

一种基于概念关联式的词义消歧方法

缪建明 张 全

(中国科学院声学研究所 北京 100190)

摘 要 词义排歧是自然语言处理中最关键也是最困难的问题之一,至今仍未得到完全有效的解决。在研究 HNC 表达汉语知识的基础上,提出了一种基于概念关联式的汉语词义消歧方法,用于处理汉语的歧义字段。该方法综合了词语概念的层次性、网络性、结构性特征,用一种统一的表示式来规范这类特征,解决了多个不同概念之间的知识关联表示问题。实验对 20 个汉语高频多义词进行了测试,平均正确率为 94%,验证了该方法的有效性。

关键词 词义消歧,概念关联式,HNC 理论

中图法分类号 TP18 **文献标识码** A

Word Sense Disambiguating Method Based on Concept Relativity

MIAO Jian-ming ZHANG Quan

(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract As one of the most important and also the most difficult problem of Chinese information processing field, word sense disambiguation has not been entirely solved until now. On the basis of our research on HNC representing Chinese knowledge, this article proposed a word sense disambiguating method based on concept relativity. The method combines various features in concepts. The features include hierarchy, network and structure etc. The uniform representation formalizes the features. In this way, the problem of knowledge relationship expression among various concepts will be solved. 20 Chinese polysemous words were tested in our experiment. The result with average precision 87% shows that the method is effective.

Keywords Word sense disambiguation (WSD), Concept relativity, HNC theory

1 引言

词义歧义现象普遍存在于各种语言中。它源于一个词语具有多个义项,而词义消歧,就是解决多义词在自然语言中的义项归属问题。由于词义分类种类繁多、无统一的分类标准、义项不易描述、义项表达方式不同,词义消歧始终是自然语言处理过程中的一个难题^[1],目前仍未能得到完全而有效的解决,使得以此为基础的多个应用领域都难以取得突破性进展。

目前词义消歧的方法主要有两个:规则消歧法和统计消歧法。近几年也有研究者提出了其它的消歧方法,但这些方法大都是上述两方法的改进或变形,本质上仍然是这两种方法^[2]。规则消歧法和统计消歧法在某种程度上提高了词义消歧的精度,但是这两种方法本身都有不可避免的缺陷。一般来说,目前词义排歧的主流方法是基于统计机器学习的方法,首先从测试语料集中提取词语的样本,统计其各种义项出现的频次等信息,从而给定这一词语不同义项的加权值,当真实语料中出现该词语的语句时,采用不同的概率模型理论对语句中的词语进行义项预测,计算得到的最大值即为该词语的义项。这种方法由于统计机器学习本身的快速、便捷的优点,

在词义排歧上被广为采用。但是统计机器学习的方法本身存在一定的缺陷,它把自然语言的运用过程简单模拟为一个随机过程,并采用随机过程的处理方法来处理自然语言。但自然语言运用过程不同于随机过程,彼此词语间具有一定的概念关联性,且不能满足概率模型理论的前提条件,若满足了大概率事件,则小概率就得不到满足,由此造成原本能够简单使用概念之间关联性来排除的义项反而不能通过这种方法排除^[3]。规则消歧法在构造规则库时既要考虑规则的完备性又要考虑其合理性,需要根据词义义项标记的不同对规则进行设计,虽然使用一条规则可提高准确率,但是人工难度较大,需要人类专家具有较高的语言学功底,由此必然造成这种方法难以进行大规模的批量化设计及使用。

因而大多数研究者强调将统计机器学习的方法与人工规则的方法相结合,充分吸取两者的长处,弥补各自的不足,从而达到理想的处理效果,但两者如何结合、怎么结合的处理仍各有不同的方法。

本文利用的概念关联式的词义排歧方法,正是试图结合人工规则方法与统计机器学习方法的优点,来解决语句中的多义词语的歧义困扰。这种方法是建立在概念层次网络(Hi-

到稿日期:2009-02-20 返修日期:2009-04-21 本文受国家 973 项目“自然语言理解的交互引擎研究”(2004CB318104),国家科技支撑计划课题“搜索引擎中的语言翻译基础研究”(2007BAH05B02-05),中科院声学所知识创新工程项目“句群理解处理理论及其应用”(O654091431),中国科学院声学研究所“所长择优基金”(GS13SJ04),中国科学院青年人才领域前沿项目(O754021432)资助。

缪建明(1977—),男,博士,主要研究方向为自然语言理解处理,E-mail:mjrm@mail.ioa.ac.cn;张全 研究员,博士生导师。

erarchical Network of Concepts,简称 HNC)理论的概念基元体系基础研究之上的。这个概念基元体系是概念基元的一个分类体系,概念基元是整个 HNC 理论概念知识表示体系的基础,它用数字串的方式对自然语言概念基元进行表述,用五元组特性对概念的多元性表现进行了描述,两者的有效结合使有限的概念基元可表达丰富的词语内涵,从而可将自然语言的词语一一映射到相对应的概念上。概念基元之间具有层次性和网络性的特点,两两相关的概念,其内涵相同或近似,相应的层次符号也相同或近似。这种表达方式显然有利于计算机对概念关联性的把握,从而有利于对自然语言的理解。不同的概念基元之间具有一定的概念关联性,包括特殊的概念同行性和普通的概念关联性两种。本文正是在两两概念之间的关联性基础上进行词义排歧研究。

2 概念关联式

概念关联式的引入是在两两概念形式化表示的基础上展开的。HNC 理论对概念的一般表示方法为:Σ{类别符号串}{层次符号串}{组合结构符号}{类别符号串}{层次符号串}。其中的“类别符号串”也叫“字母串”,用类别符号集的字母表示。五元组符号一定在其他类别符号的后面。“层次符号串”也叫“数字串”。字母串代表概念的类别特征,数字串代表概念的层次性内涵,组合结构符号代表复合概念的组合结构[4]。

任何一个概念都需要从不同侧面予以表达,我们将这一现象叫做概念的多元性表现。概念分为抽象概念和具体概念。具体概念的多元性十分复杂,比较难以给出规范化的表述。抽象概念则有所不同,其多元性表现在自然语言中有明显的迹象,这就是词性现象。英语里的词根或带有词根特色的词,可配以不同的后缀分别构成动词、名词、形容词和副词。汉语的这种词性模糊现象和西语的以形态变化表现不同词性的现象都是抽象概念多元性的生动表现。兼词性只是一种表现形式,本质在于这些概念本身具有这一变化的固有特征。实际上这是所有抽象概念的共同特征,或者说是抽象概念语言表现的最大共性,它体现了实际语言概念之间的一类关联性,称之为同行关联或同行优先,因为映射到层次网络符号,所以这些概念属于概念矩阵的同一行。抽象概念需要从动态、静态、属性、值和效应 5 个侧面加以表述,这就是抽象概念的五元组特性,简记为(v,g,u,z,r)特性,它们是抽象概念外在表现的基元。这里举一些不同五元组符号的概念实例。

专业活动 gva0 财富 rwa0
经济活动 gva2 财产 rwa2

其中,财富是人类一般专业活动的效应物,而财产是经济活动的效应物。上列两两相关的概念,其内涵相同或近似,相应的层次符号也相同或近似。这种表达方式显然有利于计算机对概念关联性的把握,从而有利于对自然语言的理解。当然,这就需要自然语言词汇映射成层次网络符号。抽象概念表达的 5 个基本侧面是自然语言概念关联性的基本表现之一。将它和概念的层次性、对比性、对偶性、包含性统称为概念同行关联。表达这些概念的具体自然语言符号可以变化万千,但对内涵或其内核相同或相似的概念,我们将赋予相同或相似的层次符号,即相同或相似的数字串。从概念矩阵来看,它们都属于概念矩阵的同一行。这样,一部分语义距离的计算问题就简化为对数字串的逐层比较问题。自然语言的词,

大多数是多义的,人在言语感知过程中必须进行大量的“多义选一”处理。也许可以说,这是交谈或阅读过程中大脑里最基础、最频繁的操作。概念矩阵的表示方式就为计算机进行类似的“多义选一”操作提供一个简明有效的途径和工具。例如:词库中“名将”一词具有两个义项,分别可以由 HNC 概念符号 u331/pa4 和 u331/pa339 来表示,在具体的语句中需要确定其对应的语义项。义项 1 表示军事领域声名显赫的人,义项 2 表示体育竞技领域声名显赫的人,两个义项分处不同的领域范畴之内,静态来看,很难确定两个义项的可能概率,而在具体句子中,依据概念关联性则很容易激活其中的一个义项。例如:“记者采访我省自行车名将张俊英还是颇费了几番周折。”句中,通过“自行车”一词的具体概念“a339i\13ckm”,我们能发现其与“名将”的义项 II 具有相似性,都有一个相同的部分为“a339”,这种相似性称之为概念同行性,表明两个概念属于同一个概念树(a3)之下,“自行车”这一概念属于“体育(a339)”这一概念的某一个具体子类,这两个概念天然具有关联性,这只是其中的一类同行概念关联性[3]。

概念之间的交互式关联知识则通过给定的交互式概念关联式来激活。两个概念节点交互式关联就是指“两个概念有交叉”,深入地说,交互式关联是同一个概念本体从不同观察角度看到的不同映像。这里所说的“不同观察角度”就是指作用效应链的 6 个基本环节[2]。例如,“死亡”这一“概念”,从过程来看,它是代谢的“谢”142;从效应来看,它是“消失”312;从状态来看,它是“减少”500412,因此,过程节点 142,效应节点 312 和状态节点 500412 是交互式关联的。同样 141,311,500411 也是交互式关联的,其上层的概念节点 14,31,50041,也就必然具有交互式关联性。这类关联性不可能从现有的熟语词库直接得到,只能依靠人工方式先建立一个框架。例如:概念关联表示式“q731\k=a339”就表示概念“比赛类型(q731\k)”,与不同概念树下的概念“体育(a339)”强交互式关联。

这两种概念关联性所带来的概念关联式构成了词义排歧的主要依据。

3 消歧算法及实现

在任意一个自然语言文本中,不同词语之间的概念相关性是不同的,从语义角度将它们称为语义相关度。词语的语义相关度是与概念的关联性有关的,一个词可以有多个义项,每个义项可以分解为不同的概念基元,这种概念基元就是用于描述一个义项的最小基本语义单位。在 HNC 概念基元符号体系中,共有 7661 个概念基元,然后用这些概念基元和组合符号一起对 HNC 词语知识库中的 40028 个汉语词语进行了语义描述。图 1 给出了一些 HNC 词语知识库的实例。

词形	音调	义项义项号/义况	组合代码	概念类别	hnc符号
走漏	zou3,lot3	3 3		v	v3ae02+u012
深入	shen1,ru3	3 1	1	u	(u8109,u811,u92
深入	shen1,ru3	2 3	1	v	v22b+(v42,v401
深入	shen1,ru3	1 1	1	v	(v8109,v201)
难住	nan2,gu3	1 0	1	uv	uv1219+j76e71
难住	nan2,gu3	2 0	1	uu	!j1uul3c21 v711
难住	nan2,gu3	3 0	1	gu	f53
怀抱	huai2,bz3	2 3	5	++g	gb00
难过	nan2,gu3	2 3	1	u	u50ae72
提出	ti2,chu3	2 0	1	vv	v902
意味	yi4,wei43	1 1	1	r	j50+jgwa30;r923

图 1 词语知识描述的部分实例

具体的消歧算法流程如下。

3.1 初始处理

输入一篇文章作为输入串,将文章切分为不同的语句串,

每一个语句串作为一个处理单元进行初始化处理。将语句串利用分词程序进行分解,形成对应的词语序列串,从词语概念库提取每一个词语对应的概念符号串,把存在多个义项的词语和包含有关键性信息的词语(包含领域概念基元符号的词语)挑选出来,进入下一步匹配计算阶段。

3.2 匹配计算

将多义项词语的概念符号串分解为多个单一义项的概念基元符号串,这些概念基元符号串与包含信息概念基元串进行逐层匹配计算(利用同行概念关联式进行计算),或者通过领域范围内给出的交式概念关联式来进行匹配计算。如果中间存在匹配的义项概念串,则形成义项概念串的匹配序列,进入下一步消歧处理阶段;如果这些义项概念符号串中不存在匹配的义项概念符号串,则进入统计算法计算模式,通过统计数据给出该义项概念在该词语中所出现的概率值,通过统计算法给出它的置信权值。由于这一部分不是本文研究的重点,作者在此不过多阐述。当然这一步骤也可以相对地调整,可以先采用统计算法计算出对应的结果,只有在置信度低于一定的数值时才采用规则的方法进行义项的匹配计算。

3.3 消歧处理

提取对应的义项概念串进行排列,同时进行必要的全语句串的一个验证过程,使得整个语句的概念符号串符合 HNC 的句类结构,从中得到最后的“句类—概念—词语”结果,实现语句级别的语义深层结构。至此,输出对应的词语义项的概念符号串结果,整个消歧算法过程结束。

4 实验及其结果

表 1 新闻语料测试结果

多义词	义项数	统计准确率	本文准确率	提高百分比
底子	6	83.25	92.64	9.39
拉扯	6	84.67	96.51	11.84
左右	5	81.62	91.05	9.43
门户	5	84.48	97.16	12.68
扑腾	5	82.94	92.38	9.44
点子	5	80.69	92.59	11.9
后身	5	78.26	85.57	7.31
昏天黑地	5	84.78	98.37	13.59
当头	4	81.68	92.18	10.5
机关	4	83.72	93.62	9.9
开放	4	82.95	96.48	13.53
出世	4	81.63	93.85	12.22
关闭	4	84.65	94.38	9.73
调理	4	82.64	93.67	11.03
分散	4	83.45	98.37	14.92
站住	4	84.16	92.28	8.12
狠心	4	82.39	93.42	11.03
教育	4	82.57	96.29	13.72
到底	4	80.27	93.67	13.4
沐浴	4	82.71	97.64	14.93

实验选取了 20 个汉语文本中经常出现的多义词作为测试对象,从新闻语料中随机找了 1011 个包含上述多义词的语句进行测试,采用上面所述的基于概念关联式的词义排歧方法在句子中进行消歧,验证了本文提出的方法的有效性。实验测试的 20 个多义词如表 1 所列。

实验获得了较好的结构及平均将近 94.11% 的准确率,比单纯用统计方法进行词义消歧的结果(82.68%)平均提高了 11.43 个百分点,提升效果还是令人满意的。

我们发现,如果在进行词语义项激活之前,准确给出语句所属文章对应的领域符号串,则消歧准确率在此基础上更能得到进一步的提高,这说明了篇章或语句的领域信息对词语义项排歧的重要作用,也为下一步准确率的提高指明了方向。

结束语 综上所述,词语义项的多义排歧,是对自然语言理解处理所面临的一个研究难点也是重点之一,它的准确率在很大程度上影响着其它领域诸如机器翻译、中文信息检索等的研究成果。本文利用 HNC 理论概念基元符号体系中概念的关联式,从多义词所蕴含的义项的概念符合串入手,结合统计方法和规则方法各自的优势,证实了这种方法为多义词的义项排歧探索了一条行之有效的途径。

下一步的工作将从丰富和完善概念的交式关联式方面入手,提高利用词语概念对语句领域判断的准确率,把它形式化后添加到初始化处理中,使得在做义项匹配计算之前能够得到对应的领域指导,以进一步提高消歧的正确率。

参 考 文 献

[1] 陈笑蓉,秦进. 基于最大熵原理的汉语词义消歧[J]. 计算机科学,2005,32(5):174-176

[2] 王广正,王喜凤. 基于知网语义相关度计算的词义消歧方法[J]. 安徽工业大学学报,2008,25(1):71-75

[3] 缪建明,张全. 词语义项的概念激活方法[C]//第九届汉语词汇语义学研讨会论文集. 新加坡: COLIPS PUBLICATAION, 2008:191-197

[4] 黄曾阳. HNC(概念层次网络)理论[M]. 北京:清华大学出版社,1998

(上接第 90 页)

[2] Biba K J. Integrity considerations for secure computer systems [R]. MA, USA: Air Force Systems Command, 1977

[3] Clark D D, Wilson D R. A Comparison of Commercial and Military Computer Security Policies[C]//Proceedings of the IEEE Symposium on Security and Privacy. CA, USA: IEEE Computer

Society, 1987: 184-194

[4] 季庆光. 高安全级操作系统形式设计的研究[D]. 北京:中国科学院研究生院, 2004

[5] Bell D E. Modeling the multipolicy machine[C]//Proceedings of the 1994 Workshop on New Security Paradigms. CA, USA: IEEE Computer Society, 1994: 2-9