基于无指导机器学习的全文词义自动标注方法[1)]

卢志茂[1,2] 刘 挺[2] 李 生[2]

[1](哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)
[2](哈尔滨工业大学计算机学院信息检索研究室 哈尔滨 150001)
(E-mail: lzm@ir.hit.edu.cn)

摘 要 为实现汉语全文词义自动标注，本文采用了一种新的基于无指导机器学习策略的词义标注方法. 实验中建立了四个词义排歧模型，并对其测试结果进行了比较. 其中实验效果最优的词义排歧模型融合了两种无指导的机器学习策略，并借助依存文法分析手段对上下文特征词进行选择. 最终确定的词义标注方法可以使用大规模语料对模型进行训练，较好的解决了数据稀疏问题，并且该方法具有标注正确率高、扩展性能好等优点，适合大规模文本的词义标注工作.

关键词 词义标注，无指导学习算法，单纯贝叶斯模型，依存文法
中图分类号 TP391

Full-words Automatic Word Sense Tagging Based on Unsupervised Learning Algorithm

LU Zhi-Mao[1,2] LIU Ting[2] LI Sheng[2]

[1](*Computer Science & Technology School, Harbin Engineering University, Harbin* 150001)
[2](*Computer Science & Technology School, Harbin Institute of Technology, Harbin* 150001)
(E-mail: lzm@ir.hit.edu.cn)

Abstract For the purpose of implementing automatic Chinese word sense tagging, this paper presents a new method for word sense disambiguation based on unsupervised machine learning strategies. Four models of word sense disambiguation are built and compared. The model with two unsupervised machine learning strategies and selecting contextual features using dependence grammar obtains the best performance. And it can be trained with large-scale corpus to deal with the problem of data sparseness. In addition, it has such characteristics as high accuracy, high speed, easy extension and so on. Thus this technique is competent for word sense tagging on large-scale real-world text.

Key words Sense tagging, unsupervised learning algorithm, naive Bayesian model, dependency grammar

1 引言

词义标注是利用词义排歧 (Word sense disambiguation, WSD) 技术来解决如何在给定上下文语境中确定多义词词义 (Sense) 的问题. 词义排歧一直是自然语言处理 (NLP) 领域

1) 国家自然科学基金重点项目 (60435020) 和国家自然科学基金项目 (60575042, 60573072) 资助
Supported by the Key Project of National Natural Science Foundation of P. R. China (60435020) and National Natural Science Foundation of P. R. China (60575042, 60573072)
收稿日期 2004-5-24 收修改稿日期 2005-11-30
Received May 24, 2004; in revised form November 30, 2005

一个重要的热点研究问题，词义自动排歧在包括信息检索、文本挖掘、机器翻译、文本分类、自动文摘等在内的许多自然语言处理系统中都有重要的应用[1,2]．

统计学方法随着语料库语言学的兴起，以及良好的词义排歧效果受到自然语言处理领域的广泛关注，并且逐渐占据了主流地位．基于语料库的统计方法根据训练语料事先是否经过人工标注又可以分为有指导的和无指导的两类．有指导的机器学习方法在词义排歧问题上取得了较好的实验效果[3,4]，但是该类方法为了克服数据稀疏问题，获得更好的学习和排歧效果，必须有规模更大的标准语料库的支持．标准语料的获得需要大量的人工标注工作，很难实现基于大规模标准语料的有导词义排歧工作，客观上也限制了该类方法的推广和应用．无指导的词义排歧方法不依赖于人工标注的语料，可以实现跨领域大规模真实语料的训练和学习，能够有效克服数据稀疏问题，所以该类方法开始引起了研究人员的重视．无指导的词义排歧方法也很多，具有代表性的有双语对齐方法[5]、机器可读词典方法[6]和向量空间模型方法[7,8] 等等．

对大规模文本进行词义标注需要采用标注精度高、扩展性好的方法．本文介绍了基于贝叶斯模型和依存文法的汉语全文词义标注新方法，该方法采用无指导的机器学习策略，可以从大规模语料中自动获取词义排歧知识．由于无需人工标注标准语料，既可以省去人工标注语料所需的大量人工劳动，又可以获得较为完备的词义排歧知识，同时采用大规模训练语料，也避免了数据稀疏问题的发生，适合对大规模真实文本进行全文自动标注．另外，由于采用无指导的机器学习方法，训练语料的来源和规模可以根据标注的需要任意调整，使得训练出来的语言模型具有良好多扩展性，适合各个应用领域 (如，新闻、体育、财经、科技、卫生等等) 的文本词义标注．

2　基于贝叶斯模型的无指导学习

歧义词的词义是隐含的，只有通过具体的上下文语言环境才能表现出来，词义与上下文的关系可以进一步形式化为二元组 $(s_k, V_{context})$ 来表示一个词义，其中 $V_{context}$ 表示 s_k 的上下文词语的集合．在文本里，上下文可以是一个句子，也可以是一个句群，甚至是一个篇章．一般来说，确定一个多义词的词义，一个句子的上下文就足够了．本文用贝叶斯模型实现一种无指导的词义排歧算法用词义的自动标注任务．

2.1　单纯贝叶斯模型

单纯模型在词义排歧过程中把歧义词的上下文作为特征变量，并且基于两个假设，其一是把句子中歧义词的所有上下文作为一个 “词袋 (Bag of words)”，忽略句子中存在任何的语法、词义结构和词汇的线性顺序；其二是词袋中的词汇相互独立．

词义排歧是一个典型的分类问题，贝叶斯模型的词义分类原理是计算在特定上下文环境下歧义词每个词义出现的概率，并选择具有最大条件概率的词义作为最可能正确的词义，该原理可用公式表示为：

如果 $P(s'|V) > P(s_k|V)$, $s_k \neq s'$, 那么就选择 s' 为最可能的词义．

其中 V 是上下文环境 (即词语集合), s_k 是歧义词的任意词义变量，P 为概率．

由于特征变量 (即上下文中的词语) 之间是相互独立的，模型中词义变量 s_k 和特征变量集合 V 之间的联合概率可用下面的公式计算：

$$P(V, s_k) = P(V|s_k) \times P(s_k) = P(\{v_j|v_j \ in \ V\}|s_k) \times P(s_k) = P(s_k) \prod_{v_j \ in \ V} P(v_j|s_k) \qquad (1)$$

公式 (1) 中，v_j 是上下文 V 中的某一个词．

贝叶斯概率模型中，设 w 为歧义词，s_k 和 s' 分别是 w 的两个词义，其中 s' 是最终被确定的词义，V 是词 w 在句子中的上下文词语．

2.2 无指导的学习算法

本文使用的无指导学习方法包括两部分内容. 其一, 利用 HowNet 构造词义向量; 其二, 利用贝叶斯模型构造汉语无指导的词义分类器.

2.2.1 机器可读词典的处理和利用

汉语全文标注系统中使用知识库 HowNet(亦称作 "知网 ", 见 http://www.keenage.com) 给出每个词语的所有可能词义.

据统计, HowNet 中包含 17,000 多个概念, 词义刻画得可谓细致入微. 如果使用这样数量巨大的词义概念集进行词义标注, 计算量难免过大. 为了方便词义标注, 需要对 HowNet 中的概念进行适当分类, 用词义类构造词义标注集, 数量小, 便于计算.

在 HowNet 中, 把若干与概念有关的义原按一定的规则组合起来 (义原集合) 解释词义概念, 而这个义原集合被称为一个 "义项 ", 由 DEF 表示 (也就是一个词义概念), 如词语 "精确 " 的一个概念解释 "DEF=aValue| 属性值, correctness| 正误, accurate| 准, desired| 良 ". 对 HowNet 的词表和概念解释进行适当的聚类处理, 可以从中构造出大小适用的词义向量.

构造词义向量过程中根据具体情况对 DEF 进行不同处理, 如通常取义项中的第一个义原, 即 DEF 的主要特征. 当主要特征是 "属性 "、"属性值 "、"数量 "、"数量值 " 时, 还要取次要特征, 即第二个义原. 本文在 HowNet 中构造的语义向量只包含 1278 个词义类, 能够表示实际语料中词语的所有语义概念, 可以作为本文进行词义标注的依据. 为了便于操作, 本文为语义向量中词义类指定了 ID 号码, 该号码在词义聚类过程中根据先后顺序自动生成. 值得一提的是, 这样处理并不是对多义词的词义进行了合并, 而是合并了 HowNet 中的编号 (NO.), 词义与 DEF 一一对应, 处理前后多义词的词义数量没有增减.

2.2.2 基于贝叶斯模型的无指导学习策略

贝叶斯模型计算歧义词各个词义 (s_k) 在特定上下文环境 ($V_{context}$) 出现的条件概率 $P(s_k|V_{context})$ 作为词义决策依据. 在计算上, 需要知道先验概率 $P(s_k)$ 和条件概率 $P(v_j|s_k)$, 然而在没有标注词义的语料 (对于词义排歧任务来说, 该语料可以称为生语料) 中无从获得歧义词的词义信息, 也就无法计算出先验概率 $P(s_k)$ 和条件概率 $P(v_j|s_k)$. 即便是在已标注词义的语料中计算出的 $P(s_k)$ 值和 $P(v_j|s_k)$ 值也会随着语料规模的增减或来源的改变而发生变化, 所以这时获得概率是词义真实分布的近似值, 应该说其本质是对词义真实分布的一种统计估算. 语料的规模越大, 这种估计值越接近于真实值, 而语料的平衡性越好, 这个估计值也就越稳定. 这种统计上的近似给本文一个启示, 如果能够在无标注词义的语料中估算出 $P(s_k)$ 值和 $P(v_j|s_k)$ 值, 就获得了词义的近似分布, 从而用于对未知词义的判断和决策. 寻找一种方法, 能够从生语料中估算出先验概率 $P(s_k)$ 值和条件概率 $P(v_j|s_k)$ 值, 是本文要解决的任务.

一般认为生语料没有可供参考的词义标注信息, 谈不上获得词义的先验知识. 那么面对这样的生语料, 对于词义知识的获取似乎就无能为力了. 其实不然, 此时生语料仍包括很多可挖掘的词义信息, 歧义词的词义分布可以通过估算, 近似地获得.

语料中没有词义信息, 歧义词的词义分布是未知的, 根据信息论中的最大熵原理, 对于未确定的分布, 应该采取均匀分布. 所以, 为了估算歧义词的概率分布, 本文提出一个假设:

假设 1. 在某一个特定的上下文环境 ($V_{context}$), 歧义词 w 的每一个词义出现的机会均等, 在统计上它们的频次相等, 即:

$$c(s_1, V_{context}) = c(s_2, V_{context}) = \cdots = c(s_k, V_{context}) = c(w, V_{context})/k \tag{2}$$

公式中 $c(s_k, V_{context})$ 表示歧义词语 w 的一个词义 s_k 和其上下文集合 $V_{context}$ 在某一实例中共同出现的频度, $c(w, V_{context})$ 为歧义词出现的频度.

对语料进行无指导学习的关键就在于, 该假设带来的均匀分布是短暂的、很不稳定的状态, 很容易被打破. 当在训练语料中孤立地考察某一个词的时候, 这种均匀的词义分布

状态才会存在. 而孤立的看待某一个词在大规模语料的统计上没有意义. 在机器学习过程中, 当统计到句子中某一个词时, 它的词义虽然采取了均匀分布, 但是语料中会有很多词语具有某个或某些相同的语义, 并且这些词义以不同的频率重复出现. 语料中数万个不同的词形, 其对应的词义类只有 1278 个, 词形重复出现的频率远不如词义重复出现的频率高, 这会导致最终的统计结果中歧义词的词义分布不但不再均匀, 甚至很不均衡, 而这种词义分布的不均衡正是真实文本的自然表现. 举个形式化的例子来说明如下:

设 w_i, w_j 是训练语料中出现在不同句子中的任意两个歧义词 (不互为上下文), 根据所给条件计算各词义出现的频度.

令 $c(w_i)=m$, $c(w_j)=n$; $i,j,m,n=1,2,3,\cdots$

w_i 有 3 个词义 s_{i1},s_{i2},s_{i3}; w_j 有 2 个词义 s_{j1},s_{j2};

则 $c(s_{i1})=c(s_{i2})=c(s_{i3})=m/3$; $c(s_{j1})=c(s_{j2})=n/2$;

如果 $s_{i3}=s_{j1}$;

则 $c(s_{i3})=c(s_{j1})=m/3+n/2$;

由于 s_{i3},s_{j1} 是相同的词义, w_i 或者 w_j 的再次出现, 增加了词义 $s_{j1}(s_{i3})$ 出现的几率, 对于 w_i,w_j 来说, 其各个词义出现的频度不再相等, 均匀分布的初始状态发生了改变. 随着语料规模的增大, 参与训练的句子增多, 这种变化将越来越显著, 最终导致歧义词的词义分布趋于真实分布 (即不均衡状态).

如果语料规模足够大, 词义和上下文词语共现的次数能表现出某种词语间固定搭配的关系, 例如形容词和名词、副词和动词、动词和名词之间等的修饰和搭配关系. 词义和上下文词语的搭配关系如果强, 其共现次数就多, 否则共现次数就少. 搭配关系的强弱反映了上下文对歧义词词义约束力的强弱, 可以帮助对词义进行正确判断. 通过对大规模语料的统计学习, 可以得到词义与其上下文词语的搭配关系, 从而自动获取汉语自动词义标注的知识.

2.2.3　具体统计学习方法

首先为参数建立二维表, 其中行数 M 等于词义的总个数, $0<M\leqslant 1278$; 列数 N 为词语的总个数, $0<N\leqslant 53335$. 第 1 列记录各个词义出现的次数 $c(s_k)$, 第 2 列到 $N+1$ 列记录 s_k 同其上下文词语 v_j 共现的次数 $c(s_k,v_j)$, 其中 $0<k\leqslant M, 0<j\leqslant N$, 进行统计学习之前将表中的各元素用 0 进行初始化.

对于某个句子出现的词语 w_i, 令它的词义为 $\{s_{i1},s_{i2},\cdots,s_{ih}\}$, 上下文为 $\{v_{i1},v_{i2},\cdots,v_{il}\}$, 其中 $i=1,2,3,\cdots,0<h<M,0<l<N$.

如果 w_i 在句子出现 1 次, 则有: $w_i: c(s_{i1})=c(s_{i2})=\cdots=c(s_{ih})=1/h$, w_i 在表中对应的 h 个词义的出现次数都加 $1/h$; $c(s_{ih},v_{i1})=c(s_{ih},v_{i2})=\cdots=c(s_{ih},v_{il})=1/h$, 表中 s_{ih} 对应的上下文出现的次数都加 $1/h$.

对大规模语料的统计结束后, 就可以利用表中的数据分别计算概率 $P(s_k),P(v_j|s_k)$, 建立词义标注阶段需要的 "$P(s_k)-P(v_j|s_k)$" 参数表.

$P(s_k),P(v_j|s_k)$ 的计算公式如下:

$$P(s_k)=\frac{c(s_k)}{\sum\limits_{k=1}^{m}c(s_k)} \tag{3}$$

$$P(v_j|s_k)=\frac{c(s_k,v_j)}{c(s_k)} \tag{4}$$

对于歧义词 w 的每个词义, 利用贝叶斯公式分别计算 Score, 并比较大小以确定最合适词义.

2.2.4　无指导学习中的 EM 算法

为了比较上述的无指导机器学习的实验效果，本文同时采用 EM 算法[7] 构造了词义排歧模型，并进行对照实验.

在生语料中也可以用 EM 算法对词义分布进行估算，从而实现无指导的机器学习方法. 初始的时候随即初始化概率参数 $P(v_j|s_k)$, 然后根据 EM 算法重新估计 $P(v_j|s_k)$. 在随即初始化之后，上下文就有了初始分类，该结果可以作为下一步迭代计算的训练数据，然后重新估计参数 $P(v_j|s_k)$, 使得模型给定数据的似然值最大. EM 算法保证每一步计算中模型似然对数的值是增加的，所以该迭代计算的终止条件是似然值不再明显的变化 (增加).

基于贝叶斯模型的 EM 算法估算 $P(v_j|s_k)$ 和 $P(s_k)$ 的过程如下：

1) 初始化

以任意值初始化参数 $P(v_j|s_k)$ 和 $P(s_k)$; $1 \leqslant j \leqslant J$; $1 \leqslant k \leqslant K$. 初始化后，概率值 $P(s_k)$ 和 $P(v_j|s_k)$ 对词语的词义和上下文进行了预分类，由此可以计算出上下文 $C_{context}$ 在给定语言模型 model 条件下的概率 $P(C|model)$. 由于单个上下文之间是条件独立的，$P(C|model)$ 等于单个上下文概率 $P(c_i)$ 的乘积，$1 \leqslant i \leqslant I$.

$$P(C|model) = \prod_{i=1}^{I} P(c_i) = \prod_{i=1}^{I} \sum_{k=1}^{K} P(c_i|s_k)P(s_k) \tag{5}$$

$$\log[P(C|model)] = \log \prod_{i=1}^{I} P(c_i) = \log \prod_{i=1}^{I} \sum_{k=1}^{K} P(c_i|s_k)P(s_k) = \sum_{i=1}^{I} \log \sum_{k=1}^{K} P(c_i|s_k)P(s_k) \tag{6}$$

2) E - 过程

根据 c_i 和 s_k 计算条件概率 $P(c_i|s_k)$.

$$h_{ik} = \frac{P(c_i|s_k)}{\sum\limits_{k=1}^{K} P(c_i|s_k)} \tag{7}$$

根据贝叶斯假设,

$$P(c_i|s_k) = \prod_{v_j \ in \ c_i} P(v_j|s_k) \tag{8}$$

3) M - 过程

根据最大似然估计重新计算 $P(v_j|s_k)$ 和 $P(s_k)$,

$$P(v_i|s_k) = \frac{\sum\limits_{i=1}^{I} \sum\limits_{c_i:\ v_j \ in \ c_i} h_{ik}}{Z_j} \tag{9}$$

其中 $\sum\limits_{c_i:\ v_j \ in \ c_i}$ 是包含 v_j 的上下文总和，Z_j 是归一化常数,

$$Z_j = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{c_i:\ v_j \ in \ c_i} h_{ik} \tag{10}$$

$$P(s_k)=\frac{\sum_{i=1}^{I}h_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{I}h_{ik}} \tag{11}$$

当 $P(C|model)$ 的值不再增大，停止迭代计算．利用计算出的 $P(v_j|s_k)$ 和 $P(s_k)$ 构造贝叶斯模型.

$$s(w_i)=\arg\max_{s_k}\Big[\log P(s_k)+\sum_{v_j\in C_i}\log P(v_j|s_k) \tag{12}$$

2.3　特征选择

Bayesian 模型在词义分类上，选择歧义词的上下文作为特征变量，并且 Naïve Bayesian 模型里没有特征选择，把整个句子中歧义词的所有上下文词语都无一例外的作为特征变量．虽然歧义词在当前句子中选择哪种词义和其左右出现的上下文有关系，但并不是所有的上下文词语都对歧义词词义变量的取值有约束力．词语之间的固定搭配有强弱之分，词义之间的约束力自然也有大小之别.

事实上句子中词语之间存在着某些固定搭配和语法关联，而贝叶斯假设却忽略了这些语法上的信息．为了弥补 Naïve 模型的不足，本文利用依存文法 (Dependency grammar, DG)[9] 来确定歧义词与其上下文中词语间存在的依存关系，选择与歧义词有依存关系的上下文词语为模型的特征向量 $\boldsymbol{v}(w_i)$, 其中 w_i 表示特征词出现在句子中第 i 个位置.

依存语法通过分析语言单位内各成分之间的依存关系揭示其句法结构，主张句子中动词是支配其他成分的中心成分，而它本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配成分．本实验对语料的依存分析结果示例如下 (例句先经过分词和词性标注).

依存分析结果用三元组表示为 $(A_i,\leftarrow,A_j)$, 其中 A_i 是支配成分，A_j 是受支配成分，直接依存于 A_i. 该例句的依存分析结果如下：

例句：[1] 小丽 /nm [2] 说 /vg [3] 她 /r [4] 昨天 /t [5] 给 /p [6] 我 /r [7] 打 /vg [8] 了 /ut [9] 一个 /m [10] 电话 /ng [11]./wj

([1] 小丽，→, [2] 说)　([2] 说，←, [7] 打)　([3] 她，→, [7] 打)　([4] 昨天，→, [7] 打)
([5] 给，→, [7] 打)　([5] 给，←, [6] 我)　([7] 打，←, [8] 了)　([7] 打，←, [10] 电话)
([9] 一个，→, [10] 电话)

依据依存文法分析的结果，该例句中我们选择 { 说，她，昨天，给，了，电话 } 作为歧义词 “打 ” 的上下文特征变量，即 $\{v(w_2),v(w_3),v(w_4),v(w_5),v(w_8),v(w_{10})\}$; 选择 { 小丽，打 } 为歧义词 “说 ” 的特征变量；选择 { 打，我 } 为歧义词 “给 ” 的特征变量．其中 “电话 ” 和 “打 ” 的语义密切相关，出现在 “打 ” 右边第 4 个位置上，而距离更近的词语 “我 ” 和 “一个 ” 被此法排除在外．经过这样的处理，歧义词 “打 ” 的特征变量从 9 个 (标点除外) 减少到 6 个，只考虑实词，特征变量进一步减少到 4 个．因为虚词对歧义词词义的判断作用不大，在实验过程中要先滤掉虚词，然后再提取特征词.

句子中各成分拥有的依存关系数量差别很大，比如动词成分同其他成分构成的依存关系一般较多，名词、形容词、副词等其他成分存在的依存关系数少一些，最少的就一个.

3　实验与结果讨论

3.1　实验与结果

为了测试该系统的词义标注的效果，实验中采用四种方法构造词义排歧模型，模型 1 采用单纯贝叶斯模型 (NB); 模型 2 采用贝叶斯模型与依存文法分析相结合的方法 (NB +

DG); 模型 3 采用 EM 算法与贝叶斯模型相结合的方法 (EM + NB); 模型 4 采用贝叶斯模型、 EM 算法与依存文法分析相结合的方法 (NB + EM + DG).

本文所使用的单纯贝叶斯模型 (模型 1: NB) 是按照前文 2.2.3 介绍的方法构建的. 模型 2 在模型 1 的基础上采用依存文法分析进行上下文特征选择. 模型 3 的概率参数估计采用 2.2.4 介绍的 EM 算法, 语言模型采用贝叶斯模型.

模型 4 的实现过程比前 3 个模型略为复杂, 参数估计分两个过程. 第一过程, 采用模型 1 的方法统计并计算出参数 $P(v_j|s_k)$ 和 $P(s_k)$; 第二过程, 把第一过程得到的参数 $P(v_j|s_k)$ 和 $P(s_k)$ 作为 EM 算法的初值进行迭代计算, 最终确定参数 $P(v_j|s_k)$ 和 $P(s_k)$ 的值. 模型 4 中也采用了依存文法分析方法对上下文特征词进行了选择.

本文的实验有两个目的, 其一探讨训练语料的规模对测试结果的影响; 其二比较各个模型的排歧性能. 为此, 本文设计了多组实验, 第一组采用 30 万词的训练语料, 分成两块, $C_{train-1}$(包含 15 万词) 和 $C_{train-2}$(包含 30 万词), 并且 $C_{train-1} \in C_{train-2}$. 测试语料为 5000 个自然句 (包含 38,928 个词语, 事先进行人工词义标注和校对作为标准评价语料), 实验结果见表 1.

表 1 四个模型的实验结果对照

Table 1 The comparison of the experimental results in four models

训练语料	词义标注的正确率 (%)sense-tagging accuracy			
	模型 1	模型 2	模型 3	模型 4
$C_{train-1}$	74.69	85.02	75.12	86.98
$C_{train-1}$	76.97	86.11	75.82	87.26
备注	测试语料为 5000 句, Baseline72.59%			

第二组采用包含 400 万词超大规模的训练语料对模型 1 进行测试, 探讨训练语料规模影响测试结果的规律, 参见表 2.

表 2 模型 1 的实验结果

Table 2 The experiment result of any size training corpus – training model 1

训练数据 (万词)	2.0	5.0	10.0	20.0	100.0	400.0
正确率 (%)	71.92	73.52	74.15	75.36	80.01	79.92
备注	测试语料为 5000 句, Baseline72.59%					

实验中为了更客观的说明文中实现的几个模型的性能, 需要有一个参照系, 即所谓的 Baseline. 由于目前对于汉语的词义标注, 还没有统一的词典资源和评价语料, 许多研究结果没有可比性. 所以本文参照文献 [10] 建议的方法, 即使用最简单的词义标注方法得到的实验结果作为 Baseline. 歧义词的各个词义在特定的语料中出现的频率大小不同, 选择最大概率词义是最简单的词义标注方法, 可以作为 Baseline. 本文使用的 5000 句测试语料具有人工标注的词义, 可以统计出歧义词各个词义的分布情况, 以此为据, 计算出 Baseline 的实验结果, 参见上面的实验数据.

第三组采用规模按照等差递增的语料 (将 30 万词的语料分成大小不等的 6 块) 训练模型 2 和模型 4, 进一步考察训练语料对这两个模型的影响, 参见表 3.

表 3 语料规模对模型的影响

Table 3 The influence of the size of training corpus on WSD model

训练数据 (万词)		5.0	10.0	15.0	20.0	25.0	30
正确率 (%)	模型 2	84.04	84.54	85.02	85.46	86.02	86.11
	模型 4	84.63	85.35	86.98	87.06	87.14	87.26
备注	测试语料为 5000 句, Baseline72.59%						

3.2 结果分析与讨论

表 1 中的数据是第一组实验的结果，由该表可以看出 4 个模型的实验结果都高于 Baseline, 说明本文采用的无指导学习策略是切实可行的. 其中模型 1 采用简单的无指导学习策略，正确率能够达到 76.97%, 高出 Baseline 4 个百分点，并且该模型随着训练语料的增大，排歧正确率显著提高，说明大规模的训练语料对改善该模型的排歧性能是很有利的. 模型 2 在模型 1 的基础上进一步采用依存文法对上下文特征词进行选择，收效十分显著，增幅近 9 个百分点，同时也证明了特征选择对改善贝叶斯模型的词义分类效果作用很大. 诚然，歧义词的词义判断需要上下文词义的支持，但并不是说所有的上下文词语都对歧义词的词义有约束作用. 一般来说，只有那些与歧义词构成语义搭配的上下文词才会提供有用信息，其他的词非但没有帮助，还容易形成噪声影响. 借助语法分析，找到词语间的依存关系，可以确定与歧义词构成搭配关系的上下文词语，以此作为特征选择的依据. 进行特征选择的模型，有效缩小了上下文窗口，一方面减少了计算量，降低了算法的时间复杂度；另一方面消除了一些噪声影响，提高了排歧的精度.

模型 3 只是采用了基于 EM 算法的无指导学习策略，实验结果和模型 1 的相差无几. 在用语料 $C_{train-1}$ 训练时要比模型 1 略强，在用语料 $C_{train-2}$ 训练时要比模型 1 略差. 模型 3 在无指导机器学习上效果和模型 1 的效果在仲伯之间，差别不大. 然而，由于 EM 达到的是局部最优，增大训练语料的规模虽然也有利于改善学习效果，但是却没有模型 1 采用的学习策略改善的幅度大. 说明，在训练数据规模很大的情况下，模型 1 的无指导学习策略要略优于模型 2 的.

模型 4 的词义排歧效果最好. 该模型综合使用了前文介绍的多种方法，词义排歧的能力要明显优于前三个模型. 前文 2.2.3 介绍的无指导学习策略是建立在假设 $c(s_1, V) = c(s_2, V) = \cdots = c(s_k, V)$ 之上的，该方法简单易行，但为了提高各词义区分度，需要大幅度提高训练语料的规模. EM 算法在训练语料规模固定的条件下可以通过多次的迭代计算寻找最优解. 模型 4 综合了两种学习策略的优点，将模型 1 计算出的参数 $P(v_j|s_k)$ 和 $P(s_k)$ 作为 EM 算法的输入，避免 EM 算法陷入局部最优，通过多次迭代计算最终确定参数 $P(v_j|s_k)$ 和 $P(s_k)$, 同时模型 4 也借助了依存文法分析手段对上下文进行了选择，实验结果比模型 2 又有明显的提高.

第二组实验 (实验结果参见表 2) 使用了不同规模的训练数据，测试结果的正确率开始一路攀升，训练数据到达 400 万词的超大规模时，正确率反而略有下降. 第三组实验 (实验结果参见表 3) 使用的训练数据在 5 万词 – 30 万词之间，规模不是很大，测试结果的正确率随着语料规模的增大而增大. 总的来说，训练语料的规模越大，所包含的语言学现象也就越丰富，蕴涵的语言学知识也就越多，词义排歧模型可以获得的有效信息的数量也就越大. 但是，当来自某特定领域的语料规模增大到一定规模，其包含的语言现象和知识就会趋于饱和，再增大规模已经毫无意义，有时甚至适得其反. 原因就在于，大语料在增加有效信息的同时，无效信息也同时增加. 当有效信息不再随着语料规模增大而增多的时候，无效信息仍然会增加，造成的干扰还会加强，从而对排歧模型造成严重的负面影响，直接导致正确率下降.

4 结束语

本文针对全文词义标注任务的特点，采用了一种新的无指导词义排歧算法. 实验中采用不同的机器学习策略，分别建立了四个模型用于词义排歧. 通过实验对比发现文中采用 2.2.3 介绍的学习策略、2.2.4 介绍的 EM 算法和依存文法建立的排歧模型具有较高的词义排歧正确率，适合对大规模文本进行词义标注工作.

由于词义排歧算法采用无指导的机器学习方法，可以从大规模语料中自动获取词义排歧所需的知识，可以在很大程度上克服有指导学习难以克服的数据稀疏的问题. 因为可以

任意扩大训练语料的规模，本实验系统适合各个专业领域 (如行政、司法、新闻、科技等领域) 文本的词义标注工作，使得该方法具有良好的领域扩展性.

借助依存文法分析手段进行特征提取具有较高的特征词压缩比，有效缩短了词义判断的时间，提高了词义标注效率.

词义自动标注是自然语言理解的一个重要环节，高效准确的词义标注方法对机器翻译、信息检索、文本分类等诸多问题都会起到积极的促进作用，另外本文实现的词义标注方法对于完成大型语料库建设工程中的文本词义标注任务更具有重要意义.

References

1 Qin B, Liu T, Li S. Multi-document summarization based on local topics identification and extraction. *Acta Automatica Sinica*, 2004, **30**(6): 905∼910

2 Nancy I, Jean V. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 1998, **24**(1): 1∼40

3 Purandare A, Pedersen T. Word sense discrimination by clustering contexts in vector and similarity spaces. In: Proceedings of the Conference on Computational Natural Language Learning, Boston, MA: Association for Computational Linguistics, 2004. 41∼48

4 Pederson T. A simple approach to building ensembles of naïve Bayesian classifiers for word sense disambiguation. In: Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA: Association for Computational Linguistics, 2000. 63∼69

5 Dagan I, Itai A. Two languages are more informative than one. In: Proceedings of the 29th Annual Meeting of Association for Computation Linguistics. Berkeley, CA: Association for Computational Linguistics, 1991. 130∼137

6 Yarowsky D. Word sense disambiguation using statistical methods of Roget′s categories trained on large corpora. In: Computation Linguistic′92. Nantas: Association for Computational Linguistics, 1992. 454∼460

7 Schutze H. Automatic word sense discrimination. *Computational Linguistics*, 1998, **24**(1): 97∼124

8 Lu S, Bai S, Huang X, Zhang J. Supervised word sense disambiguation based on vector space model. *Journal of Computer Research & Development*, 2001, **38**(6): 662∼667

9 Jason E. Three new probabilistic models for dependency parsing: An exploration. In: Proceedings of 16th International Conference on Computational Linguistics (COLING-96), Copenhagen: Association for Computational Linguistics, 1996. 340∼345

10 Manning C D, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts, London, England: The MIT Press, 1999. 229∼260

卢志茂 哈尔滨工业大学博士生，哈尔滨工程大学副教授，从事词义消歧，自然语言处理等研究工作.

(**LU Zhi-Mao** Ph. D. candidate in School of Computer Science and Technology at Harbin Institute of Technology, associate professor in School of Computer Science and Technology at Harbin Engineering University. His research interests include word sense disambiguation and natural language processing.)

刘 挺 哈尔滨工业大学计算机学院，教授，从事自然语言处理，信息检索等研究工作.

(**LIU Ting** Professor in School of Computer Science and Technology at Harbin Institute of Technology. His research interests include natural language processing and information retrieval.)

李 生 哈尔滨工业大学计算机学院，教授，博士生导师，从事自然语言处理，机器翻译等研究工作.

(**LI Sheng** Professor in School of Computer Science and Technology at Harbin Institute of Technology. His research interests include natural language processing and machine translation.)