

蒙古语兼类词词性标注的处理

那日松¹，敖其尔²

(内蒙古大学 计算机学院, 内蒙古 呼和浩特 010021)

摘要: 本文主要解析了通过最大概率值词性标注和同现概率值词性标注这两种方法对蒙古语的兼类词进行词性标注的过程, 并给出了它们详细的实验步骤和结果。根据分析过程, 可以了解这两种方法的利弊。

关键字: 词性标注; 兼类词; 最大概率值词性标注; 同现概率值词性标注

中图分类号: H085.6

文献标识码: A

一、引言

语料库加工的基本层次就是确定词性, 因此, 词性标注是自然语言处理, 包括机器翻译的基本处理步骤。词性标注的任务是为一个句子中的每个词都标上对应的词性。对于机器翻译来说, 词性标注不仅为后续句法分析奠定了基础, 而且缩小了译文选择的范围, 为译文生成做出了贡献。

但是自然语言中的词性兼类的现象很普遍。所熟悉的汉语兼类词非常复杂, 那么蒙古语也不例外。蒙古语词的应用也非常灵活, 可以充当不同的句子成分, 所以也可以说蒙古语的词性兼类也和其他语言一样特别多, 特别复杂。蒙语中词的兼类现象具体表现有形同音不同、同形同音但意义无关联、具有典型意义的兼类词、很多情况的组合等。

由于蒙古语词性兼类的普遍性, 所以兼类消歧具有重要意义。考虑到标注兼类词的重要性, 这里我通过两种方法解决兼类词的词性标注的问题。

二、标注集的确定

词性标注的第一步确定了标注集。

使用 20 个标注符号组成的标注集。

AYA:感叹词

CON:时位词

DAG:后置词

DAU:后置词

DAY:副词

DVR:模拟词

HND:情态词

HMN:量词

H0L:连词

H0U:联系动词

JIN:名词

SVL:语气词

TEN:形容词

TON:数词

TOU:代词

UIL:动词

]]:地名

=: 一个成分

[]: 人名

., ? , !, 《, : , : 逗号, 问号, 感叹号, 双引号, 分号

三、蒙古语词性标注过程

对 1250 个蒙语句子的熟语料库进行词性标注。

通过程序对熟语料进行统计,得出在这 1250 个句子中,共有 5783 个单词,52 个兼类词。那么重点就在于对这 52 个兼类词的词性标注。用什么方法才能提高对兼类词词性标注的准确率呢?下面用两种方法,一种是最大概率词性标注,另一种是同现概率值词性标注。通过用这两种方法标注这些兼类词的详细说明后,可以很容易地看出他们的利弊。

先列出所有兼类词和他们所对应的词性以及这些词性在语料库中出现的次数:

表 1 兼类词

第一个词性和出现次数	第二个词性和出现次数	第三个词性和出现次数
0CI JIN 1	0CI UIL 2	
0LAN TEN 19	0LAN JIN 4	
0R0I TEN 1	0R0I JIN 1	
0YIR_A-YIN TEN 3	0YIR_A-YIN C0N 1	
A AYA 12	A SVL 6	
ADHV UIL 1	ADHV JIN 1	ADHV HMN 1
ALI TON 11	ALI UIL 2	
ARAI DAY 9	ARAI HND 3	
ARV JIN 2	ARV C0N 3	
AY_A JIN 1	AY_A AYA 1	

B0L SVL 3	B0L UIL 42	
B0LH0R H0L 2	B0LH0R UIL 1	
BA H0L 41	BA H0U 1	
BAGATVR JIN 4	BAGATVR TEN 6	
BAYALIG JIN 1	BAYALIG TEN 2	
BVI SVL 44	BVI UIL 14	
BVSV SVL 5	BVSV TON 1	
BVYV H0L 13	BVYV SVL 2	
CAG HMN 7	CAG JIN 21	
CAG-TV HMN 3	CAG-TV JIN 3	
CAG-VN HMN 1	CAG-VN JIN 5	
DAHI/N JIN 1	DAHI/N UIL 18	
E SVL 35	E AYA 3	
EDUI UIL 2	EDUI TON 4	EDUI SVL 2
EGEGE HND 1	EGEGE DAY 1	
EMEGTEI TEN 2	EMEGTEI JIN 2	
EREGTEI TEN 2	EREGTEI JIN 2	
GAJAR HMN 6	GAJAR JIN 19	
GAJAR-I HMN 1	GAJAR-I JIN 4	
GAJAR-VN HMN 2	GAJAR-VN JIN 6	
GEM SVL 1	GEM JIN 2	
HACIN TEN 3	HACIN JIN 3	
HARI UIL 3	HARI JIN 2	
HECEGUU DAY 1	HECEGUU TEN 3	
HEDUYINEI C0N 2	HEDUYINEI DAY 1	
HOGSIN JIN 1	HOGSIN TEN 8	
JAH_A-DV C0N 1	JAH_A-DV JIN 1	
JERGE DAG 2	JERGE C0N 2	
LAB DAY 4	LAB HND 1	
NAM JIN 3	NAM TEN 4	

OGEDE C0N 2	OGEDE JIN 1	
S0NIN JIN 6	S0NIN TEN 6	
SIG DAU 21	SIG DAG 1	SIG SVL 1
SIGVD DAY 4	SIGVD TEN 1	
SIR_A TEN 3	SIR_A JIN 1	
TAL_A JIN 1	TAL_A TEN 1	
TVLA DAG 1	TVLA DAU 29	
UGEI SVL 164	UGEI UIL 85	
UGEI-BER UIL 2	UGEI-BER SVL 1	
UGEI-DEGEN SVL 2	UGEI-DEGEN UIL 3	
UGEI-YI UIL 1	UGEI-YI SVL 1	
YERU TEN 3	YERU HND 10	

（一）最大概率值词性标注

最大概率值词性标注的特点是容易实现，可以很简单地做到兼类词的词性标注，但是相对而言，正确率比较低。

具体步骤是：

1. 把每个兼类词所对应的所有词性列出来。
2. 记录下每个兼类词在语料库中出现的次数最多的词性。

注：如果兼类词的 2 个或者更多个词性在语料库中出现的次数相等，在这种情况下随机记录此兼类词的一个词性。

3. 取第二步记录下的信息，进行语料库中兼类词的标注。
4. 对标注完的兼类词进行测试。看下图：

表 2 最大概率值词性标注详细表

兼类词	总出现个数	第一个词性出现的个数和标注正确的个数	第二个词性出现的个数和标注正确的个数	第三个词性出现的个数和标注正确的个数	标注正确的总个数
OCI	3	1-0	2-2	0-0	2
OLAN	23	19-19	4-0	0-0	19
OROI	2	1-1	1-0	0-0	1
OYIR_A-YIN	4	3-3	1-0	0-0	3

A	18	12-12	6-0	0-0	12
ADHV	3	1-1	1-0	1-0	1
ALI	13	11-11	2-0	0-0	11
ARAI	12	9-9	3-0	0-0	9
ARV	5	2-0	3-3	0-0	3
AY_A	2	1-1	1-0	0-0	1
BOL	45	3-0	42-42	0-0	42
BOLHOR	3	2-2	1-0	0-0	2
BA	42	41-41	1-0	0-0	41
BAGATVR	10	4-0	6-6	0-0	6
BAYALIG	3	1-0	2-2	0-0	2
BVI	58	44-44	14-0	0-0	44
BVSV	6	5-5	1-0	0-0	5
BVYV	15	13-13	2-0	0-0	13
CAG	28	7-0	21-21	0-0	21
CAG-TV	6	3-3	3-0	0-0	3
CAG-VN	6	1-0	5-5	0-0	5
DAHI/N	19	1-0	18-18	0-0	18
E	38	35-35	3-3	0-0	35
EDUI	8	2-0	4-4	2-0	4
EGEGE	2	1-1	1-0	0-0	1
EMEGTEI	4	2-2	2-0	0-0	2
EREGTEI	4	2-2	2-0	0-0	2
GAJAR	25	6-0	19-19	0-0	19
GAJAR-I	5	1-0	4-4	0-0	4
GAJAR-VN	8	2-0	6-6	0-0	6
GEM	3	1-0	2-2	0-0	2
HACIN	6	3-3	3-0	0-0	3
HARI	5	3-3	2-0	0-0	3

HECEGUU	4	1-0	3-3	0-0	3
HEDUYINEI	3	2-2	1-0	0-0	2
HOGSIN	9	1-0	8-8	0-0	8
JAH_A-DV	2	1-1	1-0	0-0	1
JERGE	4	2-2	2-0	0-0	2
LAB	5	4-4	1-0	0-0	4
NAM	7	3-0	4-4	0-0	4
OGEDE	3	2-2	1-0	0-0	2
SONIN	12	6-6	6-0	0-0	6
SIG	23	21-21	1-0	1-0	21
SIGVD	5	4-4	1-0	0-0	4
SIR_A	4	3-3	1-0	0-0	3
TAL_A	2	1-1	1-0	0-0	1
TVLA	30	1-0	29-29	0-0	29
UGEI	249	164-164	85-0	0-0	164
UGEI-BER	3	2-2	1-0	0-0	2
UGEI-DEGEN	5	2-0	3-3	0-0	3
UGEI-YI	2	1-1	1-0	0-0	1
YERU	13	3-0	10-10	0-0	10

在语料库中所有兼类词的个数是 819。利用取最大概率值词性标注后标注正确的兼类词的个数是 615，所以正确率是： $615/819 \approx 0.75$ ，也就是 75%。

（二）同现概率值词性标注

同现概率值词性标注与最大概率值词性标注相比，更复杂一些。这里记录下兼类词后续一个词的词性。思想是：在最大概率值词性标注的基础上，对错误标注的兼类词根据记录的后续词的词性再做一次标注。所以同现概率值词性标注的正确率肯定比最大概率词性标注高。

具体步骤是：

1. 把上次标注中错误标注的兼类词取出来。
2. 记录下兼类词后续一个词的词性。
3. 根据这些兼类词和记录下的后续词性的搭配，再对这些兼类词进行标注。

4. 对标注完的兼类词进行测试。看下图：

注：图中记为“*”符号的项表示，同现概率值词性标注与最大概率值词性标注相比，正确标注的数量改进的兼类词。

表 3 同现概率值词性标注详细表

	兼类词	总出现个数	第一个词性出现的个数和标注正确的个数	第二个词性出现的个数和标注正确的个数	第三个词性出现的个数和标注正确的个数	标注正确的总个数
*	OCI	3	1-1	2-2	0-0	3
	OLAN	23	19-19	4-0	0-0	19
	OROI	2	1-1	1-0	0-0	1
	OYIR_A-YIN	4	3-3	1-0	0-0	3
*	A	18	12-12	6-3	0-0	15
*	ADHV	3	1-1	1-1	1-1	3
*	ALI	13	11-11	2-2	0-0	13
	ARAI	12	9-9	3-0	0-0	9
	ARV	5	2-0	3-3	0-0	3
*	AY_A	2	1-1	1-1	0-0	2
*	BOL	45	3-1	42-42	0-0	43
*	BOLHOR	3	2-2	1-1	0-0	3
	BA	42	41-41	1-0	0-0	41
*	BAGATVR	10	4-3	6-6	0-0	9
*	BAYALIG	3	1-1	2-2	0-0	3
*	BVI	58	44-44	14-6	0-0	50
	BVSV	6	5-5	1-0	0-0	5
*	BVYV	15	13-13	2-2	0-0	15
	CAG	28	7-0	21-21	0-0	21
*	CAG-TV	6	3-3	3-2	0-0	5
	CAG-VN	6	1-0	5-5	0-0	5
	DAHI/N	19	1-0	18-18	0-0	18
	E	38	35-35	3-3	0-0	35

*	EDUI	8	2-2	4-4	2-1	7
*	EGEGE	2	1-1	1-1	0-0	2
*	EMEGTEI	4	2-2	2-2	0-0	4
	EREGTEI	4	2-2	2-0	0-0	2
*	GAJAR	25	6-1	19-19	0-0	20
*	GAJAR-I	5	1-1	4-4	0-0	5
	GAJAR-VN	8	2-0	6-6	0-0	6
*	GEM	3	1-1	2-2	0-0	3
*	HACIN	6	3-3	3-1	0-0	4
*	HARI	5	3-3	2-1	0-0	4
*	HECEGUU	4	1-1	3-3	0-0	4
	HEDUYINEI	3	2-2	1-0	0-0	2
	HOGSIN	9	1-0	8-8	0-0	8
	JAH_A-DV	2	1-1	1-0	0-0	1
	JERGE	4	2-2	2-0	0-0	2
*	LAB	5	4-4	1-1	0-0	5
*	NAM	7	3-3	4-4	0-0	7
	OGEDE	3	2-2	1-0	0-0	2
*	SONIN	12	6-6	6-4	0-0	10
	SIG	23	21-21	1-0	1-0	21
*	SIGVD	5	4-4	1-1	0-0	5
*	SIR_A	4	3-3	1-1	0-0	4
	TAL_A	2	1-1	1-0	0-0	1
	TVLA	30	1-0	29-29	0-0	29
*	UGEI	249	164-164	85-1	0-0	165
	UGEI-BER	3	2-2	1-0	0-0	2
*	UGEI-DEGEN	5	2-1	3-3	0-0	4
	UGEI-YI	2	1-1	1-0	0-0	1
	YERU	13	3-0	10-10	0-0	10

利用同现概率值词性标注后标注正确的兼类词的个数是 664, 比起最大概率词性标注的正确标注兼类词的个数增加了 49。所以正确率是 $664/819 \approx 0.81$, 也就是 81%。

蒙古语的兼类词标注一直是一个难题, 至今还没有一个最好的解决方法。从上面的数据可以看出, 同现概率值词性标注还是可取的。如果想提高兼类词词性标注的正确率, 还需要利用其它办法, 比如同现概率和隐马尔可夫模型的结合等。

参考文献

- [1] 赵铁军. 机器翻译原理 [M]. 哈尔滨: 哈尔滨工业大学出版社, 2000.
- [2] 姚天顺, 驻靖波, 张琨, 杨莹, 等. 自然语言理解 [M]. 北京: 清华大学出版社, 2002.
- [3] 白栓虎. 基于统计的汉语语料库词性自动标注的研究与实现 [M]. 北京: 清华大学出版社, 1996.
- [4] 陈丹琪. 统计与规则相结合的英语词性标注和基本名词短语分析 [M]. 哈尔滨工业大学硕士学位论文, 1999.
- [5] 白栓虎, 夏莹, 黄昌宁, 等. 汉语语料库词性标注方法研究 [M]. 北京: 电子工业出版社, 1991.

The management of Mongolian sylleptic word labeling

Narisu¹, Ochir²

(School of Computer Science, Inner Mongolia University, Hohhot 010021)

Abstract : The process that labeled Mongolian sylleptic word with part of speech through labeling part of speech with maximal probability value and with concurrence probability value mainly analyzed in this paper. Moreover, there presented their detailed experiment step and result. Advantages and disadvantages for this two ways must be realized according to analyzing process.

Key word: labeling part of speech; sylleptic word; labeling part of speech with maximal probability value; labeling part of speech with concurrence probability

收稿日期: 2004-9-14;

基金项目: 教育部人文社会科学研究重大项目 “蒙古文信息处理平台 (MIPP) 的研究”

(02JAZJD850003); 国家自然科学基金项目 “基于 WEB 的蒙文图书信息管理系统” (60163003)

作者简介: 1. 那日松 (1980 -), 女, 蒙古族, 内蒙古兴安盟扎赉特旗人, 内蒙古大学计算机学院 2002 级研究生。主要从事蒙文信息处理和计算语言学方面的研究。2. 敖其尔 (1941 -), 男, 蒙古族, 内蒙古兴安盟人。内蒙古大学计算机学院教授, 内蒙古大学蒙古学研究中心兼职教授, 硕士生导师, 主要研究蒙文信息处理和计算语言学。