基于对数模型的词义自动消歧*

朱靖波，李 珩，张 跃，姚天顺

(东北大学 计算机科学研究所，辽宁 沈阳 110006)

E-mail: cipol@nlplab.com

http://www.nlplab.com

摘要：提出了一种对数模型(logarithm model，简称 LM)，构造了一个词义自动消歧系统 LM_WSD(word sense disambiguation based on logarithm model)。在词义自动消歧实验中，构造了 4 种计算模型进行词义消歧，根据 4 个计算模型的消歧结果，分析了高频率词义、指示词、特定领域、固定搭配和固定用法信息对名词和动词词义消歧的影响。目前，该词义自动消歧系统 LM_WSD 已经应用于基于词层的英汉机器翻译系统(汽车配件专业领域)中，有效地提高了翻译性能。

关键词：词义自动消歧；机器翻译；对数模型；自然语言处理

中图法分类号：TP18 **文献标识码**：A

词义歧义是英汉机器翻译过程中的一类很典型的歧义问题。词义自动消歧(word sense disambiguation)过程对于许多自然语言处理系统是十分有用的，包括信息检索、文本分类、机器翻译等。词义自动消歧方法基本上可以分为两大类：基于定性的方法(qualitative approach)和基于定量的方法(quantitative approach)。

基于定性的方法主要采用选择约束性规则来确定每个词汇在不同上下文中的词义选择，如基于选择性限定规则(selectional restrictions)[1]、决策树(decision trees)[2]、决策表(decision lists)[3]等等。基于定性的方法大多依赖于一些语言学知识库，如机器可读词典(machine-readable dictionary，简称 MRD)，其面临的关键问题在于规则知识库的构造及知识获取的瓶颈问题。

基于定量的方法通过计算每个词汇候选词义在上下文条件下的概率权值，选择最大概率权值的词义作为结果输出，如 Naive-Bayes 算法、基于类的方法(class-based approach)[4]、VSM(vector space model)等等。基于定量的方法面临着统计数据稀疏的问题，带标语料的人工构造是知识获取的瓶颈问题。

本文提出了一种对数模型(logarithm model，简称 LM)。为了解决其中的统计数据稀疏问题，本文采用了一种基于相似的评估技术，利用 WordNet 来计算词汇词义的相似度。在词义自动消歧实验中，我们构造了 4 种计算模型进行词义消歧，根据 4 个计算模型的消歧结果，分别讨论了高频率词义、指示词、特定领域、固定搭配和固定用法信息对名词和动词词义消歧的影响。

* 收稿日期：1999-07-20；修改日期：2000-05-10

基金项目：国家自然科学基金资助项目(69985001)；国家重点基础研究发展规划 973 资助项目(G19980305011)；国家教育部博士点基金资助项目(19990145O3)

作者简介：朱靖波(1973－)，男，浙江永康人，博士，副教授，主要研究领域为多国语机器翻译，现代汉语分析理论与方法；李珩(1975－)，男，辽宁沈阳人，硕士，主要研究领域为语料库语言学，机器翻译；张跃(1975－)，男，江苏扬州人，硕士，主要研究领域为现代汉语分析；姚天顺(1934－)，男，江苏苏州人，教授，博士生导师，主要研究领域为计算语言学，知识工程。

1 词义自动消歧

1.1 基本过程

Reifler[5]关于一个词汇与其上下文存在一定的"语义巧合"关系的思想很快成为了 WSD 的基本思想.根据一个词汇的上下文,从该词汇的多个词义定义中选择一个正确词义的过程,我们称为词义消歧(word sense disambiguation,简称 WSD)过程.一般来说,词义自动消歧过程分为两步:

第 1 步:首先确定一个词汇的所有可能的词义;

第 2 步:根据该词汇的上下文,采用某种消歧技术赋予一个正确的词义.

大多数词义自动消歧的方法通常采用日常词典(如牛津词典)、分类词典(如 WordNet)和双语词典等来详细描述词义的精确定义,同时依赖于一些词法分析,如词性自动标注等.对于词义自动消歧处理过程的第 1 步处理来说,一个词汇的所有可能的词义定义通常事先存放在系统的词典知识库中,通过对词典知识库的查找就可以获得该词汇的所有可能的词义,这一过程相当于一个检索过程.其中一个词汇的词性经过分词与词性标注后可以确定下来.词义自动消歧技术的研究焦点主要集中于第 2 步处理上.本文提出的词义自动消歧过程是在确定词性的前提下,对具有多个词义的词汇完成消歧过程.

1.2 对数模型 LM

从概率理论角度来说,词义自动消歧的过程相当于根据给定输入条件选择最大概率的词义这样一个过程.根据 Bayesian 公式,词义 s 的条件概率 $P(s|x)$ 计算公式为

$$P(s|x)=\frac{P(s)P(x|s)}{P(x)}. \tag{1}$$

在实际应用中,事先给定的上下文 x 对于所有词义来说是不变的,因此 $P(x)$ 可以忽略不计,不会影响不同词义的概率计算结果,则

$$\operatorname{argmax}P(s|x)=\operatorname{argmax}P(s)P(x|s). \tag{2}$$

词义 s 的概率 $P(s)$ 是根据训练数据中词义 s 的分布来进行计算的,因此,这种方法的性能很大程度上依赖于概率 $P(x|s)$ 的计算方法.为了使讨论具有普遍性,将输入 x 描述成由用于词义消歧的特征构成的向量表示:

$$X=(F_1=f_1,F_2=f_2,\ldots,F_n=f_n).$$

其中 F_i 表示第 i 个特征,f_i 表示第 i 个特征值.

我们可以根据词义特征的条件概率近似地计算条件概率 $P(x|s)$.假设对于给定的词义 s 来说,每个特征都条件无关,根据"Naive-Bayes 方法",条件概率 $P(x|s)$ 的计算公式为

$$P(x|s)\approx\prod_{i=1}^{n}P(F_i=f_i|s). \tag{3}$$

根据公式(2)和(3)可以得出词义 s 的条件概率 $P(s|x)$ 的计算公式为

$$P(s|x)=P(s)\prod_{i=1}^{n}P(F_i=f_i|s). \tag{4}$$

对公式(4)中的词义条件概率 $P(s|x)$ 取对数,则可以转换为

$$\ln P(s|x)=\ln\left(P(s)\prod_{i=1}^{n}P(F_i=f_i|s)\right)-\ln P(s)+\sum_{i=1}^{n}\ln P(F_i=f_i|s). \tag{5}$$

根据 Bayesian 公式,条件概率 $P(F_i=f_i|s)$ 的计算公式为

$$P(F_i=f_i|s)=\frac{P(F_i=f_i)P(s|F_i=f_i)}{P(s)}. \tag{6}$$

利用公式(6)替换公式(5)中的条件概率 $P(F_i=f_i|s)$,则

$$\begin{aligned}\ln P(s|x)&=\ln P(s)+\sum_{i=1}^{n}\ln\frac{P(F_i=f_i)P(s|F_i=f_i)}{P(s)}\\&=\ln P(s)+\sum_{i=1}^{n}\ln P(F_i=f_i)+\sum_{i=1}^{n}P(s|F_i=f_i)-\sum_{i=1}^{n}\ln P(s)\\&=\sum_{i=1}^{n}\ln P(F_i=f_i)+\sum_{i=1}^{n}\ln P(s|F_i=f_i)-(n-1)\ln P(s).\end{aligned} \tag{7}$$

假设给定上下文 x,待消歧词汇 w 具有两个词义 s_1 和 s_2,如果计算结果为

$$P(s_1|x)>P(s_2|x), \tag{8}$$

其中 $0\leqslant P(s_1|x)\leqslant 1$,$0\leqslant P(s_2|x)\leqslant 1$,则选择词汇 w 在当前上下文 x 条件下的正确词义为 s_1.

由于 $P(s_1|x)>0$,$P(s_2|x)>0$,根据对数函数 ln 的特点,同样可以得出

$$\ln P(s_1|x)-\ln P(s_2|x)>0. \tag{9}$$

利用公式(7)来计算不等式(9)中的条件概率 $P(s_1|x)$和 $P(s_2|x)$,则

$$\begin{aligned}\ln P(s_1|x)-\ln P(s_2|x)&=\sum_{i=1}^{n}\ln P(F_i=f_i)+\sum_{i=1}^{n}\ln P(s_1|F_i=f_i)-(n-1)\ln P(s_1)-\\&\quad\sum_{i=1}^{n}\ln P(F_i=f_i)+\sum_{i=1}^{n}\ln P(s_2|F_i=f_i)-(n-1)\ln P(s_2)\\&=\sum_{i=1}^{n}(\ln P(s_1|F_i=f_i)-\ln P(s_2|F_i=f_i))-\\&\quad(n-1)(\ln P(s_1)-\ln P(s_2))\\&=\sum_{i=1}^{n}\ln\frac{P(s_1|F_i=f_i)}{P(s_2|F_i=f_i)}-(n-1)\ln\frac{P(s_1)}{P(s_2)}.\end{aligned} \tag{10}$$

根据公式(10)的计算,可以得出如下结论:

$$\ln P(s_1|x)-\ln P(s_2|x)\begin{cases}>0 & \text{选择词义 } s_1\\=0 & \begin{cases}P(s_1)\geqslant P(s_2) & \text{选择词义 } s_1\\P(s_1)<P(s_2) & \text{选择词义 } s_2\end{cases}\\<0 & \text{选择词义 } s_2\end{cases}. \tag{11}$$

我们称公式(11)为对数模型.本文将利用对数模型 LM 完成词义的自动消歧过程.

1.3 基于相似的平滑技术

基于相似的平滑技术的基本假设为:当对象 x 的数据缺乏时,则根据对象 x 的所有相似对象 $x'\in S(x)$(其中 $S(x)$表示与对象 x 相似的对象集合)的概率分布来评估对象 x 的概率分布,同时考虑对象 x 与 x'之间的相似性 $Sim(x,x')$.

假设 X 为我们所考虑的对象集合,Y 为所有可能的上下文的集合$\{y_1,y_2,\ldots,y_n\}$.$C(x,y)$表示数据对$(x,y)\in\{X\times Y\}$在样本语料中出现的次数.同理,$C(x)$表示对象 $x\in X$ 在样本语料中出现的次数.条件概率 $P_{\mathrm{SIM}}(y|x)$的计算公式为

$$P_{\mathrm{SIM}}(y|x)=\sum_{x'\in S(x)}\frac{Sim(x,x')}{\sum_{x'\in S(x)}Sim(x,x')}\times\frac{C(x',y)}{C(x')}. \tag{12}$$

为了论述方便,设集合 $S(x)$ 中的所有对象与对象 x 相似.但是,如果 $S(x)$ 规模非常大,则公式(12)的计算十分耗时.在下面的实际应用中,我们将以某种限制约束 $S(x)$ 规模,要求 x 与 x' 之间的相似性大于某个阈值等等.在实际应用中,我们发现限制选择相似对象对系统性能并没有太多影响.下面需要考虑如何计算对象 x 与 x' 之间的相似性 $Sim(x,x')$.

1.4 基于 WordNet 的相似性计算

词汇相似计算过程在很多基于统计的语言模型中得到了应用,如基于手工构造的词典(Roget 辞典、EDR 等)、基于向量空间模型 VSM 等.其中基于辞典的相似性计算方法一般是基于一个假设:在辞典结构中相邻位置节点的词汇具有相似的意义.所以,给定两个词汇的相似性往往利用两者在辞典结构中相邻的路径长度来表示[6].在基于向量空间模型 VSM 的相似性计算方法中,利用向量表示给定的词汇,考虑词汇共现概率信息,计算两个向量的相似度来衡量两个词汇的相似度.向量的相似度的计算一般采用 COS 形式的距离函数,计算复杂度为 $O(N)$,其中 N 为向量的长度.

本文采用基于 WordNet 来完成给定词汇的相似性计算.由于仅考虑 WordNet 辞典结构中相邻位置的路径长度来计算词汇的相似程度,效果不是很好[7].因此,本文在计算给定词汇的相似度过程中,考虑两者在 WordNet 中位置之间的距离和它们本身的深度.

假设给定两个词汇 w_1 和 w_2,在 WordNet 中对应的词义为 s_1 和 s_2,词汇 w_1 和 w_2 词义距离 SD(w_1 和 w_2)的计算公式为

$$SD(w_1,w_2)=\frac{1}{2}\left\{\frac{Dis(s_1,ca)}{Dis(s_1,root)}+\frac{Dis(s_2,ca)}{Dis(s_2,root)}\right\}. \tag{13}$$

其中 ca 表示给定词汇 w_1 和 w_2 的词义 s_1 和 s_2 在 WordNet 中的共同祖先概念节点,Dis 函数表示计算两个概念在 WordNet 中的位置之间的路径长度.

相似性 $Sim(w_1,w_2)$ 的计算公式为

$$Sim(w_1,w_2)=e^{-SD(w_1,w_2)}. \tag{14}$$

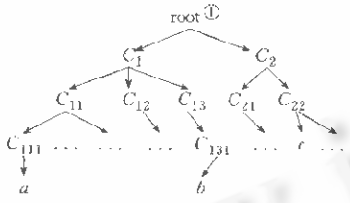

Fig. 1 Positions of nodes a, b and c in the WordNet
图1 a,b,c 在WordNet中的位置

由公式(14)可以看出,词义距离越大,相似性越小.当词义距离为 0 时,也就是说,当前两个词汇 w_1 和 w_2 属于同一个概念时,相似度为 1(绝对相似).

例如,a,b 和 c 在 WordNet 中的概念体系的位置如图 1 所示.

根据公式(13),可以计算出图 1 中的节点 a 和 b 以及 a 和 c 之间的相似度,计算过程如下:

(1) 对于 a 和 b,

$$ac=C_1,$$
$$Dis(a,ca)=3, Dis(b,ca)=3,$$
$$Dis(a,root)=4, Dis(b,root)=4,$$
$$SD(a,b)=1/2(Dis(a,ca)/Dis(a,root)+Dis(b,ca)/Dis(b,root))$$
$$=1/2(3/4+3/4)=0.75.$$

(2) 对于 a 和 c,

$$ac=root,$$
$$Dis(a,ca)=4, Dis(c,ca)=3,$$

$$Dis(a,root)=4, Dis(c,root)=3,$$
$$SD(a,c)=1/2(Dis(a,ca)/Dis(a,root)+Dis(c,ca)/Dis(c,root))$$
$$-1/2(4/4+3/3)-1.$$

结论:由于 $SD(a,b)<SD(a,c)$,则 $Sim(a,b)>Sim(a,c)$,从而得出 a 和 b 比较相似.

1.5 上下文特征的选择

特征的选取是词义自动消歧模型中很重要的一个环节.本文的词义自动消歧模型的特征选取不仅考虑词汇本身,同时也考虑一些词汇的形态变化信息、词性、固定搭配等等.特征的选取类型主要包括:

◆句子的特征主要考虑一些句子包含的名词、动词和形容词,利用句子中的名词、动词和形容词构造上下文向量 CV;

◆词典中关于词义的详细定义,特征的自动抽取主要考虑定义中包含的名词、动词和形容词,利用定义中的名词、动词和形容词构造词义向量 SV;

◆词汇的词性,输入句子根据词性自动标注过程 NAA POS,能够唯一地确定词性;

◆词汇的形态变化,包括可数/不可数、复数/单数、原形动词/过去分词/过去式/动名词等等.例如,词汇"resource"的第 1 个词义"资源",通常以复数形式出现;词汇"responsibility"表现为可数形式的时候,词义为"职责",否则为"责任、义务"等等;

◆固定搭配、短语信息.例如,词汇"resource"的第 3 个词义"精力",固定搭配为"leave sb to his/her own resources",解释为"不去打扰某人";词汇"respect"的词义为"关于",固定短语为"with respect to".

2 实验结果

在词义自动消歧实验中,目前主要对测试语料中的名词和动词进行词义消歧.系统采用准确率来评估该方法的性能,定义 CR 为准确率,N_c 表示标注正确的名词和动词个数,N 表示测试语料中名词和动词的总词数.准确率的计算公式如下:

$$CR=\frac{N_c}{N}\times 100\%. \tag{15}$$

实验中采用 4 种计算模型对测试语料进行自动词义标注,包括:

(1) 计算模型 WSD-1(word sense disambiguation 1th)

直接利用频率最高的词义(第 1 个词义)作为正确词义标注.

(2) 计算模型 WSD-2(word sense disambiguation 2th)

利用词义详细定义中的特征词汇作为"指示词"进行消歧,如果上下文中没有"指示词",则利用第 1 词义作为正确词义标注.

(3) 计算模型 WSD-3(word sense disambiguation 3th)

基于对数模型 LM 的词义自动消歧过程,其中不利用固定搭配和固定用法信息.

(4) 计算模型 WSD-4(word sense disambiguation 4th)

基于对数模型 LM 的词义自动消歧过程,其中利用固定搭配和固定用法信息参与词义自动消歧过程.如果满足固定搭配和固定用法,则直接进行词义消歧,否则利用 LM 模型进行词义消歧.

为了测试基于生语料库的词性自动标注效果,我们构造了一个包括大约有两万词次的关于汽车配件的开放性测试集,分别对测试语料中的名词和动词进行词义自动标注.

表 1 给出了 4 个模型对名词词义消歧的实验结果.

Table 1 Experimental result of noun sense disambiguation

表 1 名词词义消歧实验结果

Computational model①	Number of nouns in the test corpus② N	Number of correct sense lables③ N_C	Correct rate④ CR (%)
WSD-1	2 684	1 351	50.3
WSD-2	2 684	1 542	57.4
WSD-3	2 684	2 198	81.9
WSD-4	2 684	2 274	84.7

①计算模型,②测试语料中名词的个数,③正确词义标注个数,④正确率.

表 2 给出了 4 个模型对动词词义消歧的实验结果:

Table 2 Experimental result of verb sense disambiguation

表 2 动词词义消歧实验结果

Computational model①	Number of verbs in the test corpus② N	Number of correct sense lables③ N_C	Correct rate④ CR (%)
WSD-1	955	534	55.9
WSD-2	955	575	60.2
WSD-3	955	741	77.9
WSD-4	955	813	85.1

①计算模型,②测试语料中动词的个数,③正确词义标注个数,④正确率.

下面通过设置不同上下文的长度为 0,5,10,30(表示整个句子),得出 4 个计算模型的词义消歧结果比较,如图 2 和图 3 所示.

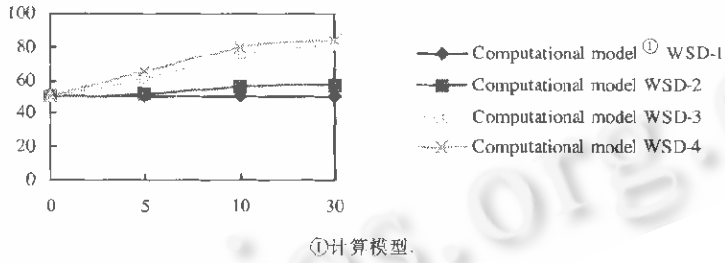

Fig. 2 Effect of length of context to noun sense disambiguation

图 2 上下文长度对名词词义消歧的影响

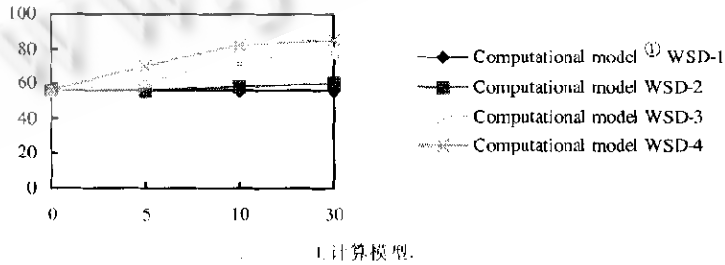

Fig. 3 Effect of length of context to verb sense disambiguation

图 3 上下文长度对动词词义消歧的影响

3 进一步讨论

从实验结果我们可以得出如下一些结论:

(1) 系统测试的语料是特定领域的语料,属于汽车配件领域.特定领域对于词性消歧具有指导作用,因为很多具有多词义的名词和动词在汽车配件语料中集中表现为某一特定的词义,其他词义很少出现.因此导致计算模型 WSD-1 的词义标注准确率比较高,名词的词义标注准确率为 50.3%,动词的词义标注准确率为 55.9%.

(2) "指示词"对于动词词义消歧的影响效果(+4.3%)没有对名词词义消歧的影响效果(+7.1%)好.而且仅仅根据这些"指示词"来实现词义消歧,效果不够理想.其中名词词义消歧准确率达到 57.4%,动词词义消歧准确率达到 60.2%.

(3) 动词的词义与上下文存在的"语义巧合"关系没有名词明显.计算模型 WSD-3 中名词词义消歧准确率(81.9%)比动词词义消歧准确率(77.9%)要高 4%.

(4) 从计算模型 WSD-4 和计算模型 WSD-3 的实验结果比较可以发现,固定搭配、固定用法对于动词词义消歧的影响效果(+7.2%)比对名词词义消歧的影响效果(+2.8%)要好.

(5) 从图 2 和图 3 中可以明显发现,全局上下文比局部上下文效果要好.

4 结束语

本文提出了一种对数模型,构造了一个词义自动消歧系统 LM_WSD.目前,该词义自动消歧系统 LM_WSD 已经应用于基于词层的英汉机器翻译系统(汽车配件专业领域)中,运行状态良好.但是,从实际应用中发现还存在一些不足之处.由于真实文本中存在很多未登录词汇,对于未登录词汇的词义标注一般采用猜测的方法,目前效果不是很理想.实际上,动词的词义消歧最好可以利用一些句法分析结果,例如,动词的名词性宾语等信息.目前在我们的系统中,首先进行词义消歧过程,然后进行句法语义分析.将来可能考虑将词义消歧过程与句法语义分析结合起来,这样有助于词义消歧的效果.

References:

[1] Katz, J. J., Fodor, K. A. The structure of a semantic theory. Language, 1963,39(2):170~210.

[2] Mooney, R. J. Comparative experiments on disambiguation word senses: an illustration of the role of bias in machine learning. In: Church, K., ed. Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing. New York: Academic Press, 1996. 82~91.

[3] Rivest, R. L. Learning decision lists. Machine Learning, 1987,12(1):299~246.

[4] Roberto, B. Towards a bootstrapping framework for corpus semantic tagging. In: Nancy, V., ed. Proceedings of the ACL-SIGLEX Workshop. Washington: Morhan Kaufmann Publishers, Inc., 1997. 66~73.

[5] Erwin, R. The mechanical determination of meaning. In: Locke, W. N., Booth, D. A., eds. Machine Translation of Languages. New York: John Wiley Press, 1995. 136~164.

[6] Kurohashi, Sadao, Nagao, Makoto. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. IEICE Transactions on Information and Systems, 1994,77(2):227~239.

[7] Stetina, Jiri, Nagao, M. Corpus-Based PP attachment ambiguity resolution with a semantic dictionary, In: Zhou, J., ed. Proceedings of the 5th Workshop on Very Large Corpora. Beijing: Tsinghua University Press, 1997. 66~80.

Logarithm Model Based Word Sense Disambiguation*

ZHU Jing-bo, LI Heng, ZHANG Yue, YAO Tian-shun

(Institute of Computer Science, Northeastern University, Shenyang 110006, China)

E-mail: cipol@nlplab.com

http://www.nlplab.com

Abstract: In this paper, a method for automatic word sense disambiguation based on logarithm model (LM) is discussed, and a word sense disambiguation system LM_WSD is implemented. In the experiments, four models are used to word sense disambiguation. Experiments showed the effect of high-frequency sense, salient words, specialized field and general usage to noun and verb word sense disambiguation. Now the system LM_WSD was applied in a word-based English-Chinese machine translation system for car fittings field, and improved the performance of the system.

Key words: word sense disambiguation; machine translation; logarithm model; natural language processing

* Received July 20, 1999; accepted May 10, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69985001; the National Grand Fundamental Research 973 Program of China under Grant No. G19980305011; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 1999014503