统计词义消歧的研究进展

卢志茂[1,2], 刘 挺[1], 李 生[1]
(1. 哈尔滨工业大学计算机学院信息检索研究室, 黑龙江哈尔滨 150001;
2. 哈尔滨工程大学计算机科学与技术学院, 黑龙江哈尔滨 150001)

摘 要: 本文参考大量的文献资料, 分析了当前国内外统计词义消歧研究中采用的多种方法和技术, 指出了统计词义消歧研究的关键问题, 并围绕关键问题阐述了统计词义消歧的研究进展, 探讨了研究中存在的问题和未来研究的重点.

关键词: 统计词义消歧; 自然语言处理; 综述
中图分类号: TP301.6 文献标识码: A 文章编号: 0372-2112 (2006) 02-0333-011

The Research Progress of Statistical Word Sense Disambiguation

LU Zhimao, LIU Ting, LI Sheng
(1. Information Retrieval Laboratory of Computer Science & Technology School, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China;
2. Computer Science & Technology School, Harbin Engineering University, Harbin, Heilongjiang 150001, China)

Abstract Abstract Many kinds of statistical word sense disambiguation (WSD) methods and technologies in home and abroad are analyzed, and lots of literatures are referred to in this paper. Key questions of statistical WSD research are pointed out, and the research progress is illustrated around them. In the final section the problems and future research emphases in WSD are explored in this paper.

Key words statistical word sense disambiguation; natural language processing; survey

1 词义消歧定义及其研究意义

词义消歧 (Word Sense Disambiguation, WSD) 是计算语言学和自然语言处理领域一个重要的研究课题, 也是近些年来该领域的热点研究问题之一. 早在上个世纪的 50年代, 解决自然语言中歧义问题的研究就引起了人们的关注[1]. 1960年 Bar-Hillel指出词语歧义问题是机器翻译 (Machine Translation) 研究所面临的主要障碍[1], 在他的著作中把词义消歧描述为不可攻克的难题, 致使从那以后的一段时间里研究人员知难而退, 逐渐放弃了对词义消歧, 甚至包括对机器翻译进行研究的希望.

从 20世纪 80年代, 随着计算技术的飞速发展, 超大容量的存储设备和强大计算能力的处理器相继出现. 同时, 以统计学习为代表的机器学习理论的日臻完善. 有了计算机软硬件技术的强大支持, 包括词义消歧在内的自然语言处理领域的各问题研究一一复苏, 并进入了崭新的发展阶段.

1.1 词义消歧的定义

自然语言处理的最终目的就是让机器像人一样理解并使用自然语言, 达到最高层次的"人-机交流". 如果要达到这个目的, 还有很长的路需要走, 很多问题亟待解决. 其中词义理解就是一个很重要的问题, 也是自然语言处理面临的难关之一.

词义消歧是一项词义理解技术, 能够让机器代替人去分析、判断和识别特定语言环境中的词义信息.

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 N 个词义中确定词义 S′. 每个词义 S_k 和上下文 C 存在或强或弱的关系 $R(S_k|C)$, 而 S′同上下文 C 的关系应该是最强的. 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, 排除非正确的词义, 最后确定 S′. 整个过程可用下面的公式描述.

$$S' = \arg\max_k R(S_k|C)$$

通过上下文来选择多义词的正确词义, 是模拟人的思维判断过程. 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, 帮助了人迅速地去推理、判断, 最终得出答案. 机器模拟人思维的过程, 在上下文中收集重要的语言信息, 提取特征词语, 来指导对多义词的歧义消解.

收稿日期: 2005-03-25; 修回日期: 2005-11-14
基金项目: 国家自然科学基金 (No. 60435020); 国家自然科学基金 (No. 60575042); 国家自然科学基金 (No. 60573072)

1.2 相关概念

研究中经常用到的两个概念,一个是“词义消歧”(WSD),一个是“词义标注”(WST),这两个概念所代表的含义差别不是很大,都是为词语打上词义标记.本文认为两个概念的差别在于:WSD是研究词义消解的方法,而研究的对象有可能是几个词语,有可能是所有的歧义词,而研究的角度和解决的问题也多种多样,例如研究WSD的数学模型,或者研究WSD模型的特征提取、数据稀疏问题,或者研究WSD模型的训练问题等等,重点在于探索解决歧义消解的新方法、新技术;而WST是利用现有的WSD技术对语料中的歧义词进行词义的标注,重点在于词义的标注任务,而不是寻求新的方法,其前提是已经有了可满足需要的WSD技术.

早期文献中用过概念“词义辨别”(Word Sense Discrimination),是指仅仅识别词语的词义,而不进行标注[2].词义辨别不需要带有词义标记的训练语料,也不标注词义,被看成是词义的聚类,属于完全无指导的词义处理方法.相反,WSD/WST要为词语标注词义,而词义标记是事先约定好的,从该角度讲,WSD/WST应该属于典型的词义分类问题[1~3].在不产生歧义的前提下,本文不刻意区分这几个概念.

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,甚至有的研究领域把词义消歧作为重要步骤或关键环节.如果词义自动消歧问题得到有效的解决,将对包括机器翻译、文本分类、自动文摘、信息检索、文本挖掘、语音识别、文语转换等在内的许多自然语言问题的研究和应用产生巨大的帮助.

除了以上列出的重要应用外,词义消歧技术还可以引入到语法分析或句法分析之中,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一个重要的推动作用.

2 统计词义消歧方法概述

随着统计学的发展以及计算机计算能力和存储能力的大幅度提升,诞生了一种新的研究方向——统计自然语言处理(Statistical Natural Language Processing, SNLP).借助统计学的思想和方法来处理词义歧义问题,将词义消歧研究带入了一个崭新的天地.统计学的方法不需要制定规则,自动在语料库、知识库中获取所需信息,来指导词义消歧.

基于统计的词义消歧需要有语言资源支持,可以使用的语言资源有机器可读词典(Computer Readable Dictionary, CRD)和语料库(Corpora).不同的语言资源实现的方法也大不相同,所以统计词义消歧又派生出基于词典的词义消歧(Dictionary-Based word sense Disambiguation)和基于语料库的词义消歧(Corpus-Based word sense Disambiguation).

基于词典的词义消歧是上个世纪80年代后期和90年代初期比较流行并很受推崇的方法.基于词典的方法有时也被称作基于知识(Knowledge-based)的方法,因为使用词典或者类似词典的知识库.由于,目前知识库建设从规模到内容还很不完善,所以进入90年代,随着语料库语言学的兴起,涌现出一大批性能卓越的统计词义消歧新方法,基于词典的方法逐渐受到冷落,并慢慢退出了词义消歧研究的历史舞台.

基于语料库的方法借助计算机强大的统计计算能力,展现出以往任何方法都无可比拟的优势,一经出现就引起了业内研究人员的广泛关注,并迅速开辟了一块计算语言学研究的新天地.十几年来,在一批批饱含激情的研究人员的共同努力下,语料库语言学获得了空前的发展.随着一批批基于语料库技术的高性能信息处理系统的产品化和市场化,语料库语言学也奠定了在自然语言处理领域的主流和先锋地位.

根据语料库的使用方法的不同,还可以把基于语料库的词义消歧方法分为两类,一类是基于实例的词义消歧方法,一类是基于概率统计的词义消歧方法.

2.1 基于词典的消歧方法

基于词典的词义消歧始于1986年,Lesk直接利用词典中词义(亦称“义项”)解释或定义来指导歧义词的词义判断[4].该方法简单易行,只需计算歧义词的各个词义在词典中的定义与歧义词上下文词语的定义之间的覆盖度,选择覆盖度最大的作为正确的词义.Lesk公布该方法的正确率在50~70%之间,不是很理想,同时他提出通过迭代计算来改进该方法的建议.1988年,Pook和Catlett提出另外一种新的改进方法:对上下文词语进行同义词扩展,从而扩大了上下文的窗口,可以增大计算覆盖度的成功率[3].

1992年,Yarowsky把主题分类方法引入了语料库,实验结果表明,分类词典中的范畴和语义与主题很好的吻合时,如词语“bass”有两个词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%)[5].

1995年Agirre等人采用WordNet的分类体系计算歧义词及其上下文词语的概念密度(Conceptual Density),选择具有最大概念密度的词义作为正确词义,正确率为80%左右[6].

基于词典的词义消歧不需要训练语料,也不需要对词典等资源进行人工处理,可以实现完全自动的消歧系统.但是现有的词典知识缺乏必要的完备性,特别是分类词典覆盖面不够宽,对词语的一般性分类在某些专门领域往往

不适用.此外,词典知识通常是静态的,难以适应语言的动态变化和发展,缺乏足够的可扩展性和灵活性.词典的这些缺陷限制了该方法的发展.

2.2 基于实例的消歧方法

基于实例的词义消歧方法与基于实例的机器翻译方法[7]十分相似,可以说该方法是受到了 MT 方法的启发.基于实例的词义消歧方法有两个关键问题,一个是词义消歧实例的获得,另一个是实例间相似度的计算.

1996年 Ng等人在其 LEXAS系统中成功地采用了基于实例的词义消歧方法[8]. Ng等人采用 WordNet定义词语的词义,综合利用多种知识源来指导词义的判断.其中包括多种语法知识,如上下文的词性知识、歧义词的用法、词语搭配关系等等.实验中, Ng选择了 191个常用的歧义词,共获得 192, 800个实例,事先对包含这些实例的句子进行人工词义标注,并将这些实例作为训练数据,测试的正确率达到 69%.

2.3 基于统计学习的消歧方法

语料库语言学已经成为当今自然语言处理的研究热点之一,基于语料库的方法也成为处理自然语言问题的主流方法.目前,词义消歧研究和其它自然语言问题一样离不开语料库的支持.

基于统计的词义消歧方法,运用统计学技术手段自动在训练语料中获取所需的知识,如歧义词与上下文词语之间的语法关系或语义关系等,并将这些"知识"用于词义的识别和判断. 1991年 Brown率先把统计模型引入了词义消歧研究[9],目前已经证实很多常见的机器学习方法,如决策树 (Decision Tree)、支持向量机 (Support Vector Machine, SVM)、最大熵 (Maximum Entropy, ME)都可以用于统计词义消歧.

机器学习方法根据训练语料的不同分为有指导机器学习 (Supervised Machine Learning)和无指导机器学习 (Unsupervised Machine Learning).有指导的词义消歧需要知道训练样本的词义类别,也就是需要训练语料中的词语词义标记 (sense label),而无指导词义消歧没有这个要求.所以有指导的词义消歧常被看作词义分类问题 (Classification Task),无指导词义消歧被看作聚类问题 (Clustering Task).

把无指导词义消歧看作词义聚类,是说连词典资源的支持都不需要了,是彻头彻尾的无指导方法,真正意义的无指导. 1998年 Schütze将训练语料中歧义词的上下文聚成若干个类,每个类别代表一个抽象"词义"(这里的"词义"与词典中的定义不要求对应),词义的识别和判断在这些类别里进行[9]. Schütze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.

有指导的机器学习方法在词义消歧问题中取得了较好的效果,但是该类方法为了克服数据稀疏问题,获得更好的学习和消歧效果,必须有规模很大的标注语料库的支持.而标注语料的获得需要耗费代价高昂的人工,很难实现基于大规模标注语料的有指导词义消歧工作,客观上也限制了该类方法的推广和应用.而无指导的词义消歧方法不依赖于人工标注的语料,可以实现跨领域大规模真实语料的训练和学习,能够有效克服数据稀疏问题,所以该类方法开始引起了研究人员的重视,该方面的研究报道也多了起来.

3 统计词义消歧的关键问题及研究现状

统计词义消歧的研究从词典资源的选择和确立,到训练和测试数据的准备,到数学模型的建立,涉及很多的中间环节.这些环节进行的好坏都会直接或间接的影响最终的实验结果.其中有些环节是至关重要的,本文称之为统计词义消歧关键问题.文献中针对这些问题提出的解决方法体现了国内外当前统计词义消歧研究的水平.

3.1 词典(或相关知识库)

除了完全无指导的方法以外,几乎所有的词义消歧系统都需要有一个词典资源,不管是普通的释义词典,还是分类辞典,只要机器可读就可以作为资源加以利用.目前比较知名的,如英文有 Roget's, Longman's, WordNet等;汉语有《同义词词林》[10]和知网 (HowNet)[11].

WordNet是目前最为成功的一部英文在线语义词典,它由普林斯顿大学认知科学实验室的 Miller, Beckwith等人,自 1985年起开发的一部在线词典数据库系统,也是一部由心理语言学家和计算机科学家共同努力下创建的独具特色的英文语义网络系统[12]. WordNet使用同义词集合 (Synset)来代表词汇概念,将英语的名词、动词、形容词、和副词组织为 Synset并描述词汇矩阵模型,即在词的形式和意义之间建立起映射关系.每一个 Synset表示一个基本的词汇概念,并在这些概念之间建立了包括同义关系、反义关系、上下位关系、整体与部分关系、关系等多种语义关系.

WordNet是完全免费的资源,已经在英语语言处理研究和应用中广泛使用,几乎成了英语语言知识库的标准.

HowNet是一个以汉语和英语的词语所代表的概念为描述对象,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.作为一个知识系统,知网是一个网而不是树,它所着力要反映的是概念的共性和个性,知网描述了下列各种关系:上下位关系 (由概念的主要特征体现)、同义关系、反义关系、对义关系、部件-整体关系、属性-宿主关系.

WordNet也有一个树状的概念层次体系,知网似乎与之很相似,但实际上二者有着本质上的差别.在 WordNet中,概念是描写词义的最小单位,每一个概念都是语义层次体系 (网络结构)中的一个结点.而在知网中,每一个概念都是通过一组义原来表示,而概念本身并不充当义原层次体系中的结点,真正的结点是义原.

在《同义词词林》中,并没有去定义或者解释任何一个词语概念,没有体现概念之间的关系,只是把汉语词汇按照词义的远近关系分成大、中、小三类.其中大类 12个,中类 94个,小类 1428个.从分类的角度讲,《同义词词林》应该是一部较完善的分类辞典.如果把同义词词群作为一个概念解释,那么《同义词词林》在词义的理解上也会为使用者提供一个实质性的帮助.

一般来说,词典的编撰都是面向某个应用领域,比如说面向大众使用的,面向计算机领域的,面向化工领域的,等等.很难找到一部通用的词典大全,可以涵盖方方面面的所有应用领域.一部机读词典满足了一个领域的需要,在另一个领域就捉襟见肘了.尤其是基于分类辞典的方法,专业性更强,问题也更严重.另外,新的词语或者词语的新的用法不断出现,这也给已有词典词源的使用带来了新的挑战.

对于自然语言的研究到底需要什么样的词典资源?需要多大规模的知识库?值得进一步的讨论和分析.如何整合和挖掘现有的词典资源,如何开发和建设新的词典,有待进一步的研究和探索.

3.2 语料库建设

语料库是统计机器学习的知识源,是统计词义消歧系统的一个重要组成部分.没有一个好的语料库来支持,再优秀的统计学习算法也不可能获得出色的学习效果.

对词义消歧的研究来说,重要的是标注词义的语料规模的大小.因为有指导的词义消歧需要有高质量、大规模的词义标注语料的支持.如果没有这样的保证而采用较小规模的训练语料,将会产生严重的数据稀疏问题,虽然可以采用某些数学平滑方法作为补救措施,但最终的词义消歧效果还不是很理想.

通常,词义消歧使用的标注语料基本上靠人工完成.建设规模足够大、质量足够好的标注语料,从人工消耗和时间开销上是不现实的.所以需要寻找新的方法,有的学者提出标注语料的自动构造方法,有的学者建议采用平行语料库(ParallelCorpora),有的学者干脆放弃使用标注语料库而改用无指导的机器学习方法.

3.2.1 增强语料算法(Bootstrapping Corpora Algorithms) 对于未标注词义的语料(简称"无标注语料",对应的是"标注语料")在学习上存在不确定性,基于无标注语料的词义消歧方法称为无指导的词义消歧.如果通过某种方法把无标注语料构造成标注语料,然后又用该语料训练词义消歧模型,此方法是有指导的还是无指导的?应该说这种方法模糊了有指导和无指导的界限.

标注语料的自动生成为有指导的词义消歧提供了一个可行的解决方案.从小规模标注语料逐步递增,最后生成规模更大的语料,一般称此方法为增强语料算法,用到词义消歧上称为增强词义消歧(BootstrappingWSD).增强语料算法最早可以追溯到 1992年,Gale等人用无标注语料和增强算法实现了一个词义消歧系统[13].他们的调查十分有限,仅有6个测试样本,并且每个样本具有 2个词义,每个样本的实例数目也很有限.

后来 Yarowsky(1995)[14]、Mihalcea(2002)[15]又在同一个领域做了进一步的研究.这些研究都是为了解决语料规模的问题,并提供高质量的训练语料.他们为增强词义消歧提供了新的解决思路,其中 Yarowsky和 Mihalcea的研究在主导思想上有共同之处,即采用标注语料作为初始种子(initial seed),通过迭代计算来增大种子的数量,最终获得规模更大的标注语料.

Yarowsky从小规模的标注语料开始迭代,实现了一种决策表方法[14].初始种子是人工按照 Roget's分类词典标注的语料.该方法获得了很大的成功,测试结果达到了 95%的精度[14].

从另一个角度,Mihalcea在生成式算法(generation algorithm)—GenCor[16]基础上进行了增强处理[15].GenCor使用的种子包括来源于WordNet的单义词、来源于WordNet中多义词解释的词义标注实例,以及人工标注语料.初始种子用来作为 Query从 Web上获得更多的实例,把返回的实例及其上下文扩充到种子集合中.同时将返回实例上下文中的词语直接进行消歧.新增的实例和上下文作为新的 Query再到 Web上搜索更多的实例,如此反复进行.这是典型的迭代算法,经过不断的扩大生产,可以自动生成规模巨大的标注语料.词的选择上严格限制在名词性的合成词或者动词的内部论元(internalargument).Mihalcea采用基于实例学习的算法构造了有指导机器学习系统,并使用两种训练语料进行了对比研究.一种训练语料是 GenCor自动生成的语料,一种是人工标注的语料.实验证明了 Mihalcea方法是成功的,测试数据中有 6个样本的实验结果超过了人工语料的结果.

3.2.2 基于双语语料的方法 双语语料中在原文(源语言)和译文(目标语言)的相同语法单位之间的对应关系,称为双语对齐(BilingualAlignment),扩展到多语,叫做多语对齐(MultilingualAlignment).对齐语料亦称为平行语料.

双语语料的词对齐研究早期主要应用在机器翻译上,译文选择是词义消歧在机器翻译上的重要应用,研究人员自然会想到用双语语料来研究词义消歧.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一个可供选择的好方法.该方向的研究屡见报道,说明基于双语语料的WSD方法比较受欢迎.

1991年,Dagan指出两种语言包含的信息比一种语言多[17],他在 1994年又探讨了使用第二种语言来帮助词义消歧的方法[18].1997年 Resnik和 Yarowsky在一篇会议论文中正式推介基于双语语料的WSD方法[19].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进步,例如,Escudero等(2000)[20]、Ide等(2002[21])、Cong Li和 Hang Li(2002)[22]、Ng等(2003)[23]、Diab(1999[24],2002[25],

2003[26], 2004[27,28])和 Bhattacharya等(2004)[29]为双语语料在WSD研究上的应用起到了积极的推动作用.

Ng等人把语言数据协会(Linguistic Data Consortium, LDC)提供的汉英双语语料应用到了词义消歧开上[23]. LDC提供的汉英语料共6块, 规模很大, 总字节达到了280MB, 并且都进行了句子对齐处理. 为了能够使其在WSD上使用, 双语语料还需要进行词对齐处理. 目前词对齐的很多, 并且已经有了一些实用的工具在网上共享, 其中Och和Ney提供的词对齐软件GIZA++[30]比较知名. Ng先采用GIZA++将语料进行词对齐, 然后对目标译文词进行了人工选择. 目标译文对源语词起到词义标注的作用, 所以通过前面处理可以获得高质量的可用于WSD的双语平行语料. Ng选择Lee和Ng在2002提出的方法[31], 用Naïve Bayes模型构造词义分类器, 测试了SENSEVAL-2中的29个名词, 英文词义解释使用WordNet的标准定义. 平行语料的实验结果(P_1)与人工标注语料的结果(M_1)进行对比, 有两个词"nature"和"sense"的正确率比较高, P_1基本超过或接近M_1, 说明平行语料在机器学习模型的训练上是比较有希望的[33].

诚然, 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. 但是该方法的使用上也面临几个重要的问题需要认真对待.

(1)对源语言进行精确的消歧需要多大规模的平行语料?

(2)高质量的词对齐语料需要人工来完成, 如果建设大规模的平行语料, 既费时又费力, 与语料的人工词义标注在时间复杂度上不相上下. 如果采用机器自动对齐, 当前的词对齐算法是否有足够高的精度?

(3)如果采用当前最好的有指导WSD系统, 训练语料分别用平行语料和人工标注语料, 哪个正确率会更高一些?

上面这几个问题还没有得到很好的回答. 可以说, 平行语料是该方法的重中之重, 虽然已经有研究人员尝试从Web上获得平行语料[32], 但是大规模平行语料还是比较难于获得. 缺乏高质量的大规模平行语料, 影响了该方向的研究进展.

Cong Li和Hang Li把前文介绍的Yarowsky方法[14], 称为"单语增强(Monolingual Bootstrapping MB)", 他们把Yarowsky方法的基本思想从单语(monolingual)扩展到双语, 即从英语到汉语, 得到一种新的词语译文消歧(Word Translation Disambiguation)方法, 称之为"双语增强(Bilingual Bootstrapping BB)".

BB使用的初始种子既要有小规模英语标注语料, 也要有汉语标注语料, 两块语料可以不对齐, 但必须是同一领域来源. 为每个语言单独建立分类器, 对无标注语料进行消歧后, 将其添加到训练数据中, 如此反复. 两种语言在翻译上的对应关系可以用双语词典来确定.

Cong Li和Hang Li通过实验比较了BB和MB两种的实验效果, 证明了BB要优于MB, 同时也说明双语语料的信息量要比单语语料多一些. 另外, 他们采用的方法最大优点在于使用了非平行双语语料, 既省去了句子对齐和词对齐的中间步骤, 又发挥了双语语料的优势.

Diab在论文中介绍了无指导的词义消歧系统SALAAM[25], 该系统利用Al Onaizan等人的GIZA程序[33]自动生成token-level的对齐(token可以是词也可以是短语), 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(data acquisition)问题提供了多语言的解决框架. 2003年Diab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表现是当前最出色的[26]. 2004年该方法用来增强阿拉伯语词义消歧系统[27], 是在多语种扩展上的一个应用范例. 近期, Diab又考查了大规模粗糙标注数据(noisily annotated data)在解决词义消歧数据获取问题所起的作用[28].

Diab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[28]. 在对比实验上, Diab借鉴Mihalcea的方法, 设计一组用人工标注语料训练的机器学习模型作为参照[15]. Diab的对比实验采用了Cabezas建立有指导的词义消歧系统UMSST[34], 测试数据的选择与Ng一样, 即SENSEVAL-2中的29个名词[23]. 实验结果说明有9个词(占总数的31%)的正确率和有指导的方法相同或者极为接近, 并且超出当前其它增强算法11个百分点[28].

Bhattacharya等人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[25]表现得更好, 而概念模型又比语义模型强出很多.

3.3 特征提取

特征提取问题是机器学习和模式分类所面临的一个重要问题. 特征提取也可以叫做特征选择, 是指从已知一组特征集中按照某一准则选择出有很好的区分性的特征子集, 或按照某一准则对特征的分类性能进行排序, 用于分类器的优化设计[35].

词语的上下文(context)就是通常所说的语言环境, 能够对自然语言处理起到举足轻重的作用. 首先, 上下文是是知识获取的来源; 其次, 在自然语言处理问题解决过程中, 上下文提供解决问题所需的信息, 尤其是在语料库语言学中, 各种机器学习方法的引入使词语的上下文成为计算语言学知识获取和问题求解过程中最为重要的资源.

对于词义消歧, 也是需要从上下文中获得词义知识. 语言学家Firth曾说过"观其伴, 而知其意[36]", 指的是一个词的词义只能通过与之相伴出现的搭配词才能加以判

断和识别. 区分词的词义需要调查词语的搭配关系和用法模式. 对人和机器而言, 利用上下文信息都是解决歧义现象的根本出发点. 一个篇章一个词义 (one sense per discourse), 歧义词的词义受到给定文本的严格限制[37]; 一个搭配一个词义 (one sense per collocation), 上下文词语同歧义词的距离、次序和语法关系对歧义词词义判断提供了强有力的线索[38].

3.3.1 上下文的有效范围 诚然, 歧义词上下文能够提供词义的约束信息, 词语间的相互作用与上下文中的位置和物理距离有直接的关系, 一般来说, 距离较远的词语间搭配关系比较弱, 对词义判断的贡献比较小. 那么, 上下文的有效作用距离有多大? 解决这个问题, 首先要保证限定范围内的特征候选词可以为词义判断提供足够的信息, 其次要保证产生的噪声足够的小. 此外还要兼顾系统运行的时间和空间上的效率. 早在 1983年 Martin认为核心词左右各取 5个词可以提供 95% 的词义搭配信息[39]. Yarowsky在实验中发现, 上下文窗口从 ±5扩大到 ±50后, 词义消歧的正确率由 86% 提高到 90%[38]. 1994年 Hughes指出上下文窗口取 ±2可以获得最佳的结果, 如果窗口再增大, 不会使有效信息有明显的增加, 反而会带来更多的噪声, 以及更大的计算开销[40]. 2001年, 鲁松通过计算发现: 基于《人民日报》统计汉语核心词最近距离左 8个位置和右 9个位置的上下文范围和基于《格林童话选》统计英文核心词最近距离左 16个位置和右 13个位置的上下文范围, 可为确定核心词提供 85% 的信息量[41]. 上下文位置信息量的信息增益确定, 从量上验证了上下文词语对核心词语呈现规律性分布, 其提出的方法对上下文语境有效范围的确定实现了很有价值的定量化描述, 克服了前人主观描述的不足, 为上下文提供了一种具有统计意义的量化解释.

3.3.2 常用的特征提取方法 确定了上下文窗口的大小后, 窗口内的词都可以作为候选特征词. 简单的作法是把窗口看成词袋 (bag of words), 忽略窗口中词语间的线性结构和距离的远近, 认为每个词语对歧义词都有约束作用, 并且作用是相互独立的. 这种方法将窗口中的所有词语都看作特征词, 应该说, 词袋的作法放弃了很多有用的信息, 没有充分利用上下文为语言模型提供更多的知识, 不利于改善系统的性能. 上下文窗口中词语间的线性结构往往蕴含着大量的知识, 而这些知识大多是对词义识别有所帮助的. 如何在有限的上下文中发掘更大的知识是非常有价值的研究课题, 也是词义消歧研究的一个关键问题.

上下文窗口中的词语因搭配关系和距离造成对歧义词的作用力强弱不均, 有必要按照作用力的强弱对候选特征词加以区分, 以便选择信息更丰富的词语作为特征词. 在统计词义消歧的研究中, 很容易想到用词语出现的频度作为特征词选择的一个标准. 例如, 选择那些与歧义词共现频率大于某一阈值的候选词作为特征词. 这样做的先决条件是基于一个假设: 与歧义词构成搭配关系的词语总是与歧义词相伴出现, 也就是共现的频率很高. 需要注意的是很多虚词或者停用词的频率也很高, 有时甚至超过特征词的频率. 在实践中这些词应该先过滤掉, 免得造成噪声干扰.

为了进一步的表征上下文词语与歧义词之间关系, 可以计算上下文词与歧义词的互信息 (Mutual Information, MI), MI能够表示上下文词与歧义语义上的相对语义距离. 互信息的计算公式如下[42]:

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

$P(w_1)$、$P(w_2)$和 $P(w_1, w_2)$分别是词语在语料库中出现的概率和共现概率. 根据互信息的计算结果, 选择互信息高于某个阈值的上下文词作为特征词语.

一个特征的优劣不仅要用其分类的性能来衡量, 有时特征提取的代价 (cost)也是需要认真考虑的因素. 1999年, 王亚东等人从实际应用角度考虑, 提出一种基于信息增益和特征提取代价综合评价函数的特征选择标准[43]. 将特征的分类性能与特征的提取代价统一考虑, 在识别过程中特征的选择与提取同时进行的方法, 并给出了基于决策的启发式算法. 王等人将该算法应用于手写汉字识别系统的特征选择问题, 实验结果表明该算法在保证识别精度的同时, 大大减少了特征提取的时间消耗, 提高了识别速度[43].

在某些语言模型中由专用模块负责特征选择, 如最大熵方法、主成分分析、粗糙集、遗传算法、神经网络等. 下面介绍几种典型的常用方法.

3.3.3 最大熵方法 采用信息论中熵的概念, 也可以评价上下文词与歧义词之间的搭配关系, 从而作为特征选择的判定标准. 如果依据最大熵原理计算各参数的概率分布, 可以实现多种有效的特征选择算法. 李涓子从上下文特征词与歧义词的搭配关系出发, 提出了利用互信息和 z-测试方法相结合的特征选择算法, 改进了基于 IIS的最大熵方法[44]. 虽然互信息可以作为描述搭配强度的物理量, 但如果特征选择直接确定选择互信息大于某一阈值的上下文信息为特征时, 对不同互信息的分布设定的阈值也不应该相同, 而这样做, 算法难以实现. 而 z测试可以将互信息的分布变换为标准的正态分布, 不论互信息如何分布, 都可以使用统一的阈值进行求解. 李采用 Kullback-Leibler距离 (交叉熵) 来测定特征所确定模型的质量, 设 p_c是由训练语料确定的概率模型, p_s为由特征集确定的模型, 则 Kullback-Leibler距离定义为:

$$D(p_c \| p_s) = \sum_{x,y} p_c(x, y) \log \frac{P_c \log(y|x)}{P_s \log(y|x)},$$

最终要找的模型 p_{fs}为: $P_s = \min D(p_c \| p_f)$. 李证明其所采用的特征选择算法所需的运算量小于 Pietra等人提出的特征选择算法所需的运算量[45]. 同时, 李将其建立的 ME模型应用于汉语词义消歧, 测试了三个歧义词, 结果正确率

较高.

2003年, Wong Fingwai和 Yang Yongsheng也用 ME方法实现了一种汉语词义消歧的方法[46]. 该方法是受到了基于 ME模型的词性标注技术[47]的启发, 采用 HowNet作为机读词典,获得了 89.39%的正确率. 出色的成绩也得益于最大熵模块在特征选择上的卓越性能.

Armando Suárez和 Manuel Palomar为英文词义消歧建立了条件最大熵模型(ConditionalME),分析了特征提取方法对消歧性能的影响[48]. Suárez和 Palomar改变了所有歧义词使用同一种信息的传统做法,替代的办法是为每个歧义词提供不同的特征集,并采用基于知识的方法(Knowledge-based Method)和基于语料的方法为歧义词分别建立 ME分类器. 实验证明, Suárez和 Palomar的方法更有效[48].

3.3.4 主成分分析　主成分分析(Principle ComponentAnalysis, PCA)是很常用的子空间分析技术和降维方法,主要用于图像分类等模式识别领域[49]. PCA是一种统计方法,其在分析中主要利用了所分析对象的内在的统计特征.因而通过 PCA变换使得提取的分类特征更加明显,使其能量向某些相对分量集中,增强随机矢量总体的确定性,分类效果也可更好[49].

2004年, DdKaiWu等人开发了一种基于非线性变换的 KernelPCA方法,并用于英文词义消歧[50]. 实验中, Wu等人参照文献中提到的当前最优秀的有指导词义消歧算法,贝叶斯模型(NB)、最大熵模型(ME)和支持向量机(SVM)建立了三个词义消歧的 baseline系统作为实验的对比,其中 SVM也是一种 kernel方法,分类效果要优于 NB和 ME.实验结果参见表 1.

表 1　四种词义消歧系统的实验结果[50]

WSD Model	NB	ME	SVM	KernelPCA
Accuracy	63.3%	63.8%	65.2%	65.8%

四个 WSD系统中, NB没有特征选择的过程,仅仅单纯的控制上下文的窗口; ME如前文所述是一种经典的特征提取算法,有特征选择和参数估计过程;支持向量机也有一个特征选择的算法过程. 实验证明, Wu等人使用的 KernelPCA方法在特征选择上更为有效,分类结果最为出色[50].

3.3.5 粗糙集　粗糙集(Rough Sets)是由波兰数学家 Pawlak提出来的一种分析数学理论,该理论在分类意义上定义了模糊性和不确定性的关系[49]. 采用粗糙集理论可以对待分类的特征进行适当的约简,减低特征空间的维数.显然,降维是一个典型的特征提取过程,而粗糙集是应用很广泛的特征提取技术.

粗糙集理论的要点是将分类与知识联系在一起,并使用等价关系来形式化地表示分类.利用定义在数据集合 U上的等价关系 R对 U的划分作为知识,而对知识不确定程度的测量则是对被分析数据整体的处理之后自然获得的.这样,粗糙集理论无需对知识或数据的局部给予主观评价,即粗糙集理论对不确定性的描述相对客观.粗糙集理论为数据分类提供了一个有效的处理工具.

陈清才等人(2001年)为歧义词建立 N维的词语向量空间[51]. 如果词空间维数过高,计算量难免过大,为了在保持语言模型词义辨别能力的前提下尽可能的降低词向量空间的维数,陈引入粗糙集技术对此空间进行约简.词空间的坐标词,即上下文的候选特征词,属性简约的过程就是特征词提取的过程.实验证明,由经过粗糙集进行属性约简的词空间建立的语言模型具有更好的词义辨识能力[51].

如何选择主分量用作分类特征一直是主成分分析(PCA)方法的研究重点.虽然 PCA方法能用 K-L变换优化,但不能适应所有分类目标. 2004年,宋余庆等人使用粗糙集将待分类特征属性进行适当约简,提取最小核(Core)用于分类,这样可使提取到的特征向量是最佳分类向量,从而弥补了 PCA的不足[52].

3.4 数据稀疏

数据稀疏一直是统计自然语言处理技术无法回避的问题,如何解决数据稀疏问题也是衡量自然语言处理系统的一个重要方面.

产生数据稀疏问题的主要原因有两个,一个是特征维数,一个是训练语料的规模,其中语料的规模是根本原因.特征维数是指特征的数量,如果特征维数高,就需要统计更多的实例.语料的实例覆盖度一般与语料规模成正比关系,对高维特征的统计需要规模更大语料支持,否则就会出现严重数据稀疏现象.对于语料来说,规模越大包含的信息就会越多,如果在语料中训练语言模型的参数,规模小就会有某些需要统计的信息统计不到,从而出现数据稀疏问题,语料规模越小,稀疏问题就会越严重.

解决数据稀疏问题应该针对产生的原因选择合适方法.对于特征维数过高的问题,主要考虑如何降维,减少特征的数量.在不影响模型效果的前提下,采用有效的降维方法是克服数据稀疏问题的主要出发点,另外在条件允许的情况下适当增加语料的规模也是收效显著的解决方案.对于训练语料规模过小造成的数据稀疏问题,解决的办法也只有增加语料的规模了.在有指导的学习中,需要用到标注语料,而这样语料的规模往往很小,所以数据稀疏问题也很严重.前文介绍了几种自动生成标注语料的办法,这些方法只在某种程度上改善了语料规模的现状.一般来说,无指导的机器学习方法因为不需要对训练语料进行标注,语料规模可以根据需要任意扩大,所以数据稀疏问题不是该类方法的主要矛盾,与有指导的方法相比之下,更受欢迎.

3.4.1 数据平滑　数据平滑(Data Smoothing)技术是解决数据稀疏问题最常用的数学方法.平滑技术可以保证语言模型中的任何概率参数不为零,又可以将低概率(包括零概率)调高,高概率调低,使模型参数概率分布趋向更加均匀.在语言建模上,数据平滑技术十分重要,因为模型的

训练语料不可能无限的增大,数据稀疏问题在所难免,只不过在有的问题上严重,在有的问题上不严重.例如,前文提到的特征维数高,数据稀疏问题就严重,维数低就不严重;有指导的方法严重,无指导的方法不严重.

实用的数据平滑技术很多,如加法平滑(Additive Smoothing)、Good-Turing估计、Jelinek-Mercer平滑、Katz平滑、Church-Gale平滑、Average-Count平滑等.

3.4.2 无指导的机器学习 一般来说,统计上的零概率事件是不可避免的,而平滑技术借助数学方法避免了零概率的发生,使语言模型的计算过程得以顺利地进行,可见平滑技术是建立语言模型所必需的.有的平滑技术甚至在某种程度上还改善了机器学习的效果,提高语言模型的性能,但是平滑技术不可能大幅度提高模型的知识获取能力.通常情况下,训练语料规模的增大给语言模型带来的改善要比平滑技术大得多.毕竟训练语料是学习的对象,是知识获取的源泉,如果训练语料很小,再优秀的平滑技术也不会获得好的学习效果.为此,研究人员还得从其它角度入手,消除或削弱小概率事件对语言模型的消极影响.

本文3.2介绍的有关标注语料的获取技术,从有指导机器学习的训练语料规模入手,探索克服数据稀疏问题的方法.无论是从单语自动生成标注语料,还是以平行语料作为训练源,都使得方法本身带有了无指导的特征,所以从事该研究的人员称使用了这些技术的词义消歧为无指导的词义消歧.其中无指导的方法以平行语料为多.双语词义消歧除了前文提到的,还有 Paul Buitelaar 他在 2001年发表了有关德语的无指导词义消歧研究论文[53],随后2002,2003年又配合他人共同研究了词义消歧和词义标注问题,特别是专门领域(Specific Domain)的无指导方法和词义消歧[54].在国内,李涓子和黄昌宁提出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[55].但是,这些针对训练语料的技术一般要用到标注语料作为初始种子,或者用目标语言中对应的词语译文,还没有彻底摆脱人工标注语料和人工选择译文的困扰,仍然属于在有指导的框架下解决数据稀疏问题.

黄昌宁和李涓子依照《同义词词林》的语义分类体系,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,构成词义的分类器,实现一种代价最小的无指导学习算法,并且实验结果令人满意[56].

鲁松用词向量概念为歧义词一一建立词向量,并参照 Schütze的词义聚类方法[2]建立了向量空间模型(VSM),结合《同义词词林》实现了一种无指导的汉语词义消歧方法[57].鲁通过词矩阵的定义和计算上下文中词语在刻画该词语时的重要性(计算词语权重),可以在词向量空间对词语进行精确定位,这与仅依靠词语共现频率的形式化方法相比显得更有效.实验中选择 10个常用的歧义词作为消歧对象,并为每个歧义词人为指定若干义项词语,实验正确率最高的达到 92.13%,最低的为 72.83%,平均值为 83.13%[57].由于使用了义项词语,模型的训练不必使用标注语料,也是一种典型的无指导方法.

3.5 机器学习方法

统计词义消歧需要机器学习方法在相关知识源(语料、词典、Ontology等)获取词义判断的知识,可见在此机器学习方法是必不可少的.词义消歧问题与词性标注问题一样可以看作分类问题,许多用于分类的学习算法都可以用于词义消歧.前文提到的多种统计机器学习方法,如决策树、贝叶斯模型、神经网络、支持向量机等都是词义消歧研究中常用到的机器学习方法.同一种分类算法对于不同的分类问题,效果往往大不相同,特别是不同的分类算法的分类效果和执行效率差别更大,这是值得关注的问题.

机器学习方法如此的重要,那么哪种分类算法更适合于词义消歧?然而,文献中并没有较为直接的报道.似乎很少有研究人员关注如何选择机器学习算法的问题.自从 Brown把统计学习方法引入到自然语言处理领域以来[9],统计自然语言处理的研究已经有了十几年的历史,多种机器学习方法在自然语言处理领域的各个问题上可谓是遍地开花,得到了广泛采用.张刚等人[58,59]做过一些实验,考查隐马尔科夫模型与贝叶斯模型,神经网络与单纯贝叶斯网络在汉语词义消歧上的应用情况,对比分析几个典型机器学习方法消歧能力.当然,同样的方法应用于不同的问题域,就会有不同的情况;同一个问题采用不同的建模方法,就会有不同的结果,……不同的实验对象、不同的实验手段、不同的实验结果,林林总总,很难说孰是孰非、谁优谁劣,即便是同一种问题,同一种方法,又因为采用的训练数据和测试数据的不同导致了实验结果的不可比性.

诚然,不同的机器学习方法在效果上存在着差别,但是每种机器方法在采用了合适的建模策略,如适当的训练语料的规模,适当的特征提取方法,适当的数据平滑方法等等,都会获得较好的词义消歧效果.

4 未来研究的重点

如前所述,词义消歧的研究论文大量涌现,研究的热潮可谓是如火如荼.然而,就词义消歧问题本身来讲,统计词义消歧的研究目前还没有取得划时代的进展或重大突破.研究大都分散在词义消歧的各个点上,选择的视角和方法都难免带有一定的局限性.

统计词义消歧需要借助统计的手段在语料库或者知识库中发现和获得词义信息或知识,这里有两个重要的入手点,即统计方法(机器学习)和知识源(语料库或词典),既要有好的学习手段,又要有好的知识源.对于词义消歧,要么看作分类问题,要么看作聚类问题,但不管是分类还是聚类,都已经有很多成熟并且实用的学习算法可供选择.可以说,确定哪种统计学习算法更有效,虽然在研究上也十分重要,但是它并不是词义消歧研究上最紧要的问题.而围绕词义消歧知识源的研究才是最紧要的也是最重

要的问题,也是目前国内外相关研究的热点.

正如前文提到的那样,有指导的学习方法效果好,但是需要人工标注训练语料.研究中发现消歧的正确率与训练语料的规模成正比[59],然而很难获得规模足够大的标注语料,尤其是包含歧义词所有词义实例的语料更是难上加难.无指导的学习算法,不需要标注语料,但是很难达到有指导方法的精度.矛盾的焦点在于知识源上,因为知识源的不完备、不充分导致了难以克服的数据稀疏问题.选择有指导的方法,为了缓解和克服数据稀疏问题,就要考虑如何获得规模足够大,质量足够好的标注语料库;为了摆脱数据稀疏问题的困扰,选择无指导的方法,又要考虑如何在未标注词义的知识源中发现可用于词义判断的知识.数据稀疏就是知识源带来的矛盾,一个统计词义消歧无法回避的问题.

前文描述的语料增强算法和基于平行语料的方法在自动获得标注语料方面给了我们很好的启发,为解决数据稀疏问题提供了一个新的思路.但是,目前该方向上的相关研究工作还没有大面积展开,尤其是在国内尚属空白.需要进一步加强该方面的研究,以期取得实质性的进展.

探索解决数据稀疏问题的方法,只是统计词义消歧研究的一个方面,除此之外还要研究获取词义知识的方法.研究词义知识的获取方法,就是研究如何在有限的资源条件下,获得更多有利于词义判断的信息或者知识.这里包括无指导学习中的知识获取问题,也包括特征提取问题.解决该问题也需要在知识源上下功夫,并且还更多地涉及到了语言学的知识,这就又回到了词义消歧的本原——语言学问题.目前这方面的研究还不多,沿着该方向深入地研究下去,将对最终解决词义消歧问题产生积极的推动作用.

相对国外的研究,国内词义消歧起步比较晚.比较早的有刘小虎[60]和李涓子[61],后来研究逐渐多起来.方法上也从基于规则的方法迅速过渡到以概率统计为主的新方法上,如荀恩东[62]、朱靖波[63]、鲁松[41][57]和杨尔弘[64].可以说,对于汉语的词义消歧研究,需要做的工作更多.其中建设一部像 WordNet那样广为接受和采用的统一词典资源,构造类似 SENSEVAL[34]的标准测试数据,是亟待完成的任务.

5　结论

词义消歧研究的历史还很短,特别是统计词义消歧刚刚走过十几年的发展历程,词义消歧还是一个比较年轻的研究方向.在短短的时间里,就涌现出大量的词义消歧研究报告和学术论文,尤其是近五年,论文的数量和质量都有了很大的提高,标志着词义消歧的研究进入了快速的发展阶段.

统计学方法在词义消歧中的应用迅速普及,出现大量以统计机器学习为核心技术的词义消歧方法.

可参考的文献中,绝大部分是讨论英文词义消歧问题,其它语种的研究在数量上十分有限,反映了语种之间在研究水平上存在的差距.

国内外的研究表明,统计词义消歧技术还局限在实验室水平,距离实用化阶段还有很长的一段路要走.国内的研究需要吸收国际上先进的研究思想和经验,大胆创新,迅速提高研究水准.

参考文献:

[1] Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art [J]. In Computational Linguistics, 1998, 24(1): 1-40.

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

[3] C D Manning, H Schütze. Foundations of statistical natural language processing[M]. The MIT Press, Cambridge, Massachusetts, London, England, 1999. 229-260.

[4] Michael E Lesk. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone[A]. In Proceedings of the SIGDOC Conference[C]. Association for Computing Machinery, New York, 1986. 24-26.

[5] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora[A]. In COLING 14[C]. Nantes, 1992. 545-460.

[6] Eneko Agirre, Rigau German. A proposal for word sense disambiguation using conceptual distance[A]. Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing[C]. Bulgaria, 1995.

[7] S D Richardson, W B Dolan, et al. Overcoming the customization bottleneck using example-based MT[A]. In Proceedings, Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter[C]. Association for Computational Linguistics Toulouse, France, 2001. 9-16.

[8] H T Ng, H B Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach [A]. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics[C]. Santa Cruz, California, 1996. 40-47.

[9] Peter F Brown, Stephen A. Della Pietra, et al. Word-sense disambiguation using statistical methods[A]. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics[C]. Berkeley. 1991. 264-270.

[10] 梅家驹, 竺一鸣, 等. 同义词词林[M]. 上海: 上海辞书出版社, 1996.

[11] 董振东. HowNet[DB/OL]. http://www.keenage.com. 2002.

[12] George A. Miller. (Ed.) WordNet: An on-line lexical database[J]. International Journal of Lexicography,

1990, 3(4): 235-312.

[13] W A Gale, K W Church, D Yarowsky. Using bilingual materials to develop word sense disambiguation methods [A]. Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation[C]. Montréal, Canada, 1992. 101-112.

[14] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods[A]. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics[C]. Cambridge, MA. 1995. 189-196.

[15] R Mihalcea. Bootstrapping large sense tagged corpora [A]. Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC) [C]. Las Palmas, Canary Islands, Spain, 2002.

[16] R Mihalcea, D I Moldovan. An automatic method for generating sense tagged corpora[A]. In Proceedings of AAAI-99[C]. Orlando, FL, July 1999. 461-466.

[17] I Dagan, A Itai, Schwall U. Two languages are more informative than one[A]. In Proceedings of the 29th Annual Meeting of the ACL [C]. Berkeley, California, 1991. 130-137.

[18] Ido Dagan, Alon Itai. Word sense disambiguation using a second language monolingual corpus[J]. Computational Linguistics, 1994, 20(4): 56-596.

[19] Philip Resnik, David Yarowsky. A perspective on word sense disambiguation methods and their evaluation[A]. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How [C]. Washington, 1997. 79-86.

[20] G Escudero, L Marquez, G Rigau. Boosting applied to word sense disambiguation[A]. In Proceedings of the 12th European Conference on Machine Learning[C]. Barcelona, 2000. 129-141.

[21] N. Ide, T. Erjavec, and D. Tufis. Sense discrimination with parallel corpora[A]. In Proceedings of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions[C]. Philadelphia, PA. 2002. 54-60.

[22] Cong Li, Hang Li. Word translation disambiguation using bilingual bootstrapping[A]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C]. Philadelphia, PA. 2002. 343-351.

[23] H Tou Ng, B Wang, et al. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study[A]. In Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics[C]. Sapporo, Japan. 2003: 455-462.

[24] Philip Resnik, Olsen M. Broman, Mona Diab. Creating a parallel corpus from the book of 2000 tongues[J]. Computers and the Humanities, 1999, 33: 1-2. 129-153.

[25] Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora[A]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C]. Philadelphia, PA. 2002. 255-262.

[26] Mona Diab. Word Sense Disambiguation Within a Multilin-gual Framework[D]. PhD Thesis University of Maryland College Park, USA, 2003.

[27] Mona Diab. An unsupervised approach for bootstrapping arabic word sense tagging[A]. Proceedings of Arabic Based Script Languages, COLING 2004[C]. Geneva, Switzerland, 2004.

[28] Mona Diab. Relieving the data acquisition bottleneck in word sense disambiguation[A]. In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics[C]. Barcelona, Spain. 2004. 303-310.

[29] I Bhattacharya, L Getoor, Y Bengio. Unsupervised sense disambiguation using bilingual probabilistic models [A]. In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics[C]. Barcelona, Spain. 2004. 287-294.

[30] Franz Josef Och, Hermann Ney. Improved statistical alignment models[A]. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics[C]. Hong Kong, China. 2000. 440-447.

[31] Yoong Keok Lee, Hwee Tou Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation[A]. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing[C]. Philadelphia, PA, USA. July 6-7, 2002. 41-48.

[32] Philip Resnik. Mining the Web for bilingual text[A]. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics[C]. College Park, Maryland, USA, 20-26 June, 1999. 527-534.

[33] A Al-Onaizan, J Y. Curin, et al. Statistical Machine Translation, Final Report, JHU workshop. http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps[DB/OL]. 1999.

[34] Clara Cabezas, Philip Resnik, and Jessica Stevens. Supervised Sense Tagging using Support Vector Machines [A]. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) [C]. Toulouse, France, 2002.

[35] 陈彬, 洪家荣, 等. 最优特征子集选择问题[J]. 计算机学报. 1997, 20(2): 133-138.

[36] J R Firth. A Synopsis of Linguistic Theory 1930-1955. In Studies on Linguistic Analysis [M]. London: Blackwell.

1957. 101- 126.
[37] W Gale, K Church, D Yarowsky. One sense per discourse[A]. In Proceedings of DARPA Speech and Natural Language Workshop [C]. Morgan Kaufmann, 1992. 233- 237.
[38] David Yarowsky. One sense per collocation[A]. In ARPA Human Language Technology Workshop[C]. Princeton, NJ, 1993. 266- 271.
[39] W Martin, et al. On the processing of text corpus[A]. In Hartmann Red. Lexicography: Principles and Practice [C]. New York, 1983. 56- 64.
[40] John Hughes. Automatically acquiring a classification of words [D]. PhD dissertation. Paris: University of Leeds, 1994.
[41] 鲁松, 白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报, 2001, 24(7): 742- 747.
[42] K. Church, P. Hanks. Word association norms, mutual information, and lexicography [J]. Computational Linguistics, 1990, 16(1): 22- 29.
[43] 王亚东, 郭茂祖等. 一种基于信息增益与费用评价函数的特征选择准则[J]. 计算机研究与发展. 1999, 36(7): 788- 793.
[44] Dekai Wu, Weifeng Su, Marine Carpuat. A kernel PCA method for superior word sense disambiguation[A]. In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics [C]. Barcelona, Spain. 2004. 637- 644.
[45] 陈清才, 王晓龙. 一种基于词矢量的汉语语义量化模型[J]. 计算机研究与发展, 2001, 38(2): 208- 212.
[46] Wong Pingwai, Yang Yongsheng. A maximum entropy approach to HowNet based Chinese word sense disambiguation[DB/OL]. http://www.cs.ust.hk/~hltc/semanet02/pdf/wong.pdf. 2003.
[47] 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[J]. 软件学报, 1999, 10(3): 258- 263.
[48] A. L. Berger, Pietra S. Della, et al. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 40- 72.
[49] Adwait Ratnaparkhi. A maximum entropy approach for Part of Speech Tagging[A]. In Proceedings of the First Empirical Methods in Natural Language Processing Conference[C]. Philadelphia, USA, 1996. 133- 141.
[50] M. Saiz-Noeda, A. Suárez, M. Palomar. Semantic pattern learning through maximum entropy-based WSD technique[A]. In Proceedings of COLING-2002[C]. Taipei, 2002.
[51] R. O. Duda, P. E. Hart, et al. Pattern classification and scene analysis[M]. New York: John Wiley & Sons, Inc, 1998.
[52] 宋余庆, 罗永刚, 等. 应用主分量分析与粗糙集处理的特征提取[J]. 计算机工程与应用, 2004, 22: 48- 50.
[53] Paul Buitelaar, Jan Alexandersson, et al. An unsupervised semantic tagger applied to german[A]. In Proceedings of Recent Advances in NLP (RANLP) [C]. Tzigov Chark, Bulgaria, September 2001. 5- 7.
[54] Diana Steffen, Bogdan Sacaleanu, Paul Buitelaar. Domain specific sense disambiguation with unsupervised methods[A]. In: Kunze, C. & Lemnitzer, L. & A. Wagner. Anwendungen des deutschen Wortnetzes in Theorie und Praxis. Tagungsband des 1. GermaNet-Workshops des GLDV-AK Lexikografie[C]. Tübingen, Germany, October 9th/10th, 2003.
[55] 李涓子, 黄昌宁. 基于转换的无指导词义标注方法[J]. 清华大学学报(自然科学版), 1999, 39(7): 229- 234.
[56] 黄昌宁, 李涓子. 词义排歧的一种语言模型[J]. 语言文字应用, 2000, 3: 85- 90.
[57] 鲁松, 白硕, 等. 基于向量空间模型中义项词语的无导词义消歧[J]. 软件学报, 2002, 13(6): 1082- 1089.
[58] 张刚, 刘挺等. 隐马尔可夫模型和 HowNet 在汉语词义标注中的应用[J]. 计算机应用研究, 2004, 10(增刊): 67- 69.
[59] 卢志茂, 刘挺, 等. 神经网络和贝叶斯网络在汉语词义消歧上的应用对比[J]. 高技术通讯, 2004, 8: 15- 19.
[60] 刘小虎. 英汉机器翻译中词义消歧方法研究[D]. 博士学位论文. 哈尔滨: 哈尔滨工业大学, 1998.
[61] 李涓子. 汉语词义消歧方法研究[D]. 博士学位论文. 北京: 清华大学, 1999.
[62] 荀恩东, 李生, 等. 基于汉语二元同现的统计词义消歧方法研究[J]. 高技术通讯, 1998, 10: 21- 25.
[63] 朱靖波, 李绪, 等. 基于对数模型的词义自动消歧[J]. 软件学报, 2001, 12(9): 1405- 1412.
[64] 杨尔弘, 张国清, 等. 基于义原同现频率的汉语词义排歧方法[J]. 计算机研究与发展, 2001, 38(7): 834- 837.

作者简介:

卢志茂　男, 1972年生于黑龙江宾县, 哈尔滨工业大学计算机系博士生, 副教授, 1994年毕业于山东大学, 获理学学士学位, 1996年毕业于哈尔滨工业大学, 获工学硕士学位, 主要研究方向: 中文信息处理, 词义消歧, 知识获取.
E-mail: lzm@ ir.hit.edu.cn.