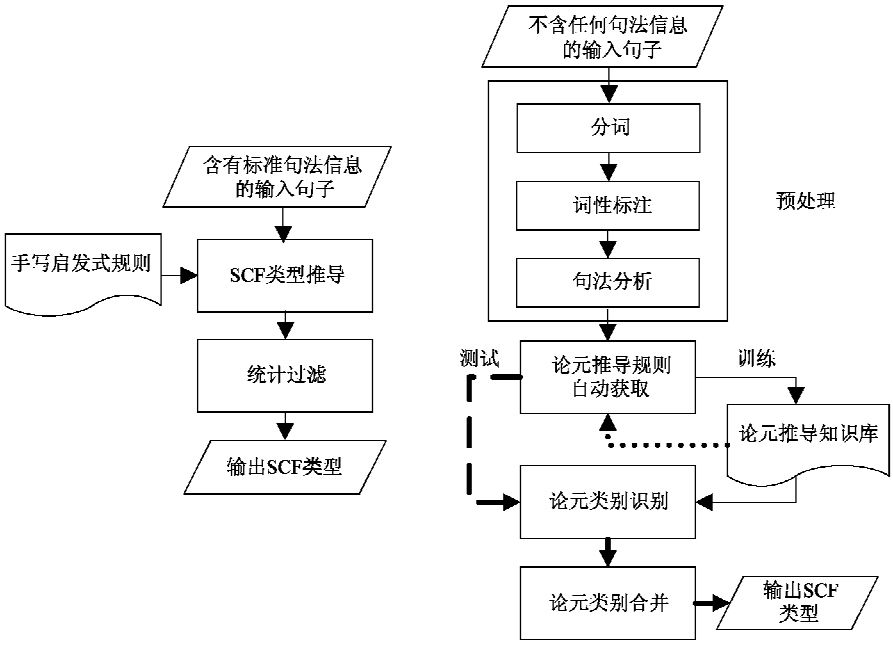
表2 汉语动词SCF的变量型论元成分

论元	定义
NP	名词性论元成分
VP	谓语或谓语中心词以外的一般动词引导的论元
QP	趋向动词构成的论元成分
BP	结果动词构成的论元成分
PP	介词或方位词短语论元成分
BAP	介词"把"引导的论元成分
BIP	介词"被"、"让"、"由"等引导的表被动的论元
TP	表时间段的论元成分
MP	独立数量词短语论元成分
JP	形容词、副词或结构助词"得"引导的修饰性论元
S	内含小句型论元成分
NP	名词性论元成分

而如果以SCF中的论元作为分析单元,可以把问题转化为一个新的"多对一"映射$g:T\rightarrow A$,其中输入T是每个论元对应的相关句子成分,也可以说是Token串(见表1);而A是论元的集合(见表2),这里的$\|A\|$只有12。通过映射g,我们把原问题即对整句SCF类别的判断,转化成为对句子每个子部分相应论元的判断。在我们的数据中,原问题的每个SCF类可使用的平均句子数为451.81句,而问题转化后,每个论元对应的Tokens的数目为17366,多元分类的每个类别可用的平均样本数目较原问题增加了近34.5倍,为分类器的性能提升提供了必要的样本支持,同时把需要判定的类别数从137减少到了12,降低了多元分类器设计的复杂度。

传统的SCF获取过程必须以含有完全正确句法信息的句子作为输入,以人工事先给定的论元推导规则作为启发式信息进行SCF类别的推导,而且推导之后必须用统计过滤的方法除去一些因为规则集的低覆盖率和不一致性造成错误分析的句子,这就造成传统获取方法召回率较低(图1(a))。而我们的方法可以直接以真实应用中的汉语句子作为输入,经过预处理(汉语分词、词性标注和句法分析),

不需要任何人工校正。我们的论元推导规则自动获取算法可以容忍预处理结果中含有的句法错误,自动发现新的论元推导规则。我们的方法也不同于传统方法直接应用这些规则进行推导,而是以具有相同右部的规则的左部作为训练数据,相同的右部作为分类类别,应用间隙加权子序列核函数的多元支持向量机(support vector machine,SVM)分类方法,从统计的角度在特征空间内给出当前输入对应的每一种论元类别的概率。将一个输入句子所有的论元类别结果合并在一起,可以得到此输入句子对应的SCF类别(见图1(b))。



(a)传统SCF分析方法　　(b) 自动获取方法

图1　本文方法与传统SCF获取方法的比较

3　基于间隙加权子序核函数SVM多元分类

根据统计学习理论,学习模型的实际风险由经验风险值和置信范围值两部分组成。而由Vapnik提出的支持向量机(SVM),以训练错误最小为约束条件,以置信风险最小为优化目标,所以SVM具有很好的推广能力。

近年来,基于结构的核函数被广泛应用在自然语言处理方面[12-14]。在序列核中,核函数利用的结构化信息是两个样本字符串的公共子序列数。两个字符串的公共子序列越多,就认为两个字符串越相近。在基本的子序列核之上,我们进行了改进,引入了间隙加权子序列核(weighted gap subsequence kernel),该核函数基于一个简单的原理,即对某个公共子序列,其在原串中间隔越短,说明该序列的结构越紧密,因此也可以期望这种方式在计算两个字符串的距离时有着更重要的作用。由于间隙加权子序列核函数可以考虑更多跨距离或者远距离的依赖关系,在应用中往往具有更好的效果。

我们假设$\sum$为一个有穷汉语单词集,$s = s_1 s_2 \cdots s_{|s|}$(i.e.$s_i \in \sum$,$1 \leqslant i \leqslant |S|$)是$\sum$上的词串,而$I = [i_1, i_2, \cdots, i_n]$,$1 \leqslant i_1 < i_2 < \cdots < i_n \leqslant |S|$是一个偏序集合,表示子串的下标,例如一个长度为$n$的子串可以表示为$s_{i_1} s_{i_2}, \cdots, s_{i_n}$,需要明确指出的是这里的子串不需要一定是连续的。例如"今天 天气 非常 不错"是一个单词串s,下标集i为[2,4],那么子串$s[i]$是"天气不错"。我们定义$l(i)$为单词串s的长度,$l(i) = i_n - i_1 + 1$,那么两个子序列s和t的间隔加权核函数可以表示为

$$K_n(s,t) = \sum_{u \in \Sigma^n} \sum_{i:s[i]=u} \sum_{j:s[j]=u} \lambda^{l(i)+l(j)} \qquad (1)$$

其中,$\lambda \in [0,1]$是对非连续子序列的惩罚因子,第一个求和公式表示对所有长度为n子序列进行求和。公式(1)在特征空间内定义了一个有效的内积计算方法。下面简要介绍一下阶的概念。在子序列

核中,核的阶数指计算距离时使用的子序列的长度[15]。p 阶核指对两个单词串取所有长度为 p 的公共子序列进行计算得到的核函数。对一个单词符串,p 阶核对应的特征空间向量可表示为

$$\Phi_u^p(s) = \sum \lambda^{l(i)}, \quad u \in \sum p \tag{2}$$

那么子序列核也就是 p 阶的公式(1)又可以表示为

$$K_p(s,t) = < \Phi^p(s), \Phi^p(t) > = \sum_u \Phi_u^p(s) \cdot \Phi_u^p(t) \tag{3}$$

而使用公式(3)后,将输入空间转化为特征空间,那么目标函数可以表示为

$$h(x) = \mathrm{sign}(\sum_{i=1}^{l} y_i \alpha_i K_p(x, x_i) + b) \tag{4}$$

为了使间隙加权子序列核函数对不同长度词串的计算结果都具有可比性,我们采用公式

$$\hat{K}(s,t) = \frac{K(s,t)}{\sqrt{K(s,s)K(t,t)}} \tag{5}$$

对核函数进行归一化。

由间隙加权子序列核定义的特征空间具有 $|\sum|^p$ 的维度,直接在特征空间中计算具有极高的时间复杂度和空间复杂度。但核函数的特性之一就是在样本空间而非特征空间进行特征空间的内积计算,序列核则是使用一种递归的方法进行内积的计算。使用巧妙的动态规划方法可以将时间复杂度降低到 $O(p \cdot |t| \cdot |s|)$。

4 实验

4.1 数据

为了将汉语动词 SCF 应用于实践,我们从实际应用中收集了 61898 个不带有任何信息的汉语句子,并人工为每个句子标注相应的动词 SCF 类别。而后通过哈工大机器智能与翻译研究室的分词词性标注系统和完全句法分析系统,为每个句子添加了句法结构信息。由于目前汉语完全句法分析在开放测试中整体性能偏低,加之分词和词性标注阶段的错误累计,造成这些句子被标注的句法信息含有的错误噪音非常多。而我们的论元推导规则自动获取方法只对用以切分句子的核心动词的句法信息相对敏感,而对其他类型的噪音错误,如果语料充分,则可以克服,所以我们更关注句子核心动词部分的噪音情况。我们收集的所有句子经过上述的处理后,有 6.31%的句子在对包含核心动词短语的处理上有噪音错误发生,以句子为单位的噪音分布细节见表 3。其中核心动词在分词阶段就出现错误的句子占 2.77%;核心动词在词性标注阶段出现错误的句子占 5.01%;根据预处理阶段标注的句法信息,在识别核心动词时,出现错误的句子有 1.77%;而包含核心动词的短语的类别出现错误的占 4.03%。由此可见噪音情况还是很严重的。

表 3 核心动词相关错误分布

错误类型	噪音分布(%)
核心动词切分错误	2.77
核心动词词性标注错误	5.01
非核心动词误识别错误	1.77
核心动词短语类型标注错误	4.03

4.2 论元类别分析结果

将所有搜集的 61898 个汉语句子按照 9:1 的比例分成训练集和测试集。我们的论元推导规则自动获取方法平均获取 41536 条规则(去重复后)。其中大多数的映射规则集中在“NP”、“BAP”、“QP NP”和“PP”4 个论元类型,它们占到了整个获取规则数量的 80%以上。根据这些自动学习得到的论元推导规则,我们共进行了 3 组论元类别识别实验:

(A) 依据传统方法直接应用获取的规则(噪音数据上自动获取的)进行论元类别的推导;

(B) 依据我们的方法,但是不使用任何核函数的 SVM 模式识别方法;

(C) 应用基于间隙加权子序核函数的 SVM 模式识别方法。

10-交叉验证的论元分析实验结果见表 4,通过比较实验结果和实际计算量,实验 C 采用的阶数 p 最终被确定为 3。

表 4 SCF 自动分析结果

	实验 A	实验 B	实验 C
论元类别识别精度(%)	55.16	82.07	93.43

由于自动生成的论元推导规则的左部遗留了相当多的句法噪音错误,而且单纯使用这些规则进行匹配时,规则集的一致性和覆盖率都不尽人意,所以实验 A 的论元识别精度比较差;而以统计的角度,重新审视这些规则集,将单纯的规则匹配推导转化为以同一类的所有规则为训练数据,将原规则推导问题转化为模式识别问题后,同样的规则集,由于使用的方法的改变,大幅提高了论元识别的性能。采

用了基于SVM模式识别方法的实验B较单纯使用规则匹配推导的实验A性能约提升了48.79%,这说明我们提出的使用规则的新方式,可以较好地解决规则集原本存在的低覆盖率和不一致性问题。而加入基于间隙加权子序核函数后,实验C的论元识别的精度,又提升了约13.84%,这说明我们提出的基于间隙加权子序核函数,由于可以使用跨距离的依赖关系,因而更加适合这种不规则并且含有噪音错误单词串的相似度比较。

我们又将基于间隙加权子序核函数的论元识别结果应用于整句SCF类别的分析。实验结果见表5。我们将文献[10]和文献[16]中提到的结果作为对比实验的基准。文献[10]的结果是目前基于规则推导的SCF获取的最好性能,而文献[16]是目前以整句为特征基于机器学习方法的SCF自动分析的最好结果。其中我们的方法是直接以句法分析的结果为输入,并且没有采用任何额外的语言学知识,而以往工作采用的输入含有的句法信息是经过人工校验过的,并且需要人工预先给定的规则作为启发式信息。这3种方法采用的是完全相同的数据集。

表5 SCF自动分析结果

实验	精确率(%)	召回率(%)	F测度(%)
本文方法	87.64	99.49	93.19
以往工作1[10]	76.94	83.83	80.24
以往工作2[16]	71.42	71.24	69.83

从结果中,可以看到本文方法的召回率比较高,这是因为我们的论元推导规则自动获取算法自动地在训练数据上抽取了大量的规则,规则数量远远大于人工预先给定的规则集,保证了规则集的覆盖率。而且通过转化对这些规则的使用方法,以有相同推导类别的一类规则作为训练数据,从统计的角度在特征空间内给出当前输入究竟和哪一类规则描述的论元类别更加相似,大幅提高了规则集的一致性。也就是说,传统的方法只是判断当前输入可以应用哪些规则,而我们的方法实质上是判断当前输入与哪一类规则描述的论元类别更相似,较规则推导的方式,我们的方法极大地提高了规则推导结果的一致性。

5 结 论

我们使用机器学习的方法来扩展传统的论元规则推导机制。将规则的右部作为分类类别,将规则左部作为问题输入空间,通过间隙加权子序列核函数,根据特征空间内规则之间的相似度作为判断采用哪类规则的依据。由于问题的转换以及子序列核函数的特性,使得汉语动词次范畴论元分析结果得到了大幅提升,也提高了动词次范畴的整体分析性能,为将次范畴应用于实际提出了新的解决方法。

参考文献

[1] Chomsky N. Aspects of the Theory of Syntax: [Ph. D dissertation]. Cambridge: MIT Press, 1965. 16-20

[2] Korhonen A. Subcategorization Acquisition: [Ph. D dissertation]. Cambridge: Trinity Hall University of Cambridge, 2001. 29-77

[3] Brent M R. Automatic acquisition of subcategorization frames from untagged text. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA, 1991. 209-214

[4] Chrupala G. Acquiring Verb Subcategorization from Spanish Corpora: [Ph. D dissertation]. Barcelona: Universitat de Barcelona, 2003. 5-71

[5] Gamallo P, Agustini A, Gabriel P L. Using cocomposition for acquiring syntactic and semantic subcategorization —— Unsupervised lexical acquisition. In: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, USA, 2002. 34-41

[6] Preiss J, Briscoe T, Korhonen A. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007. 912-919

[7] 韩习武. 汉语动词次范畴化自动获取技术: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2006. 3-60

[8] Han X W, Zhao T J, Yang M Y. FML-based SCF predefinition learning for Chinese verbs. In: Proceedings of the International Joint Conference of NLP, Hainan, China, 2004. 115-122

[9] Han X W, Zhao T J, Qi H, et al. Subcategorization acquisition and evaluation for Chinese verbs. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 2004. 723-728

[10] Han X W, Zhao T J. Two-fold filtering for Chinese subcategorization acquisition with diathesis alternations as heuristic information. *Computational Linguistics and Chinese Language Processing*, 2006, 11(2): 101-114

[11] 韩习武. 从真实语料中自动获取汉语动词次范畴化信息. 计算机工程与应用, 2005, 19:1-5

[12] Zhang M, Che W X, Aw A, et al. A grammar-driven convolution tree kernel for semantic role classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007. 200-207

[13] Moschitti A, Quarteroni S, Basili R, et al. Exploiting syntactic and shallow semantic kernels for question/answer classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007. 776-783

[14] Diab M, Moschitti A, Pighin D. Semantic role labeling systems for arabic using kernel methods. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Ohio, USA, 2008. 798-806

[15] Cancedda N, Gaussier E, Goutte C, et al. Word-sequence kernels. *Journal of Machine Learning Research*, 2003, 3: 1059-1082

[16] 王俊. 面向真实语料的汉语动词次范畴化自动获取的研究: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2006:46-48

Arguments analysis of Chinese verb subcategorization based on weighted gap subsequence kernel function

Zhu Conghui, Zhao Tiejun, Han Xiwu*, Zheng Dequan

(Ministry of Education-Microsoft Key Laboratory of NLP and Speech, Harbin Institute of Technology, Harbin 150001)

(* School of Computer Science and Technology, Heilongjiang University, Harbin 150080)

Abstract

The paper proposes a new arguments analysis method for Chinese verb subcategorization to improve the present performance of analyzing Chinese verb subcategorization frames (SCFs). The method introduces the weighted gap subsequence kernel function into the analysis, and treats the left part and the right part of the traditional rules as training samples and related categories respectively, transforming the originally rule-derived problem into the machine-learning problem. This new kernel can take more cross-distance features. Compared with the rule based methods, the weighted gap subsequence kernel based method improves the precision of argument type analysis from 55.16% to 93.43% on syntactic noisy data. The analyzing performance of whole sentence is also much improved.

Key words: Chinese verb subcategorization frame (SCF), argument analysis, active learning strategies, weighted gap subsequence kernel