第16卷 第3期
2006年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.3
Mar. 2006

一种新的汉语词义消歧方法

闫 蓉,张 蕾
(西北大学 计算机科学系,陕西 西安 710069)

摘 要:针对自然语言处理领域词义消歧这一难点,提出一种新的汉语词义消歧方法。该方法以《知网》为语义资源,充分利用词语之间的优先组合关系。根据优先组合库得到句中各个实词与歧义词之间的优先组合关系;将各实词按照优先组合关系大小进行排列;计算各实词概念与歧义词概念之间的相似度,以判断歧义词词义。实验结果表明该方法对于高频多义词消歧是有效的,可作为进一步结构消歧的基础。

关键词:词义消歧;优先组合关系;词关联;相似度;知网
中图分类号:TP391 **文献标识码**:A **文章编号**:1005-3751(2006)03-0022-04

New Chinese Word Sense Disambiguation Method

YAN Rong, ZHANG Lei
(Department of Computer Science, Northwest University, Xi'an 710069, China)

Abstract: WSD(word sense disambiguation)is a difficult field in natural language processing. This paper puts forward a new Chinese WSD method. This method imposes HowNet as the semantic knowledge resource, meanwhile fully uses PCR between words. Firstly, get PCR between the words through PCR database. Secondly, according to the size of the PCR, these notional words are put in order. Lastly, calculate the similarity degree between each concept of notional words and the concept of a word of different meaning, to obtain the right meaning of the word of different meaning. The experiment result proves this method to be effective for high frequency WSD. It can be the foundation of further structure disambiguation.

Key words: WSD; preference combination relation; word association; similarity; HowNet

0 引 言

词义消歧(WSD)就是让计算机能够处理和识别词的不同词义(sense),是为了解决自然语言中同形异义词在不同上下文环境中的义项标注问题。多义词分布的普遍性和无规律性决定了多义词消歧成为多种应用中的关键问题和难点之一。王惠提出了基于《现代汉语语法信息词典》和《现代汉语语义词典》利用汉语词义之间的多级组合特征进行词义消歧[1],但其方法仅对名词进行词义消歧。余晓峰等人提出一种简单无指导的词义消歧方法,只是单纯通过词语之间相似度计算来判断歧义词词义,而没有利用汉语词语之间的诸多关系,对一些高频歧义词的消歧没有达到预期的效果[2]。

文中提出一种新的词义消歧方法,利用汉语词语之间的优先组合关系,与词语之间语义计算相结合,以《知网》为语义资源,判断歧义词词义。实验结果表明这种方法对于高频多义词消歧是有效的。

收稿日期:2005-07-01
基金项目:陕西省教育厅专项科研基金资助项目(HD01302)
作者简介:闫 蓉(1979—),女,内蒙古鄂尔多斯人,硕士研究生,主要研究领域为人工智能及自然语言理解;张 蕾,博士,教授,主要研究领域为人工智能及自然语言理解。

1 词义消歧的基本思想

1.1 基本思想

汉语中有些词尽管具有类似的句法结构或语义,但是各自却存在着更为适宜的不同语境,即词与词之间存在着许多优先组合关系(PCR)。词义和词的分布之间具有密切的关系。一个词无论包含多少种意义,在一定语句中起作用的往往只是其中某一个意义。且词的不同意义往往会在句法或词汇搭配层面上表现出不同的组合特征,并且词与词之间存在着许多优先组合关系。例如,名词和名词间紧密的语义联系,形容词和名词组成的特定修饰关系,动词和名词的固定搭配等。有些多义词,其内部的不同意义虽然语法功能基本相同,但在句中出现时,所组合的词语却完全不同,即多义词在表现不同意义时候,与之进行组合的词语之间存在着优先关系。这种组合关系与多种因素有关,最重要的是不同词语之间的词义制约。当一些相互有关系的事物在词义中得到反映时,这些词就能够互相结合;反之,如果本来就是一些互相之间没有联系的事物,或它们的联系还没有在词义中得到反映,这些词就不能组合。多义词在表现不同意义的时候必定是和不同的词一同出现的,则根据和该多义词一起出现的其它词,就可以相对准确地判断出该多义词的真实意义。且在一般情况下,歧义词与它上下文中前面与后面一个或几个实词

的关系最为密切,因此,完全可以根据这几个实词来确定出该词的词义。

1.2 词义消歧步骤

现有的汉语词类自动标注的正确率已经达到96%以上,因此对于多义词的不同意义属于不同的词类,计算机完全可以借助语料中的词类标注判断其正确意义。王惠在对200万字的《人民日报》语料(1998年1月)统计中发现,22744个名词中共有多义词2196个,其中意义词类不同的有592个,占27%[2],说明仅仅利用词类标记就可以消除超过1/5的汉语歧义。文中提出的词义消歧方法是在确定词语词性的前提下,针对多义词为不同意义且属于相同词类进行的。消歧步骤如下:

a.读入源语言文本。对源语言文本进行分词和词类标注。

b.将分词和词类标注完毕的文本进行初始词义标注。

c.对词类不同且意义不同的多义词,根据词类标注信息进行消歧,得到待词义标注的文本。

d.词义消歧过程。将含有歧义词的文本取出,依据优先组合库,得到句中各实词与歧义词优先组合关系,并将这些实词按照优先组合关系大小排列,利用《知网》作为语义资源,计算各实词概念与歧义词概念之间的相似度,选择相似度取最大值所对应的歧义词概念,作为歧义词的词义,完成词义消歧,结束。

下面给出基于优先组合关系的汉语词义消歧模型,见图1。

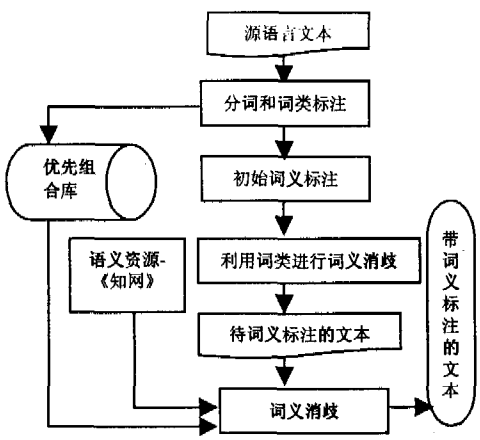

图1　基于优先组合关系的汉语词义消歧模型

2 词义消歧

歧义消解的前提是为歧义词选择恰当的上下文,一个多义词的具体语义一般受到一定语境或上下文关联词的限制。文中选择歧义词所在句子作为上下文。

2.1 语义资源——《知网》

《知网》是一个以汉英双语为代表的概念以及概念的特征为基础的,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。它是一个网状的有机的知识系统[3],它提供了词之间的聚合关系,提供了一个较详细的词间联系的网络。采用不易分割的意义的最小单位——义原对所有概念进行定义,从而使词具有很强的可计算性。《知网》知识库包括:a.中英双语知识词典;b.义原分类源文件;c.知网管理工具;d.知网说明文件等。文中相关的《知网》知识库文件主要有a和b。

2.2 优先组合库的获取

通过词对之间的词关联计算获得词语优先组合信息的优先组合库。选择清华大学与北京语言大学语言信息处理研究所提供的标注好的语料库(约191万字),统计得到各词的优先组合库。优先组合库中将各个词按照其词性的不同,将与其关系紧密的词按照互信息大小,进行链式存储。

所谓相关信息,统计学中又称互信息(mutual information),是用来衡量两个汉字串的相关程度,计算窗口中各个词对的相关程度。考虑多义词 x 和词 y,相关信息 $I(x;y)$ 就反映了两个词之间的词相关程度,其计算公式为:

$$I(x,y)=\log_2[\frac{P(x,y)}{p(x)\cdot P(y)}] \tag{1}$$

若 x,y 分别表示两个不同的单词,则 $I(x,y)$ 体现了多义词 x 和词 y 信息的相关程度,其中 $P(x),P(y)$ 分别表示 x 和 y 在语料中出现的概率,$P(x,y)$ 是 x 和 y 同时在语料中出现的概率。

2.3 应用语义资源《知网》进行词义消歧

词义和词的分布之间具有密切的关系。一个词无论包含多少种意义,在一定语句中起作用的,往往只是其中某一个意义。多义词在表现不同的意义的时候必定是和不同的词一同出现的,则可以根据和该多义词一起出现的其它词,即关联度高的词,就可以相对准确地判断出该多义词的真实意义,并且词的不同意义往往会在句法或词汇搭配层面上表现出不同的组合特征,并且词与词之间存在着许多优先组合关系。

下面给出利用优先组合关系进行词义消歧的过程。

定义　实词概念相似度是指一个实词的某个概念与另外一个实词的某个概念之间相似的程度。

设歧义词 W 有 n 个概念:$k_1,k_2,\cdots,k_n(n\geqslant 2)$。将句子中除歧义词外的 T 个实词(包括名词、量词、代词、动词、形容词、成语、简称略词、习用语等)取出。依据优先组合库,得到这 T 个实词与歧义词的优先组合关系大小,并按照与歧义词优先组合关系的大小逐一进行排列,得 $W_1,W_2,\cdots,W_m(1\leqslant m\leqslant T)$。设这 m 个实词分别有 $r_1,r_2,\cdots,r_m$ 个概念($r_i\geqslant 1,1\leqslant i\leqslant m$),其中实词 $W_i(1\leqslant i\leqslant m)$ 的 r_i 个概念分别为:$k_1^i,k_2^i,\cdots,k_{ri}^i$。

两个义原之间的语义距离如下[4]:

$$\mathrm{Sim}(P_1,P_2)=\frac{\alpha}{d+\alpha} \tag{2}$$

其中 P_1 和 P_2 表示两个义原;d 是 P_1 和 P_2 在义原层次体系中的路径长度,是一个正整数;α 是一个可调节的参数。

两个概念语义表达式的整体相似度记为[4]:

$$\mathrm{Sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{Sim}_j(S_1,S_2) \quad (3)$$

其中，S_1、S_2 为两个实词概念，$\beta_i(1\leqslant i\leqslant 4)$ 是可调节的参数，且有：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。$\mathrm{Sim}_1(S_1,S_2)$、$\mathrm{Sim}_2(S_1,S_2)$、$\mathrm{Sim}_3(S_1,S_2)$ 与 $\mathrm{Sim}_4(S_1,S_2)$ 分别表示第一独立义原描述式、其他独立义原描述式、关系义原描述式与关系符号义原描述式，这四个部分的相似度。

则此歧义词 W 的 n 个概念中与实词 $W_i(1\leqslant i\leqslant m)$ 的 r_i 个概念相似度最大者记为[4]：

$$\mathrm{Max}(W,W_i)=\mathrm{Max}\left(\underset{1\leqslant a\leqslant n,1\leqslant b\leqslant r_i}{\mathrm{Sim}}(k_a,k_b^i)\right) \quad (4)$$

而此歧义词 W 的 n 个概念与另外 m 个实词 $W_1,W_2,\cdots,W_m$ 的 $r_1+r_2+\cdots+r_m$ 个概念相似度最大者记为[4]：

$$\mathrm{Max}(W)=\mathrm{Max}\left(\underset{1\leqslant i\leqslant m}{W,W_i}\right)=\underset{1\leqslant i\leqslant m}{\mathrm{Max}}\left(\underset{1\leqslant a\leqslant n,1\leqslant b\leqslant r_i}{\mathrm{Sim}}(k_a,k_b^i)\right)=\underset{1\leqslant a\leqslant n,1\leqslant b\leqslant r_i,1\leqslant i\leqslant m}{\mathrm{Max}}(\mathrm{Sim}(k_a,k_b^i)) \quad (5)$$

取 $\mathrm{Max}(W)$ 所对应的歧义词的某个概念 $k_j(1\leqslant j\leqslant n)$ 作为结果输出[4]，即：

$$k_j=\arg\mathrm{Max}(W) \quad (6)$$

2.4 利用优先组合关系进行词义消歧的实例说明

在本次实验中，公式(2)与公式(3)中的几个参数的取值如下：

$$\alpha=1.6,\beta_1=0.5,\beta_2=0.2,\beta_3=0.17,\beta_4=0.13$$

(1)输入经过分词、词性标注和初始词义标注且含有歧义词的待词义标注句子：

参加/v 十五大/j 后/f，/w 李/nr 国安/nr 在/p 北京/ns 自费/d 买/v 了/u 600/m 多/m 册/q 十五大/j 精神/n 宣传/vn 材料/n 。/w

其中待词义标注词：材料

(2)从优先组合库中抽取与“材料”有优先组合关系的实词，与句中各实词进行语义匹配，得到一个与歧义词有优先组合关系的实词串，该实词串按照与“材料”的优先组合关系大小进行排列(只列出前几个优先组合关系)：

参加/v 十五大/j 后/f，/w 李/nr 国安/nr 在/p 北京/ns 自费/d (买/v)3 了/u (600/m 多/m 册/q)2 (十五大/j 精神/n)4(宣传/vn)1 材料/n 。/w

其中数字编号标识优先组合关系。

(3)将实词串中的各实词取出，对每个实词概念与歧义词概念进行相似度计算，选择相似度取最大值所对应的歧义词概念，作为歧义词的词义，完成词义消歧。给出消歧后歧义词“材料”的词义。

NO. =009408

参加/v 十五大/j 后/f，/w 李/nr 国安/nr 在/p 北京/ns 自费/d 买/v 了/u 600/m 多/m 册/q 十五大/j 精神/n 宣传/vn 材料/n 。/w

材料

DEF=information|信息

2

其中 NO. 为概念编号，“材料”为所选出的歧义词，“DEF=information|信息”为“材料”在句子中的正确概念，“2”表示这个概念在“材料”一词所有的概念中是第 2 个。

3 实验结果与分析

3.1 实验结果

选择由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的标注语料库《人民日报标注语料》作为试验语料，其标注集采用北京大学的汉语词性标注集。以 2 个高频多义词“材料”和“精神”作为测试实例，在语料库中随机抽取了含有“材料”和“精神”的歧义句子共 578 句进行测试，其中 200 句进行封闭测试，378 句进行开放测试，来对文中提出的方法进行分析和评价。

●材料的三个义项：

(1)可以直接造成成品的东西，如建筑用的砖瓦、纺织用的棉纱等；

(2)可供写作或参考的事实或文字资料；

(3)比喻适宜做某种工作的人。

●“精神”的三个义项：

(1)人的思想意识、思维活动和一般心理状态；

(2)主要意义、宗旨；

(3)表现出的活力(以上两个多义词的义项均取自《辞海》)。

系统采用准确率来评估该方法的性能，定义 CP 为准确率，Nc 表示标注正确的个数，N 表示测试语料中的总词数。准确率的计算公式如下：

$$CP=Nc/N\times 100\%$$

表 1 给出“材料”和“精神”的测试结果。

表 1 “材料”和“精神”测试的实验结果

词例	词义	各词例优先组合词示例	封闭测试准确率(%)	开放测试准确率(%)
材料	S_1	建筑～，金属～，一种～，过去的～，供应～，～买了，～充足，～用途	80.2	72.4
	S_2	收集～，人事～，一批～，书上的～，找到了～，～丰富，～来源，～生动		
	S_3	唱歌的～，军人的～，聪明的～		
精神	S_1	振奋～，革命～，高尚～，一种～，奋斗～，～焕发，～分裂，～面貌	90.1	88.6
	S_2	文件～，领会～，核心～，主要～，传达～，～明确		
	S_3	眼睛很～，振作～，一点儿～，干活的～，老人的～，～好，～的恢复		

3.2 实验结果分析

用贝叶斯网络的无指导方法[5]对相同语料及相同词例进行开放测试，并与文中提出的基于优先组合关系的有指导方法进行比较，结果见表 2。

表 2 基于优先组合关系方法和贝叶斯网络方法开放测试比较结果

词例	基于优先组合关系方法准确率(%)	贝叶斯网络方法准确率(%)
材料	72.4	70.3
精神	88.6	71.9

根据表 2 可知，文中提出的词义消歧方法取得了比较

好的消歧效果。

实验结果表明,通过判断句中各实词与歧义词的优先组合关系,一方面,只计算与歧义词有优先组合关系词的概念与歧义词概念的相似度,即有针对性地选择进行相似度计算的词语,避免将歧义词与句中关联程度低的词进行相似度计算,从而减小了计算的工作量。另一方面也是最重要的方面,就是充分利用了词义与词的分布之间的关系,使得消歧准确率有所提高,尤其对于高频多义词的词义消歧效果更是明显。分析排歧错误实例,发现造成排歧错误的主要原因是:

(1)当多义词在表示不同意义的同时,与其词关联度高的词却是相同的。如"材料"的第一个意义和第二个意义,与表示时间的词"过去的"的词关联度就很接近,在这种情况下就难以判断其真实含义。

(2)当多义词的几个意义都比较接近时,这种情况就显得尤为严重。如:"精神"的第一个意义和第三个意义就很接近。"振奋"一词都与这两个意义的词关联度很高。这里的解决方案是适当增大窗口的大小来达到解决问题的目的。

4　结束语

本方法充分利用词义与词的分布之间的关系,抽取词语之间的优先组合特性,来判断歧义词词义,是一种简单有指导的词义消歧方法。实验结果表明该方法对高频歧义词的消解是有效的,对现实的机器翻译系统具有一定的实用价值,可以作为进一步结构消歧的基础。

文中所提出的词义消歧方法避免了规则知识库构造,但该方法面临着统计数据稀疏的问题。今后的工作是在词义消歧过程中适当增加语法功能描述,将词义消歧过程与句法语义分析结合起来。

参考文献:

[1] 王　惠.基于组合特征的汉语名词词义消歧[J].Computational Linguistics and Chinese Language Processing,2002,7(2):77-88.

[2] 余晓峰,刘鹏远,赵铁军.一种基于《知网》的汉语词语词义消歧方法[A].第二届学生计算语言学研讨会论文集[C].北京:[出版者不详],2004.128-133.

[3] 董振东,董　强.知网简介(2000).http://www.keenage.com.2000.

[4] 刘　群,李素建.基于《知网》的词汇语义相似度计算[J].Computational Linguistics and Chinese Language Processing,2002,7(2):59-76.

[5] 卢志茂,刘　挺,丁江伟,等.基于依存分析和贝叶斯网络的无指导汉语词义消歧[J].高技术通讯,2004,14(2):7-11.

(上接第21页)

工具、Linux微内核(包含内存管理、进程管理及事务处理)、初始化进程。如果想使系统的功能更加完善并同时保证系统的小型化,可以在嵌入式Linux系统中添加相应的硬件驱动程序、应用程序等。当然,为进一步加强系统功能,还可添加文件系统、TCP/IP网络堆栈、磁盘等[5]。

3　嵌入式Linux内核主要支持的功能

首先,作为操作系统,以下几种功能是必不可少的:

* 处理器支持
* 内存管理
* 进程管理
* 文件系统管理

其次,还可根据应用增加以下一些功能:

* 模块支持
* 网络支持
* MISC binary library支持
* 能源管理
* TCP/IP协议支持
* 提供网络、串口等设备驱动程序支持

4　嵌入式Linux的不足之处

由于Linux操作系统本身采用的内存管理技术是虚拟内存的交换技术,因此,这样必定会大大降低一个实时系统的性能。再加上,由于Linux的代码是开放的,所以在设计一个嵌入式系统时,会通过削减一些不必要的功能来提高系统的性能,但是这样做也很有可能带来一些比较严重的bug。

5　结束语

虽然Linux存在一些不足之处,但嵌入式Linux具有占用内存空间小、启动速度快、稳定性好、支持多任务多线程的特点,并且其源码开放的特性使得Linux必将成为嵌入式系统开发的一个高效、实用的操作系统。

参考文献:

[1] 许海燕,付　炎.嵌入式系统技术与应用[M].北京:机械工业出版社,2002.

[2] 张海峰,张宏海,张士平.嵌入式Linux系统[J].微计算机信息,2004,20(1):74-75.

[3] 王学龙.嵌入式Linux系统设计与应用[M].北京:清华大学出版社,2001.

[4] 许德民.操作系统原理——Linux篇[M].北京:国防工业出版社,2004.

[5] 赵　炯.Linux内核完全注释[M].北京:机械工业出版社,2004.