基于词典的语料库词义标注

DICTIONARY BASED CORPUS SENSE TAGGING

肖 航

XIAO HANG

新 加 坡 国 立 大 学 中 文 系

NATIONAL UNIVERSITY OF SINGAPORE

2009

基于词典的语料库词义标注

DICTIONARY BASED CORPUS SENSE TAGGING

肖 航

XIAO HANG

新加坡国立大学中文系

硕士学位论文

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF ARTS
DEPARTMENT OF CHINESE STUDIES

NATIONAL UNIVERSITY OF SINGAPORE
2009

Acknowledgments

To all the people who helped me with this thesis, I want to express my sincere appreciation.

I am deeply indebted to my supervisor, Dr. Wang Hui, who plays a role of an instructor in research and a friend in life to all her students. She spent numerous times on guiding me to shape the ideas in this thesis.

Thanks go to all the teachers in department of Chinese studies. Their eximious teaching help me go deeper in linguistic research. Thanks also to Mdm Fong Yoke Chan, Mdm Tan Sah Mui, and Mdm Quek Geok Hong for their earnest dedication.

I would owe my sincere thanks to friends who have accompanied me in these years. I got a lot from their intellectual inspirations. Their friendships enrich my study and life. They are Mr. Qin Shaokang, Mr. Lin Jinzhan, Mr. Bai Xiaopeng, Ms. Xu Tingting, Mr. Wang Yuelong, Ms. Liu Zengjiao, Ms. Koh Yi Ting, and Ms. Cheong Zheng Yin.

Lastly, I express deepest gratitude to my wife, Dr. Yang Lijiao, and my 5-month-old son, Xiao Yiyang. They are treasures in my life. In the days I am far away from home, their encouragement gave me the confidence to finish the study.

目 录

统计表一览

插图一览

Abstract

This study is aiming to build a word sense tagged Chinese corpus. Corpus sense tagging is a procedure to add semantic tags to target corpora, so it is part of the research of word sense disambiguation.

This study employs Contemporary Chinese Dictionary as the semantic system for sense tagging. The key point of dictionary based sense tagging is the sense distinctions which affect the precision of sense distinguishing both by annotator and by WSD software. This study manages to probe into the effects of different linguistic disambiguation cues extracted from dictionary to practice word sense tagging.

This thesis consists of nine chapters. Chapter 1 is the literature review of the field. Chapter 2 introduces the research questions, methodology and materials of the study.

In Chapter 3 to 6, the article describes the function of disambiguation cues such as Pinyin, part-of-speech and collocation in corpus sense tagging. Chapter 3 indicates that Pinyin could be used as an effective cue to distinguish the homographs. Chapter 4 points out that part-of-speech could be employed to discriminate the ambiguities, since a word with multi part-of-speech has multi-sense in the dictionary. Chapter 5 probes into the capability of

collocation used as sense disambiguation cue. Based on the sense tagged corpus, Chapter 6 further investigates word sense frequency distributions and estimates the baseline of dictionary based corpus sense tagging.

Based on the full sense tagging procedure, Charter 7 and 8 further analyzes the difficulties summed up in automatic and manual sense tagging. Charter 7 investigates the difficult points found in automatic sense tagging which are mainly related to the context similarities of different senses. Charter 8 tries to analyze the difficulties that obstruct the annotator to get higher consistency. According to manual tagging experience, the features of dictionary sense distinction have serious influences in inter annotator agreement. Charter 9 concludes the study, and then points out the main limitations of the study.

On the basis of the practice of sense tagged corpus construction, the study supports that a high precision word sense tagging procedure is realizable by using hybrid linguistic disambiguation cues; however, from the analysis of difficult parts in sense tagging, more studies in dictionary sense distinctions should be done to improve tagging consistency.

基于词典的语料库词义标注

引 言

本文研究的目的是构建词义标注语料库（Sense Tagged Corpus）。语料库词义标注是给语料中的多义词标注正确的词义，为语料库添加词汇语义标记。Leech（1993）指出词义标注是最实用的语义标注。词义标注语料库是机器翻译、信息检索等自然语言处理系统的基础性资源，在语言研究、词典编纂等方面也有重要应用。例如，Sinclair 等（1991）提出在 COBUILD 词典编纂中利用词义标注语料库统计得到词义频率信息编排义项。

本文使用“华文教材语料库”中的中小学语文教材作为语料库，总字数约为 200 万字。语料已经经过分词和词类标注。本文的研究将为这个语料库添加词义标记。

语料库词义标注研究存在采用词典（语言词典）、语义词典和面向计算机的词义知识库等几种不同的方式。本文的研究选择传统语言词典——《现代汉语词典》作为词义体系。《现代汉语词典》在释义上具有代表性，是语言研究中使用最为广泛的汉语词典之一，选择《现代汉语词典》作为词义体系可以提高词义标注语料库的在语言研究、词典编纂等方面的应用价值。

语料库词义标注是从词形到词义的词义识别过程。从词义标注来看，语料中的一个词形下主要存在三种类型的词义判断困难。首先是同形词的区分，如何识别两个词形相同的不同词的词义存在困难；其次是兼类词的区分，按照词典分词类立义项的规则，兼类词即多义词，因此识别词义需要通过词类进行；第三是词典多义项词的不同义项之间的区分，义项之间的可区分程度影响词义标注的准确率。

对人而言，多义词词义的识别主要依靠对多义词的词义进行语义上的理解；对计算机而言，词义识别需要关注的是多义词语义上的差别如何表

现为语言使用上的差异。本文从语料库词义标注出发，对从词形到词义的消歧线索进行分类，探讨如何通过读音区分词义不同的词、如何通过词类区分兼类词的词义、如何通过词语搭配确定词义，以及如何通过利用常用义标注提高词义消歧的下限值，然后对人工标注和机器消歧中遇到的典型难点进行总结分析，并提出改进意见。

词义消歧的可实现性有两种看法。一方面，由于词典义项划分没有明确的规则，造成了多义词不同义项之间的区分特征不明显。人在判断这种类型的多义词词义时存在困难，机器消歧由于缺乏消歧线索也很难实现高准确率。由此，一部分专家认为词义消歧是一个难以完成的任务。另一方面，从真实语料出发，多义词在语言使用上具有明显特点，通过读音、词类、搭配、义频分布等信息的辅助，多义词的歧义可以得到大幅度的消减，没有理论上那么严重。因此，面对具体语料库的词义消歧可以达到较高的准确率。本文将通过语料库词义标注实践探讨通过语言使用识别词义的可能性以及词义区分困难对词义标注的具体影响。

第一章　文献回顾

在自然语言中，词语多义是普遍的现象[1]。因此，如何识别文本中多义词的词义成为自然语言处理的一个重要课题。词义消歧（Word Sense Disambiguation，WSD）研究就是为解决这个问题而产生，其任务一般定义为根据语境为文本中的多义词选择合适的词义。词义消歧最早于1950年代作为机器翻译的一个任务被提出。Weaver（1955）指出实现翻译的前提是知道词语在当前语境下的词义，因此机器翻译系统必须具有词义识别能力。Wilks & Stevenson（1996）指出词义消歧是自然语言处理的基础性工作，是许多基于内容理解的自然语言处理任务的必要环节。

词义标注语料库建设和词义消歧研究是关联在一起的。词义标注需要利用自动消歧方法，没有自动消歧方法的辅助，大规模语料库的词义标注是难以实现的；另一方面，词义消歧研究也需要词义标注语料库，词义标注语料库提供的多义词不同词义的使用信息对提高词义消歧准确率至关重要。Veronis（2001）认为没有大规模词义标注语料库支持，词义消歧研究不会有本质的进步。

一个完整的词义消歧系统通常包含四个方面的内容（Dagan & Itai，1994）：1）知识来源；2）知识获取方法；3）消歧模型；4）结果评测。模型的四个组成部分中，知识来源是词义消歧系统基础性的部分，主要有下述几种类型（Agirre & Edmonds，2006）：1）词典（指传统语言词典），例如，LDOCE（Longman Dictionary of Contemporary English）、OALD（Oxford Advance Learner Dictionary）、CED（Collins English Dictionary）、辞海、现代汉语词典等；2）语义词典（义类、百科词典），例如，Roget's International Thesaurus、同义词词林（梅家驹，1983）等；3）面向机器的词义知识库，

[1] 根据 Zipf 的研究，词语多义的具有心理上的基础。Zipf 认为人类行为，包括语言交际行为，都遵守最小精力付出原则（Principle of Least Effort）。在语言交际过程中，出于省力，说话人总是希望用少的词汇就能表达多的意思，而听话人则希望说话人使用更多的词汇表达不同的意思以使信息的歧义变小。因此词汇的语义数量与其使用频率有密切关系，使用频率越高的词有越多的语义数。见 Zipf, 1949, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley Press.

例如，WordNet[2]、EuroWordNet[3]、知网（Hownet）[4]等，是语义词典的一种类型；4）词义标注语料库，例如SEMCOR等。知识来源的类型对词义消歧系统的消歧线索选择、消歧方法有很大影响。

传统语言词典与语义词典的主要差别是语言词典的词义是平面化的，没有层级关系，不同词的词义之间没有联系。而语义词典的词义往往是层级化的，不同词之间的词义通过逻辑语义关系（如部分—整体、上下位等）互相关联，形成一个语义网络。对词义标注来说，这些差别影响到消歧线索的选择。基于语言词典的词义消歧研究的重点是怎么利用词典释义、例证等提供的信息。Wilks et al.（1988）、Cottrell（1989）、Church et al.（1991）、Arriola（1996）、Guthrie et al.（1996）、Ng & Lee（1996）、Prince（1997）、Wilks & Stevenson（1998）、Stevenson & Wilks（2001）、Dang et al.（2002）、Marquez et al.（2004）、Krek et al.（2005）在这方面做了很多工作，提出了词典信息的提取和利用的多种方法。

1.1 词义标注语料库研究

目前已经建设的词义标注语料库主要以采用词义知识库WordNet为主，著名的有SemCor语料库、SenseVal语料库和DSO语料库等。采用传统语言词典进行词义标注的语料库数量很少，不成规模。

SemCor词义标注语料库[5]由美国普林斯顿大学Miller等人在WordNet的基础上创建（Miller，1993）。所标注的语料来源于Brown语料库。SemCor使用WordNet 1.6作为词义体系，标注语料中的名、动、形、副等实词的词义。随着WordNet的发展，词义标注自动映射到最新版本（目前是3.0）。最

[2] WordNet 是由美国普林斯顿大学的心理学家、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面向计算机的英语词典。WordNet 是一个覆盖范围宽广的英语词汇语义网，名词、动词、形容词和副词各自被组织成一个同义词的网络。每个同义词集合(Synset)都代表一个基本的语义概念，并且这些集合之间也由各种关系连接。详见《WordNet: An Electronic Lexical Database》(Fellbaum, 1998)。WordNet webpage: wordnet.princeton.edu .

[3] EuroWordNet webpage: http://www.illc.uva.nl/EuroWordNet/

[4] HowNet webpage: http://www. keenage. Com .

[5] SemCor 语料库可通过下述网址下载：http://www.cse.unt.edu/~rada/downloads.html

新版本的SemCor标注了234,136词次的词义信息。

SenseVal[6]词义标注语料库有-1、-2、-3、-4四个版本。1998年建设的SenseVal-1使用HECTOR词典作为词义体系，标注了约40个多义词的词典义项。从-2开始，SenseVal采用WordNet作为词义体系。SenseVal-2标注了71个多义词的词义，共7957词次。SenseVal-3是使用最多的版本，一共标注了57个多义词，其中有20个名词、32个动词、5个形容词。57个多义词的总词次是7860个，平均义项数是6.47个。

DSO词义标注语料库是新加坡国立大学的Ng和Lee于1996年建设的[7]，包含191个高频名词动词的词义标注语料，共标注192,800词次，所用的词典是WordNet 1.5。语料来源自Brown语料库和华尔街日报语料。所选择的191个词都是高频的歧义情况复杂的名词和动词。其中名词121个、动词70个。191个词占语料库所有名词和动词频次的20%。

汉语的词义标注语料库建设起步较晚，主要有北京大学汉语词义标注语料库（Chinese Word Sense Tagging Corpus, STC）[8]。该语料库由北京大学计算语言学研究所建设，所选语料是2000年1～3月和1998年1月的人民日报，共计642万字，所用词典是该所开发的《现代汉语语义词典》。该语料库标注了966个多义名词和动词的义项。其中名词794个、动词168个。（金澎，2008）

与现有研究不同的是，本文使用传统的汉语语言词典——《现代汉语词典》作为词义体系。使用语言词典作为词义体系，与使用强调词义关系的语义词典在标注的方式和方法方面存在很多不同，需要更多地关注词义划分、词语释义方面的问题。相比为自然语言处理服务的基于WordNet等的词义标注，基于语言词典的词义标注语料库对词义研究、词典编纂、语

[6] SenseVal 是基于 WordNet 的词义消歧评测，主要评测计算机词义消歧程序的性能。截止2007年，SenseVal评测已经开展四届。详见网站：http://www.senseval.org

[7] Ng H. T. and Lee H. B. (1996) Integrating Multiple Knowledge Sources to Disambiguate Word Sense : An Exemplar-Based Approach. In : Proceedings of ACL-1996. pp. 40-47.

[8] Wu Yunfang, Jin Peng, Zhang Yangsen & Yu Shiwen. (2006). A Chinese Corpus with Word Sense Annotation. In Proceeding of Computer Processing of Oriental Languages. (pp. 414-421). 语料库更多信息参见 Web page: http://icl.pku.edu.cn

言教学等有更为突出的应用价值。

1.2 词义标注的基本方法

语言词典可供了丰富多样的词义信息，是词义消歧研究的重要知识来源。1990 年代以来大量的传统词典已经有了机器可读版本，为词义消歧研究提供了便利。词义消歧研究是从利用语言词典开始的。

M. Lesk（1986）最早提出直接利用词典释义进行词义消歧的尝试。他提出的基于多义词的词典释义文本重叠程度比较的算法被称为 Lesk 算法。这种算法的理论依据主要有两个：1）一个词的词义和该词的词典释义词之间存在明显的语义联系；2）上下文中连续出现的词通常具有语义上的密切联系。Lesk 算法的基本过程是：1）从机器可读词典中读取语料中多义词的所有义项及其释义；2）计算多义词释义文本中实词的重叠程度；3）选择具有最多词语重叠的义项作为该词在文本中的词义。

由于经常出现意义上有关联紧密的一组多义词，释义文本中并没有重叠词、不能匹配上的情况，Lesk 算法总体准确率不高（约 40%）。另外，释义文本的重叠词，有时并不能指示意义，造成判断错误（冯志伟，2004）。Guthrie et al.（1991）、Cowie et al.（1992）、Kilgarriff & Rosensweig（2000）、Banerjee & Pedersen（2002）都对 Lesk 算法做过进一步的研究，进行了局部改进。

语料库词义标注根据标注词的不同，可分为部分词词义标注和全词（all-words）词义标注两种类型。全词词义标注是指标注语料中的全部多义词（一般限定于实词）的词义，标注的词汇量大。本文的研究属于语料库实词全词词义标注。

Stevenson & Wilks（2000）的语料库标注实验[9]是一个典型的语料库全词词义标注过程。他们提出全词词义标注有两个重要前提条件：1）满足语言处理的大词汇量及其释义；2）可用于词义消歧的知识。经过比较，他们选择使用传统的语言词典——朗文当代英语词典（LDOCE）进行标注实验。

[9] Stevenson M. and Wilks Y., 2000, Large Vocabulary Word Sense Disambiguation. 见《*Polysemy: Theoretical and Computational Approaches*》(Edited by Y. Ravin and C. Leacock, Oxford University Press, 2000)一书第 9 章，页 161-177。

实验仅限于标注实词（名、动、形、副）的词义。

开始标注前，首先利用 LDOCE 词典词类信息进行同形词和多义词的区分。在英语词典中，不同词性的词通常分列为不同的词目，例如单词“fast”在 LDOCE 中有四个词目，词性分别为副词、形容词、动词和名词。这种情况称为有多个同形词。一个单词的一个或多个同形词都有可能是单义项或多义项。由此，词典中的词语歧义在知道其词性的情况下可被分为三种类型：1）无歧义，一个词的所有同形词都是单义；2）部分歧义，一个词的多个同形词中有些单义，有些多义，若确定词性，可完全识别该词性下的单义词的词义；3）完全歧义，一个词的所有同形词都是多义项。根据调查，在 LDOCE 的词条中，有 34%是多义词，有 12%的词具有同形词（由于每个同形都有词义，所以同形词一定多义）。88%的多义词可以在同形层面上消歧，在知道词性的情况下，95%的多义词可以在同形的层面上达到不同程度的消歧。例如单词“fast”的四个同形词中，副词有 7 个义项，形容词有 14 个义项，动词和名词只有一个义项。可见，“fast”的多义仅限于副词和形容词，做动词和名词时是单义的，无需消歧。

具体的词义消歧过程如下：1）通过词类标注进行同形层面上的词义区分；2）使用 Lesk 算法根据多义词的词典释义文本比较进行词义消歧；3）利用 LDOCE 的主题代码（Subject Code）[10]通过语义相似性进行词义消歧；4）利用 LDOCE 中的“选择～限制”信息进行词义消歧；5）利用上下文进行词义消歧，消歧方法采用 Yarowsky（1992）提出的基于 Roget's 词典的统计模型，根据上下文词的义类联系判断词义，上下文为关键词左右各 50 个词；6）通过决策算法综合上述各步骤的消歧结果，给出多义词最可能的词义。

Stevenson 和 Wilks 的标注方法是语料库词义标注的一个基本模型，具有典型性，大部分其他标注系统都采用类似的步骤，主要的区别在于利用的资源和具体算法不同。从上述词义消歧过程来看，所利用的词典信息主要包含释义文本、搭配词和义类信息。

[10] 在 2006 版的 LDOCE 中，subject code 称为 signposts。Signposts 是一个词或短语，与释义文本相关联，用于辅助快速查找相关的词义（Signposts help to guide you to the meaning you want, if a word has a lot of different meanings.）。

现有的研究在如何利用各种消歧线索进行歧义消解方面做了很多工作。但对各种消歧线索所能消解的歧义的类型和区分程度方面研究较少。本文的研究将弥补这方面的不足。

1.3 词义标注的主要难点

基于词典的词义标注的主要难点之一是由于词义区分困难带来的标注的不一致问题。不同义项之间缺乏明确的区分线索造成了词义判断上的困难，并且由于没有客观的参照标准，使得母语人在多义词词义判断上也存在明显的不一致。目前的词义标注语料库多采用以多位标注者之间的一致性（inter-annotator agreement，IAA）作为标注准确率值。词义标注的不一致性很大程度上影响了词义消歧的可实现程度。许多学者都在这方面做了研究（Gale et al.，1992；Veronis，1998，2000，2001；Kilgarriff & Plamer，2000；Ide & Wilks，2006）。

Veronis（1998）对词义标注的不一致问题进行了实验分析。他针对多义词的认知和识别问题先后做了两个实验。实验采用法语词典 Petit Larousse[11]作为词义体系，标注者是 6 个大学四年级语言学专业的学生。实验一研究词语多义的认知情况，调查了标注人对 600 个多义实词（名动形各 200 个）的是否存在多义的认识。结论是 73.0%的词被认为只有一个意思，这与词典不符，所有的实验词在词典中均有多个义项。实验二调查词义标注过程中的词义可区分性问题。取标注人均认为多义词的 60 个词（名动形各 20 个），看在具体语料中（每个词各 60 个语句），标注人能否准确识别出多义词的正确词义。结论是标注人之间的不一致现象非常明显。通过实验，Veronis 认为对自然语言处理来说，词典的词义划分太细，不利于词义区分，主要问题在于：1）词典的释义包含了太少的可用于判断的线索；2）词典的义项区分并没有考虑事实上的词义分布；3）标注者在标注过程中普遍认为义项之间太模糊，缺乏足够的区别信息。Veronis 认为这个问题不仅仅存在于实验所用的词典中，在其他词典、包括其他语言的词典中也广泛存在，虽然现在新的词典开始提供更多的句法、搭配和典型句信息，

[11] 法国拉鲁斯出版社出版的一本常用的法语词典。

然而对面向计算机的应用而言，这些信息仍然不够系统和准确。

从 Veronis 等人的研究可以看到，在基于词典的语料库词义标注中，词义消歧所能达到的准确率与词典义项的可区分度密切相关。Kilgarriff（1992，1997，1998，2004）、Dolan（1994）、Ng（1997）、Wilks（1997）、Wilks & Stevenson（1997）、Veronis（1998，2000，2001）、Kilgarriff & Koeling（1999）、Ng et al.（1999）、Krishnamurthy & Nicholls（2000）、Palmer（2000）、Palmer et al.（2004）、吴云芳&俞士汶（2006）等都分析了词典的义项区分问题及其对词义标注的影响，指出词义的区分程度影响人及计算机程序对词义的判断。

Kilgarriff（1992，1997，1998）从词典编纂的角度讨论了词义区分对词义消歧的影响。他认为对词义消歧而言，若不能对词义体系进行清晰的定义，词义消歧研究无法向前发展；根据词典标注词义，势必要了解词典的释义过程，特别是释义的规则和义项分立的原则，而从目前的词典释义本身来看，词义标注很难达到高准确率。词义消歧研究中，细粒度（fine-grained）和粗粒度（coarse-grained）是和义项区分密切相关的两个概念，影响词义消歧的可实现程度（Kilgarriff，1992；Tufi et al.，2004；Ide & Wilks，2006；吴云芳等，2006）。Kilgarriff（1992）指出了人不能很好区分LDOCE中的“细粒度”的词义。Kilgarriff & Koeling（1999）认为如果词典学要给出真正的词义，就需要提供词义之间的区分线索，缺乏足够的区分线索基于词典的语料库词义标注或词义消歧是很难实现的[12]。

词义标注的不一致性是词义标注语料库建设必须面对的问题。在这方面，本文的工作是分析语料标注遇到的标注不一致的多义词的类型及其造成不一致的原因。

词义标注的另外一个突出的难点是词典释义是否符合词语的语言使用情况。具体表现为词典给出的词义并不能覆盖语料库中所有的词语使用情况，无法对语料中的所有词标注准确的词典义项（Veronis，2000；Ide & Wilks，2006）。

在词义与语言使用方面，语言学家 Meillet 指出词语的意义是通过语言

[12] If we are to know what word senses are, we need operational criteria for distinguishing them.

使用确定的（The sense of a word is defined only by its language uses.）[13]；Wittgenstein 也指出“没有所谓词义，只有具体的使用，……，词语的意义就是它如何在语言中使用”（There are no senses, but only usages……the meaning of a word is its use in the language.）[14]。而现阶段的词典中，普遍存在着词语释义不能覆盖所有语言使用的情况。根据 Wierzibicka（1989）的理论，确定一个词在具体语境中的准确含义是困难的，词典学家也只能给出一个词的核心意义（Core Sense），无法对一个词的全部意义进行解释。Wierzibicka 认为词典学家在解释一个词的词义时，通常只能找到一个原型（prototype），通过对原型的描述解释词义，而词汇在原型之外的变化并不能得到全面的解释。词典并不是为完备的列出词义而设计的。

1.4 词义消歧的可实现性

Kilgarriff（1993，1997）从词典释义出发，认为词义区分是困难的，词义区分的困难使得高准确率的词义消歧难以实现。Wilks（1997，2000）对 Kilgarriff 的观点进行了反驳。

Wilks 从多义词词义在语料中的具体使用和分布角度探讨了词义标注的可行性。他通过对 Gale 等人（1992）提出的“One Sense Per Discourse”[15]、Yarowsky（1993）提出的“One Sense Per Collocation”[16]两个观点进行分析，指出从语料本身出发，其歧义性并没有那么严重。Gale 等人（1992）指出 94%的多义词在一个语篇中只出现一个意思，Yarowsky（1993）认为可以利用搭配信息通过统计方法实现较高的准确率（95%）。Wilks 据此认为 Kilgarriff 的论断仅基于词典释义的复杂性本身，并未考虑具体的语言使用，因此高估了词义消歧的难度。

---

[13] 转引自 Veronis, J. (2001). Sense tagging: does it make sense. In *Proceeding of the Corpus Linguistics Conference-2001*.

[14] 同上。

[15] That word tokens in text tend to occur with a small number of senses than often supposed and, most specifically.

[16] In a single discourse a word will appear in one and only one sense, even if several are listed for it in a lexicon, at a level of about 94% likelihood for non-monosemous words (a figure that naturally becomes higher if the monosemous text words are added in).

从上述争论可以看出，词义消歧的可实现性存在两种情况，一种是理论的，一种是实践的。

理论上的词义消歧困难主要是通过对词典义项划分和词语释义的观察得到的。Kilgarriff（1993）所指出的词义的难以区分问题是面向词典的词语释义而言的，探讨了词义区分的理论上的困难。从词典释义的具体情况来看，确实普遍存在多义词的不同义项之间缺乏区分线索的情况。对很多多义词，我们很难找到可以明确的规则将不同义项区分开来。因此，对义项之间词义区分不明显、缺乏区分线索的多义词，机器消歧和人工标注都难以实现高准确率。

词义消歧具有可实现性的结论是从语料库词义标注实践得出的。Wilks（1997，2000）更为关注不同词义在使用上的区分，认为从具体语料出发，高准确率的词义消歧是可以实现的。由于在具体语料中，多义词的词义使用和分布具有规律，如果有效利用这些规律，词义消歧能够达到很高的精度。因此，理论上的词义区分问题可以通过实践过程来解决。

长期以来，词义消歧研究领域以计算机专家为主，研究对集中于消歧算法及其实现上。在词义消歧评测中，对词义区分存在的困难经常采取回避问题的态度。例如，2007年召开的SenseVal-3词义消歧评测会议，在动词的词义消歧评测中，由于考虑到WordNet中动词的词义区分很细，消歧准确率普遍不高，因此SenseVal-3放弃WordNet中的动词部分，转而采用颗粒度较粗的WordSmith[17]中的动词词义来做消歧评测。

词义区分的本质是语言学问题。2000年以来，词义消歧研究领域开始越来越多地从语言学角度讨论多义词词义区分问题，集中于讨论词义消歧需要什么样的词义体系、如何为词义消歧研究提供合适的词义体系等。

本文的研究同时关注词义区分的理论和实践两方面的内容。一方面，尝试通过词义标注过程分析读音、词类、搭配等信息对于降低歧义消解复杂度的作用及其可区分程度；另一方面，本文通过总结人工标注和机器消歧过程中遇到的困难，分析词典释义对词义标注带来的影响。由于语言词

[17] WordSmith 由英国利物浦大学 Mike Scott 设计，牛津大学出版社出版的一款软件，主要具备检索、单词列表、主题词统计等功能，是最常用的语料库研究工具之一。详见：http://www.lexically.net/wordsmith/version4/index.htm .

典自有其面向的对象，仅从词义消歧的角度讨论义项该如何划分是不充分的。所以本文研究的目的并不在于探讨义项分立问题，而是分析现有的多义词义项区分在具体语料中的表现。希望通过对词义标注过程中从词形到词义的词义判断过程做出量化分析，说明各个步骤涉及到的词义区分问题的性质以及可能的解决方法。

第二章　本文的研究

2.1 研究内容

本文的研究建立在利用《现代汉语词典》对大规模汉语语料库进行多义词词义标注，建设词义标注语料库的基础上。从语料库词义标注来看，多义词的词义区分表现为两种类型。第一种是外在的区分，通过语言、语法等属性体现出来，表现为同一词形的不同词义具有明显不同的特征，例如读音不同、词类不同、搭配不同或使用频率上的不同等；第二种是内在的区分，具体表现在词典划分义项的原则和义项间的区分特征上。

在多义词的界定上，本文从语料库词义标注出发，采用广义的多义词概念，即同一词形具有多种词义可能的均视为多义词。因此，除词典的多义项词外，本文所指的多义词还包括同形带来的多义、兼类带来的多义情况，如图 1 所示：

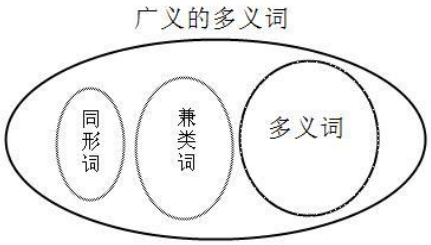

图 1：广义的多义词

在词义消歧的过程中，三种类型都具有歧义，需要通过词义消歧过程判断其词义。但是三种类型的多义词的消歧方式存在很大不同。同形词和兼类词通过外部特征如读音、词类来进行区分就可达到很好的消歧效果，同一词类下的多义项词则需要通过搭配、上下文及分析义项间的区别性特征来消歧。

语料库词义标注一般分为计算机自动词义消歧和人工校对（人工标注）两个过程。首先进行的是计算机自动根据消歧线索判断并选择正确词义的词义消歧过程，其次是通过人工校对来提高词义标注准确率的人工标注过程。反映在自动消歧和人工标注中的词义区分困难具体主要表现在四个方

面：1）针对义项划分，词典往往不能给出大部分多义词的义项分立原则；2）词典的义项划分并不能完全覆盖词语在语言中的使用情况；3）不同义项之间缺乏足够的区分信息；4）词典中大量存在细粒度的词义，带来准确区分上的困难。

基于词典的词义标注语料库建设如何面对并克服这些困难，其核心就是要探讨多义词词义的区分过程。本文研究如何从语料中出现的一个词形出发，首先通过拼音、词类等信息削减词语的歧义复杂程度，再通过搭配及上下文信息判断其词义。基于已建设的词义标注语料库，本文研究如何利用词义标注语料对词义区分情况进行量化分析，为词义消歧程序提供更多的可用信息，从而提高自动消歧的正确率。

本文的主要研究内容可以总结为下述五点：1）读音对词义的区分；2）词类对词义的区分；3）词义的频率分布对词义消歧的作用； 4）自动消歧中难以区分的多义词的类型及其原因；5）影响人工标注准确率的多义词的类型及其性质。

2.2 研究材料

在研究材料方面，本文的研究使用新加坡国立大学中文系王惠博士主持开发的“华文教材语料库”中的初中和小学语文教材部分，总字数约为200万字（含字母、数字和汉字，汉字数为约170万字）。详如表1所示。

表1：本文研究所用的语料

※语料内容		
小学语文课本：5套		
教材名称	出版社	版本
新课标《义务教育课程标准实验教科书》1-11册	人民教育出版社	2006版
新课标《义务教育课程标准实验教科书》1-11册	江苏教育出版社	2005版
《九年义务教育六年制小学课本》1-12册	江苏教育出版社	2001版
《九年义务教育五年制小学试用课本》1-10册	北京师范大学出版社	1999版
《九年义务教育六年制小学试用课本》1-10册	广东教育出版社	1991版
初中语文：3套		

新课标《义务教育课程标准实验教科书》13-18 册	人民教育出版社	2001 版
新课标《义务教育课程标准实验教科书》13-18 册	江苏教育出版社	2005 版
新课标《义务教育课程标准实验教科书》13-18 册	语文出版社	2003 版
※语料基本信息		
课文数：2237 篇		
总字符数：1986803 个		
总字数：1711493 个，其中有 4963 个不同汉字。		
总词数：1143120 个，其中有 52798 个不同词。		

“华文教材语料库”建设的主要目的是为词汇教学、词典编纂等服务，是一个经过深度加工的、高质量的标注语料库。在进行词义标注之前，初中和小学语文教材已经完成了分词和词性标注。分词和词性标注语料示例如下：

夕阳/n 染/v 红/a 了/u 西边/f 的/u 天空/n 。/w 一片片/mq 晚霞/n ，/w 倒映/v 在/p 清流/a 如/v 镜/n 的/u 小/a 河/n 里/f ，/w 像/v 开/v 了/u 一/m 大/a 朵/q 一/m 大/a 朵/q 鸡冠花/n 。/w

*词间以空格作为分界，“/”号后是词类标记；

*词类标记符号的意义见表 12 语料库词类标记集。

本文的研究将为其增加词汇语义标记。词义标注语料示例如下：

夕阳/n#xī/yang^1 **染/v#rǎn^1** **红/a#hóng^1** 了/u#le 西边/f#xī/bian 的/u#de 天空/n#tiān/kōng 。/w 一片片/mq#yī/piàn/piàn 晚霞/n#wǎn/xiá ，/w 倒映/v#dào/yìng 在/p#zài 清流/a#qīng/liú **如/v#rú^2** 镜/n#jìng 的/u#de **小/a#xiǎo^1** 河/n#hé 里/f#li ，/w 像/v#xiàng^3 **开/v#kāi^1** 了/u#le 一/m#yī **大/a#dà^1** 朵/q#duǒ 一/m#yī **大/a#dà^1** 朵/q#duǒ **鸡冠花/n#jī/guān/huā^2** 。/w （注：^号后为词典义项编号）

*示例语料中多义实词已经标注了该词在《现代汉语词典》中的义项编号；

*完整的标注词义的语料样本见附录：词义标注语料样例。

语料库词义标注在语义体系和词典资源的选择上有多种不同做法。本文选择在释义方面具有代表性的、使用最为广泛的《现代汉语词典》（2005版）作为词义体系。《现代汉语词典》是汉语语言研究、研究教学等使用最为广泛的词典。2005版《现代汉语词典》与之前版本比较，增加了词语的词类标记，这对语料库词义标注是非常重要的。

本文研究过程中，除使用《现代汉语词典》作为词义体系外，还参考《现代汉语规范词典》（李行健主编，外语教学与研究出版社、语文出版社联合出版，2004年1月第一版）的词语释义作为比较分析对象。英文的词义部分参考 *Longman dictionary of contemporary English*[18]（朗文当代英语词典，简称LDOCE）。此外，《现代汉语搭配词典》（梅家驹主编，汉语大辞典出版社出版，1999年12月第一版）中的词语搭配信息也作为材料应用到研究中。

2.3 研究方法

本文的研究建立在词义标注语料库建设的基础上。

词义标注语料库建设的具体过程是：1）从词形出发，利用读音区分大部分同形词；2）利用词类区分多义词的不同（语法）词义；3）利用搭配对词义的指示关系进行自动消歧；4）通过常用义/非自由义预标注提高标注准确率（提高词义消歧的下限值）；5）通过人工标注提高词义标注的准确率。总体流程如图2所示。

[18] Summers, D. (2006). Longman dictionary of contemporary English (New ed.). Harlow, Essex: Pearson/Longman.

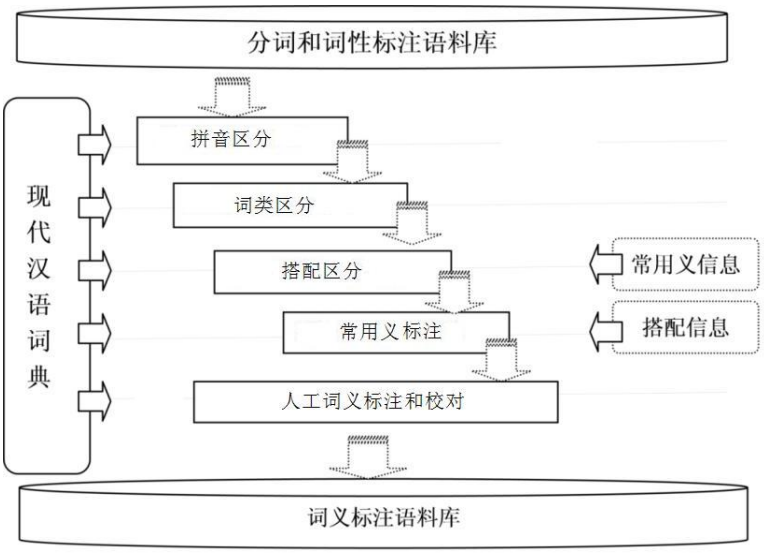

图 2：语料库词义标注流程

在标注词的范围上，本文主要探讨实词的词义标注，仅限于名词、动词、形容词三种词类。现代汉语词典中的形容词包含属性词和状态词两个小类，在语料中分别对应为区别词和状态词。虚词由于通常只具有语法意义，不是词义消歧的重点，本文中不作为词义标注对象。

在词汇量上，本文采取大词汇量词义标注的方式，标注语料库中全部多义实词的词义。语料中的词语组成情况如表 2 所示。

表 2：语料中的词语组成情况

语料中的词语组成情况			
语料库总词数	52798 个	总词频	1143120 次
※实词（名动形）	37603 个	占总词数约 71%	占总频次的 63%
※多义实词	5514 个	占实词数的 15%	占实词频次的 43%
※多义名词	2758 个		占多义实词频次的 28%
※多义动词	2473 个		占多义实词频次的 59%
※多义形容词	945 个		占多义实词频次的 13%

注：词数指词语的数量；频次指词在语料库中的出现次数；因为兼类的关系，名动形总数之和大于实词数量。

词义标注的人工校对由六个以汉语为母语的、语言学专业的研究生完成。校对方式是由校对者根据《现代汉语词典》的释义，通过多义词的语料上下文判断自动消歧给出的义项的正误，并为标注错误重新选择合适的词典义项。校对者通过上下文判断词义的方式如表 3 所示。

表 3：词义标注人工校对表

保证/v　　bǎozhèng				
①担保；担保做到：我们～提前完成任务。				
②确保既定的要求和标准，不打折扣：～产品质量丨～科研时间。				
句编号	左窗口	词	义项	右窗口
#1	信贷资金的目的出发，	保证	2	重点，区别缓急，坚持
#2	保证货源，保证质量，	保证	1	及时向外商供货，有意
#3	分布的维管束称叶脉，	保证	1	叶内的物质输导。
#4	全世界作出庄严承诺，	保证	1	香港继续她的生活方式
#5	均高出岷江的洪水位，	保证	2	内江与外江相互分隔。
#6	业务，为了保证货源，	保证	2	质量，保证及时向外商
#7	办法、规定外，并	保证	1	恪守下列条款：
#8	叔叔来之前，我们得	保证	2	那东西完好无损。
#9	使洪水顺畅排出，从而	保证	2	进入灌区的水量不致为患
#10	卵石甩向飞沙堰，从而	保证	2	宝瓶口和下游灌渠不致淤塞

表 3 中，首先给出的是校对词的词类、拼音和义项信息。其后的表格中，分别给出了包含目标词的语料库的所有语句。语句按目标词的位置分为左右两个部分，即上文和下文，分别是表格中的左窗口和右窗口列。义项列给出的是已经标注的义项号，以根据读音、词类、搭配等信息推测的可能义项为基础，辅以该词的最常用义预标注。

校对过程中，碰到模糊的不能确定的多义词允许选择多个义项，并按照可能性的高低顺序排列。这部分词将作为标注过程中的难点词进行专门处理，在第七章和第八章中进行专门分析。此外，遇到不能对应词典中任一义项的词和通过上下文语境不能准确判断的词都通过规定的符号做标

识，由后续的步骤来处理。

2.4 研究目的及意义

本文研究的目的是构建词义标注语料库。词义标注语料库是机器翻译、信息抽取、信息检索等自然语言处理系统的基础性资源，在语言研究等方面也有重要应用。《现代汉语词典》是汉语语言研究中使用最为广泛的词典之一，而目前还没有以《现代汉语词典》作为词义体系的标注语料库。本文的研究所建设的语料库在词义消歧、词典编纂、词汇教学等方面都具有重要的应用价值。

在对词义消歧的贡献方面，本文的研究有助于从充分利用各种消歧线索的角度来提高机器消歧的准确率。本文尝试通过分析语料库词义标注的完整过程，对各消歧线索对词义消歧的作用做出量化说明，并对词义区分困难的多义词进行分析，从而提出有针对性的消歧方法。语料库词义标注是从词形到词义的词义判断过程。现有的词义消歧研究通常只关注同一词类下的多义项词的歧义消解问题，对语料库词义标注的完整过程分析不够，对读音、词类、搭配、义项频率分布等信息对歧义消解的作用和程度缺乏量化分析。本文的研究对基于词典的语料库词义标注有重要的参考作用。

第三章　读音区分词义

读音对区分词义的作用主要体现在对同形词（homographs）的区分上。本文采用广义多义词的概念（见图 1），同一词形具有多种词义可能的均视为多义词。同形词也属于词语多义的一种特殊情况。词典中，同形词和多义词的处理方式不同，同形词单列为不同的词目，分别进行释义；多义词在一个词目下列出多个不同义项。对词义标注来说，同形词和多义词都需要通过消歧过程辨别其词义。

同形词和多义词的联系非常紧密，涉及词语意义的历时演变和多义词的认知语义框架等多个研究课题(符淮青，2004)。从性质上看，同形词是书写形式相同的两个不同词；多义词是一个词具有多个不同的词义，几个词义彼此不同而相互关联（张博，2004，2008)。同形词和多义词在消歧方式上的最大分别就是大部分同形词可通过其读音来区分。

3.1 读音对区分词义的作用

相同书写形式的词，只要读音不同必定伴随着意义上的不同（李尔钢，2006)。例如，“好事”，读“hǎoshì”时意思是“好的事情；有益的事情”，做“爱管闲事”义时读作“hàoshì”，两词词形相同而读音、意义不同，是同形异义异音的两个词。有时候读音上差异也可能很小，例如意思是方位的“东西（dōngxī)”和泛指事物的“东西 dōngxi”，读音上的区别只是后一个字读轻声。同形异音现象在英语中也存在，例如单词 present，读['preznt]时意思是“礼物、礼品”，读[pre'znt]时意思是“赠送”，单词的词形完全一样，但发音不同、意义不同，是同形的两个词。从根源上看，同形的两个词读不同的音是语言的使用者从区分的需要出发，使用音变的方式对同一词形的不同意义进行区别的结果（张博，2005)。语音上变化是一种明显的标记，指出了这是同形的两个词，而不是一词多义关系。

汉语中部分词语存在同形异音异义现象。同形异音且词类相同的一组词(例如：转动/v#zhuǎndòng、转动/v#zhuàndòng)和同形异音但词类不同的一组词(例如：精神/n#jīngshén、精神/a#jīngshen)都被称作同形异音异义

词。读音区分词义的重点是根据读音区分同形异音异义词。

同形异音异义词可分为单音节和多音节两种。单音节同形异音词是同形异音异义现象的主体，如上文举过的“背（bēi/bèi）、长（cháng/zhǎng）、倒（dǎo/dào）”等。除由轻声产生的多音情况外，多音节同形异音一般都带有由单音节同形异音词充当的词素。多音节异音同形词是比较特殊的一类词，这类词语有两个或两个以上的音节，同时又有两个或两个以上的读音，并同时对应两个或两个以上的意义，如教学（jiāoxué）和教学（jiàoxué）、琢磨(zhuómó)和琢磨（zuómo）等等。再如，动词“转动”，在“电机正在转动”中，意思为“物体以一点为中心或以一直线为轴作圆周运动”，″转动″在这里应读成“zhuàndòng”，不能读成“zhuǎndòng”；在“他的腰部转动自如”中，意思是“身体某部分自由活动”，“转动”在这里只能读成“zhuǎndòng”，不可以读成“zhuàndòng”。

词的读音作为词的外在属性，与词义内容之间存在着相互作用。不同的语音形式会对词义的表达起到区别作用，例如普通话中的轻声现象。

轻声现象也反映了通过读音区分词义的作用。轻声是指在语音序列中有许多音节常常失去原有的声调，而读成一个又轻又短的调子。轻声现象构成是多音节同形异音词的重要因素。例如表 4 中的词在读轻声的情况与读原调词义不同。

表 4：轻声区分词义的词

序号	词	读音	《现汉》释义
1	地道	di4dao4	在地面下掘成的交通坑道（多用于军事）。
		di4dao5	①真正是有名产地出产的。②真正的；纯粹。③（工作或材料的质量）实在；够标准。
2	照应	zhao4ying4	配合：呼应
		zhao4ying5	照料
3	东西	dong1xi1	①东边和西边。②从东到西（距离）。
		dong1xi5	①泛指各种具体的或抽象的事物。②特指人或动物（多含厌恶或喜爱的感情）。
4	兄弟	xiong1di4	哥哥和弟弟

		xiong1di5	①弟弟。②称呼年纪比自己小的男子（亲切口气）。③谦辞，男子跟辈分相同的人或对众人说话时的自称
5	精神	jing1shen2	①指人的意识、思维活动和一般心理状态。②宗旨；主要的意义：领会文件的～。
		jing1shen5	①现出来的活力。②活跃；有生气。

读音还具有区分词类的作用。汉语有一部分单音节词可通过读音的不同区分其词类，如表 5 中的词。

表 5：读音区分词类的词

ID	词	读音 1	词类 1	读音 2	词类 2
1	长	chang2	形容词	zhang3	动词
2	称	chen1	名词	chen4	动词
3	冠	guan1	名词	guan4	动词
4	好	hao3	形容词	hao4	动词
5	乐	le4	动词	yue4	名词
6	丧	sang1	名词	sang4	动词
7	咽	yan1	名词	yan4	动词
8	中	zhong1	介词	zhong4	动词

从词义消歧的角度看，由于词典中多义词的不同词类分立为不同的义项，因此读音也就间接地区分了词义。

3.2 为语料库标注词语拼音

为语料库标注词语的拼音的目的是利用读音来区分同形词。

语料库拼音标注通过编写拼音标注程序自动标注实现。自动标注可以达到约 98%的准确率，难点是单字节多音词的拼音标注，如长（cháng）和长（zhǎng）。自动标注后采用人工对多音词进行校对的方式提高准确率。目前语料库拼音标注的准确率达到 99.9%以上。这为通过拼音区分同形词

打下了良好的基础。

拼音标注采用的格式是在词类标注后加“#”号作为引导符，然后再加具体拼音。标调的方式有两种，分别是传统的上标调方式（如“太阳/n#tài/yang”）和数字后标调（太阳/n#tai4/yang2）方式。数字标调方式中，阴平、阳平、上声、去声和轻声分别用数字 1、2、3、4、5 表示。本文中涉及的所有拼音都采用这两种标调方式。

经过拼音标注的语料如下所示：

• 传统上标调方式

我们/r#wǒ/men 吃/v#chī 过/u#guò 晚饭/n#wǎn/fàn ，/w 热气/n#rè/qì 已经/d#yǐ/jīng 退/v#tuì 了/y#le 。/w 太阳/n#tài/yang 落/v#luò 下/vd#xià 了/u#le 山坡/n#shān/pō ，/w 只/d#zhǐ 留下/v#liú/xià 一/m#yī 段/q#duàn 灿烂/a#càn/làn 的/u#de 红霞/n#hóng/xiá 在/p#zài 天边/s#tiān/biān 。/w

• 数字后标调方式

我们/r#wo3/men5 吃/v#chi1 过/u#guo4 晚饭/n#wan3/fan4，/w 热气/n#re4/qi4 已经/d#yi3/jing1 退/v#tui4 了/y#le5。/w 太阳/n#tai4/yang2 落/v#luo4 下/vd#xia4 了/u#le 山坡/n#shan1/po1，/w 只/d#zhi3 留下/v#liu2/xia4 一/m#yi1 段/q#duan4 灿烂/a#can4/lan4 的/u#de5 红霞/n#hong2/xia2 在/p#zai4 天边/s#tian1/bian1 。/w

拼音标注的核心问题是多音字、词的标注。统计发现，语料库中共出现 4958 个不同的汉字，其中多音字有 728 个，约占汉字数的 15%。若考虑频率情况，4958 个汉字的总出现频次是 1,707,656 次，728 个多音字的出现频次是 743,540 次，多音字占语料总字数的比例约为 43.54%。在词语层面，多音字的问题可以得到很好地解决。绝大部分多音字在词语中不多音，而多音词只占语料词数的很小的一部分。

多音问题在大部分情况下都不影响词义的识别和标注，词义标注是针对词形进行的，只通过词形即可直接对应获得词典的释义信息。拼音标注中的核心问题是多音词的识别和标注，因为多音词一定是同形词。

多音词可分为两种，即单字节多音词和多字节多音词。单字节多音词

占多音词的绝大多数，多字节多音词在语料库中所占比例不大，但对区分词义来说，这两类多音词都需得到重视，其最大的原因就是多音词往往都是高频词。典型的多音词及其在语料库中出现的次数如表 6 所示：

表 6：语料库中的多音词

ID	词	拼音 1	出现次数	拼音 2	出现次数
1	长	chang2	1921	zhang3	1852
2	背	bei1	155	bei4	191
3	倒	dao3	146	dao4	65
4	调	diao4	18	tiao2	11
5	转	zhuan3	659	zhuan4	91
6	大气	da4/qi5	26	da4/qi4	24
7	地方	di4/fang1	452	di4/fang5	362
8	精神	jing1/shen2	339	jing1/shen5	6

语料库中，多音词数量为 101 个词，209 词次。这 101 个多音词的总词频 88,407，占语料库总词频的 7.8%。从数量和词次来看，多音词是非常有限的。但如果考虑到出现次数，通过上述数据可以看出，多音词普遍高频，其将近 8%的总词频说明了通过拼音标注来区分词义的理论上的重要意义。

3.3 可通过拼音区分词义的词

根据对词义标注所使用的《现代汉语词典》第五版的统计，词典共收录同形异音词约 216 个。本文只考虑名、动、形三种词类的情况，因此读音不同的同形词数量要少一些，共计 84 个，占名动形总词数的约 1%。最多的是有 3 个读音的同形词，一共有 5 个词形，其他都是两读的同形词。因为有一部分多音词属于轻声和原调两读，两种读法皆可，不属于同形词范畴，例如，“出来、过、过去、来、去、起、起来、上来、下来”等。84 个读音不同的同形词中，在语料库中出现的语料中的约为 48 个。从音义关系来看，同一词形读不同音的目的是为区分其不同词义，因此，除少数读轻声和文白异读情况外，同形异音词的不同读音一定有不同的意义。本文

的研究在词语读音上以现代汉语词典为依据，不考虑文白异读问题，此外，也不计入小部分高频趋向动词，例如“来（lai2/lai5）”，可读轻声并且不区分词义的情况。语料库中可通过拼音准确区分的同形词及其出现次数如表 7 所示。

表 7：语料中可通过拼音区分词义的同形词

ID	词形	拼音	频次	义项数
1	挨	ai2	0	3
		ai1	64	1
2	扒	ba1	31	4
		pa2	0	3
3	背	bei4	191	6
		bei1	155	2
4	绷	beng4	0	1
		beng1	14	4
5	长	chang2	577	2
		zhang3	557	2
6	场	chang2	0	1
		chang3	6	1
7	大气	da4/qi5	26	3
		da4/qi4	24	2
8	当	dang4	0	4
		dang1	279	3
9	倒	dao4	65	3
		dao3	146	4
10	地方	di4/fang5	3	2
		di4/fang1	811	3
11	调	diao4	18	6
		tiao2	11	2
12	钉	ding4	0	2
		ding1	19	2
13	东西	dong1/xi5	8	2
		dong1/xi1	8	1
14	分	fen4	0	1
		fen1	229	4
15	供	gong4	0	2
		gong1	73	2
16	和	he2	6	1
		he4	10	2
17	降	jiang4	23	2
		xiang2	0	2
18	角	jiao3	51	3
		jue2	0	2
19	尽	jin4	126	3
		jin3	0	1
20	禁	jin4	4	1
		jin1	0	1
21	精神	jing1/shen5	6	3
		jing1/shen2	339	2
22	卷	juan3	70	3
		juan4	6	3
23	卡	ka3	3	3
		qia3	3	4
24	开通	kai1/tong5	4	2

		kai1/tong1	4	2
25	看	Kan4	3683	6
		kan1	8	2
26	搂	lou3	41	1
		lou1	0	3
27	落	la4	0	3
		luo4	400	6
28	闷	men4	9	1
		men1	21	4
29	蒙	meng2	39	2
		meng1	0	2
30	抹	ma1	3	2
		mo3	50	3
		mo4	0	2
31	拧	ning2	27	2
		ning3	27	3
32	耙	ba4	0	2
		pa2	1	2
33	刨	bao4	0	2
		pao2	5	1
34	劈	pi3	0	3
		pi1	43	4
35	片子	pian4/zi5	0	2
		pian1/zi5	6	3
36	漂	piao3	0	2
		piao1	39	2
37	切	qie4	0	2
		qie1	70	2
38	撒	sa3	97	2
		sa1	1	2
39	散	san4	69	3
		san3	0	1
40	挑	tiao3	1	5
		tiao1	159	4
41	吐	tu4	0	2
		tu3	71	3
42	应	ying4	0	2
		ying1	155	2
43	扎	za1	64	1
		zha1	0	2
44	炸	zha2	45	1
		zha4	0	2
45	占	zhan4	59	2
		zhan1	0	1
46	转	zhuan4	72	2
		zhuan3	112	2
47	着	zhao1	0	1
		zhao2	246	4
		zhuo2	0	1
48	重	chong2	1	2
		zhong4	168	4

表 7 列出了语料中出现的 48 个同形异音词。这 48 个词（词形）的总频次为 9672 次，平均每词（词形）出现约 203 次。从平均频次来看，这部分同形异音词都属于语料中的高频词。语料中的大部分同形异音词不同读

音的分布出现很大差异，其中一个读音高频，另外的读音低频是普遍现象。频次最高的词是动词“看”，读 kàn 时出现 3683 次，读 kān 时出现 8 次（读 kān 但不单独成词的不算），不同读音的出现次数差距悬殊。又如，词形“地方”，读 dìfāng 时出现 811 次，而读 dìfang 时只出现 3 次；词形“卷”，读 juǎn 时出现 70 次，读 juàn 时只出现 6 次。同形异音的两个词使用频率最接近的是“长”，读 cháng 时出现 577 次，读 zhǎng 时出现 557 次。

从义项数量来看，48 个同形异音词的现代汉语词典义项总数是 240 个，平均每个词的义项数约为 4.9 个。例如，词“背”读 bèi 时有 6 个义项，读 bēi 时有 2 个义项。现代汉语词典中的多义名词、动词和形容词总数为 8263 个，这些词的平均义项数是 2.33 个。相比之下，可见同形异音词大多为意义复杂的多义词。

3.4 通过拼音区分后的单义词

对区分同形异音词词义而言，拼音有效降低了同一词形下的义项数。

对其中的一部分词来说，经拼音区分后，可完全确定义项。所有读音下都是单义，即该词形有多个读音但每个读音下只有一个义项，例如，场（chǎng/cháng）、禁（jìn/jīn）等。

对另一部分词，确定读音后，可以使得其中一个读音下不再有多义项，成为单义。表 6 中的 48 个多音词，有 15 个词的其中一个读音下只有一个义项，成为经拼音区后的单义词。例如，挨（āi）、分（fèn）、闷（mèn）、散（sǎn）、着（zhuó），见表 8。

表 8：拼音区分后的单义词

ID	词形	拼音	频次	义项数
1	挨	ai1	64	1
2	东西	dong1xi1	8	1
3	和	he2	6	1
4	禁	jin4	4	1
		jin1	0	1
5	闷	men4	9	1
6	刨	pao2	5	1
7	散	san3	0	1
8	扎	za1	64	1
9	炸	zha2	45	1
10	占	zhan1	0	1

3.5 小结

在本文的研究中，出于对拼音区分的同形词在性质上属于不同词的考虑，首先说明读音对区分词义的作用。但在词义标注过程中，优先利用词类信息进行消歧。当通过拼音和词类可以到同样的消歧效果时，词类优先，因此真正通过拼音信息进行消歧的多义词少于理论上能通过拼音消歧的多义词的数量。

对语料库词义标注来说，以读音作为消歧线索可以有效的降低词义消歧的复杂度。多音词多是高频词，101 个多音词的总词频约占语料库 8%，体现了利用读音消歧的潜力。本文只标注名动形等实词的词义，通过读音消歧的同形异音的实词只占词频的 2%。

通过对语料中同形异音词的读音和义项情况可以看出，拼音只是降低了词义的区分的复杂度，对大部分同形词而言，拼音并不能唯一确定它们的词义。语料中的 48 个同形异音词中，只有 18 个词在确定读音的情况下不再多义。

然而，尽管拼音对多义词词义消歧和标注的总体贡献有限，但从音形义的关系和词义消歧出发，对拼音区分词义的能力和性质做出量化分析是有价值的，它帮助我们建立起从词形到读音再到词义的完整的区分系统。多义词的词义消歧是一个极为复杂的问题,拼音作为词语的外在形式之一，对降低词义区分复杂度的作用应该被肯定和利用。

第四章　词类区分词义

英语词典具有标注词类的传统，汉语词典中的词类标注则起步较晚。二十世纪八十年代以前出版的汉语语文词典，一般都没有标注词性，特别是实词的词性。这个问题在近年有所改变，2004 年出版的《现代汉语规范词典》和 2005 年出版的《现代汉语词典》（第五版）都推出了带词类标注的版本，采取按词性分立义项的做法，词性不同则分为不同的义项。对词义消歧研究来说，词典怎么给多义词标注词类是一个需要考虑的重要问题，关系到多义词的认定。若词典按分词类立义项的方式处理词类标注，那么多义项即是需要消歧的多义词；若词典将词类归到释义的下位，那么认定多义词就不能只靠义项数，还需再判断义项下的词类情况。

本文研究所参照的《现代汉语词典》采用为不同词类单立义项的处理方式，因此本文中的多义词就同时包括语法功能相同或不同的多义情况。

4.1 词类对区分词义的作用

与以前的版本不同，《现代汉语词典》第五版为义项标注词类，这就意味着一个兼类词有几个词类就至少有几个义项，即多词类就多义项。（苏宝荣，2002；李尔钢，2006；周荐，2007）例如，“规范”兼有名动形三个词类，现代汉语词典分别对三个词类进行释义；“幸福”可做名词和形容词，现汉的两个义项分别对应名词和形容词。

这种通过词类来区分义项的词在现代汉语词典的多义词中大约占 20%。从词义消歧的角度看，词类区分在这种类型的多义词中效果最为明显，部分多义词通过词类区分，因在同一词类下不存在多个义项，从而成为单义词。现代汉语词典中，经词类区分后的单义词如表 9 所示：

表 9：词类区分后的单义词

ID	多义词	词类	义项数
1	安定 āndìng	形容词	1
		动词	1
2	安慰 ānwèi	形容词	1
		动词	1
3	帮工 bānggōng	名词	1
		动词	1
4	保守 bǎoshǒu	形容词	1

		动词	1
5	倡议 chàngyì	名词	1
		动词	1
6	陈设 chénshè	名词	1
		动词	1
7	道德 dàodé	形容词	1
		名词	1

8	丰富 fēngfù	形容词	1
		动词	1
9	机密 jīmì	形容词	1
		名词	1
10	杂 zá	形容词	1
		动词	1

词典为不同词类设立不同义项对词义消歧是非常有利的。词类作为语法功能的分类，反映了词语在语言使用上的情况。这种反映在语言使用上的特点可以用于词义的识别。

词典释义中体现了对词的语法功能的解释，因此同样作为多义词，不同的词具有不同的多义上的性质。根据多义词的多个义项在语法语义上的不同性质,语料库中的多义词按照现代汉语词典释义可分为如下三种类型：

第一种类型的多义词是不受语法功能影响的多义词。这部分词只有一个词类，并且在该词类下多义，表现为其多个义项都属同一词类。例如，多义词“计算”有 3 个词典义项，且均为动词，其词义中没有语法功能带来的词义。其他例词见表 10。在语料库中，这部分多义词占主要部分，总数是 4242 个，占全部多义词（5406）的 97%。

表 10：第一种类型的多义词

词	词性	义项数	词	词性	义项数
矮	形容词	3	饱满	形容词	2
爱	动词	4	变	动词	3
把戏	名词	2	气候	名词	3

第二种类型的多义词的多义受语法功能的影响。这部分词虽有多个义项，但各义项分属不同的词类，每个词类下都不具有多义项，例如“规范”有 3 个词典义项，义项①是“标准”作名词用，义项②是“合乎标准”，作形容词用，③是“使合乎标准”，作动词用。三个义项分属名、动、形三个

不同词类，但在其可充当的任一语法功能下不存在多义。其他例词见表 11。

表 11：第二种类型的多义词

词	词性	义项数	词	词性	义项数
矮	形容词	3	饱满	形容词	2
爱	动词	4	变	动词	3
把戏	名词	2	气候	名词	3

第三种类型的多义词是复合的多义词，这部分词有多个义项，其中一部分义项属同一词类，而同一词类下也存在多义项，兼有前两种类型多义词的特征。其他例词见表 12。

表 12：第三种类型的多义词

词	词性	义项数	词	词性	义项数
矮	形容词	3	饱满	形容词	2
爱	动词	4	变	动词	3
把戏	名词	2	气候	名词	3

词义消歧研究的重点是同一词类下多义问题，若同一词类下不存在多义，则可通过词类来直接消歧。因为判断一个词的语法功能要比识别其词义容易的多，所以确定词类下不多义的词不需要通过消歧算法来判断词义。上述的第二种类型的多义词是最典型的可通过词类区分词义的词，第三种类型的多义词也有一部分在确定词类下是单义的，识别出这部分词对降低词义消歧的难度，提高消歧总体正确率有重要帮助，体现了通过词类区分词义对词义消歧和标注的重要作用。

4.2 通过词类区分词义和语素义

词典的词类标注有助于进行词义和语素义的区分。因为只有能够单独使用的词才有词类这个性质，不能单用的语素就不具备词类属性。词典中单字字头的释义部分通常同时包括语素义和词义。通过词类标注可以有效区分出其中的词义，忽略掉不能单独使用的语素义，减少义项的数量，从

而极大降低词义消歧的复杂程度。

通常所说的语素义是指在合成词好固定结构中语素的意义（包括成词语素和不成词语素）。符淮青（2004）指出词典要分辨词义和语素义，词义是词可独立运用的意义，而语素义只能存在于它所构成的词或语中。从词典释义看，单字（词）的意思可能为语素义也可能为词义。有的字表示一个意思，这个意思只是语素义，如“永”；也可以同时是语素义和词义，如“大”；有的字表示多个意思，这些意思都是语素义或词义，也可以同时是语素义或词义。

《现代汉语词典》考虑了词义与语素义的不同，词义标注其词类，而语素义则没有词类标记，例如“民”有五个义项，除第 5 个义项做姓氏时标注为名词外，其他四个义项都是语素义，不标注词类。

在词义标注的过程中区分词义和语素义是非常必要的。用于词义标注研究的语料通常是经过了词语切分和人工校对的，已经经过了是否为词的判断，特别是本文研究所用的语料库在进行词语切分和标注后，还专门根据现代汉语词典的情况专门处理了词与非词的区分问题，除通过词类标记（语素标为 g）明确表示的情况外，语素不会作为独立的词在语料中存在，可以保证语料中带词类标记的分词单位都可以作为词来处理。在词义标注中，真正需要消歧对象的是词义，语素义将在词或语之中自动得到消解，不需要专门处理。

对词义标注而言，词类标注对词与语素的区分有重要意义，大幅度减少了多义词或义项的数量，从而降低了词义消歧的复杂程度。例如，“明”在《现代汉语词典》中有 9 个义项，其中只有义项②“明白、清楚”和义项⑨“明明”标注词类，两个义项分别可作形容词和副词，其他的 7 个义项都是语素不能作为词使用。这样，由于词义消歧以词为单位，消歧过程中不需要考虑词的语素义，从而减少了可能的义项数。上例中，“明”在语料中作为一个词出现时，判断其词义，只需考虑它在上下文中是义项②和⑨中的那一个，而不需考虑其他 7 个语素义。

4.3 为语料库标注词类信息

词类信息对词义消歧用重要的意义，对大部分词义消歧系统来说，词类信息都是不可缺少的（Wilks & Stevenson，1998）。目前词类的自动标注已经可以达到很高的准确率（汉语 95%以上，英语可以更高）。本文研究所使用的语料库的词类标注进行自动词类标注后，还经过高质量的人工校对，准确率达到要求。为了在词的层面和现代汉语词典取得对应，在词语切分层面语料库还专门进行了一致性处理。

语料库词类标注所使用的词类标记集如表 13 所示。

表 13：语料库词类标记集

ID	词类标记	词类名称
1	a	形容词
2	Ag	形容词性语素
3	b	区别词
4	c	连词
5	d	副词
6	Dg	副词性语素
7	e	叹词
8	f	方位词
9	g	语素
10	h	前接成分
11	i	习用语
12	j	缩略语
13	k	后接成分
14	l	习用语
15	m	数词
16	Mg	数词性语素
17	n	名词
18	Ng	名词性语素
19	nr	人名
20	ns	地名
21	nt	机构团体名
22	nz	专名
23	o	拟声词
24	p	介词
25	q	量词
26	r	代词
27	s	处所词
28	t	时间词
29	Tg	时间性语素
30	u	助词
31	v	动词
32	Vg	动词性语素
33	x	其他
34	y	语气助词
35	z	状态词

本文涉及到所有词类标记都遵守表 13 中的代码规范。本文研究的对象

——语料库实词的词类标记分别为：名词（n）、动词（v）、形容词（a），形容词还包含区别词（b）和状态词（z）两个小类。

4.4 兼类造成多义在多义词中的比例

《现代汉语词典》第五版中共有多义的名词、动词、形容词 8263 个，其中兼类（至少两个词类）的约 2085 个，占多义实词总数的 25.23%，可见兼类是多义词的主要特征之一。现汉中的实词兼类情况如表 14 所示：

表 14：现汉中的实词兼类情况

ID	兼类	词数	例词
1	名、动、形	56	错、规范、活动、麻烦、讲究
2	名、动	1231	帮、备份、编辑、代表、负担
3	名、形	389	单、本分、典型、高明、卫生
4	动、形	409	多、保守、充实、端正、团结

从表 14 可以看出，同时兼有名动形三个词类的词是实词中的少数（2.69%），最典型的兼类情况是名词、动词兼类，占全部兼类词的 59.04%，超过一半。而名词—形容词、动词—形容词兼类的数量大致相当，各占总数的 19%左右。

兼类词的不同词类在使用频率上存在很大差异，这通过语料调查可以清晰看出。例如，“规范”可做名动形三个词类，三个词类在语料中的出现比例分别为“名词 38%、动词 25%、形容词 38%”，各个词类的出现频率相差不大；而“圆”也可做名动形三个词类，三个词类在语料中的出现比例分别为“名词 4%、动词 8%、形容词 88%”，各个词类的出现频率相差巨大，形容词是其最主要词类。详如表 15 所示。由于不同词类意味着不同词义，由此也可以看出多义词的不同义项在文本中的实际分布也是很不均衡的。

表 15：语料中主要兼类词不同词类的使用频率

ID	词	频次	词类		
			名词	动词	形容词

			次数	比例(%)	次数	比例(%)	次数	比例(%)
1	革命	165	8	4.9	155	93.9	2	1.2
2	规范	24	9	37.5	6	25.0	9	37.5
3	花	780	637	81.7	95	1.2	48	6.2
4	活动	332	214	64.5	11	3.3	5	1.5
5	卷	72	6	8.3	64	88.9	2	2.8
6	麻烦	44	7	15.9	5	11.4	32	72.7
7	圆	111	4	3.6	9	8.1	98	88.3
8	生活	1092	632	57.9	460	42.1		
9	画	627	196	31.3	431	68.7		
10	工作	597	157	26.3	449	75.2		
11	希望	408	133	32.6	275	67.4		
12	科学	334	295	88.3			39	11.7
13	自由	195	49	25.1			146	74.9
14	理想	164	118	72.0			46	28.0
15	困难	133	45	33.8			88	66.2
16	灰	42	18	42.9			24	57.1
17	破	231			102	44.2	129	55.8
18	丰富	185			31	16.8	154	83.2
19	感动	94			74	78.7	20	11.3
20	充实	29			16	55.2	13	44.8

4.5 词类区分词义在语料库中的表现

通过统计，本文使用的 200 万字的教材语料总词数为 52,798 个，其中共有 37,603 个实词（名动形）。实词中在《现代汉语词典》中有多个义项的多义词共有 5,514 个，约占总数的 15%。从频率方面看，全部实词的总词次是 717,566 次，其中多义实词的总出现次数是 308,008 次，约占总词次的 43%。可以看出多义词虽然数量在语料词语中不占多数，但出现频率

却很高。语料中超过 40%的实词是多义词。

语料中通过拼音和词类能够唯一确定义项的词语约为 1301 个词，占多义词总数的 23.7%；考虑词语出现频率，这 1301 个词的出现次数总和约为 47,356，占多义词总出现次数的 15.3%。表 16 列出了部分经词类区分后的单义词。

表 16：语料中词类区分后的单义词

ID	词	读音	词类	语料频次	义项号
1	爱好	ai4/hao4	n	20	2
2	安定	an1/ding4	a	12	1
	安定	an1/ding4	v	5	2
3	安慰	an1/wei4	a	3	1
	安慰	an1/wei4	v	46	2
4	把握	ba3/wo4	n	3	3
6	包裹	bao1/guo3	n	9	2
	包裹	bao1/guo3	v	4	1
7	包装	bao1/zhuang1	n	4	2
8	保险	bao3/xian3	a	1	2
	保险	bao3/xian3	n	6	1
9	比喻	bi3/yu4	n	52	1
	比喻	bi3/yu4	v	22	2
10	报道	bao4/dao4	n	37	2
	报道	bao4/dao4	v	52	1

4.6 小结

词类是降低词义消歧的复杂度的最为有效的消歧线索。词类消歧的专门研究，通常只以同一词类下的多义作为消歧目标。本文从词义标注出发，关注从词形到词义的词义识别整体过程，对通过词类进行词义消歧的可能性进行了量化分析。

研究发现，通过词类标注，我们可以大幅度缩小多义词的词义范围。在词类信息的帮助下，对一部分多义词，可以减少其义项的数量；对另一部分词来说，通过词类可以完全判断其词义。语料库中有 15.3%的多义词经过词类区分后可确定词义，这体现了词类信息对词义消歧的重要作用。

第五章　搭配区分词义

搭配信息在自然语言处理中的重要性越来越受到重视。在词义消歧过程中，搭配信息具有重要的意义，几乎所有的词义消歧系统都通过各种方式利用搭配进行多义词的歧义消解。

5.1 通过搭配区分词义

词义消歧的任务是为多义词选择合适的义项，而一词多义常常与词语的搭配不同有关。Leech（1974）指出一个词会因经常与这个词同时出现的一些词的意义产生出搭配意义。Firth（1957）提出“观其伴，知其义”[19]的思想，说明词的意义与其搭配词是紧密相关的。

在利用搭配区分词义方面，Yarowsky（1993）提出词义消歧研究中的一个重要假设：One sense per collocation。这个假设认为多义词在搭配中总是保持同样的意思，因此可以通过搭配确定词义。林杏光（1999）提出“在信息处理中，要解决多义词的义项选择问题，必须通过词语搭配”。在现有的词义消歧方法中，搭配信息都扮演了重要角色，是最常用的消歧线索之一。

计算语言学研究中，搭配倾向于定义为一组同现频度高的词语组合。比较典型的是 Benson 等人（1989）对搭配作的定义：搭配是一种任意的、可重复出现的词语组合，具有如下四个主要属性：1）搭配是重复出现的，偶然出现的词语组合不是搭配；2）搭配是任意的，可能是自由的也可能是不自由的；3）搭配通常具有固定的结构；4）搭配与领域有关，例如某些专业领域的习惯用语。

从词义消歧的角度看，搭配应满足两个条件：1）组合中的词义是确定的；2）词语组合应表现出高频的特点。具备条件1是搭配的必要条件，不能确定词义的搭配无法用于以区分词义为目标的词义消歧研究。具备条件2才能使利用搭配区分词义具有可操作性。

通过搭配区分词义必须解决搭配信息的获取问题。目前，搭配信息的

[19] You shall know a word by the company it keeps.

获取途径主要有两种，一种是利用搭配词典，另一种是通过计算机程序自动抽取。词义消歧任务的性质要求搭配信息必须是分义项给出的。根据对汉语搭配词典的调查，现有的搭配词典中只有极少数是分义项给出搭配的，如本文所利用的《现代汉语搭配词典》。搭配词典用于词义消歧的另外一个条件是必须采用类似的义项体系。可见，利用现有搭配词典作为搭配信息来源的困难。此外，人工编纂的搭配词典在搭配词选择上主观性明显。Smadja（1994）对 Oxford English Dictionary 进行分析，认为词典中人工编辑的搭配的正确率只有4%。因此，通过计算机程序提取搭配信息是词义消歧领域常用的一种方法。在汉语搭配信息的自动抽取方面，孙茂松等（1997）的研究具有重要参考价值。

搭配信息的自动抽取通常是通过观察词语一定范围的前后文进行。Martin（1983）提出："统计测试表明，95%以上的搭配信息可以通过考察-5和+5范围内的词获得"。孙宏林（1998）认为不同语言（如英语、汉语等）、不同词类（名动形等）的由于语义重心的不同，观察窗口也应当有所区别。他通过语料库抽取并统计了名词、动词和形容词的搭配词语的分布情况，得出这三类词的搭配词语的最佳观察窗口是名词[-2, +1]、动词[-3, +4]、形容词[-1, +2]。

本文研究中，搭配的抽取采用计算两个词的互信息的方法。根据 Church（1991）的研究，任意两个词之间关系的疏密程度可以用它们之间的互信息（Mutual Information）来计算，计算公式如下：

$$mi(w, w_i) = \log_2 \frac{p(w, w_i)}{p(w)p(w_i)}$$

其中，w 和 w_i 是任意两个词，$p(w, w_i)$ 是是两个词在上下文中共现的概率；p(w)和 p( )分别是 w 和 w_i 分别是两个词各自独立出现的概率。可以看出，互信息公式的基本含义是两个词一起出现的概率和各自单独出现的概率的比值。互信息值越高说明两个词越趋向于一起出现，从而形成搭配。互信息可用于说明上述 Benson 搭配定义的前两个属性，即搭配是重复出现的和搭配既可能是自由组合也可能是约束组合。

5.2 不同义项的搭配词实例分析

前文指出，利用搭配区分词义的前提条件是多义词不同义项具有不同搭配。本文尝试通过词义标注语料分析义项之间可通过搭配进行词义指示的可能性和区分程度。首先进行多义词的不同义项的搭配抽取，然后通过分析不同义项的可搭配词语来判断搭配对词义的区分能力。由于语料库规模有限，再加上多义词不同义项的频率分布有明显的差异，所以只能选取少数典型词语做实例分析。

搭配抽取采用互信息方法，观察窗口采用孙宏林（1998）提出的以名词[-2, +1]，动词[-3, +4]，形容词左边[-1, +2]作为观察窗口。

下文对多义词“发现、准备”等的搭配情况做具体分析。

1）多义动词“发现”的搭配信息分析

根据《现代汉语词典》第五版，“发现”只有动词词性，有两个义项，分别为：①“经过研究、探索等，看到或找到前人没有看到的事物或规律：～新的基本粒子丨有所发明，有所～，有所创造”。②“发觉：这两天，我～他好像有什么心事。”

通过互信息公式分别对“发现”的两个义项以[－3，+4]为观察窗口，计算其与上下文词语的互信息，取频次大于 3，互信息大于 2 作为条件。搭配词例举如下：

•“发现”的义项①左右窗口部分搭配词及互信息

ID	词	位置	MI
1	斯石英/n	[+]	5.90
2	外形/n	[+]	5.51
3	夸克/n	[+]	5.17
4	发明/v	[+]	4.84
5	化石/n	[+]	4.80
6	善于/v	[-]	4.63
7	规律/n	[+]	4.51
8	从中/d	[-]	4.51
9	鲸/n	[-]	4.43

ID	词	位置	MI
10	科学家/n	[-]	4.17
11	思考/v	[+]	3.49
12	观察/v	[-]	3.45
13	科学/n	[-]	2.92
14	研究/v	[-]	2.87
15	新/a	[+]	2.51
16	人类/n	[+]	2.40
17	新/a	[-]	2.28
18	发现/v	[+]	2.16

注：[-]表示在左边搭配词，[+]表示在右边的搭配词。下表同。

•“发现”的义项②左右窗口部分搭配词及互信息

ID	词	位置	MI
1	放羊/v	[-]	6.59
2	醒来/v	[-]	4.98
3	草丛/n	[+]	4.94
4	不少/a	[+]	4.88
5	目标/n	[+]	4.77
6	教科书/n	[-]	4.76
7	忽然/d	[-]	4.16
8	花瓣/n	[+]	4.01
9	事物/n	[-]	3.62
10	旁边/f	[+]	3.58
11	这时/r	[-]	3.57
12	里面/f	[+]	3.56
13	突然/a	[-]	3.54
14	原来/t	[+]	3.51
15	羊/n	[+]	3.48
16	容易/a	[-]	3.33
17	才/d	[-]	3.31
18	敌人/n	[-]	3.06
19	前面/f	[+]	3.06
20	时候/n	[-]	3.05
21	后来/t	[-]	2.78
22	自己/r	[+]	2.78
23	那里/r	[+]	2.65
24	眼睛/n	[-]	2.59

通过动词“发现”的搭配词及互信息表，可以清楚的观察到三个重要现象：

•“发现”的两个不同词义之间的搭配词有着很大的不同。义项①的搭配词具有明显的特点，体现为自然科学名词、重要地理文化名称、专业术语的互信息较高；而义项②的搭配词明显更为生活化，体现为日常生活名词、常见地点的互信息较高。这种搭配上的不同可以验证搭配可区分词义的观点。

•“发现”的不同词义搭配词的语义指向性不同。义项①的搭配词具有较为明确的语义指向性，具体表现为容易对搭配词进行分类，从而归纳出搭配词的特征。义项②的搭配词明显没有的类别性，不容易总结其规律。由此可以看出，多义词的不同词义可由搭配特征指示的程度不同，有些词义容易通过搭配进行识别，有些搭配不易通过搭配识别。

• 从不同义项的搭配词来看，可以验证 Yarowsky 所指出的“One Sense

Per Collocation”的论断。两个义项的搭配词，除虚词外，很少有重叠的情况，这说明不同的词义倾向于有不同的搭配。

2）多义动词“准备”的搭配信息分析

根据《现代汉语词典》第五版，动词“准备”共有两个义项，分别是：①“预先安排或筹划：精神～｜～发言提纲｜～一个空箱子放书。”，②“打算：春节我～回家｜昨天我本来～去看你。”

对“准备”的两个义项以[－3，+4]为观察窗口分别计算其与上下文词语的互信息。互信息较高的词列表如下：

•“准备”的义项①左右窗口搭配词及互信息

ID	词	位置	MI
1	就绪/v	[+]	7.89
2	待命/v	[+]	7.45
3	出征/v	[+]	6.76
4	随时/d	[-]	6.28
5	事先/d	[-]	6.14
6	迎接/v	[+]	5.84
7	发言/v	[+]	5.66
8	探究/v	[-]	5.58
9	精心/a	[-]	5.47
10	随时/d	[+]	5.43
11	命令/n	[-]	5.22
12	分钟/q	[-]	5.11
13	任务/n	[-]	4.89
14	工作/n	[+]	4.63
15	作/v	[-]	4.61
16	客人/n	[-]	4.49
17	分钟/q	[+]	4.13
18	准备/v	[-]	3.75
19	准备/v	[+]	3.75
20	做/v	[-]	3.68
21	工作/v	[+]	3.60
22	一切/r	[-]	3.52
23	注意/v	[-]	3.45
24	好/a	[-]	3.35
25	好/a	[+]	3.22

•“准备”的义项②左右窗口搭配词及互信息

ID	词	位置	MI
1	回国/v	[+]	5.41
2	半夜/n	[+]	5.10
3	离开/v	[+]	4.37
4	举行/v	[+]	4.23
5	国/n	[+]	4.09
6	回家/v	[+]	4.08
7	正/d	[-]	3.87
8	回去/v	[+]	3.85
9	学生/n	[-]	3.57

10	出去/v	[+]	3.56
11	这儿/r	[+]	3.40
12	送/v	[+]	3.21
13	边/n	[-]	3.14
14	当/p	[-]	3.04
15	忽然/d	[+]	3.01
16	于是/c	[-]	2.87

17	往/p	[+]	2.48
18	笑/v	[-]	2.35
19	再/d	[+]	2.13
20	做/v	[+]	2.07
21	时候/n	[+]	2.06
22	去/v	[+]	1.73
23	到/v	[+]	1.39

通过动词“准备”的搭配互信息表可以观察到：

•“准备”的两个不同词义之间的搭配词的区分比较明显，但是很难找到简单的规则将二者进行有效的区分。义项①“预先安排或筹划”和义项②“打算”从语义上看非常接近，这也体现在二者的搭配上。例如，“准备+出国”这样的搭配，我们很难通过搭配词直接判断搭配的义项，需要更多的上下文信息进行辅助选择。

•从语法的角度看，“准备”的义项②“打算”后面基本上只能跟动词，但通过基于统计的搭配抽取方式很难实现。也就是说，基于统计的搭配信息对动词“准备”的词义确定不易取得好的效果。另一方面，人工编制的《现代汉语搭配词典》中“准备”的搭配词，也呈现了和统计方法类似的情况，义项①词典给出了 31 个搭配词，如“～考试、～发言”等，义项②词典只给出了 4 个非严格意义的搭配，分别是“～明日启程、～回家过年、～给他写封信、～申请出国”。这也从某种意义上说明，义项②很难通过其搭配词区分出来。

通过对“发现”、“准备”等词的分析，可以认为搭配对区分词义具有明显的作用。但是存在相当一部分多义词或义项不能通过搭配进行有效区分。

5.3 可区分词义的搭配信息类型

根据对词义标注语料的分析，可用于区分词义的搭配信息主要可以分为下面几种类型。

1）实词类搭配词

通常所有的搭配词都是指实词。下文给出了人工标注的多义动词“上”的部分实词搭配。

上/v shang4	
①由低处到高处	～山、～楼、～车
②到；去(某个地方)	～街、～工厂、～单位、～哪儿、～那儿
⑥把饭菜等端上桌子	～饭、～菜、～茶、～汤
⑧把一件东西安装在另一件东西上	～刺刀、～螺丝、～子弹
⑩登载；电视上播映	～报、～电视、～电台、～报纸、～头条
⑿到规定时间开始工作或学习等	～班、～课、～工、～学

2）虚词类搭配词

虚词搭配更接近于语法的性质。对利用搭配来区分词义，虚词搭配往往可以取得好的效果，因为虚词常常高频，可以达到更高的文本覆盖率。其缺点是只有少部分词具有不同义项具有不同的虚词搭配这个特点。虚词搭配具体有下面几种类型：能不能加着、了、过；能不能加的、地、得；能不能加数量词等。

以动词“长（zhǎng)”为例，《现汉》释义如下：

① 生：～锈｜山上～满了青翠的树木。

② 生长；成长：杨树～得快｜这孩子～得真胖。

③ 增进；增加：～见识｜～力气｜吃一堑，～一智。

通过语料分析发现，只有义项②后面可以加“得”，义项①和③不能；义项①和②后面可以加助词“着”，义项③不能；义项①和③可以加助词“了”，义项②不能；只有义项①后面可以加介词“在”，义项②和③不能。

可见通过对这种虚词性搭配的判断可以很大程度上区分词的不同义项。

3）词类作为搭配信息

词类作为搭配信息的是将语法功能相同的一类词作为搭配对待。由于多义词的不同义项在使用上具有不同特点，词类作为搭配信息义项的判断

有很好的作用。

例如，多义动词“爱/v”有四个动词义项，词类对义项有很好的区分作用。“爱/v”的《现汉》释义如下：

① 对人或事物有很深的感情：～祖国丨～人民。

② 喜欢：～游泳丨～劳动丨～看电影。

③ 爱惜；爱护：～公物丨～集体荣誉。

④ 常常发生某种行为；容易发生某种变化：～哭丨铁～生锈。

通过语料分析发现，义项①和③后面可以跟名词，义项②和④不能加名词；义项②和④后面可以跟动词，义项①和③不能；义项④后面可以跟形容词，其他三个义项都不能。

目前的词义消歧研究中，对搭配信息的利用一般局限在实词类搭配上。从语料数据来看，虚词类搭配和词类也具有突出的作用，应当得到重视。

5.4 不同义项的搭配词重叠程度比较

搭配区分词义的一个重要前提就是多义词的不同义项具有不同的搭配，不同义项具有不同搭配的多义词才具有通过搭配区分词义的性质；不同义项在搭配词语上具有相似性或有大量重叠的，很难通过搭配来区分词义。下文将通过多义词的不同义项的搭配词的重叠情况分析这一点。

上文已经对多义词“发现”的两个义项的搭配进行了抽取和分析。下面通过比较搭配词的重叠程度来看是否两个不同义项有不同的搭配及其重叠程度。两个义项都取互信息为 5 以上的搭配，不限定出现频率，不区分搭配词在前或后的情况，义项①有 141 个搭配词，义项②有 188 个搭配词。通过比较发现义项①和义项②的搭配词只有两个相同，都是副词，分别是“及早/d”和“无意中/d”。从这个比较可见发现的义项①和义项②的搭配词具有高度的异质性。因此可以判断，多义词“发现”的两个义项之间的具有高区分度。

通过多义词“发现”不同义项搭配词的比较， Yarowsky 提出的“One Sense Per Collocation”在语料库中是可被验证的，“发现”的不同搭配词往往都对应不同的词义。但对另外一些多义词，例如“表 26：不同义项上下

文相似度高的多义词”的词，搭配信息在自动消歧中的作用非常有限。

5.5 利用搭配区分词义的局限

本文研究过程中，主要利用《现代汉语词典》词语释义的例证部分给出的搭配和少量人工专门制作的分义项的搭配词表进行消歧，对搭配信息的利用程度不高。下文尝试通过《现代汉语搭配词典》中提供的搭配信息来说明搭配能够多大程度用于词义消歧。具体的检测方法是计算搭配词典给出的搭配词在语料库中出现的次数及比例。

仍以多义词“发现”为例，《现代汉语搭配词典》中搭配的两个义项给出的搭配词分别是：

【义项 1】：

——前搭配词：突然、偶然、无意中、最近、早已

——后搭配词：油田、新矿、新大陆、新的行星、野人、猿人遗址、运动规律、元素周期律

【义项 2】：

——前搭配词：及时、最近、善于

——后搭配词：敌人、问题、毛病、情况、线索、秘密、他有心事、别人的优点、反常的现象

搭配词典中“发现”的两个义项的搭配除“最近”一词相同外，其他都不相同。

搭配能否起到有效的区分词义作用的要素之一是搭配词应当在相关语料中高频，否则由于覆盖率有限，消歧的效果也就有限。表 17 给出了“发现”的搭配词在含“发现”的语句中的出现情况：

表 17：“发现”的搭配词在语料库中出现的情况

ID	搭配词	位置	频次
1	突然	［前］	14
2	偶然	［前］	3
3	无意中	［前］	3
4	最近	［前］	2
5	早已	［前］	3
6	油田	［后］	1
7	新矿	［后］	0
8	新大陆	［后］	0
9	新的行星	［后］	0

10	野人	［后］	0
11	猿人遗址	［后］	0
12	运动规律（规律）	［后］	6
13	元素周期律	［后］	0
14	及时	［前］	0
16	善于	［前］	8
17	敌人	［后］	9
18	问题	［后］	12
19	毛病	［后］	2

20	情况	［后］	6
21	线索	［后］	1
22	秘密	［后］	3
23	他有心事	［后］	0
24	别人的优点（优点）	［后］	1
25	反常的现象（现象）	［后］	2
总频次：			76

***非词的搭配，如“别人的优点”取“优点”作为搭配词。**

根据语料库数据，多义词“发现”的出现频次为 586 次，搭配词典给出的搭配词在语料库中只出现 76 次，假设所有出现搭配词典给出的搭配词的语句都能通过搭配词确定词义，这个比例也只有 13%。这个数字说明利用搭配词典提供的搭配词只能进行小范围的词义消歧。本文还对“准备、学习”等多义词进行了检验，实验结果相近。

上述实验中，主要搭配词的低频说明了利用搭配区分词义的具有实际应用上的局限，不仅搭配词典如此，自动抽取的搭配也存在此问题。

利用搭配作为词义消歧线索一方面要求多义词具有不同义项具有不同搭配这个特征，另一方面还要求少量搭配词具有高的文本覆盖率，这在一定程度上限制了利用搭配区分词义在大规模语料库词义标注中的作用。

5.6 小结

搭配是最重要的词义消歧线索之一，几乎所有的词义消歧系统都通过各种方式利用搭配进行多义词的歧义消解。本文通过抽取多义词不同义项的搭配词并进行分析，说明了搭配对区分词义的作用和局限。

由于条件所限，本文的研究只使用了词典释义的例证部分提供的搭配信息和人工标注的搭配进行消歧。词义标注过程中真正利用搭配进行词义判别的比例并不高。利用义项信息进行词义消歧的一个重要条件是必须分

义项给出搭配词，对多个义项都适用的搭配词是不能用于区分词义的。而分义项给出搭配的搭配词典目前还较少。《现代汉语搭配词典》属于分义项给出搭配词的词典，对词义消歧而言是很好的资源，但由于本文的词义系统只基于《现代汉语词典》，而《搭配词典》的义项划分和《现代汉语词典》存在很大的不同，不能直接利用其分义项的搭配信息。人工构造的搭配信息只是出于主观判断，也存在明显的局限。

本文研究过程中，词义标注语料库尚未完成，不能用于抽取搭配进行词义消歧。词义标注语料库建设完成后，对利用搭配进行词义消歧研究将有极大帮助。

第六章　常用义标注与义频分布

根据对多义词在真实语料下的词义情况分析，绝大部份多义词的义项频率分布是不均衡的。具体表现为只有个别义项高频，其他义项低频。因此，出于提高标注效率和准确率两方面的考虑，语料库词义标注的一个重要过程就是预先标注多义词的最常用义项。

6.1 义频分布对词义消歧的作用

从词义消歧的角度看，义项的频率分布情况也是可区分词义的一种消歧线索。例如，多义词具有一个或者多个高频义项将会影响消歧方法的选择，词义消歧必须以高频义项为重点等等。其中最重要的是通过常用义预标注来提高词义消歧的下限值。

Gale、Church 和 Yarowsky 于 1992 年提出了词义消歧程序所能达到的上限（Upper Bound）和下限（Lower Bound）概念[20]，用于衡量词义消歧程序的性能。词义消歧的上限是指人所能达到的水平。词义消歧的下限一般是指为文本中的多义词选择其最常用义项进行词义标注所能达到的准确率。下限也通常称为词义消歧的基准线（Baseline）。

对不同性质的多义词，词义消歧的下限可能不同。如果一个多义词只有两个出现概率相似的义项，那么 90%的消歧正确率可以算是高水平；如果多义词两个义项的出现频率相差很大，例如 9 比 1，那么只能达到 90%正确率的算是低水平的消歧程序。从多义词义项性质的角度看，义项区分度低的词往往上限就低；义项频率分布有显著差别的词，更容易达到高的消歧下限。可见，词义消歧的上限值与词典义项划分与释义关系紧密；而下限值的高低则与多义词的义项频率分布密切相关。

为了充分利用大部分多义词的义项频率分布不均这个特点，通过预标注来提高词义消歧的准确率，本文采用人工选择多义词的最常用义的方式

[20] Gale W., Church K., Yarowsky D., Estimating upper and lower bounds on the performance of word-sense disambiguation programs, In Proceedings of the 30th annual meeting on Association for Computational Linguistics table of contents, Newark, Delaware, 1992, P249 - 256

对语料库进行词义预标注。虽然标注人可能很难准确判断一部分多义词的哪个词义更常用，存在误判的情况，但从对标注结果的检验来看，总体准确率较高，可以达到 70%左右，详见表 18。预先标注多义词常用义对提高词义自动标注的准确率非常重要。

表 18：人工标注常用义的准确率

常用义标注	词数	总词频	准确率
多义实词（名动形）	2525	177181	69%
※两个义项	1771	71116	80%
※三个义项	456	37605	70%
※四个义项	125	15098	66%
※五个义项及以上	112	53362	55%

6.2 词义消歧的下限值估计

词义消歧的下限值是指标注最常用义所能达到的词义消歧准确率。

表 19 给出了统计得到的为语料库中所有多义词标注准确的最常用义所能达到的词义标注准确率，即词义消歧的下限值。

表 19：词义消歧的下限值

常用义标注	词数	总词频	下限值
多义实词（名动形）	2525	177181	79%
※两个义项	1771	71116	88%
※三个义项	456	37605	81%
※四个义项	125	15098	74%
※五个义项及以上	112	53362	67%

综合表 18 和表 19 的数据，可以估计出基于词典的汉语词义消歧的平均下限值。一般认为，基于人工标注最常用义和基于语料库统计两种方式所能达到的准确率数据可以作为词义消歧的平均下限值。从本文研究所用的词典和语料来看，这个值应当介于 70%～80%之间。

根据观察，不同的语料类型会影响下限值，本文研究所采用的“华文教材语料库”中文学类作品比重较大，词义表现也比较丰富，因此下限值可能偏低。如果是比较规整的新闻体裁语料，下限值应当高于 75%。需要注意的是本文的标注仅限于名、动、形三种词类，如果加入副词等词类这个数值会有变化。但由于名、动、形是主要的多义词，其平均数值可以用来代表全体多义词。

6.3 常用义预标注的可行性

在没有大规模词义标注语料库支持的情况下，人工标注常用义是必须的过程。人工标注常用义是否可行可以通过具体数据进行分析。表 20 分别给出了词义标注语料库中 2525 个多义词人工给出的和统计得到的最常用义对应词典义项序号的概率。

表 20：常用义对应的词典义项

义项	人工标注的		统计得到的	
	常用义	比例	常用义	比例
第一个义项	1663	66%	1606	64%
第二个义项	665	26%	669	26%
第三个义项	112	5%	143	6%
其他义项	85	3%	107	4%

从表 20 我们可以看到，超过 60%的多义词的最常用义为其第一个词典义项。对所有词只标注其第一个词典义项就可达到 60%以上的总体准确率。从数据看，人工标注常用义的时候总体上贴近常用义的事实分布，但人在标注的时候稍微更倾向于第一义项。事实上，常用义不是第一义项的情况比标注人判断的要多一些。

需要注意的是虽然人工给出的常用义与词典常用义分布情况接近，但人工给出的常用义的准确率却有待提高。从表 18、表 19 的常用义消歧准确率来看，人工给出常用义的准确率离基于统计的、准确的常用义还有将近 10%的可提高空间。所以对语料库词义标注工程实践来说，多投入人力、

时间在常用义的标注上是值得的。因为常用义标注面对的是数量固定的多义词表，对这些词表进行精加工，可以大幅度的提高词义自动标注的准确率，从而减轻后期的人工校对的工作量。

根据语料统计，人工标注的常用义对大部分多义词来说都符合其事实上的分布，但也存在人工判断错误的情况。人工判断错误的原因主要有两种：1）一部分多义词的具有多个高频义项，判断哪个最常用非常困难；2）人往往不具备对词义使用情况的客观认识，造成判断错误。

容易因为第一种原因造成常用义不易判断的多义词可参照表 23 和表 24 给出的具有两个及以上高频义项的词。表 21 例举了因第二种原因造成的人工常用义标注错误的多义词。

表 21：常用义人工标注错误的词

多义词	《现代汉语词典》词义	人工判断常用义	语料库义频统计数据	
			最常用	详细分布
方向/n	①指东、南、西、北等。②正对的位置；前进的目标。	①	②	①出现 25 次 ②出现 101 次
知识/n	①人们在社会实践中所获得的认识和经验的总和。②指学术、文化或学问。	①	②	①出现 12 次 ②出现 163 次
介绍/v	①使双方相识或发生联系。②引进；带入（新的人或事物）。③使了解或熟悉。	①	③	①出现 10 次 ②出现 2 次 ③出现 205 次
早/a	③时间在先的。④比一定的时间靠前。	③	④	③出现 10 次 ④出现 169 次
严肃/a	①（神情、气氛等）使人感到敬畏的。②（作风、态度等）严格认真。	②	①	①出现 75 次 ②出现 1 次
作用/n	②对事物产生某种影响的活动。③对事物产生的影响；效果；效用。	②	③	②出现 13 次 ③出现 81 次

注：上表中给出的是《现代汉语词典》的多义词的词义，语素义未列入。

表 21 说明了人工标注常用义的困难。表格中的所有词都具有义项频率分布不均衡的特点，具有标识出最常用义的理论上的可能性。但由于人往往不具备对词义使用情况的客观认识，标注多义词的最常用义不是对所有词都容易实现的。

下文将具体分析语料库中的多义词义项分布情况。

目前在汉语中，关于义项分布的全面研究还很少。本文给出基于词义标注语料库的义项分布数据，对词义分布研究具有重要的参考价值。

6.4 只有一个高频义项的多义词分析

1）只出现一个高频义项的多义词

多义词义项频率分布的一种极端情况是虽然有多个词典义项但除一个高频义项外其他义项不出现。具有这种分布特征的词约占语料多义词的 35%。

考虑语料库规模的影响，低频的多义词有些义项未出现与其词频有直接关系，因此本文只分析在语料库中出现 50 次以上的多义词。表 22 例举了语料库中只出现一个义项的多义词。

表 22：只出现一个义项的多义词

ID	多义词	词类	语料库义项数[*1]	词典义项数[*2]	义项频率分布	语料频次[*3]
1	发生	v	1	2	①_273	273
2	唱	v	1	2	①_273	273
3	脸	n	1	4	①_267	267
4	先生	n	1	3	②_267	267
5	风	n	1	2	①_253	253
6	朋友	n	1	2	①_252	252
7	响	v	1	2	②_206	206
8	情况	n	1	2	①_185	185

9	洗	v	1	4	①_182	182
10	清楚	a	1	2	①_171	171
11	眼	n	1	4	①_168	168
12	车	n	1	2	①_167	167
13	近	a	1	2	①_155	155
14	大地	n	1	2	①_154	154
15	交流	v	1	2	②_147	147
16	总理	n	1	3	①_146	146
17	年轻	a	1	2	①_142	142
18	真	a	1	2	①_140	140
19	摘	v	1	3	①_139	139
20	白菜	n	1	2	②_50	50

注：*1 语料库义项数是指该词类的多义词在语料库中出现的义项总数

*2 词典义项数是指该词类的多义词在《现代汉语词典》中的义项（不含语素义）总数

*3 语料库频次是指该词类的多义词在语料库中出现的次数

对词义消歧与词义标注来说，确定这部分词的意义在于，只需要选择或标注常用的那个义项就可达到极高的正确率。

从词义消歧的角度看，词义的使用应当成为词义划分时需要考虑的一个重要因素，可以减少“伪多义词”的产生。

从词典编纂的角度看，义项不出现在真实语料中反映了词典义项划分和释义的一些问题。根据对《现代汉语词典》的分析，语料库中多义词只出现一个义项的原因可总结为如下三种：

1）其他义项不常用。例如：

交流 jiāoliú

①交错地流淌 ：涕泪～丨河港～。

②彼此把自己有的供给对方：物资～丨文化～丨～工作经验。

多义词“交流”在语料库中只出现义项②。义项①“交错地流淌”基本

不使用，所以语料库只出现一个义项。

2）多义词的其他义项实际上为语素义而不是词义，不作为单独的词使用。例如：

车 chē

①陆地上有轮子的运输工具：火～｜汽～｜马～。

③机器：开～｜～间。

名词“车”的义项③“机器”一般是不作为词使用的，与其作为语素义的义项②“利用轮轴旋转的工具”性质相近。

3）义项区分困难。人工标注时不能有效区分开多个义项，全部倾向性地标注为其中一个义项。例如：

白菜 báicài

①一年生或二年生草本植物，叶子大，花淡黄色。是常见蔬菜。品种很多，有大白菜、小白菜等。 ②特指大白菜。

多义词“白菜”的义项②是特指义，其词义范围要小于义项①，但对人来说是最熟悉的意思，根据生活经验，“白菜”一般都指义项②。文本往往不能提供足够的信息使得人工标注时有效区分开两个义项，并准确识别出义项①，造成了义项①在语料中不出现的情况。根据对语料的观察，发现绝大多数情况下，义项①和②是不能区分的，例如“他看见小白兔挑着一担白菜，给老山羊送来”，不能确定其所指，但根据生活常识，校对者这里选择了义项②。

除上述三种情况外，语料库的性质也会影响到词义的出现情况。例如：

朋友 péngyou

①彼此有交情的人。

②指恋爱的对象：姑娘多大了，有～了没有？

多义词“朋友”在语料库中只出现义项①。本文研究采用的是中小学语文教材语料库一般不涉及恋爱话题，影响到义项②的出现。此外，义项②倾向于口语化，在书面语语料库中可能会被其他同义词代替，也会影响其使用。

2）出现多义项但只有一个高频义项的多义词

多义词不同义项在频率分布上不均衡，个别义项高频，其他义项低频是普遍现象。若以出现多个义项并且有一个义项的出现频率超过 80%作为衡量标准，语料中约有 31%多义词符合只出现一个高频义项这个特点；若以 70%作为衡量标准，则有 41%的多义词满足条件。

表 23 列出了部分有多个义项但只有一个高频义项的多义词。

表 23：出现多义项但只有一个高频义项的多义词

ID	多义词	词类	语料库义项数*1	词典义项数*2	义项频率分布	语料频次*3
1	世界	n	3	5	①_460 ③_23 ⑤_4	487
2	老	a	2	8	①_448 ⑤_14	462
3	在	v	2	5	②_441 ③_1	442
4	笑	v	2	2	①_440 ②_2	442
5	学	v	2	2	①_368 ②_37	405
6	变	v	3	3	①_376 ②_13 ③_1	390
7	画	v	2	2	①_43 ②_338	381
8	讲	v	4	4	①_332 ②_12 ③_7 ⑤_27	378
9	跳	v	4	4	①_316 ②_9 ③_27 ④_5	357
10	美	a	2	2	①_303 ③_42	345
11	白	a	2	2	①_330 ②_1	331
12	远	a	2	3	①_307 ③_18	325
13	了解	v	2	2	①_275 ②_19	294
14	边	n	4	5	①_1 ②_283 ③_3 ⑦_6	293
15	研究	v	2	2	①_269 ②_21	290
16	道	v	2	2	③_277 ④_11	288
17	老人	n	2	2	①_260 ②_5	265
18	同志	n	2	2	①_10 ②_239	249
19	历史	n	3	4	①_208 ②_29 ④_10	247
20	快	a	3	3	①_235 ⑤_3 ⑥_1	239

注：*1 语料库义项数是指该词类的多义词在语料库中出现的义项总数

*2 词典义项数是指该词类的多义词在《现代汉语词典》中的义项（不含语素义）总数

*3 语料库频次是指该词类的多义词在语料库中出现的次数

“**义项频率分布”列中**,“①_193”是指义项①出现了 193 次，其他以此类推。

对词义标注来说，只出现一个高频义项的多义词一般只采用常用义标注就可达到高准确率。上表中的 35 个义项频率分布差异大的多义词只通过常用义标注就可达到高准确率。具体来看，这 35 个多义词的总出现次数是 10,161 次，各自的最常用义的出现次数之和是 9,535 次。若标注正确的常用义，即每个词都标注出现次数最高的那个义项，其正确率为 $9535 / 10161 \times 100 \approx 94\%$。

以上数据只分析在语料库中出现两个以上义项的多义词，没有加入在语料库只出现一个义项的多义词。若把上一节中分析的只出现一个义项的多义词加入（如表 21），这种只有一个高频义项的多义词通过常用义标注的正确率将会有更大的提高（超过 95%）。

6.5 有多个高频义项的多义词分析

1）有两个以上高频义项的多义词

多义词不同义项在频率分布上不均衡的另外一种情况是多个义项中有两个或以上的义项同时高频常用，其他义项不常用。若以有两个义项的出现频率都在 40%以上作为衡量标准，语料中有约 9%的多义词同时出现两个高频义项；若以 30%作为衡量条件，则比例提高到约 20%。这类词的高频义项间的区分是提高词义消歧正确率的难点之一。表 24 例举了部分有两个以上高频义项的多义词。

表 24：有两个以上高频义项的多义词

ID	多义词	词类	语料库义项数*1	词典义项数*2	义项频率分布	语料频次*3
1	天	n	4	6	①_174 ⑩_24 ③_154 ⑦_32	384
2	爬	v	3	3	①_166 ②_195 ③_19	380
3	落	v	5	6	①_244 ⑩_9 ②_92 ⑥_8 ⑨_5	358

4	问题	n	4	4	①_102 ②_219 ③_3 ④_16	340
5	指	v	4	4	②_2 ③_191 ⑤_2 ⑥_84	279
6	受	v	4	4	①_109 ②_136 ③_24 ④_1	270
7	掉	v	4	7	①_100 ②_2 ③_2 ⑤_153	257
8	黑	a	3	3	①_170 ②_49 ⑤_5	224
9	提	v	5	7	①_96 ②_14 ④_66 ⑥_3 ⑦_27	206
10	留	v	5	6	①_81 ③_13 ⑤_71 ⑥_2 ⑦_37	204
11	动	v	5	5	①_72 ②_42 ③_36 ④_11 ⑤_4	165
12	力量	n	4	4	①_83 ②_37 ③_26 ④_14	160
13	客人	n	3	3	①_69 ②_27 ③_30	126
14	坏	a	3	3	①_20 ②_79 ⑤_27	126
15	文字	n	3	3	①_31 ②_3 ③_89	123

2）所有义项都高频的多义词

多义词的义项分布主要呈不平衡状态，少数义项高频，部分义项低频、甚至不出现。其例外就是义项频率分布差异很小的，或者说所有义项都高频的多义词。这类词极大程度地压低了词义消歧的下限值。如果以没有任何一个义项的出现频率低于 30%作为衡量标准，语料中义项分布平均的多义词约占 15%；若以 40%作为衡量标准，则比例降低到约 8%。表 25 列出了部分具有义项频率分布平均的多义词。

表 25：所有义项都高频的多义词

ID	多义词	词类	语料库义项数[*1]	词典义项数[*2]	义项频率分布	语料频次[*3]
1	长	v	2	3	①_310 ②_138	448
2	送	v	3	3	①_204 ②_112 ③_106	422
3	怕	v	2	3	①_183 ③_140	323
4	准备	v	2	2	①_179 ②_89	268
5	表示	v	2	2	①_81 ②_182	263
6	有关	v	2	2	①_86 ②_167	253

7	一般	a	2	2	①_110 ③_106	216
8	月	n	2	3	①_63 ②_137	200
9	充满	v	2	2	①_56 ②_134	190
10	环境	n	2	2	①_83 ②_55	138
11	名字	n	2	2	①_86 ②_41	127
12	抬	v	2	3	①_77 ②_37	114
13	扔	v	2	2	①_51 ②_53	104
14	消息	n	2	2	①_49 ②_53	102
15	贴	v	2	3	①_50 ②_50	100
16	深刻	a	2	2	①_63 ②_35	98
17	亲切	a	2	2	①_56 ②_39	95
18	隔	v	2	2	①_43 ②_39	82
19	楼	n	2	2	①_38 ②_42	80
20	开放	v	2	2	①_43 ②_36	79

6.6 小结

通过对语料多义词义项频率分布的统计分析，可以看到义项频率分布不均衡是一个普遍的特点。这个特点对词义消歧有十分重要的意义。正如Wilks（1997）所指出的，考虑到词义的具体语言使用，词义消歧并没有根据对词典词义区分进行分析所看到的那么复杂。充分利用多义词的义项频率分布特点，建立一个实用的、高准确率的词义消歧系统是可以实现的。

义项频率分布对词典编纂、词汇教学也有重要意义，这体现了词义标注语料库在语言研究方面作为基础性资源的作用。

第七章　自动消歧的难点

目前的自然语言处理研究现状决定了计算机自动词义消歧的核心是基于统计的概率方法的运用，而非基于对文本进行语义上的理解。统计方法容易实现较高的正确率，但局限也非常明显，准确率达到一定程度后就很难再有实质性提升。此外，对大部分高频常用、语义复杂的多义词，统计方法的表现也相对较差。本章中我们将通过对词义标注语料的数据分析探讨计算机自动词义消歧的难点和可能的提高途径。

7.1 自动消歧的准确率估计

根据对语料库中 2525 个多义词的统计，通过读音、词类、搭配和常用义进行词义消歧的总体准确率情况如表 26 所示。

表 26：自动消歧的准确率

	词数	正确率
多义实词（名动形）	2525	73%
※名词	1075	73%
※动词	1097	69%
※形容词	353	80%

需要特别指出的是，表 26 给出的准确率不是全语料库词义消歧的准确率，不包含通过拼音、词类区分后不再有歧义的多义词的情况。若包含读音、词类等信息可完全区分的多义词，全语料库词义消歧的准确率约为 83%。因为本文研究采用《现代汉语词典》作为词义体系，而目前还没有其他的基于《现代汉语词典》的全词词义消歧准确率介绍，所以无法做横向的准确率比较。

从语料库数据来看不同性质的多义词在准确率方面差异很大，能达到高准确率的主要是义项数少、义项频率分布差异大的多义词。

按义项数分类的多义实词标注准确率如下所示：

1）两个义项的多义词

多义实词（名动形）	词数	正确率
两个义项的多义词	1782	78%
※只出现一个义项	750	91%
※两个义项都出现	1032	74%

2）三个义项的多义词

多义实词（名动形）	词数	正确率
三个义项的多义词	470	69%
※只出现一个义项	106	86%
※出现两个义项	174	76%
※三个义项都出现	190	66%

3）三个义项以上的多义词

多义实词（名动形）	词数	正确率
三个义项以上的多义词	273	60%
※所有义项全部出现	77	53%
※只出现部分义项	196	67%

从上述数据可以看出，机器词义消歧的准确率与多义词义项的多少成反比关系，义项数越多的多义词越难以达到高的准确率。此外，还与多义词的义项分布有关，一个多义词的不同义项在文本中的出现情况差距越大就自动消歧的准确率就越高。其主要原因就是语料库经过常用义标注，这与前文讲过的词义消歧的下限情况是吻合的。

7.2 自动消歧高准确率词的特点

目前，基于概率的方法是自动词义消歧研究的主流。而概率方法主要词义表现在文本中的形式特征，例如搭配不同、语法功能不同等等。

从这个角度说，计算机词义消歧要达到高的准确率的前提条件是多义词的不同词义必须在语言使用上有明显的不同。两个或多个不同词义在使

用上没有明显区别的多义词是自动词义消歧的难点。此外，出于计算复杂度的考虑，计算机进行词义消歧时只能判断一个长度有限的上下文，通常是一个或多个有限的句子，如果不同词义的区分信息需要通过更大的上下文范围才能提供，也会造成词义消歧的困难。另外一种难以区分的情况是词义消歧时需要进行语义上的正确理解，目前的计算机程序很难达到这一点。

自动词义消歧能够达到高准确率的多义词，一般都具有不同义项具有不同的使用这个特点。例如，多义动词 “笑”。“笑/v”之所以容易消歧是因为其不同义项在语言使用上有明显的不同。“笑/v”的义项①“露出愉快的表情，发出欢喜的声音”和义项②“讥笑”具有语法上的明显差别。义项①是不及物的，后面不能跟宾语，义项②是及物的后面可以跟宾语，并且绝大多数情况下都需要有宾语；义项①后面可以加“了、着、得”等助词，义项②一般不能；义项①有重叠和离合使用形式，如“笑笑、笑了笑、笑一笑”，义项②没有。

下面是语料库中“笑/v”的部分例句：

1）义项①：

春天像小姑娘，花枝招展的，笑着，走着。
影子也高兴地笑了。
爸爸哑着嗓子，拉起我的手笑笑说："我怎么能够去？"
他笑了笑，似乎早就料到我会提出这样的问题。
邱士力笑得咧开嘴，

2）义项②：

我还暗地里笑他，以为他总是崇拜偶像。
留下三个太阳脸晒黑了，大家笑她像非洲人。
你以为它真是盛酒的金罍吗？它会笑你呢。
我想喊他等等我，却又怕他笑我胆小害怕；
你们笑什么？笑你们自己！.....你们这些人呀！

通过观察“笑/v”的两个义项的例句，我们可以看到义项①和义项②在使用上明显不同。此外，“笑/v”的两个义项在频率分布上也有巨大的差异，语料库中义项①出现 422 次，义项②只出现了 20 次。从纯粹提高消歧

准确率的角度看，标注义项①为常用义，再把义项②按照其使用上的特点识别出来，“笑/v”的词义消歧就可达到99%以上的准确率。

另外一类容易达到高的自动消歧准确率的词是通过有限的搭配词就可以很好的区分开不同的义项。例如多义名词“背景”。“背景/n”有四个义项，分别是：①舞台上或电影、电视剧里的布景，放在后面，衬托前景；②图画、摄影里衬托主体事物的景物；③对人物、事件起作用的历史情况或现实环境；④指背后倚仗的力量。“背景/n”的义项①和②，按本文前一章的分析，是划分非常细的义项，人在区分这个类型的义项时会遇到不易区分的困难，但在搭配上也还是有各自的特点。义项③和④是典型的可通过搭配来实现高消歧准确率的义项。“背景/n”在语料库中一共出现了60次。通过分析这60个出现“背景”这个词的所有语句的分义项的搭配词情况，我们发现其最高频的义项③可以很好地通过搭配来识别。指示义项③的搭配词及其出现次数如下：时代～（13次）、文化～（4次）、社会～（3次）、写作～（3次）、～材料（3次）、历史～、创作～、知识～、人物～、生活～、～资料。通过这些搭配词，语料库中出现的多义词“背景”的义项③可以被有效的全部识别出来。

需要注意的是不同义项具有不同使用特点的多义词只是占少数，对大部分多义词来说，只根据语法上的特点是很难完全区分开不同义项的。

7.3 自动消歧的难点分析

词义自动消歧的本质是利用多义词的不同义项在语言使用上的不同特征进行词义的识别及判断。这些区分特征既有形式的又有语义的。语义上的区分特征主要是语义之间的选择和限定（Selectional Restrictions），例如谓词论元方法（Resnik，1997）。“选择—限制”方法用于词义消歧的语言学基础建立在语义之间的约束体现在词义的选择上。这种方法尝试通过词义间的“选择—限制”来消除不正确的词义，从而选择正确词义。谓词论元方法就是“选择—限制”方法的一种，利用谓词与论元之间语义的选择关联性进行词义消歧。这种基于语义的方法的缺点是通常只适用于基于语义词典的词义消歧，对语言词典的词义消歧很难操作。另外还需要大规模

的人工构造知识库，所以实际应用有限。本文研究的重点不在于具体的词义消歧方法，且只基于《现代汉语词典》这部传统的语言词典，并未利用其他资源，因此基于语义的方法不是本文关注的对象。

二十世纪九十年代以来，在自然语言处理领域，基于统计的概率方法占据主导地位。概率方法的特点是只需要数学模型和语料库就可实现好的效果。概率方法的核心是通过形式上的区别来进行语义的识别和判断。对词义消歧这个任务来说，这种形式上的区别就是多义词的不同词义在语言使用上的不同特点。因此计算机自动词义消歧的难点是形式不能区分或不能完全区分的多义词。

上文指出，容易实现高的自动消歧准确率的多义词是不同义项在语言使用上具有明显不同的词。同理，自动消歧的低准确率词就是不同义项之间区分更多的是基于纯语义，而较少反映在语言使用上的词。这部分词难以消歧的原因具体表现在不同义项的上下文相似、搭配词大量重叠或是必须进行完全的语义理解。

7.4 不同义项上下文相似性高的词

计算机自动词义消歧的核心是计算不同词义在语言使用上的差异，主要反映为计算不同义项的语境词及频率。为考察多义词不同义项的上下文相似性，本文设计了一个算法，计算不同义项上下文词语的重叠程度。算法大致分为预处理和计算两个过程。

预处理过程首先是选择出至少有两个义项高频义项的词，如果不存在两个同时高频的义项，无法进行基于统计的对比分析；其次是提取不同义项的上下文词频表，分开存放；第三，对词频表进行修正，去除虚词和低频词，因为低频词具有偶然性，不能作为典型的上下文同现词。

计算过程的重点就是求两个义项的上下文词的交集，以两个义项的上下文同现词的平均重叠个数作为上下文相似性的值。公式表示如下：

$$w_1 = C_{s1} \cup C_{s2} / N_{s1} \qquad w_2 = C_{s2} \cup C_{s1} / N_{s2}$$

$$f_{simi} = (w_1 + w_2)/2$$

说明：C_{s1}是义项 1 的同现词；C_{s2}是义项 2 的同现词；N_{s1}是义项 1 的

同现词的个数；N_{s2}是义项 2 的同现词的个数；w_1是义项 1 与义项 2 的同现词相同的个数与义项 1 同现词个数的比例；w_2是义项 2 与义项 1 的同现词相同的个数与义项 2 同现词个数的比例；f_{simi}是义项 1 和义项 2 同现词的平均相似度。

表 27 给出了根据上下文相似度公式计算出的不同义项上下文相似度高的部分多义词。

表 27：不同义项上下文相似度高的多义词

ID	多义词	相似度	义频分布
1	环境/n	53.6	①_83 ②_55
2	困难/a	53.13	①_66 ②_17
3	欣赏/v	52.73	①_143 ②_11
4	喊/v	52.53	①_190 ②_13
5	玩/v	51.48	①_176 ②_19
6	中间/n	48.79	②_31 ③=129
7	借/v	48.45	①_103 ②_35
8	长/v	47.43	①_310 ②_138
9	织/v	47.23	①_37 ②_14
10	有关/v	46.37	①_86 ②_167
11	土地/n	46.3	①_110 ②_16
12	社会/n	45.93	①_41 ②_236
13	作用/n	45.83	②_13 ③=81
14	扔/v	44.67	①_51 ②_53
15	出现/v	44.59	①_203 ②_76
16	满足/v	44.56	①_36 ②_24
17	命运/n	44.45	①_61 ②_26
18	精神/n	44.34	①_208 ②_19
19	菜/n	43.41	①_35 ③=255
20	青年/n	43.25	①_42 ②_84

根据对语料例句的分析，上表中不同义项上下文相似度高的多义词普遍存在着义项区分困难的情况。例如：

1）方向/n　fāngxiàng

①指东、南、西、北等：在山里迷失了～。

②正对的位置；前进的目标：军队朝渡口的～行进。

邓稼先的一生是有　方向　、有意识地前进的。

大家都朝他指点的　方向　看。

当板块向一个或另一个　方向　运动时，大陆也随之一起运动。

千万不能动摇自己的信念和　方向　。

顺着叮当的斧凿声传来的　方向　，我们转进一条巷子。

它用尾巴保持平衡，调整　方向　。

只要　方向　不错，怎么走都可以。

小男孩跳下站台看那个　方向　，电车没有来。

分析：名词“方向”的两个义项意思相近，并且其不同义项很难通过上下文形式上的特征，例如搭配来区分，在句法上也看不出区别，是计算机很难消歧的多义词。

2）亲切/a　qīnqiè

①亲近：亲密：他想起延安，像想起家乡一样～。

②形容热情而关心：老师的～教导。

他转过脸，　亲切　地望着他这位朋友，它也微笑着望着他。

这一条我看后却觉得很　亲切　，而且"与有荣焉"。

伐木工人正在回忆他　亲切　的笑语。

当地群众　亲切　地称这些舟船为"生命之舟"。

作者以　亲切　动人的笔墨，记录了社会生活的一面，

及至看到了林场，这种　亲切　之感更加深厚了。

江主席　亲切　地拍了拍我的肩膀。

分析：形容词“亲切”的两个义项不容易通过上下文词语进行有效区分。

3）有关/v　yǒuguān

①有关系：～方面｜～部门｜这些问题都跟哲学～。

②涉及到：他研究了历代～水利问题的著作

请搜集　有关　资料，写成一篇作文，

有条件的话，可以浏览以下　有关　昆虫知识的网站：

根据这些线索，借阅　有关　的图书。

了解我国古今与土地　有关　的影响比较大的改革。

可将　有关　的成果或设想用恰当的语言写成书面介绍。

总指挥李继耐率　有关　领导乘车随后。

参观过一个　有关　《红楼梦》的展览。

你能根据自己对　有关　文章的理解回答这些问题吗？

分析：词典虽然给出了动词“有关”的不同义项的几个搭配词，但从整个语料库来看，“有关”不容易通过上下文同现词来消歧，主要原因之一

是可以跟“有关”同现的词语数量太大，典型搭配在其中只占非常小的比例。因此，计算机无法充分提取并利用同现词来消歧。

7.5 必须通过理解语义进行消歧的词

目前的自动消歧方法绝大多数是基于概率的，基本不通过理解文本语义来消解词语多义，并且从目前自然语言处理研究的水平来看，对文本进行准确的语义理解也是很难实现的。因此，如果多义词的义项区分必须通过语义理解来实现，那么很难通过自动消歧达到高准确率。例如：

1） 借 jiè

①暂时使用别人的物品或金钱；借进：向图书馆～书丨跟人～钱。

②把物品或金钱暂时供别人使用；借出：～书给他丨～钱给人。

本级的学生会干事到我寓里来了，要 借 我的讲义看。

巴萨尼奥 借 了你三千块钱，现在拿六千块钱还你好不好？

被子一拿出来，我方才明白她刚才为什么 借 的道理了。

不肯

他总是那么诱人，能 借 什么东西给他，似乎是件很荣幸的事。

“风雨这么大，不会有人来 借 书了吧。”小王想。

不要误会，我不是拿东西，是把 借 你的橡皮放回去。

他不好意思拒绝，于是就和主席讲好：只 借 一个月，到期归还。

他的藏书很多，别人向他 借 书要办借书手续；他自己从书架上拿走一本书，也要办理一下手续。

分析：按照词典释义，动词“借”的不同义项的核心是动作的方向性，一个是借入，一个是借出。这种方向性的识别必须通过对涉及到的物品、金钱的所有者及其流转的方向的判断得到。人可以通过对文本的语义理解来实现消歧，基本不会出现判断错误，但对计算机来说，这种语义信息很难识别出来，因此低准确率是无法避免的。

2）趴/v pā

①胸腹朝下卧倒：～在地上射击。

②身体向前靠在物体上；伏：～在桌子上画图。

小女孩子 趴 在船边，用两只小手淘着水玩。

狗 趴 在地上吐出红舌头，

他头也不回地跑回自己的房间， 趴 在床上哭了起来。

指导员 趴 在离我不远的地方，一动也不动，

邱少云像千斤巨石一般， 趴 在火堆里一动也不动。

小花狗懒洋洋地 趴 在楼梯上。

扭头看，这小家伙竟 趴 在我的肩头睡着了，

分析：动词“趴”的两个义项的区分是语义的，按照释义，要准确区分两个义项需要理解“趴”在语境中的准确的动作方式，这是目前的消歧程序很难做到的。

7.6 小结

目前大部分词义消歧系统对词典义项之间的关系信息利用不够，在做不同算法的消歧效率的实验时，消歧对象通常只集中于词义间距离较大的多义词或者同形词。从词义区分的特点来看，消歧实验系统应当给出存在不同的义项间关系的多义词的实验数据，才能提高消歧效率的实际意义。因此，对词义消歧程序来说，充分利用多义词义项之间的关系，对不同类型的多义词采取不同的消歧算法是提高消歧准确率的一个可行的途径。例如，对义项间存在包含关系和重叠关系的多义词采用不同的消歧策略，对存在义项包含关系的多义词，根据应用的需要，只需要标注词义范围大的那个义项就可以了。

第八章　人工标注的难点

本章通过总结人工标注词义过程中碰到的困难，对多义词词义区分的难点进行总结，尝试根据词义区分的特点对这些难点进行分类，并分别通过语料数据进行分析。

8.1 人工标注过程中的难点分析

从词典角度看，多义词是具有两个或两个以上义项的词。多义词所具有的义项之间的地位通常并不是平等的，其中至少有一个义项是基本的、常用的，其他的义项一般是由这个义项直接或间接地发展转化来的，因此在语义上存在着或多或少的联系。与同形词相比，多义词的核心就在于的其各个词义之间存在意义上的必然联系。词典在进行多义词义项划分和释义时，主要的依据就是多个意义间的区别，意义区别到了一定程度后，就作为独立的义项划分出来；其次，就是考虑用法上的不同，意义相近而用法明显不同的也作为独立的义项划分出来。通过对词典义项划分的分析，按照词典释义中的“Core Sense”理论，多义词的义项间的词义区别可以通过图 3 来表示。

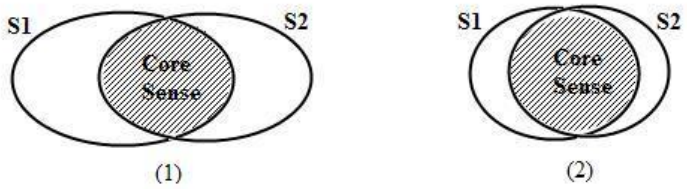

图 3：多义词义项间的词义区别

图 3 中，⑴的两个义项虽然共有一个核心词义，但二者各自独有的意义所占比重很大，因此义项间的区别较为显著；⑵的两个义项也有共同的核心词义，但两个义项分别独有的意义所占比重低，因此义项间的区别较不明显。

在人工标注词义的过程中，判断两个义项之间的区别性特征是主要的出发点。这种区别性特征既包括语义也包括语法。对人而言，图 3 中，⑴的两个义项间的区别较为明显，就更容易实现高的标注准确率；⑵的两个

义项之间的区别较不明显，就更容易带来标注错误和标注的不一致性。

一般来说，词典划分义项并没有统一的原则，并不是客观的，都带有词典学家对词义的主观认识。这从不同的辞书具有不同的释义体系体现出来。一部词典确定的多义词，在另一部词典中可能只有一个义项；一部词典划分出两个义项的，另一部词典可能划分为三个；一部词典划分出这两个义项的，另一部词典可能划分为那两个义项。由于这种义项划分的主观性，词典中普遍存在对义项间的区别特征把握不一致，不同多义词采用不同标准的情况。因此就出现了有些多义词的差别较大的两个义项未划分开，而有些多义词的两个区分并不十分明显的词义划分为两个义项的情况。这种不一致性反映到词义标注过程中，就存在一部分义项不易区分和判断的现象。对词义标注来说，只有可以明确区分的词义才能得到准确的标注，对区分不清晰的词义，校对者往往只是进行倾向性的判断，其判断依据通常带有主观色彩。这正是造成语料库词义标注不一致问题的最主要原因。

通过对标注语料数据以及《现代汉语词典》多义词义项划分和释义进行分析，可以总结出多义词的不同义项之间的逻辑关系存在几种基本模式，即多义词义项之间在词义上存在着相离、重叠（交叉）、包含等多种情况。这里为说明义项之间的逻辑关系，不考虑多义词各义项共有的词义核心，各义项的词义只取其核心词义之外的部分。义项关系示例如图 4。

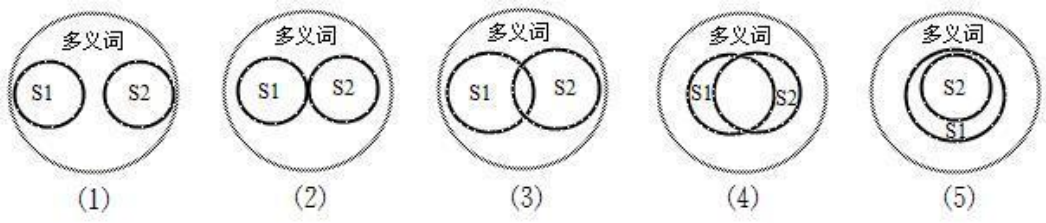

图 4：多义词义项关系的几种类型

图 4 中，⑴表示两个义项的距离大，义项可区分性很好。这种情况下词义不易区分的主要原因在于两个距离大的义项之间存在着义项缺失的情况，在标注时发现多义词的词义既不属于 A 也不属于 B，难以归到现有的释义中。⑵表示两个义项有很好的区分性，在词义标注时往往不易造成困扰。但这只是一种绝对的状态，释义过程中一般不会出现，多义词的不同义项之间总是存在着或多或少的关联。因为兼类而划分出不同义项的多义

词可以认为是满足⑵的一种特殊情况。⑶和⑷都表示两个义项存在语义上的重叠。⑶表示两个义项之间存在小的语义重叠现象，如果语料中的多义词词义落在这个重叠区间内而辞书又没有给出区分原则的话，就会造成判断上的困难。⑷表示两个义项的语义大部分重叠，此时若无明确区分线索，很难标注准确义项。⑸表示两个义项之间存在包含关系，即一个义项的语义可以完全涵盖另一义项，例如释义时泛指和特指现象。词义标注时，遇到义项间存在包含关系时，精确判断成了难题，从逻辑上说，标注范围的大的义项总是对的，但又缺乏精确性。标注者很容易在这个问题上出现处理不一致的情况。

词义的不易区分体现在了词义标注语料库建设的过程中，下文将根据语料标注操作实践，对人工标注过程中的难点进行分析。

8.2 词典没有提供足够的区分线索

消歧线索是词义消歧研究常用的一个概念，指的是用于区别词义的语言信息。对人而言，进行词义判断时同样需要区分线索。这种区分线索既有语法的又有语义的，人总是尝试通过对词典释义和例证的理解把握两个义项区分的要素，并以此进行词义判断。因此，如果词典释义和例证不能给出足够的区分信息时，人会因为无法把握义项区分要素而造成判断困难。此外，有些义项的区分主要是基于所指的不同，而这种不同一旦不能通过上下文得到时，词义的人工标注也存在困难，不能得到准确判断。

1） 地形/n#di4/xing2

①地理学上指地貌。 ②测绘学上地貌和地物的统称。

又因各地 地形 和距离海洋远近不同，气候复杂多样。

他不但多次到香山勘察 地形 ，攀登峰顶，俯览周围环境，

他较好地考虑了陵园与周围环境、 地形 的结合，

他们利用险要的 地形 ，把冲上来的敌人一次又一次地打了下去。

群峰竞秀的峰林石山 地形 是热带地方独有的。

分析：“地形”的两个义项的释义中缺乏足够的区分线索，并且没有提供例证，使得人工标注时很难准确区分两个义项。

2） 弟弟/n#di4/di5

①同父母(或只同父、只同母)而年纪比自己小的男子。

②同辈而年纪比自己小的男子：叔伯～。

喂，云雀 弟弟 ，叽叽喳喳说些什么？

小 弟弟 ，好玩呢，洋铜鼓，洋喇叭，买一个去。

他下了车，把小 弟弟 抱进了车里。

我和 弟弟 常常在草地上玩耍。

可怜的小利比，我的好 弟弟 ！我死了你怎么办呢？

分析：“弟弟”的两个义项的区分难点在于难以根据上下文判断是否严格符合义项①的释义。例如，习惯上“堂弟”和“表弟”也都可以称为“弟弟”，在语料中往往没有足够的语境使之得到准确的区分。另外，由于现汉没有将“小弟弟”作为词收入，而其中的“弟弟”不能认为属于义项①或②。“弟弟”做称呼语使用时普遍存在这种情况。因此，现汉对“弟弟”的释义可能存在义项缺失问题，可能应补充泛指男孩这个义项。

3） 充分/a#chong1/fen4

①足够(多用于抽象事物)：你的理由不～｜准备工作做得很～。

②尽量：～利用有利条件｜必须～发挥群众的智慧和力量。

然后根据自己的爱好和特长， 充分 发挥联想和想像

诗与其他文体相比较，能更为 充分 地显示作者的品格和情怀。

创作要 充分 发挥集体的聪明才智

我现在可 充分 体会出游子 12 的心境了。

我能 充分 了解他一生的遗憾，

对话开始时要 充分 肯定家长对自己的教育和关爱。

分析：“充分”的释义存在问题，义项②释义为“尽量”，而尽量在现汉中的词性是副词。从语料来看，义项②的使用也都是副词用法，因此造成义项区分上的困扰。

4） 一定/b#yi1/ding4

①规定的；确定的：要按～的程序进行操作。

②固定不变；必然：文章的深浅跟篇幅的长短，并没有～的关系。

③特定的：～的文化是～社会的政治和经济的反映。

④相当的：我们的工作已经取得了～的成绩丨这篇论文具有～水平。

根据自己的特长和爱好，在 一定 的场合，大大方方地推荐自己。

天气变化异常复杂，看云识天气毕竟有 一定 的限度。

命令登月舱升到 一定 的绕月轨道与飞船对接。

写出游丝的时候，笔尖运行要有 一定 的跳跃。

相信你对中国戏曲有了 一定 的了解。

自传可以是生平事迹的实录，也可以具有 一定 的文学性。

一旦达到 一定 的密度，奇怪的现象就发生了

分析：“一定”作为区别词的上述四个义项之间缺乏足够的区分线索，不易区分。

提高释义的信息量是解决消歧线索不足的一个重要方法。释义是词典描述词义的主要手段，也是人理解词义的主要途径。目前词典中释义文本短至只有一个词的情况还普遍存在，给准确理解词义带来了困难。例如，动词“停”有四个义项，释义分别是①“停止”、②“停留”；③“停放；停泊”；④“停当”。从语料看，义项的划分可以覆盖其具体使用，但因为释义信息很短，很难准确理解各个义项词义的范围，造成区分困难。例如“一只鸟飞过来停在树上”，其中的“停”很难判断其准确义项。究其原因，本文认为是义项的释义信息不足导致的，人无法通过释义准确判断各个义项的词义范围，造成了区分困难。因此，提高释义的信息量、提供更多的区分线索有助于解决词义难以区分问题。

例证属于词典给出的词语的使用信息。除释义文本外，例证是区分词义的重要手段。词典中，有些词没有提供例证造成了区分困难。例如前文中的例词 “地形”的两个义项：①“地理学上指地貌”；②“测绘学上地貌和地物的总称”，没有给出例证。由于普通人缺乏对地形学和测绘学的区别的认识，造成了区分困难，若词典给出两个义项的不同例证，可以极大程度上提高义项的可区分程度。再如上例中的动词“停”，若词典能将经常使用的类似“鸟停在树上”这样的用法在例证中举出，义项判断上的困难就可以很大程度得到消解。

8.3 义项缺失影响词义标注

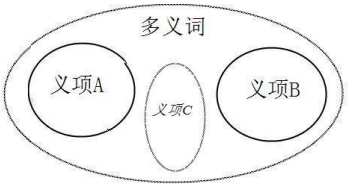

图 5：词典释义存在义项缺失的情况

多义词的义项之间距离大是指两个义项之间区分明显，由于上下文语义连贯性的约束，义项距离远的词在上下文语境方面往往也存在着明显的不同。对词义消歧来说，义项距离远的词比较容易实现高准确率的消歧；对人工标注来说，两个距离远的义项也容易判断。但是一部分多义词的义项距离大是因为词典释义时的义项缺失造成的。这种情况下，会对人工标注者造成困扰，对语料给出的多义词出现无法判断词义的情况，即出现多义词在语料上下文的词义既不属于 A 也不属于 B 的情况。此时应考虑是否增加一个新的义项 C。

1） 送/v#song4

①把东西运去或拿去给人：～报｜～信｜～饭。

②赠送：奉～｜老师～我两本书。

③陪着离去的人一起走：把客人～到大门外｜～小孩儿上学。

后来，我很不乐意地被爸爸 送 进了学校，

少年代表 送 上一封信和礼物，

你把多少人马渡过彼岸，你把滚滚流水 送 向远方，

微风 送 来阵阵花香。

今天将由中国制造的火箭，从这里 送 上太空。

他们提防有人给苇塘里的人 送 来柴米。

就请准备一本毕业纪念册，给你的同学 送 上你的照片和心语

当他满了两周岁的时候，我们决定把他 送 托儿所了。

白发的母亲 送 枪给儿子，去打击日本侵略者；

左边，是渡江前夕，农民运军粮、妇女 送 军鞋等热烈支援前线的场面。

它们常常直竖着身子坐着，用前爪往嘴里 送 东西吃，

评选出有价值的调查报告 送 给新闻单位或环保部门。

钱学森突然被联邦调查局非法逮捕，送到特米那岛关押了15天。

有一天开明书店送了几册新出版的《音乐入门》来。

各地民众、海外华侨也纷纷来电声援，并送来了大批慰问品。

分析：“送”是典型的义项之间区分困难的多义词。义项①和义项②之间存在某种程度的交叉。在“他亲自送礼物到家来给我”这样的语句中，送同时包含两个意思，既有“赠送”的意思，又有“把东西运去或拿去给人”的意思，二者结合在一起，很难区分。事实上，在很多情况下，做“赠送”解时，“送”也可以包含有“把东西运去或拿去给人”的意思。此外，类似“把卫星送入太空、他把孩子送进学校读书、把东西送到嘴里”这样的句子中，很难为其中的“送”选择合适的义项，这里，“送”存在义项缺失的情况。

2）转/v#zhuan4

①旋转：轮子～得很快。

②绕着某物移动；打转：～圈子丨～来～去。

牛顿得意地转着风车，大家也夸奖他做得好。

两个人走马灯似的转了三四圈，

狡猾的狐狸眼珠子骨碌一转，扯着嗓子对老虎说：

想来会有不少不发光的行星绕着它们转的吧。

塔克兴奋起来，随着乐曲疯狂地转啊转啊，

留学生会馆的门房里有几本书买，还值得去一转；

希望能找到一块值得保存的文物作纪念，但转了半天一无所获，

分析：“转”的义项①和义项②区分过细，二者在意义上存在很大程度重叠，不易准确区分。《现代汉语规范词典》在释义时将二者合并成一个义项，从词义区分的角度看更为可取。此外，“转”还存在义项缺失问题，例如“去商场转转、转了半天一无所获”这样的句子中，“转”的意思两个义项都不符合，似应增加“闲逛”这样的释义作为单独的义项。

3）发现/v#fa1/xian4

①经过研究、探索等，看到或找到前人没有看到的事物或规律：～新的基本粒子丨有所发明，有所～，有所创造。

②发觉：这两天，我～他好像有什么心事。

关键是要善于从美的事物中 发现 美，并用美的语言表现美。

遥测 发现 目标！雷达发现目标！

试修改自己最近的一篇作文， 发现 有书写不合规范的标点符号，就仔细改正。

傍晚时，他在一条小河边 发现 了一片灯心草丛。

他们 发现 了一个活着的动物，可是很难把它称做人。

读书就是要善于 发现 问题，善于思考，

一位朋友在这个可怜人的床上 发现 一张便条，

分析：从释义看，“发现”的两个义项具有很好的区分度，人很容易分开这两个义项。但从语料例句来看，“发现”只解释为这两个义项是不充分的。上面给出的语料例句中的“发现”都很难被认定符合义项①或②，似应增加“找到”作为独立的义项。《现代汉语搭配词典》设“找到”作为义项，将现汉的义项①合并到“找到”这个义项中，从词义区分的角度看也是可行的处理方式。

8.4 义项之间存在重叠关系

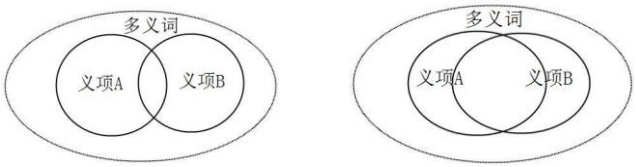

图 6：义项间存在重叠关系

词义之间存在着语义间的关联是多义词的重要性质，因此辞书划分出的多义词不同义项之间也必然存在语义上的重叠现象。此时义项间的区分需要通过其他的原则，例如搭配信息来明确。若缺乏其他区分信息，重叠部分的词义就会造成词义标注过程中的判断困难。

如果两个义项的大部分意义都是相近的，各自只有少部分区别性的词义，人在区分两个义项时，看到的都是其语义相同的部分，语义的区别不易被准确辨别出来，就会造成消歧困难。对部分义项重叠程度高的义项，辞书编纂过程中进行适当的合并是较好的处理方式。若大部分人都认识不到两个义项间的差别，义项分立就是不必要的。

下面语料库通过语料实例分析这个现象。分析所用全部语句都来自于

本文研究所使用的语料库。

• 名词

1） 饭店/n#fan4/dian4

①较大而设备好的旅馆：北京～。

②饭馆。

懒惰的 饭店 服务员往往是最令人满意，最优秀的。

第二天，老杨把那束鲜花还给了 饭店 。

有了这张卡片，他在 饭店 吃住就方便多了。

可是现在只要凭一张卡片，在 饭店 用餐，在旅馆住宿，

在面海而立的 饭店 里品尝来自大海的珍肴，又是另一番情趣。

分析：“饭店”两个义项的概念义不同，指两个不同的又关联紧密的事物。其难以区分的原因通常在于难以根据上下文判断准确的所指。

2） 孤儿/n#gu1/er2

①死了父亲的儿童：～寡母。

②失去父母的儿童：～院。

它们仿佛成了失掉了母亲的 孤儿 ，不久就会微笑不下去。

一个偶然的机会，他收 孤儿 韩子奇为徒。

我的愿望没能实现，从此我就成了没有母亲的 孤儿 。

孔繁森用献血所得的营养费，帮助这三个 孤儿 上学读书。

很久很久以前，有个 孤儿 跟着哥哥嫂子过日子。

分析：义项划分过细，义项①应当是古汉语中的意义，不宜单独作为义项收入，否则容易因为语境信息不足造成义项区分困难。

3） 节日/n#jie2/ri4

①纪念日，如五一国际劳动节等。

②传统的庆祝或祭祀的日子，如清明节、中秋节等。

我独自一人，倾听着田野的 节日 音乐会，

甜甜的一杯春酒，是 节日 的珍品，

小镇上锣鼓喧天，鞭炮声声，充满了 节日 的气氛。

今天是我的 节日 ，妈妈一定会给我买那条裙子。

给这条饱经沧桑的小街增添了 节日 的喜庆气氛。

每当 节日 到来，天安门广场更是花团锦簇。

分析：“节日”的两个义项划分过细，真实文本往往难以提供足够的信息用于区分两个义项。从语料看需要一个泛指义项来笼统指称，否则要准确区分义项①和义项②非常困难。

• 动词

1） 出现/v#chu1/xian4

①显露出来：比赛开始前半小时运动员已经～在运动场上了。

②产生出来：近年来～了许多优秀作品。

天上挂什么云，就将 出现 什么样的天气。

在太阳和月亮的周围，有时会 出现 一种美丽的七彩光圈

还有一种云彩常 出现 在清晨或傍晚。

即便是在阳光普照的时候，也难免 出现 短暂的阴云。

这些古老的爬行动物在南极的 出现 ，说明恐龙确实遍布于世界各地。

然而它们只 出现 在沙子被强烈挤压的地方。

分析：从释义上看，“出现”的两个义项一个指“已有的东西显露”；一个指“新的东西产生”。但在真实语料中，这二者往往不能得到准确的区分，不同人容易有不同的观点，人工标注一致性很差。这两个义项的区分，需要用到一定的生活常识或科学知识。

2） 发表/v#fa1/biao3

①向集体或社会表达(意见)；宣布：～谈话丨～声明。

②在刊物上登载(文章、绘画、歌曲等)：～论文。

你随时可以把自己写的文章在网上 发表 ，和别人交流。

二是发表在作者的个人网站上；三是 发表 在其他网站的留言版上。

许多同学喜欢写作文，因为写作中可以自由地 发表 自己独特的见解和感受。

写一篇简单的议论文， 发表 你的看法。

新闻是目前世界上 发表 量最大、受众最多的文体，

可以向有关部门提交调查报告，也可以正式 发表 调查报告。

分析：“发表”的两个义项都有表达观点意见的意思，其核心区别是方式和途径不同。两个义项的释义存在交叉的现象，例如“在报纸上发表声

明、在会议上发表论文”等等，此时义项不易区分。

3）够/v#gou4

①数量上可以满足需要：钱～不～？｜老觉得时间不～用。

②达到某一标准或某种程度：～格｜～条件|绳子～不～长？

也就是刚刚 够 生活罢了。

"天呀！回家后 够 我受的了！

玩 够 了时总要带些纪念品回去。

笋芽儿看看这儿，看看那儿，怎么也看不 够 。

让瑞恩通过自己干活来攒 够 2000 元，实在是太困难了。

分析：“够”的两个义项存在着释义交叉的问题，数量上满足需要也可以理解为达到某种标准或程度，难以区分的原因是因为义项区分过细，使得校对人在标注词义时不易准确判断。事实上，在《现代汉语规范词典》中，编者将这两个义项合并为一个，即“满足或达到需要的数量、标准等”。

4）救/v#jiu4

①援助使脱离灾难或危险：～命｜挽～｜营～｜搭～｜抢～。

②援助人、物使免于(灾难、危险)：～亡｜～荒｜～灾｜～急。

上帝既然用神力把我从死亡里 救 出来，一定也会救我脱离这个境地。

这才把费了多少年心血建造的其他宫殿 救 了下来。

看你不道得 10 舍施了茶汤，便又 救 了我们热渴。

到 2 号上午 10 点，二号冲锋舟共 救 起六十四人，

村民们在事实面前服了：治理大山工程 救 了全村。

分析：“救”的两个义项释义大部分重叠，都是“使脱离（免于）灾难、危险”，因此存在判断上的困难。从“救”的具体使用来看，其义项②可以大部分归并到义项①，重新解释为“使灾难、危险终止”，这样能更好的区分开两个义项，也有利于词义辨别。

5) 准备/v#zhun3/bei4

①预先安排或筹划：精神～｜～发言提纲｜～一个空箱子放书。

②打算：春节我～回家｜昨天我本来～去看你，因为临时有事没去成。

准备 从在校学生中招聘一批小记者。

她说 准备 到我家去一次，问我什么时候在家。

准备 等下次收到粮食时把粮食捣成面粉来做面包。

但他日夜盼着， 准备 着要造一栋有高台阶的新屋。

敌人的舰队正停泊在港里， 准备 一有顺风就驶出港口。

说是许多高中部的学生都 准备 在假期里出去勤工俭学。

父亲就是这样 准备 了大半辈子。

拿破仑正 准备 发起一次决定性的攻击。

归途中，他满怀希望， 准备 为国效劳。

分析：“准备”的义项①和义项②的释义存在交叉问题，“打算”与“预先安排”和“筹划”存在意义上的关联，因此不易准确区分，上述例句即体现了这一点。此外，在“我正准备去美国”这样的句子中，两个义项都符合语义，需要通过其他的语境信息来做倾向性的判断。

• 形容词

1) 单纯/a#dan1/chun2

①简单纯一；不复杂：思想～｜情节～。

②单一：～技术观点｜～追求数量。

是 单纯 的日子，也是多变的日子，

他是这样复杂，又是这样 单纯 ；

在一片 单纯 明亮的背景前突然出现一座长桥，

电视如果只是 单纯 地剥夺我们的时间，也就没有那么可恶了，

在其道德品质方面，也许比 单纯 的才智成就方面还要大。

分析：“单纯”的义项①和义项②在释义上有交叉，“单一”和“简单纯一”在意义上相近，因此造成义项区分困难。

2) 基本/b#ji1/ben3

①根本的：～矛盾｜～原理。

②主要的：～条件｜～群众。

科学的广义概念是知识， 基本 要素也是知识。

对科学而言，寻求规律是最本质、最 基本 的任务。

简·爱的人生追求由两个 基本 "旋律"构成。

语段既是词、句的综合运用，又是文章的 基本 结构，

由此我们可以了解，为什么 基本 知识上的突破是不常有的事情。

可以知道一篇文章或一本书的 基本 内容，

分析：两个义项在语义上相近，具有很大的重叠部分，即根本的也可能是主要的，主要的也可能是根本的。从语料词义标注来看，义项的判断带有很强的主观性，校对者在其中加入了自己对事物的认识，因此词义标注一致性非常低。

3) 整齐/a#zheng3/qi2

①有秩序；有条理；不凌乱：～划一丨服装～丨步伐～。

③外形规则、完整：山下有一排～的瓦房。

④大小、长短相差不多：出苗～丨字写得清楚～。

相反地，它的头部比例 整齐 ，

在碑身背面，一行行镏金字 整齐 地排列着，

周围镶嵌着不大 整齐 的石头，石头上长着青苔。

我真想拿把剪刀替它们剪一剪，因为太不 整齐 了。

球体上有一道道从上到下、排列 整齐 的纹理。

又慢又不 整齐 ，于是人们发明了锯；

整齐 的前刘海下面嵌着一对大眼。

分析：“整齐”的三个形容词义项之间语义联系紧密，存在意义上的交叉，因此在语料中存在不易区分的情况。

4） 地道/a#di4/dao5

①真正的；纯粹：她的普通话说得真～。

②(工作或材料的质量)实在；够标准：他干的活儿真～。

您试试这个！刚装来的， 地道 英国造，又细又纯！

东西真 地道 ，传家的玩艺！

我们都是 地道 老好人！

这些烟草长大后，就成了一株株 地道 的人造发光植物。

分析：“地道”的义项①和义项②划分过细。两个义项不仅意义相近，而且用法相似，在词义标注过程中不易准确区分。从词义区分的角度看，

两个义项应当合并为一个。

5）经典/a#jing1/dian3

③著作具有权威性的：马列主义～著作｜～作家。

④事物具有典型性而影响较大的：～影片。

本单元四篇文章都堪称 经典 作品，

若干年后这样的旋律会成为享誉世界的 经典 音乐作品，

阿炳的音乐成为许多世界级交响乐团的 经典 演奏曲目。

《二泉映月》已成为代表中国民族音乐的世界性 经典 曲目。

读 经典 作家的作品，

分析：“经典”的两个形容词义项区分过细，两个义项在语义存在很大部分的重叠，“具有权威性”往往意味着“具有典型性而影响较大”。正是这种语义上的重叠，造成词义标注过程中的区分困难。

解决由词义重叠带来的难以区分问题的一个有效方法是重组词义区分不明显的义项。通过对《现代汉语词典》、《现代汉语规范词典》和《现代汉语搭配词典》的观察，可以发现三部词典在词义的划分上存在很多不同之处。义项的合并是其中最典型的差异，即一部词典分为两个义项的，另一部词典处理为一个义项。而这些经过合并的义项常常存在词义区分不明显的问题。例如，形容词“地道”，义项②“真正的；纯粹”和义项③“（工作或材料的质量）实在；够标准”存在区分不明显的问题。在“他普通话说的真地道”和“他能说地道的普通话”这样的句子中，人工标注的一致性非常低，说明人不能很好的区分开这两个义项。将义项进行适当的合并是可行的解决方法。

当然，语言词典自有其面向的对象，仅从词义消歧的角度讨论义项该如何划分是不充分的。Kilgarriff（1997）指出词义消歧研究不能期待一个适合于所有应用的词义体系，提出可以根据应用需要对词典释义体系进行适当改造的思想，这对解决基于词典的语料库词义标注中的词义区分困难问题是一种可行的办法。

8.5 义项之间存在包含关系

义项间的包含关系是指多义词的多个义项中，某个义项在语义上包含另外一个义项。这种情况在名词中最为多见。

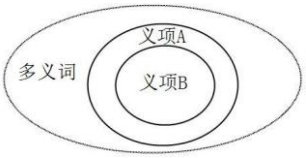

图 7：义项间存在包含关系

下面是根据语料实例对存在义项包含关系的多义词所做的分析。

1）大雁/n#da4/yan4

①鸿雁(鸟名)。　②泛指雁类。

大雁 是春天的使者。

后来它向 大雁 学飞行，向老鹰学打猎，也都是如此。

有了守夜的雁， 大雁 就能防备打雁的人来打了。

一只 大雁 从远处慢慢地飞来，边飞边鸣。

可是，他们抬头一看， 大雁 早已飞得无影无踪了。

从飞行整齐的 大雁 那里，我懂得了纪律的重要。

分析：很难区分语料中的“大雁”是专指“鸿雁”还是泛指雁类。特指和泛指的区分困难是人工标注时会碰到的普遍问题。

2）对面/n#dui4/mian4

①对过儿：他家就在我家～。　②正前方：～来了一个人。

宏儿没有见过我，远远的 对面 站着只是看。

我的床是光谱太空舱后面的一堵墙， 对面 的地板上有一台通气扇。

他坐在你的 对面 、毫无卖弄意味地与你侃侃而谈，

小白杨挺直了腰杆，迎接 对面 扑来的风沙，

他的照相至今还挂在我寓居的东墙上，书桌 对面 。

分析：“对面”的两个义项都指示方位，但义项①的范围要大于义项②。从意义上看，二者是一种包含的关系。在标注词义时，这成了难以区分的原因。

3）干部/n#gan4/bu4

①国家机关、军队、人民团体中的公职人员（士兵、勤杂人员除外）。

②指担任一定的领导工作或管理工作的人员：工会～｜科室～。

两位 干部 一看陶影说得这样宁静，反倒有些无措。

我到家属基地去，那里的一位 干部 带我们去看托儿所。

于是，党员和 干部 挨家挨户地做工作，讲道理。

常有学生、 干部 、街道积极分子到我们这儿参观

这是 干部 职工的新住房。

分析：从释义上看，“干部”的两个义项所指的对象群体存在不同，但两个义项明显存在着包含关系，义项①是义项②的一个特指。因此在语境中两个义项很难得到准确的区分。

4） 劳动/n#lao2/dong4

①人类创造物质或精神财富的活动：体力～｜脑力～。

②专指体力劳动：～锻炼。

原野到处有一种鸣叫，天空清亮透明， 劳动 的声音从这头响到那头。

为了创建圆明园，曾经耗费了两代人的长期 劳动 。

这是他一年来没日没夜 劳动 的成果，

背诵是一种艰苦的 劳动 ，光靠理解还远远不够。

没有奴隶的 劳动 ，哪里可能有什么金字塔。

我的 劳动 没有白费，所以我感到幸福。

分析：“劳动”的两个义项是包含关系，义项①包含义项②。义项②为专指义，不易区分出来。

5） 马路/n#ma3/lu4

①供车马行走的宽阔平坦的道路。 ②泛指公路。

你要注意自己在 马路 上的行为。

在街头，常有各色衣服像"万国旗"一样晾晒在 马路 边，

一辆辆汽车，奔驰在宽阔的 马路 上；

北京有许多又宽又长的柏油 马路 。

可是你不难发现，不少行人却在 马路 上横冲直撞，

分析：“马路”的两个义项存在包含关系，义项①包含义项②。此外，义项②释义也有问题，这里“马路”应当是专指义，指公路这种特殊的道

路。在语料中，两个义项不易区分。

多义词义项间的包含关系给词义标注带来很大的困扰。由于一个义项的语义可以完全涵盖另一义项，准确标注词义成了一个难题，标注者很容易在这个问题上出现处理不一致的情况。对词义标注来说，这个问题可以通过制定规则来解决。首先提取词典中存在义项包含关系的多义词表，然后规定该表中多义词的义项标注方法，为标注者提供直接的参照标准。

8.6 小结

人工词义标注中遇到的困难总是和词典的义项划分联系在一起。义项划分是词典编纂中的难题。对基于词典的词义标注来说，词典的义项划分是否颗粒度适当、义项间是否有合适的区分特征是决定词义标注正确率和一致性的关键。

第九章　结语

本文通过基于《现代汉语词典》的词义标注语料库建设，对词义标注过程中，从词形到词义的词义判断过程进行了分析，说明了读音、词类、搭配、义项频率分布等信息对词义消歧的作用。本文还对词义自动消歧和人工标注中遇到的难点进行了归纳总结，通过语料数据分析了造成词义区分困难的具体原因，并提出解决方法。

首先，以读音作为消歧线索可以有效的降低词义消歧的复杂度。多音词多是高频词，101 个多音词的总词频约占语料库 8%，体现了利用读音消歧的潜力。但从利用拼音准确区分词义来看，其作用是有限的。读音区分词义的重点在于同形词的区分。对大部分同形词而言，拼音可以降低歧义复杂度，但不能唯一确定它们的词义。词义消歧是一个极为复杂的问题，拼音作为词语的外在形式之一，对词义消歧的作用应该被肯定和利用。

其次，词类是降低词义消歧的复杂度的最为有效的消歧线索。通过词类标注，我们可以缩小多义词的词义范围。在词类信息的帮助下，对一部分多义词，可以缩小其义项的数量；对另一部分词来说，通过词类可以完全判断其词义。语料库中有 15.3%的多义词经过词类区分后可确定词义，这体现了词类对词义消歧的重要作用。

第三，搭配是最重要的词义消歧线索之一。本文通过抽取多义词不同义项的搭配词并进行分析，说明了搭配对区分词义的作用和局限。利用搭配信息进行词义消歧的一个重要条件是必须分义项给出搭配词。而目前分义项给出搭配的搭配词典和义项标注语料库都很缺乏。本文研究所建设的词义标注语料库对利用搭配进行词义消歧研究将有极大帮助。

第四，义项频率分布不均衡是多义词的一个普遍特点，充分利用义项的频率分布特点对提高词义消歧的准确率有重要作用。本文研究中，通过人工标注多义词的最常用义即可达到 69%的准确率。可见，充分利用多义词的义项频率分布特点，高准确率的词义消歧系统不是不可实现的。

由于词义区分不明显的多义词的存在，到一定程度之后，词义消歧的准确率很难进一步提高。要解决这个问题，需要对多义词词义的区分特点进行具体分析。本文通过对消歧过程进行分析，总结了自动消歧中的两个

主要难点：一个是一部分多义词的不同义项的上下文相似度很高，对利用上下文信息消解歧义造成了困难；一个是一部分多义词的义项区分需要通过纯语义理解进行。这两个难点给目前主流的以依靠形式特征消解歧义的词义消歧系统造成困难。如果对不同类型的多义词采取有针对性的消歧方法，词义消歧准确率可以进一步得到提高。

本文还通过总结人工标注过程中遇到的难点，对多义词词义难以区分的情况做了分析。多义词义项之间缺乏区分线索和义项之间存在的语义上的重叠和包含关系是造成人工标注困难的主要原因，使得人工标注存在大量的不一致现象。要解决这个问题，必须从词典的义项划分和释义入手。如果词典能够给出更多的用于区分义项的线索，例如更充分的释义、更充足的例证信息，将有助于提高人工标注词义的一致性。

本文的研究从语料库词义标注出发，从实践上探讨了拼音、词类等信息对词义消歧的作用；从理论上看，词义标注语料库建设过程中反映出的词义不易区分的问题对词典编纂、词汇研究等都有重要参考价值。

目前，汉语的词义消歧研究中，基于语言词典的词义标注语料库建设研究非常有限，本文在这方面做了尝试。有了大规模词义标注语料库的支持，对多义词的不同义项分别进行基于语言使用的描述成为可能，这将推动词义消歧研究的发展。

本文的研究对词典编纂也可以起到很好的帮助作用。根据对词典义项在语料库中进行覆盖程度检查，词典的义项划分对词义在语言中的具体表现注重不够。事实上，没有大规模的标注词典义项的语料库的支持，词典编纂也很难通过利用现代信息技术在义项划分的一致性和可靠性方面取得进步。语料库用于词典编纂是词典研究的一个发展方向，本文研究所建设的词义标注语料库对促进基于语料库的词典编纂无疑具有重要意义。

本文的研究主要存在两个方面的不足。其一是对搭配区分词义的量化研究不够。因为通过搭配区分词义的前提条件是具备多义词不同义项的搭配信息。本文研究过程中，词义标注语料库尚未完成，分义项的搭配信息不能通过统计得到，因此采用《现代汉语搭配词典》和人工标注的搭配进行词义消歧。从具体实践来看，上述两种类型的搭配信息有很大的局限性，仅能覆盖大规模语料库中的一小部分词语，通过搭配消解歧义的准确率不

高。其二，本文对词义标注中的难点进行了分析，并总结了难以消歧的多义词的类型，但由于词义标注语料库建设是和论文写作同步完成的，因此论文未能充分利用标注语料对全部多义词进行更为全面的分析。当然，本文研究的重点在于说明从词形到词义的词义标注过程中，各种语言信息所能起到的作用，并对词义消歧的难点进行分类分析，对词义消歧方法的研究仅限于构建词义标注语料库，没有进行深入探讨。

附录 1　词义标注语料样例

说明：　1）以#号引导的行为格式行，包含语料来源信息。

2）语料中标注了词类、拼音和多义词义项编号。

3）格式为：词语的第一个“/”号后是词类标记；“#”号后是拼音；“^”号后是义项编号；拼音采用数字后标调方式。

4）单义词、虚词、名动形以外的实词不标注词义。

#语料编号：XXYWBSD03039

#篇名：九年义务教育五年制小学试用课本(语文)(北京师范大学出版社)

#期数：第六册

#时间：1999

#作者：“五四”学制教材总编委会

#版面：第 4 课

#标题：雨后春笋

4/m 雨后春笋/i#yu3/hou4/chun1/sun3

春 雨 /n#chun1/yu3 过 后 /n#guo4/hou4 ， /w 一 /m#yi1 夜 /q#ye4 之 间 /f#zhi1/jian1 。/w 竹林/n#zhu2/lin2 里/f#li5 冒/v#mao4^1 出/vd#chu1 一/m#yi1 片 /q#pian4 嫩 /a#nen4^1 笋 /n#sun3 ， /w 显 得 /v#xian3/de5 生 机 勃 勃 /i#sheng1/ji1/bo2/bo2 。/w 笋/n#sun3 怎么/r#zen3/me5 会/v#hui4^5 从/p#cong2 土 /n#tu3^1 里/f#li5 钻/v#zuan1^2 出来/v#chu1/lai2 呢/y#ne5 ？ /w 是/v#shi4 人们 /n#ren2/men5 播/v#bo1^2 下/vd#xia4 了/u#le5 种子/n#zhong3/zi5^1 吗/y#ma5 ？ /w 不/d#bu4 是/v#shi4 。/w 是/v#shi4 种子/n#zhong3/zi5^1 自己/r#zi4/ji3 落 /v#luo4^1 在/p#zai4 地/n#di4^2 上/f#shang4 长/v#zhang3^1 出来/v#chu1/lai2 的 /u#de5 吗/y#ma5 ？/w 也/d#ye3 不/d#bu4 是/v#shi4 。/w

笋/n#sun3 是/v#shi4 竹子/n#zhu2/zi5 初/a#chu1 出土/v#chu1/tu3^2 的/u#de5 幼芽/n#you4/ya2 。/w 竹子/n#zhu2/zi5 是/v#shi4 靠/v#kao4^4 竹鞭/n#zhu2/bian1 上 /f#shang4 的 /u#de5 芽 /n#ya2^1 繁 殖 /v#fan2/zhi2 的 /u#de5 。 /w 竹 鞭

/n#zhu2/bian1 的/u#de5 每个/r#mei3/ge4 节/n#jie2^1 上/f#shang4 长/v#zhang3^1 着/u#zhe5 许多/m#xu3/duo1 须/n#xu1 ，/w 那/r#na4 是/v#shi4 竹子/n#zhu2/zi5 的/u#de5 根/n#gen1^1 ，/w 有的/r#you3/de5 节/n#jie2^1 上/f#shang4 会/v#hui4^B5 长/v#zhang3^1 出/vd#chu1 一/m#yi1 两/m#liang3 个/q#ge4 幼芽/n#you4/ya2 。/w 这些/r#zhe4/xie1 幼芽/n#you4/ya2 有的/r#you3/de5 向/p#xiang4 横/n#heng2^8 里/f#li5 长/v#zhang3^2 ，/w 成为/v#cheng2/wei2 竹鞭/n#zhu2/bian1 ；/w 有的/r#you3/de5 发育/v#fa1/yu4 成/v#cheng2^3 笋/n#sun3 ，/w 向上/v#xiang4/shang4 破土而出/l#po4/tu3/er2/chu1 ，/w 长/v#zhang3^2 成/v#cheng2^3 新/a#xin1^1 的/u#de5 竹子/n#zhu2/zi5 。/w

笋/n#sun3 在/p#zai4 冬天/t#dong1/tian1 开始/v#kai1/shi3^2 孕育/v#yun4/yu4 ，/w 入冬/v#ru4/dong1 以前/f#yi3/qian2 发育/v#fa1/yu4 得/u#de5 又/d#you4 肥/a#fei2^1 又/d#you4 壮/a#zhuang4^1 ，/w 这/r#zhe4 就/d#jiu4 是/v#shi4 冬笋/n#dong1/sun3 ，/w 挖/v#wa1 出来/v#chu1/lai2 可以/v#ke3/yi3^2 做/v#zuo4^1 菜/n#cai4^3 吃/v#chi1^1 。/w

春天/t#chun1/tian1 来/v#lai2^1 了/u#le5 ，/w 气温/n#qi4/wen1 回升/v#hui2/sheng1 ，/w 沉睡/v#chen2/shui4 了/u#le5 一/m#yi1 冬/Tg#dong1 的/u#de5 笋/n#sun3 开始/v#kai1/shi3^2 萌动/v#meng2/dong4^1 。/w 遇到/v#yu4/dao4 一/m#yi1 场/q#chang2 好/a#hao3^1 雨/n#yu3 ，/w 它们/r#ta1/men5 就/d#jiu4 会/v#hui4^5 争先恐后/i#zheng1/xian1/kong3/hou4 地/u#di5 冒/v#mao4^1 出/vd#chu1 地面/n#di4/mian4^1 来/vd#lai2 。/w

雨/n#yu3 后/f#hou4 的/u#de5 春笋/n#chun1/sun3 不但/c#bu4/dan4 长/v#zhang3^2 得/u#de5 多/a#duo1^5 ，/w 而且/c#er2/qie3 长/v#zhang3^2 得/u#de5 很/d#hen3 快/a#kuai4^1 。/w 有人/r#you3/ren2 做/v#zuo4^3 过/u#guo4 测定/v#ce4/ding4 ，/w 以/p#yi3 毛竹/n#mao2/zhu2 为/v#wei2^2 例/n#li4^1 ，/w 它/r#ta1 在/p#zai4 拔/v#ba2^1 节/n#jie2^1 时/n#shi2^1 ，/w 一/m#yi1 天/q#tian1 一/m#yi1 夜/q#ye4 可以/v#ke3/yi3^2 长/v#zhang3^2 一/m#yi1 米/q#mi3 。/w 如果/c#ru2/guo3 你/r#ni3 搬/v#ban1^1 个/q#ge4 小/a#xiao3^1 凳/n#deng4 坐/v#zuo4^1 在/p#zai4 那里/r#na4/li5 ，/w 甚至/c#shen4/zhi4 可以/v#ke3/yi3^2 听到/v#ting1/dao4 它/r#ta1 长/v#zhang3^2 高/a#gao1^1 时/n#shi2^1 发出/v#fa1/chu1^1 的/u#de5 叭/o#ba1 叭/o#ba1 的/u#de5 声音/n#sheng1/yin1 。/w 今天/t#jin1/tian1

它/r#ta1 比/p#bi3 你/r#ni3 矮/a#ai3^1 ，/w 明天/t#ming2/tian1 它/r#ta1 高/a#gao1^1 你/r#ni3 一/m#yi1 头/n#tou2^1 。/w

人们/n#ren2/men5 用/v#yong4^1 “/w 雨后春笋/i#yu3/hou4/chun1/sun3 ”/w 来/v#lai2^7 形容/v#xing2/rong2 大量/m#da4/liang4 出现/v#chu1/xian4^2 的/u#de5 、/w 发展/v#fa1/zhan3^1 十分/d#shi2/fen1 迅速/a#xun4/su4 的/u#de5 新生事物/l#xin1/sheng1/shi4/wu4 ，/w 是/v#shi4 很/d#hen3 有/v#you3^2 道理/n#dao4/li3^2 的/u#de5 。/w

附录 2　语料库多义词表

说明：

1. 本表列出了语料库中词频 20 次以上的、出现至少两个义项的多义词 1028 个；

2. 表中，义项数指该词在语料库中出现的义项的个数；频次指该词在语料库中的出现次数；

3. 本表按音序排列。

词	词类	义项数	频次
矮	a	2	64
爱	v	4	578
安	v	2	33
扒	v	3	31
白	a	2	388
摆	v	3	136
拜	v	3	46
班	n	4	156
搬	v	2	117
办	v	3	104
办公室	n	2	27
帮	v	2	168
包	n	3	22
包	v	4	81
宝贝	n	2	33
饱	a	2	48
饱满	a	2	26
保留	v	3	22
保证	v	2	39
报	v	4	28
抱	v	3	249
暴雨	n	2	33
爆发	v	2	25
悲剧	n	2	33
背	n	2	99

词	词类	义项数	频次
背	v	2	155
背	v	4	92
背景	n	3	60
本子	n	2	26
笨	a	2	25
比	v	4	129
笔记	n	2	21
笔尖	n	2	31
边	n	4	349
编	v	4	62
编写	v	2	34
扁豆	n	2	20
变	v	3	426
表	n	2	54
表面	n	2	51
表示	v	2	271
表现	v	2	205
别	v	3	36
冰冷	z	2	23
兵	n	2	76
拨	v	2	36
玻璃	n	2	108
博士	n	3	51
搏斗	v	2	37
薄	a	2	58

词	词类	义项数	频次
补	v	3	52
补充	v	2	40
不错	a	2	68
不行	a	3	35
不行	v	3	42
不幸	a	2	77
不止	v	2	33
不足	v	3	28
布置	v	2	36
擦	v	4	111
材料	n	4	161
采	v	2	77
菜	n	2	67
参加	v	2	176
参考	v	2	70
藏	v	2	90
草	n	2	244
草地	n	2	226
草莓	n	2	21
层次	n	3	62
插	v	2	124
查	v	2	52
茶	n	2	71
差	a	3	38
缠	v	2	52
产	v	2	29
长	v	3	557
尝	v	2	65
场面	n	2	30
抄	v	3	37
超过	v	2	69
吵	v	2	24
车子	n	2	42
扯	v	2	37
沉	v	3	66

词	词类	义项数	频次
沉静	a	2	26
沉重	a	2	81
闯	v	4	31
撑	v	5	57
成	v	4	1192
成分	n	2	54
成立	v	2	41
呈	v	2	61
城	n	3	100
程度	n	2	66
吃	v	4	1474
持	v	2	20
翅膀	n	2	200
充分	a	2	57
充满	v	2	232
冲	v	3	193
抽	v	4	92
出	v	9	2560
出发	v	2	88
出门	v	2	51
出现	v	2	319
处	n	2	209
处理	v	4	56
触动	v	2	20
穿	v	5	422
传	v	4	129
窗口	n	4	49
创造性	n	2	31
吹	v	4	347
词	n	3	165
凑	v	3	33
粗	a	5	55
粗糙	a	2	28
粗壮	a	2	26
存	v	4	22

词	词类	义项数	频次
搭	v	5	87
搭配	v	2	27
答应	v	2	108
打	v	21	855
打发	v	4	21
打开	v	2	127
打量	v	2	28
大	a	3	3163
大伯	n	2	44
大陆	n	2	56
大气	n	2	24
大嫂	n	2	27
大雁	n	2	63
代表	n	4	107
代表	v	2	60
代价	n	2	34
带	v	8	789
带领	v	2	41
单	b	2	25
单位	n	2	110
弹	v	3	56
淡	a	2	32
当	v	2	279
挡	v	2	60
刀	n	2	91
倒	v	3	65
到	v	3	6002
道	n	2	52
道	v	2	468
道理	n	3	153
得	v	5	258
登	v	2	84
蹬	v	2	26
等于	v	2	34
瞪	v	2	50

词	词类	义项数	频次
低	a	3	152
低头	v	2	36
滴	v	2	60
底	n	2	59
抵	v	2	25
地	n	8	735
地方	n	2	811
地面	n	2	126
地区	n	3	81
点	n	5	120
点	v	7	183
电话	n	3	116
电视	n	3	130
店	n	2	50
掉	v	4	294
跌	v	2	33
顶	v	7	56
顶峰	n	2	36
定	v	4	87
丢	v	3	88
动	v	5	341
动静	n	2	22
动手	v	2	54
动摇	v	2	20
冻	v	3	59
斗争	v	2	82
抖	v	2	46
抖动	v	2	21
读	v	3	812
读书	v	3	329
杜鹃	n	2	31
渡	v	2	44
断	v	5	124
锻炼	v	2	42
堆	v	2	51

词	词类	义项数	频次
队	n	2	38
对	v	3	124
对手	n	2	27
对象	n	2	91
对照	v	2	27
蹲	v	2	110
多	a	2	1509
发	v	8	302
发表	v	2	114
发出	v	2	199
发挥	v	2	41
发起	v	2	24
发现	v	2	628
发展	v	2	259
法	n	2	47
法律	n	2	74
翻	v	6	179
翻滚	v	2	21
翻身	v	3	28
反	v	2	47
反射	v	2	34
反应	v	2	21
反映	v	2	88
犯	v	2	36
饭	n	3	162
方向	n	2	144
放	v	12	742
飞	v	4	902
肥	a	3	38
沸腾	v	3	27
分	v	3	226
分开	v	2	25
分离	v	2	20
风暴	n	2	29
风雨	n	2	39

词	词类	义项数	频次
锋利	a	2	24
缝	n	2	27
扶	v	2	58
浮现	v	2	20
符号	n	2	27
负	v	3	36
附	v	2	29
该	v	3	236
改	v	3	83
改变	v	2	91
改写	v	2	21
盖	v	4	133
干	v	2	234
干净	a	3	74
赶	v	6	148
感觉	v	2	77
感情	n	2	251
感染	v	2	20
高	a	3	820
高潮	n	2	27
高峰	n	2	26
高贵	a	3	42
高粱	n	2	33
高尚	a	2	29
搞	v	2	70
告别	v	3	68
哥	n	2	27
搁	v	3	25
歌唱	v	2	103
革命	v	2	155
隔	v	2	116
个人	n	2	82
个体	n	2	21
给	v	2	2314
根	n	2	77

词	词类	义项数	频次
工程	n	2	95
工夫	n	2	61
工具	n	2	85
工作	n	3	157
功夫	n	2	32
功课	n	3	59
攻击	v	2	29
供	v	2	73
沟	n	3	21
够	v	3	109
古人	n	2	61
谷子	n	2	30
骨头	n	2	40
鼓	v	2	55
故事	n	2	572
瓜	n	2	42
挂	v	4	270
拐	v	2	23
关	v	5	71
关系	n	4	198
管	v	5	150
管理	v	3	51
惯	v	2	25
灌	v	2	39
光	a	2	49
光明	a	2	24
广大	a	3	36
归	v	2	39
规则	n	2	47
鬼	n	2	23
桂花	n	2	55
滚	v	2	68
果实	n	2	46
过	v	4	1245
过后	n	2	26

词	词类	义项数	频次
孩子	n	2	1343
害	v	3	38
含	v	3	88
喊	v	2	240
好	a	9	2259
好处	n	2	77
好听	a	2	31
号	n	4	59
喝	v	2	290
合	v	4	57
黑	a	3	359
黑暗	a	2	96
痕迹	n	2	43
哼	v	2	23
横	a	2	25
横	v	2	33
红	a	2	539
红领巾	n	2	27
红旗	n	2	23
吼	v	3	24
厚	a	2	75
呼	v	3	22
呼唤	v	2	64
葫芦	n	2	64
花	a	2	48
花	n	5	637
化	v	3	38
画	v	2	431
怀	v	2	24
怀疑	v	2	34
坏	a	3	153
欢迎	v	2	90
环境	n	2	174
换	v	3	187
黄瓜	n	2	31

词	词类	义项数	频次
恢复	v	2	55
辉煌	a	2	36
回	v	3	553
回头	v	3	78
会	n	2	52
会议	n	2	43
混	v	3	28
活	a	2	47
活	v	2	163
活动	v	4	113
火	n	3	198
机关	n	2	25
积极	a	2	47
急	a	4	98
挤	v	3	133
计算	v	2	40
记	v	2	172
寂寞	a	2	34
寄	v	2	61
加	v	4	142
加工	v	2	29
加入	v	2	46
加油	v	2	20
家	n	2	892
价值	n	2	101
驾	v	2	27
架	v	4	50
架子	n	2	23
尖	a	3	36
简单	a	2	139
见	v	5	1223
建	v	2	83
建立	v	2	59
健康	a	2	97
将军	n	2	127

词	词类	义项数	频次
讲	v	4	400
降	v	2	23
交	v	3	67
浇	v	3	48
骄傲	a	2	75
角	n	3	51
角度	n	2	61
角落	n	2	39
角色	n	2	23
脚步	n	2	120
搅	v	2	23
教育	v	2	84
接	v	5	129
接触	v	2	26
接受	v	2	101
节奏	n	2	47
结	v	3	25
结构	n	2	149
结合	v	2	94
姐妹	n	2	24
解	v	2	33
解放	v	2	86
解释	v	2	94
介绍	v	3	218
借	v	2	235
紧	a	4	135
紧张	a	3	91
劲	n	2	58
进	v	3	931
进攻	v	2	39
经	v	2	37
经济	n	3	64
惊	v	2	72
惊醒	v	2	24
精神	n	2	339

词	词类	义项数	频次
井	n	2	123
景观	n	2	23
净	a	2	24
境界	n	2	59
静	a	2	62
旧	a	3	98
救	v	2	105
举	v	3	174
巨人	n	3	98
卷	v	2	64
决定	v	2	164
觉得	v	2	647
军	n	2	27
开	v	9	1031
开辟	v	3	29
开放	v	2	87
开花	v	2	80
开始	v	2	541
开展	v	2	62
看	v	5	3683
考	v	2	43
考察	v	2	54
靠	v	4	313
靠近	v	2	31
可怜	a	2	121
可笑	a	2	27
克隆	v	2	88
客人	n	3	135
课	n	4	98
空气	n	2	193
控制	v	2	57
口	n	4	161
口袋	n	2	86
扣	v	5	25
枯	a	2	24

词	词类	义项数	频次
苦	a	2	93
挎	v	2	24
跨	v	3	58
快	a	3	401
宽大	a	2	26
宽广	a	3	28
宽阔	a	2	44
困难	a	2	88
垃圾	n	2	27
拉	v	10	437
喇叭	n	2	24
来	v	6	4589
烂	a	3	20
浪花	n	2	74
捞	v	2	24
老	a	5	645
老汉	n	2	29
老人	n	2	328
累	v	2	20
冷	a	2	151
离	v	3	272
理	v	2	40
理会	v	5	28
力	n	2	24
力量	n	4	191
历史	n	3	281
厉害	a	2	78
立	v	5	130
利用	v	2	143
荔枝	n	2	89
连接	v	2	42
脸	n	3	615
脸色	n	3	65
练	v	2	68
练笔	v	2	29

词	词类	义项数	频次
两岸	n	2	62
亮	a	2	104
亮	v	2	84
亮光	n	2	29
了	v	2	225
了不起	a	2	41
了解	v	2	304
临	v	2	65
灵魂	n	4	88
灵活	a	2	44
领	v	2	79
令	v	2	211
溜	v	3	32
流动	v	2	38
留	v	5	241
楼	n	2	109
漏	v	3	20
路	n	2	458
路线	n	2	28
轮廓	n	2	23
萝卜	n	2	52
落	v	6	400
落地	v	2	21
马车	n	2	42
骂	v	2	83
卖	v	3	293
满	a	2	500
满	v	2	35
满足	v	2	69
慢	a	2	102
毛病	n	2	36
冒	v	2	141
没有	v	5	1708
美	a	2	448
门	n	5	404

词	词类	义项数	频次
门槛	n	2	32
朦胧	a	2	37
猛烈	a	2	22
蒙	v	2	39
密切	a	2	21
棉花	n	2	40
勉强	a	4	29
面	n	4	105
面貌	n	2	24
描	v	2	34
灭	v	2	39
民间	n	2	47
民族	n	2	161
名词	n	2	45
名字	n	2	218
明亮	a	2	71
明媚	a	2	23
鸣	v	2	24
命	n	2	45
命运	n	2	95
摸	v	4	215
模样	n	3	65
磨	v	3	71
魔鬼	n	2	42
抹	v	3	50
末	n	2	38
墨水	n	3	21
目标	n	2	93
目光	n	3	133
沐浴	v	3	31
牧场	n	2	21
拿	v	6	882
难得	a	2	21
难过	a	2	72
难受	a	2	46

词	词类	义项数	频次
闹	v	4	69
年代	n	2	89
念	v	2	157
捏	v	2	75
凝固	v	2	22
凝聚	v	2	21
扭	v	5	44
浓	a	2	54
浓郁	a	2	21
弄	v	4	180
趴	v	2	64
爬	v	3	447
怕	v	2	392
拍	v	3	163
拍打	v	2	22
抛	v	3	55
咆哮	v	2	38
跑	v	5	791
培养	v	2	59
培育	v	2	36
配	v	4	32
捧	v	2	110
碰	v	3	110
劈	v	2	43
皮	n	3	57
脾气	n	2	26
屁股	n	2	21
篇幅	n	2	24
骗	v	2	33
漂	v	2	39
品格	n	2	21
品质	n	2	38
品种	n	2	34
平	a	2	71
平衡	a	2	28

词	词类	义项数	频次
评	v	2	23
破	v	6	102
破坏	v	4	60
扑	v	3	142
葡萄	n	2	59
蒲公英	n	2	34
朴素	a	4	36
齐	a	3	21
旗帜	n	3	27
起	v	8	1674
起伏	v	2	50
起身	v	3	34
气	n	5	239
气	v	2	42
气息	n	2	50
气象	n	3	27
前后	n	2	72
钱	n	4	544
浅	a	2	42
欠	v	2	30
强	a	3	116
抢	v	2	69
桥梁	n	2	93
巧	a	3	31
亲	v	2	20
亲切	a	2	101
亲人	n	2	81
勤	a	2	25
青年	n	2	181
轻	a	4	115
轻快	a	2	34
清	a	3	199
清新	a	2	47
情感	n	2	85
情绪	n	2	44

词	词类	义项数	频次
请	v	3	706
秋风	n	2	34
求	v	3	70
球	n	3	60
曲	n	2	21
曲折	a	2	40
取	v	2	131
去	v	7	4262
圈	n	2	105
全	a	2	180
缺	v	2	47
群众	n	2	105
燃烧	v	2	35
染	v	2	50
让	v	5	1330
绕	v	3	86
惹	v	3	36
热	a	2	118
人	n	8	5599
人格	n	2	22
人家	n	2	100
人心	n	2	20
认	v	3	21
任	v	2	45
扔	v	2	113
日	n	6	78
日子	n	3	190
容	v	2	34
容易	a	2	195
柔和	a	2	36
揉	v	2	40
肉	n	2	147
如	v	3	783
入	v	2	206
软	a	4	56

词	词类	义项数	频次
撒	v	2	97
洒	v	2	52
赛	v	2	30
嗓子	n	2	58
扫	v	3	56
色彩	n	2	134
杀	v	2	108
晒	v	2	105
山沟	n	2	21
山水	n	3	50
闪	v	4	158
扇	v	2	25
上	v	11	1701
上帝	n	2	128
上升	v	2	41
上下	n	2	71
上学	v	2	80
上游	n	2	26
烧	v	3	125
少	v	2	34
少年	n	2	172
社会	n	2	343
射	v	3	84
身份	n	2	30
深	a	6	231
深沉	a	3	40
深厚	a	2	22
深刻	a	2	99
深切	a	2	24
神	n	4	75
升	v	2	95
生	v	4	252
生长	v	2	115
生活	n	2	632
生活	v	2	460

词	词类	义项数	频次
生机	n	2	34
生平	n	2	30
胜利	v	2	117
省	v	2	25
盛	a	2	21
盛	v	2	40
失败	v	2	58
失掉	v	2	26
师傅	n	2	56
石板	n	2	43
石榴	n	2	36
时	n	2	1520
时间	n	3	625
时节	n	2	25
使	v	3	1130
驶	v	2	28
世界	n	6	894
市	n	2	50
市场	n	2	44
事	n	4	870
事情	n	2	300
事业	n	2	93
收	v	7	142
手段	n	2	26
守	v	3	38
受	v	4	294
瘦	a	2	62
叔	n	2	28
熟	a	5	89
数	n	2	40
数	v	2	157
数字	n	2	32
摔	v	5	66
甩	v	3	48
水	n	2	1257

词	词类	义项数	频次
水流	n	2	35
说	v	4	6744
说法	n	2	39
说话	v	2	272
说明	v	2	193
丝	n	2	36
司令	n	2	29
思想	n	2	212
死	a	3	30
松	v	2	20
耸	v	2	21
送	v	3	487
速度	n	2	98
塑造	v	2	44
酸	a	3	41
算	v	6	227
随	v	3	110
随便	a	2	54
缩	v	2	25
所在	n	2	31
塌	v	2	21
台	n	2	52
抬	v	2	192
太太	n	2	26
太阳	n	2	538
态度	n	2	100
贪婪	a	2	21
探	v	2	37
糖	n	3	43
烫	v	2	23
讨	v	2	41
套	v	2	32
提	v	5	257
题目	n	2	98
体育	n	2	24

词	词类	义项数	频次
天	n	4	490
天才	n	2	30
天地	n	2	72
天堂	n	2	44
天下	n	2	74
天真	a	2	36
田地	n	2	29
甜	a	2	83
填	v	2	40
挑	v	2	159
条件	n	3	133
跳	v	4	440
贴	v	2	110
听	v	3	1182
停	v	3	335
停顿	v	2	47
挺	v	2	49
通	v	4	48
通过	v	2	42
同学	n	2	736
同志	n	2	475
痛快	a	3	36
头	n	4	943
头脑	n	2	35
投	v	6	80
投入	v	2	40
透	a	2	50
透	v	2	80
突出	v	2	37
涂	v	3	30
土地	n	2	200
吐	v	3	71
推	v	6	157
腿	n	2	209
退	v	5	84

词	词类	义项数	频次
吞	v	2	36
拖	v	3	102
脱	v	3	81
完	v	3	566
玩	v	2	242
挽	v	2	22
晚	a	2	66
王国	n	3	26
网	n	3	65
往来	v	2	24
忘记	v	2	116
望	v	2	619
威胁	v	2	35
为	v	3	762
围绕	v	2	63
尾巴	n	2	215
位置	n	3	68
味	n	3	54
味道	n	2	46
喂	v	2	51
温和	a	2	45
文化	n	2	191
文件	n	3	27
文章	n	2	393
文字	n	3	156
稳定	a	2	22
问	v	4	1100
问题	n	4	469
卧	v	2	38
乌云	n	2	59
污染	v	2	38
屋	n	2	136
无力	v	2	38
无聊	a	2	26
武器	n	2	56

词	词类	义项数	频次
舞	v	3	39
物质	n	2	64
西瓜	n	2	63
吸	v	3	66
吸收	v	3	47
牺牲	v	2	55
袭击	v	2	26
戏剧	n	2	75
细	a	7	162
细腻	a	2	22
细致	a	2	30
狭窄	a	3	22
下	v	12	1171
先生	n	2	583
掀	v	2	31
鲜明	a	2	75
线	n	3	135
陷入	v	2	36
献	v	2	38
乡	n	3	24
乡亲	n	2	45
香	a	4	74
香甜	a	2	21
想	v	4	1918
向	v	2	114
像	v	2	2416
消化	v	2	21
消灭	v	2	41
消息	n	2	144
小	a	2	4678
小朋友	n	2	83
小人	n	2	35
效果	n	2	54
笑	v	2	770
写	v	2	1665

词	词类	义项数	频次
写法	n	2	25
谢	v	2	46
心	n	2	496
心理	n	2	65
心思	n	3	25
心脏	n	2	41
欣赏	v	2	167
新	a	4	742
新鲜	a	4	74
信	n	2	296
信	v	2	85
信号	n	2	21
信息	n	2	159
星期	n	2	67
行	n	2	33
行动	v	2	55
形态	n	3	33
形体	n	2	21
形象	n	2	146
醒	v	2	135
凶	a	2	24
修	v	6	78
修饰	v	2	31
修养	n	2	21
秀才	n	2	21
许	v	3	30
悬	v	4	24
选	v	2	97
学	v	2	443
学科	n	2	22
学生	n	2	227
学习	v	2	632
压	v	4	92
压力	n	3	29
压迫	v	2	31

词	词类	义项数	频次
烟	n	2	126
严肃	a	2	88
严重	a	2	78
研究	v	2	310
颜色	n	2	204
眼光	n	2	56
扬	v	2	32
养	v	4	109
样	n	2	35
样子	n	3	218
腰	n	2	121
摇晃	v	2	30
摇篮	n	2	33
咬	v	3	177
要	v	8	4549
一般	a	2	264
一定	b	4	110
一面	n	2	21
一时	n	2	40
移植	v	2	21
艺术	n	2	247
意见	n	2	139
意思	n	4	267
意义	n	2	152
因素	n	2	25
音响	n	2	22
引	v	4	81
引导	v	2	22
应	v	2	155
英雄	n	2	199
迎	v	2	90
影	n	2	41
影子	n	3	127
硬	a	2	64
拥	v	2	20

词	词类	义项数	频次
涌	v	2	98
用	v	3	2459
优雅	a	2	21
游	v	2	176
有	v	7	8921
有关	v	2	272
宇宙	n	2	130
语气	n	2	60
语言	n	2	338
玉米	n	2	36
园	n	2	28
远	a	2	402
约	v	2	27
月	n	2	330
运动	n	2	25
运动	v	3	47
砸	v	2	39
在	v	5	525
在乎	v	2	31
在于	v	2	63
早	a	3	219
造	v	2	167
责任	n	2	73
沾	v	3	43
展开	v	2	152
占	v	2	59
占领	v	2	33
战斗	v	2	96
战士	n	2	213
站	n	2	53
招	v	5	27
招呼	v	3	25
照	v	3	220
照顾	v	3	45
遮	v	3	48

词	词类	义项数	频次
折	v	2	56
针	n	2	37
震	v	2	35
震动	v	2	21
争	v	2	53
争取	v	2	27
征服	v	2	30
整	a	2	170
整齐	a	3	53
正	a	3	20
正面	n	3	24
支	v	3	20
支撑	v	2	22
支持	v	2	68
知识	n	2	267
织	v	2	53
直	a	4	68
止	v	2	32
指	v	4	298
指导员	n	2	43
制度	n	2	28
制造	v	2	58
治	v	3	51
质量	n	2	34
中	v	2	74
中间	n	3	175
中心	n	4	180
钟	n	2	35
种子	n	2	100
重	a	2	167
重复	v	2	31
主持	v	2	28
主动	a	2	25
主人	n	2	127

词	词类	义项数	频次
主题	n	2	92
主席	n	2	261
主意	n	2	48
住	v	3	1048
抓	v	4	196
转	v	2	112
转	v	2	72
转身	v	2	72
装	v	2	221
撞	v	4	84
追	v	2	114
追求	v	2	84
准备	v	2	284
着	v	4	246
仔细	a	2	211
子	n	2	25
字	n	3	729
综合	v	2	53
总	a	2	67
走	v	9	2657
足够	v	2	31
组织	n	3	36
祖先	n	2	57
钻	v	3	181
嘴	n	3	254
嘴巴	n	2	55
罪	n	2	21
尊重	v	2	43
作	v	3	361
作用	n	2	151
坐	v	4	908
做	v	8	1527
做人	v	2	30
做事	v	2	25

参考书目

英文书目

Books：

Agirre, E., & Edmonds, P. G. (Eds.). (2006). *Word sense disambiguation : algorithms and applications*. Dordrecht: Springer.

Antal, L. (1963). *Questions of meaning*. Muton: The Hague.

Cottrell, G. W. (1989). *A connectionist approach to word sense disambiguation*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Fellbaum, C. (1998). *WordNet: An electronic lexical database and some of its applications*. MIT Press.

Garside, R., Leech, G., & McEnery, T. (1997). *Corpus annotation: linguistic information from computer text corpora*. Longman.

Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K. N., & Ward, N. (2000). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall.

Leech, G. N. (1974). *Semantics*. Harmondsworth: Penguin.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Ravin, Y., & Leacock, C. (Eds.). (2000). *Polysemy: theoretical and computational approaches*. Oxford University Press.

Saint-Dizier, P., & Viegas, E. (Eds.). (1995). *Computational lexical semantics*. Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Summers, D. (2006). *Longman dictionary of contemporary English* (New ed.). Harlow, Essex: Pearson/Longman.

Wierzbicka, A. (1989). *Semantics culture and cognition*. Oxford: Oxford University Press.

Zipf, G. K. (1939). *The psycho-Biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press.

Articles:

Arriola, J., Artola, X., & Soroa, A. (1996). Automatic extraction of lexical information from an ordinary dictionary. In *Proceedings of EURALEX-1996*, Goteborg, Sueden.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. *Lecture Notes in Computer Science*, pp.136-145.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: a guide to word combinations*: John Benjamins Publishing.

Church, K., Gale, W., Hanks, P., & Kindle, D. (1991). Using statistics in lexical analysis. In Z. U. (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*: Lawrence Erlbaum.

Ciaramita, M., Hofmann, T., & Johnson, M. (2003). Hierarchical semantic classification: word sense disambiguation with world knowledge. In *proceedings of the International JointConference on Artificial Intelligence*.

Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proceedings of Coling-92*.

Cunningham, H., Stevenson, M., & Wilks, Y. (1998). Implementing a sense tagger within a general architecture for language engineering. In *proceedings the The 3rd Conference on New Methods in Language Engineering*.

Dagan, I., & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4), pp.563-596.

Dang, H. T., Chia, C., Palmer, M., & Chiou, F. D. (2002). Simple features for Chinese

word sense disambiguation. In *Proceedings of COLING2002*.

Firth, J. (1957). Modes of meaning. *Papers in Linguistics* (Vol. 1957, 1934-1951). London: Oxford University Press.

Gale, W., Church, K. W., & Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language-1992*.

Guthrie, J., Guthrie L., Wilks, Y. & Aidinejad, H. (1991) Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of ACL 29*, pp.146-152.

Guthrie, L., Pustejovsky, J., Wilks, Y., & Slator, B. (1996). The role of lexicons in natural language processing. *Communications of the ACM, 39*(1), pp.63-72.

Hearst, M., & Schutze, H. (1993). Customizing a lexicon to better suit a computational task. In *Proceedings of the SIGLEX Workshop*.

Ide, N. (2000). Cross-Lingual sense determination: can it work? *Computers and the Humanities, 34*(1), pp.223-234.

Ide, N., & Veronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time. In *Proceedings of KB&KS Workshop, Tokyo*.

Ide, N., & Veronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics, 24*(1), pp.2-40.

Ide, N., & Wilks, Y. (2006). Making sense about sense. In Agirre & Edmonds (Eds) *Word sense disambiguation* (pp. 47-73). Dordrecht: Springer.

Kilgarriff, A. (1992). Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities, 26*(5), pp.365-387.

Kilgarriff, A. (1997). "I Don't Believe in Word Senses". *Computers and the Humanities, 31*(2), pp.91-113.

Kilgarriff, A. (1997). What is word sense disambiguation good for? In *Proceedings of*

the Arxiv preprint cmp-lg/9712008.

Kilgarriff, A. (1998). The hard parts of lexicography. *International Journal of Lexicography, 11*(1), pp.51-54.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? *Text, Speech and Dialogue* , pp.103-111.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics, 29*(3), pp.333-347.

Kilgarriff, A., & Koeling, R. (1999). An evaluation of a lexicographer's workbench incorporating word sense disambiguation. *Computational Linguistics and Intelligent Text Processing*, pp.109-121.

Kilgarriff, A., & Palmer, M. (2000). Introduction to the special issue on SENSEVAL. *Computers and the Humanities, 34*(1), pp.1-13.

Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities, 34*(1), pp.15-48.

Krek, S., Gorjanc, V., & Stabej, M. (2005). Dictionaries, corpora and word-formation. *Meaningful Texts: The Extraction Of Semantic Information From Monolingual And Multilingual Corpora*.

Krishnamurthy, R., & Nicholls, D. (2000). Peeling an onion: the lexicographer's experience of manual sense-tagging. *Computers and the Humanities, 34*(1), pp.85-97.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing, 8*(4), pp.275-281.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC conference*, pp.24-26.

Li, W., & Lu, Q. (2005). Integrating collocation features in Chinese word sense disambiguation. In *Proceeding of 4th Sighan Workshop on Chinese Language Processing*.

Lin, C., & Ahrens, K. (2005). How many meanings does a word have? Meaning estimation in Chinese and English. *Language Acquisition, Change and*

Emergence: Essays in Evolutionary Linguistics.

Màrquez, L., Taulé, M., Padró, L., Villarejo, L., & Martí, M. A. (2004). On the quality of lexical resources for word sense disambiguation. *Advances in Natural Language Processing*, pp.291-302.

Martin, W., Al, B., & Sterkenburg, P. (1983). On the processing of a text corpus. In R. Hartmann (Ed.), *Lexicography: Principles and Practice*, Academic Press, pp. 77-87.

Miller, G., Leacock, C., Tengi, R., & Bunker, T. (1993). A semantic concordance. In *Proceeding of ARPA Workshop on Human Language Technology*.

Ng, H. T. (1997). Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: What, why and how?*

Ng H. T., Lee H. B., (1996) Integrating multiple knowledge sources to disambiguate word sense : An Exemplar-Based approach. In *Proceedings of ACL-1996*, pp.40-47.

Ng, H. T., Lim, C. Y., & Foo, S. K. (1999). A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources*.

Niu, Z. Y., Ji, D. H., & Tan, C. L. (2004). Optimizing feature set for Chinese word sense disambiguation. In *Proceeding of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Palmer, M. (2000). Consistent criteria for sense distinctions. *Computers and the Humanities, 34*(1), pp.217-222.

Palmer, M., Babko-Malaya, O., & Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of Workshop on Scalable Natural Language Processing*.

Pedersen, T. (2002). A baseline methodology for word sense disambiguation. *Lecture Notes in Computer Science*, pp.126-135.

Pedersen, T., & Mihalcea, R. (2005). Advances in word sense disambiguation (Tutorial at ACL2005). In *Proceeding of ACL2005*.

Prince, V. (1997). An intelligent lexicon for contextual word sense discrimination. *Applied Intelligence, 7*(2), pp.125-146.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceeding of the ACL SIGLEX Workshop*.

Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceeding of the SIGLEX Workshop on "Tagging Text with Lexical Semantics: Why, What and How"*.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), pp.97-123.

Smadja, F. (1994). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*(1), pp.143-177.

Stevenson, M., & Wilks, Y. (2000). Large vocabulary word Sense disambiguation. In Ravin & Leacock (Eds.) *Polysemy: theoretical and computational approaches*: Oxford University Press.

Stevenson, M., & Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics, 27*(3), pp.321-349.

Tufi, D., Ion, R., & Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceeding of Coling2004*.

Veronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *the Programme and advanced papers of the Senseval workshop*.

Veronis, J. (2000). Sense tagging: Don't look for the meaning but for the use. *Computational Lexicography and Multimedia Dictionaries (COMLEX'2000)*, pp.1-9.

Veronis, J. (2001). Sense tagging: does it make sense. In *Proceeding of Corpus Linguistics Conference'2001*.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, 32(2), pp.73-89.

Weaver, W. (1955). Translation. In Locke W. N. and Booth A. D. (eds.), *Machine Translation of Languages: Fourteen Essay*, pp.143-172.

Wilks, Y. (1997). Senses and texts. *Computers and the Humanities, 31*(2), pp.77-90.

Wilks, Y. (2000). Is word sense disambiguation just one more NLP task? *Computers and the Humanities, 34*(1), pp.235-243.

Wilks, Y., Fass, D., Guo, C. M., McDonald, J. E., Plate, T., & Slator, B. M. (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceeding of the 12th conference on Computational Natural Language Processing*.

Wilks, Y., & Stevenson, M. (1996). The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?, *Technical Report CS-96-05*.

Wilks, Y., & Stevenson, M. (1997). Sense tagging: Semantic tagging with a lexicon. In *Proceeding of the SIGLEX Workshop Tagging Text with Lexical Resources*.

Wilks, Y., & Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering, 4*(02), pp.135-143.

Wilks, Y., & Stevenson, M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. In *Proceeding of COLING-ACL'98*.

Wu, Y. f., Jin, P., Zhang, Y. s., & Yu, S. w. (2006). A Chinese corpus with word sense annotation. In *Proceeding of Computer Processing of Oriental Languages*. *Beyond the Orient: The Research Challenges Ahead*, pp.414-421.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceeding of Coling-92*.

Yarowsky, D. (1993). One sense per collocation. In *Proceeding of the workshop on Human Language Technology*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceeding of the 33rd annual meeting on Association for Computational Linguistics*.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities, 34*(1), pp.179-186.

中文书目

专书：

符淮青 《词典学词汇学论文集》（北京：商务印书馆，2004）。

符淮青 《词义的分析和描写》（北京：外语教学与研究出版社，2006）。

郭锐 《现代汉语词类研究》（北京：商务印书馆，2002）。

李尔钢 《词义与词典释义》（上海：上海辞书出版社，2006）。

李行健主编 《现代汉语规范词典》（北京：外语教学与研究出版社、语文出版社，2004）

林杏光 《词汇语义和计算语言学》（北京：语文出版社, 1999）。

梅家驹 《同义词词林》（上海：上海辞书出版社，1983）。

梅家驹 《现代汉语搭配辞典》（上海：汉语大辞典出版社，1999）。

李如龙、苏新春编 《词汇学理论与实践》（北京：商务印书馆，2001）。

苏宝荣 《词义研究与辞书释义》（北京：商务印书馆，2000）。

王惠 《现代汉语名词词义组合分析》（北京：北京大学出版社，2004）。

张志毅、张庆云 《词汇语义学》（北京：商务印书馆，2001）。

中国社会科学院语言研究所词典编辑室编 《现代汉语词典（第5版）》（北京：商务印书馆，2005）。

论文：

冯志伟 <词义排歧方法研究>,《术语标准化与信息技术》，2004(1)，页31-37。

黄彬 <义项划分的依据与标准>,《辞书研究》，2004(5)，页31-36。

金澎、吴云芳、俞士汶 <词义标注语料库建设综述>,《中文信息学报》，2008，22(3)，页17-23。

李尔钢 <兼类词的义项设置和词性标注问题>,《辞书研究》，2006(3)，页14-24。

苏宝荣 <汉语语文辞书的词性标注及其对释义的影响>,《辞书研究》，2002(2)，页1-11。

孙宏林、黄昌宁 <词语搭配在文本中的分布特征>,《中文信息处理国际会议论文集 (Vol. 236)》（北京：清华大学出版社，1998）。

孙茂松、黄昌宁、方捷 <汉语搭配定量分析初探>,《中国语文》，1997(1)，页29-38。

王惠 <机器翻译中基于语法、语义知识库的汉语词义消歧研究>,《广西师范学学报(自然科学版)》，2003，21(1)，页86-93。

吴云芳、金澎、郭涛 <基于词典属性特征的粗粒度词义消歧>,《中文信息学报》，2007，21(2)，页3-8。

吴云芳、俞士汶 <信息处理用词语义项区分的原则和方法>,《语言文字应用》，2006(2)，页126-133。

张博 <现代汉语同形同音词与多义词的区分原则和方法>,《语言教学与研究》，2004(4)，页36-45。

张博 <影响同形同音词与多义词区分的深层原因>,《宁夏大学学报 (人文社会科学版)》，2005(1)，页5-11。

张博 <现代汉语复音词义项关系及多义词与同音形词的分野>,《语言研究》，2008(1)，页11-18。

周荐 <兼类词词性与多义词义项关系试说>,《辞书研究》，2007(3)，页37-46。