学校编码：10384　　分类号密级
学号：10220080150382UDC

博士学位论文

现代汉语名词歧义度研究

Research on Ambiguity grade of the Noun in Modern Chinese

李安

指 导 教 师 ： 苏新春　教授
学 科 名 称 ：

论文提交日期：2012　年　月
论文答辩日期：2012　年　月
学位授予日期：2012　年　月

答辩委员会主席：
评　　阅　　人：

2012 年 6 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（）课题（组）的研究成果，获得（）课题（组）经费或实验室的资助，在（）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（）1.经厦门大学保密委员会审查核定的保密学位论文，于年　月　日解密，解密后适用上述授权。

（）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　日

摘 要

歧义度是本文的核心论题，我们对其理论背景、定义、计算方法和影响因素做了系统论述。歧义度源于计算机词义消歧，词义消歧指用计算机自动给多义词标注一个确定的义项，是自然语言处理领域一个十分重要又十分困难的课题，也是当前研究的热点问题。我们注意到不同多义词的消歧结果差异巨大，有的多义词可以轻易得到很高的消歧正确率，有的词则难以实现。多义词这种体现在词义消歧难易程度的差异是由其内在语义属性决定的，这种内在属性就是歧义度。

在词汇语义学视角下，歧义度可以看做多义词的一个客观状况，它体现多义词义项间组合关系差异大小，在更深层次上体现了义项间语义关系亲疏远近的差异，多义词义项组合关系趋同则其歧义度高，反之就低，歧义度差异及其内在因素正可以从语义及其分布两个层面上解读多义词义项的关系。描写现代汉语名词的歧义度、分析造成词汇间歧义度差异的原因、探究歧义度研究的应用价值是这篇论文要解决的三个主要问题。

第一章主要介绍了本文选题的依据、研究对象、方法、步骤、特色、意义。

第二章主要介绍了词义消歧的进展与问题，着重论述了歧义度的提出、计算、在词汇语义学中的位置和基础理论。

第三章介绍了歧义度研究的课题设计，实现了 1352 个多义名词的词义消歧和歧义度计算，分析了多义名词区别性形式特征的内容和功能。统计发现歧义度最高为 100%，最低为 0，差异巨大，论文从义项间语义关系及其对应的区别性形式特征两个方面相结合的方法分析了这种现象的成因，这也是后面几章的主要内容。

第四章以统计数据为基础回答了语义距离与歧义度的关系。语义距离表示多义词义项间语义关系的亲疏远近，具体表现为多义词义项在语义分类词典内义类上的远近关系及概念语义相似性大小。统计发现，语义距离与歧义度之间存在密切负相关关系，根据歧义度与语义距离，将多义词义项间关系分为同义近义关系、同义类关系、跨义类关系三种类型，从宽计算同义近义关系可以并入同义类关系，同义类、跨义类是两种最重要的语义关系，本文分别建构了不同的分析框架，解

释其内部词歧义度差异的原因。

第五章提取了同义类词 [职业领域]、[构造]、[附属]等十几种重要的区别性义素及其与之对应的区别性形式特征，分析了义项间的区别性义素种类多少对歧义度的影响。

第六章针对跨义类词语义距离过大，不适于直接使用义素分析方法的特点，提取了语义相似、相关、无关三种语义关系模型，从义项历时联系、认知语言学角度解释了其歧义来源，分析了其内部不同词歧义度的差异和原因。

第七章以多义词语义距离和歧义度理论为基础提出了机用词典义项粒度设置的原则和方法，分析了其对歧义度的影响。“现汉”有义项粒度过细的特点，有 24.10%的词语义距离为 1，42.79%的词为同义类词，这些词义项间区别不明显，歧义度却很高，如去掉语义距离为 1 的词平均歧义度将由 46.20%变为 40.54%，去掉所有同义类词歧义度将变为 35.68%。

第八章提出了歧义度分析对词义消歧的启示，认为在研究中应该重视词义属性，应该在分类的基础上有所侧重地分别研究。

第九章对全文的主要结论做了简单总结，简述了研究的不足和后续研究计划。

本文在词汇语义学层面上回答了词义与形式特征的关系，尝试了词义统计研究方法，构建了将词义与其形式特征相互印证的方法；在词义消歧领域，通过对歧义度和多义词义项关系的研究，一定程度上解决了机用词典义项粒度问题，提出了分类逐步解决词义消歧课题的思路，提出了用歧义度解决词义消歧算法评测难的方法。在研究过程中坚持以实际应用推动理论建构的思路，不盲从已有的研究模式，在继承前人研究的基础上提出了歧义度的新概念并建构理论系统解决了相关问题。

关键词:歧义度 词义消歧 名词多义词 机用词典

Abstract

Ambiguity grade is the core topic of this article. In the research, we systematically discussed and analyzed its theoretical background, definition, calculation methods and influence factors. Ambiguity grade derives from computer word sense disambiguation which means automatic sense tagging to polysemous words by computer. We note that there are huge differences of disambiguation results among different polysemous words, while some can easily get high disambiguation accuracy rate, some of them cannot. These differences of disambiguation difficulty degree are decided by polysemous words' intrinsic semantic property, that is to say ambiguity grade.

Ambiguity grade can be seen as an objective situation of polysemous words under the perspective of lexical semantics; it reflects differences of word sense combination relations, and at a deeper level, embodies distance differences of word sense semantic relations. The more consistent combination relations become, the higher ambiguity grade is, and it's the same in reverse.From semantic and its distribution, ambiguity grade differences and its internal factors can interpret word sense relation of polysemous words well. To describe ambiguity grade of noun in modernChinese, to analyze causes of different ambiguity grade among words as well as explore application value of ambiguity grade study are three main issues to be discussed in the article.

Chapter one mainly discusses the basis of topic selection, research objects, research methods, research procedures, characteristics and significance of this article.

Chapter two reviewed the development and problems of disambiguation, and focuses on discussing the proposition and calculation of ambiguity grade, as well as its position and basic theory in lexical semantics.

Chapter three introduces the design of research on ambiguity grade. We conducted word sense disambiguation of 1352 polysemous words and described ambiguity grade of each word, meanwhile analyzed the content and function of polysemous words 'distinctive characteristics. Based on statistical data, there is great difference of word sense disambiguation grade. The highest disambiguation grade of difficulty is 100%, and the lowest is 0. The article analyzes this phenomenon from

two aspects: semantic relations between word senses and its corresponding forms of distinctive characteristics. These are also the main content of later chapters.

Chapter four mainly interprets the relation of semantic distance and ambiguity grade based on statistical data. Semantic distance refers to close or distance relation among word senses of polysemous words, concretely includes word sense distance relation of polysemous words in semantic classification dictionary and concept semantic similarity. The result shows a close negative correlation between semantic distance and ambiguity grade. According to ambiguity grade and semantic distance, word sense relation of polysemous words can be divided into three classifications: synonymous and near synonymous relation, same category relation and cross-type relation. Through lenient calculating, synonymous and near synonymous can be classified into the same category relation. Same category and cross-type relation are most important semantic relations. The article constructs different analytical frameworks separately in order to explain why differences exist in ambiguity grade of internal words.

Chapter five extracts a dozen important distinction sememes, such as [occupational area], [structure], [attachment] and corresponding forms of distinctive characteristics from same category words. The article analyzes the number of distinction sememes in word senses and its influence on ambiguity grade.

Chapter six analyzes the semantic distance of cross-type words are two far, which is not suitable for directly using sememe analysis methods. Against this, the article proposes three semantic relation models: semantic similarity, semantic relevant, and semantic non-linked in order to interpret ambiguous source in terms of diachronic relation of word senses and cognitive linguistics, as well as discuss the differences and reasons exist in ambiguity grade of different internal words.

Chapter seven introduces sense granularity's setting principles and methods of dictionary for computer and influence to ambiguity grade based on semantic distance of polysemous words and ambiguity grade theory. The sense granularity in 'Modern Chinese Dictionary' is too fine, 24.10% words' semantic distance is 1, 42.79% words are same category words. The distinction between word senses is not obvious, but ambiguity grade is quite high. For example, if we remove the words whose semantic distance is 1, ambiguity grade will be changed from 46.20% into 40.54%; if we remove all same category words, ambiguity grade will be changed into 35.68%.

Chapter eight discusses the revelation to disambiguation by analyzing ambiguity

grade. Word sense property should be taken seriously and discussed separately based on classification.

Chapter nine briefly summarizes the main conclusions of this article, proposes the limitation of current research and further research plan.

The article interprets relation between word sense and formal feature on the level of lexical semantics. By using word sense statistics, we constructed a method of mutual confirming word sense and its formal feature. In field of word sense disambiguation, through research on ambiguity grade and word sense relation of polysemous words, the issue about sense granularity of dictionary for computer could be solved to some extent. The thinking of classically and gradually solve word sense disambiguation, and the method of using ambiguity grade to solve word sense disambiguation arithmetical valuating difficulty are expounded in this article. In the process of research, theoretical framework was constructed based on practice. We didn't blindly follow existing study model, but succeed to propose a new concept of ambiguity grade, to construct theory framework as well as to systematically solve correlative issues.

Key words: ambiguity grade; word sense disambiguation; Noun; polysemous word; dictionary for computer

目 录

Table of Contents