# 基于蒙古语名词语义网的同形词歧义消除研究*

哈斯[1]　布音其其格[2]

（1.内蒙古师范大学，内蒙古 呼和浩特 010000）

**摘要：**蒙古文同形词歧义消除问题是蒙古文信息处理的难点之一。本文提出了基于蒙古语名词语义网的同形词歧义消除方法，设计实现了同形词歧义消除算法,最后给出了语料库中同形词歧义消除实验的设计过程及结果分析。

**关键词：**蒙古文；名词语义网；同形词；歧义消除

**中图分类号：**TP391　　**文献标识码：**A

# Research on the Homonyms Disambiguation Based on Mongolian Nouns Semantic Network

Hasi[1]　Buyinqiqige[2]

(1. Inner Mongolia Normal University, Huhhot ,Inner Mongolia, 010022,China ; 2. Huhhot Nationalities College, Huhhot , Inner Mongolia, 010051,China)

**Abstract:** Mongolian homographs disambiguation problem is one of the difficulties of the Mongolian information processing. This paper puts forward a method for eliminating homonyms ambiguity based on Mongolian nouns Semantic network, achieving the design and implementation of the homograph disambiguation algorithm. Finally, the design process and experimental results of the corpus of homograph disambiguation experiment are provided.

**Keywords:** Mongolian;Nouns Semantic network;Homonyms;Disambiguation

.

## 1　引言

自然语言中“歧义”是一个普遍存在的现象，又是比较难处理的问题。自然语言歧义问题事实上是词义和词形之间矛盾的问题。同一词形对应于两个或两个以上词义或结构时，就无可避免地产生了歧义，因此我们把歧义又称“同形歧义”。

蒙古文的歧义表现于它的语言单位，可分别为字体、音标、字节、字、构词助词、格

* **收稿日期：**2016-05-03　　**定稿日期：**2016-05-27

**基金项目：**国家自然科学基金项目《蒙古语词汇语义网研究》（批准号:61363053）;内蒙古自治区 2014 年度蒙古语言文字信息化专项扶持项目《蒙古文 MOOC 教学平台研发及基础资源建设》及内蒙古师范大学计算机与信息工程学院科技创新团队项目资助。

**作者简介：**作者一哈斯（1976——），男，教授，蒙古文信息处理；布音其其格（1974——），女，讲师，词汇学。

助词、词、连接法等等。例如：音标方面有（ A/E/N， T/D， D/T 等），字节方面有（ — ， — 等）歧义[1]。同形词歧义方面比如蒙古文中的《 》（0R0I）这个词有“山顶”的顶和“早晚” 的晚两种含义，只从词形上是无法确定其含义的。同形词的歧义消除问题已经涉足词汇语法属性和语义属性领域了。30 多年的蒙古文信息处理实践说明，无论是基础研究，还是应用开发，都离不开对词汇进行语法属性和语义属性的描述。没有一个基于词汇的语法、语义属性描述体系，就无法满足深层次的语言信息处理需求。

作为蒙古语语义属性描述体系的一个重要组成部分，《蒙古文同形词词典》的建立与应用是本研究的一部分。在蒙古语的语义研究中，通过语料库进行研究已经成为主要的手段。基于语料库的同形词研究不仅要统计同形词的词形出现频率，更重要的是同一词形分别以不同词义出现的频率。这样才能准确统计同形词按不同词义出现的概率，为搭配词库的应用，机器翻译等提供概率统计方面的帮助[4]。

## 2 蒙古文同形词信息词典

同形词歧义研究工作中为了更清楚地表示蒙古文同形词的不同词义形式，内蒙古大学淑琴博士研究设计了蒙古文同形词信息词典，其中包括“蒙古文词形”（MONGGOL） 、“拉丁转写”（GALIG)、“词类”（UGSAIMAG）、“分类标记”（ILGAHV）、 “汉语词义”（HITAD）等字段。

以蒙古文词形 举例为如下：

| 蒙古文词形 MONGGOL | 拉丁转写 GALIG | 词类 UGSAIMAG | 分类标记 ILGAHV | 汉语词义 HITAD |
|---|---|---|---|---|
| | 0I | Ne2 | | 记性 |
| | 0I | Is | | 喂(招呼声) |
| | 0I | Ve2 | A | 跳起 |
| | 0I | Ve2 | B | 坠落 |
| | 0I | Ne2 | | 林 |
| | 0I | Ne1 | | 生日 |
| | 0I | Ne2 | | 嫌恶感 |

从中可以看出蒙古文的 (0I) 这个词形有名词、动词等不同词类形式，具有兼类词性质，同时同一个词类情况下有不同词义现象。这就说明词形和词类确定不了其词义，还得进一步区分同一种词类时究竟是哪个词义来使用。为此同形词词典中添加了分类标记（ILGAHV）字段来区分以上情况。如 (0I)这个词的词类为动词 Ve2 情况下如果其分类标记（ILGAHV）为 A 则表示跳起，B 则表示坠落。

蒙古文同形词歧义问题如果把同形词的词形、词类的基础上能够准确标注其分类标记（ILGAHV），则歧义自然就能够消除了。

基于语料库的同形词研究中首先要求同形词歧义消除问题，即上述分类标记（ILGAHV）的正确标注是关键问题。对于大规模的语料库当然需要一个能够自动标注分类标记（ILGAHV）的功能。

## 3 基于蒙古语名词语义网的歧义消除方法

蒙古语名词语义网是一种语言知识库，其建立的目的就是为自然语言理解与处理服务。蒙古语名词语义网提供词汇语义查询功能以外还有一个特点是解决词汇歧义问题。通过词汇概念的形式化描述和概念之间语义关系的简明结构使得蒙古语名词语义网成为词义消歧的主要词典资源。特别是名词继承体系，其严格的继承关系使得在语言分析过程中，句子中的词汇歧义，只要通过内部结构自身就可以利用概念密度的计算，加以消除。

例如蒙古文的句子：

（食堂没地方，我在饭馆吃了饭 。）

句子中“ ”（地方）是有歧义的词。不经过分析，单单依靠词汇语义网，就可以对“ ”（地方）进行消歧。

“ ”（地方）有以下六种不同的意思：

（地球）

（庄稼）

（地方）

（长度单位）

（地板）

（指原因的副词）

以上含义分列在不同的语义树上，其中一个树的部分如图 1 所示。

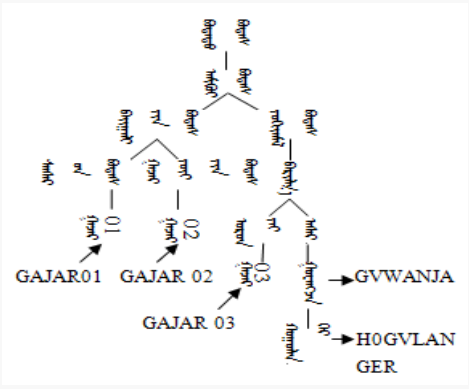



图 1　词汇“ ”（地方）的语义树

只要计算一下“ ”（食堂）和“ ”（地方）之间的距离，直观上就可以知道，“ 03”（GAJAR03）是离“ ”（HOGOLAN GER）最近的，应选“地方”的“ 03”（GAJAR03）词义。

## 4　语料库中同形词歧义消除实验

### 4.1 歧义消除算法

蒙古文的同形词歧义消除过程中可利用上下位语义树。如蒙古文的 （CIHI）这个

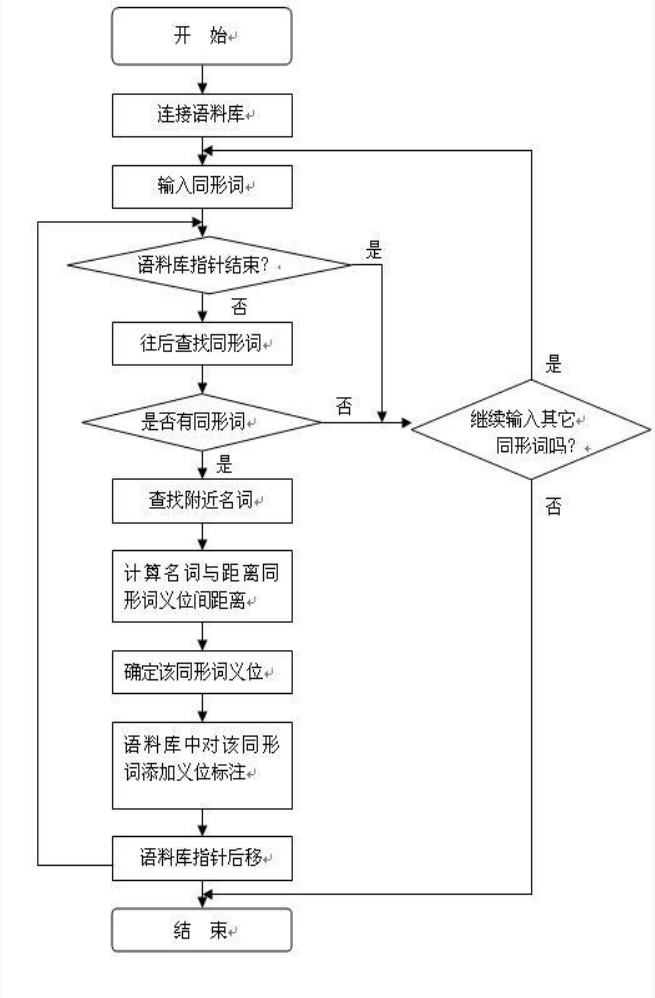
词形对应到三个名词，一是耳朵，二是树苗的苗子，三是器皿两侧的耳子。在同形词词典中以上三个词条分别对应到分类标记 A、B、C。因此只要语料库中遇到的 （CIHI）能够自动识别并准确进行 A、B、C 就可以完成同形词的歧义消除工作。

为了统计自动识别算法的结果，首先在 26 万词条（已完成语法信息标注）的语料库中人工方式对每一个 （CIHI）进行了分类标记的标注。接下来调用自动标注算法进行了一遍标注工作。最后对以上两次标注工作结果进行对比，进而评价自动标注算法的性能。

整个歧义消除算法（或者现在已经对应到分类标记自动标注算法）是在以句子为单位的语义环境中的进行歧义判断的。通过同形词与同句名词之间的语义关系计算最终判断歧义。因此所考虑的名词必将是跟同形词共处一个句子，并且要句法关系相近的词汇才行。本文中查找名词时考虑到了与同形词间的距离，选择计算的是与同形词距离最近的名词。

语料库中进行以上同形词的歧义消除过程的算法为如图 2 所示：

图 2　基于词汇语义网的语料库中同形词歧义消除算法流程图




4.2 实验设计

同形词词汇歧义消除实验的设计方案跟前面的多义词歧义消除实验大体相同。本文中直接利用了基于搭配词库的同形词歧义消除实验中的初始数据。

歧义消除试验中共选择了 8 个词形，20 个名词。下面是在 26 万词条（已完成语法信息标注）的语料库中利用语义网进行词义消除情况。

8 个词形，20 个名词的同形词表的信息为如下表 1：

表 1　同形词词词汇歧义消除实验单词信息表

| ID | MONGGOL | GALIG | UGSAIMAG | ILGAHV | HITAD |
|---|---|---|---|---|---|
| 195 |  | OI | Ne2 | A | 记性 |
| 199 |  | OI | Ne2 | B | 林 |
| 200 |  | OI | Ne1 | C | 生日 |
| 201 |  | OI | Ne2 | D | 嫌恶感 |
| 268 |  | ORO | Ne1 | A | 床 |
| 269 |  | ORO | Ne1 | B | 位 |
| 270 |  | ORO | Ne2 | C | 迹 |
| 488 |  | AGVR | Ne2 | A | 蒸汽 |
| 489 |  | AGVR | Ne2 | B | 生气 |
| 616 |  | ANGGI | Ne1 | A | 班级 |
| 617 |  | ANGGI | Ne1 | B | 阶级 |
| 1455 |  | CIHI | Ne1 | A | 耳 |
| 1456 |  | CIHI | Ne1 | B | （器皿两侧的）耳子 |
| 1457 |  | CIHI | Ne1 | C | 秧儿 |
| 2986 |  | HELE | Ne1 | A | 语言 |
| 2987 |  | HELE | Ne1 | B | 舌 |
| 2669 |  | HOTA | Ne1 | A | 浩特* |
| 2670 |  | HOTA | Ne1 | B | 城 |
| 3156 |  | HOLOSO | Ne2 | A | 工钱 |
| 3157 |  | HOLOSO | Ne2 | B | 汗 |

### 4.3 实验过程

（1）第一步：将同形词按每个义位添加到蒙古文上下位关系语义树中，即给每个义位进行 SynsetID 标注，并指出其上位 SynsetID。 例如蒙古文的 （0I）在 《同形词典》里有 4 个义项，对应到 4 个义位。因此将其看成是 4 个单词来添加到语义树里。

SynsetI 标注结果为如图 3 所示：

| ID | MONGGOL | GALIG | UGSAIMAG | ILGAHV | HITAD | synset_id | hypernym | hypertree |
|---|---|---|---|---|---|---|---|---|
| 195 |  | OI | Ne2 | A | 记性 | 105326073 | 105325039 | ->105325039->105295659->100020729 |
| 199 |  | OI | Ne2 | B | 林 | 107926765 | 107925401 | ->107925401->107470940->100026769 |
| 200 |  | OI | Ne1 | C | 生日 | 114388390 | 114299333 | ->114299333->114299149->114257468 |
| 201 |  | OI | Ne2 | D | 嫌恶感 | 107043607 | 107042000 | ->107042000->100021668->100020333 |
| 268 |  | ORO | Ne1 | A | 床 | 102719813 | 103281101 | ->103281101->103280711->103443493 |
| 269 |  | ORO | Ne1 | B | 位 | 100556725 | 100553013 | ->100553013->100389883->100026194 |
| 270 |  | ORO | Ne2 | C | 迹 | 104293634 | 104290583 | ->104290583->103941718->104387207 |
| 488 |  | AGVR | Ne2 | A | 蒸汽 | 110765993 | 110684168 | ->110684168->110701049->110681095 |
| 489 |  | AGVR | Ne2 | B | 生气 | 100714423 | 100713313 | ->100713313->100712691->100701215 |
| 616 |  | ANGGI | Ne1 | A | 班级 | 100831838 | 100831015 | ->100831015->100389883->100026194 |
| 617 |  | ANGGI | Ne1 | B | 阶级 | 107492073 | 107463651 | ->107463651->100026769 |
| 1455 |  | CIHI | Ne1 | A | 耳 | 105014060 | 104994211 | ->104994211->104992592->104919813 |
| 1456 |  | CIHI | Ne1 | B | （器皿两面三刀侧的） | 103347214 | 103357984 | ->103357984->102634643->103746082 |
| 1457 |  | CIHI | Ne1 | C | 秧儿 | 110807937 | 110807729 | ->110807729->112333068->100014510 |
| 2669 |  | HOTA | Ne1 | A | 浩特* | 107687598 | 107685760 | ->107685760->107493031->107470450 |
| 2670 |  | HOTA | Ne1 | B | 城 | 108019865 | 107975544 | ->107975544->108030730->108104509 |
| 2986 |  | HELE | Ne1 | A | 语言 | 106479855 | 105899749 | ->105899749->100028764->100028549 |
| 2987 |  | HELE | Ne1 | B | 舌 | 104996105 | 104992592 | ->104992592->104919813->108797461 |
| 3156 |  | HOLOS( | Ne2 | A | 工钱 | 112527854 | 112522505 | ->112522505->112520120->112519607 |
| 3157 |  | HOLOS( | Ne2 | B | 汗 | 105096534 | 105095511 | ->105095511->105089633->104960498 |

图 3　同形词 SynsetID 标注结果

(2)第二步：语料库中查找所有上述同形词，先进行人工标注（标注其义位编号），然后调用上述算法进行自动标注方式完成歧义消除，即确定句子中的同形词究竟是对应到多个义位中的哪一个。进行歧义标注的语料库为如图 4 所示：

图 4 语料库中同形词歧义标注结果

| ID | GALIG | ILGAHV | UGSAIMAG | MILGAHV | mingci |
|---|---|---|---|---|---|
| 3974 | 0R0 | B | 0R0/Ne2 | 2 | SVRVGCI |
| 4443 | H0TA-DV | A | H0TA/Ne1-DV/Fc21 | 2 | AJIL |
| 6181 | HELE-BER | A | HELE/Ne1-BER/Fc51 | 1 | HOMON |
| 6287 | H0TA-DV | A | H0TA/Ne1-DV/Fc21 | 2 | H0DEGE |
| 2174 | CIHI | A | CIHI/Ne1 | 2 | HVLVSV |
| 2647 | CIHI | A | CIHI/Ne1 | 1 | CIRAI |
| 2986 | ANGGI | B | ANGGI/Ne1 | 1 | KOLONI |
| 3002 | ANGGI | B | ANGGI/Ne1 | 2 | ORON |
| 3006 | ANGGI-YIN | B | ANGGI/Ne1-YIN/Fc11 | 1 | CILOGELELTE |
| 3045 | ANGGI-TAI | B | ANGGI/Ne1-TAI/Fc61 | 2 | VLVS |
| 3064 | ANGGI-TAI | B | ANGGI/Ne1-TAI/Fc61 | 2 | VLVS |
| 3602 | 0R0 | B | 0R0/Ne1 | 2 | SAGVDAL |
| 3974 | 0R0 | B | 0R0/Ne2 | 2 | SVRVGCI |
| 4443 | H0TA-DV | A | H0TA/Ne1-DV/Fc21 | 2 | AJIL |
| 6181 | HELE-BER | A | HELE/Ne1-BER/Fc51 | 1 | HOMON |
| 6287 | H0TA-DV | A | H0TA/Ne1-DV/Fc21 | 2 | H0DEGE |

### 4.4 实验总结

（1）第一步：对所得结果进行统计。

通过程序运行最终共对 1013 个单词进行了歧义标注，结果为表 2 所示：

表 2　语料库中同形词歧义消除结果

| 序号 | GALIG | 次数 | 正确次数 | 错误次数 | 准确率 |
|---|---|---|---|---|---|
| 1 | 0I | 89 | 79 | 10 | 88.80% |
| 2 | AGVR | 64 | 37 | 27 | 57.80% |
| 3 | ANGGI | 332 | 187 | 135 | 56.30% |
| 4 | CIHI | 91 | 30 | 61 | 33% |
| 5 | HELE | 166 | 92 | 74 | 55.40% |
| 6 | H0TA | 190 | 106 | 84 | 55.80% |
| 7 | HOLOSO | 43 | 10 | 33 | 23.30% |
| 8 | 0R0 | 38 | 17 | 21 | 44.7% |
|  | 合计 | 1013 | 558 | 445 | 55.1% |

（2）第二步：对统计结果进行分析

分析结果后发现，错误标注的主要原因有以下几方面：

① 语义网中名词的同义词集合 ID 标注有不准确的情况 如果同义词集合 ID 标注合理准确将会提高准确率；

② 自动标注算法运行过程中所找到的名词跟当前词（同形词）不在同一语义块中，导致无法计算距离；

③ 第一个同形词 0I 的标注结果准确率相对较好的原因是语义计算的名词大部分都是该词常用搭配词，进而提高了歧义消除效率。

④ 同形词 （CIHI）的标注结果准确率较差的原因是总共 91 次标注当中 1）出现了 74 次， 3）出现了 17 词，而 2）没有出现。但是在自动标注结果中 1）标注了 19 次， 2）57 次， 3）15 次，将多数 1）误认为 2）。经过分析后发现主要是这些标注错误的句子中没有出现与 1）搭配关系很近或语义关系密切的词出现，进行语义计算的词汇大部分都是其它的一些词汇。如 1）与 （眼睛）、 （鼻子）， （脸）等词汇共同出现则 都能够非常准确的标注。同样 3）也是，与 （庄稼）、 （树）、 （杨树）等词汇计算距离时都能准确判断其义位。

因此依靠语义网进行语义计算，完成歧义消除时找到语义树上距离相近的词汇很关键。随着蒙古文句法处理技术的深入，结合短语标注等技术可以较准确地判断与同形词计算距离的名词。这样不仅提高准确率，还可以降低算法中查找名词的时间复杂度。

## 5 结论

词汇语义网络是词汇语义计算的非常重要的工具。目前基于 WordNet 等各类语种词汇语义网络的应用比比皆是。蒙古语名词语义网的研究课题目前刚刚起步，本研究初步尝试了基于蒙古语名词语义网的同形词歧义消除工作。下一步我们将进一步优化词汇语义网的框架结构,完善词汇语义数据库的信息,补充动词和形容词等其它词类信息的同时要提高语义网的应用性能。

## 参考文献

学报：[2011 年第 2 期] 哈斯. 基于搭配词库的蒙古文同形词歧义消除[J]. 内蒙古师范大学学报（自然科学版）. 2011 年第 2 期.

学报：[2009 年第 1 期] 哈斯、淑琴.同形同音词词典中分类标志的自动标注法[J].中国蒙古学.

论文：Beckwith R，Miller G A ，Tengi R. Design and Implementation of the WordNet Lexical Databaseand Searching Software[J] . Specification of WordNet. 1993 .

论文：Fellbaum C. WordNet: an Electronic Lexical Database [M]. MIT Press.1999 .

论文：Miller G A. An on line lexical database[J].International Journal of Lexicography .1990.3（4）:235 - 244.

论文：Kamps J. Visualizing WordNet Structure[C] .ICGW 2002. India. 2002 .

论文：Hasi、Nasun-urt . The Automatic Construction Method of Mongolian WordNet Noun Sets of Synonyms[C].The 4th International Conference on Intelligent Networks and Intelligent Systems. Kunming. China. 2011.

论文：Hasi、Nasun-urt. The Automatic Construction Method of Mongolian Lexical Semantic Network Based

on WordNet[C] . The 5th International Conference on Intelligent Networks and IntelligentSyst. Tianjin. China. 2012.
学报：［2008 年第 6 期］朱虹，刘扬.词汇语义知识库的研究现状与发展趋势[J].情报学报.
中文信息学报：［2010 年第 2 期］黄居仁，谢舒凯，洪嘉馡等.中文词汇网络_跨语言知识处理基础架构的设计理念与实践[J].中文信息学报.
论文：那顺乌日图.蒙古文信息处理概述[C].The second China-Japan Natural Joint Processing Research Promotion Conference， Peking .2002 年 10 月.
内蒙古大学学报：［2002 年第 5 期］那顺乌日图.关于面向信息处理的蒙古文义研究[J].内蒙古大学学报.
学术会议：德·萨日娜、那顺乌日图.蒙古文语义信息词典的初步构建[C]. 第十届全国计算语言学学术会议.中国计算机语言学研究前沿进展（2007-2009）. 2009 年.
中文信息学报：[2011 年第 6 期] 那顺乌日图.蒙古文语言知识库的建立与应用[J].中文信息学报.
中文信息学报：[2009 第 7 期] 王石，曹存根.一种 WordNet 概念自动翻译方法[J].中文信息学报.
中文信息学报：[2011 年第 4 期] 赵小兵，邱莉榕，赵铁军.多民族语言本体知识库构建技术[J].中文信息学报.
学位论文：[2012 年 6 月] 李慧.蒙古文语义知识词典的研究与实现[D].内蒙古大学硕士学位论文.

作者联系方式：
姓名：哈斯
地址：内蒙古师范大学计算机与信息工程
学院
邮编：010000
电话：13848113707
电子邮箱：hasi@lmani.edu.cn

作者照片：

哈斯

布音其其格