中 文 信 息 学 报
第 20卷 第 3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

文章编号: 1003 - 0077(2006)03 - 0006 - 08

# 基于 AdaBoost MH算法的汉语多义词消歧 *

刘风成,黄德根,姜 鹏

(大连理工大学 计算机科学与技术系,大连 116024)

**摘要**:本文提出一种基于 AdaBoost MH算法的有指导的汉语多义词消歧方法,该方法利用 AdaBoost MH算法对决策树产生的弱规则进行加强,经过若干次迭代后,最终得到一个准确度更高的分类规则;并给出了一种简单的终止算法中迭代的方法;为获取多义词上下文中的知识源,在采用传统的词性标注和局部搭配序列等知识源的基础上,引入了一种新的知识源,即语义范畴,提高了算法的学习效率和排歧的正确率。通过对 6个典型多义词和 SENSEVAL3中文语料中 20个多义词的词义消歧实验,AdaBoost MH算法获得了较高的开放测试正确率(85.75%)。

**关键词**:人工智能;自然语言处理;词义消歧;AdaBoost MH算法;多知识源

**中图分类号**: TP391 **文献标识码**: A

## Chinese Word Sense Disambiguation with AdaBoost.MH Algorithm

LIU Feng-cheng, HUANG De-gen, JIANG Peng

(Department of Computer Science, Dalian University of Technology, Dalian 116024, China)

**Abstract:** An approach based on supervised AdaBoost.MH learning algorithm for Chinese word sense disambiguation is presented. AdaBoost.MH algorithm is employed to boost the accuracy of the weak decision stumps rules for trees and repeatedly calls a learner to finally produce a more accurate rule. A simple stopping criterion is also presented. In order to extract more contextual information, we introduce a new semantic categorization knowledge which is useful for improving the learning efficiency of the algorithm and accuracy of disambiguation, in addition to using two classical knowledge sources, part-of-speech of neighboring words and local collocations. AdaBoost.MH algorithm making use of these knowledge sources achieves 85.75% disambiguation accuracy in open test for 6 typical polysemous words and 20 polysemous words of SENSEVAL3 Chinese corpus.

**Key words:** artificial intelligence; natural language processing; word sense disambiguation; AdaBoost.MH algorithm; multiple knowledge sources

## 1 引言

词义消歧(Word Sense Disambiguation,简称 WSD)一直是自然语言处理研究领域十分重要的问题和研究热点之一[1],在机器翻译、信息检索、自动文摘、知识挖掘等自然语言处理领域均具有重要的应用价值。

近几年,国内外研究人员将统计学和机器学习引入到词义消歧的处理中,提出基于语料库的多义词处理方法(Corpus Based Approach, CBA)。一般来说,有指导的消歧方法要比无指导

* **收稿日期**: 2005 - 05 - 26 **定稿日期**: 2005 - 10 - 26

**基金项目**:国家自然科学基金资助项目(60373095; 60373096)

**作者简介**:刘风成(1978—),男,在读硕士,主要研究方向为自然语言处理.

的方法[2,3]有更好的效果。许多标准的有指导的学习算法被应用于词义消歧的模型中,如贝叶斯分类[4]、基于信息论的方法[5]、神经网络算法[6]等。

有指导的 AdaBoost MH算法是提高预测学习系统能力的有效工具,在 POS[7]和文本分类[8]中得到成功应用。AdaBoost MH算法的主要思想是把多个不太准确的弱规则合并为一个高精度的分类规则。弱规则由一个独立的称为弱学习器 (Weak Learner)的过程产生,这些弱规则组合得到最终规则,即通过学习得到的分类规则。

利用 AdaBoost MH算法的思想,在设计词义排歧的学习模型时,只要给定足够的数据和一个能可靠地提供中度正确的弱假设的弱学习器,它就获得了理论上的保证,而不需要设计一个能产生高准确率的学习器。

Schapire and Singer[9]提出的 AdaBoost MH算法适合多类多标签的分类问题。对于 WSD,由于在每个给定的上下文中多义词的词义是唯一确定的,因此 WSD只是单标签的分类问题。本文对 AdaBoost MH算法进行了调整,提出了面向单标签的 WSD的加强算法,该算法对简单决策树产生的分类规则进行加强,经过若干次迭代后,最终得到一个准确度更高的分类规则,即为最终的消歧模型;为获取多义词上下文中的知识源,本文在采用传统的词性标注和局部搭配序列等知识源的基础上,引入了一种新的知识源,即语义范畴。同时给出了一种简单的终止 AdaBoost MH算法中迭代的方法。

## 2 面向 WSD的 AdaBoost MH算法

### 2.1 算法描述

本文在算法描述中沿用了 Schapire and Singer[9]中的一些表示符号。

设样本集 $S = \{(x_1, y_1), \cdots\cdots (x_m, y_m)\}$,其中,$x_i \in X$, $X$为训练集,$m = |X|$; $y_i \in Y$, $Y$为词义标签集,记 $k = |Y|$。样本 $(x, y)$为单一实例 $x$和该实例对应的词义标签 $y$。

AdaBoost MH维持了样本集 $S$上一个 $m \times k$的权重分布 $D$,初始状态下,分布 $D$的权值是相同的。

令 $D_t$为第 $t$次迭代后的分布,$h_t: X \times Y \rightarrow R$为分布 $D_t$上获得的弱规则,该规则由弱学习器产生。$h_t(x, l)$表示对词义标签 $l \in Y$是否应该赋给实例 $x$的一种预测,其值 $|h_t(x, l)|$反映了这种预测的可信度。

弱规则的产生是一个序列式的学习过程。在每次迭代中,运用下面的调整公式调整每个样本的权值,使每次输入弱学习器的样本集具有不同的权重,让弱学习器集中学习那些使用前一规则最难以预测的样本 $(x, y)$。

给定 $l \in Y$, $y_i \in Y$,引入符号 $y_i[l] \in \{-1, +1\}$,若 $y_i$与标签 1一致,$y_i[l]$的值为 1,否则 $y_i[l]$的值为 -1,调整公式定义如下:

$$D_{t+1}(i, l) = \frac{D_t(i, l)\exp(-y_i[l]h_t(x_i, l))}{Z_t} \quad (1)$$

其中 $1 \leq i \leq m$, $Z_t$为归一化因子 ($Z_t$的计算公式参见 2.2部分), Schapire and Singer[9]已经证明,算法误差的最大值为: $\prod_{t=1}^{T} Z_t$ ($T$为总的弱规则数),因此为减少总误差,在选取 $Z_t$时应使其最小化。从公式中可以看出,若 $h_t(x, l)$的预测可信度较好,则 $D_t(i, l)$的值是增加的,反之,$D_t(i, l)$的值是减少的,且 $D_t(i, l)$的调整幅度与 $|h_t(x, l)|$成比例。

由于 WSD为单标签分类,因此最终形成的合并假设为唯一的标签 l,而且这个标签满足 $f(x, l)$最大化。

面向 WSD的 AdaBoost MH算法如下:

1) 输入样本集 $S$

2) 初始化: $D_1(i, l) = 1/mk;$ $\quad 1 \leq i \leq m, 1 \leq l \leq k$

3) 训练过程:循环学习 $T$次

a) 把 $D_t$传给并调用弱规则学习器;

b) 获得弱规则 $h_t: X \times Y \rightarrow R$

c) 利用调整公式调整矩阵 $D$的权值

4) 求 $f(x, l) = \max_t \sum_{t=1}^{T} h_t(x, l)$,即为最终形成的合并假设。

## 2.2 弱学习器的设计及 $Z_t$的选取

本文采用简单决策树作为弱学习器,其中树的叶子定义了实例空间 $X$的一个划分,弱学习器做出的假设是基于该划分的。对于划分中任一划分块 $X_j$,若任意的 $x, x' \in X_j$,总有 $h(x) = h(x')$,也就是说, $h(x)$做出的预测,仅与给定实例 $x$所属的划分块有关。

对于任意给定的特征 $p$ (特征的选取参见下面第 3部分),可以将实例空间 $X$划分为 2个划分, $X1$为包含 $p$的实例集, $X0$为不包含 $p$的实例集。

一个简单的弱规则形式如下:

$$h(x, l) = \begin{cases} c_{0l} & p\text{不是 } x\text{的特征} \\ c_{1l} & p\text{是 } x\text{的特征} \end{cases} \tag{2}$$

其中 $c_{jl}$ ($j \in \{0, 1\}$)是一个实数。

一个弱规则产生的过程如下,在每次迭代中,对每一个特征,弱规则学习器计算预测误差;预测标准为,某一实例属于某一分类,当且仅当实例中包含此特征。最小化分类误差的特征被选择为产生此次迭代过程中弱规则的特征。

Schapire and Singer[9]已经证明,对于给定的迭代 $t$与 $c_{jl}$,选取 $c_{jl} = 0.5 \times \ln(v_{+1}^{jl} / v_{-1}^{jl})$,可以使 $Z_t$最小。这时有:

$$Z_t = 2 \sum_{j \in \{0, 1\}} \sum_{l \in Y} \sqrt{v_{+1}^{jl} v_{-1}^{jl}} \tag{3}$$

其中 $v_{+1}^{jl}$ ($v_{-1}^{jl}$)的计算如下:在给定分布 $D_t$的情况下,对于每一可能的词义标签 1,在 $j \in \{0, 1\}$和 $b \in \{-1, +1\}$的情况下,有:

$$v_b^{jl} = \sum D_t(i, l) \parallel x_i \in X_j \cap y_i[l] = b \parallel \tag{4}$$

也就是说, $v_{+1}^{jl}$ ($v_{-1}^{jl}$)为 $X_j$中词义标签为 (不为) 1的训练实例在分布 $D_t$中的权值和。

在每次迭代中,对每个实例,求解 $Z_t$。选取 $Z_t$最小的实例所对应的 $h(x, l)$为本次迭代的弱规则。在计算 $c_{jl}$时,如果 $v_{+1}^{jl}$或 $v_{-1}^{jl}$非常小或接近零,会导致 $c_{jl}$的值是很大或者无穷大。这样,这种大的预测值会引起数值问题。为避免这种情况的发生,需要对 $c_{jl}$采用平滑值:

$$c_{jl} = 0.5 \times \ln\left(\frac{v_{+1}^{jl} + \varepsilon}{v_{-1}^{jl} + \varepsilon}\right) \text{其中 } \varepsilon > 0 \tag{5}$$

由于 $v_{+1}^{jl}$和 $v_{-1}^{jl}$的值域为 $[0, 1]$,因此这种平滑对 $c_{jl}$造成的影响的上限为:

$$\frac{1}{2} \ln\left(\frac{1 + \varepsilon}{\varepsilon}\right) \approx \frac{1}{2} \ln\left(\frac{1}{\varepsilon}\right) \tag{6}$$

本文采用的平滑系数为: $\varepsilon = 1/(m \times k)$。

## 3 知识源

### 3.1 相邻词的词性标注 (POS)

词性标注资源共使用以下 7个特征: $P_{-3}$, $P_{-2}$, $P_{-1}$, $P_0$, $P_{+1}$, $P_{+2}$, $P_{+3}$。

其中 $P_{-i}$ ($P_{+i}$)代表多义词 $w$ 左 (右)边第 $i$ 单词所对应的词性, $P_0$ 代表多义词 $w$ 所对应的词性。

例如,为区分"材料"一词在下面句子中的词义:

以 /p/Kb05 树叶 /n/Bh11 、/wp/-1 彩布条 /n/Bq01 和 /c/Kc01 香烟盒纸 /n/Bp17 为 /v/Ja01 材料 /n/Ba06 的 /u/Kd01 贺卡 /n/Bp18 ,/wp/-1 做工 /v/Hj21 虽 /c/Kc04 简 /a/Ed37 ,/wp/-1 其 /r/Ba10 情 /n/Df04 却 /d/Ka33 浓 /a/Eb12 。 /wp/-1 例句 (1)

上下文中包含的 POS资源有:

$$P_{-3} = c,\ P_{-2} = n,\ P_{-1} = v,\ P_0 = n,\ P_{+1} = u,\ P_{+2} = n,\ P_{+3} = wp$$

### 3.2 局部搭配信息

局部搭配 $C_{i,j}$ 指多义词 $w$ 上下文中局部的连续的单词序列信息。 $i$, $j$ 分别代表多元序列的起始和结束位置。实验中,多元信息主要考虑了以下 9个特征:

$$C_{-1,-1},\ C_{1,1},\ C_{-2,-2},\ C_{2,2},\ C_{-3,-3},\ C_{3,3},\ C_{-2,-1},\ C_{-1,1},\ C_{1,2}$$

例如,对于例句 (1),考虑的局部搭配中 $C_{-1,-1}$ 和 $C_{-2,-1}$,对应的特征为:

"$C_{-1,-1}$ =为"和"$C_{-2,-1}$ 香烟盒纸 为"

### 3.3 语义范畴信息

上下文中词汇的语义范畴大体上确定了这个上下文的语义范畴,并且上下文的语义范畴可以反过来确定词汇的哪一个语义被使用[10]。Walker[11] 和 Yarowsky[12] 在词义消歧的研究中分别提出了基于义类辞典的消歧算法。

《同义词词林》[13] 是当前汉语信息处理中的一部机器可读的一类词典。全书把词义分为大、中、小类三级,共分为 12个大类, 94个中类, 1428个小类。词典中给每个指定的词一个或多个语义码。如:词"觉悟"具有的一个语义码为"Ga15",词"材料"有三个语义码,分别为:"Ba06"、"Dk17"和"Al03"。

实验中考虑的语义范畴信息主要包括以下 6个特征:

$$d_{-3},\ d_{-2},\ d_{-1},\ d_{+1},\ d_{+2},\ d_{+3}$$

其中 $d_{-i}$ ($d_{+i}$)代表多义词 $w$ 左 (右)第 $i$ 单词所对应的语义码信息。

例如,对于例句 (1),考虑的语义信息包括以下特征:

$$d_{-3} = \text{Kc01},\ d_{-2} = \text{Bp17},\ d_{-1} = \text{Ja01},$$
$$d_{+1} = \text{Kd01},\ d_{+2} = \text{Bp18},\ d_{+3} = -1$$

其中"Kc01"、"Bp17"等为相应词汇的语义码。

## 4 实验

### 4.1 语料库

#### 4.1.1 人民日报语料

实验中使用的人民日报语料 (见表 1)来源于哈工大信息检索研究室提供的《同义词词林扩展版》和北大人民日报 2000年半年的电子版。《扩展版》中使用的语料为北大人民日报 1998年上半年的电子版 (带有词性标注),约 35万词语。《扩展版》在原来人民日报的基础上,按照《同义词词林》中定义的词的语义分类原则,在语料中增加了词的义类信息。因《扩展

版》中多义词语料有限,本文从人民日报 2000年半年的语料中抽取部分含有多义词的语料,按照《同义词词林》中定义的语义分类原则,进行人工义类标注。

4.1.2 SENSEVAL3中文语料

SENSEVAL是语义系统评估方面的国际性平台。该平台公开发布了一些评测语义系统的国际标准语料。SENSEVAL3中新增加了中文语料,语料格式按照 SENSEVAL的标准格式发布。SENSEVAL3中文语料 (表 2)共包括 20个多义词,其中训练语料为 793句,测试语料为 380句。

### 4.2 实验评测及结果

为评测 AdaBoost MH算法的学习效果,对多义词的学习结果进行测试,测试分为封闭测试和开放测试。由于本 WSD系统对每个测试实例总是输出一个唯一的值,因此其召回率和正确率总是相同的。这里,对测试结果的好坏只使用正确率作为衡量的标准,正确率定义如下:

$$\text{正确率}\frac{\text{正确排歧的实例个数}}{\text{可排歧的实例个数}} \tag{7}$$

表 1 实验中使用的人民日报语料及实验结果 (交叉实验 5次且封闭测试正确率 95%以上)

| 多义词 | 句子数 | 词义数 | 特征数 | POS | 开放测试正确率 (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 | 5 | 平均值 |
| 材料 | 349 | 2 | 2919 | n | 85.29 (70) | 91.17(600) | 86.76(50) | 86.76(600) | 85.29(150) | 87.05 |
| 地方 | 89 | 3 | 918 | n | 84.21(150) | 78.94(500) | 78.94(500) | 78.94(600) | 84.21(600) | 81.05 |
| 发表 | 157 | 2 | 1449 | v | 93.10(50) | 93.10(70) | 89.66(50) | 86.21(100) | 93.10(50) | 91.02 |
| 为 | 212 | 2 | 1836 | v | 90.00(50) | 82.50(600) | 85.00(500) | 87.50(200) | 82.50(300) | 85.50 |
| 要 | 540 | 4 | 4632 | v | 81.48(650) | 83.33(800) | 87.96(500) | 85.18(750) | 80.55(700) | 83.70 |
| 到 | 234 | 4 | 2388 | v | 83.33(50) | 88.09(500) | 80.95(50) | 88.09(600) | 83.33(500) | 86.19 |
| 平均 | | | | | | | | | | 85.75 |

注:表中开放测试正确率括号内的数值为获得此正确率的迭代次数。

表 2 实验中使用的 SENSEVAL3中文语料及实验结果

| 多义词 | 训练语料数 | 测试语料数 | 词义数 | 特征数 | POS | 迭代次数 | 封闭测试正确率 | 开放测试正确率 |
|---|---|---|---|---|---|---|---|---|
| 把握 | 31 | 15 | 4 | 386 | nvvn | 30 | 100% | 73.3% |
| 包 | 76 | 36 | 8 | 888 | nnrqv | 50 | 96.05% | 50% |
| 材料 | 20 | 20 | 2 | 352 | n | 100 | 100% | 80% |
| 冲击 | 28 | 13 | 3 | 360 | vnv | 80 | 96.43% | 84.6% |
| 传 | 28 | 14 | 3 | 362 | v | 300 | 100% | 64.28% |
| 地方 | 36 | 17 | 4 | 421 | bn | 200 | 100% | 70.59% |
| 分子 | 36 | 16 | 2 | 435 | n | 70 | 97.22% | 75% |
| 活动 | 36 | 16 | 5 | 426 | avvn | 150 | 100% | 68.75% |
| 老 | 57 | 26 | 6 | 616 | Ngaandj | 500 | 100% | 61.54% |
| 路 | 57 | 28 | 6 | 605 | nnrq | 300 | 100% | 64.29% |
| 没有 | 30 | 15 | 3 | 386 | dv | 200 | 100% | 66.67% |
| 起来 | 40 | 20 | 4 | 513 | v | 150 | 97.5% | 75% |
| 钱 | 40 | 20 | 4 | 506 | nnr | 600 | 100% | 75% |
| 日子 | 48 | 21 | 3 | 560 | n | 80 | 100% | 66.67% |
| 少 | 42 | 20 | 5 | 510 | Ngaadjv | 750 | 100% | 60% |
| 突出 | 30 | 15 | 3 | 387 | aadv | 500 | 100% | 53.33% |

续表

| 多义词 | 训练语料数 | 测试语料数 | 词义数 | 特征数 | POS | 迭代次数 | 封闭测试正确率 | 开放测试正确率 |
|---|---|---|---|---|---|---|---|---|
| 研究 | 30 | 15 | 3 | 380 | nvvn | 70 | 96.67% | 73.33% |
| 运动 | 54 | 27 | 3 | 601 | nnzvvn | 100 | 100% | 70.37% |
| 走 | 49 | 24 | 5 | 565 | vvn | 500 | 100% | 62.5% |
| 做 | 25 | 12 | 3 | 304 | v | 600 | 100% | 58.33% |
| 平均 | | | | | | | 99.19% | 67.68% |

4.2.1 人民日报语料实验

为保证测试结果的客观性,实验中对语料进行交叉验证,从中随机抽取 80%作为训练语料,其余 20%作为开放测试语料,进行 WSD实验。然后再重复这个的选择过程,每次选择不同的 20%作为测试语料,其余的为训练语料,取这若干次实验结果的平均值作为最终的结果。

本文对 6个典型的多义词 (见表 1)分别进行了实验。实验结果见表 1。

4.2.2 SENSEVAL3中文语料实验

在 SENSEVAL3中文语料上进行的 WSD实验情况见表 2。同时本文将 AdaBoosMH的实验结果和 Zheng-Yu Niu and Dong-Hong Ji[14]的实验结果进行了比较。Zheng-Yu Niu and Dong-Hong Ji实验中使用的是贝叶斯 WSD算法,使用的语料同为 SENSEVAL3的中文语料。对比实验模型中选取的上下文特征与 Zheng-Yu Niu and Dong-Hong Ji中选择的上下文特征保持一致,即 POS特征和一定窗口的词袋信息,对比情况见表 3。

从表 2中的数据可以看出,在 SENSEVAL3上得到的测试结果低于表 1中的实验结果。这主要是由于 SENSEVAL3中的语料偏少导致数据不足、特征空间过于稀疏的缘故。AdaBoosMH在给定的数据不充分、弱假设过度复杂或弱假设太弱的情况下,不能表现出很好的性能,这一点与理论一致[9]。但是,从作为语义评测的平台的角度来看,SENSEVAL3的中文语料还是具有它的可行性的。从表 3的数据中可以看出,AdaBoost MH算法模型的优于贝叶斯算法模型,高出 7.28个百分点。

表 3 AdaBoost MH与贝叶斯在 SENSEVAL3中文语料上的对比

| 算法模型 | 开放测试平均正确率 (%) |
|---|---|
| AdaBoost MH算法模型 | 67.68% |
| 贝叶斯算法模型 | 60.40% |

4.3 算法中迭代次数的确定

表 1的迭代次数一列中的数据表明:在获得较高的正确率的情况下,算法对每个词进行学习的迭代次数是不同的。这是由于每个词具有不同的特征属性和语料中包含的句子对表征该词的特征属性的贡献能力不同而致。

图 1数据表明测试结果的准确率与迭代次数的多少并不总是成正比。每个词语有一个最佳的迭代次数。Schapire and Singer[9]指出迭代次数过多,AdaBoost MH有可能发生过适应。另一方面从系统的使用角度来看,迭代次数增多,意味着学习所需的时间和保存学习结果所需的空间的增加。因此需要在实际应用中,需要确立一个合适的迭代次数,也即终止迭代的条件。

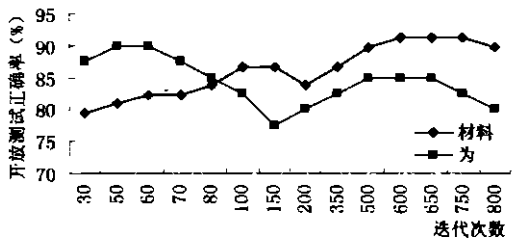


图 1 迭代次数的确定

实验中采用了一个简单的终止条件,即当封闭训练的正确率达到 95%以后,开放测试的第二个极大值点处终止迭代过程。选择两个极值中的较大者的迭代次数作为最终的迭代次数。

### 4.4 语义信息对排歧效果的影响

在获取多义词上下文信息的过程中,我们使用了的多种知识源:词性标注、局部搭配序列和语义范畴。前两部分知识源在以往的 WSD的研究中,已被多次使用,并显示出比单独的词袋信息更好的效果,本文在以往的知识源中加入了语义范畴信息。为观察语义信息在排歧过形成中对结果的影响,本文以"材料"一词的语料为例,在相同的语料情况下(随机从语料中选取 285句作为训练语料,剩余的作为开放测试语料共 63句),对"材料"一词进行 WSD实验。实验对比情况见图 2、图 3。

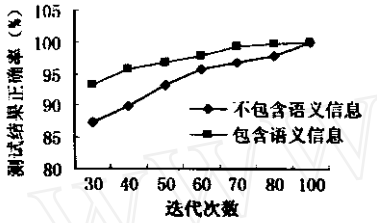


图 2 语义信息对封闭测试的影响

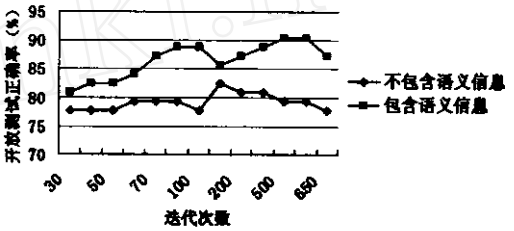


图 3 语义信息对开放测试的影响

从图 2和图 3中数据可以看出,语义范畴信息加快算法的学习速度和提高排歧的正确率方面的效果比较明显。对于封闭测试的情况,标注了语义范畴信息的学习模型的测试正确率达到 100%需要的迭代次数比未加入语义范畴信息的模型减少了近 20余次,并且在相同的迭代次数下,加入语义范畴信息的模型的测试正确率比未加入语义范畴信息的模型平均高出 2%～7%;同样对于开放测试,不仅在同样的迭代次数下,标注了语义范畴信息的学习模型的测试正确率比未加入语义范畴信息的模型提高了 2%～10%,而且标注了语义范畴信息的学习模型的最好结果比未标注语义范畴信息的模型高出约 5%。

## 5 结论与进一步研究

本文在汉语词义消歧中引入有指导的 AdaBoost MH分类方法,通过对多义词上下文的多知识源(词性标注、局部搭配序列和语义范畴)进行学习,获得了正确率较高的分类器。实验结果表明 AdaBoost MH算法具有较强的学习能力和较高的排歧正确率(开放测试正确率平均为 85.75%);结合每个多义词所具有的特征属性的不同性和系统的实用性,给出了一种简单实用的终止算法中迭代的方法。

与其它有指导的学习算法一样,AdaBoost MH算法需要在带词义标注的训练语料中获取知识。从表 1的数据可以看出,语料的规模和质量(指语料所提供的上下文信息对排歧的贡献)对排歧结果和系统的效率有一定的影响:高正确率高效率的排歧系统有赖于规模适当而且质量较高的语料。但是,在人工标注的过程中,建立这种量与质兼有的大规模语料是很难的[14]。因此,如何利用其他语义资源(如 WordNet[15]和 HowNet[16])和相关方法[15,17,18]来自动获取大规模的标注语料是进一步需要研究的课题;此外,本文在学习过程中获取的特征限制在多义词前后 3个窗口内,这会导致上下文中一些对排歧有用的信息没有被获取到,如果只是简单的扩大窗口,又会引入过多的噪音,影响排歧效果,因此如何有效的利用多义词上下文的信

息也是今后需要进一步研究的课题。

## 参 考 文 献:


[1] N. Ide, J. Veronis, Introduction to the special Issue on Word Sense Disambiguation: The State of the Art[J]. Computational Linguistics, ACL, 1998. 24(1).

[2] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods[A]. In: the 33rd Annual Meeting of ACL[C]. Massachusetts, 1995: 181 - 188.

[3] 李涓子,黄昌宁,杨尔弘. 一种自组织的汉语词义排歧方法[J]. 中文信息学报, 1999, 13(3): 1 - 8.

[4] H. T. Ng, Exemplar-based Word Sense Disambiguation: Some Recent Improvements[A]. In: proceeding of the 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP, 1997.

[5] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense disambiguation using statistical methods[A]. In: proceedings of the 29th conference on Association for Computational Linguistics[C]. California, June 1991, 264 - 270

[6] G. Towell, E. M. Voorhees, Disambiguating Highly Ambiguous Words[J]. Computational Linguistics, ACL, 1998. 24(1).

[7] S. Abney, R. E. Schapire, Y. Singer. Boosting Applied to Tagging and PP-attachment[A]. In: proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Proceeding and Very larger Corpora[C]. 1999.

[8] R. E. Schapire, Y. Singer, BoosTexter: A Boosting-based System for Text Categorization[J]. Machine Learning. 2000. 39: 135 - 168

[9] R. E. Schapire, Y. Singer, Improved Boosting Algorithms Using Confidence-rated Predictions[J]. Machine Learning. 1999. 38: 297 - 336

[10] Christopher D. Manning and hinrich Schutze. Foundations of statistical natural language processing[M]. Cambridge: MIT Press, 1999.

[11] Walker, E. Donald, Knowledge resource tools for accessing large text files. In: proc. First Conference of the UW Centre for the New Oxford English Dictionary: Information in Data[C]. Waterloo, Canada. Nov. 6 - 7, 1995.

[12] Yarowsky, David. Word-sense disambiguation using statistical models of Roget's categories trained on larger corpora[A]. ACL, 1992. 454 - 460

[13] 梅家驹,等. 多义词词林[M]. 上海: 上海辞书出版社, 1996.

[14] Zheng-Yu Niu. and Dong-Hong Ji. Optimizing Feature Set for Chinese Word Sense Disambiguation[A]. SENSEVAL-3: Third International Workshop on the Evaluation of Systems[C]. Barcelona, Spain, July, 2004.

[15] H. T. Ng, Getting Serious about Word Sense Disambiguation[A]. In: proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What and How?"[C], 1997.

[16] G. A. Miller, R. Beckwith, C. Fellbaum, et al. Five Papers on WordNet[J]. Special Issue of International Journal of Lexicography. 1990.

[17] 董振东. 知网[E13/OL]. http://www.keenage.com 2000.

[18] R. Mihalcea, I Moldovan. An Automatic Method for Generating Sense Tagged Corpora[A]. In: proceedings of the 16th National Conference on Artificial Intelligence[C], 1999.

[19] Eneko Agirre, Olatz Ansa, Eduard Hovy and David Martinez. Enriching Very larger ontologies using the WWW[A]. In: proceedings of the Ontology Learning Workshop[C], Berlin, 2000.