基于隐最大熵原理的汉语词义消歧方法*

张仰森，黄改娟，苏文杰
北京信息科技大学 智能信息处理研究所，北京 100192
E-mail: bistu_syz@163.com

摘 要：本文针对最大熵原理只能利用上下文中的显性统计特征构建语言模型的特点，提出了采用隐最大熵原理构建汉语词义消歧模型的方法。在研究了《知网》中词语与义原之间的关系之后，把从训练语料获取的文本上下文中的词语搭配信息转换为义原搭配信息，实现了基于义原搭配信息的文本隐性语义特征提取方法.在结合传统的上下文特征后，应用隐最大熵原理进行文本中多义词的词义消歧。实验结果表明，采用文中所提方法对 10 个多义动词进行词义消歧，正确率提高了大约 4%。

关键词：隐最大熵原理；文本隐性特征；义原搭配信息；词义消歧

The Method of Chinese Word Sense Disambiguation Based on Latent Maximum Entropy Principle

Zhang Yangsen, Huang Gaijuan, Su Wenjie
Beijing Information Science & Technology University, Beijing 100192
E-mail: bistu_syz@163.com

Abstract: the Maximum entropy We present a new approach for Chinese word sense disambiguation based on latent maximum entropy principle(LME), LME is different from Jaynes' maximum entropy principle that only use the context statistical characteristics to construct language model. After studying the relationship between the words and the sememes in Hownet, we convert the words collocation that obtained from the context of training corpus into sememes collocation, and realize the extracting of text latent semantic features based on sememe collocations. On the basis of combining traditional context features, we use the latent maximum entropy principle to disambiguate polysemy words. Experimental results show that the method proposed in this paper for the sense disambiguation of 10 polysemous verbs word, the correct rate increased by about 4%.

Keywords: latent maximum entropy principle; text latent features; sememes collocation information; word sense disambiguation

1 引言

多年以来，关于汉语词汇的语义消歧研究一直是自然语言处理领域的热点，其中，基于最大熵原理[1]的词义消歧模型由于可以将多种上下文特征集于统一的模型框架之中，词义消歧效果比较好，受到学界的广泛应用。最大熵模型与 n-gram 模型相比，能够获取和使用自然语言多个方面的信息特征，将多种特征信息集成于一个模型之中，与朴素贝叶斯模型、决策树等统计语言建模方法相比，有无需独立性假设及自动特征权重确定的优点。但其主要缺点是只能处理显性统计特征信息，对那些自然语言中经常遇到语义和句法信息无法进行处理。

为了将自然语言中人们不能直接观察到的隐性特征，如语义信息或语法结构引入最大熵方法，本文提出基于隐最大熵原理的词义消歧方法，将语义搭配特征等隐性特征与显性统计特征等一同引入一体化的指数性概率框架模型之中，以提高汉语词汇的语义消歧正确率。

* 基金项目：国家自然科学基金(60873013，61070119)；北京大学计算语言学教育部重点实验室开放课题基金(KLCL-1005)；北京市属市管高等学校人才强教计划资助项目（PHR201007131）

2 隐最大熵原理[2]

在 Jaynes 提出的最大熵模型（ME 模型）中[1]，由于使用的是显性统计特征，因此，在进行模型参数估计时，可以使用最大似然估计法来计算训练语料中的概率分布，对模型参数进行估计。然而，对于真实的自然语言来说，除了词语、词性标注等显性统计特征以外，还有句法和语义特征，如何将句法语义特征融入最大熵模型以提高模型的效率，shaojun wang 等于 2002 年提出了隐最大熵原理[2]，提出了将句法语义信息融入模型的方法。

设 $X\in\Phi$ 是概率为 p(X)的完全数据，Φ 为一自然语言，$Y\in\Psi$ 是可观察的非完全数据，Ψ 表示词、句、文档等，并且 Y=Y(X)是一个从 Φ 到 Ψ 的多对一映射，丢失的数据在文档级为语义内容，在语句级为句法结构。

设 P(Y)表示 Y 的概率，P(X|Y)为给定 Y 条件下的 X 的条件概率。则：

$$P(Y)=\sum\Phi(Y)p(X)$$

这里，$\Phi(Y)=\{X: X\in\Phi, Y(X)=Y\}$,并且 p(X)=p(Y)p(X|Y)

具有隐变量的最大熵原理的问题是从一组允许的概率分布中选择一个模型 p，使其具有最大的熵：

$$H(p)=-\sum_X p(X)\log p(X) \tag{1}$$

服从 $$\sum_X p(X)f_i(X)=\sum_{y\in\Psi}\overline{p}(y)\sum_{X\in X(Y)}p(X\mid Y=y)f_i(X) \quad i=1,2,\ldots,N \tag{2}$$

这里 $\overline{p}(y)$ 是一组可观察的训练样本 $y_1,y_2,\ldots,y_c$ 的经验分布，由 $\overline{p}(y)=\frac{C(y)}{C}$，$C(y)=\sum_{i=1}^{C}\delta(y,y_i)$ 为训练样本中 y 的出现次数,$f_i(x)(i=1,2,\ldots,N)$为一组特征，p(X|Y=y)将层次依赖结构引入了统计模型。注意，某些特征是可观察数据 Y 的函数，即 $f_j(X)=f_j(Y)$。在这种情况下，约束条件被消解为通常的 $\sum_{Y\in\Psi}p(Y)f_j(Y)=\sum_{y\in\Psi}\overline{p}(Y=y)f_j(Y=y)$。

对隐变量没有约束，最大熵的解将把相等的概率分配到各隐变量上去，如果没有丢失数据，则问题将被简化为 Jaynes 模型，因此，（2）式比 ME 具有更一般的描述。

3 特征表示与特征提取

随机过程的输出与上下文信息 x 有关，但在建立语言模型时，如果考虑所有与 y 同现的上下文信息，则建立的语言模型会很繁琐，而且从语言学的知识上来讲，也不可能所有的上下文信息都与输出有关。所以在构造模型时，只要从上下文信息中选出与输出相关的信息即可，称这些对输出有用的信息为特征。

特征表示由两部分构成，一部分是目标类的上下文语境 x，另一部分是目标类 y。为了让模型能够理解特征，可以使用特征函数来表示（x,y）的特性。定义一个{0,1}域上的二值函数来表示特征：

$$f_i(x,y)=\begin{cases}1, & \text{if}(x,y)\text{满足某些条件}, i=1,2,3,\ldots,n\\ 0 & \text{否则}\end{cases} \tag{3}$$

特征的选择与提取可通过特征模板的方法来实现，在设计模板时可将影响多义词词义的上下文距离信息以及特定位置上的词性信息考虑进来。一般考虑的因素有：（1）特征类型，包括词形（Word）、词性(Pos)、词形+词性；（2）窗口大小，包括语句中当前词前后的 n 个词；词形特征表示使用 Word+Index 的形式，词性特征表示法与词形类似。这里 Word 用字母 W 表示，Index 为特征词相对于当前词的位置。本文中所设计的特征模板如表 1 所示。

表 1　特征模板设计

模板 ID	模板形式	模板含义
1	Word(0)	当前词
2	Word(+1)	当前词右边第一个词
3	Word(-1)	当前词左边第一个词
4	Word(+2)	当前词右边第二个词
5	Word(-2)	当前词左边第二个词
6	Pos(0)	当前词词性
7	Pos(+1)	当前词右边第一个词词性
8	Pos(-1)	当前词左边第一个词词性
9	Pos(+2)	当前词右边第二个词词性
10	Pos(-2)	当前词左边第二个词词性

利用特征模板所得到的候选特征集合比较大，需要采用特征筛选方法从中筛选出对输出影响较大的特征。本文采用特征频次和互信息相互结合的特征选择方法进行特征筛选。依据特征模板进行特征提取过程如算法 1 所示。

算法 1　特征提取算法

Step1: 从第一句开始扫描语料库;

Step2: 循环特征模板中的特征列表，利用当前模板开始匹配特征并进行提取，命名为 feature;

Step3: 查看特征文件中 feature 是否存在，如果已经存在，特征数目加 1，转到 Step2; 如果不存在，将 feature 写入特征文件，转到 Step2;

Step4: 是否扫描到语料库结尾，如果是，结束; 否则，转到 Step1 继续扫描。

（1）特征频次筛选法。特征频次筛选法就是计算特征集中每个特征出现的次数，并根据实验需求设定一个阈值，把出现次数较少的特征舍弃[3]。

（2）互信息选择法。互信息是用来衡量两个变量之间的相关度的量[4]。词义消歧中可以使用互信息来表示特征词与多义词之间的相对语义距离。计算公式如下：

$$MI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{4}$$

P(w1)、P(w2)和 P(w1,w2)分别是词语在语料库中出现的概率和共现概率。根据计算结果，选择满足一定互信息要求的特征。

4　基于义原搭配信息的文本隐性特征提取

利用上述设计的特征模板提取的上下文特征属于显性统计特征，是比较容易获取的，如果上下文的窗口宽度选择的比较大的话，其特征数量将是相当大的，参数空间也会非常大，使建模的工作量增大。所抽取的特征反映的是上下文中词语与当前词之间的词语搭配特征，而更深一层次的语义特征被忽略了。借助《知网》，词语搭配之间更抽象一层关系能够被抽取出来，这就是义原搭配信息[5]。为了避免算法过于复杂，本文只考虑从动宾结构中抽取义原搭配特征。将动宾结构中的两个词语之间的二元搭配组合转变为多个义原之间相互制约的多元组合。这样就丰富了文本特征所涵盖的语义信息。

表 2 中给出了义原搭配的例子。多义动词为“吃”，可能的宾语为“老本”、“利息”、“面包”、“饭”、“汽油”等。在传统最大熵模型中，这些搭配信息都会被考虑到。但如果借助《知网》，就能够抽取出义原搭配的信息，获取到语义搭配特征。表 2 中词义 ID 是表示当前多义动词“吃”的

义项编号。表中第三列和第四列分别是从《知网》中抽取出来的动词和宾语的义原信息。这种义原信息可以反映出上下文的语义搭配特征，大大较少最大熵模型的特征数量，缩小参数空间，优势是显而易见的。

表2　义原搭配示例

动词	宾语	词义ID	动词主义原	义原搭配信息	词义所对应的隐性特征
吃	老本	2	absorb|吸收	fund|资金	资金
吃	老本	2	absorb|吸收	finance|金融	金融
吃	利息	2	alive|活着	wealth|钱财	钱财
吃	面包	4	eat|吃	food|食品	食品
吃	饭	4	eat|吃	edible|食物	食物
吃	汽油	5	exhaust|损耗	material|材料	材料

义原搭配信息能够表征语义特征，但如何获取和存储语义搭配特征就成为关键。下面以动宾结构短语为例，给出获取和构建义原搭配信息数据库的方法，如算法 2 所示。在本算法中暂不考虑动宾结构中动词和名词均为多义词的情况。

算法2　义原搭配信息数据库的构建算法

Step1: 从训练语料中抽取动宾结构搭配词语，作为义原搭配信息抽取的对象。

Step2: 在《知网》知识库中查找动词条目。以“展开”为例，查找“W_C=展开”，若存在，判断词性是否为动词，即 G_C 的值是否以“V”开始，若是，则跳到下一步；若不是，则返回 step2 继续查找；若文件结束，则返回。

Step3: 在“DEF”中读取动词概念中的第一义原，记作 Verb_DEF。如果动词在《知网》知识库中具有多个概念，则抽取训练语料中与动词所标注词义相一致的概念所在的义原。

Step4: 在《知网》知识库中查找名词条目。以“地图”为例，查找“W_C=地图”，若存在，判断其词性是否为名词，即 G_C 的值是否以“N”开始，若是，则执行下一步；若不是，则继续执行查找；若文件结束，则返回。

Step5: 在“DEF”中读取名词概念中的第一义原、领域义原和主体义原，分别记作 Nouns_Sememe_First，Nouns_Sememe_Domain，Nouns_Sememe_Host。如果领域义原或主体义原不存在，则赋值空串。

Step6: 更新数据库操作。将 step2 和 step5 中所抽取的信息插入到数据库中。

Step7: 如果还存在未处理的动宾结构搭配词语跳转 step2，否则，结束。

生成的义原搭配信息将被存储于 MySQL 数据库中。数据库建立完成之后，义原搭配信息所在的数据表中第三列（Verb_Word）为多义动词原型；第五列（Feature_Verb_Sememe）为多义动词的义原信息；第六列（Feature_Nouns_Sememe）为多义动词的义原搭配信息；最后一列（Sence_ID）为动词的义项标示。将最后两列按照一定的格式，输出到文本文件中，就可以作为隐性特征供词义消歧模型来使用。

5　基于隐最大熵原理的词义消歧实现

最大熵模型的缺点是它只考虑了目标词所在上下文中的显性特征[6]。隐最大熵模型是在最大熵模型基础上考虑了隐性特征，将显性特征和隐性特征相结合应用于消歧模型。本文通过《知网》从词语搭配中所获取义原搭配是一种语义搭配特征，它将最大熵模型的特征空间变成了语义类的特征空间，从而使参数空间大大缩小，提高了最大熵参数估计算法的效率和词义消歧的准确率。

本文设计并实现了一个词义消歧实验系统，该系统包括三个模块：机器学习模块、词义消歧

模块、结果评测模块。

机器学习模块主要包括文本预处理、特征提取、模型参数计算等操作。文本预处理主要的功能是去除停用词和非法字符等。特征提取包括显性特征提取和隐性特征提取。显性特征依据算法 1 按照所设计的特征提取模板来实现，隐性特征的提取则依据《知网》，按照依据算法 2 实现。模型参数的训练使用隐最大熵原理来实现，输出的模型参数信息将保存在文本文件中供下一步中的预测模块来使用。

词义消歧模块用来对待消歧的文本进行词义消歧。本模块中文本预处理过程与机器学习模块相同。特征提取模块提取多义动词所在上下文的特征词语，频次和互信息相结合的方法进行来进行特征筛选，同时提取该多义动词的宾语，并获取机器学习模块所获得的义原搭配信息，最后根据模型参数与所选特征，计算出该多义动词的可能词义。

结果评测模块是通过将机器标注的语料与人工标注的语料进行比较，对词义标注模型与算法的性能进行评价。

6 实验结果与分析

6.1 系统实现工具与实验语料的选择

所实现的词义消歧系统使用 Java 语言开发，开发环境是 Eclipse。数据库使用 MySQL3.5 版本。数据库设计工具使用 MySQL Workbench5.0。

我们选取了由北京大学计算语言所和富士通公司共同制作的 2000 年 11 月和 12 月《人民日报》基本标注语料作为实验语料。其中 50 天的语料作为模型参数训练语料，剩下 10 天语料作为测试语料。

为了使实验简单，我们从确定的语料中选取 10 个多义动词进行实验。选取目标多义词的原则如下：

（1）目标词应当具有多于一个词义；

（2）应当选取出现次数较多的动词，一般来说，出现的次数越多越好；

（3）多义词的某一词义在所有词义中所占的比重不应当太大。比如，某个动词有 3 个词义，其中一个词义所占比重达 90%，其他两个词义只占 10%，剩余两个词义的区分将变得十分困难。

选定的多义动词及其在语料中出现的次数如表 3 所示。所选动词的词义数目为 2、3、4，在统计词义的过程中，我们发现，所用的北大人民日报基本标注语料库的义项数与《知网》中所列义项数并不完全一致。我们以人民日报语料所标的义项为准。

表 3　多义词表

目标词	词义数目	训练集中出现的次数	测试集中出现的次数
用	3	412	88
表示	2	1120	135
发动	3	97	15
出	4	336	41
补	3	65	13
想	4	271	42
要	3	196	26
让	2	1238	176
写	2	624	110
发表	2	864	105

6.2 实验结果与分析

系统采用准确率、召回率和 F 值对实验结果进行评测。对测试语料去除义项标注后，进行义项标注的测试。实验结果按未使用义原搭配信息和使用义原搭配信息来进行分类。实验系统运行结果如表 4 所示。

表 4　多义动词消歧结果

目标词	词义数目	未使用义原搭配准确率	使用义原搭配后准确率	使用义原搭配后召回率
用	3	0.750	0.792	0.74
表示	2	0.811	0.852	0.82
发动	3	0.814	0.861	0.81
出	4	0.721	0.701	0.64
补	3	0.810	0.852	0.78
想	4	0.795	0.790	0.72
要	3	0.875	0.895	0.85
让	2	0.833	0.885	0.84
写	2	0.75	0.845	0.85
发表	2	0.880	0.933	0.89

上面的表格显示出不同多义词在使用义原搭配信息和不使用义原搭配信息情况下的正确率对比。从数据表 4 中可以看出：

（1）使用义原搭配隐性特征后，系统词义消歧的平均准确率为 84.06%，比未使用义原搭配信息前提高了大约 4 个百分点。

（2）系统对义项数目较少的多义词，消歧结果较好，比如“发表”、“表示”、“发动”等，而当多义词义项数目较多时，消歧的结果稍差。分析原因主要有两点：a）对于某些词，如“发表”，在人民日报的语料中有其固定的搭配。《人民日报》不是小说，一些拟人、虚构等手法在人民日报中并不会出现。人民日报语料中更多的是关于政治、事实的报道，一些固定搭配可能对词义消歧产生较大影响。比如：例句①：表示/v!1　亲切/a　的/ud　问候/vn　！/wt

例句②：按照/p　“/wyz　三/m　个/qe　代表/v!2　”/wyy　的/ud　要求/n

对于例句①和例句②中“表示”、“代表”的消歧，固定搭配将会起到关键性作用。b）当多义词义项数目较多，而在训练集或测试集中出现的次数较少时，由于语料的不充分造成的准确率不高。

（3）有少量词在使用义原搭配信息后并未呈现出较好的结果，比如“出”，“想”，分析其原因，可能的因素有两个方面：一是多义词在语料中出现的次数较少造成的，二是可能多义词词义较多，系统抽取义原搭配信息的结果会导致其中某两个词义或多个词义出现义原搭配相同或相似的情况，对词义消歧产生混淆作用，从而导致消歧的准确率下降。

参 考 文 献

[1] Jaynes, E. T. Information Theory and Statistical Mechanics. Physical Review, 1957, 106 (4): 620-630.

[2] Shaojun Wang, Dale Schuurmans, Yunxin Zhao. The Latent Maximum Entropy Principle. IEEE International Symposium on Information Theory, 2002: 182-185.

[3] 代六玲，黄河燕，陈肇雄．中文文本分类中特征抽取方法的比较研究[J]．中文信息学报，2004，18(01)：26-32.

[4] 王国胤，于洪，杨大春．基于条件信息熵的决策表约简[J]．计算机学报，2002，25(07)：759-766.

[5] 郭充，张仰森．基于《知网》义原搭配的中文文本语义级自动查错研究．计算机工程与设计，2010.9,Vol.31(17): 3924-3928.

[6] 张仰森．基于最大熵模型的汉语词义消歧与标注方法[J]．计算机工程，2009(9)：15-18.