

# 互联网视角下蒙古语本体半自动扩充方法研究

李叶青<sup>1</sup> 邱莉榕<sup>2</sup> 赵小兵<sup>2</sup>

(1. 内蒙古财经大学 呼和浩特 010051; 2. 中央民族大学 信息工程学院 北京 100081)

**〔摘要〕** 蒙古语本体资源建设关系蒙古语在语义层面研究的未来发展,在蒙古语自然语言处理中占有重要地位。构建蒙古语本体,合理组织蒙古语的知识概念,将促进蒙古语信息化的进一步发展。蒙古语本体扩充的半自动方法,采用增量法,利用基于人工构建指导本体和自主学习富集的方法来扩充蒙古语本体,旨在增加关于本体论实例的本体知识。在该方法的指导下,可以初步构建较为合理的蒙古语本体。在实验中使用有监督方法构建的本体进行指导,可以扩充出更多合理的蒙古语本体知识,是一种有效的本体构建方法,可以为构建蒙古语语义词典奠定基础。

**〔关键词〕** 语义本体; 蒙古语; 本体学习

**〔中图分类号〕** H212 **〔文献标识码〕** A **〔文章编号〕** 1005-8575 (2015) 01-0142-04

## 一、引言

互联网是一个巨大的信息资源库,人们可以从互联网上获取大量的信息和知识。允许用户与互联网交互并充分发挥用户主动性,从而更容易地查找、共享和组合信息是语义 Web<sup>①</sup>的主要目的。因而,网络知识获取成为当前研究的热点。

由于互联网的大量使用和自然语言处理技术的发展,本体<sup>②</sup>以使用结构化的、机器和人都能理解的方式来组织和共享知识为契机,成为许多知识密集型应用的重要组成部分。<sup>[1]</sup>

本体最为著名并被广泛引用的定义是由

Gruber 提出的“本体是概念模型的明确的规范说明”。<sup>[2]</sup>通俗地讲,本体是用来描述某个领域甚至更广范围内的概念以及概念之间的关系,使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义。<sup>[3]</sup>本体学习(ontology learning)的目标是利用机器学习和统计自然语言处理等技术,自动或半自动地从已有的数据资源中获取期望的本体。<sup>[3]</sup>

自提出基于本体学习的知识获取方法,英语和汉语的知识获取方法都获得迅速的发展。少数民族自然语言处理目前则处于知识获取方法的基础技术研究阶段,所以在民族语言本体方面的研

(收稿日期) 2014—06—24

**〔作者简介〕** 李叶青(1966—),女(蒙古族),辽宁彰武县人,内蒙古财经大学副教授,主要研究方向为自然语言处理。

邱莉榕(1978—),女,山东济南市人,中央民族大学信息工程学院副教授、博士,研究方向为自然语言处理。

赵小兵(1967—),女(蒙古族),内蒙古呼和浩特市人,中央民族大学信息工程学院教授,博士生导师,研究方向为自然语言处理。

**〔基金项目〕** 本文得到内蒙古自然科学基金项目“蒙语语义本体知识表示技术研究”(编号: 2013MS0901)的资助。

① 语义 Web: 是计算机业和互联网业对网络下一阶段发展所做出的术语化定义,其基本含义即基于网络建立任何微小数据的连接,这种连接不仅仅局限于网页。(百度百科术语介绍,语义 Web, <http://baike.baidu.com/view/854337.htm>)

② 本体: 本体是指概念化的明确的规范说明。其目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇间相互关系的明确定义。(百度百科术语介绍,本体, <http://baike.baidu.com/subview/29987/6379240.htm>)

究目前并不多，赵小兵等<sup>[4]</sup>对多民族语言本体知识库<sup>①</sup>构建进行过研究，邱莉榕等<sup>[5]</sup>对藏语中基于上下位关系本体构建的方法做过介绍。除此之外，在其他少数民族语言中并未见到更多的相关研究成果。从互联网上扩充蒙古语本体的技术仍然没有被广泛应用。原因有以下几点：

1. 与英文或中文网站相比，没有数量可观的、经常更新的蒙古语网站可供下载。
2. 与大多数少数民族语言一样，蒙古语语料资源数量很少，知识采集人员在获取知识源时面临数据稀疏这样巨大的问题。
3. 缺乏一定规模的蒙古语电子词典，阻碍了蒙古语知识库构建的发展进程。

为了实现蒙古语本体的自动或半自动扩充，本文采用从收集的文档、字典中学习本体的构建方法。另一方面，在创建和丰富本体过程中，彻底消除人为干预，在目前尚不可行，因此，本文

提出一种半自动的方法。这种半自动方法是一个逐渐丰富的过程，可以从一个固定的概念中自动地选择候补的实例，并由领域专家来进一步丰富本体。

本文提出的本体扩充方法包含人工构建指导本体知识和自主学习扩充本体技术。这项工作的主要创新点如下：

1. 提出一种用于构建蒙古语本体的半监督学习方法。
2. 使用基于概念相似度计算的方法建立扩展概念集，并能借助互联网，从网络文档中提取实例，扩充本体中的实例知识。

## 二、本体扩充方法

本文提出的基于人工构建和自主学习的本体扩充方法基本过程如图1所示：

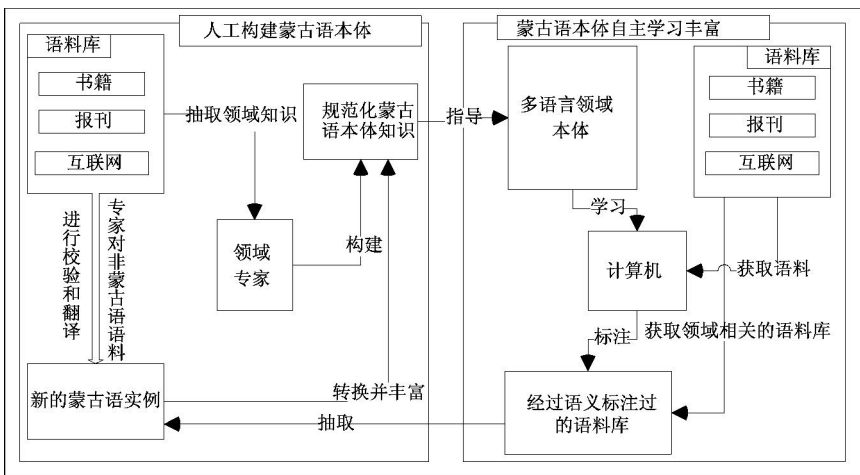


图1 基于人工构建和自主学习的本体扩充方法示意图

可通过以下两步来完成半自动本体扩充：首先，领域专家从已建立的语料库<sup>②</sup>中获取一定规模的蒙古语本体资源；然后，使用已构建的小规模本体资源指导计算机采用机器学习方法自动扩充本体知识。

### （一）人工构建本体过程

由领域专家构建初始本体，并且需要由领域专家根据得到的新实例进行校验和翻译，扩充本

体知识。

1. 构建初始蒙古语语料库资源。
2. 领域专家从语料库中抽取合适的蒙古语概念，构建初始蒙古语本体。
3. 对自主学习过程生成的实例进行校验和翻译（如果实例是从非蒙古语本体中获得），对构建的本体进行扩展。

### （二）自主学习丰富过程

① 知识库：知识库是知识工程中结构化、易操作、易利用、全面有组织的知识集群，是针对某一（或某些）领域问题求解的需要，采用某种（或若干）知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识片集合。（百度百科术语介绍，知识库，<http://baike.baidu.com/view/7976.htm>）。

② 语料库的三个基本认识：（1）语料库中存放的是在语言实际使用中真实出现过的语言材料；（2）语料库是以电子计算机为载体承载语言知识的基础资源；（3）真实语料需要经过加工（分析和处理），才能成为有用的资源。（百度百科术语介绍，语料库，<http://baike.baidu.com/view/686705.htm>）

1. 从已构建的蒙古语本体中选择概念。
2. 用多种语言本体将选定的概念扩展为概念集合。这一步的目的就是扩充概念，因为与英语和汉语本体相比，蒙古语本体的数量相当有限，无法满足应用需求。
3. 根据概念集合，从语料库中选择相关的语义。这一步利用已有概念集对语料库进行标注。这是一种基于概念集的方法，可以利用由当前本体明确导出的概念来对文档进行自动标注。

这称为概念定义。<sup>[6]</sup>

4. 根据自主学习法提取新的实例集合。
5. 计算候选实例属于概念集合的概率，概率的计算是通过概念相似度的计算来实现。
6. 结合语料库资源，对候选实例进行排序。

### (三) 应用实例

这一部分，通过一个简单具体的例子来说明。如图2所示，从蒙古语本体中选择一个概念——“足球”。

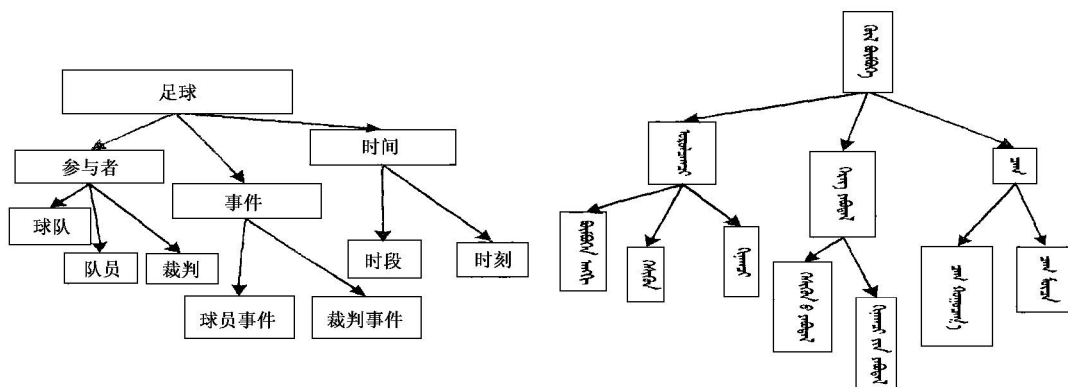


图2 蒙古语的初始本体

利用在互联网上已经发布的“足球”本体作为一个多种语言本体，用多种语言本体来扩展这个概念集合。接着采用算法实现概念的获取、相似度的计算，扩展概念，希望能够在互联网上

搜索更多相关实例。图3所示为领域专家描述的这些新搜索到的概念的所有语义关系。最后，根据 W3C OWL2 标准<sup>①</sup>，本体专家将从中提取出概念和实例来创建一个蒙古语本体。

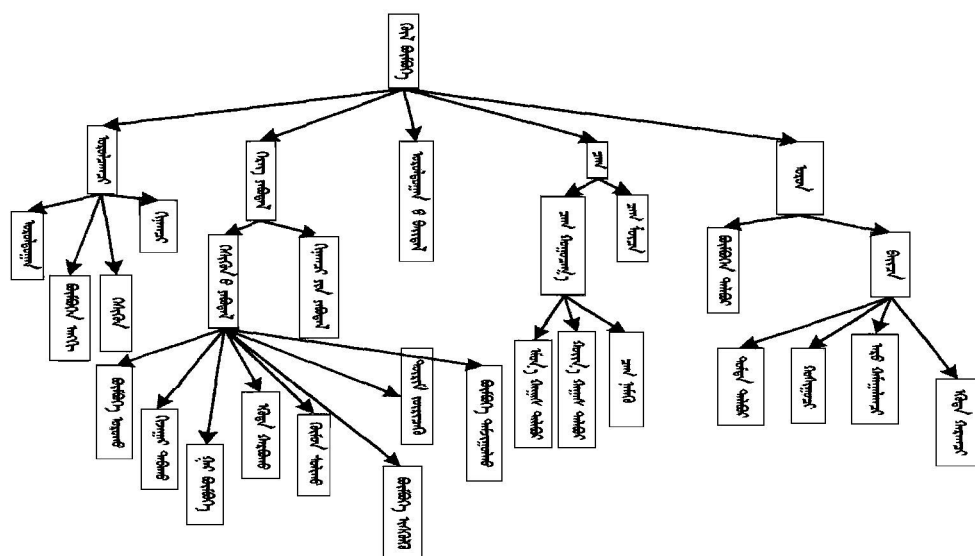


图3 蒙古语中已学习的概念本体

① W3C OWL2 标准：是 W3C 组织在 2012 年 12 月推出的本体语言的推荐标准。OWL2 是 W3C 语义 Web 工具包，允许用户对特定领域知识进行刻画，并通过工具来管理和查询这些语义信息。（中国官方网站，W3C，<http://www.chinaw3c.org/archives/68/>）

### 三、小结与展望

通过万维网可以更加高效、快捷地获取信息和知识。<sup>[7]</sup>一般可以用两种模式来区分本体扩充的方法:一种是使用依赖于术语结构的模式;另一种是使用语境特征的模式。<sup>[8]</sup>

本文提出了一种本体扩充方法,该方法由人工构建和从文本中自主学习富集两部分组成,目的在于扩充蒙古语本体实例的本体知识。论文中

采用了一个简单但真实的例子来验证该方法,初步结果显示该方法是一种实用的解决方案。

本文中的研究成果是我们正在进行的研究工作的一部分,建立蒙古语电子语义词典是我们的研究目的。所建立的语义词典包含丰富的语义信息,可以为蒙古语信息检索和智能服务提供知识支持。下一步,希望能将蒙古语本体扩充过程中的人为操作减少,尽可能实现本体知识的自动获取和扩充。

#### (参考文献)

- (1) A. G. Valarakos and G. Paliouras. Enhancing ontological knowledge through ontology population and enrichment. Engineering Knowledge in the Age of the Semantic Web [M]. Springer Berlin Heidelberg, 2004, 3257: 144 - 156.
- (2) Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications [Z]. Knowledge Acquisition, 5 (2): 199 - 220, 1993.
- (3) 杜小勇,李曼,王珊. 本体学习研究综述 [J]. Journal of Software, 2006, 17 (9): 1837—1847.
- (4) 赵小兵,邱莉榕,赵铁军. 多民族语言本体知识库的构建技术 [J]. 中文信息学报, 2011, 25 (4): 71—74.
- (5) 邱莉榕,翁彧,赵小兵. 藏语语义本体中的上下位关系模式匹配算法 [J]. 中文信息学报, 2011, 25 (4): 45—49.
- (6) 孔敬. 本体学习: 原理、方法与相关进展 [J]. 情报科学, 2006, 25 (6): 657—665.
- (7) F. Amardeilh, P. Laublet and J. -L. Minel. Document annotation and ontology population from linguistic extractions [Z]. In Proceedings of the 3rd international conference on Knowledge capture, October 2 - 5, 2005, Banff, Alberta, Canada.
- (8) H. T. Tanev and B. Magnini. Weakly supervised approaches for ontology population [Z]. EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3 - 7, 2006, Trento, Italy.

## Semiautomatic Expansion Method of Mongolian Ontology in the Perspective of Internet

1. LI Ye-qing, 2. QIU Li-rong, 2. ZHAO Xiao-bing

- (1. Inner Mongolia University of Finance and Economics, Hohhot, Inner Mongolia 010051;
2. School of Information Engineering, Minzu University of China, Beijing, 100081)

**[Abstract]** The construction of Mongolian ontology resources plays an important role in the development of Mongolian semantic research. To construct Mongolian ontology resources and to organize Mongolian concepts can motivate the development of Mongolian information processing. In this paper, we propose a semiautomatic method to extend ontology resources. Considering the usability and adoption, our method utilizes an incremental methodology to populate ontology in Mongolian based on automatic learning. Guided by this method, preliminary and reasonable Mongolian language ontology can be built, and using ontology built by supervision method as guidance in experiments, more reasonable Mongolian ontology can be extended. This is an effective way to build ontology which can lay a foundation for the future compilation of Mongolian semantic dictionaries.

**[Key words]** semantic ontology; Mongolian; ontology learning

(责任编辑 宝玉柱)