文章编号: 1003-0077(2008)03-0016-08

词义标注语料库建设综述

金 澎,吴云芳,俞士汶

(北京大学 计算语言学研究所,北京 100871)

摘 要: 词义消歧的关键问题是缺少大规模、高质量的词义标注语料库。本文分别从语料选取、词典选择、标注规模和标注质量等方面介绍已经建成的较有影响的若干词义标注语料库。在自动构建词义标注语料库的方法中,本文集中介绍 bootstrapping 策略在语料库建设方面的应用以及利用双语对齐语料库开展的相关研究。最后,针对词义标注语料库建设存在的问题提出自己的分析和思考。

关键词: 计算机应用;中文信息处理;词义消歧;词义标注语料库;平行语料库;bootstrapping

中图分类号: TP391 **文献标识码**: A

Survey of Word Sense Annotated Corpus Construction

JIN Peng, WU Yun-fang, YU Shi-wen

(Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: The bottleneck of word sense disambiguation (WSD) is lack of large scale, high-quality word sense annotated corpus. In this paper, several word sense annotated corpus are introduced in the aspects of corpus coverage, dictionary, tokens, word types and the inter annotator agreement, involving English, Chinese and Japanese. As for the auto and semi-auto construction methods, this papers focuses on bootstrapping methods and word-aligned parallel corpus based approaches. And finally, some issues in the word sense annotated corpus construction are pointed and possible solutions are given.

Key words: computer application; Chinese information processing; word sense disambiguation; word sense annotated corpus; parallel corpus; bootstrapping

1 概述

词义消歧(Word Sense Disambiguation, WSD)长期以来一直是自然语言处理的热点难题,在机器翻译[1]、信息检索[2,3]等领域均有重要的应用价值。而词义标注语料库的建设对 WSD 研究有着重要的意义: Ng 指出,WSD 的中心任务是建设一个大规模的词义标注语料库来训练有指导的机器学习模型[4]。Veronis 认为,没有大规模的词义标注语料库,WSD 研究不会有本质的进步[5]。

词义标注语料库是指,根据某个词典对多义词各个义项的定义,在真实语料上标注多义词的正确义项。理想中的词义标注语料库应该具有规模大、覆盖广和准确度高等特点。语料的规模是指已经标注所有多义词的出现总次数(token),所选语料库本身的规模也有一定的参考价值。语料的覆盖是指标注的单词词形(word type)的个数,也即词典中列举的多义词被标注的比例或个数。标注的质量通常用标注一致程度(Inter Annotator Agreement, IAA)①来衡

收稿日期: 2007-07-10 **定稿日期**: 2008-04-09

基金项目: 国家 973 计划资助项目(2004CB318102);国家自然科学基金资助项目(60703063);国家 863 计划资助项目(2007AA01Z198)

作者简介: 金澎(1977—),男,博士生,主要研究方向为计算语言学、词义消歧;吴云芳(1973—),女,博士,主要研究方向为计算语言学、语料库语言学;俞士汶(1938—),男,教授,博导,主要研究方向为计算语言学。

① 也称作 ITA(Inter Tagger Agreement)。

量。IAA 的简单计算如下:

$$IAA \equiv A/N \quad (1)$$

其中 N 是该词已标注的总次数; A 是各个标注者(通常是两个)相互认同的次数。这样计算的缺点是没有考虑到不同标注者偶然一致的情况。根据 Kappa 统计量来计算的 k 值定义如下[6]:

$$k = \frac{p_a - p_e}{1 - p_e} \quad (2)$$

$$p_e = \sum_{j=1}^{M} \left(\frac{C_j/2}{N} \right)^2 \quad (3)$$

$$p_a = IAA \quad (4)$$

其中 M 是目标词 w 的义项个数; C_j 是两个标注者标注为义项 j 的次数之和。通常认为 k 值超过 80 %就是高质量的标注[7]。

另外,词典的选择也是衡量词义标注语料库质量的一个重要指标。本文将从词义标注语料库建设的时间、机构、词典、语料库来源、标注方法、标注规模和质量等方面介绍目前已建成和正在建设的词义标注语料库。

2　人工构建的词义标注语料库

采用人工方法进行大规模词义标注语料库建设是目前通行的方法。本部分重点介绍英文和中文的词义标注语料库,对其他语种仅做简单介绍。

2.1　英语词义标注语料库

2.1.1　Semcor 语料库

该语料库由普林斯顿大学于 1993 年由 Miller 负责完成[8]。所用语义标注体系是 WordNet1.6。而 WordNet 也正是由其负责完成的。在 WordNet 中,用同义词集合(Synset)来表示概念。一个多义词,将在多个不同的 Synset 中出现。根据 WordNet 对义项的区分在完成词性标注后的 Brown 语料库上进行标注。共标注词次(token)超过 200 000 个。分布于 Brown 语料库中的 352 个文件,其中 186 个文件(共 359 732 词次)的所有实词(名词、动词、形容词和副词)全部被标注(192 639 词次)。另外的 166 个文件(316 814 词次),只标注了其中的动词(41 497 词次)。该语料库可以免费下载,并提供了相应的查询工具,但是并未见到关于 IAA 的报告。

该语料库是目前最大的英语词义标注语料库。尽管如此,Miller 认为该语料库规模太小,仍不足以据此设计一个健壮的、高准确率的词义消歧系统。

在 Semcor 上开展的研究很多,几乎所有的针对所有词(all-words)的英文 WSD 研究都会基于该语料进行[9～13]。

2.1.2　DSO 语料库

词义标注(Defence Science Organisation,DSO)语料库由新加坡国立大学于 1996 年由 Ng 负责完成[14]。所用词典是 WordNet 1.5,语料来自 100 万词 Brown 语料库和 250 万词华尔街时报(WSJ)。由该大学 12 个语言学专业的本科生,用一年时间标注完成。覆盖英语中最常见且歧义性最大的 191 个词(其中名词 121 个,平均 7.8 个义项;动词 70 个,平均 12 个义项)。这 191 个词各覆盖所有多义名词和动词出现的 20 %。

共计标注 192 800 词次(分别是 Brown 语料库的 50 个文件共 7 119 词次; WSJ 的 6 个文件共 14 139 词次)。其中名词 113 000 词次,动词 79 800 词次。每个多义词最多达 1 500 个例句。其负责人估计标注的错误率大约在 10～20 %。随机抽取和 Semcor 中相同的 5 317 词次,两者的标注相同率为 57 %。随机选择 30 315 句,用 Kappa 统计量得到的 k 值是 57 %[6]。该语料库已经加入 LDC(编号: LDC97T12)。

基于该语料库的研究表明[15],这 191 个多义词,都不符合"一文一义"的假设[16]。另外,在包含多义词出现超过 2 次的文件中,有 39 %的文件不符合这个规律。本文认为,这和高频、歧义性大的选词策略密切相关。

2.1.3　SENSEVAL-1 语料库

1998 年在英国的 Sussex 大学举办了首次词义消歧国际评测(SENSEVAL-1)。该评测由 ACL 的 SIGLEX 负责。其英语语料是从牛津大学于 1993 年建成的 HECTOR 语料库中抽取部分语料组成的。抽取后用 HECTOR 词典重新标注,标注者均为词典编纂专家。选择 35 个多义词,涉及名词、动词、形容词和 5 个词性不确定的词。标注的总词次为 8 448 个。

作为国际上首次开展的词义消歧评测(2007 年更名为 SemEval-2007),该语料的意义在于提供了公开评测数据,并且可以免费下载。标注者把 HECTOR 中的义项标注映射到 WordNet 且标注质量较高(k 值超过 80 %),在此后的研究中多次使用[6,17,18]。自此以后的历届评测中,绝大部分的评测语料都可以免费下载,极大地推动了词义消歧相关研究。

2.1.4 SENSEVAL-2 语料库

Kilgarriff 组织了于 2001 年进行的第二次评测中的英语采样词任务[19]。词典是 WordNet1.7,语料选自 BNC-2 和 Penn TreeBank。标注的方法是先由两个标注者进行平行标注,他们标注不一致的交给第三方审查,如果第三方同意其中某个初始标注者的标注则赋予该义项;否则再交给另一个人审查,直到有两个以上标注者意见统一为止(这种标注方法为绝大多数手工标注者采用)。共选取 71 个多义词(27 个动词,15 个形容词,29 个名词),平均每个词 7.8 个义项。标注 7 957 词次,IAA 为 85.5%。其中形容词的 IAA 是 83.4%,名词的 IAA 是 86.3%。该任务共 27 支队伍参加,提交系统 27 个。需要注意的是,动词部分的语料是和"所有词"任务在一起的。基于其上的研究有文献[17,20]等。

Palmer 负责组织英语所有词任务[21]。所用词典是 WordNet1.7。语料来自 Penn TreeBank,共标注 2 387 个词次,其中动词 554 个、名词 1 067 个、形容词 465 个、副词 301 个。比赛中不提供训练语料。共 21 支队伍参加,提交系统 21 个。与 Semcor 一样,几乎所有的进行所有词消歧研究的实验,都会用到该数据集[10~13]。

2.1.5 SENSEVAL-3 语料库

Mihalcea 组织了 2004 年进行的第三次评测英语采样词任务[22]。词典选择:名词和形容词义项来自 WordNet1.7.1,动词义项根据 WordSmyth 确定。之所以这样做,是因为 WordNet 中动词的义项区分过细。选用的语料是 BNC。为增大语料库规模,组织者在网上募集自愿者来进行词义标注。

所选多义词分别是 20 个名词,5 个形容词和 32 个动词,共计 57 个,每词平均 6.47 个义项。共标注 11 804 词次,其中 7 860 个作为训练样例,3 944 个作为测试样例。语料的 IAA 是 67.3%,根据 Kappa 统计量得到的 k 值分别是 0.58(micro-K)和 0.35(macro-K)。共 27 支队伍参加,提交系统 47 个。因标注质量并不高,后续相关研究并不多[23]。

本次评测中的所有词语料,由宾州大学提供[24]。所用词典是 WordNet1.7.1。语料选自两篇华尔街时报和一个 Brown 语料库的文件,题材分别为社论、新闻报道和科幻文章,共计约 5 000 个单词。共标注 2 212 个词次。语料的 IAA 是 72.5%(其中动词为 67.8%,名词为 74.9%,形容词为 78.5%)。16 支队伍参加,提交系统 26 个。相关研究见文献[13,23]。

2.2 汉语词义标注语料库

2.2.1 北京大学词义标注语料库

Wu 详细描述了北京大学计算语言学研究所建设的词义标注语料库[25]。所选语料是 2000 年 1~3 月和 1998 年 1 月 1~10 日的《人民日报》(共计 642 万字)。在词义标注前已经完成切词和词性标注。所用词典是北大计算语言学研究所研制的现代汉语语义词典(Chinese Semantic Dictionary, CSD)。该词典基于《现代汉语语法信息词典》[26]开发,从词的组合关系出发,进行词义区分和描述。词典采用"属性—值"的描述方法,如表 1 所示。

表 1 现代汉语语义词典关于词条"想"的描述

词语	词类	拼音	义项	同形	释义	语义类	子类框架	配价数	主体	客体	ECAT	WORD	例句
想	v	xiang3	1		思考	心理活动	NP	2	人类	抽象事物	V	think	~办法
想	v	xiang3	2		推测,认为	心理活动	VP	1	人类		V	suppose	我~他今天不会来
想	v	xiang3	3		希望;打算	心理活动	VP	1	人类		V	want	我~去杭州一趟
想	v	xiang3	4		想念	心理活动	NP	2	人类	人	V	miss	我~妈妈了

义项标注由中文系的 1 名博士和 1 名博士生,1 名计算语言学方向的博士生和 1 名有多年语料库标注经验的工作人员负责,已完成情况如表 2 所示。IAA 为 84.8%。目前标注工作仍在进行之中。

其中 1998 年 1 月 1~10 日的《人民日报》词义标注语料可以免费下载(http://www.icl.pku.edu.cn)。

该语料库将在北京大学正在研制的"综合性语言知识库"中扮演重要角色:把现有语言数据资源无缝整合,填补其各构成成分之间的"缝隙(gap)"。粗粒度的词义标注语料库以"词语"+"词类"+"同形"为轴连接了标注语料库和语义词典;细粒度的词义标注语料库以"词语"+"词类"+"同形"+"义项"为轴连接了标注语料库和语义词典。这就是以词义

为主轴把标注语料库与词典知识库连接起来的基本构思。进一步还可以把中文概念词典(Chinese Concept Dictionary, CCD)集成进来[27]。

表 2　北大词义标注语料库情况说明表

	CSD		词义标注语料库
词性	多义词(个)	平均义项(个)	标注词次(个)
名词	794	2.14	20 664
动词	168	3.41	45 538
合计	962	2.36	66 202

2.2.2　台北"中研院"语料库

该语料库由台北"中研院"的黄居仁教授负责。语料选自台北"中研院"语料库。选择"中频"多义词,且词的义项在 3～5 个。截至 2004 年 9 月,历时 3 年共标注 107 078 词次,IAA 接近 92.6 %。

另外,台北"中研院"、哈尔滨工业大学分别为 SENSEVAL-2 和 SENSEVAL-3 提供了中文评测语料,复旦、清华和山西大学等都进行过词义标注语料库建设,囿于篇幅,本文不多做介绍。

2.3　其他语种词义标注语料库

除上面介绍的英语和汉语词义标注语料库外,还有捷克语、罗马尼亚语、韩语、日语、土耳其语、巴斯克语、西班牙语等等。本文仅对日语语料库做简单介绍。EDR语料库由日本电子辞书研究院(Japan Electric Dictionary Research Institute, EDR)负责。语料全部是新闻报道,约 200 000 个日语句子。词义来源于 EDR 概念词典,对所有的实词(约 20 万)进行标注。没有看到标注总词次和标注一致率的报道。除此以外,还标注了语义角色。基于该语料库所做的研究见文献[28,29]等。

另一个日语词义标注语料库是 NTT 的 Hinoki[30]。该语料库既标注了词义也标注了语义角色。所用的词典是 NTT 的日语语义词典 Lexeed。该词典按照熟悉程度把日语单词分为 7 级,只选择熟悉程度大于等于 5 的词入选该词典,共计 28 000 个。对该词典的统计表明,越不熟悉的词越倾向于单义。需要说明的是这里的熟悉程度(familiarity)并不是使用频次,而是来自一个心理测试。

标注的语料有两方面的来源: 一个是词典 Lexeed 中本身的定义和例句(定义和例句中所用的词也仅限于该词典中出现的词);一个是新闻(Mainichi)。标注前都作了词性标注。标注时,每 5 个人一组,共有 3 组。涉及多义词 9 835 个,平均每个词有 2.88 个义项。共标注 818 814 词次,其标注一致率 IAA 为 78.7 %。

最后对上面介绍的词义标注语料库,总结为表 3。

表中学术影响部分,为本文根据语料库在目前 WSD 研究中被引用的情况、是否免费等因素所给出的个人评价。

表 3　词义标注语料库一览表

语料库名称	规　模	覆　盖	标注一致性(IAA)		词　典	影　响
			IAAraw	k		
Semcor	234 136	所有实词	—	—	WordNet1.6	★★★★☆
DSO	192 800	191 个名词和动词	80 %	57 %	WordNet 1.5	★★★☆
SENSEVAL-1	8 448	35 个实词	—	80 %	HECTOR	★★★★☆
SENSEVAL-2(采样词)	7 957	73 个实词	—	85.5 %	WordNet1.7	★★★★☆
SENSEVAL-3(采样词)	11 804	57 个	67.3 %	58 %	WordNet1.7 WordSmyth	★★★★
SENSEVAL-3(所有词)	2 212	名词、动词、形容词	—	72.5 %	WordNet1.7	★★★★
北大计算语言所	66 202	942 个名词、动词	84.8 %	—	CSD	★★★
Hinoki	818 814	9 835 个	78.7 %	—	Lexeed	★★★

3 自动构建词义标注语料库研究

人工建设一个大规模、高质量的词义标注语料库是一个耗时耗力的语言工程。一直以来,都有研究者尝试用自动或半自动的方法进行建设。本文主要介绍 bootstrapping 方法和基于双语对齐语料库所做的研究。

3.1 Bootstrapping 方法

该方法的基本思想是,人工标注的语料作为种子,以此为基础,利用一个或多个监督分类器,自动地迭代扩大标注语料库。较早的研究是 Yarowsky 采用决策表分类器,利用“一文一义”[16]的规则[LU1],针对同形词(Homograph)进行词义消歧实验[31]。

Mihalcea 在多义(Polysemous)的层面上,利用互联网,基于 bootstrapping 的思想,设计一个生成算法[32]。该生成算法由下面三步组成:

第一步:用人工标注的语料创建一个种子集合。包括以下人工标注语料:SemCor,从 WordNet 中提取的语料等。

第二步:用这些种子语料作为查询请求,搜索互联网。获得包含这些请求的前 N 个网页。

第三步:对包含该查询的网页片段进行消歧。把消歧后的网页片段加入种子集合,返回第二步。

具体实现时,要求第一步中的种子语料满足以下限制:1)至少包含两个开放词类的单词;2)两个开放词类中至少一个已经标注义项;3)目标词是名词短语的一部分或者有动宾、主谓关系。例如,对于多义名词“channel”,初始种子集合为{“fiber optic channel”、“river channel”、“channels in the surface”、“water channel”、“channel of expression”、“calcium channel”、“sports channel”}。同时,要求第三步中,进行消歧的词应该和查询中已经标注词义的词具有以下关系之一:词形相同;同义关系;上下位或兄弟关系。如果只是为了针对某一个特定的词进行消歧,则只需要使用“词形相同”这一关系即可。

针对上面提到的“channel”的例子,利用相同的消歧程序和 SENSEVAL-2 的测试集合,用基于 bootstrapping 方法建成的标注语料库作为训练集合达到的性能,要优于利用 SENSEVAL-2 提供的训练数据达到的性能。

3.2 基于双语对齐语料的自动构建

究竟什么是“词义”,一个词应该有几个义项,这几个义项分别是什么,应该如何刻画等等,这些词汇语义学的问题在语言学界也没有一致意见[33~35]。上面介绍的英语词义标注语料库绝大多数使用 WordNet,但是 WordNet 被人诟病其义项区分的颗粒度太小,以至于人工标注时,标注者有时都不能达成一致[31]。在自然语言处理的应用中,也不容易把握词义区分的颗粒度(WSD 因此被批评为一个孤立的自然语言处理问题)。而一个词对齐(word aligned)的双语平行语料库,就是一个词义标注语料库:不同的翻译对应着不同的“义项标注”。这样不仅避免了词义区分(word sense discrimination)的纷争,而且可以直接为机器翻译服务。文献[36]较早建议使用双语平行语料库来进行词义消歧研究。

Ng 利用 GIZA++对 6 个中英平行语料库进行词对齐[4]。接下来,手工完成翻译对到目标语义项的映射(也可通过双语词典自动完成)。仍以“channel”为例,如果对应的中文翻译是“水渠”或者“排水渠”,则都对应到 WordNet1.7 的同一个 Synset(描述为:“A passage for water”)。在 SENSEVAL-2 的 29 个名词上进行实验,义项个数由 WordNet 中的 5.07 个减少到 3.52 个。其中 7 个词变成了单义词,实际只有 22 个多义词。在消歧程序和测试集合不变的情况下,初步的实验结果表明对于绝大多数歧义词,用人工标注的训练集要好于双语对齐语料库的结果。Ng 进一步分析认为领域相关和某些义项的训练语料过少(有些义项甚至在平行语料库中没有出现)是导致这一结果的两大原因。通过把训练语料和测试语料重新分组以消除领域因素的影响,两者的差距由 0.189 降至 0.14。进一步去掉测试集在平行语料库中出现过少的语料后,两者的差距降至 0.065。由此可见,词对齐的双语语料库可以作为建设词义标注语料库的一条有效途径。

该方法面临的主要问题是缺乏大规模的词对齐平行语料库。由此引起的问题是某些义项对应的翻译在对齐语料中根本没有出现。为缓解这一问题,文献[20]提出使用汉语单语语料库和一个汉—英双语词典来构建词义标注样例(注:这些样例只是和特定义项密切相关的实词的集合,并非真实语料)。另外,由于多个义项对应同一个翻译词,必将导致比单语消歧的粒度更粗,从而实验结果不具备可比性。进一步在大规模的 Brown 语料库上对 800 个常用

多义名词进行实验。要求和 SENSEVAL-2 中采用完全相同的义项区分,对没有出现的翻译对,采用"加权替换"策略。结果表明,消歧准确率非常接近最好的系统(相差 0.8 %)[37]。

Tufis 同时利用了词聚类和多语言的 WordNet (BalkaNet, EuroWordNet)在一部被译成 6 种语言的小说上进行实验[38]。利用平行语料库进行 WSD 研究见文献[39,40]等。

双语对齐语料库造成多义词的义项减少,会给诸如信息检索等应用带来问题。比如"病毒",在汉语中分别指"比病菌更小的病原体"和"有害的计算机程序",而两种义项对应的英语翻译都是"virus"。这也是基于双语对齐语料库进行词义标注语料库建设面对的一个难题。

4　分析与思考

目前几乎所有的词义标注语料库都是采用人工标注。尽管已经开展了自动或半自动标注方法的研究,但由于各种原因,研究成果并不尽如人意。以下针对词义标注语料库建设和应用中存在的问题,做简单讨论。

4.1　语料库规模小

采用人工标注进行词义标注语料库建设的缺点是耗时和一致性差,并且很难做到大规模。英语词义标注语料库中标注最多的 Semcor 语料库也仅有 20 万词次。如何引入半自动,甚至是全自动的方法来加速词义标注语料库的建设已经成为一个重要的研究课题。

在保证高标注准确率的前提下,让机器自动完成尽可能多的标注词次,是目前比较可行的半自动建设大规模词义标注语料库的途径。

Jin 利用决策表具有消歧准确率高的优点[41,42],根据大量的无标注语料上的词聚类结果,进行决策表扩展。实验结果表明在几乎不降低准确率的前提下,召回率得到大幅度提高(从 37 %提高到 57 %)。这种方法可以有效地加速词义标注语料库的建设[43]。

4.2　语料分布不平衡

除了标注规模小外,另外一个问题是语料分布不平衡。即便在一个规模较大的语料库中,也会有一些低频的多义词从未出现,或者是高频多义词的某些低频义项从未出现。在 2.2.1 节介绍的北京大学词义标注语料库中,其词典描述的 794 个多义名词中,仅 485 个(60.93 %)在该语料库中出现。在这 485 个名词中,只以一个义项出现的有 237 个。只有 248 个(占 31.16 %)多义词在这三个月的《人民日报》中表现为真正的多义词。

为平衡标注语料库的分布,在不增加人工标注工作量的前提下,可采用主动学习(active learning)的方法,自动选择信息量更为丰富的或可能是低频义项的未标注语料提供给标注者。Dang 和 Chen 分别在细粒度和粗粒度的英文语料上进行了实验,后者的结果更为乐观[44,45]。

最后,标注的一致性校对仍然采用人工方法。如何利用机器学习来自动发现语料标注中的不一致,从而改善标注的质量,也是亟待解决的一个难题。目前这方面的研究尚未看到相关报道。

4.3　词义标注语料库的应用

由于目前的标注语料库规模较小,只能用于词义消歧算法的评测研究。利用已有词义标注语料库训练得到的标注器,尚未在应用系统中使用。但文献[1]利用在 SENSEVAL3 词义标注语料库上证明性能很高的消歧模型,集成到统计机器翻译系统 Hiero 的解码过程中。实验用的语料 NIST MT 2002 的汉英语料,系统的 BLEU-4 值从原来的 29.73 提高到 30.30。该词义消歧模块从输出额外的翻译词和纠正已有翻译中的错误两方面改善翻译结果。

如何更好地把 WSD 集成到相关的自然语言处理应用系统中,是 WSD 研究者亟待解决的难题。

5　结论

词义标注语料库作为词义消歧研究的基础性资源,已经经过了十几年的建设。无论英语、汉语还是日语等都有了自己的词义标注语料库。特别是从 1998 年以来开展的国际评测,大大促进了词义消歧研究和词义标注语料库建设。但是建设一个大规模、高质量的词义标注语料库是一个耗时耗力的语言工程。而传统的手工标注由于其固有的耗时、耗力和不一致等缺点,以致目前的词义标注语料库规模和质量都不足以训练得到一个可以应用的词义消歧系统。

探求半自动、甚至自动地建设大规模词义标注

语料库的策略、模型、算法,显得极为迫切。本文认为,bootstrapping 的方法是半自动建设大规模词义标注语料库的有效方法,而利用互联网资源对于获取多义词的低频义项出现具有现实意义。

最后,如何利用大量的无标注语料以改善消歧效果,虽然很早就开始研究[31],但并没有本质进展。随着半监督学习研究的深入,我们期望词义标注语料库建设的研究能从中受益。

参考文献:

[1] Y. S. Chan, H. T. Ng and D. Chiang, Word Sense Disambiguation Improves Statistical Machine Translation [A]. In: Proceedings of the ACL-2007 [C]. 33-40.

[2] C., Stokoe, M. P., Oakes, J. Tait, Word Sense Disambiguation in Information Retrieval Revisited [A]. In: Proceeding of the ACM SIGIR 2003 [C]. 159-166.

[3] 闵金明, 孙乐, 张俊林. 重新审视跨语言信息检索 [J]. 中文信息学报, 2006, 20(4): 33-40.

[4] H. T. Ng, B. Wang and Y. S. Chan, Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study [A]. In: Proceedings of the ACL-2003 [C]. 455-462.

[5] J. Veronis, Sense tagging: Does it Make Sense? [A]. In: The Corpus Linguistics'2001 Conference [C]. 2001.

[6] H. T. Ng, C. Y. Lim and S. K. Foo, A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation [A]. In: Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources [C]. 1999. 9-13.

[7] J. Carletta, Assessing Agreement on Classification Tasks: The kappa statistics [J]. Computational Linguistics, 1996, 22(2): 249-254

[8] G. Miller, C. Leacock, R. Tengi and T. Bunker, A Semantic Concordance [A]. In: Proceedings of ARPA Workshop on Human Language Technology [C]. 1993.

[9] M. Stevenson and Y. Wilks, The Interaction of Knowledge Sources in Word Sense Disambiguation [J]. Computational Linguistics, 2001, 27(3): 321-349.

[10] D. McCarthy and J. Carroll, Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences [J]. Computational Linguistics, 2003, (29):4, 641-654.

[11] D. McCarthy, R. Koeling, J. Weeds and J. Carroll, Finding Predominant Word Senses in Untagged Text [A]. In: Proceedings of ACL [C]. 2004.

[12] U. Kohomban and W. S. Lee, Learning Semantic Classes for Word Sense Disambiguation [A]. In: Proceedings of ACL [C]. 2005. 34-41.

[13] S. Brody, R. Navigli and M. Lapata, Ensemble Methods for Unsupervised WSD [A]. Proceedings of ACL [C]. 2006.

[14] H. T. Ng and H. B. Lee, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach [A]. In: Proceedings of ACL [C]. 1996. 40-47.

[15] R. Krovetz, More Than One Sense Per Discourse [A]. In: Proceedings of the ACL-SIGLEX SENSEVAL Worshop [C]. 1998.

[16] W. Gale, K. Church and D. Yarowsky. One Sense per Discourse [A]. In: Proceedings of the DARPA Speech and Natural Language Workshop [C]. 1992.

[17] D. Wu, W. Su and M. Carpuat, A Kernel PCA Method for Superior Word Sense Disambiguation [A]. In: Proceedings of ACL [C]. 2004.

[18] M. Palmer and H. T. Dang, Making Fine-grained and Coarse-grained Sense Distinctions, Both Manually and Automatically [J]. Natural Language Engineering, 2007, (13):137-163.

[19] A. Kilgarriff, English Lexical Sample Task Description [A]. In: Proceedings of ACL-SIGLEX SENSEVAL-2 workshop [C]. 2001. 17-20.

[20] X. Wang and J. Carroll, Word Sense Disambiguation Using Sense Examples Automatically Acquired from a Second Language [A]. In: Proceedings of EMNLP [C]. 2005.

[21] M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. T. Dang. English tasks: All-words and verb lexical sample [A]. In: Proceedings of the SENSEVAL-2 workshop [C]. 2001. 21-24.

[22] R. Mihalcea, Timothy Chklovski and Adam Kilgarriff, The SENSEVAL 3 English Lexical Sample Task [A]. In: Proceedings of ACL-SIGLEX SENSEVAL-3 workshop [C]. 2004. 25-28.

[23] R. Mihalcea, Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling [A]. In: Proceeding of HLT/EMNLP [C]. 2005. 411-418.

[24] B. Snyder and M. Palmer, The English All-Words Task [A]. In: Proceedings of ACL-SIGLEX SENSEVAL-3 workshop [C]. 2004. 41-43.

[25] Y. Wu, P. Jin, Y. Zhang and S. Yu. 2006. A Chinese Corpus with Word Sense Annotation [A]. In: Proceeding of ICCPOL'06 [C]. 2006. 414-421.

[26] 俞士汶,等. 现代汉语语法信息词典详解(第二版) [M]. 北京: 清华大学出版社, 2003.

[27] 俞士汶,等. 汉语词汇语义研究及词汇知识库建设 [A]. 第七届汉语词汇语义学研讨会 [C]. 2006.

[28] A. Fujii, Corpus-Based Word Sense Disambiguation [D]. Tokyo Institute of Technology, 1998.

[29] K. Shirai and T. Yagi. Learning a Robust Word Sense Disambiguation Model Using Hypernyms in Definition Sentences [A]. In: Proceeding of COLING

[C]. 2004.
[30] T. Tanaka, F. Bond and S. Fujita, The Hinoki Sensebank - A Large-Scale Word Sense Tagged Corpus of Japanese [A]. In: Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora [C]. 2006. 62-69.
[31] D. Yarowsky, Un-supervised Word Sense Disambiguation Rivaling Supervised Methods [A]. In: Proceedings of ACL [C]. 1995. 189-196.
[32] R. Mihalcea, Boostrapping Large Sense Tagged Corpora [A]. In: Proceedings of the 3rd International Conference on Languages Resources and Evaluations [C]. 2002.
[33] A. Kilgarriff, I don't believe in word senses [J]. Computers and the Humanities, 1997, (31): 91-113.
[34] 符淮青. 现代汉语词汇(增订本)[M]. 北京: 北京大学出版社, 2004.
[35] 徐国庆. 现代汉语词汇系统论 [M]. 北京: 北京大学出版社, 1999.
[36] P. Resnik and D. Yarowsky, A Perspective on Word Sense Disambiguation Methods and Their Evaluation [A]. In: Proceedings of The ACL-SIGLEX Workshop Tagging Text with Lexical Semantics [C]. 1997. 79-86.
[37] Y. S. Chan and H. T. Ng, Scaling Up Word Sense Disambiguation via Parallel Texts [A]. In: Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005) [C]. 1037-1042.
[38] D. Tufis, R. Ion and N. Ide, Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering [A]. In: Proceedings of COLING [C]. 2004.
[39] C. Li and H. Li, Word Translation Disambiguation Using Bilingual Bootstrapping [A]. In: Proceedings of ACL [C]. 2002. 343-351.
[40] M. Diab and P. Resnik, An Unsupervised Method for Word Sense Tagging using Parallel Corpora [A]. In: Proceedings of ACL-2002 [C]. 255-262.
[41] D. Yarowsky, One Sense Per Collocation [A]. In: Proceeding of ARPA Human Language Technology workshop [C]. 1993.
[42] D. Yarowsky, Hierarchical Decision Lists for Word Sense Disambiguation [J]. Computers and the Humanities. 2000, (1): 179-186.
[43] P. Jin, X. Sun, Y. Wu and S. Yu. Word Clustering for Collocation-Based Word Sense Disambiguation [A]. In: Proceedings of the 8th International Conference on Intelligent Text Processing and Computational Linguistics [C]. 2007, 267-274.
[44] H. T. Dang, Investigations into the Role of Lexical Semantics in Word Sense Disambiguation [D]. University of Pennsylvania, 2004.
[45] J. Chen, Towards High-Performance Word Sense Disambiguation by Combining Rich Linguistic Knowledge and Machine Learning Approaches [D]. University of Pennsylvania, 2006.

书讯(合订本)

2007 年《中文信息学报》合订本已出,还有少量过刊合订本,详细定价如下:

出版年份	定价(元)	出版年份	定价(元)
1997	30	2003	55
1998	30	2004	65
1999	55	2005	70
2000	55	2006	85
2001	55	2007	100
2002	55	—	—

愿购者(邮购需加 15 %的邮资费),请按以下地址汇款:

邮编: 100190　　通信地址:北京 8718 信箱《中文信息学报》编辑部

电话: 010-62562916　　E-mail: cips@iscas.ac.cn