ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.6, June 2010, pp.1287−1295
doi: 10.3724/SP.J.1001.2010.03591

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于特征选择和最大熵模型的汉语词义消歧*

何径舟[1,2], 王厚峰[1,2+]

[1](北京大学 信息科学技术学院 计算语言学研究所,北京 100871)

[2](北京大学 计算语言学教育部重点实验室,北京 100871)

Chinese Word Sense Disambiguation Based on Maximum Entropy Model with Feature Selection

HE Jing-Zhou[1,2], WANG Hou-Feng[1,2+]

[1](Institute of Computational Linguistics, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China)

[2](Key Laboratory of Computational Linguistics (Ministry of Education), Peking University, Beijing 100871, China)

+ Corresponding author: E-mail: wanghf@pku.edu.cn

He JZ, Wang HF. Chinese word sense disambiguation based on maximum entropy model with feature selection. *Journal of Software*, 2010,21(6):1287−1295. http://www.jos.org.cn/1000-9825/3591.htm

Abstract: Word sense disambiguation (WSD) can be thought as a classification problem. Feature selection is of great importance in such a task. In general, features are selected manually, which requires a deep understanding of the task itself and the employed classification model. In this paper, the effect of feature template on Chinese WSD is studied, and an automatic feature selection algorithm based on maximum entropy model (MEM) is proposed, including uniform feature template selection for all ambiguous words and customized feature template selection for each word. Experimental result shows that automatic feature selection can reduce feature size and improve Chinese WSD performance. Compared with the best evaluation results of SemEval 2007: task #5, this method gets MicroAve (micro-average accuracy) increase 3.10% and MacroAve (macro-average accuracy) 2.96% respectively.

Key words: maximum entropy model; classification feature; automatic feature selection; Chinese word sense disambiguation

摘 要: 词义消歧是自然语言处理中一类典型的分类问题.在分类中,特征的选择至关重要.通常情况下,特征是由人工选择的,这就要求特征选取者对于待分类的问题本身和分类模型的特点有深刻的认识.分析了汉语词义消岐中特征模板对消歧结果的影响,在此基础上提出一套基于最大熵分类模型的自动特征选择方法,包括针对所有歧义词的统一特征模板选择和针对单个歧义词的独立特征模板优化算法.实验结果表明,使用自动选择的特征,不仅简化了特征模板,而且提高了汉语词义消歧的性能.与 SemEval 2007:task #5 的最好成绩相比,该方法分别在微平均值 MicroAve(micro-average accuracy)和宏平均值 MacroAve(macro-average accuracy)上提升了 3.10%和 2.96%.

* Supported by the National Natural Science Foundation of China under Grant Nos.60675035, 60973053, 90920011 (国家自然科学基金); the Beijing Municipal Natural Science Foundation of China under Grant No.4072012 (北京市自然科学基金)

Received 2008-10-10; Revised 2009-01-20; Accepted 2009-02-24

关键词: 最大熵模型;分类特征;自动特征选择;汉语词义消歧
中图法分类号: TP391 文献标识码: A

词义消歧(word sense disambiguation)是自然语言处理的重要研究内容,对机器翻译、信息检索和文本摘要等诸多应用有着直接影响.早在 20 世纪 60 年代,语义障碍(semantic barrier)就成为引发机器翻译危机的导因之一.当时,Bar-Hillel 给出的著名例子“John was looking for his toy box. Finally he found it. The box was in the **pen**. John was very happy”就是关于“pen”的词义消歧问题[1].

多义词在自然语言中非常普遍.在特定的上下文中,多义词的词义是明确的.词义消歧就是根据特定的上下文,确定多义词明确词义的过程.上面提到的英语单词 pen 作为名词具有两个词义,其一表示“笔”,其二表示“围栏”.虽然作为前一个词义的使用频度非常高,但在例子所给的上下文中,其词义恰巧是后者.汉语中也有大量的多义词,如“中医”,可以表示“医生”,也可以表示“医学”.

近些年来,将机器学习方法用于词义消歧的研究受到了广泛关注.比较典型的研究有 Pedersen 的多 Naïve Bayes 分类器集成的词义消歧[2].在 SemEval-2007 Task5 国际词义消歧评测中,参赛系统几乎都用到了学习方法,包括有指导最大熵分类模型 MEM(maximum entropy model)[3]和 Naïve Bayes 分类方法[4],半指导的带标传播(label propagation)方法以及无指导的方法[5].与其他应用问题一样,全指导分类和半指导分类方法取得的词义消歧效果更为明显[3−9].在汉语方面,近来也十分关注机器学习方法,例如,全昌勤等人研究了多分类器集成的词义消歧[10];吴云芳等人在多分类器集成方面做了更细致的工作[11];刘风成等人则用到了 AdaBoost.MH 分类方法[12].

分类是机器学习中的一类重要问题.分类的基本思想是先训练分类函数 f,然后将待分类的对象以特定的表示 x 作为输入,并通过分类函数计算相应的输出值 y 作为分类结果,即 $y=f(x)$.如果将多义词的每个词义作为一个类,那么,根据上下文确定具体词义的过程实际上就是分类.

分类需要考虑两个基本问题:

(1) 如何有效地描述分类对象,即如何通过输入 x 来表示待分类对象的特征;

(2) 如何选择合适的分类模型,并通过训练得到分类函数 f.

目前,已经出现了很多有用的分类方法,包括生成式分类方法(如 Naïve Bayes)和判别式分类方法(如 MEM 和 SVM);从分类器的构成看,有单个分类器的分类和多个分类器集成的分类(如 Boosting);从训练过程中使用的数据是否带有标记来看,又分为全指导、半指导(如 EM,Self-Training 和 Co-Training)和无指导方法.无论是哪种分类方法,都共同面临着输入 x 的表示问题.一般而言,输入 x 以特征向量形式表示.向量中的每一特征代表了分类数据的某个属性或属性组合.

为了有效分类,需要精细地挑选特征.首先,需要选择具有区分性好的特征.只有使用了有区分性的特征描述对象,才可能有效地区分不同对象;其次,在保证区分性的情况下,需要控制特征空间的维度.较低的维度有助于减少数据的稀疏和降低计算复杂度.

特征选择是机器学习的一项重要研究内容,在不同的应用中已有相应的讨论[13,14].但大部分的研究主要集中在降维,如主成分分析(principal component analysis,简称 PCA)和隐性语义标引(latent semantic index,简称 LSI).降维的主要目的是减小特征空间,降低时空复杂性.然而,降维在减少复杂度的同时,也很可能会降低可区分性.

在针对词义消歧的分类处理中,人们通常以待消歧的目标词为中心,在一定的上下文窗口内先定义特征模板,然后从实际的训练数据中获取特征,形成特征向量.模板的构成成分一般是窗口内的词、词性、词的 n-gram,词与词性的组合以及词性的 n-gram 等.大量的实验表明,窗口大小和模板类型的变化将导致分类结果的变化.因此,对分类问题而言,构造有效的特征模板是非常重要的环节.在构造特征模板的过程中,人们需要有一定的语言学知识,并进行大量的语料观察,甚至通过多次实验才能确定模板的构成形式,这无疑是繁琐的.特别是对不同多义词而言,影响词义的窗口大小和模板构成可能是不一样的.如果完全通过人工对每个多义词分别构造

合理模板,工作量将是巨大的.

为此,本文基于最大熵分类模型[15],研究了汉语词义消歧中特征模板的自动选择方法.实验结果表明,使用自动选择的特征模板,可以得到更好的词义消歧效果.

本文第 1 节介绍最大熵方法以及不同窗口和不同特征模板类型对汉语词义消歧的的影响.第 2 节介绍了特征选择的基本原则和实现算法.第 3 节进行实验评测,并比较不同特征选择方法对汉语词义消解性能的改进.最后是结论.

1 窗口大小和特征模板项对结果的影响

1.1 最大熵方法

最大熵方法的基本思想是建立与已知事实(训练数据)一致的模型,对未知因素不作任何假设——尽可能地保持均匀分布.最大熵方法的最大优点是能方便地引入有用特征.

给定一组样本集合 $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$,其中 x_i 表示特征向量,y_i 表示相应的分类结果,$1\le i\le n$.最大熵模型以指数形式计算条件概率:

$$p(y\mid x)=Z_\lambda(x)\exp\left(\sum_i \lambda_i f_i(x,y)\right) \tag{1}$$

其中, $Z_\lambda(x)=1\Big/\sum_y \exp\left(\sum_i \lambda_i(x,y)\right)$.

式(1)中,f_i 为特征;λ_i 表征了特征 f_i 的重要性,称为参数.$Z_\lambda(x)$为归一化因子.参数的估计方法包括 GIS,IIS 和 LBFGS 等.

建立和使用最大熵模型的过程包括:先定义特征模板,然后获取训练样例集并根据特征模板从训练集中抽取特征集,最后训练模型(估计模型参数)以及利用模型完成分类任务.本实验使用张乐开发的最大熵工具包(http://homepages.inf.ed.ac.uk/s0450736/maxent toolkit.html).参数估计选用 LBFGS,迭代次数为 200 次.

1.2 特征对结果的影响

本文实验数据来源于 SemEval2007:(Task #5)(http://nlp.cs.swarthmore.edu/semeval/tasks/index.php).表 1 给出了数据的说明.

Table 1 Evaluation data on SemEval2007 Task #5

表**1** SemEval2007 Task #5 评测数据

	#Average senses	#Training instances	#Test instances
19 nouns	2.45	1 019	364
21verbs	3.57	1 667	571

在词义消歧中,特征模板的选择与窗口大小有关.表 2 给出了基于窗口大小的特征模板.

Table 2 Window feature template

表**2** 窗口特征模板

Feature type	Feature template	Description
Independent word and pos	$W_i(-L\le i\le L,i\ne 0)$	Words in the window
	$P_i(-K\le i\le K,i\ne 0)$	Part of speech (pos) for each word in the window
Combination	$W_iW_{i+1}(-L\le i\le L-1)$	2-gram words in the window
	$P_iP_{i+1}(-K\le i\le K-1)$	2-gram pos in the window
	$W_iP_i(-L\le i\le L,i\ne 0)$	Pair of word and its pos in the window

在表 2 中,W 表示词,P 表示词性,K 表示在目标词左(右)边的词性个数,L 则表示目标词左(右)边的词个数.由于词和词性的组合有相同下标,因此,W_iP_i 有时也会简记为 WP_i.以“中医”为例:

例 1:现任/v[中国/ns 中医/n 研究院/n 长城/nz 医院/n]nt 院长/n 的/u 周/nr 文志/nr 教授/n.

在(3,2)窗口模板中,将得出以下特征:$W-2$=现任,$W-1$=中国,$W1$=研究院,$W2$=长城,$P-3$=*NULL*(表示

空),*P*–2=v,*P*–1=ns,*P*1=n,*P*2=nz,*P*3=*n*,*W*–2*W*–1=现任+中国,*W*–1*W*0=中国+中医,*W*0*W*1=中医+研究院,*W*1*W*2=研究院+长城,*P*–3*P*–2=*NULL*+v,*P*–2*P*–1=v+ns,*P*–1*P*0=ns+n,*P*0*P*1=n+n,*P*1*P*2=n+nz,*P*2*P*3=nz+n,*WP*–2=现任/v,*WP*–1=中国/ns,*WP*1=研究院/n,*WP*2=长城/nz.

图 1 给出了(*K*,*L*)不同取值情况下在 SemEval2007:Task #5 数据上的测试结果.从中可以看到,窗口的大小对结果有着明显的影响.

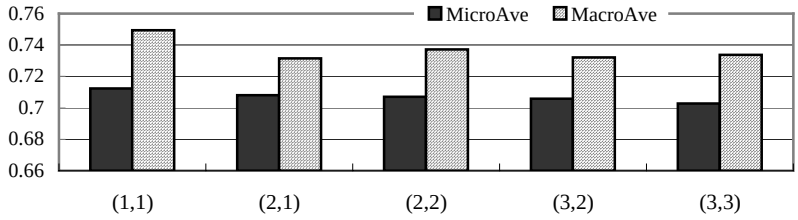

Fig.1 Effect on result within different window sizes

图1 窗口大小对结果的影响

图 1 中的实验结果以宏平均 MacroAve(macro-average accuracy)和微平均 MicroAve(micro-average accuracy)方式给出,计算公式分别为

$$MicroAve=\sum_{i=1}^{N}m_i\Big/\sum_{i=1}^{N}n_i\ ,\ MacroAve=\sum_{i=1}^{N}p_i\Big/N\ ,\ p_i=m_i/n_i \tag{2}$$

其中,*N* 表示歧义词个数,m_i 和 n_i 分别表示第 *i* 个词的正确消歧次数和出现次数,p_i 称为第 *i* 个词的正确率.

在上述基础上,本文进一步研究了特征模板构成的不同对结果的影响.我们在表 2 基础上固定(*K*,*L*)为(2,2),再增加表 3 所示的 3 类模板项,形成了增强特征模板.在例 1 中,主题特征有:*WB*=现任,*WB*=中国,*WB*=研究院,*WB*=长城,*WB*=医院,*WB*=院长,*WB*=的;这类特征实际上是词袋特征,不考虑位置因素.实体特征为:InEntity=ture,EType=nt,目标词特征为:*W*0=中医,*P*0=n,*WP*0=中医/n.

我们在(*K*,*L*)=(2,2)的窗口特征基础上,分别增加这 3 类特征项进行了实验评测,得到表 4 所示的 4 组实验结果.可以看到,使用不同类型的特征模板项,词义消歧的结果是不同的.

Table 3 Extended feature templates

表3 增强特征模板

Feature type	Feature template	Description
Window feature	(*K*,*L*)=(2,2)	See Table2
Topic	*W*[–5,5]	Words in window [–5,5]
Entity	InEntity, EType	Being in a named entity and the entity type
Target word	W_0,P_0,WP_0	Target word, its pos and the combination

Table 4 Effect of different feature template within the same window on result

表 4 相同窗口长度下不同特征模板项取舍对结果的影响

	Window (2,2)	+Entitiy	+Topic	+Target word
MicroAve	0.707 0	0.714 4	0.715 5	0.702 7
MacroAve	0.737 1	0.739 7	0.753 4	0.742 5

此外,不同词的词义消歧对窗口大小和特征项的要求也不相同.为了说明这一现象,我们从测试数据中选取了 6 个词:“中医”、“推翻”、“气息”、“吃”、“机组”和“菜”,分别取模板窗口(*K*,*L*)为(1,1),(2,2),(3,3),并按窗口特征模板进行了实验.图 2 给出了这些词在 3 个不同窗口下词义消歧的正确率,每个词的正确率计算见公式(2).

在图 2 中,“中医”、“推翻”、“气息”这 3 个词在窗口大小为(1,1)时结果最好;“吃”和“机组”在取(2,2)窗口时最好;而“菜”则在窗口为(3,3)时最好.

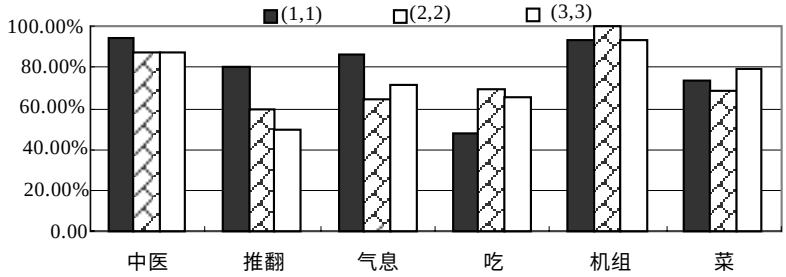

Fig.2 Test results of six ambiguous words with different feature templates

图 2 6 个歧义词在不同特征下的测试结果

上述 3 组实验的结果表明,因窗口大小变化导致特征模板项变化对词义消解会产生影响,不同的词对特征的要求也不完全相同.这也说明,预先设定统一的、固定不变的特征模板是不合理的,需要针对具体词选择具体特征.

2 特征自动选择算法

2.1 特征选择的基本原则

针对前面的现象,本文研究了特征模板的自动选择问题,即从一个较大的备选特征(模板)中选取一个子集作为实际使用的特征(模板).

通常,在选择子集时,需要考虑以下因素:

(1) 有效性:选择的子集应该比全集和其他子集具有更好的分类效果,这是本文主要关注的.

(2) 简洁性:特征子集应当尽量的简洁,以便节省系统资源,减小运行的复杂性.

(3) 针对性:特征的选择应该针对具体的对象,对象变化后,对特征的选择应作相应变化.

特征可以通过手工方法进行选择.即对于备选特征(模板),人工加入或者删除某些模板项,然后通过实验进行评估.然而,手工选择特征既费时又费力.原理上讲,每个模板项都可以被选择,也可以不被选择.如果备选的模板项有 m 个,那么,组合数多达 2^m-1 种(假设不能为空);而且,在手工选择的过程中,并没有一个系统的、定量的标准作为选择依据,这些都加大了手工选择的难度和选择的工作量.

2.2 统一特征(模板)选择算法

为了选择合理的子集,需要对所有子集进行枚举.在备选方案多达 2^m-1 的情况下,即使通过自动方式选择,复杂度也将相当可观.为了提高效率,本文采用贪心策略来逐步扩展.主要思想如下:

已选特征模板项从空开始,每次从备选集中选择一条特征模板项加入.在已经选取的模板项基础上,给每个尚未选取的模板项评分,每轮评分后选取得分最高的那个模板项;重复至多 m 次,这样就可以选择一个子集,同时,将原先 $O(2^m)$量级的复杂度降低为 $O(m^2)$量级.

特征与分类模型 M 是密切相关的,相同的特征对不同分类模型的影响并不完全相同.因此,特征的选择需要与分类模型结合进行.本文使用了最大熵作为分类模型.在选择特征(模板)项过程中,使用了语料 T 用于训练,语料 D 用于评估作为特征选择依据,其中,语料 T 和语料 D 是对原始训练语料按比例划分得到的,并未使用测试语料.

算法 **1**. 基于增强特征模板的自动特征选择算法:统一特征模板选择.

Input: *T*, *D*, *FT*.
Initialize: *FS*=∅,Scores=∅,DropTimes=0,PreScore=0.
for *k*=1 to |*FT*|:
    R=∅
    for each f_i in *FT*

use $FS \cup \{f_i\}$ as feature template for M to learn a classifier C_i on T
Compute p_i and $R = R \cup \{p_i\}$
end for
$i' = \text{argmax}_i R$
$FS = FS \cup \{f_{i'}\}$ //add $f_{i'}$ to FS
$FT = FT \setminus \{f_{i'}\}$ // delete $f_{i'}$ from FT
Scores[k] = $p_{i'}$
if Score[k]>PreScore then
PreScore=Score[k], DropTimes=0
else if Score[k]<PreScore then
PreScore=Score[k]
DropTimes+=1
end if
if DropTimes=2 and Scores[k]<δ then end for
end for
$k' = \text{argmax}_k$ Scores
$FS = FS[1..k']$
return FS

假设从备选特征模板项集合 FT 中选择一个子集 FS 作为最后的特征模板,选择过程按逐步扩展方式进行.设 FS 是一个有序的模板项序列,Score 序列用于记录 FS 中每个模板项的评分,初始值都为空.对于 FT 中每条尚未选入 FS 的模板项 f_i,算法将以 $FS \cup \{f_i\}$作为特征模板,利用 M 在 T 上训练出一个分类器 C_i,再在 D 上计算分类器 C_i 的得分 p_i,并追加到集合 R,然后再从 R 中选取 p_i 最高的模板项 f_i 加入 FS,并记录 f_i 的评分.重复这一步骤,直到 FT 中所有的模板项加入 FS.最后,从 Score 中找到最高评分对应的下标 k,将 FS 中前 k 个模板项作为算法输出返回.

算法 1 给出了选择过程.p_i 值的计算综合考虑宏平均和微平均值:

$$p_i = \frac{2 \times MicroAve \times MacroAve}{MicroAve + MacroAve} \tag{3}$$

在实际选择过程中,并不需要检查所有的模板项,当算法经过若干轮后,如果选取的模板项评分持续下降(本实验中持续下降的次数取 2)并且已经下降到某一阈值δ就可以终止.

2.3 针对每个歧义词的优化特征(模板)选择算法

上面介绍的方法是针对所有歧义词选择统一特征(模板)项.通过图 2 所示的实验结果不难看出,不同词对于特征模板项的需求是不一样.因此,有必要针对每个歧义词进行特征优化,形成优化特征模板.

在针对歧义词集合 WordList 中的各个歧义词选择特征模板项时,p_i 值就是各个词的正确率,见公式(2),其计算只与该词的评估语料相关.由于每个歧义词的训练语料 T 和评估语料 D 的规模非常有限,当按单个词选择特征模板项时,容易出现数据稀疏和过拟合问题,表现为由于评估语料 D 规模太小导致模板项选择的偶然性增加,同一轮选择中有可能出现多个模板项打分一样,而最后选择的模板项数量过少等现象.为了避免这些情况,我们对算法 1 作如下改进:

(1) 对于每个歧义词 W_j,针对该词的语料 T_j 和语料 D_j 合并为 T_j',并采用 N-fold 交叉训练(本实验中 N=4),将 T_j'分成 N 份,每次用 N–1 份训练模型,1 份进行评估,一共重复 N 次.这里需要说明的是,用于每个歧义词 W_j 的语料 T_j 和语料 D_j,是根据评测任务中所提供的训练数据划分产生的;

(2) 对备选特征模板项事先规定选择优先级,当发生打分一样时优先级高的模板项被优先选择(本实验中的优先级排序为:主题词特征,实体特征,目标词特征,以位置区分的单个特征,组合特征);

(3) 当 W_j 的选择结果模板少于某个阈值ω时,用算法 1 的统一特征模板替代该词的特征模板 FS_j.

3 实验与分析

我们仍然选择 SemEval2007:Task #5 Multilingual Chinese_English Lexical Sample Task 语料为评测数据.

在统一模板的特征选择过程中,我们将 SemEval2007:Task #5 提供的训练数据按照 3:1 划分为语料 T 和语料 *D*;在分别按词进行的优化特征选择过程中,我们将对应词的训练数据按 4-fold 交叉训练选择.所有的实验都使用相同的测试数据,即 SemEval2007:Task #5 提供的测试数据进行评测.

首先,我们在窗口特征模板(表 5)上进行实验,窗口大小为(*K*,*L*)=(2,2).

同时,我们针对增强特征模板进行了测试.这是在窗口特征模板基础上,通过增加主题特征、实体特征和目标词特征形成的,形成方式见表 3.

为了进行比较,本文还在增强特征模板基础上,通过人工方式精选了一组特征模板,见表 6.

在增强特征基础上,针对测试数据中的所有歧义词,我们使用算法 1 选取了统一特征模板(没有使用阈值δ限制).经过选择后得到的特征模板项见表 7.可以看到,自动选择的模板项数大为减小.

Table 5 (2,2) window feature templates

表 **5** (2,2)窗口特征模板

Feature type	Feature template	Description
Independent feature	*W*–2,*W*–1,*W*1,*W*2	Words in position –2,–1,1,2
	P–2,*P*–1,*P*1,*P*2	Pos of each word mentioned above
Combination	*W*–2*W*–1,*W*–1*W*0,*W*0*W*1,*W*1*W*2	2-gram words in the window
	P–2*P*–1,*P*–1*P*0,*P*0*P*1,*P*1*P*2	2-gram pos in the windows
	WP–2,*WP*–1,*WP*1,*WP*2	Pair of word and its pos in the windows

Table 6 Manually selected feature templates

表 **6** 手工特征模板

Feature type	Feature template
Independent feature	*W*–1, *W*1, *P*–1, *P*1
Combination	*W*–1*W*0, *W*0*W*1, *P*–1*P*0, *P*0*P*1
Topic	*W*[–5,5]
Entity	InEntity, EType

Table 7 Selected feature templates based on all disambiguation words

表 **7** 自动选择的统一特征模板

Feature type	Feature
Independent feature	*W*–1,*W*1,*P*–1,*P*1
Combination	*P*–1*P*0
Topic	*W*[–5,5]
Entity	InEntity, EType

在增强特征基础上,分别针对每个歧义词优化了特征模板项的选择,表 8 给出了一部分词的特征模板.

Table 8 Selected feature templates based on individual word

表 **8** 按词自动选择的独立特征模板

Word	Feature template
中医	*Tp*, *P*1, *P*0*P*1
使	*Tp*, *W*1, *Tg*, *Et*, *W*–1, *P*2, *P*1*P*2
儿女	*P*–1, *Tg*, *Et*, *W*–1, *W*1, *WP*–1
出	*Tp*, *P*–1, *P*1*P*2, *WP*2, *Et*, *W*1, *W*1*W*2, *Tg*, *W*–1, *P*–2, *WP*–1, *WP*1, *P*–2*P*–1
动	*W*1, *P*1
动摇	*P*2, *Tp*, *Tg*, *Et*, *W*2, *P*1, *W*–1, *P*–1, *P*0*P*1
叫	*P*1, *W*1, *Et*, *WP*1, *P*2, *Tp*, *P*1*P*2
……	
震惊	*WP*1, *Tg*
面	*Tp*, *W*–1, *P*2, *WP*–1

表 8 中,*Et* 表示实体特征(InEntity, EType),*Tp* 表示主题特征(即 *W*[–5,5]),*Tg* 表示目标词特征(W_0,P_0,WP_0).可以看出,有些词语的特征模板项过少.在本实验中特征项小于 5 的词(如表 8 中的中医、动、震惊、面),模板替换为表 7 中的统一特征模板.

针对上述各类特征模板,我们分别进行了评测,仍然按照 MicroAve 和 MacroAve 两种标准度量,评测结果如图 3 所示.

在图 3 中,Baseline 是 SemEval2007 Task5 评测中第 1 名的结果[3],其 MicroAve 和 MacroAve 分别为 71.66%和 74.92%;若采用简单的窗口特征模板,消歧效果并不理想,MicroAve 和 MacroAve 分别为 70.70%和 73.71%,不及 Baseline;如果将窗口模板扩展为增强特征模板,结果上升为 72.19%和 75.39%.好于 Baseline.在增强模板的基

础上进行手工选择特征模板,结果为 73.69%和 76.85%,有比较明显的提升,但这种改进仍然是不够的.如果采用自动特征选取的方法选择统一特征模板,MicroAve 和 MacroAve 分别达到 73.90%和 77.41%;按每个词选取独立特征模板,则提高到了 74.76%和 77.88%,相比 SemEval2007 Task5 评测中第 1 名,MicroAve 和 MacroAve 值分别提升了 3.10 和 2.96 个百分点.

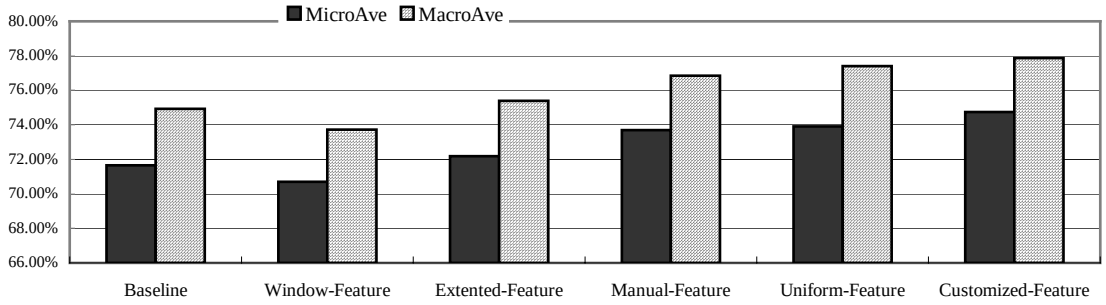

Fig.3 Evaluation results with different feature templates

图 3 不同特征下的测试结果

图 3 所给的 6 组实验都是用最大熵模型实现的.因为所选特征不同,结果出现了差异.图中左边的 4 组特征(Baseline,窗口特征、增强特征和手工特征)其实都是人工选择的.其中,窗口特征对应的测试结果较差,手工特征对应的结果相对较好.手工特征是我们在已知测试数据答案的基础上,通过反复实验得到的.在真正的应用(或评测)中,待测数据的答案是未知的,这就使手工选择特征缺乏明确依据.通过自动方式选择特征,只需事先给出合适的备选特征(主要是窗口大小)即可.这不仅减少了人工的工作量,更重要的是,有利于控制改进消歧的效果.图 3 表明,两种自动选择方法在结果上都有改进.

此外,吴云芳等人研究了 3 种不同分类模型的 9 种集成方法[11].他们同样在 SemEval2007 Task5 上作了评测,但只报道了宏平均值,名词最高为 73.9%,动词最高为 75.8%(如果名词和动词一起计算,将低于 75.8%).本文在名词和动词一起计算的情况下,两种自动特征模板选择方式得到的宏平均值均超过了 77% .这进一步说明,除了选择高性能的分类模型外,特征的选择也非常重要.

4 结束语

自然语言处理的很多任务都可以归结为分类问题,影响分类性能的一个重要因素是特征选择.本文所给的多个实验表明,特征模板的选择对于词义消歧的结果具有明显影响.

本文使用最大熵分类模型,针对汉语的词义消歧,重点研究了特征(模板)的选择问题.介绍了基于最大熵模型的特征模板自动选择方法,基本思想是通过综合打分,渐近式地扩展特征模板项.

本文提出了两种模板选择算法:其一,针对所有歧义词(测试数据有 19 个名词,21 个动词),设计了统一特征模板的选择算法;其二,针对每个歧义词,实现了优化的特征模板选择算法,这一算法在特定情况下引入统一特征模板进行修正.实验结果表明,两种特征选择都改进了词义消歧的性能,后者的效果尤为明显.

特征模板选择不仅有助于压缩特征模板项的数目,降低计算的复杂性,而且还可以改善词义消歧的性能.这表明,自动选择特征是很有作用的.

本文提出的特征选择方法可以作为一套框架与很多其他分类方法结合使用.下一步我们将研究特征选取方法与半指导分类方法的结合,同时,也将研究在褒贬分析、文本分类等方面的应用.

References:

[1] Bar-Hillel. The present status of automatic translations of languages. Advances in Computers, 1960,1:91−163.

[2] Pedersen T. A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation. In: Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL). 2000. 63−69. http://www.d.umn.edu/~tpederse/pubs/naacl00.pdf

[3] Xing Y. SRCB-WSD: Supervised Chinese word sense disambiguation with key features. In: Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007). 2007. 300−303. http://aclweb.org/anthology-new/S/S07/S07-1065.pdf

[4] Yee KO. CITYU-HIF: WSD with human-informed feature preference. In: Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007), 2007. 109−112. http://aclweb.org/anthology-new/S/S07/S07-1020.pdf

[5] Jin P, Wu YF, Yu SW. SemEval-2007 Task 5: Multilingual Chinese-English lexical sample. In: Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007). 2007. 19−23. http://aclweb.org/anthology-new/S/S07/S07-1004.pdf

[6] Mihalcea R. Co-Training and self-training for word sense disambiguation. In: Proc. of the CoNLL 2004. http:// www.cse.unt.edu/~rada/papers/mihalcea.conll04.pdf

[7] Mihalcea R, Chklovski T, Killgariff A. The Senseval-3 English lexical sample task. In: Proc. of the 3rd Int'l Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3). 2004. http://www.cse.unt.edu/~rada/papers/mihalcea2.senseval04.pdf

[8] Pham TP, Ng HT, Lee WS. Word sense disambiguation with semisupervised learning. In: Proc. of the 20th AAAI Conf. on Artificial Intelligence (AAAI-2005). 2005. http://www.comp.nus.edu.sg/~nght/pubs/aaai05-wsd-ssup.pdf

[9] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995). 1995. 189−196. http://www.cs.jhu.edu/~yarowsky/acl95.ps

[10] Quan CQ, He TT, Ji DH, Yu SW. Word sense disambiguation based on multi-classifier decision. Journal of Computer Research and Development, 2006,43(5):933−939 (in Chinese with English abstract).

[11] Wu YF, Wang M, Jin P, Yu SW. Ensemble of classifiers for chinese word sense disambiguation. Journal of Computer Research and Development, 2008,45(8):1354−1361 (in Chinese with English abstract).

[12] Liu FC, Huang DG, Jiang P. Chinese word sense disambiguation with AdaBoost.MH Algorithm. Journal of Chinese Information Processing, 2006,20(3):6−13 (in Chinese with English abstract).

[13] Xu Y, Li JT, Wang B, Sun CM. A category resolve power-based feature selection method. Journal of Software, 2008,19(1):82−89 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/82.htm [doi: 10.3724/SP.J.1001.2008.00082]

[14] Vincent Ng, Claire Cardie. Weakly supervised natural language learning without redundant views. In: Proc. of the HLT-NAACL. 2003. 94−101. http://www.hlt.utdallas.edu/~vince/papers.hlt-naacl03.pdf

[15] Berger AL, Pietray SAD, Pietray VJD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):1−36.

附中文参考文献:

[10] 全昌勤,何婷婷,姬东鸿,余绍文.基于多分类器决策的词义消歧方法.计算机研究与发展,2006,43(5):933−939.

[11] 吴云芳,王淼,金澎,俞士汶.多分类器集成的汉语词义消歧研究.计算机研究与发展,2008,45(8):1354−1361.

[12] 刘风成,黄德根,姜鹏.基于 AdaBoost.MH 算法的汉语多义词消歧.中文信息学报,2006,20(3):6−13.

[13] 徐燕,李锦涛,王斌,孙春明.基于区分类别能力的高性能特征选择方法.软件学报,2008,19(1):82−89. http://www.jos.org.cn/1000-9825/19/82.htm [doi: 10.3724/SP.J.1001.2008.00082]

何径舟(1985－),男,江苏南通人,硕士生,主要研究领域为自然语言处理.

王厚峰(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器学习.