蒙古文 WordNet 名词同义词集合构建算法*

哈斯 那顺乌日图

内蒙古大学 蒙古学学院，呼和浩特 010021

内蒙古师范大学 计算机与信息工程学院，呼和浩特 010022

E-mail: hasi@imnu.edu.cn

摘 要： 蒙古文名词同义词集合的自动建立是研发“蒙古文 WordNet”名词子网时首要完成的基础工作。本文提出了一种从中英文 WordNet 转换生成蒙古文 WordNet 名词同义词集合的方法，并设计实现了蒙古文 WordNet 名词同义词集合的生成维护系统。论述了蒙古文 WordNet 名词同义词集合的构造扩充原则和词义消岐等应用前景。

关键词： WordNet；蒙古文 WordNet；名词；同义词集合

The Construction Algorithm of Mongolian WordNet Noun Sets of Synonyms

Hasi Nasun-urt

Academy of Mongolian Studies, Inner Mongolia University, Huhhot 010021

Computer and Information Engineering College, Inner Mongolia Normal University, Huhhot 010022

E-mail: hasi@imnu.edu.cn

Abstract: Automatic construction of Mongolian noun sets of synonyms is the fundamental work we have to accomplish first when we develop the noun subnet of Mongolian WordNet. This article proposed an approach of transforming Chinese or English WordNet to Mongolian WordNet noun sets of synonyms, and also designed and implemented the maintenance system of the Mongolian WordNet noun sets of synonyms. Some application prospects such as extension principles for noun sets of synonyms and disambiguation algorithm are discussed.

Keywords: WordNet; Mongolian WordNet; noun; sets of synonyms

1 前言

近年来，学者们在面向信息处理的蒙古语语义研究方面做过不少的探索，这些探索涉及到名词的语义分类、动词的语义分类和形容词的语义分类、格框架、配价理论等方面，这些研究都取得了一定的成果。但是，机器翻译系统及其他自然语言处理各种系统，通常需要一部包括语义知识的电子词典为计算机自动分析提供更全面、深入的语义信息。虽说《蒙古语语法信息词典》中涵盖语义信息，但它完全不是语义词典，因此目前迫切需要构建蒙古语语义信息词典。另外，当前面向信息处理的语义研究对丁蒙古语动词、名词、形容词的内部语义关系，动词与名词、名词与名词、名词与形容词之间的语义关系不够深入。因此，面向信息处理蒙古语语义研究的核心内容是运用语义网、概念依存理论来分析词语与词语之间的语义关系，建立语义关系网络。

语义关系网络方面最具有代表性的是英文 WordNet。该词网是普林斯顿大学认知科学实验室开发的一部在线词典数据库系统，是基于英文的词汇语义网络系统 • WordNet 将英文的名词、动词、形容词和副词组织为同义词集合（synset s），每一个集合表示一个基本的词汇概念，并在这些词汇概念间建立了包括同义关系、反义关系、上位关系、下位关系、部分关系以及完全关系等多种词汇语义关系。目前，WordNet 被成功地用于词义消歧、语言学自动处理、双语及多国语机器翻译、检索系统等一系列语言工程，被普遍认为是用于计算语言学、文本分析和许多相关领域的最重要的资。国内也有基于英文 WordNet 的中文 WordNet 研究实现先例。

* 本文承国家自然科学基金项目“《蒙古语语义信息词典》的设计与实现”（项目号：60873084）和内蒙古大学创新人才培养项目的资助。

我们应充分吸收现有的研究成果，在语法知识库的基础上构建语义知识库是弥补蒙古语语义研究不足的一个有效途径。蒙古文信息处理用的词汇语义研究的任务是从计算机处理的需求出发，全面而深入地揭示词汇的各种语义关系，使之便于对意义进行计算。因为关系是词汇语义的灵魂。这里所指的语义关系是概念与概念之间的关系，概念的属性和概念属性之间的关系。运用计算语言学和计算语义学的理论和方法，自动构建一个面向信息处理的语义网络，提供强有力的、适用的语义资源是解决蒙古文信息处理所面临的燃眉之急。名词作为蒙古语三大“开放性”词类之一，对它进行语义关系网络建立是构建蒙古语 WordNet 的基础性探索。为了给计算机自动分析和自动生成提供更全面、更深入的语义知识，建立一项与语言工程应用相结合的、面向语义知识库的名词语义网络是“蒙古文 WordNet”的一项基础研究。其中名词同义词集合的建立是该课题的重要组成部分，是我们首要完成的分支工作。

2 蒙古文名词同义词集合的建造

2.1 蒙古文名词同义词集合的实现方法

以《蒙古语语法信息词典》中的名词概念为基础，利用 WordNet 概念间关系以半自动方式创建一个适用于蒙古文信息处理的蒙古文名词同义词集合。首先从《蒙古语语法信息词典》中抽取名词概念，从 WordNet 中查找对应的语义框架及概念间关系，并将其移植为蒙古文 WordNet 框架，然后在这个框架中植入蒙古语词汇，逐步转换生成蒙古文 WordNet 名词子网。《蒙古语语法信息词典》中有些名词概念是 WordNet 中不存在的，这样的词汇概念可以手工方式将其添加到语义网中，这样的结点往往是处于较低层次上的叶子结点，。这就可以保证在较高的概念层次上与 WordNet 兼容，在较低的概念层次上具有最大限度的灵活性。

2.2 蒙古文名词同义词集合的建造过程

WordNet 的构架是以同义词集合（synset）为组织内容的，因此建立蒙古文同义词集合，以 synset 为单位构造语义网是跟其他语言的 WordNet 兼容的必要前提。为了充分利用前辈们的研究成果，研究中以《蒙古语语法信息词典》的名词词汇为基础，对其进行 synset ID 标注。其主要思路是：

（1）将依照《蒙古文同义词词典》，建立蒙古文同义词表，表中每条记录为一组同义词，共有 2 千多条记录。

（2）对《蒙古语语法信息词典》的每一个名词进行判断是否有同义词，即在《蒙古文同义词词典》中找到该词则将其标注为有同义词。

（3）从《蒙古语语法信息词典》中选一个未进行 synset ID 标注的名词，并读取其中英文对照词。

（4）以中英文对照词到中英文 WordNet 中找对应的同义词集合，找到则读取该 synset ID，并给对应蒙古文名词标注此 synset ID。英文 WordNet 的 synsetID 取值范围是在 100000000 至 400000000 内，中文 WordNet 的 synsetID 取值范围是在 600000000 至 900000000 内。中文同义词集合 ID 是表示相同概念的英文同义词集合 ID 的第一位加了 5 之后生成的。比如英文词汇 thing 的 synsetID 为 100002056，而中文词汇“物”的 synsetID 为 600002056，故蒙古文词汇“BVDAS”（物）的 synsetID 设置为 100002056。

（5）若《蒙古语语法信息词典》中还有未进行 synset ID 标注的名词，则跳到(3)继续标注。

考虑到《蒙古文语法信息词典》中蒙古文名词对应的中英文词汇标注不够完整或有些不准确情况，对这些词可利用《达尔罕词典》等其他工具查找其对应的汉语词汇，再在中文 WordNet 词汇表中找对应的词汇。通过以上两种方法完成了 8 千多个词汇的 synset ID 标注工作，剩余的词汇如“SIYANBeI”（鲜卑）等词汇中文 WordNet 词汇表中没有，则可以手工方式设置其 synsetID。为

了不跟中英文同义词集合冲突，将这些需要手工设置的 synsetID 从 m00000001 开始自动编号，m 表示蒙古语独有词汇概念，如“BISILIG”(奶食品的一种)是中英文 WordNet 词汇库中不存在的。

3 蒙古文名词同义词集合的进一步完善

同义词集合（sets of synonyms）确实是 WordNet 词库的基石，也是 WordNet 构成一个义类词典的根本所在。所以名词子网的第一项工作是对《蒙古文语法信息词典》中的所有蒙古文名词设置其同义词集合 synsetID。并对同义词集合（synset）添加说明性的注释。这跟传统的词典情况类似。不过一个 synset 不等于词典中的一个词条。尤其是词典中的一个词条可能是个多义词(polysemous word)，它就会包含多个解释，而一个 synset 只包含一个注释。

关于词库词汇学习方式中最著名的也是最重要的心理语言学事实之一是，人们对一些词语比另一些词语更熟悉。对一个词语的熟悉度在许多方面会有所表现：阅读速度，理解速度，易于回忆，使用概率，等等。这些方面的影响如此普遍地存在，以至于那些希望研究词语其他性质的实验者，即便付出极大的努力，也很难将不同词语的熟悉度程度视作一样。换言之，词库的初衷是反映心理语言学原则，如果在词库中忽略词语的熟悉度在上述表现上的差异，将是不可想象的。

为将词语熟悉度的差异反映到 WordNet 中，我们给每个词形式（Word Form）添加了一个熟悉度的标记指数。使用频率通常被认为是熟悉度的最好体现。那些扮演着重要的句法角色的封闭类词语是使用频率极高的词语，不过，甚至在开放类词语中，使用频率上也存在着较大差异。使用频率通常被假定为跟熟悉度的差异相关，或者干脆就用前者来解释后者。词频数据在一些技术文献中可以查到，但是，对于 WordNet 这样规模的词库来说，原有的词频数据还是不够的。Thorndike 和 Lorge(1994)出版了基于 500 万词文本语料库的统计结果的词频表，不过他们只报道了 3 万常用词的结果。此外，他们对词的定义是两个空格间的字符串，因此他们对同形异义字（homograph）的统计是不可靠的，比如他们的结果无法说明 lead 这个词作为名词和作为动词出现的频率有什么差别。Francis 和 Kucvera(1982)用他们自己的句法类标记来标明词语的词性，不过他们报告的结果仅仅是从包含 1,014,000 个单词的文本中得到的结果(含有 50400 个词形，其中包括许多专有名词)。因此这个结果对反映非常用词的频度是不够的。(通常的语速为 120 词/分钟，因此 100 万词大约相当于 140 个小时的话语，或者一个人两周所说的话）。

WordNet 专家们用另外的办法来表示熟悉度。Zipf(1945)的研究表明，词语出现的频率跟多义性是相关的。平均来说，频度越高的词语，在词典中也就有越多的不同意义。心理语言学一项令人感到有趣的发现（Jastrezembski，1981）是，多义性似乎预示了人们访问大脑词库的时间，就好像一个词的频度所能起到的作用那样。因此，WordNet 不用词语的出现频度来指示熟悉度，而用多义性来反映熟悉度。词语义项数可以从一部在线词典中得到。如果那些不在这部词典中出现的词语被指派熟悉度指数值为 0，对于词典中收录的词语，则根据词语的义项数来指派熟悉度指数(比如 1、2、3、……)，那么，这样的数值就可以为各个词类中的每一个词指派一个。因此，对于 WordNet 中的每个词形式，都用一个整数值来记录该词形式（作为名词、动词、形容词、副词使用时）的义项数。

我们在研究中利用《蒙古文多义词信息词典》对蒙古语名词多义词进行熟悉度的标记。除此之外多义词可按照其基本词汇意义被列入相应的同义词集合之外，还可以按照其他义项添加到别的同义词集合中，这样能使蒙古文名词同义词集合信息更加丰富。

4 名词同义词集合的应用

一个名词通常只有一个直接上位词，因而编词典的人用这个上位词来定义该名词；一个名词通常不只一个下位词，因而编词典的人一般很少罗列这些下位词。WordNet 中名词的组织方式正

好利用了名词的这种上下位关系。所以上下位关系是 WordNet 名词子网中的主要组织依据，有着很重要的应用价值。如一些动词的搭配选择限制也表明名词上下位关系的重要性。比如动词“喝”的直接宾语可以是“饮料”的任何一个下位词。这也暗示有关名词的上下位关系的知识应该以一种人们能够快速访问和搜索到的方式存储。

构造蒙古文 synset 集合后，以自动转换方式从中文 WordNet 把词汇上下位关系，反义关系，整体部分关系等语义关系可以添加到蒙古文名词 synset 集合中，进而较方便地构造蒙古语名词语义网的主框架。尽管不同民族语言之间存在着差异，但从概念角度来说，人们对世界的认识还是相通的、相近的、甚至是相同的。因此利用中文 WordNet 的语义关系来构造蒙古语名词语义网的框架是较好的途径。

WordNet 名词子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。本研究中建立 WordNet 名词子网的查询应用功能以外主要的一个特点是想解决名词歧义问题。通过名词概念的形式化描述和概念之间语义关系的简明结构使得名词语义网成为词义消歧(Word Sense Disambiguation，WSD)的主要词典资源，并可能在其他语义分析中得到应用，如机器翻译等项目中进一步利用。

例如蒙古文的句子：

H0GVLAN GER-TU GAJAR UGEI B0LHVR BI GVWANJAN-DU BVDAG_A IDEBE.

（食堂没地方，我在饭馆吃了饭 。）

句子中“GAJAR”（地方）是有歧义的词。有以下六种不同的意思：

1) HOMON TOROLHITEN-U AMIDVRA/N 0R0SI/JV BAYI/G_A DELEHEI-YIN BOMBORCEG（地球）

2) HOROSO，SIR0I，TARIYA/N GAJAR（庄家）

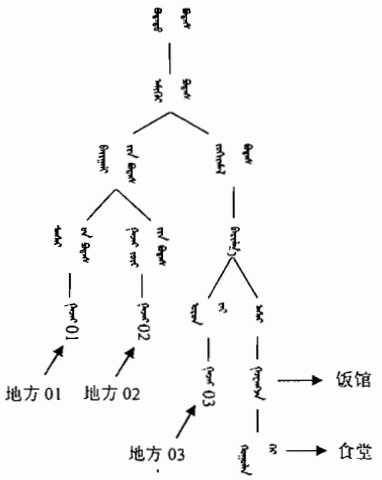
3) 0R0N NVTVG，0R0N BAYIRI（地方）

4) VRTV-YIN NIGECI.NIGE GAJAR NI JAGV TABI/N H00S ALDA(ARBA/N TABV/N YIN)BVYV HAGAS KIL0MetR-TEI TENGCE/N_E.BAYIGVLG_A BVYV TEGUN-U D0T0RAHI NIGE NIGECI（长度单位）

5) $ALA HUSER（地板）

6) SILTAGAN-V DAYIBVRI UGE（指原因的副词）

以上含义分列在不同的名词语义树上，只要计算一下“H0GVLAN GER”（食堂）和“GAJAR”（地方）之间的距离，直观上就可以知道，“GAJAR”3）是最近的，应选“GAJAR”3)”中词：“地方”的词义。如图所示:



5 结束语

目前蒙古文名词 synset 集合的构造基本完成，有些特殊词汇需要手工方式进行 synset ID 的设置。已确定 synset ID 的词汇以 synset 为组织单位开始进行语义关系的标注工作。蒙古语名词子网的研究是蒙古文 WordNet 的一个初步探索，蒙古文 WordNet 的构建是长期的、动态的、工程量极大的项目，同时涉及到多个学科，多方面的技术。相对英文 WordNet 或中文 WordNet，蒙古文 WordNet 的研究工作刚刚起步。由于时间关系和研究尚未完全成熟，目前的工作中还有很多有待完善的地方。我们在下一步工作，即名词同义词集合语义关系研究中将同义词集合信息不断完善。

参 考 文 献

[1] 德力格尔玛.《蒙古语语义研究》. 辽宁民族出版社, 2001.

[2] 那顺乌日图. “关于面向信息处理的蒙古语义研究”. 内蒙古大学学报, 2002 年第 5 期.

[3] 那顺乌日图. “蒙古文信息处理概述”. the second China-Japan Natural Joint Processing Research Promotion Conference, Peking, 2002. 10.

[4] 王惠、詹伟东、俞士汶. “现代汉语语义词典” 的结构及应用.语言文字应用, 2006 年第 2 期.

[5] 额尔顿朝鲁. “面向信息处理的蒙古语动词研究”. 内蒙古大学博士学位论文, 2005 年.

[6] 海银花、那顺乌日图. “面向蒙古语语义信息词典的名词语义分类体系”. 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.

[7] 德萨日娜. “蒙古语语义词典的数据库建设”. (收入《中文计算技术与语言问题研究—第七届中文信息处理国际会议论文集》. 电子工业出版社, 2007 年 9 月, 合著).

[8] 德萨日娜. “关于蒙古语语义分析的思考”. 内蒙古社会科学（汉文版），2004 年 3 期.

[9] 德萨日娜、那顺乌日图. “蒙古语语义信息词典的初步构建”. 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.

[10] 巴达玛敖德斯尔. “面向信息处理的蒙古语词语分类体系研究”. 中央民族大学学报(哲学社会科学版), 2004 年第 3 期.

[11] 林八鸽. “蒙古语常用名词语义研究”. 中央民族大学硕士学位论文, 2005 年.

[12] 张俐、李晶皎、胡明涵、姚天顺. “中文 WordNet 的研究及实现”. 东北大学学报(自然科学版)第 24 卷第 4 期, 2003 年 4 月.

[13] 董振东、董强. “知网和汉语研究”. 当代语言学, 2001. 1.

[14] 姚天顺、张俐、高竹. “WordNet 综述”. 语言文字应用, 2001 年 3 月第 1 期.