

# 蒙古语词类标注系统—AYIMAG

华沙宝

内蒙古大学蒙古语文研究所

**摘要:**本文介绍对蒙古语语料库进行词类标注的两个基本步骤,着重阐述了蒙古语词类标注用产生式规则和词类共现矩阵。其中,产生式规则是根据蒙古语各类构词附加成分和构形附加成分与词干结合规律编排的,它适合于充分利用蒙古文的粘着型特征;词类共现矩阵是对人工标注词类的含10万个单词的语料进行各种词类搭配出现情况统计得到的,它是本标注系统处理兼类词词类归属的主要措施。

**关键词** 蒙古语 词类标注 产生式规则 词类共现矩阵

## The Mongolian Part of Speech Tagging System—AYIMAG

Huashabao

Mongolian Language Institute, Inner Mongolia University

**Abstract** This paper introduced the two establishing steps of the Mongolian part of speech tagging system. And it introduces emphatically the production rule - set and the matrix with the co - occurrence probabilities of Mongolian parts of speech used in the system. Among them, the production rule is compiled and arranged according to the combination rule between various word-formation suffixes and word stems. The rule is suitable to make full use of the agglutinative features of Mongolian language. The matrix with the co-occurrence probabilities of Mongolian parts of speech is produced after the investigation of the corpus containing 100,000 words considering speech tagging words. It is the main measurement of the tagging system as well as the process dealing with the ambiguity of word category.

**Keywords** Mongolian, Part of speech tagging, Production rule - set, The matrix with the co - occurrence probabilities of Mongolian parts of speech

### 1. 引言

随着计算机软、硬件的迅速发展,近年来语料库语言学研究有了新的气色。通讯网络技术和电子出版技术的迅速发展,对多渠道获取语料库生语料提供了极大的方便。在很短的时间内,能够搜集几百万甚至上千万词级的电子文本已经不是什么稀奇的事情。然而,语料库并非资料的任意堆积,它的语料必须经过专门加工,使之具备人们可以利用它进行各种统计、分析、检索、归纳的性能。语料的加工,使生语料变成成熟语料,是语料库建设成败的关键一环。对语料的加工,用于各种不同目的的语料库,虽然各自都有不同的内容和要求,但是,词类标注一项似乎是共同的。

从70年代起,美国人对Brown语料库进行词类标注以来,人们对语料库的词类自动标注进行了大量的研究,提出了基于规则的和基于统计处理的算法。他们通过人工标注、机器统计、提取带词类标记的词频统计表和词类共现频度矩阵、建立概率计算模型等手法,较满意地完成了对语料库自动标注词类的工作。当初对Brown语料库进行词类标注时,准确率只是

77%多,而后来对 LOB 语料库进行词类标注时,准确率确达到了 96%以上。我国语料库建设发展的也很快。清华大学 1992 年研制的汉语词类自动标注系统的正确率达到了 96.8%。北京大学、山西大学等对汉语语料库进行词类标注,也取得了令人鼓舞的成就。

1991 年 12 月,我们完成了国家社会科学基金资助项目—100 万词级的《现代蒙古语文数据库》<sup>①</sup>,对一部分语料做了词类标注。1996 年—1998 年期间,我们承担国家自然科学基金资助项目《对蒙古语语料库的词性标注与统计》<sup>②</sup>,对蒙古语语料库的词性标注工作进行了专题研究。本文就该项目的研究成果,着重介绍实现蒙古语词类标注系统—AYIMAG 的基本步骤和系统所采用的主要方法。

## 2. 蒙古语语料库的词类标注

各个自然语言的词类标注都有其各自的特点,英语与汉语不同,蒙古语也不同于英语和汉语。如汉语是孤立型语言,汉语词类自动标注系统一般采用词典标注、规则标注、统计标注等多种手法,其中统计标注法占的比重较大。蒙古语是粘着型语言,其形态变化非常丰富,更适合于以规则标注法作为实现词类自动标注的主要途径。另一方面,与其它语言一样蒙古语也有相当多的兼类词,附加成分与非附加成分词素同形的现象也很普遍,这些因素加大了标注工作的难度。

对蒙古文语料库的词类标注,我们采取了分两步走的措施。第一步,先对非兼类词和带有动词构形附加成分的兼类词进行词类标注。其中,非兼类词的词类标注主要靠词典和基于词干与附加成分搭配的构词规律。标注独立出现的兼类词的当前词类,是本课题研究的难点。因为,具有词类标注的上下文会给确认兼类词的当前词类提供更多的信息。譬如,可以直接运用蒙古文有关词类的语法理论,还可以统计归纳各种词类搭配共现的频度等等。所以,我们把标注独立出现的兼类词问题留给第二步去处理。

### 2.1 产生式规则集

我们根据蒙古语各类词和构词、构形附加成分的结构特征,根据词干与附加成分的结合规律,归纳了一部规则。这部规则主要由附加成分规则、词序特征规则等组成。附加成分规则部分是一组基于蒙古语各类构词附加成分和构形附加成分与词干结合规律的产生式规则集。

我们用一部非兼类词词典和上述产生式规则来完成非兼类词的词类标注。其中,非兼类词词典是一个随机文件,用二分法查找,构词附加成分按其语法功能分若干个组,分布在几个二维数组之内,分别用顺序查找法和筛选法访问。

产生式规则集主要以①专用符号、②特殊单词、③附加成分、④词干、⑤词干变化、⑥词干与附加成分的匹配、⑦附加成分与附加成分之间的匹配条件作为识别当前单词词类的依据。例如:标注句子

CILUGELEGDE/HU TERE UY-E - YIN [ ]ARSLAN - V BOHO AMI/N TEJIGEL - I

① 该项目由确精扎布教授主持,项目组成员还有华沙宝、齐德华、吉仁尼格、那顺乌日图等。

② 国家自然科学基金资助项目,批准号为 69563002。由笔者主持,吉仁尼格、那顺乌日图、娜仁通拉嘎共同研究。



## 2.2 处理新单词

我们称那些即不在词典又超出本系统智能推理部分的单词为“新单词”。新单词,大多数为复合词、外来语和专用名词。遇到新单词,在其标注词类的位置上统一做记号“YYY”,留做后续编辑对象。同时用一个文件专门记录这些新单词,以便将来根据它们的使用频度,编排到适当的词典之中。

## 2.3 第一步实验结果

为了说明本系统第一步的效果,在这里提供一分实验数据。例如,初中语文第六册(蒙文),共有 19330 个单词。其中通过非兼类词词典确定词类的单词有 15258 个、用产生式规则确定词类的单词有 598 个、独立出现的兼类词有 2781 个、超出以上范围的新单词有 693 个。可见第一步完成的比例占 82% 以上,在词类标注的全部过程中占主导地位。

## 2.4 第二步的介绍

经过第一步的处理,非兼类词和一些非独立出现的兼类词得到了正确的词类标注。但是,独立出现的兼类词,特别是有一些出现频率较高的兼类词所占的比例相当大。这就是说,要提高蒙古文词类标注系统的自动化程度,提高它的实用性,处理兼类词问题是关键。由于确定兼类词的当前词类与该单词的具体环境等多方面因素有着密切的联系,做到完全准确的自动标注几乎是不可能的。<sup>①</sup>

本系统的第二步,实际上就是对独立出现的兼类词进行词类标注的过程。我们选择 0LAN、B0L 等 10 个出现频率较高的兼类词,仍用规则来确定它们的当前词类<sup>②</sup>。由于第一步为第二步提供了一个带有词类标注的环境,制定这 10 个高频兼类词词类标注规则的条件约束就宽松多了。譬如,独立出现的兼类词 0LAN,紧接在它后面的单词,除了 BOGODEGER、DAYAGAR 等少许几个单词以外,其它所有带名词、形容词、动词、数词、量词、连接词、副词词类标记的单词时,它一定是形容词 0LAN(多的),而不是名词 0LAN(大家)。这就是说,有了带有词类标注的环境,利用蒙古语有关词类序列理论的机会就更多了。对于这 10 个高频兼类词,我们用特殊的上下文与一般性的词类序列理论模式相结合的方法编写了词类标注规则。对其他独立出现的兼类词,我们都用统计方法来选定它们在当前语境下的词类。做为统计方法的基础,我们对含有 10 万条单词的语料做了词类标注并进行了严格的把关审核。根据这部分语料,建立了蒙古语词类共现矩阵。搜集了一千多条兼类词,建立了一部兼类词词典,对每个词条都附上了它所有可能取到的词类标记,以便将通过词类共现矩阵计算,从中选择适合当前环境的最佳词类。

<sup>①</sup> 在《现代蒙古语数据库》中,对独立出现的兼类词都是通过人机对话方式标注的。

<sup>②</sup> 它们在 100 万词级《现代蒙古语数据库》语料中出现的次数都超出了 3,000 次。

## 2.5 有关第二步的一个参考数据

为测试第二步的效果,我们从《现代蒙古语文数据库》语料中随意选择一段含 49991 个单词的语料进行了词类标注。其中,41700 个单词经过第一步处理,得到了正确的词类标注。在第二步处理过程中,用规则处理的兼类词,即上述 10 个高频兼类词出现 1149 次,用共现矩阵处理的兼类词出现 7142 次。经过核实发现,用共现矩阵处理的结果之中有 1181 处错误。另外,在这篇语料中遇到新单词 1515 次。这些新单词,除了 209 个复合词和少许专用名词、外来语以外,几乎 2/3 是由于书写不规范而产生的。实验结果表明,蒙古文词类标注系统—AY-IMAG 的准确率目前已经达到 94.6%。如果进一步完善,增补一些有关拼写规范化和构词方面的规则,那么,该系统的准确率就有可能达到 97%~98%。这里顺便说明,蒙古文词类标注系统—AYIMAG 对具有相同词类的同形词没有进一步做更详细的标注<sup>①</sup>。

## 3. 结束语

实践经验告诉人们,若语料没有词类标注,则语料库语言学研究就无法深入进行。也就是说,要进行语料库语言学研究,就必须先走对语料的词类标注这一步。对蒙古语语料库进行词类标注是一项巨大的工程。完成这项工程,一要准,二要省。保证这两点,就要求词类标注系统具备高度的自动化功能。在蒙古语词性标注系统中,产生式规则和蒙古语共现矩阵法对确保系统的准确性和提高系统的自动化程度方面所起到的作用是非常积极的。

迄今为止,在国内外进行的词类标注系统都是针对英、俄、德等屈折型语言或汉语等孤立型语言进行的,尚未发现对类似蒙古语这种粘着型语言进行词类自动标注的研究。英、汉等其它语言的自动标注经验固然对蒙古语自动标注研究有所参考价值,但由于语言类型的差异和蒙古语本身的语法规律,不可能直接套用他人的经验。只有紧紧抓住蒙古语本身的内在规律,才会使蒙古语词类标注系统更加完善。我们用蒙古语词类标注系统—AYIMAG,对含有 100 万单词的语料做了词类标注。标注结果表明,用蒙古语词类共现矩阵方法来选定兼类词归属是一种可取的途径。如果进一步调整蒙古语规则标注方法与蒙古语词类共现矩阵方法之间的有机关系,将会再度提高蒙古语词类标注结果的准确率。

## 参考文献

- [1]. 黄昌宁, 语料库语言学,《中国计算机用户》,1990 年第 11 期。
- [2]. 白栓虎, 汉语词切分及词性自动标注一体化方法,《计算语言学进展与应用》,陈力为、袁奇主编,清华大学出版社,1995。
- [3]. 周强、俞士汶, 一个人机互助的汉语语料库多级加工处理系统 CCMP,《计算语言学进展与应用》,陈力为、袁奇主编,清华大学出版社,1995。

<sup>①</sup> 在各种蒙古文辞典里,对词类相同的同形词用角码(或者用圈码、阴阳码)标注,以示它们之间的区别。在《现代蒙古语文数据库》中,通过人机对话方式也做了类似的标注。