基于最大熵原理的汉语词义消歧*)

陈笑蓉　秦　进
(贵州大学信息与计算机科学学院　贵阳550025)

摘　要　词义消歧是自然语言处理中亟待解决的一个关键问题,本文提出一种基于最大熵模型的有监督的机器学习方法,用于汉语词义消歧。该方法综合了词标记、词性、主题等上下文特征,并用一种统一的表示方法规范化特征形式,解决了多种不同特征之间的融合和特征的知识表示。实验对20个汉语高频多义词进行了测试,平均正确率为87%,验证了该方法的有效性。

关键词　词义消歧,最大熵模型,有监督机器学习

Maximum Entropy-Based Chinese Word Sense Disambiguation

CHEN Xiao-Rong　QIN Jin
(College of Information Computer Science, Guizhou University, Guiyang 550025)

Abstract　Word sense disambiguation is a crucial problem to be solved in NLP. A supervised machine learning method is proposed in this paper, which is applied in word disambiguation in Chinese. The method combines various features in context to disambiguate word senses. The features include annotations of words, parts of speech and subjects etc. And a uniform representation formalizes the features. In this way, the problem of synthesis among various features and knowledge representation of feature will be solved. 20 Chinese polysemous words are tested in our experiment. The result with average precision 87% shows that the method is effective.

Keywords　Word sense disambiguation, Maximum entropy models, Supervised machine learning

1　引言

词义歧义普遍存在于各类语言。词义消歧,即是解决多义词在上下文中的归属问题。由于词义分类种类繁多、无统一标准、词义间关系错综复杂、不易描述,词义消歧始终是自然语言处理中的一个难题。

从机器学习的角度看,有监督的词义消歧是个统计分类问题。设待消歧词为 W,词义集合为 $S_i(I=1,2,\cdots,n)$,当前上下文为 X,若 W 的词义为 S_i,则词义消歧将 X 归为支持 S_i。分类的数据为 X,待求的是 $Pr(S_i|X)$,这里 Pr 为概率。

在机器学习的分类问题中,数据通常用一系列(可能大量)特征来刻画。词义消歧中可利用的特征,是上下文信息中能够支持消歧的信息点,包括词搭配、词性标记、关键词、语法关系、领域信息等。这些特征数目多、来源不同(语法、语用)、相关性强(如词性标记序列和局部语法结构),且不同词的有效特征集之间区别也很大。因此,词义消歧要求语言模型能将来自不同知识源的信息(这里指特征)结合起来为不同词分别构造分类器。

针对这一需要,通常做法是,按照不同知识源分别建立语言模型,再用插值法或 BackOff 法将各分类模型组合得到一最终模型。这样的最终模型内部并不一致,对各分模型的利用也达不到最优。

在最大熵方法看来,各特征从不同方面刻画了待建模型必须满足的性质,每个特征都是施加于待建模型上的一个约束集。依最大熵方法建立单一模型,即最大熵模型,既同时满足各特征的约束,内部又一致。而且,最大熵模型能够方便地加入各种新特征,即使这些新特征缺乏知识的支持。

在西语中,用最大熵模型进行词义消歧已有广泛研究。在汉语中,这方面的研究尚少。本文详细介绍了自然语言处理中的最大熵原理,建立最大熵模型的方法,最大熵模型如何用于汉语词义消歧。最后,建立实验系统对最大熵词义消歧的效果进行了验证,并给出总结和展望。

2　ME(最大熵)框架

2.1　ME 原理

熵这个概念,源于物理学,指系统的混乱程度。熵理论在统计物理学中获得成功,并引起了各个学科的关注。Shannon 将熵概念引入了信息论,称为信息熵。信息论中,用最大信息熵原理,信息论解决了一大类在先验知识不充分的条件下进行决策或推断的问题,如谱估计、图像重建等。以下提到的熵和最大熵原理,指的都是信息论中的熵和最大熵原理。

统计自然语言处理中,视自然语言中各现象的产生为一随机过程,要求建立的统计模型对该过程具有一定的预测能力。通常情况下,预测是根据已知,推测未知,要点是正确和客观。对应到统计自然语言处理中,已知的是事先收集的训练样本数据,未知的是将来的测试样本,"正确"是待建模型与已知样本不矛盾,"客观"是对未知的样本不做任何假设。这种预测的要点是,满足一定限制的前提下,保留尽可能多的不确定性。

从信息论的角度看,视自然语言为一未知信源,则已知样本体现了有关该信源的不完全的知识。待建模型一方面要引入这些不完全的知识,另一方面要避免对未知的知识施加任何主观猜测。这正体现了最大熵原理。

最大熵原理,从直觉上讲,就是将已知的知识建模,对未知的不做任何假定。换言之,给定事实的集合,选择一个模型,该模型与所有已知事实一致,对未知的可能发生的事实,模型所赋予的概率分布均匀。最大熵原理的动机是保留尽可能多的不确定性。

*)基金项目:贵州省自然科学基金资助项目。

2.2 最大熵模型的参数估计

最大熵模型的实施包括两个步骤:特征选择和参数估计。特征选择的任务是选出对模型有表征意义的特征,参数估计用最大熵原理对每一个特征进行参数估值,使每一个参数与一个特征相对应,以此建立所求模型。特征选择本身不属于最大熵原理的内容,且与特定应用有关,后面会针对词义消歧专门讲解。这里专讲参数估计。

在自然语言处理这一随机过程中,所有最终输出值构成了语言学类别有限集 Y。对于每个 $y\in Y$,其生成的均受上下文信息 x 的影响和约束。已知与 y 有关的所有上下文信息组成的集合为 X,则模型的目标是,给定上下文 $x\in X$,计算输出为 $y\in Y$ 的条件概率 $p(y|x)$。

最大熵模型的输入是从语料中抽取的训练样本集:

$$T=\{(x_1,y_1),\cdots,(x_N,y_N)\}$$

这里$(x_i,y_i)(i=1\cdots N)$的含义如前所述。用先验概率分布表达这些数据:

$$\tilde{p}(x,y)=freq(x,y)/N$$

$freq(x,y)$表示(x,y)在观测数据中出现的次数。

y 的生成与其上下文 x 的部分信息有关,从 x 中找出对 y 的取值有用的知识才是模型所追求的目标。这有用的部分知识也就是最大熵方法所要寻找的特征。

每个特征须表示成一个二值约束函数的形式。每个约束函数与特定的类别 y' 相联系,并取等式(1)的形式,其中 $cp:X\rightarrow\{\text{true},\text{false}\}$为文本中可观察的特征,又称上下文谓词:

$$f(x,y)=\begin{cases}1 & \text{if } y'=y \text{ and } cp(x)=\text{true}\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

训练样本中 f 相对于分布 $\tilde{p}(x,y)$的期望值为:

$$\tilde{p}(f)\equiv\sum_{x,y}\tilde{p}(x,y)f(x,y) \quad (2)$$

f 关于待建模型 $p(y|x)$的期望值为:

$$p(f)=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \quad (3)$$

限制这个期望值与训练样本中 f 的期望值相等,即:

$$p(f)=\tilde{p}(f) \quad (4)$$

(2)、(3)、(4)结合得到:

$$\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \quad (5)$$

称式(5)为一个约束等式,在式(5)约束下的模型 $p(y|x)$ 包含了特征 f,即基于观测数据对模型施加了单个特征约束。

假设有 N 个特征函数 $f_i,i=1,\cdots,N$。表示 N 种有关特征的随机模型应满足所有 N 种模型的约束。这些约束可表示为:

$$C=\{p\in P|p(f_i)=\tilde{p}(f_i),i\in\{1,2,\cdots,N\}\} \quad (6)$$

满足上式约束的模型有很多。根据最大熵原理,应选取概率均匀分布,即使条件熵 $H(p)$取最大的模型 p^*:

$$H(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x)$$

$$p^*=\arg\max_{p\in C}H(p)$$

已有成熟算法求解 p^*,这里不再赘述。

2.3 词义消歧 ME 模型中的特征表示

ME 模型可用于自然语言处理中的许多方面,如文本分类、文本校对等。场合不同,ME 模型的特征也不同。在加入最大熵模型前,这些特征必须先表示成二值约束函数的形式,如式(1)。为了将特征形式规范化,对特征做以下分析。

特征描述的是上下文中某区域中某特性具有的属性。这里的区域可以是一个点,如待消歧词的前接位置,也可以是一个区间,如宽度为3的上下文窗口。属性是一个值,也是值集合中的一个元素。据此,用下列4种表达式将特征的表现形式统一起来:

$info(x,i)=a$

$info(x,i)\in C$

$info(x,i,j)=a$

$info(x,i,j)\in C$

上面4种表达式中,i、j 都是上下文窗口中相对于待消歧词的偏移,可取非零的正负整数值。a 表示属性的值,C 表示属性值的集合。

以下分别解释这4种表达式的用处:

$info(x,i)=a$ 表示出现于上下文位置 i 处的属性值为 a。如 $POS(x,-1)=$ 'adj' 表示待消歧词前接词的词性标记为 'adj'。

$info(x,i)\in C$ 表示出现于上下文位置 i 处的属性值为属性值集合 C 的元素。如 $POS(x,-1)\in\{$'adj','$verb$'$\}$表示待消歧词前接词词性标记为 'adj' 或 '$verb$' 的词。

$info(x,i,j)=a$ 表示某个上下文范围 i 至 j 处出现了属性值 a。如 $POS(x,-2,2)=$'adj' 表示待消歧词前后两个邻接词中有词性标记为 'adj' 的词。

$info(x,i,j)\in C$ 表示某个上下文范围 i 至 j 处的属性值中出现了属性值集合 C 的元素。如 $POS(x,-2,2)\in\{$'adj','$verb$'$\}$表示待消歧词前后两个邻接词中有词性标记为 'adj' 或 '$verb$' 的词。

使用时,用这4种表达式替换式(1)中的 $cp(x)$即可。需要指出的是,这里对特征表现形式的分析也适合于其它应用中的最大熵特征。

3 实验系统

实验选取20个汉语文本中经常出现多义词作为测试对象,采用上面建立的最大熵模型对实验文本中的多义词进行消歧,验证本文提出的方法的有效性。模型输入每个多义词 w 的 N 个训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$,其中$(x_i,y_i)$表示在 x_i 的上下文中 w 应取义项 y_i。

实验中采用的特征涵盖了上下文词和词性等语言信息。

3.1 实验数据

1. *词表* 采用《同义词词林》中的义类代码作为多义词的每个义项的编码,例如"人"有 Aa01、Ab02、Dd17、De01、Dn03共5个义项。

2. *训练语料* 训练样本选自1998年1月一个月的已经经过切分、标注的《人民日报》熟语料,用于抽取特征和训练特征的权值。手工标注每个训练样本中的多义词的义项编码。

3. *测试语料* 2000年《人民日报》生语料,从中选择需要测试的词语的测试句子。

3.2 实验结果及分析

1. *实例分析* 我们以一个例子来说明消歧的过程。例如有下面的测试句子,考虑其中的多义词"人",其义类编码集合 Y={ Aa01, Ab02, Dd17, De01, Dn03}。

何况弘扬科学精神,催人奋发进取,本身就是一种重要的正确舆论导向。

在训练阶段提取了1223个特征,经过特征选择后还有878个特征。经过训练得到每个特征的权值。然后计算出在这个句子中,"人"取各个义项的概率,如下:

$p(\text{Aa01}|x)=0.3821$,

$p(\text{Ab02}|x)=0.2011$,

$p(\text{Dd17}|x)=0.1189$,

$p(\text{De01}|x)=0.1619$,

$p(\text{Dn03}|x)=0.1359$,

所以在当前上下文的情况下，“人”应取编码为 Aa01的义项，完成歧义消除。

表1 多义词消歧结果

多义词	义类数	样本数	特征数	正确率
人	5	6499	1223	81.39
等	4	3130	1227	86.12
说	3	3482	1065	89.48
和	9	6519	1739	85.25
是	5	5837	1379	94.67
中	11	7083	1435	86.47
对	12	3851	1159	86.01
要	9	3813	837	86.57
与	6	2456	1027	84.30
在	7	7875	1499	87.27
地	7	4428	1144	89.97
以	6	4293	1225	90.45
将	6	2118	826	92.12
都	5	2218	891	92.76
还	5	2442	1088	86.12
把	12	1538	877	80.37
下	15	2866	873	92.61
于	6	2943	949	93.91
了	6	6524	1419	80.81
上	17	4984	1267	93.24

2. 实验结果 实验测试的20个多义词如表1第1列所示，第2列是每个多义词从《同义词词林》中获取的义类数，第3列是训练阶段的样本数，第4列是提取的特征数，最后1列是消歧的正确率。每个多义词使用了200个测试例。

实验获得了较好的结果，平均将近88%的准确率。从实验结果我们可以看出并不是训练阶段提取的特征越多越好，一个适度的特征数才能取得较好的正确率。

结语 至此，我们详细介绍了本文提出的基于最大熵原理的汉语多义词消歧方法。本文介绍了最大熵的原理，讨论了在最大熵框架下的特征选择，用一种统一的形式解决了各种特征的表示问题，综合利用各种语言信息进行消歧。实验证明本文提出的方法是有效可行的。

下一步的工作需要放在对汉语语言知识的研究上，怎样获取新的汉语语言知识，把它形式化后加入到模型中去，进一步提高消歧的正确率。

参 考 文 献

1 屈刚，陆汝占. 基于特征的汉语词性标注模型. 计算机研究与发展，2003，40(4)：556～561
2 李涓子，黄昌宁. 语言模型中一种改进的最大熵方法及其应用. 软件学报，1999(3)
3 李涓子，黄昌宁，杨尔弘. 一种自组织的汉语词义排歧方法. 中文信息学报，13(3)
4 A Maximum Entropy Approach to Natural Language Processing
5 A Maximum Entropy Approach to Adaptive Statistical Language Modeling
6 A Maximum Entropy-based Word Sense Disambiguation system

（上接第173页）

配失败则返回空记录，若进一步查找，则需用户手动更改查询条件，系统不具备自动推理功能。本文在基于本体和知识的论文检索体系的基础上，设计了一种智能化的论文检索算法：在

输入：查询关键词
输出：Φ或者满足条件的论文

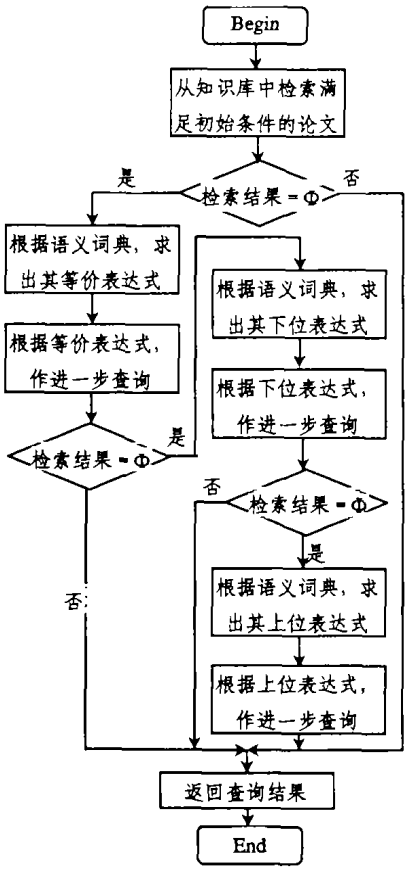

图4 智能化论文检索算法流程图

一般匹配失败的情况下，借助于本体和语义词典对论文关键词语义的刻画，系统能够自动寻找另一条合理的路径，进一步查找。

假设要检索有关“本体论在信息集成方面应用”的相关论文，先根据用户的检索要求，提取检索原始关键词，即“本体论”和“信息集成”。由于论文数据库中没有“精确”包含这两个关键词的相关论文，因此，基于传统信息检索技术的查询结果则为空。但根据语义词典，可知“本体论”和“信息集成”与其它关键词之间存在有丰富的语义联系，如“本体论”和（“本体”、“本体理论”、“信息知识本体论”、“元数据”、“语义网络”）之间存在同义关系等，因此，系统则根据关键词的语义关系（等价关系、上下位关系），自动更改查询条件，做进一步的查询。其具体的检索算法框图描述如图4。

结束语 有关本体论的研究在计算机界受到越来越多学者的重视，然而关于本体论的应用却还处在雏形阶段，没有统一的定义和固定的应用领域。本文在分析现有论文检索机制缺陷的基础上，探讨了将本体论应用于论文检索的新思路，建立了描述论文的本体，提出了基于本体论的论文检索系统模型，并在此基础上，研究了智能化论文检索算法，利用这种检索机制能有效提高系统的检索性能。

参 考 文 献

1 Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing. AI Magazine, 1991, 12(3):36～56
2 Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering, 1998, 25(122):161～197
3 Zhang N, Chen H, Wang Y, et al. Odaies: ontology-driven adaptive Web information extraction system. IAT 2003. 454～460
4 Buttler David, Liu Ling, Pu Calton. A Fully Automated Object Extraction System for the World Wide Web. In: Proc. of the 21st Intl. Conf. on Distributed Computing Systems, 2001. 361～371