基于信息增益改进贝叶斯模型的汉语词义消歧

范冬梅[①]　卢志茂[①]　张汝波[①]　潘树燊[②]
[①](哈尔滨工程大学计算机学院　哈尔滨　150001)
[②](哈尔滨工业大学　哈尔滨　150001)

摘　要：词义消歧一直是自然语言处理领域的关键问题和难点之一。通常把词义消歧作为模式分类问题进行研究，其中特征选择是一个重要的环节。该文根据贝叶斯假设提出基于信息增益的特征选择方法，并以此改进贝叶斯模型。通过信息增益计算，挖掘上下文中词语的位置信息，提高贝叶斯模型知识获取的效率，从而改善词义分类效果。该文在 8 个歧义词上进行了实验，结果发现改进后的贝叶斯模型在消歧正确率上比改进前平均提高了 3.5 个百分点，改进幅度较大，效果突出，证明了该方法的有效性。

关键词：词义消歧；自然语言处理；信息增益；贝叶斯模型
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2008)12-2926-04

Chinese Word Sense Disambiguation Based on Bayesian Model Improved by Information Gain

Fan Dong-mei[①]　Lu Zhi-mao[①]　Zhang Ru-bo[①]　Pan Shu-shen[②]
[①](*Harbin Engineering University*, *Harbin* 150001, *China*)
[②](*Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: Word Sense Disambiguation (WSD) is one of the key issues and difficulties in natural language processing. WSD is usually considered as an issue about pattern classification to study, which feature selection, is an important component. In this paper, according to Naïve Bayesian Model (NBM) assumption, a feature selection method based on information gain is proposed to improve NBM. Location information concealed in the context of ambiguous word is mined through information gain, to improve the knowledge acquisition efficiency of Bayesian model, thereby improving the word-sense classification. The eight ambiguous words are tested in the experiment. The experimental results show that improved Bayesian model is more correct than the NBM an average of 3.5 percentage points. The accuracy rise is bigger and the improvement effect is outstanding. These results prove also the method put forward in this paper is efficacious.

Key words: Word sense disambiguation; Natural language processing; Information gain; Naïve Bayesian model

1　引言

语言中的歧义现象一直困扰着信息处理技术的研究和发展。词义消歧(Word Sense Disambiguation, WSD)技术用来解决如何在给定上下文语境中确定歧义词词义(sense)的问题。词义消歧一直是自然语言处理(natural language processing)领域一个重要的热点研究问题[1]，词义自动消歧在包括信息检索、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。

统计词义消歧(Statistical Word Sense Disambiguation, SWSD)方法[2]，运用统计学技术手段自动在训练语料中获取所需的知识，如歧义词与上下文词语之间的语法关系或语义关系等，并将这些“知识”用于词义的识别和判断。很多常见的机器学习方法，如决策树、贝叶斯模型、神经网络、支持向量机、最大熵、遗传算法等，都在统计词义消歧上获得了成功的应用。

统计学方法给 WSD 带来的推动作用引起了自然语言处理领域的广泛关注，并且逐渐成为词义消歧的主要研究方法。统计词义消歧需要借助统计的手段在语料库或者知识库中发现和获得词义信息或知识，这里有两个重要的入手点，即统计方法(机器学习)和知识源(语料库或词典)，既要有好的学习手段，又要有好的知识源[3]。

SWSD 借助统计学习的方法从语料库中自动获取语言信息，自动学习词义判断所需的知识。语料有标注词义的、未标注词义的两种，这两种都可以作为 SWSD 的训练数据。训练数据是否带有词义标记，决定了词义消歧实现的思想和方法的差异，也把 SWSD 分成了有指导的词义消歧方法和无指导的词义消歧方法[4]。

2007-06-04收到，2008-09-23改回
国家自然科学基金(60575042，60603092)和国家教育部博士点专向基金(20070217043)资助课题

2　相关工作

2.1 特征选择与词义消歧

特征选择，是指从已知一组特征集中按照某一准则选择出有很好的区分性的特征子集，或按照某一准则对特征的分类性能进行排序，用于分类器的优化设计。

词语间的相互作用于上下文中的位置和物理距离有直接的关系。为了表征上下文词语与歧义词之间关系，引入语言知识可以帮助选择对词义判断更有帮助的词语充当特征词，也可以通过统计的方法计算语义相关度来选择特征词语。

2.2 机器学习模型的改进

通过特征选择来改善分类算法的效果是研究分类问题的一个重要的入手点。本文选择贝叶斯模型进行改进，是因为该模型是一种十分常用的分类方法，广泛应用于模式识别领域，同时在计算上它又简洁、高效，容易实现。对比其他经典的分类算法，例如 BP 神经网络、支持向量机(SVM)、决策树(DT)和最大熵方法(ME)，贝叶斯模型也毫不逊色，甚至会超越。

上下文词语对歧义词词义的约束集中体现在特定的句法关系和词语搭配上，文献[5]介绍把基于依存分析的句法分析技术用于汉语词义消歧的特征选择，获得了较好的效果。

通过依存文法分析对贝叶斯模型的特征进行了选择，实验结果表明其方法行之有效。依存文法分析手段找到了那些与歧义词构成强搭配的上下文词语，从而在计算上剔出了大量的无用信息，不仅提高了计算效率，同时也改善了消歧效果。但是该方法没有考虑词语的位置信息，对于选择出来的特征词语一视同仁，没有区分对词义判断的贡献大小。而且自动依存文法分析过程也许要占用很大的时间开销，从整体上增大词义消歧系统的时间复杂度。

本文力求在不增大太多计算量的前提下改善词义消歧的精度，尝试使用信息增益计算为贝叶斯模型进行特征选择，根据实验结果证实该方法在汉语词义消歧上收效显著。

3　贝叶斯模型的改进思想

单纯贝叶斯模型(Naïve Bayesian Model, NBM)是求解在给定条件下决策事件的最大条件概率，用于歧义消解就是在上下文窗口中考虑特征词语对歧义词词义的决策作用。根据贝叶斯模型，正确词义在给定的上下文环境里出现的概率(后验概率)最大，即贝叶斯决策规则形式如下：

if $P(s' \mid C_{\text{context}}) > P(s_k \mid C_{\text{context}})$, $s_k \neq s'$ then decide s'

其中 C_{context} 是上下文环境(即词语集合)，s_k 是歧义词的任意词义变量，s' 是正确词义，P 为概率。贝叶斯决策规则具有最小的误差概率。

3.1 贝叶斯假设

为了计算似然函数 $P(C_{\text{context}} \mid s_k)$ 的值，需要用到贝叶斯假设(Bayesian Assumption)。贝叶斯假设是指刻画事物特征的属性彼此条件独立，用于词义判断，即假设各个上下文特征词语是相互条件独立的。因此，$P(C_{\text{context}} \mid s_k)$ 可以计算如下：

$$P(C_{\text{context}} \mid s_k)=P(\{v_j|v_j \text{ in } C_{\text{context}}\}|s_k)=\prod_{v_j \text{ in } C_{\text{context}}} P(v_j|s_k) \quad (1)$$

其中 v_j 是上下文 C_{context} 中的第 j 词语。

贝叶斯词义分类器的建立和实现是建立在贝叶斯假设的基础之上的，这个假设虽然不符合语言的使用规律和实际情况，但是尽管如此，该假设不仅保证模型的顺利建立，也获得了出人意料的分类效果。贝叶斯模型为了计算上的需要提出上述假设，来忽略实际语言应用中词语间的相互依赖关系，该假设不仅保证模型得以顺利实现，也为模型的进一步改进埋下了伏笔。本文也通过实验发现，对 NBM 的特征参数集进行优化和选择会收到良好的改进效果。

一般认为，词语间形成的固定搭配都有着相对稳定的位置关系，所以通过位置信息，可以在某种程度上衡量上下文词语的作用。如果在贝叶斯模型的计算上加入位置信息，那将会有助于改善模型的词义辨识能力。那么，如何计算词语的位置信息呢？前文介绍的最大熵方法能够估算特征词语的权重，可以很好地衡量词语间作用的强与弱，但是该方法在计算上略显复杂，如果引入贝叶斯模型，将会大幅度加大模型运行的时间开销。本文试图利用信息增益(Information Gain, IG)的计算方法对词语位置信息进行量化，以期获得对贝叶斯模型的改良。

3.2 基于信息增益最大原则的改进思想

1986 年，Quinlan 的以信息熵作为启发函数的决策树归纳学习算法 ID3，采用信息增益最大原则进行特征的选择[6]。信息增益(IG)是信息论中比较重要的一个计算方法，该方法能够估算系统中新引入的特征所带来的信息量，即信息的增加量。通过该计算，实现了 ID3 算法，使得决策树分类方法获得改良，并得到了广泛应用。

本文根据 NBM 的实现条件和特点，决定引入特征词语的位置信息，以便提高模型的词义辨识能力。该位置信息借助信息增益来加以量化，并根据信息增益最大原则对特征集进行优化，增加那些对歧义词词义判断贡献更大的上下文词语的权重，以突出它们对词义判断的作用。使用 IG 衡量位置信息对词义判断的贡献，获得实验所需的位置权重值，进而达到特征优化的目的。

3.2.1 信息增益的计算方法　采用熵的方法可以估算系统在添加新信息前后的不确定性，通过不确定性的变化可以估算信息的增益程度。首先把上下文中的每一个词作为一类，计算整个上下文环境(Context)的统计不确定性，即熵 $H(\text{Context})$。然后计算在已知某个相对位置的前提下，整个上下文环境的不确定性，即条件熵 $H(\text{Context} \mid V_p)$，$V$ 表示某个指定位置出现的词语集合。这里熵和条件熵的计算方法如下：

$$H(\text{Context}) = -P(\text{Context}) \times \log P(\text{Context}) = -\sum_{v \in \text{context}} P(v) \times \log P(v) \quad (2)$$

$P(v)$ 是词语 v 在训练语料中出现的频率，计算方如下：

$$P(v) = \frac{C(v)}{\sum_j C(w_j)} \quad (3)$$

其中 $C(w_j)$ 为语料库中的某个词语出现的频度。条件熵 $H(\text{Context} \mid V_p)$ 按照式(4)计算。

$$H(\text{Context} \mid V_p) = \sum_{v_j \in V_p} P(v_j) \times H(\text{Context} \mid v_j) \quad (4)$$

其中 $P(v_j)$ 是词语 v_j 在指定位置出现的概率。

式(2)与式(4)的差值显示了该指定位置为上下文环境提供的信息量，即代表了信息的增益量，所以信息增益(IG)可以计算如下：

$$\text{IG}_p = H(\text{Context}) - H(\text{Context} \mid V_p) \quad (5)$$

根据式(5)计算的信息增益是针对特定的实验语料进行的，该值取决于语料库和歧义词的选择。如果把信息增益作为位置权重，需要假设上下文词语对歧义词词义的约束能力由远及近逐渐增大，即距歧义词最近的词语作用最大，这符合人类的一般认知过程。计算结果和大量基于统计学习的实验都验证了该假设是可行的。

综上所述，本文试图通过式(5)计算出来的 IG 可以直接作为上下文词语的权重，并引入到基于贝叶模型的词义消歧模型上，以期获得改良效果。

3.2.2 贝叶斯模型的改进 将上下文位置信息引入单纯贝叶斯模型(Naïve Bayesian Model, NBM)，位置权重作为模型参数参与计算。位置权重采用前文计算出的信息增益量(IG)，词义的决策规则修改如下：

$$s' = \arg\max_{s_k} \left[\log P(s_k) + \sum_{v_j \text{ in } C_{\text{context}}} \log P(v_j \mid s_k) \times \text{IG}_j \right] \quad (6)$$

其中 IG_j 是上下文词语 v_j 出现位置的权重，体现了位置信息对词义判断所起的作用。根据信息增益量的高低，可以为特征词语附加位置信息，为词义判断提供更多的帮助。通过信息增益的计算，在模型的实现上不仅突出了特征词语的位置信息，同时又收到了特征提取的效果。

4 实验设计与结果分析

为了验证前文提出的统计词义消歧改进方案，本文只须将改进前后的模型在词义消歧实验上进行对比就可以了。所以在实验上设计上，本文只关注实验结果的相对值，即此模型改进前后消歧正确率的增加幅度。

4.1 实验的消歧对象

本文采用有指导的贝叶斯模型建立词义分类器，进行实验比较分析消歧模型的改进方案。为了实验结果不受人为参与的影响，本实验采用人造歧义词技术构造训练数据和测试数据。

早在 1992 年，Schütze 在各自的论文中就分别介绍了人造歧义词(伪词)的构造和使用方法。伪词(pseudoword)是按照一定规则把多个词语组合在一起构成人造歧义词[7]。构成伪词的每一个词语代表伪词的一个词义，所以这个词语必须是单义的，否则又会出现二义性。例如伪词 w_q(w_q | {计算，指导})由单义词语“计算”和“指导”组合而成，具有两个词义，即词义 S_1=“计算”，S_2=“指导”。

本文的实验目的是比较词义消歧模型改进前后的实现效果，重点不在于歧义词本身消歧难度的大小，也不在于训练语料规模的大小，只须实验语料具有很好的一致性就可以了。因此，实验中本文选择了使用人造歧义词。在构造人造歧义词时，选择若干不同词性的实词(必须是单义词)，此番构造的伪词词义变化范围在 2~5，考虑了更多的情况，参见表 1。

表 1 人造歧义词的构造表

伪词	词义数	伪词素
w_1	2	人民/政府
w_2	2	干部/群众
w_3	2	每/重要
w_4	2	亿/技术
w_5	3	全国/人民/政府
w_6	3	群众/亿/技术
w_7	4	每/重要/群众/亿
w_8	5	每/重要/群众/亿/技术

4.2 实验设计方案

因为在实验中使用了伪词技术，所以实验数据的规模容易做得很大。本文从人民日报 1998 年的电子版抽取实验语料，为每个伪词提供 7,000 个实例，其中 5,000 个作为训练数据，2,000 作为测试数据。人造歧义词的词义频度比控制在 1:1，这样，歧义词的训练规模的大小因歧义词的词义数量不同而不同。其中 w_8 最少，但也保证每个词义有 1,000 个训练实例。

上下文窗口的控制按照(−10，+10)，如果歧义词的左边或者右边的词语数量不足 10 个，则以空位(NULL)补齐，不另外扩大取词范围。

计算上要考虑位置信息，位置信息只取决于距离歧义词(w_q)的相对位置，例如($-n$，…，−2，−1，w_q，+1，+2，…，$+m$)，不管哪个词语，只要出现在同一个位置上，就使用同一个位置权重，这个权重使用式(4)计算。

没有做特征优化的 NBM 作为本文实验的参照模型，实验结果参见表 2。

表中 BM_{IG} 代表经过信息增益改进过的贝叶斯模型。表中正确率的计算方法如下：

$$P(\text{Correct}) = \frac{C(\text{正确标注的数量})}{C(\text{标注的总数})} \times 100\% \quad (7)$$

表 2 信息增益改进贝叶斯模型的正确率(%)

模型	NBM	BM_{IG}
w_1	79.79	82.03
w_2	78.26	82.56
w_3	90.33	92.84
w_4	91.60	95.01
w_5	62.25	64.96
w_6	83.95	87.75
w_7	75.35	78.79
w_8	69.51	74.56
平均	78.88	82.31

4.3 实验结果分析

本文的实验中分别选择 NBM 和 BM_{IG} 建立了两个词义消歧的分类器，为了方便比较，把实验结果的正确率增加幅度(Amplitude)做成柱形图，参见图 1。

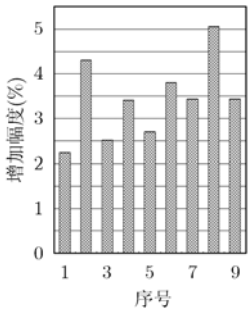

图 1 模型改进后实验结果的增加幅度

Amplitude 的计算公式如下：

$$P(\text{Amp}) = \frac{P_{\text{NBIG}} - P_{\text{NB}}}{P_{\text{NB}}} \times 100\% \tag{8}$$

其中 P(Amp)代表正确率的增幅，P_{NBIG} 表示采用改进模型的正确率，P_{NB} 表示采用 NBM 模型的正确率。

图中 1~8 歧义词序号是 8 个人造歧义词，序号 9 是消歧结果的平均值。从图中可以很清晰地看到改进后的模型在词义消歧正确率上出现了明显的提升。增幅最大的超过了 5.05 个百分点，小的也超过了 2.24 个百分点，增幅的平均值达到了 3.43 个百分点。

通过实验结果的对比，可以看到贝叶斯模型经过特征选择，消歧正确率提高明显。实验表明，在贝叶斯模型上通过信息增益计算特征词语的位置权重后，提高了消歧模型的词义辨识能力。该实验也证明上下文特征词语的位置信息对歧义词词义的判断有着很重要的作用，本文通过计算信息增益，深入挖掘上下文线性顺序中的语言信息，在词义消歧实验上收效显著。

5 结束语

本文通过对比实验发现，单纯贝叶斯模型(NBM)在没有使用位置信息的前提下也一样具有较为理想的表现，探讨如何进一步改善贝叶斯分类器的性能颇具研究价值。

为了改进贝叶斯模型，本文针对贝叶斯假设的特点，在实验中采用了基于信息增益最大原则的方法来估算上下文词语的位置信息，并以此为基础实现了一种改进的贝叶斯词义分类器。实验证明该方法切实可行，使得词义消歧的正确率获得了进一步的提升。

参 考 文 献

[1] Zhu Jingbo and Hovy Eduard. Active learning for word sense disambiguation with methods for addressing the class imbalance problem[C]. The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007: 783–790.

[2] 卢志茂, 刘挺, 李生等. 统计词义消歧的研究进展[J]. 电子学报, 2006, 34(2): 333–343.
Lu Zhi-mao, Liu Ting, and Li Sheng. The research progress of statistical word sense disambiguation. *Acta Electronica Sinica* [J], 2006, 34(2): 333–343.

[3] 吴云芳, 金澎, 郭涛. 基于词典属性特征的粗粒度词义消歧[J]. 中文信息学报, 2007, 21(2): 3–8.
Wu Yun-fang, Jin Peng, and Guo Tao. Coarse-grained word sense disambiguation using features described in the lexicon. *Journal of Chinese Information Processing*[J], 2007, 21(2): 3–8.

[4] 陈浩, 何婷婷等. 基于K-menas聚类的无导词义消歧. 中文信息学报[J], 2005, 19(4): 10–16.
Chen Hao and He Ting-ting, *et al.*. An unsupervised approach to word sense disambiguation based on clustering by K-means. *Journal of Chinese Information Processing* [J], 2005, 19(4): 10–16.

[5] 卢志茂, 刘挺, 张刚等. 基于依存分析改进贝叶斯模型的词义消歧. 高技术通讯[J], 2003, 13(5): 1–7.
Lu Zhi-mao, Liu Ting, and Zhang Gang, *et al.*. Word sense disambiguation based on dependency relationship analysis & Bayesian model. *Chinese High Technology Letters*[J], 2003, 13(5): 1–7.

[6] Quinlan J R. *Induction of decision trees. Machine Learning*[J]. 1986, 1(1): 81–106.

[7] Schütze H. Context space. In Robert Goldman, Peter Norvig. Eugene Charniak, and Bill Gale (eds.). Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, Menlo Park, CA. AAAI Press, 1992: 113–120.

范冬梅：　女，1972 年生，博士生，研究方向为自然语言处理、词义消歧.

卢志茂：　男，1972 年生，教授，博士生导师，主要研究方向为自然语言处理、搜索引擎.

张汝波：　男，1963 年生，教授，博士生导师，主要研究方向为人工智能、语音识别.