ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.8, August 2009, pp.2138−2152
doi: 10.3724/SP.J.1001.2009.03566

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

无监督词义消歧研究*

王瑞琴[1,2+], 孔繁胜[1]

[1](浙江大学 人工智能研究所,浙江 杭州 310027)

[2](温州大学物理与电子信息工程学院,浙江 温州 325035)

Research on Unsupervised Word Sense Disambiguation

WANG Rui-Qin[1,2+], KONG Fan-Sheng[1]

[1](Artificial Intelligence Institute, Zhejiang University, Hangzhou 310027, China)

[2](College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China)

+ Corresponding author: E-mail: angelwrq@163.com

Wang RQ, Kong FS. Research on unsupervised word sense disambiguation. *Journal of Software*, 2009,20(8): 2138−2152. http://www.jos.org.cn/1000-9825/3566.htm

Abstract: The goal of this paper is to give a brief summary of the current unsupervised word sense disambiguation techniques in order to facilitate future research. First of all, the significance of unsupervised word sense disambiguation study is introduced. Then, key techniques of various unsupervised word sense disambiguation studies at home and abroad are reviewed, including data sources, disambiguation methods, evaluation system and the achieved performance. Finally, 14 novel unsupervised word sense disambiguation methods are summarized, and the existing research and possible direction for the development of unsupervised word sense disambiguation study are pointed out.

Key words: word sense disambiguation; unsupervised word sense disambiguation; natural language processing; semantic understanding

摘 要: 研究的目的是对现有的无监督词义消歧技术进行总结,以期为进一步的研究指明方向.首先,介绍了无监督词义消歧研究的意义.然后,重点总结分析了国内外各类无监督词义消歧研究中的各项关键技术,包括使用的数据源、采用的消歧方法、评价体系以及达到的消歧效果等方面.最后,对 14 个较有特色的无监督词义消歧方法进行了总结,并指出无监督词义消歧的现有研究成果和可能的发展方向.

关键词: 词义消歧;无监督词义消歧;自然语言处理;语义理解

中图法分类号: TP391 **文献标识码**: A

* Supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No.Y1080372 (浙江省自然科学基金)
Received 2008-07-06; Accepted 2009-01-14

1 词义消歧(word sense disambiguation)的基础知识及研究意义

1.1 词义消歧的定义

词汇的歧义性是自然语言的固有特征.词义消歧根据一个多义词在文本中出现的上下文环境来确定其词义,作为各项自然语言处理的基础步骤和必经阶段被提出来.所谓的词义消歧是指根据一个多义词在文本中出现的上下文环境来确定其词义.形式化地,令词语 w 具有 n 个词义,w 在特定的上下文环境 C 里只有 S'是正确的词义,词义消歧的任务就是在这 n 个词义中确定词义 S'.每个词义 S_K 和上下文 C 都存在或强或弱的联系,记为 $R(S_K,C)$,其中 S'与上下文 C 的关系应当是最强的.词义消歧技术通过分析和计算 W 出现的上下文 C 和每个词义 S_K 之间的关系 R,排除干扰词义,最后确定 S'.整个过程可用下面的公式来描述:

$$S' = \arg\max R(S_K, C) \tag{1}$$

上下文中的某些词语限定了多义词的词义,正是这些词的存在,可以帮助人们迅速地去推理和判断,最终得到答案.自动词义消岐研究的是机器模拟人类思维的过程,在上下文中收集重要的语义信息,提取特征词语来指导对多义词的歧义消解.词义消歧问题曾一度被认为是一个计算机无法攻克的难题[1],致使从那以后的一段时间里,研究人员逐渐放弃了对词义消歧的研究.但随着计算技术的飞速发展,超大容量的存储设备和具有强大计算能力的多核处理器相继出现,包括词义消歧在内的自然语言处理领域的各种问题研究一一复苏,并进入了崭新的发展阶段,词义消歧逐渐成为计算语言学和自然语言处理领域中的一个重要研究课题,也是近些年来该领域的热点研究问题之一.

1.2 词义消歧的分类

每个分类问题都会根据分类依据的不同而得到不同的分类结果,词义消歧也不例外.根据消歧知识来源的不同,词义消歧方法可分为基于知识的方法和基于统计的方法,基于知识的消歧一般又细分为基于规则的方法和基于词典的方法.基于知识库的消歧方法主要是依赖语言学专家的语言知识构造知识库,通过分析多义词所在的上下文,选择满足一定规则的义项.知识库的类型包括专家规则库、词典、本体、知识库等.基于统计的方法则以大型语料库为知识源,从标注或未标注词义的语料中学习各种不同的消歧特征,进而用于词义消歧.

按照消歧过程有无指导,词义消歧分为有导消歧和无导消歧.前者利用已标注了词义的大型语料库来提取特定词义的特征属性,利用机器学习方法生成分类器或分类规则对新实例进行词义判定;后者则从原始的数据文集或机器可读字典中获取词义的相关特征,对新实例进行词义判定.所以,有指导的词义消歧常被看作词义分类问题,无指导词义消歧被看作聚类问题.

按照消歧结果的评价体系,词义消歧分为独立型评估和应用型评估.独立型评估是指不依赖于应用领域,使用一组标准的测试集,独立评价词义消歧性能.应用型评估不单独地评价词义消歧的效果,而是考察其对实际自然语言处理系统最终目标的贡献,比如,词义消歧在机器翻译系统中对翻译性能的影响、在信息检索中对搜索性能的改善情况等等.

1.3 词义消歧研究的意义

词义消歧是对词的处理,属于自然语言理解的底层研究,在许多高层次的研究和应用上,词义消歧都大有用武之地.词义消歧并不是自然语言处理的最终目的,而是自然语言处理中不可缺少的一个环节,歧义问题的解决将会带动至少下列自然语言处理领域的新进展:

- 机器翻译:在机器翻译中,要让计算机进行准确的译文选择,一个重要的前提条件就是能够在某个特定上下文中自动排除歧义,确定多义词的词义.所以,词义消歧从 50 年代初期开始机器翻译研究起就一直备受计算语言学家的关注.
- 信息检索:一个拼写正确的词汇通常包含许多词义,在特定的查询上下文中,很多词义是不相关的.在一个特定的查询中,用户只对其中一个词义感兴趣,因此只需检索和那个词义相关的文档,而当前基于关键字的搜索引擎就面临检索包含相关词义文档而过滤掉无关词义文档的大难题.据统计,在信息检索中引

入部分多义词消歧技术以后,可使其整个系统的正确率由 29%提高到 34.2%,取得较为明显的改善.

- 主题内容分析和文本处理:如文本分类、信息抽取、自动文摘和辅助写作等文本处理任务,只有对文本中的多义词进行消歧,明确单词所表示的概念,才能正确分析文本及句子的概念和主题.
- 语音处理和文语转换:这类任务往往同时涉及语音和文字的处理,语音识别中同音字的识别、语音合成中语音的校正以及文字的处理都离不开词义消歧.
- 语法分析或句法分析:帮助解决语法的歧义问题,降低语法分析难度,改善语法分析效果.

总之,词义消歧是计算语言学和自然语言处理领域的基础研究课题,提高词义消歧的研究水平,提供高质量的词义消歧技术,对包括机器翻译、信息检索、文本分类等在内的众多研究领域都会有重要的推动作用.

2 无监督词义消歧方法概述

无监督词义消歧按照消歧数据源的不同分为基于知识的方法和基于统计的方法两大类.本节将分门别类地讨论当前国内外各类主流的无监督词义消歧方法,从消歧过程中使用的数据源、采用的消歧技术、评估体系和消歧效果 4 个方面进行阐述,研究各类消歧方法使用的关键技术及其消歧性能,指出各自的优缺点及改进方案,特别地,对那些具有代表性的消歧算法将进行详细论述.

2.1 基于知识的无监督词义消岐

基于知识的无导词义消岐进一步被划分为基于规则的方法和基于词典的方法.早期人们所使用的词义消歧知识一般是凭人手工编制的规则,由于手工编写规则费时、费力,存在严重的知识获取的瓶颈问题,20 世纪 80 年代以后,语言学家提供的各类词典成为人们获取词义消歧知识的一个重要知识源.

2.1.1 基于机读词典的词义消歧

机读词典提供了有关词汇用法及词义描述的丰富知识,是早期词义消歧的主要知识来源.最早利用机器可读字典实现无监督词义消歧的研究始于 1986 年的 Lesk 方法[2].Lesk 利用词典中词义的解释或定义来指导多义词在上下文中的词义判定.该方法简单易行,只需计算多义词的各个词义在词典中的定义与多义词上下文词语的定义之间的词汇重叠度,选择重叠度最大的词义作为其正确的词义即可.Lesk 分别用 3 个机器可读词典(Webster's 7th Collegiate,Collins English Dictionary 和 Oxford Advanced Learner's Dictionary of Current English)对一组多义词实例进行了词义消歧测试,正确率在 50%~70%之间.随着 Lesk 方法的提出,无监督词义消歧逐渐流行起来.研究者对 Lesk 方法进行了各种改良,总体思想是进一步扩展词义的定义描述,使得词汇重叠的几率增加.Wilks[3]对 Longman 字典(Longman Dictionary of Contemporary English,简称 LDOCE)中每个词义的定义添加了与其定义词汇同现频率较高的其他词汇(同现频率的高低使用该词典的所有定义条目统计得到),如此将词典中的所有定义进行了扩展之后,大大提高了定义词汇重叠的概率.Pook 等人[4]提出一种改进方案,对上下文词语进行同义词扩展,从而扩大了上下文窗口的大小.实验结果表明此方法可以增加词义消歧的覆盖率.Dagan[5]和 Gale[6]则利用双语对照词典来帮助多义词消歧.

2.1.2 基于义类词典的词义消歧

义类词典的编排与传统词典有很多不同之处.它是按照词语含义编纂的辞典,把相类似的词语放在相同的目录下,使得查找同类或同义词更加方便、快捷.义类词典有助于我们提高用词的准确性.Roget's Thesaurus[7]和 WordNet[8]是常用的英语义类词典.Yarowsky(1994)[9]和卢志茂等人[10]利用 Roget's 词典进行词义消歧;Voorhees[11]和Resnik[12]从不同角度利用WordNet中的上下位关系、同义关系进行英语的词义消歧探索.《同义词词林》[13]和知网(HowNet)[14]是最常用的汉语义类资源.汉语词义消歧研究从 20 世纪 90 年代以后才开始.陈浩等人[15,16]使用HowNet作为知识源,利用聚类技术进行词义消歧.李涓子[17]和中国科学院计算技术研究所的鲁松[18]都采用《同义词词林》进行无指导的词义消歧,李涓子在大规模语料库中自动获取任意同义词集中单义词的同现实词,按照同现实词的词义分辨能力对它们加权,构成词义分类器,实现一种代价最小的无指导学习算法;鲁松则把待消歧的多义词的上下文视为查询,把与该多义词某个义项具有相同、相似或相关语义范畴的词语的上下文视为文档,从而用信息检索中的向量空间模型来解决词义消歧问题.

2.1.2.1 WordNet简介

WordNet 是从 1995 年开始,在普林斯顿(Princeton)大学认知科学实验室(Cognitive Science Laboratory)的心理学教授 George A. Miller 的指导下,由 Princeton 大学的心理学家、语言学家和计算机工程师联合设计的一种基于认知语言学的在线英语词典.由于它包含了语义信息,所以有别于通常意义下的字典.WordNet 根据词条的意义将它们分组,那些具有相同意义的词条称为同义词集(synset),每个 synset 代表一个潜在的概念(concept),一个多义词将出现在与其各词义对应的多个同义词集中.它不仅把单词以字母顺序排列,而且按照单词的意义组成一个“单词的网络”.

WordNet 的开发有两个目的:其一,它既是一个词典,又是一个辞典,从直觉上讲,它比单纯的词典或辞典更加有用;其二,支持自动的文本分析以及人工智能应用.WordNet 是完全免费的资源,其数据库及相应的软件工具的开发都遵从 BSD 许可协议,可以自由地下载和使用,亦可在线查询和使用.WordNet 已经在英语语言处理的研究中得到了广泛应用,几乎成了英语语言知识库的标准.

2.1.2.2 基于WordNet的词义消歧

WordNet 中包含了丰富的语义知识,包括词义的定义描述、使用实例、结构化的语义关系、词频信息等,所有这些信息都可以用于词义消岐.

(1) 基于定义描述的词义消岐

WordNet 为其中的每个同义词集都提供了简短、概要的定义描述和使用实例,如 bus#1: autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle--(a vehicle carrying many passengers; used for public transport; “he always rode the bus to work”).Satanjeev[19]使用 WordNet 来代替传统的机读词典,对原始 Lesk 算法进行改进,提出了 Adapted Lesk 算法.该算法使用 WordNet 中的多种语义关系来扩展词义的定义描述.在计算两个同义词集的相关度时,不同于原始 Lesk 算法的简单计数,Satanjeev 对多字短语分配了较高的权重(短语字数的平方),以突出其在相关性判断中的重要性.Satanjeev 使用了一种全局消歧策略,即消歧时不是独立确定每个词汇的词义,而是以整个句子作为处理单位,对每种词义的不同组合,计算整体相关性,取相关性最高的组合中的各个词义作为相应词汇的词义.使用 WordNet 1.7 和 Senseval-2 lexical sample 数据集的消歧测试结果为 noun:32.2%,verb:24.9%,adjective:46.9%,平均消歧精度为 32.3%.Chen 和 Yin 提出了 AALesk 算法[20],是对 Adapted Lesk 算法的进一步优化.在消歧的过程中,AALesk 算法考虑了 WordNet 中定义的全部语义关系,并根据各自的重要性分别为每种关系分配权重,从而使消歧过程不仅依赖于统计理论,而且以一种语义指导的方式进行.对于一个词汇的多个词义,Chen 等人采用平行执行多个 AALesk 算法来消除那些无关的词义组合,进而加快消歧速度.使用 WordNet 1.7 和 Senseval-2 lexical sample 数据集的测试结果为 noun:32.6%,verb:25.1%,adjective:47.2%,平均消歧精度为 32.6%;使用 WordNet 2.0 和 Senseval-2 lexical sample 数据集的测试结果为 noun:33.3%,verb:26.2%,adjective:47.5%,平均消歧精度为 33.4%.Ledo-Mezquita[21]提出合并 Lesk 方法和大型词汇资源库(如同义词词典、WordNet 本体)进行词义消歧的方法.对一个多义词汇,首先用 Lesk 方法计算其每个词义的值;然后通过同义词词典或 WordNet 本体找到每个词义的同义词、上义词、下义词等相关词,再利用 Lesk 方法计算这些词义的值;最后将这些词义值与其各自的权重相乘,再与源词义值加权求和,得到最终的词义值,词义值最大者为岐义词的确切词义.使用共包含 4 287 词义的 872 个词语的测试集得到 63%的消歧精度,而使用相同的数据集,原始 Lesk 算法的消歧精度仅为 50%.

(2) 基于概念区域密度的词义消岐

基于概念区域密度的词义消岐的基本思想是:将本体层次结构根据待消歧词的各个词义划分成互相独立的子结构,每个子结构以歧义词的每个词义作为根节点,分别计算各子结构的概念密度来判定目标词的词义.Agirre[22]最先提出使用概念密度进行词义判断,充分利用概念间的概念距离生成定义良好的概念密度公式,基于 WordNet 中覆盖广泛的名词层次结构对名词进行词义判断.该方法是一个完全自动化的无监督词义消歧方法,给定处于子结构根节点处的词义概念 c,$nhyp$ 和 h 分别表示子结构中节点包含的下义概念节点的平均数和子结构的高度,当此子结构中包含 m 个目标词与上下文词汇的词义时,c 的概率密度为

$$CD(c,m)=\sum_{i=0}^{m-1}nhyp^i \Big/ \sum_{i=0}^{h-1}nhyp^i \tag{2}$$

公式(2)中的分子表示包含 m 个词义标记的子结构的估计区域大小,分母表示子结构区域的真实大小.利用 SemCor 数据集进行测试得到了 76.04%的消歧精度和 23.21%的召回率.Rosso[23]对原始的概念密度方法进行扩展,提出了一种计算概念密度的新方法,对于公式(2)中考虑的每个节点的平均子节点个数,WN1.6 与 WN1.4 中的不一致,所以 Rosso 决定只考虑子结构中由目标词和上下文词汇的词义路径决定的相关分支,而忽略那些无关的分支,提出使用子结构中包含的相关词义个数与子结构中的词义总数的比值计算概率密度的基础公式;Rosso 发现如果不考虑词义的词频信息,概率密度法有可能会选择低频词义,在大多数情况下这是错误的,这是由多义词词义使用的严重偏斜决定的,所以他将词频信息添加到基础公式中对消歧进行调节;另外,为了提高包含相关词义较多的子结构的权重,还增加了词汇重叠调节因子;最后,考虑到位于越低层次的子结构中的词义粒度就越细、词义间的相似性越大,将概率密度函数作为子结构深度的增函数,称这种情况为聚类深度关联(CDC).调整后的概率密度公式为

$$CD(M,nh,f,depth)=M^{\alpha}\times(M/nh)^{\log f}\times(depth(cl)-avgdepth+1)^{\beta} \tag{3}$$

其中 M 表示子结构中包含的相关词义个数,nh 表示子结构包含的词义总数,f 是子结构对应的词义在 WordNet 中的词频,cl 为子结构的根节点,$depth(cl)$为子结构的深度,$avgdepth$ 为平均子结构深度,α和β为调节因子.利用 SemCor 数据集,不采用任何相关调节技术,当上下文窗口大小为 2 时,消歧效果最佳,得到 81.48%的精度、60.17%的召回率和 73.18%的覆盖率;使用了 CDC 技术后,效果未见明显改善.将上下文窗口增大到 4 时,召回率和覆盖度有所提高,分别为 61.27%和 77.87%,当上下文窗口增大到 6 时,召回率仍保持在 60%左右,精度有些微降低,但覆盖率得到的明显的增加.

Davide[24]在 Rosso 的基础上,对概念密度方法作了进一步的改进.Davide 考虑到词汇的领域信息对消歧的影响,对消歧公式作了进一步修改,在概念密度公式中添加了互领域权重(mutual domain weight,简称 MDW)调和项.在子结构中,如果某词汇与目标词汇具有相同的领域属性,则它们的互领域权重为二者领域权重(权重与频率成正比)的乘积,考虑到 WordNet Domains 中的 Factotum 领域过于一般化,当词汇对的领域属性均为 Factotum 时,其互领域权重要降低一个数量级($\times 10^{-1}$):

$$CD(M,nh,w,f,C)=M^{\alpha}\times(M/nh)^{\log f}+\sum_{i=0}^{|C|}\sum_{j=1}^{k}MDW(w_f,c_{ij}) \tag{4}$$

$$MDW(w_f,c_{ij})=\begin{cases}0, & \text{if } Dom(w_f)\neq Dom(c_{ij})\\ 1/f\times 1/j, & \text{if } Dom(w_f)=Dom(c_{ij})\wedge Dom(w_f)\neq \text{“}Factotum\text{”}\\ 10^{-1}\times(1/f\times 1/j), & \text{if } Dom(w_f)=Dom(c_{ij})\wedge Dom(w_f)=\text{“}Factotum\text{”}\end{cases} \tag{5}$$

其中,C 表示上下文词汇向量,k 为上下文词汇 c_i 的词义个数,c_{ij} 表示词汇 c_i 的第 j 个词义.利用 SemCor 数据集进行测试得到了 78.33%的消歧精度和 62.60%的召回率,如果在计算领域相关性时不考虑杂类,则得到的精度和召回率分别为 80.70%和 59.08%,由此可见,尽管 Factotum 领域没有对整体的消歧任务提供有用的信息,但是由于采用了与词义频率成比例的权重分配方案,Factotum 有助于对大量使用常用词义的名词进行消歧.鉴于早期 Lesk 利用定义描述文本进行词义消歧的思路,Davide 也考虑用词汇在 WordNet 中的定义描述来进一步改善消歧精度,在概念密度公式中添加了定义描述权重(gloss weight,简称 GM)调和项:

$$CD(M,nh,w,f,C)=M^{\alpha}\times(M/nh)^{\log f}+\sum_{i=0}^{|C|}\sum_{i=1}^{k}GW(w_f,c_{ij}) \tag{6}$$

$$GW(w_f,c_i)=\begin{cases}0, & \text{if } c_i\notin Gl(w_f)\\ 0.3, & \text{if } c_i\in Gl(w_f)\end{cases} \tag{7}$$

其中,c_i 表示第 i 个上下文,w_f 表示待消歧词的第 f 个词义,$Gl(w_f)$函数用于返回 w_f 的定义描述中的非停用词.由于 WordNet 中关于词义的描述包括两部分:定义描述和实例描述,相应地定义了两个定义描述权重函数 $Gl_d(x)$和 $Gl_s(x)$分别返回各部分中的非停用词.利用 SemCor 数据集进行的测试表明,利用定义实例进行消歧的精度

(80.12%)比利用定义描述本身(79.85%)或综合定义和实例的消歧精度(79.31%)更高,所以他们决定使用另外的机器可读词典来扩展 WordNet 的定义实例部分,文献[24]中采用了 Cambridge Advanced Learner's Dictionary (CALD).利用 Senseval-3 All-Words Task 数据集进行测试,该方法达到了业界最好的消歧效果(79.78%),然而利用 Senseval-3 AWT 的测试结果却降低了大约 10 个百分点(64.72%),原因在于,新加入定义描述中的词汇存在偏题现象,提高外来词典与 WordNet 定义匹配时的阈值可以解决这个问题.

(3) 基于结构化语义关系的图论式词义消岐

WordNet 中各同义词集之间通过各种语义关系产生互连.这些语义关系包括:上下位关系(hypernym/hyponym)、组成成分关系(meronym/part-of)、相似关系(synonym)、反义关系(antonym)等.其中,上下位关系是最常用的语义关系,它将 WordNet 中的同义词集组织成树状的概念层次体系结构.图论(graph theory)是数学的一个分支,以图为研究对象,用点代表事物,用连接两点的线表示相应的两个事物间具有某种关系,当人们的研究对象在结构上具有内在的、结构化的联系时,常常可以采用图论的方法解决问题.

Mihalcea[25]提出一种在描述语义依赖性的图中使用随机行走策略进行词汇序列消歧的方法.首先将待消歧的词汇序列通过一个语义连接图表示出来,图中的顶点为词汇的语义标签,边为语义节点之间的语义依赖关系;然后在此图中运行一个随机行走算法,使用 PageRank 算法迭代计算每个节点的重要性值;最后算法将收敛到一组节点标签的静态概率值,这些值用于为词汇序列中的每个词汇确定其最可能的词义.该方法仅使用词典定义,达到了 54.2%的消歧精度,远远超过了在这之前的同类方法[26−30].其主要原因在于,在词汇序列标注的过程中考虑了序列词义之间的整体依赖性.结构化语义互连(structural semantic interconnection,简称 SSI)[31]是目前效果最佳的半监督词义消歧技术之一,它属于基于知识的结构化模式识别问题.首先,为消歧词和上下文词汇的每个词义建立一个结构化的图形描述,词义描述的信息来自于多个数据源(WordNet 2.0,Domain labels,WordNet glosses,WordNet usage examples,Dictionaries of collocations),各数据源之间通过人工或自动化的方法进行集成;然后,用一套语法规则来生成各种有意义的关联模式,并为每个关联模式分配权重;最后根据歧义词的每个词义及其上下文信息与这些规则的匹配情况进行消歧.需要指出的是,原 SSI 方法中的数据源包括了标注文集(SemCor,LDC-DSO),去掉这一数据源之后,SSI 才可称得上是一项无监督的词义消歧技术.SSI 算法的执行是一个迭代的过程,初始化输入是一系列同现词汇 T、未决词汇集 P 和相关词义 I(I 中包含单义词的词义或固定词义,在没有单义词或固定词义的情况下,对歧义度最小的词汇的词义作初始假设,然后这一过程被克隆执行 m 次);在迭代阶段,将 P 中的某元素 t 确定词义并将其移到 I 中,要求 t 的至少 1 个词义与 I 中的词义存在语义关联,当 P 为空或某次迭代 P 中元素的个数未减时算法停止,修改后的 I 为算法的输出.SSI 性能的测试采用了独立型测试和应用型测试相结合的方式.在独立型测试中,采用了测试集 Senseval-3 All-Words Task,实验结果得到 60.40%的消歧精度和 60.40%的召回率,在无监督词义消歧的参与测试者中达到了最高精度;在应用型测试中,将 SSI 算法应用到多种语义消歧问题中,包括自动本体构建、文本的语句排列消歧和定义描述词汇的消歧.在自动本体学习的任务中,分别对 4 个领域的本体学习进行测试,得到了 56%~88%的消歧精度;在语句排列任务的消歧实验中,SSI 达到了 86.84%的精度和 82.58%的召回率,且随着上下文窗口的增大,精度和召回率都有所提高,因为增大窗口大小意味着包括了更多的语义互连信息;在 Senseval-3 Gloss 定义描述的消歧实验中,SSI 达到了非常高的消歧精度(82.6%),但召回率相对较低(32.3%),在同任务竞争者中名列第二.

Agirre[32]提出了两种图论式算法 HyperLex 和 PageRank 用于词义消歧.首先,以大型文集为数据源构建目标词汇的同现图,图中节点表示与目标词同现的词汇,两个词汇在同一个文本段落中出现视为同现,如果图中的两个节点对应的词汇出现在同一个文本段落中,则将相应的节点用边连接起来,根据词汇共现的相对频率为边赋权值.然后,采用 HyperLex 和 PageRank 算法找出图中的 hub 节点.对于 HyperLex 算法,在每一步中,算法寻找图中相对频率最高的节点.如果找到的节点大于给定的阈值,就将其作为 hub.当某个节点被作为 hub 后,其邻居节点就失去了作为 hub 的资格.当找到的节点的相对频率低于规定阈值时算法终止.也可以使用 PageRank 算法在共现图中寻找 hub 节点.PageRank 是一个迭代式的算法,它使用随机行走策略标注图形中所有节点的 page rank 值.节点的 page rank 值不仅与推荐的节点个数有关,而且也与推荐节点本身的 page rank 值相关.一旦代表目标

词义的 hubs 确定之后,则将其到目标词之间的连接边权值置为 0,计算整个图形的最小生成树 MST(minimum spanning tree),使用此 MST 进行词义消歧:对于目标词的每个实例,在 MST 中寻找此实例的上下文词汇,每个上下文词汇获得一系列 rank 值,称为 hub 向量,合并目标词上下文词汇的 hub 向量,将最大分支的向量元素对应的 hub 作为目标词的词义 hub.Agirre 采用了两种特别的测试方法,一种方法是将算法返回的 hub 映射到词汇的 WordNet 词义上去计算消歧算法的精度,另一种方法是用黄金准则以聚类形式评价算法返回的词义的精确度,经过一系列参数调整的过程,在 Senseval-3 sample 测试实验中,消歧精度达到 64.6%,比 MFS(most frequent sense)方法提高了 10 个百分点,比最好的有监督算法仅低 8 个百分点.HyperLex 算法与 PageRank 算法性能相当,但 PageRank 算法的参数更少.

Sinha[33]提出了一种基于图形的、通过度量语义相似性进行词义消歧的方法.与文献[32]类似,首先将待消歧的词汇序列通过一个语义连接图表示出来,图中的顶点为词汇的语义标签,边为语义节点之间的语义依赖关系.为了标明每个依赖关系的程度,使用 6 种基于 WordNet 计算词义相关性的方法(Leacock & Chodorow, Lesk,Wu & Palmer,Resnik,Lin,Jiang & Conrath)来计算边的权重,另外,通过考虑图中节点与其他节点之间的关系,使用 4 种基于图形的中心算法(Indegree,Closeness,Betweenness,PageRank)来度量图形节点的重要性.在 Sensevel-2 和 Sensevel-3 English all-words 数据集上的测试表明:经过词义相似性度量和基于图形的中心算法的优化组合可以得到与当今最好的无监督词义消歧势均力敌的性能(Sensevel-2:precision:58.83%,recall:56.37%,F:57.57%;Sensevel-3:precision:61.90%,recall:36.10%,F:62.80%).该方法由于通盘考虑整个图形的互相依赖信息进行整体消歧,使得此算法比单独考虑每个词汇的消歧更具有吸引力.Navigli[34]提出一种基于图结构的无监督词义消歧来缓解大规模词义消歧中的数据获取瓶颈,采用一系列分析图结构连接性的方法来识别图中的重要节点,进而确定相关词义.Navigli 使用的度量图结构连通性的方法包括局部方法和全局方法两种,局部方法(如 In-degree Centrality,Eigenvector Centrality(PageRank,HITS),Key Player Problem(KPP),Betweenness Centrality 和 Maximum Flow)通过分析图节点与其他节点间的连接关系来确定每个图节点的相关度;全局方法则将图结构作为一个整体而不是以单个节点来考虑问题.Navigli 使用了 3 种著名的全局测量方法:Compactnee, Graph Entropy 和 Edge Density.在 SemCor Corpus 测试集下,KPP 算法在所有的连通性度量算法中达到了最好的消歧效果(precision/recall:40.5%),在 Senseval-3 all words task 测试集下,也是 KPP 算法的消歧效果最好(precision/recall:noun:61.9%,adjective:62.8%,verb:36.1%).实验结果表明,局部连通算法优于全局连通算法,选择合适的连通性测量方法会使消歧性能有所提高.

2.1.3 基于领域信息的词义消歧

语义领域是指人类谈论的范围,如政治、经济、运动、娱乐等,这些领域范围都有各自的术语,并且其中的词汇具有一致性,一个领域是指一组相互间有强语义关联的词汇,利用词义的领域信息进行词义消歧近年来逐渐引起人们的重视.

Proctor 在 1978 年指出,在 LDOCE 词典[35]的编撰中,为了标记词汇的使用而引入的主题领域编码(subject field codes)就是领域属性的一个近似.最早试图将领域信息引入词义消歧的是 Cowie Guthrie 等人[36].他们将 LDOCE 词典中的主题领域编码利用模拟退火技术应用于词义消歧.Yarowsky[37]把主题分类法引入了语料库,实验结果表明,当分类词典中的范畴和语义与主题很好地吻合时,如词语 bass 有两个语义分别属于音乐范畴和动物范畴,正确率很高(99%~100%);当语义涉及到几个主题时,实验效果通常很差,如“interest”的“advantage”语义涉及音乐、娱乐、空间探索和金融多个领域,语义之间缺乏主题独立性,所以正确率偏低(<50%).在专家领域环境中,基于主题领域的词义消歧和词义调节的特定问题被 Basili 等人[38]和 Comaroni 等人[39]成功地解决,Gonzalo 等人[40]进一步强调了与 WordNet 的 synset 相关的领域信息在消歧中的重要性.

2.1.3.1 WordNet Domains简介

2000 年,Magnini 等人[41]开发了 WordNet Domains[42],WordNet Domains 是通过在 Princeton English WordNet 中添加领域标签而形成的一个扩展库,对 WordNet 1.6 中的每个同义词集,人工标注至少 1 个领域标签,这些领域标签总共有大约 200 个,并以层次结构的形式进行组织,一个领域可能包括来自于不同 WordNet 子结

构的不同句法种类的同义词集;领域还可以将同一个词汇的多个词义分成相似的聚类,以减少 WordNet 中的词义划分粒度,降低歧义程度.WordNet Domains 提供了一种自然的方式来建立词义之间语义关系的联系,这种语义关系可以很好地应用于自然语言处理任务中.

2.1.3.2 基于WordNet Domains的词义消歧

WordNet Domains 在词义消歧中的重要性在过去的十几年内引起了一批研究者的关注.Magnini 等人[43]提出了词汇领域消歧(word domain disambignation,简称 WDD)方法.它是词义消歧的一个变种,要求对文本中的每个词汇标注其领域标签而不是词义标签,词汇领域消歧可用于那些粗粒度的词义消歧应用中,如信息检索和基于内容的用户建模.Vossen[44]提出将领域相关信息合并到同义词集中进行词义消歧.Bernardo Magnini[45]详细阐述了领域属性在词义消歧中的作用.他指出,利用词义的领域标签可以在词汇之间建立语义关联,进而用于词义消歧.为了进行词义消歧,Magnini 将所有领域属性用 43 个处于领域层次结构上层的、通用的领域词进行概括,将那些无法概括其领域属性的词义视为杂类,用 Factotum 进行标识.这样的描述在没有损失相关信息的基础上达到了一个良好的抽象程度,同时避免了在那些不能很好地代表文本的领域上运用学习算法所作的无用功.Magnini 根据 One domain per discourse 的启发式思想,为歧义词的每个词义生成一个 42 维的领域向量,称为词义向量,同时为歧义词的上下文也生成其领域向量,称为文本向量,通过比较各词义向量和文本向量间的相似性来确定目标词的词义.词义向量采用一种无监督的生成方法,即将 WordNet Domains 和 Semcor 作为数据源,在 WordNet Domains 相应领域的位置将词义向量的元素值置为 1,否则置为 0,向量的长度则正比于词义在 Semcor 中的频率值.特别地,如果词义的领域为杂类,则其词义向量在每一位元素上的值均为 1,并根据其长度 1 进行规格化.而上下文文本向量的生成是由上下文中的词汇向量组合而成的,组合时充分考虑了每个词汇的位置信息.该方法在 Senseval-2 的参与测试中达到了最好的消歧精度:all-words 和 lexical-sample 任务的精度分别为 75%和 66%,但其召回率较低.与其他参与者的比较可以看出,领域信息对 all-words 任务的改善较为明显,究其原因在于,all-words 任务中的文本相对较长,可以提供准确的上下文环境,进而得到一致的领域信息;对于 lexical-sample 任务,一方面,因为其上下文相对较短,无法提供足够的判断环境,另一方面,由于 Factotum 领域标签的大量存在,导致消歧效果进一步降低,另外,词义向量的无指导生成方法简单、粗糙也影响了消歧效果.

2.1.4 基于百科知识库的词义消歧

2.1.4.1 Wikipedia简介

维基百科 Wikipedia[46]是一部从 2001 年开始的、由数千名志愿者集体构建的、到目前为止内容最多、范围最广、更新最快、完全开放的世界网络百科全书.Wikipedia 是一项基于 Wiki 技术的多语言百科全书协作计划,也是一部用不同语言写成的网络百科全书,其目标及宗旨是为全人类提供自由的百科全书——用多国语言书写的全世界知识的总和.截至 2007 年 12 月,维基百科条目数第一的英文维基百科已有 210 万个条目,而所有 253 种语言的版本共突破 900 万个条目,总用户也超越 1 000 万人.

Wikipedia 的每个文档都是关于一个特定主题或事件的详细解释.除了文本性的内容之外,Wikipedia 还包含了许多结构化的信息,内容相关的文档之间存在丰富的链接结构,每篇文章都被划分到一定的类型中,整个 Wikipedia 形成一个有向无环的类型网络结构.从 2004 年开始 Wikipedia 允许通过类型进行结构化的访问,这些结构化的语义信息都可以用来进行相关度的判断.与 WordNet 相比,Wikipedia 的覆盖范围更广、知识更全面、内容更新更快,是一个集广泛性与结构化为一体的用于自然语言处理的理想资源.

2.1.4.2 基于Wikipedia的词义消歧方法

从 2001 年开始,陆续出现了利用 Wikipedia 这一丰富的信息源进行自然语言处理的研究.Mihalcea[47]研究了一种利用 Wikipedia 知识库建立词义标注集的新方法,在无监督词义消歧界脱颖而出,引起了人们的重视.Wikipedia 中的相关文档之间通过显式的超链接连接了起来,这些超链接可以看作是对相应概念的词义标注,对于有歧异的实体来说,这是很可贵的性质.对一个多义词,首先在 Wikipedia 库的文档中寻找链接标记中含有此词汇的所有段落,比如对于“bar”一词,我们在 Wikipedia 中找到了如下链接型标注:[[musical_notation|bar]];然后通过提取链接标志最左端的部分收集给定歧义词的所有可能的词义标签,对于上面的链接型标注,得到

“bar”的一个可能的词义标签“musical_notation”;最后手工将这些词义标签映射到相应的 WordNet 词义上去,这样一个词义标注文集就建立起来了.有了词义标注实例集就可以用有监督的学习方法进行词义消歧了,首先在预处理步骤中,将标注文本进行分词、词性标注、去停用词等操作,然后以一定的窗口大小从歧义词的上下文中提取其局部和全局特征,并使用一些机器学习方法得到分类模型对新实例进行消歧.利用 Senseval-2 和 Senseval-3 测试集对 49 个多义名词进行消歧得到了 84.65%的精度,比 MFS 方法和 Lesk 方法分别提高了 44 个和 30 个百分点,由此可见,基于 Wikipedia 进行词义消歧是切实可行的.

2.2 基于统计的无监督词义消歧

近年来,随着计算机存储容量和运算速度的飞速提高,通过使用各种机用资源和大规模语料库,计算机能够自动获得各种动态的搭配知识及其统计数据,因而,词义消歧研究中涌现出许多基于语料库统计的方法.在无监督词义消岐领域,统计语料库是未经词义标注的生语料库,所以说,基于统计的无监督词义消岐是一种知识贫乏的、不依赖于任何外部数据源的词义消歧方法,它具有很强的可移植性和鲁棒性,是一种真正意义上的无监督消歧.目前,基于统计的无监督词义消岐大致可以划分为基于聚类、基于双语预料和基于 Web 统计 3 大类.

2.2.1 基于聚类的词义消岐

基于聚类的方法是无监督统计词义消岐中的一项最为常用的技术.Schutze[48]将训练语料中歧义词的上下文聚成 10 个类,每个类别代表一个抽象词义,词义的识别和判断在这些类别里进行.Schutze 的方法容易对付那些词典里查不到的词语,尤其是 Internet 上的新词语层出不穷,常规的词义消歧方法解决不了的问题,而采用该方法可以改善信息检索的效果.Deerwester[49]、Dumais[50]、Landauer[51]使用 Latent Semantic Analysis(LSA)方法进行词义聚类:使用“词-上下文”共现矩阵的形式表示数据集,其中行对应词类型,列对应上下文(上下文可以是短语、句子或段落),矩阵元素表示行对应的词汇在列对应的上下文中出现的次数.首先使用 Singular Value Decomposition(SVD)技术对此矩阵进行降维操作,然后使用 Cosine 函数计算行向量间的语义距离,最后根据语义距离进行词汇的聚类.Landauer 使用 LSA 方法,针对 TOEFL 考试中辨别同义词的题目进行了测试,得到了 74%的正确率.Burgess 和 Lund[52,53]使用 Hyperspace Analogue to Language(HAL)方法行词义聚类:使用“词-词”共现矩阵的形式表示数据集,矩阵元素表示行、列对应的词汇在一个固定长度的上下文窗口内同现,矩阵元素的值随着词汇间距离的增大而减小.每个词用一行和一列值表示,行中的值代表在上下文窗口内该词出现在相应列元素词后面的次数,同理,列中的值代表在上下文窗口内该词出现在相应行元素词前面的次数,这样,每个词汇就有两个上下文向量,最后,将词的行向量和列向量进行连接生成该词的上下文向量.首先使用 Multidimensional Scaling(MDS)技术对此矩阵进行降维操作,然后使用欧式距离计算向量端点之间的距离,最后根据语义距离进行词汇的聚类.Burgess 和 Lund 使用 Usenet newsgroup postings 词汇集进行实验发现,HAL 对于词汇类型和词性的辨别能力均非常准确.Lin 和 Pantel[54]使用 Clustering by Committee(CBC)方法来发现聚类,CBC 方法同样使用“word-context”共现矩阵来表示上下文词汇,此时的上下文是一系列词汇(聚类之前需要先进性分词操作),聚类的结果是与目标词内在词义相关的上下文词汇,即目标词词义的同义词集合.CBC 进行词汇聚类分为 3 个阶段:首先生成同现矩阵,矩阵元素表示目标词与特定上下文词汇在某给定文集中的互信息量,代表词汇间的相似度,相似性最大的 top-k 个上下文词汇作为第 2 阶段的输入;对于初始阶段产生的相似词汇,CBC 使用相同的共现矩阵寻找每个相似词汇的相似词汇,然后使用 average link clustering 将这些词汇进行聚类,并为每个产生的聚类分配一个相似性值,最相似聚类中的词汇形成 committee,如此迭代下去,直到一系列最终的 committee 产生,每个 committee 都是特征化目标词各词义的一系列词汇元素.Lin 和 Pantel 将 CBC 方法产生的聚类词汇与 WordNet 中的同义词集进行了比较,采用转化数(将 CBC 产生的聚类词义转化到相应的 WordNet 同义词集的步数)作为度量标准,得到了 60%~65%的聚类质量,性能超过了同时期的所有词汇聚类法.另外,Pedersen 与 Bruce[55,56]、Purandare 与 Pedersen[57]也运用聚类算法实现了词义消岐,这里不再作介绍.

2.2.2 基于双语语料的词义消岐

基于双语语料的词义消岐在无监督统计词义消岐中备受欢迎.词对齐语料是结构化很强的数据,目标语的词可以看作源语中对应词的词义标记,这样的平行语料就具有了标注语料的功能,使用双语语料从另一个角度

为 WSD 的训练语料建设提供了一种可供选择的好方法.

Dagan[5]在 1991 年指出,两种语言包含的信息比一种语言多,他在 1994 年[58]又探讨了使用第 2 种语言来帮助词义消歧的方法.Resnik 和 Yarowsky[59]在一篇会议论文中正式推介基于双语语料的 WSD 方法.最近几年,公开发表的有关双语词义消歧的学术论文无论在数量上还是在质量上都有了较大的进步,例如,Escudero 等人[60]、Ide 等人[61]、Cong Li 等人[62]为双语语料在 WSD 研究上的应用起到了积极的推动作用.Ng 等人[63]把语言数据协会(linguistic data consortium,简称 LDC)提供的汉英双语语料应用到了词义消歧上,用 Naïve Bayes 模型构造词义分类器,测试了 SENSEVAL-2 中的 29 个名词,将平行语料的实验结果 P 与人工标注语料的结果 M 进行对比,P 基本超过或接近 M,说明平行语料在机器学习模型的训练上是比较有希望的.1999 年,Diab[64]介绍了无指导的词义消歧系统 SALAAM.该系统自动生成 token-level 的对齐,能够同时自动生成英、德、法和西班牙语言的词义标注语料,因此为解决词义消歧的数据获取问题提供了多语言的解决框架;2003 年,Diab[65]对 SALAAM 作了进一步的改进,认为改进后的 SALAAM 作为一个无指导的系统,在 SENSEVAL-2 英语全文词义消歧任务上的表现是当前最出色的;2004 年,Diab[66]将该方法用于增强阿拉伯语词义消歧系统,这是在多语种扩展上的一个应用范例;同年,Diab[67]使用 SALAAM 自动生成了规模较大的标注语料,然后用该训练语料来增强有指导的 WSD 系统.Bhattacharya 等人[68]充分利用了大型知识库 WordNet 的语义和概念体系来确定两个概率模型(分别是语义模型和概念模型)的结构,模型建立后,用通行的 EM 算法训练概率参数,实验结果表明,Bhattacharya 等人建立的语义模型在词义消歧上比 Diab 实现的 SALAAM 系统表现得更好,而概念模型又比语义模型要强很多.在国内,李涓子和黄昌宁[23]提出的基于转换的汉语词义消歧的无指导方法也具有一定的代表性.

2.2.3 基于 Web 的词义消歧

World Wide Web 不仅是当今最丰富、领域最广泛的自然语言文本资源,而且 Web 内容的增长和更新快速、及时,另外,由于因特网上的 Web 页面内容是由无数人编辑而成的,将其作为词义消歧数据源的有效性因此得到了保证.目前存在一些强大的基于关键字的 Web 搜索引擎(如 Google),可以用于基于 Web 的词义消歧.

Klapaftis[69]提出了利用 Web 统计进行词义消歧的方法.首先将包含待消歧词的句子提交给 Google 搜索引擎,取返回的前 4 个文档,将这些文档进行分词和词性标注后称为集合 U;然后对消歧词的每个词义标签,在 WordNet 中检索其所有距离不超过 3 的语义相关词汇,形成对应的语义词汇列表,统计相关词汇在 U 中出现的频率,根据相关词汇与目标词的距离计算各语义词汇列表在消歧中的重要性权值,然后合并以上计算得到的语义关系词汇列表值,得到每个词义的整体消歧值,对消歧词的每个词义都计算了 TSS 之后,取 TSS 值最大的词义作为待消歧词的最终词义.使用 Semcor 2.0 的前 10 个文件并进行参数调整后得到 65.91%的消歧精度.YANG Che-Yu[70]提出了另一项使用 Web 统计进行词义消歧的技术.给定待消歧词及其上下文词汇,首先,在 WordNet 中搜索它们各自的词义,然后,应用集合代数和布尔操作技巧,基于 WordNet 的上下文层次结构,对每个找到的词义建立其概化和泛化表示,这是一种概念的规则化表示形式;对于所有的词义组合对(一个来自于消歧词,另一个来自于上下文词汇),将其分别用以上形式表示之后,提交给 Yahoo 搜索引擎,使用规格化的命中数作为词义对之间的相关性值,将消歧词的每个词义与上下文词汇的相关性值求和,和值最大的词义为消歧词的最终词义.从 Semcor corpus 中随意选出 4 个文本文件进行测试,得到 72%~85%的消歧精度.

3 结论与展望

本文对无监督词义消歧进行了大量的研究工作,从数据源的使用、采用的消歧手段、评估体系、消歧效果等方面详细分析了当前国内外各种主流无监督词义消歧采用的关键技术和研究进展,并指出研究中存在的问题和未来的研究重点,为该方向的研究者提供了较为全面的参考.现将 14 个具有代表性的无监督词义消歧算法进行总结(见表 1),表 1 给出了每种算法的提出时间、所属的类别、使用的数据源、采用的关键消歧技术、使用的测试集以及消歧效果.

Table 1 Comparison on part of typical unsupervised word sense disambiguation algorithms

表 1 部分代表性无监督词义消歧算法比较

Categories			Algorithms	Data sources	Key techniques	Test sets	Precision (%)	Recall (%)
Knowledge-Based WSD	Dictionary-Based	Gloss description-based	Lesk[2]	Machine-Readable dictionary	Gloss overlapping	Short samples	50~70	100
			Satanjeev[25]	WordNet 1.7	Adapted lesk	Senseval-2 lexical sample task	32.3	100
			Chen[20]	WordNet 1.7	AALesk	Senseval-2 lexical sample task	32.6	100
				WordNet 2.0			33.4	100
		Concept density-based	Agirre[22]	WordNet 1.4	Concept_Density	SemCor	76.04	23.21
			Rosso[23]	WordNet 1.6	Concept_Density; Semcor_Frequency; WN_Domains	Nouns in 19 randomly selected SemCor files	81.48	60.17
			Davide[24]	WordNet 1.6	Concept_Density; Semcor_Frequency; WN_Domains; WN_Samples	Senseval-2 all-Words task	80.12	59.96
						Senseval-3 AWT	74.06	52.03
		Graph theory-based	SSI[31]	WordNet; Machine-Readable dictionary	WN_Dmains; WN_Gloss; Collocations; WN_Samples	Sensevel-3 all-words task	60.40	60.40
						Senseval-3 gloss disambiguation	82.60	32.30
						Sentence collocations	86.84	82.58
						Automatic ontology learning	87.50	87.50
			Agirre[32]	WordNet	PageRank; HyperLex	Senseval-3 lexical sample task	64.60	64.60
			Sinha[33]	WordNet	Semantic similarity; Graph centrality	Sensevel-2 all-words task	58.53	56.37
						Sensevel-3 all-words Task	61.90	36.10
			Navigli[34]	WordNet; EnWordNet	Graph connectivity	SemCor	31.80	31.80
						Senseval-3 all-words task	40.50	40.50
	Domain-Based		Magnini[45]	WordNet; WN_domains	Domains; Semcor_Frequency	Senseval-2 all-words task	75	Lower
						Senseval-2 lexical sample task	66	Lower
	Wikipedia-Based		Mihalcea[47]	Wikipedia	Links; Pipe Links	49 nouns in Senseval-2 & Senseval-3	84.65	84.65
Statistics-Based WSD	Web-Based		Klapaftis[69]	The Web	Google; WordNet taxonomy	First 10 files of SemCor 2.0	65.91	94.32
			YANG[70]	The Web	Yahoo; WordNet taxonomy	4 randomly selected SemCor 2.0 files	72~85	100

对于基于知识的方法,知识的来源可以是规则库、词典、本体和其他知识库.来自于各种知识库的词义描述对消歧具有积极作用.机读词典首先被研究者们所重视,并成为 20 世纪 80 年代词义消歧工作的主要知识源.基于机读词典消歧的典型方法是:利用单词在词典中不同义项的定义,计算多义词各词义的定义和上下文词汇的词义定义重叠度,选择重叠度最大者作为当前词义.遗憾的是,这种方法的正确率不高,主要原因在于:(1) 传统基于机读词典的方法没有充分利用词典中的短语、示例等信息;(2) 机读词典中词义定义语句一般较短,以至于很多情况下,无论歧义的哪一种词义的定义与上下文单词的定义的重叠度均为 0;(3) 在实际应用中,不可避免地组合爆炸也限制了方法的使用.词典主要为人工使用而非机器开发而创,而且本身难免存在不协调之处,带来知识自动抽取的困难.如何将机读词典转化为机循词典(machine tractable dictionary),从中学习词义消歧知识并提高效率是基于机读词典方法的重点和难点.

WordNet 是一个综合性强、覆盖范围广、语义知识丰富的义类词典,又是一个免费资源,近年来将 WordNet 作为知识源进行词义消歧在无监督词义消歧领域一直占据主导地位,出现了多种成功的消歧方法.其中基于概念区域密度进行消歧的方法通过综合使用 WordNet 和 WordNet Domains 知识库,从不同的层面(定义、语义关系、词频信息、领域属性)来确定词义,突破了传统的基于 WordNet 中单一知识源进行词义消岐的局限,达到了前所未有的好效果.受该方法的启发,下一步的研究方向应该考虑合理利用有效知识库中的各项知识源进行多方位的消歧,进而获得更好的消歧精度,同时,多知识源的使用可以互补有无,增加消歧的召回率.另外,图论式结构化语义互连方法(SSI)取得了相当高的消歧精度.尽管它是一种半监督的消歧方法,但其中的很多原理也值得研究者借鉴,比如词义描述的生成、采用的消歧方案等.最后需要指出的是,WordNet 中对词义的划分粒度过细,

以至于利用 WordNet 进行词义消歧的效果很不理想,其实,一些实际的自然语言处理应用对词义的辨别粒度不需要很细,所以下一步的研究方向可以考虑将 WordNet 中的词义进行语义聚类,得到粗粒度的词义划分,降低词汇的歧义程度.

词义的领域信息或主题分类信息对词义消歧具有积极的作用,但是目前普遍采用的 WordNet Domains 领域知识库并不完备,只有部分词义被划分到特定领域,大多数词义作为通用领域词,其领域标签为“Factotum”,对消歧没有任何帮助,所以使用领域信息进行消歧的召回率很低,要想更好地使用这类有用信息,必须首先完备领域知识库,这又是一个长久而艰巨的任务.

Wikipedia 百科知识库的日益壮大和完善促使其作为自然语言处理的丰富知识源.从 2001 年开始,陆续出现了一批利用 Wikipedia 这一丰富的知识源进行自然语言处理的研究.前面介绍了一种利用 Wikipedia 知识库中的超链接信息建立词义标注集进而进行词义消歧的方法[53]是一项利用 Wikipedia 知识库进行词义消歧的研究.事实上,由于 Wikipedia 出现时间很短,利用 Wikipedia 进行的研究也相对较少,而 Wikipedia 是一种不可多得的、丰富的知识资源,下一步应考虑综合利用 Wikipedia 中的各种知识源(如文档描述、超链接、文档类型、消歧页等)进行基于知识的词义消歧,随着 Wikipedia 的不断发展和完善,可以预计,基于 Wikipedia 的消歧将会达到更好的效果.

基于统计的词义消歧方法使用大型语料作为消歧的数据源,其正确率比基于词典和基于规则的方法有明显的提高,但语料库通常都具有领域局限性,如何获得通用语料库是一个亟待解决的问题.双语词义消歧方法在训练语料的建设上另辟蹊径,用双语语料作为机器学习的知识源,获得了良好的应用效果,但是大规模平行语料还是比较难获得的.缺乏高质量的大规模平行语料,影响了该方向的研究进展.Web 这一丰富的世界知识库对于统计词义消歧来说是一个绝好的数据源,可以解决现有的知识获取瓶颈问题,Web 作为一个无词义标注的大型语料库,具有信息量丰富、覆盖面广、信息更新快等优点,此方法的出现开辟了自然语言处理的新纪元.已有方法只利用了搜索引擎的命中率这一项知识源进行消歧,进一步探索如何利用 Web 知识库中各项信息源(页面内容、超级链接、访问日志等)进行词义消歧是一个很有前途的研究方向.

经过多年的尝试,越来越多的研究者倾向于综合多种方法进行集成消歧,通过组合多种知识源和多种消歧手段,获得更好的消歧性能.知识源的组合扩展了消歧的可用知识,多种方法的组合可以有针对性地解决不同的歧义现象.未来的研究应该从这方面入手,对各种无监督词义消歧算法进行集成,扬长补短,进一步提高消歧精度,同时综合考虑计算复杂度的问题.

无监督词义消歧的研究历史比较短,只有二十几年的发展历程,还是一个比较年轻的研究方向.近年来,大量无监督词义消歧方面的学术论文的涌现标志着该领域的研究进入蓬勃发展的阶段.但是在可参考的文献中,讨论英文词义消歧问题的占绝大多数,对其他语种词义消岐的研究较少,特别是中文词义消岐还处于起步阶段.由于人类进行词义判断的过程是一个综合利用词法、语法和背景知识进行消歧的过程,对于计算机来说,要真正有效地提高词义消歧的水平,不仅需要获取词的释义和分类信息,而且更重要的是,综合利用现有的语法和语义资源,在词类划分基础上,增加词义的语法功能和语义搭配的分析,从多种知识源中提取词义间相互区别的分布特征,从而完成词义的判定.

References:

[1] Nancy I, Jean V. Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics, 1998,24(1):1−40.

[2] Lesk M. Automated sense disambiguation using machine readable dictionaries: How to tell a pine cone from all ice cream cone. In: Proc. of the S1GDOC Conf. New York: Association for Computing Machinery, 1986. 24−26.

[3] Wilks Y, Stevenson M. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report, CS-96-05, Sheffield: University of Sheffield, 1996.

[4] Pook SL, Catlett J. Making sense out of searching. Technical Report, Sydney: AT&T Bell Laboratories, 1988.

[5] Dagan I, Itai A, Markovitch S. Two languages are more informative than one. In: Proc. of the 29th Annual Meeting of Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 1991. 130−137.

[6] Gale WA, Church KW, Yarowsky D. A method for disambiguation word senses in a large corpus. Computer and the Humanities, 1993,26(5-6):415−439.

[7] Miller GA. WordNet: A lexical database for the English language. Communications of the ACM, 1995,38:39−41.

[8] Roget PM. Roget's thesaurus. 1852. http://dictionary.reference.com

[9] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: Proc. of the 32nd Annual Meeting of Association for Computational Linguistics. Las Cruces: Association for Computational Linguistics, 1994. 88−95.

[10] Lu ZM, Liu T, Li S. The research progress of statistical word sense disambiguation. Acta Electronica Sinica, 2006,34(2):333−343 (in Chinese with English abstract).

[11] Towell G, Voorhees EM. Disambiguating highly ambiguous words. Computational Linguistics, 1998,24(1):125−145.

[12] Resnik PS. Selection and information: A class-based approach to lexical relation [Ph.D. Thesis]. Pennsylvania: University of Pennsylvania, 1993. 23−54.

[13] Mei JJ, Zhu YM, Gao YQ, Yin HX. Tong Yi Ci Ci Lin. Shanghai: Shanghai Lexicographical Publishing House, 1983 (in Chinese).

[14] Dong ZD. HowNet Knowledge Database. 2002 (in Chinese). http://www.keenage.com/

[15] Chen H, He TT, Ji DH. An unsupervised approach to word sense disambiguation based on HowNet. Journal of Chinese Information Processing, 2005,19(4):10−16 (in Chinese with English abstract).

[16] Chen H, He TT, Ji DH. Unsupervised approach to word sense disambiguation based on Hownet. Mini-Micro Systems, 2005,26(10): 1846−1849 (in Chinese with English abstract).

[17] Li JZ, Huang CN. Unsupervised word sense tagging method based on transformation rules. Journal of Tsinghua University (Science and Technology), 1999,39(7):117−121 (in Chinese with English abstract).

[18] Lu S, Bai S, Huang X. An unsuptervised approach to word sense disambiguation based on sense-words in vector space model. Journal of Software, 2002,13(6):1082−1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/1082.htm

[19] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Proc. of the 3rd Int'l Conf. on Intelligent Text Processing and Computational Linguistics. Mexico City, 2002. 17−23.

[20] Chen YQ, Yin J. Sense rank AALesk: A semantic solution for word sense disambiguation. In: Wang L, Jin Y, eds. Fuzzy Systems and Knowledge Discovery. LNAI 3614, Berlin, Heidelberg: Springer-Verlag, 2005. 710−717.

[21] Ledo-Mezquita Y, Sidorov G, Cubells V. Combined Lesk-based method for words senses disambiguation. In: Proc. of the 15th Int'l Conf. on Computing (CIC 2006). Washington: IEEE Computer Society, 2006. 105−108.

[22] Agirre E, Rigau G. A proposal for word sense disambiguation using conceptual distance. In: Mitkov R, Nicolov N, eds. Proc. of the 1st Int'l Conf. on Recent Advances in NLP. 1995. 162−171.

[23] Rosso P, Masulli F, Buscaldi D, Pla F, Molina A. Automatic noun disambiguation. In: Proc. of the 4th In'l Conf. on Computational Linguistics and Intelligent Text Processing. Mexico City: Springer-Verlag, 2003. 273−276.

[24] Buscaldi D, Rosso P, Masulli F. Integrating conceptual density with WordNet domains and CALD glosses for noun sense disambiguation. In: Proc. of the4th Int'l Conf. on Espana for Natural Language Processing (EsTAL). LNAI 3230, Alicante, 2004. 183−194.

[25] Mihalcea R. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In: Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2005. 411−418.

[26] Litkowski KC. Use of machine readable dictionaries in word sense disambiguation for Senseval-2. In: Proc. of the ACL/SIGLEX Senseval-2. Toulouse: Association for Computational Linguistics Special Interest Group, 2001.

[27] McCarthy D, Koeling R, Weeds J, Carroll J. Using automatically acquired predominant senses for word sense disambiguation. In: Proc. of the ACL/SIGLEX Senseval-3 Workshop. Barcelona: Association for Computational Linguistics, 2004.

[28] Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 1991,17(1):21−48.

[29] Galley M, McKeown K. Improving word sense disambiguation in lexical chaining. In: Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). Acapulco, 2003. 1486−1488.

[30] Mihalcea R, Tarau P, Figa E. PageRank on semantic networks with application to word sense disambiguation. In: Proc. of the 20th Int'l Conf. on Computational Linguistics (COLING). Morristown: Association for Computational Linguistics, 2004.

[31] Navigli R, Velardi P. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE

Trans. on Pattern Analysis and Machine Intelligence, 2005,27(7):1075−1086.

[32] Agirre E, Martinez D, de Lacalle OL, Soroa A. Two graph-based algorithms for state-of-the-art WSD. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). Barcelona: Association for Computational Linguistics, 2006. 583−593.

[33] Sinha R, Mihalcea R. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: Proc. of the IEEE Int'l Conf. on Semantic Computing (ICSC). Washington: IEEE Computer Society, 2007. 363−369.

[34] Navigli R, Lapata M. Graph connectivity measures for unsupervised word sense disambiguation. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI). 2007. 1683−1688.

[35] Longman Dictionary of Contemporary English. Beijing: Foreign Language Teaching and Research Press, 2003. http://www.ldoceonline.com/

[36] Cowie J, Guthrie J, Guthrie L. Lexical disambiguation using simulated annealing. In: Proc. of the Workshop on Speech and Natural Language. Morristown: Association for Computational Linguistics, 1992. 359−365.

[37] Yarowsky D. Word sense disambiguation using statistical models of Roget's categories train on large corpora. In: Proc. of the 14th Conf. on Computational linguistics. Morristown: Association for Computational Linguistics, 1992. 545−460.

[38] Basili R, Rocca MD, Pazienza MT. Contextual word sense tuning and disambiguation. Applied Artificial Intelligence, 1997,11(3): 235−262.

[39] Comaroni JP, Beall J, Matthews WE, New GR. Dewey Decimal Classification and Relative Index. Albany: Forest Press, 1998.

[40] Gonzalo J, Verdejio F, Peters C, Calzolari N. Applying EuroWordNet to cross-language text retrieval. Computers and the Humanities, 1998,32(2-3):185−207.

[41] Magnini B, Cavaglia G. Integrating subject field codes into WordNet. In: Proc. of the 2nd Int'l Conf. on Language Resources and Evaluation (LREC-2000). 2000. 1413−1418.

[42] WordNet domains. Princeton University, 2007. http://wndomains.itc.it/

[43] Magnini B, Strapparava C. Experiments in word domain disambiguation for parallel texts. In: Proc. of the ACL Workshop on Word Senses and Multilinguality. Morristown: Association for Computational Linguistics, 2000. 27−33.

[44] Vossen P. Extending, trimming and fusing WordNet for technical documents. In: Proc. of the NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Morristown: Association for Computational Linguistics, 2001.

[45] Magnini B, Strapparava C, Pezzulo G, Gliozzo A. The role of domain information in word sense disambiguation. Natural Language Engineering, 2002,8(4):359−373.

[46] Wikipedia, the Free Encyclopedia. http://www.wikipedia.org/

[47] Mihalcea R. Using wikipedia for automatic word sense disambiguation. In: Computational Linguistics (NAACL). Morristown: Association for Computational Linguistics, 2007. 196−203.

[48] Schütze H. Automatic word sense discrimination. Computational Linguistics, 1998,24(1):97−123.

[49] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41(6):391−407.

[50] Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 1997,104(2):211−240.

[51] Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes, 1998,25(2):259−284.

[52] Burgess C, Lund K. Modeling parsing constraints with high-dimensional context space. Language and Cognitive Processes, 1997, 12(2-3):177−210.

[53] Burgess C, Lund K. The dynamics of meaning in memory. In: Dietrich E, Markman A, eds. Cognitive Dynamics: Conceptual Representational Change in Humans and Machines. 2000. 117−156.

[54] Lin DK, Pantel P. Concept discovery from text. In: Proc. of the 19th Int'l Conf. on Computational Linguistics (COLING). Morristown: Association for Computational Linguistics, 2002. 577−583.

[55] Pedersen T, Bruce R. Distinguishing word senses in untagged text. In: Proc. of the 2nd Conf. on Empirical Methods in Natural Language Processing. 1997. 197−207.

[56] Pedersen T. Knowledge lean word sense disambiguation. In: Proc. of the 15th National Conf. on Artificial Intelligence. Morristown: Association for Computational Linguistics, 1998. 800−805.

[57] Purandare A, Pedersen T. Word sense discrimination by clustering contexts in vector and similarity spaces. In: Ng HT, Riloff E,

eds. Proc. of the Conf. on Computational Natural Language Learning. Boston: Quinlan Publishing, 2004. 41−48.

[58] Dagan I, Itai A. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 1994,20(4): 563−596.

[59] Resnik P, Yarowsky D. A perspective on word sense disambiguation methods and their evaluation. In: Light M, ed. Proc. of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How. Morristown: Association for Computational Linguistics, 1997. 79−86.

[60] Escudero G, Màrquez L, Rigau G. Boosting applied to word sense disambiguation. In: Proc. of the 12th European Conf. on Machine Learning. Berlin, Heidelberg: Springer-Verlag, 2000. 129−141.

[61] Ide N, Ejavec T, Tufis D. Sense discrimination with parallel corpora. In: Proc. of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. Morristown: Association for Computational Linguistics, 2002. 54−60.

[62] Li C, Li H. Word translation disambiguation using bilingual bootstrapping. In: Proc. of the 40th Annual Meting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2002. 343−351.

[63] Ng HT, Wang B, Chan YS. Exploiting parallel texts for word sense disambiguation: An empirical study. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2003. 455−462.

[64] Diab M, Resnik P. An unsupervised method for word sense tagging using parallel corpora. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2002. 255−262.

[65] Diab M. Word sense disambiguation within a multilingual framework [Ph.D. Thesis]. Maryland: University of Maryland College, 2003.

[66] Diab MT. An unsupervised approach for bootstrapping Arabic word sense tagging. In: Proc. of the Arabic Based Script Languages, COLING 2004. Morristown: Association for Computational Linguistics, 2004.

[67] Diab M. Relieving the data acquisition bottleneck in word sense disambiguation. In: Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 303−310.

[68] Bhattacharya I, Getoor L, Bengio Y. Unsupervised sense disambiguation using bilingual probabilistic models. In: Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2004. 287−294.

[69] Klapaftis IP, Manandhar S. Google & WordNet based word sense disambiguation. In: Proc. of the 22nd ICML Workshop on Learning & Extending Ontologies. New York: Association for Computing Machinery, 2005.

[70] Yang CY. Word sense disambiguation using semantic relatedness measurement. Journal of Zhejiang University (Science A), 2006, 7(10):1609−1625.

附中文参考文献:

[10] 卢志茂,刘挺,李生.统计词义消歧的研究进展.电子学报,2006,34(2):333−343.

[13] 梅家驹,竺一鸣,高蕴琦,殷鸿翔.同义词词林.上海:上海辞书出版社,1983.

[14] 董振东.知网.2002. http://www.keenage.com/

[15] 陈浩,何婷婷,姬东鸿.基于 *k*-means 聚类的无导词义消歧.中文信息学报,2005,19(4):10−16.

[16] 陈浩,何婷婷,姬东鸿.基于 MDL 聚类的无导词义消歧.小型微型计算机系统,2005,26(10):1846−1849.

[17] 李涓子,黄昌宁.基于转换的无指导词义标注方法.清华大学学报(自然科学版),1999,39(7):117−121.

[18] 鲁松,白硕,黄雄.基于向量空间模型中义项词语的无导词义消歧.软件学报,2002,13(6):1082−1089. http://www.jos.org.cn/1000-9825/13/1082.htm

王瑞琴(1979−),女,内蒙古鄂尔多斯人,博士,讲师,主要研究领域为语义挖掘,智能信息检索.

孔繁胜(1946−),男,教授,博士生导师,主要研究领域为机器学习,数据挖掘,CAD/CG.