词义消歧研究的现状与发展方向*)

The State-of-the-Art and Future Aspects of Word Sense Disambiguation

李 生 张 晶 赵铁军 姚建民

(哈尔滨工业大学计算机学院 哈尔滨150001)

Abstract This paper outlines the state-of-the-art of word sense disambiguation(WSD)in the following aspects: the application of WSD, representation of disambiguation knowledge, approaches and their evaluations. The paper analyzes current widely-applied WSD approaches, compares their merits and shortages. The direction of further researches in WSD field has been explored.

Keywords Word sense disambiguation, Evaluation, State-of-the-art, Research direction

1 词义消歧及其应用

词义是词汇在一定的语言环境下反映的特定语言现象。它能够明确地表达该词汇在该语境下表达的语义属性如感知、行为和情绪等;表达该词汇与相关词汇之间的关系;并且表达该词汇所特有的知识及常识性的知识。通过词义,人们将能运用自己的思维描述该语言现象,对其进行推理,或者为指代词从上下文中找到指代物。

在自然语言中,一个词汇往往存在多个词义,称为词的多义性。例如:Bank 有"银行"、"河岸"的意思。但是当词汇处于一定的语言环境,则只有唯一的意思。例如:"He slipped down the bank"中,bank 的意思是"河岸"。词义消歧就是使计算机自动为词汇选择正确意思,是自然语言处理领域中词汇级别上的最大难题。

词义消歧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。其应用至少包括下述领域[1]:

机器翻译:机器翻译为源语言文本找到目标语的对等文本。研究表明:平均每个英语词汇对应大约2.33个汉语译文[2];每个俄语词汇存在大约三个英文译文[1],所以词汇的歧义现象是它面临的最大难题之一。词义消歧为源语言多义词找到目标语对应词汇。

信息检索:通过关键词查找信息时,人们只需要得到与该词汇某一个词义相关的文本。比如,在通过单词"bank"查找财经资料时,词义消歧模块使系统检索出词义为"银行"的文本,忽略词义为"河岸"的文本。

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。只有对文本中的多义词消歧,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和语音合成中语音的校正,文本的处理都离不开词义消歧。

另外,当词义消歧被引入语法分析或句法分析中,将在一定程度上有效遏制语法或句法的歧义现象,缩小分析的范围,从而改善分析性能。

由于词义消歧服务于不同自然语言处理任务,不同任务存在不同的要求。比如:机器翻译侧重于用目标词的不同来区分词义,信息检索则强调词表达的概念及与其它词的近似度。

词义消歧是自然语言处理中非常困难的问题,甚至曾被认为计算机不能解决。但通过研究,一些词义消歧方法已经在一定范围取得了很好的效果,就我们所知,除了 Stetina[3],Mihalcea[4]等少数系统外,多数系统或者方法只针对某些词类实现词义消歧。

2 消歧知识的表示

词汇是语言的构成单位,词汇的使用需要遵循一定的语言规范。词义消歧是基于语言的知识工程,要想使计算机具有象人一样的词义辨识能力,就必须把人使用的语法、语义和语用等语言知识存储在计算机中,并把人理解语言的过程形式化。词义消歧中常使用的语言学知识可归纳为以下五种:

(1)词性:多义词的词义有时分属不同词性,多义词的词性标注称为兼类消歧。对于兼类消歧是否隶属

*)本课题得到国家自然科学基金资助(批准号:69775017)和国家"八六三"高技术研究发展计划基金项目资助(863-306-zt03-06-3和863-306-zd13-04-4)。

于词义消歧，研究者们看法不一致。但兼类消歧无疑是词义消歧的前端。

(2)词汇之间的关系：如同义、反义、部分和整体、类属(ISA)、肯定和否定等。

(3)词义定义：词义定义通常说明了该词义的出现条件和区别于其他词义的属性。

(4)语义原语：语义原语是自然语言中能用来说明其他词汇的词汇集合。

(5)语义之间的关系：按照语义属性聚类，得到词汇的统计特征用于消歧。例如概念间的依存关系。

(6)其他知识：如修辞学等。

按照乔姆斯基的理论模式，人的语言知识基础包括语法规则和词典两个部分。但语言的最大特点是开放性，词汇和语法均非封闭集合，处于变化中。语言学知识的表示方式将影响词义消歧的效果和效率，常用的表示有：

(1)详细描述词汇在各个词义下的合法使用规范，形成选择约束集合。消解歧义时，通过匹配词汇的用法，选择合适的词义。这种方式下，需要相对深入的语义分析。

(2)描述词汇之间的关联和参数选择[5]。例如：在"pot"的"COOKING VESSEL"词义和"dishwasher"的"the MACHINE FOR WASHING DISHES"词义之间构造关联，则可以对句子"John put the pot in the dishwasher"中"pot"消歧。

(3)建立网络描述词汇和语义之间的关系，可以通过激活和抑制的网络节点搜索机制选择词义[6]。

3 歧义消解的上下文选择

从上面的叙述，我们看出：语言学知识描述了词汇间的关系，歧义的产生源于词汇所涉及的领域、所处的结构等因素，所以，消解歧义的前提是为歧义词选择恰当上下文。

上下文的概念是专门针对多义词而言的，所有歧义的消解都依赖于多义词上下文提供的信息。

对于上下文对歧义词词义的影响，目前主要有两种观点。一种观点认为，上下文中每个词汇都对歧义词词义有影响，需要组合上下文提供的所有特征用于消歧，在这种观点的驱动下，人们常用 Bayes 数学模型，采用最小错误概率函数选择词义。另一种观点认为上下文提供的特征对于确定歧义词词义的贡献有差别，因而需要根据与歧义词距离、句法关系、搭配等关系选择使用上下文特征，这种观点下，人们较多地采用互信息理论作为歧义消解的原理。

选择上下文时，考虑较多的是上下文和待消歧词的距离、句法关系、搭配、语义关联等。因此，上下文的选择常分为三种情况：

(1)选择歧义词所在句子的一部分作为上下文，如句子"Have the students who missed the exam take it today."，通过局部的上下文"the exam"将能确定"miss"的词义。

(2)选择歧义词所在的整个句子作为上下文，如句子"It's hard to eat pizza with chopsticks."，只有查看全句，才能确定"hard"的词义。例如 Stetina[9]使用含有语义标注的 Brown 语料库，运用整个句子作为待消歧词的上下文，消歧正确率在80.3%左右。

(3)选择与歧义词有关的句群作为上下文。例如为句子"Why are you wearing THAT"? 中的指代词"that"消歧，需要结合与本句相关的其他句子获得词义。

到目前为止，人们较多研究的是小范围上下文信息、局部信息、领域知识，其它上下文信息应用的研究才刚刚起步。但是 Gale 等人的研究表明，大范围的上下文，即使扩大到一万个词的上下文，对词义消歧仍然有用[7]。

4 词义消歧方法综述

在上下文范围确定之后，消歧方法的选择是歧义消解的关键，也是研究最多的问题。知识获取是计算语言学领域所面临的最大瓶颈，选择消歧方法的目的在于选择有效的方法、以获得有助于确定词义的上下文特征或者知识。根据获取知识的方法，消歧的方法可以分四类：

(1)基于词典的消歧。机读词典和义类词典提供了有关词汇用法及词义的丰富知识，是词义消歧的主要知识来源。

从 Amsler 等人的研究开始，机读词典为研究者们所重视，并成为八十年代词义消歧工作的主要知识源。基于机读词典的典型方法是：利用单词在词典中不同义项的定义、计算歧义词的各词义的定义和上下文词汇的词义定义覆盖量[8]，选择覆盖量最大者作为当前词义。

遗憾的是，这种方法的正确率大约为50—70%。主要原因在于：一、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；二、机读词典中词义定义语句一般较短，以致于很多情况下，无论歧义的哪一种词义的定义与上下文单词的定义覆盖均为零；三、在实际应用中，不可避免的组合爆炸也限制了方法的使用。另外，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。

义类词典中的知识也被广泛用于词义消歧。和机读词典表达词义的方法不同的是，义类词典按照词义将词汇组织成层次结构，提供单词之间的关系。WordNet 和 Roget's 是最著名的英语义类词典，《同义词词林》和知网[9]是常用的汉语义类资源。一种基于 WordNet 分类体系的经典方法是计算歧义词及其上下文的概念密度，选择具有最大概念密度的词义作为当前词义[10]，该方法无需标注，正确率80%左右。

(2)基于规则的消歧。依赖语言专家的语言知识，构造规则库描述语言知识，分析歧义词及其上下文，选择满足规则条件的词义。规则通常描述限制歧义词修饰的成分或修饰歧义词的成分。例如，为动词书写其主语、宾语以及其他修饰语的语义特征的规则；为形容词和介词书写规则描述其修饰的名词或短语的语义特征。既使上下文存在多个多义词，往往也只有一种词义组合不违反规则限制，系统就选择这组词义。

规则使系统很快具有大量的语言知识，因此，该方法一直被广泛地应用，特别是基于转换的机器翻译系统。如哈尔滨工业大学的汉英双向机器翻译系统(BT863系统)，采用分析生成一体化的策略，在分析规则的右部就含有选择词义的操作，一旦该规则被调用，规则右部的词义消歧操作被自动执行。CMU 的 KANT 系统通过词法、语法消歧规则及人机交互方法进行词义消歧。此外，SYSTRAN，EUROTRAN 等著名的翻译系统都采用基于规则的方法。

由于规则通常由专家组织，因此有很大的主观性，知识不完备，而且难以应付领域的变化。如何维持规则库的一致性和可扩充性，是该方法需要关注的问题。

(3)基于语料库的方法。它的出现开辟了自然语言处理的新纪元。近年来逐步占据主导地位，目前的词义消歧研究已经离不开语料库的支持。从语料标注的角度看，分为有指导的方法和无指导的方法。前者从含有词义标注的熟语料中收集消歧知识，后者从无词义标注的生语料中收集消歧知识。

基于语料库方法以语料库作为知识源，核心是从语料库自动或半自动学习决定单词词义的上下文，从方法上看，可分为基于统计和基于实例两类方法。

基于统计的方法从标注或未标注的语料中统计支持歧义词用作不同词义时的上下文证据，这些证据用来对新输入句子的歧义词消歧。常统计的特征有词之间或词义之间的搭配。根据使用的语料库性质的不同，可分为：基于源语言的统计方法、基于目标语言的统计方法、基于双语的统计方法。尽管实验规模不大，但该方法正确率比基于规则和基于词典的方法有明显提高。

基于源语言的统计方法考查歧义词与上下文特征如词汇、语义特征、语素、复杂特征等的同现。Hearst 统计语料库同现情况，辅以上下文的语法语义特征，结合有指导和无指导的方法尝试对5个名词的消歧，正确率多在80%-90%左右[11]。刘小虎建立多上下文特征的词义消歧统计模型，对歧义词"interest"消歧测试的正确率达到80%[12]。Yang[9]在语料中统计 HowNet 中定义的义素同现频率情况，建立互信息评价函数，对汉语词汇作消歧。通过对人民日报语料实验，消歧效果良好，封闭测试正确率75%，开放测试正确率71%。其严重不足之处在于需要手工书写规则，人工标注大量的词义，因此难以大规模使用。

基于目标语同现的统计方法主要思想为：假设源语言句子 S 中有歧义词 SW_1，它有两个目标语译文：TW_{11} 和 TW_{12}，从 S 中选取与 SW_1 搭配的非歧义词 SW_2，其译文为 TW_2；在目标语统计的基础上，比较概率 $P(TW_{11}|TW_2)$ 和 $P(TW_{12}|TW_2)$，如果 $P(TW_{11}|TW_2)>P(TW_{12}|TW_2)$，则 SW_1 选择译文 TW_{11}，否则选择 TW_{12}[13]。该方法前提是源语言中单词的不同词义可以用不同目标语表示。这种方法具有两个优点：一、无需对语料库进行词义标注；二、较好的领域适应性。赵铁军等人在此基础上优化算法，用于哈尔滨工业大学 BT863—2英汉机译系统，译文选择的正确率为75%[14]。

Gale[7]应用大型并行英法语料库，采用统计方法对六个英语单词研究词义消歧：duty、drug、land、language、position 和 sentence。训练阶段首先利用词汇对齐的语料库，用与英语词汇对应的法语译文为英语歧义词标注词义，形成大量词义标注的实例；然后，计算不同的词义下特征词出现的频率；再计算各个特征词对消歧的贡献；测试阶段为输入句子中的歧义词选择一个词义，在该词义下，输入句中所包含的特征词的贡献总和最大。实验正确率在82%和86%之间。该方法需使用庞大的实例集，要求每个词的每个词义至少150个实例。

对于小规模的词义消歧而言，有指导的统计方法往往能获得高正确率。但由于缺乏大量已标注的语料库，难以大规模应用。针对这个问题，Mihalcea[4]曾提出利用 Internet 资源的解决办法。另外，由于获取对齐语料的技术尚不完善，语料库词义的自动标注也远没有达到实用的要求，所以当前词义消歧研究大量集中于如何利用未标注的单语语料库或双语不对齐的语料库。

基于实例的统计方法核心是实例间相似度度量。如果二者距离小，它们就相似。比较成功的如 Hwee Tou Ng 的 LEXAS 系统[15]。该方法应用 WordNet 提

供的词义定义，综合利用多种知识实现词义消歧，例如：上下文的词性、歧义词的词法、同现词及一些句法关系等。通过人工标注191个常用的歧义词词义，得到192,800个实例。在训练阶段，首先对由含有歧义词 W 的句子构成的语料标注词义，系统从统计与歧义词距离为3左右的上下文和歧义词本身词性、动宾句法关系和词形信息，构成特征矢量。对名词消歧时，还抽取作该名词谓语的动词。这些特征值组成的序列构成 W 的一个实例，每个训练句子提供一个训练实例。在测试阶段，LEXAS 从包含 W 的新句子中抽取出同样的特征矢量构成 W 的测试实例。通过比较训练实例，选择与测试实例最匹配的训练实例对应的词义为 W 的词义。通过在华尔街语料上测试，消歧平均正确率69%。Karov[16]通过循环地结合使用词汇之间及上下文之间的相似度，甚至可以处理数据稀疏问题，消歧正确率92%。

(4)将各种方法综合利用的方法。经过多年尝试，越来越多的研究者们倾向于综合多种方法以消歧。该方法组合多种知识和多种方法，获得更好的消歧性能。知识源的组合扩展了消歧可能用到的知识；多种方法的组合可以有针对性地解决不同的歧义现象。上面介绍的 LEXAS 系统就是一个例子。Stevenson 使用机读词典中提供的词义，结合词性、词义定义、主题类代码和选择约束构成词义消歧决策，用于语料库的词义标注，正确率94%[17]。黄昌宁和李涓子依托《同义词词林》的语义类体系，应用无指导的学习方法使用大规模语料库，构造语义类的"分类器"，用于词义消歧[18]。刘颖等人在不同层面将不同的规则，如配价搭配规则、属性制约规则和结构制约规则等，和基于马尔可夫模型的统计方法结合消解歧义，消歧正确率在92%左右[19]。恰当地引入机器学习算法也将改善词义消歧效果。例如荀恩东使用以消歧矩阵为计算背景的贪心算法选择译文[20]。

5 词义消歧结果的评价

多年来，人们在词义消歧领域做了大量的研究工作，列举实验数据以证明某方法或某系统的消歧效果。但由于测试条件不规范，人们难以定量地对比和评价各种消歧方法和系统。随着消歧研究日益增多，如何客观比较和评价词义消歧方法越来越成为值得关注的问题。总的来说，造成评价困难的主要因素有两个：

(1)测试的词汇集、训练语料及测试语料不统一。Brown 语料和 Penn Treebank 常被用作词性标注和句法分析等自然语言处理任务的训练和测试语料，计算结果的正确率。但目前几乎没有为研究者们广泛接纳的词义标注语料，人们自备语料测试词义消歧实验结果的正确率，得到的正确率因训练语料和测试语料的不同而不同，这样导致了现有的实验数据多数不具备可比性。

(2)词义正确与否的评价标准不同。自然语言处理中常用精确率作为实验结果的评价标准。其定义如公式1所示：

$$\%精确率=\frac{正确标注的数量}{标注的总数量} \quad (1)$$

精确率是一种完全匹配的评价方式，由于不同的自然语言处理任务对词义颗粒度的要求不同，完全匹配的程度不能完全说明词义消歧系统性能的好坏。因而，仅仅使用精确率参数，不足以评价词义消歧系统和方法的优劣，消歧性能的好坏还需要考查所选词义与正确词义之间的语义距离。Resnik 建议使用交叉熵(cross entropy)作为另一个评价参数[21]，交叉熵被定义为：

$$-\frac{1}{N}\sum_{i=1}^{N}\log_2 \Pr_A(cs_i|w_i,context_i) \quad (2)$$

其中，N 是测试用的词汇数量；$\Pr_A$ 是采用算法 A 后多义词 w_i 在上下文 $context_i$ 中获得正确词义的概率。

既然词义消歧非自然语言处理的最终目标，而只是为达到某目的的一个中间过程，那么词义消歧可以采取两种评价标准：(1)独立评价词义消歧，不依赖于应用领域；(2)不单独地评价词义消歧的效果，而是考察其对实现系统最终目标的贡献。比如词义消歧在机器翻译系统中，对翻译性能的作用。

总结 词义消歧问题是自然语言处理领域的一个重要难题。二者有着几乎一样长的历史。本文从词义消歧及其应用、消歧知识的表示、常用方法及效果评价等角度描述词义消歧的现状。

本文列举了大量当今普遍采用的词义消歧方法，如基于词典的方法、基于规则的方法、基于语料库的方法和综合的方法，但这种分类也不是绝对的。

到目前为止，许多方法仍处于探索的起始阶段。无论哪种方法都没有很好地解决词义消歧问题，词义消歧还有待继续深入研究。从某种意义上说，词义消歧又回到了研究早期的经验主义方法和基于语料的方法。有了越来越充足的资源，消歧效果大为改善，同时，人们越来越多地探讨语义在词义消歧中的应用。尽管如此，词义消歧在今后相当长的时间里，仍是自然语言处理领域的难题之一。另外，如何在自然语言处理系统中将词义消歧与其他技术结合起来，提高系统的整体性能，也需要进一步探索。

参 考 文 献

1 Ide N , Véronis J . Introduction to the Special Issue on

(下转封四)

(上接第 98 页)

Word Sense Disambiguation: The State of the Art. Computational Linguistics, 1998, 24(1): 1～40
2 Wu Dekai, Xia Xuanyin. Large-scale Automatic Extraction of an English-Chinese Translation Lexicon, Machine Translation, 1995, (3-1): 285～313
3 Stetina J, Kurohashi S, Nagao M. General Word Sense Disambiguation method based on a full sentential context. In Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop (Montreal, Canada, July 1998)
4 Mihalcea R. Word Sense Disambiguation and its Application to Internet Search. Master Thethis of the Graduate Faculty of the School of Engineering and Applied Science Southern Methodist University, May 1999
5 Resnik P. Selectional preference and sense disambiguation. In: Proc. of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How? (Washington DC, April 1997)
6 Kozima H, Furugori T. Similarity between words computed by spreading activation on an English dictionary. In: Proc. of the 6th Conf. of the European Chapter of the Association for Computational Linguistics (EACL-93, Utrecht). 1993. 232～239
7 Gale W, Church K, Yarowsky D. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In: Proc. Fourth Int Conf. on Theoretical and Methodological Issues in Machine Translation. 1992. 101～112
8 Lesk M. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to tell a pine cone from an ice cream cone. In: Proc. of the 1986 SIGDOC Conf. Toronto, Canada, June, 1986. 24～26
9 董振东, 等. Abailable at: http://www.keenage.com
10 Eneko A, Rigau G. A proposal for word sense disambiguation using conceptual distance. In: Proc. of the 1st Int Conf. on Recent Advances in Natural Language Processing, Velingrad, 1995
11 Hearst M A. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. In: the Proc. of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research. Oxford. October. 1991
12 刘小虎, 李生, 赵铁军. 词义消歧的统计模型选择. Communications of COLIPS. 1997, 7(2): 69～75
13 Dagan I, Itai A. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 1994, 20(4): 563～569
14 赵铁军, 荀恩东, 陈斌, 刘小虎, 李生. 基于目标语统计的译文选择研究. 应用基础与工程科学学报, 1999, 7(1): 101～110
15 Ng H, Lee H. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In: Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96) (Santa Cruz, 1996)
16 Karov Y, Edelman S. Similarity-based Word Sense Disambiguation. Computational Linguistics, 1998, 24(1): 41～59
17 Stevenson M, Wilks Y. Combining Weak Knowledge Sources for Sense Disambiguation. In: Proc. of the Int Joint Conf. for Artificial Intelligence(IJCAI-99). 1999
18 黄昌宁, 李涓子. 词义排歧的一种语言模型. 语言文字应用, 2000, 8(3): 88～90
19 刘颖, 张祥, 刘群, 王斌. 用综合方法来处理汉英机器翻译中的歧义. 计算机研究与发展, 1999, 36(7): 59～62
20 荀恩东, 李生, 赵铁军. 基于汉语二元同现的统计词义消歧方法研究. 高技术通讯, 1998, 10(8): 21～25
21 Resnik P, Yarowsky D. A perspective on word sense disambiguation methods and their evaluation. ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington, D. C., USA. April, 1997

计算机科学

(1974 年 1 月创刊)
第 28 卷第 9 期 (月刊)
2001 年 9 月 25 日出版

中国标准刊号: ISSN 1002-137X / CN50—1075/TP

定价: 12.00 元　国外定价: 5 美元
邮发代号: 78—68
发行范围: 国内外公开

主管单位: 国家科学技术部
主办单位: 国家科技部西南信息中心
编辑出版: 《计算机科学》杂志社
重庆市渝中区胜利路 132 号　邮政编码: 400013
电话: (023) 63500828　E-mail: jsjkx@swic.ac.cn
主　　编: 朱宗元
印 刷 者: 重庆科情印务有限公司
总发行处: 重庆市邮政局
订 购 处: 全国各地邮政局
国外总发行: 中国国际图书贸易总公司 (北京399 信箱)
国外代号: 6210—MO