隐马尔可夫模型和贝叶斯模型词义消歧对比研究

丁江伟 刘挺 卢志茂 李生
哈尔滨工业大学计算机学院
(哈尔滨工业大学 321 信箱，哈尔滨 150001)
E-mail:{djw, tliu, lzm, ls}@ir.hit.edu.cn

摘 要：词义消歧是自然语言处理中的一个难点和热点问题。现阶段，多义词消歧的研究大多采用几个有代表性的歧义词作为研究与测试的对象，与实际应用还存在一定的距离，作者针对真实的应用情况，对大规模文本进行了词义消歧研究。本文比较了两个经典的统计模型解决大规模的词义消歧难题的优缺点，一阶隐马尔可夫模型考察了邻接的上下文，有些时候距离歧义词较远的词语往往对词义的确定起着至关重要的作用，所以这种方法的消歧正确率比较低，开放测试在 85%左右；单纯贝叶斯概率模型的消歧方法在抽取上下文特征时加大了上下文的窗口，使与多义词消歧相关的信息充分考虑进来，这种方法的开放消歧正确率最高可达 92%，消歧效果明显。由此证明了贝叶斯模型词义消歧的有效性和比较优势。

关键词：词义消歧，自然语言处理，隐马尔可夫模型，单纯贝叶斯模型

WSD Based on Hidden Markov Model & Naive Bayes Model Contrastively Research

Ding Jiangwei Liu Ting Lu Zhimao Li Sheng
School of Computer Science and Technology of HIT
(Harbin Institute of Technology 321 Box, Harbin 150001)
E-mail:{djw, tliu, lzm, ls}@ir.hit.edu.cn

Abstract: Word Sense Disambiguation has always been a key problem and one of the difficult points in Natural Language Processing. Presently, only some ambiguous words are selected as disambiguated objects in many word sense disambiguation researches, which have great limitations in real application. In this paper, differently, large-scale real texts are researched applying supervised word sense disambiguation approach based on statistical model. By a mass of contrastively experiments based on Hidden Markov Model and Naive Bayes Model, we find that the accuracy of the former model is only 85.05% in open test but 92.00% of the latter, substantiating the wonderful performance of Naive Bayes Model.

Keywords: Word Sense Disambiguation, NLP, Hidden Markov Model, Naïve Bayes Model

0 引言

信息的主要载体是自然语言，一词多义是它的一个普遍现象，我们对语料库的统计发现，汉语中多义词的出现频率在 0.40 左右[1]。一词多义问题给自然语言的机器理解带来了困难，解决方法就是对多义词进行词义消歧（Word Sense Disambiguation, WSD）。词义消歧，我们定义为在给定上下文语境的条件下由机器自动、高效地确定多义词义项的技术。词义消歧技术在自然语言处理（Natural Language Processing, NLP）领域是一个重要的热点研究问题，对于包括信息检索、文本挖掘、机器翻译、文本分类和自动文摘等的许多自然语言处理系统都十分有用。例如在信息检索中，利用词义消歧技术可使整个系统的检索正确率提高 3.2 个百分点，使检索结果更令用户满意[2]。

在多义词的上下文范围确定之后，消歧的方法选择就成为了歧义消解的关键，选择消歧方法的目的在于获得有助于确定多义词词义的上下文特征或者知识。知识获取是计算语言学（Computational Linguistics, CL）领域所面临的最大瓶颈[3]。根据获取知识的方法，消歧的方法可以分成基于语法语义规则的方法和基于语料库的统计方法两类。基于规则的方法主要通过约束性规则来确定多义词在上下文中的词义，这需要一个具有完备性、一致性、可扩充性和对开放领域适应性的知识库。基于语料库的统计方法通过计算给定文本中词汇语义在多义词上下文中的概率权重，选择具有最大概率权重的语义作为最佳结果输出，该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[4][5][6][7][8]。统计学方法随着语料库语言学的兴起，以其良好的词义消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。

隐马尔科夫模型（Hidden Markov Model, HMM）自引入到计算机应用学科以来，在语音识别、图像处理和自然语言处理等诸多领域得到了广泛的应用，利用该模型进行词义消歧也取得了不错的效果。国外的一些学者在词义消歧技术中应用贝叶斯分类器（Bayes Classifier, BC）同样效果显著[9][10]。本文分别对一阶隐马尔可夫模型和单纯贝叶斯模型进行了大规模文本消歧试验，进而根据实验结果对比分析了两种模型的优缺点。

1 语义资源——HowNet

HowNet（知网）是发布在网上的一个知识资源。作为一个新型知识库，HowNet 描述概念、概念之间关系和概念所具有的属性之间关系，并力求反映出概念的共性和个性。HowNet 的各个部件构成一个有机的知识系统，采用有限的形式化的可计算的释义集合表示所有的概念，便于在自然语言处理中应用。

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）解释概念，而这个义原集合称之为一个义项，用一个编号（NO.）标识。由于 HowNet 对词语概念刻画的很细，这个 NO.并不唯一，词语 w 的同一个定义（DEF）可能有不同的编号，所以在词义

消歧计算前，为了方便标注我们根据词语的 DEF，为每个义项给定一个语义号码（SenseNo）。SenseNo 按照义项的 DEF 在 HowNet 中出现的先后自动顺序生成，如“打”有两个义项，见表 1.1，并根据具体情况对 DEF 进行适当处理，如通常取 DEF 的主要特征，即 DEF 的第一项，当主要特征是“属性”、“属性值”、“数量”、“数量值”时，还要取次要特征，即第二项。这样可以建立一个“Word—SenseNo”数据库。

表 1.1 语义号码示例

Hownet	NO.=017645 W_C=打 G_C=V E_C=~架，~斗，~仗 W_E=attack G_E=V E_E= DEF=fight|争斗	NO.=017203 W_C=打 G_C=V E_C= W_E=project G_E=V E_E= DEF=send|发送
SenseNo	12	604

在该数据库中同一个词（多义词）的不同义项具有不同的 SenseNo，而不同的词（如同义词）可以有相同的 SenseNo。我们共统计出 1278 个义项，这样用有限的词语释义空间可以表示所有的概念，可以作为我们进行词义标注的依据。

2 两个经典的词义消歧模型

2.1 隐马尔可夫模型

一个隐马尔可夫模型是一组有限的状态，其中的某一个状态可以以一定的概率转移到另外的状态（终止状态除外），而且在转移时产生输出，能产生的输出是有限的，输出也是以一定的概率产生的。它的形式化描述是 $HMM=\langle S,W,A,B,\pi\rangle$。应用在词义消歧问题中的隐马尔可夫模型可以定义为：

1) S 表示模型中的状态，N 是其的状态数。在词义消歧中，状态就是知网（HowNet）中抽取的语义标号，所有独立的语义标号定义为 $S=\{S_1,S_2,...,S_N\}$，用 q_t 来表示一个句子中第 t 个词的语义，它可能是 1278 个词性标记中的任一种，即 $S=\{S_1,S_2,...,S_{1278}\}$，$N$=1278

2) W 表示每个状态的观察值，M 表示每个状态上对应的可能的观察值的数目。记作：$W=\{w_1,w_2,...,w_M\}$。词义消歧中观察值是指语料中的词，M 的大小就是一个句子中包含的单词数目。

3) 状态转移概率矩阵 $A=\{a_{ij}\}$。此矩阵中的各元素在词义消歧中表示为某一语义向其它各个语义转移的概率，即：

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i) = \frac{\text{从}S_i\text{转移到}S_j\text{的次数}}{\sum_{j=1}^{1278}\text{从}S_i\text{转移到}S_j\text{的次数}},\ 1 \le i \le 1278$$

可见，在词义消歧问题中确定的转移概率矩阵是一个1278×1278的矩阵。

4) 观察值概率分布矩阵$B = \{b_j(k)\}$，词义消歧中即为单词概率分布矩阵。其中$b_j(k)$表示在S_j语义下，t时刻出现单词w_k的概率，我们把它称为词汇概率或发射概率，即

$$b_j(k) = P(t\text{时刻出现}w_k \mid q_t = S_j) = \frac{\text{单词}w_k\text{取}S_j\text{语义的次数}}{\text{语料中语义}S_j\text{出现的总次数}}\quad 1 \le j \le 1278,\ 1 \le k \le M$$

5) 初始状态分布矢量$\pi = \{\pi_i\}$，词义消歧中表示在$t = 1$时刻单词出现语义S_i的概率，即处于句首的单词出现语义S_i的概率。

$$\pi_i = P(q_1 = S_i) = \frac{\text{句首单词取语义标记}S_i\text{的次数}}{\sum_{i=1}^{1278}\text{句首单词取语义标记}S_i\text{的次数}}\quad 1 \le i \le 1278$$

在给定的模型下，要求从一定观察值序列的所有可能的状态中，选取概率最大的作为最终的状态序列。

2.2 单纯贝叶斯模型（Naïve Bayes Model,NBM）

贝叶斯分类的方法通过歧义词上下文窗口来判断歧义词的词义，上下文中的每一个词语提供了确定词义的潜在的有用信息。贝叶斯方法没有特征抽取，而是利用从所有特征得到的信息来进行判断。在进行判断时，利用了贝叶斯规则如公式(2-1)所示，它可以使发生错误的概率达到最小。

贝叶斯规则使每一个词的判断错误概率达到最小，从而使整个句子的错误概率也降低到最小。在贝叶斯规则中$P(s_k \mid c)$一般比较难以计算，但可以根据贝叶斯公式(2-2)来计算。

$$P(s' \mid c) > P(s_k \mid c)\ \ s_k \ne s' \quad (2\text{-}1)$$

$$P(s_k \mid c) = \frac{P(c \mid s_k)}{P(c)}P(s_k) \quad (2\text{-}2)$$

在实际的应用中我们常常采用单纯贝叶斯分类器，单纯贝叶斯分类器以其高效、能够利用多种分类特征的能力在机器学习中得到广泛的应用。若歧义词为w，其上下文为c，上下文中的词语为v_j，则单纯贝叶斯假设表示如公式(2-3)所示：

$$P(c|s_i)=P(\{v_j|v_j \ in \ c\}|s_i)=\prod_{v_j in c}P(v_j|s_i) \qquad (2\text{-}3)$$

根据单纯贝叶斯假设，分类器的判定规则可以如公式(2-4)表示：

$$s'=\arg\max_{s_i}[\log P(s_i)+\sum_{v_j in c}\log P(v_j|s_i)] \qquad (2\text{-}4)$$

3 实验及结果分析

3.1 实验流程

我们从《人民日报》随机抽取了中等长度的句子 6300 句（统计表明每句平均 8.5 词），把这 6300 句分为两部分，3500 句用于模型的训练，另外 2800 句用作测试，保证没有重叠。训练数据再按 500 句等差递增，分为 500 句、1000 句、1500 句、2000 句、2500 句、3000 句、3500 句等 7 组数据。对于训练语料和测试语料要进行分词处理，并且采用半自动词义标注的方法根据 HowNet 确定每一个汉语词汇的所有可能词义得到可靠的训练集和测试集。

整个实验分为使用单纯贝叶斯模型和隐马尔可夫斯模型的两部分来完成，并且每部分的实验过程又分为两个阶段，即训练阶段和测试阶段。训练阶段，根据训练集里的上下文信息，计算出词义消歧模型所使用的参数；测试阶段，分别使用一阶隐马尔可夫和单纯贝叶斯消歧模型，赋予测试语料中多义词一个正确的语义。

3.2 实验结果

训练数据共有 7 组，可以训练出 7 组概率参数。采用单纯贝叶斯模型分别作开放测试

表 3.1　两个模型的实验结果

一阶隐马尔科夫模型（HMM）			单纯贝叶斯模型（NBM）		
训练数据（单位：句）	封闭测试正确率（%）	开放测试正确率（%）	训练数据（单位：句）	封闭测试正确率（%）	开放测试正确率（%）
500	85.59	82.35	500	99.24	87.35
1000	87.66	83.16	1000	99.15	89.37
1500	87.85	83.74	1500	98.85	90.51
2000	87.93	84.36	2000	98.69	91.07
2500	88.40	84.69	2500	98.58	91.55
3000	88.47	84.88	3000	98.48	91.71
3500	88.46	85.05	3500	98.41	92.00
说明	开放测试 2800 句				

和封闭测试：对照实验，我们采用一阶隐马尔可夫模型，训练了 7 组概率参数，同样作了开放和封闭测试，也得到 7 组实验数据，见表 3.1。

表中正确率按下面公式计算：P(Correct) =∑Correct SenseNo/∑SenseNo

文本词义标注结果示例：（其中数字为语义号）

例 1. 昨天/174 是/673 彼得/-1 打/12 来/877 的/1003 骚扰/304 电话/3 。/-1

例句 1 用一阶隐马尔可夫模型进行词义标注，结果错误。

例 2. 昨天/174 是/673 彼得/-1 打/604 来/877 的/1003 骚扰/304 电话/3 。/-1

例句 2 用单纯贝叶斯模型标注，结果正确。

其中“打” 是个很典型的多义词，有 28 种义项。例 1 中“打”标注的 SenseNo 为 12，代表义原集合“fight|争斗”、类似的有“打架，~打斗，~打仗，~打敌人，~打死，~打伤，~打得好”等等，英语解释为“attack”；在例 2 中标注的 SenseNo 为 604，代表“send|发送”，类似的有“打电报，打手电，打信号，打连发，打炮”等。

3.3 实验结果的几点讨论

我们对 7 组实验的结果用分布曲线来表示，见图 3.1，曲线 A1、A2 是一阶隐马尔可夫模型的开放和封闭测试结果，曲线 B1、B2 是单纯贝叶斯模型的开放和封闭测试结果。

通过图示的曲线走势，可以有如下几点结论：

1) 单纯贝叶斯模型实验的封闭测试时，词义标注的正确率很高，最高点趋近 100%；并且随着测试语料的增加由于语言现象越来越丰富，正确率缓慢降低；

2) 隐马尔可夫模型实验封闭测试时，随着训练语料的增加正确率升高，并且趋于平缓，最高点接近 90%；

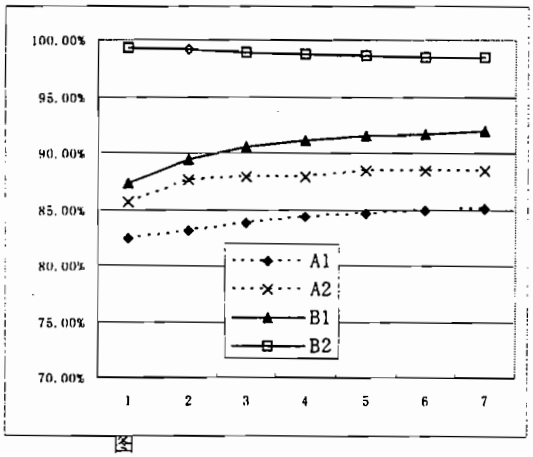

图 3.1 实验结果比较

3) 封闭测试，两个模型的正确率差别很大，并且都随着训练语料规模的增大变化趋于平缓；

4) 开放测试中，基于单纯贝叶斯模型的正确率明显优于隐马尔可夫模型的正确率；

5) 两个模型的开放测试的正确率随着训练语料规模的增大而提高，分别以封闭测试的正确率为上限。

在单纯贝叶斯模型封闭测试中，使用较小规模 的训练语料，词语 w 的各个语义的上下文信息都被考虑进来，训练时的语义 sk（正确语义）在消歧计算中的 score(sk) 往往很容易成为最大值，所以实验结果会获得相对很高的正确率。但是随着训练语料规模的增大，w 各个语义的上下文数目不断增大，不同语义会出现部分相同上下文的现象，而且这种现象会越来越严重，在一定程度上增加了词义消歧的难度，从而导致正确率有所平缓。对于开放测试实

验，训练语料规模越大，不同词语 w 各个语义的实例出现的几率会增大，模型所能学习到的有效上下文特征值也就会越多，从而提高消歧结果的正确率。

由一阶隐马尔可夫模型的假设我们知道，输出观察值概率只与当前状态有关，转移概率只与前一个状态有关。实际上我们可以认为多义词的上一个邻接词是包含了对消歧有用的全部信息。在隐马尔可夫模型消歧的封闭测试中，随着训练语料的增加，语料所含有的语言现象越来越丰富，表现为正确率上升。开放测试也是如此。显然隐马尔可夫模型的假设过于草率，只考虑了对消歧有用的一部分信息。所以无论是封闭测试的 88.46%还是开放测试的 85.05%其消歧正确率并不高，远没有单纯贝叶斯模型（封闭测试最高 92.00%）的表现好。

4 结束语

本文采用单纯贝叶斯和隐马尔可夫两种模型在同一大规模训练集和测试集上作了词义消歧的对比试验。隐马尔可夫模型假设只有多义词的邻接词对消歧起作用，单纯贝叶斯模型则充分考虑了上下文的影响，所以理论上讲单纯贝叶斯模型在消歧效果上应该优于隐马尔可夫模型。试验结果证明了这一点，开放测试 2800 个例句，一阶隐马尔可夫模型消歧正确率最高只有 85.05%，单纯贝叶斯模型则达到了 92.00%。单纯贝叶斯模型实验效果令人满意。

由于本方法需要在训练语料中获得知识，大规模的训练语料又很难获得，如何充分利用有限的资源，提高训练参数的可靠性，是一个值得关注的问题。另外贝叶斯消歧模型中的上下文窗口是以句子为单位的，不可避免的使与该多义词语义无关的信息被考虑了进来，对消歧效果产生了负面的影响，怎样利用语法、语义信息有效减小消歧窗口是下一步要重点考虑的问题。

参 考 文 献

[1] 鲁松，白硕，黄雄，张健．基于向量空间模型的有导词义消歧．计算机研究与发展．2001，38(6)：663-667

[2] Hinrich, S., Pedersen, J., Information retrieval based on word senses. In: Proc of the 4th Annual ymposium on Document Analysis and Information Retrieval. Las Vegas, NV, 1995，161-175.

[3] 荀恩东，李生，赵铁军．基于汉语二元同现的统计词义消歧方法研究．高技术通讯．1998，10

[4] 杨尔弘，张国清，张永奎．基于义原同现频率的汉语词义排歧方法．计算机研究与发展．2001，38(7)：834-837

[5] Yarowsky, D. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In Proceedings. COLING-92. Nantes, 1992: 454-460

[6] 刘小虎．英汉机器翻译中词义消歧方法的研究．哈尔滨工业大学博士学位论文．1998

[7] Rada Mihalcea and Dan Moldovan. A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation. International Journal on Artificial Intelligence Tools. 2001, 10(1): 5-21

[8] 陈丹琪．统计与规则相结合的英语词性标注和基本名词短语分析．哈尔滨工业大学硕士学位论文．1999

[9] Rada Mihalcea and Dan Moldovan. A Method for Word Sense Disambiguation of Unrestricted Text. In Proceedings of ACL '99. Maryland, NY, 1999: 152-158

[10] 付国宏．汉语句法歧义消解的统计方法研究．哈尔滨工业大学博士学位论文．2000