基于统计语言模型的蒙古文词切分*

侯宏旭[1,2,3] 刘 群[1] 那顺乌日图[2] 牧仁高娃[2] 李锦涛[1]

[1](中国科学院计算技术研究所 智能信息处理重点实验室 北京 100190)
[2](内蒙古大学 计算机学院 呼和浩特 010021)
[3](中国科学院研究生院 北京 100190)

摘 要 通过对蒙古文词切分技术的分析,利用规则作为切分的基础,提出一种统计和规则相结合的蒙古文词切分方法.这种方法利用蒙古语统计语言模型作为排歧依据,使用的语言模型有基于词性的语言模型和 Skip-N 语言模型.其词切分准确率比基于规则的系统有较大提高.

关键词 蒙古语,词切分,语言模型,词干词缀
中图法分类号 TP 391

Mongolian Word Segmentation Based on Statistical Language Model

HOU Hong-Xu[1,2,3], LIU Qun[1], Nasanurtu[2], Murengaowa[2], LI Jin-Tao[1]
[1](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2](*School of Computer Science, Inner Mongolia University, Huhhot* 010021)
[3](*Graduate University of Chinese Academy of Sciences, Beijing* 100190)

ABSTRACT

Based on the analysis of Mongolian segmentation technique and the rules used as the foundation of word segmentation, a hybrid word segmentation method is proposed. It uses Mongolian statistical language model to eliminate the ambiguity in Mongolian word segmentation. A POS language model and a Skip-N language model are used, and an experiment system is thus created. The experimental results are better than those of the system based on rules.

Key Words Mongolian, Word Segmentation, Language Model, Etyma and Suffix

1 引 言

蒙古语属于黏着型语言.蒙古语的构词、构形都是通过在词干后面缀接不同的词尾来实现的.而且还可以不断缀接,因此蒙古语词法形态变化丰富且复杂[1-2].

在基于语料库的机器翻译系统中,我们需要将

* 内蒙古自然科学基金项目(No. 200607010805)、国家 973 计划前期研究项目(No. 2007CB316503)资助
收稿日期:2008-03-03;修回日期:2008-06-05
作者简介 侯宏旭,男,1972 年生,副教授,主要研究方向为中文信息处理.E-mail:cshhx@imu.edu.cn.刘群,男,1966 年生,研究员,主要研究方向为自然语言处理.那顺乌日图,男,1959 年生,教授,主要研究方向为计算语言学.牧仁高娃,女,1982 年生,硕士研究生,主要研究方向为计算语言学.李锦涛,男,1962 年生,研究员,主要研究方向为数字媒体处理技术.

原始的语料库切分成具有词干词缀信息以及词性信息的语料库[3].

蒙古语词类包含"无词形变化的词"和"有词形变化的词".其中有词形变化的词又可分为"体词"和"动词",体词包含名词、形容词、代词、数词等.体词主要有数、格、领属、级等范畴,而动词关系到时、态、及形动词、副动词等变化形式[4].

蒙古语构形的附加成分承载着大量的语法信息,如果把蒙古语词只作为一个整体来处理,就损失了大量的语法、语义信息[5-7].而且由于收录词的形态变化而派生的新词,蒙古语词典的规模也会变得非常大.例如,在词典中,如果要列出"IR_E"的所有形式,那可能要列出几百种变化,而且即使如此也不一定能穷举出所有的变化.因此,在传统的语言学词典中只收录一个条目,即"IREHU"这个原始现在形动词形式.如果没有好的蒙古语词切分工具,这样的词典在机器翻译中恐怕没有任何用处.因此,如果我们把蒙古语词切分成词干、词缀,并且在每个构成成分后面标注其属性,那么词典中只需要有"IR_E"这个词干,系统就可得到所有以它为词干的词的含义了.

蒙古语词切分在某些方面和汉文的分词有些相像,但是也具有其特点[8].我们利用规则作为切分的基础可大大简化词切分的复杂度.

2　基于规则和词典的词切分

对蒙古文词切分研究得不多,目前较有成效的方法是基于规则和词典相结合的方法[1-2].

2.1　词典方法

在基本的词切分系统中,我们通过查词典的方式查到一个词是由哪些词干和词缀构成的,这种方法就是基于词典的词切分方法.

在基于词典的词切分中,词典中列出了所有可能的词切分方案.所以词典的词切分可达到非常高的准确率.

但是,词典也有缺点,具体如下.

1)词典方法不能穷举所有词的所有变化形式.词典方法只能给出常用词的切分形式.如果要列出所有的词的话,词典将会非常庞大,所需要的人力物力也是不能支持的.

2)词典方法仍然存在二义性.有很多都具有两种或以上的切分方法,这样的切分方法还需要其它方法进行排歧.

2.2　基于规则的方法

由于基于词典的方法对于未登录词的处理无能为力,因此我们希望通过另外的方法解决这一问题.

事实上,词的切分还是有规律可循的.例如,假设我们不知道"0CIBA"的切分方法,但是我们知道词缀中有"BA",这样我们就可以猜测"0CIBA"是否可以切分成"0CI"+"BA".这样我们就可以建立这样的规则:

$$*BA \Rightarrow * + BA\ .$$

这样的情况确实存在,因此我们可通过查询词缀表,找到所有可能的词缀,进行切分.

但是,实际情况并没有这么简单,把词切分成词干、词缀的时候会发生词形的变化.例如,前面提到的"IREGSEN"在切分成词干/词缀形式的时候变成了"IR_E+GSEN",即前面的部分发生了变化.因此我们对前面的规则形式进行修改,例如

$$*EGSEN \Rightarrow *\_E + GSEN,$$

这样的话就可在完成词切分的同时还原词干的词性.这样,我们编写了400多条词切分规则.这些规则是在文献[1]、[9]中提到的规则基础上修改得到的.利用这些规则,我们可用它生成所有可能的切分.但是这些切分并不一定都是合法的.

基于规则的切分同样存在两个问题.

1)切分的错误.并不是所有符合规则的切分都是合乎蒙古语构形法的,例如"0CIBA"可以切分成"0CI/Ve + BA/Fs14"或"0CI/Ne + BA/Fs14"(斜线后面的是词干或词缀的词性标记).其中,+BA/Fs14是一个动词后缀,只能跟在动词后面.也就是说,虽然0CI既可能是动词(Ve)也可能是名词(Ne),但是在这里,只能切分成"0CI/Ve + BA/Fs14".那么,如何解决这个问题呢?一种方法是通过建立新的规则来去除这样的问题.例如建立一个词性限定规则

$$*BA \Rightarrow */V + BA/Fs14\ ,$$

这里我们可以看到由于限定条件"*/V"出现,切分后词干部分必须是动词(V).

2)切分的二义性.并不是因为采用规则以后就可以完全排除二义性.在规则库中,我们可以看到这样两条规则:

@@*VHAN|*UHEN→*VV|*UU/! +HAN|HEN/Fa1,

@@*HAN|*HEN→*N/! +HAN|HEN/Fa1,

这里可能会出现两个规则都适用的问题.解决方案是采用最大匹配,这样,我们每次只匹配长度最大的那个规则.即如果后缀是"VHAN"的时候,我们只匹配第一条规则,而放弃第二条.

但是这样还是不能完全解决问题.在实际应用中,我们发现,事实上同一个词还是可能存在两个以

上合法的切分结果. 例如,YAGAHIGSAN 既可能是动词“YAGAHI/Vt + GSAN/Ft11”,也可能是副词“YAGAHIGSAN/Db”. 我们需要添加规则来去除这样的二义性. 但是由于这样是依赖于上下文的,需要建立更复杂的上下文模型. 因此,和基于词典的词切分模型一样,需要一些其它的手段去排除这样的歧义.

2.3 基于词典和基于规则方法的缺点

综合前面对基于词典和基于规则方法的分析我们可以看到,无论是哪种方法,都存在以下共同的问题:1)无法涵盖所有的切分情况,2)对歧义的情况很难处理. 在文献[1]、[9]中提到,在一个 1 870 个词的测试语料上测试结果的准确率大约是 86%,而且这些词中仅有 0.02% 是未登录词. 可以想象,如果对随机篇章中的内容进行切分,准确率还会降低.

3 基于统计语言模型的词切分

由于无论使用哪种方法,我们都需要采用进一步的手段来修正切分的结果. 在这里我们将使用统计语言模型[10].

统计语言模型中最常用的模型是 n 元语法,我们这里也使用 n 元语法及其改进形式来进行蒙古语的切分.

3.1 蒙古语语言模型

通过对蒙古语言模型构建方法的研究,我们发现,可以从以下 3 个层面构建不同的语言模型.

1)词的语言模型. 语言模型的训练以词为基础,其基本单位是词.

2)词干/词缀的语言模型. 语言模型训练的基础是词干和词缀. 利用这样的方法建立的语言模型可反映词干词缀的关系.

3)基于字母的语言模型. 在字母级别上建立的语言模型可用作文字识别上的校对.

从这里我们可以看到,如果要在词切分中利用语言模型,需要在词干词缀上建立语言模型. 但是,事实上,单纯使用 n 元语言模型有一个缺点,就是仅能表达相邻 n 个词(实际上,在基于蒙古文词干、词缀的模型上是更小的词干、词缀)的依赖关系,对于更长距离的依赖无法表示. 因此,我们在此基础上又增加了两种语言模型:1)Skip-N 模型[11]. Skip-N 模型是利用相隔 n 个单词的依赖关系训练得到的. 2)词性的语言模型.

3.2 Skip-N 语言模型

通常使用的 n 元语言模型在表达长距离的依赖关系的时候存在明显不足,尤其是当蒙古文词被切分成若干片段后,这样的长距离依赖更加明显. 因此,提出一种 Skip-N 语言模型,用于描述这种更长距离的依赖关系:

$$P(w_i \mid w_1 w_2 \cdots w_{i-1}) = \prod_{j=1}^{k} P(w_i \mid w_{i-j}).$$

Skip-N 语言模型实际上是二元语言模型的一种变化形式,其引入距离变量 k,用于描述相距 k 个词的长距离依赖关系. 由于随着 k 的增大,Skip-N 语言模型的计算量增加很快,在实验中,我们设定 k 为 9. 该语言模型采用 Katz 平滑技术[12] 进行平滑.

经过实验我们可以看到,针对蒙古文的词切分问题,Skip-N 语言模型是较有效的.

3.3 基于词性的语言模型

事实上,我们发现除了词语上的关联关系外,蒙古语的词在进行切分以后还存在词性上的关联关系. 正如前面的例子中看到的“0CIBA”为什么只能切分成“0CI/Ve + BA/Fs14”是因为“Fs14”是一个动词的后缀,只能跟在动词词干后面. 也就是说,“Ve”词性和“Fs14”词性有一定的关联关系. 因此,我们提取出训练语料中的词性信息,在其上训练一个词性的语言模型. 词性语言模型的训练语料是从全部训练语料中去掉词只保留词性得到的. 利用这个训练出三元语言模型.

通过一些初步的实验,我们采用两种不同的词性语言模型.

1)严格的词性模型:

Rb Sf Rb Zx Fc21 Nn Fc11 Nt.

2)忽略词干二级词性的模糊词性模型:

Rb Sf Rb Zx Fc21 N Fc11 N.

实验中模糊词性模型去掉名词(N)、动词(V)这些词干词性的二级词性. 经过实验,我们发现,采用模糊的词性模型得到的切分效果要更好. 因此,后面的实验数据中给出的词性模型就是以模糊的词性模型建立的.

3.4 词切分算法

基于语言模型的词切分算法的核心是语言模型. 工作可以分成两个部分:1)生成所有可能的切分;2)评价生成的结果,选出最终的分词结果.

我们的方法是通过切分规则表来生成所有可能的切分. 需要注意的是,这里的规则和前面提到的基于规则的词切分方法中的规则不同. 这里的规则要求要比基于规则的方法更宽泛,也就是说,我们不必去关心切分后的结果是否正确,只需要给出所有的可能性就可以了. 具体的哪些正确、哪些不正确由语

言模型确定. 从复杂度上来看是用规则作为切分依据的方法能够生成的切分情况最少、复杂度最低. 有了候选的词切分结果以后,就可以利用语言模型来进行评价:

$$\log P(S) = \lambda_1 \log P_1(S) + \lambda_2 \log P_2(S) + \lambda_3 \log P_3(S).$$

这里的 3 个语言模型分别如下:

1) $P1$:基于词干 / 词缀的 n 元语言模型;

2) $P2$:基于词干 / 词缀的 SKIP-N 语言模型;

3) $P3$:词性的语言模型.

这里有3个权值. 权值可通过最小错误率训练[13] 得到.

4　实验结果

由于基础数据的不足,我们实验中制作了一个 3 万 8 千句的蒙古语单语语料,约含 33 万蒙古语词.

我们另外选择 500 个蒙古文句子作为测试集,并人工编写这些句子的参考答案. 每个句子有一个参考答案.

评价方法采用准确率(prec)和召回率(recall)以及 F_1 值. 其定义如下:

$$prec = \frac{\text{正确的切分单元个数}}{\text{切分出的单元个数}},$$

$$recall = \frac{\text{正确的切分单元个数}}{\text{参考答案中切分出的单元个数}},$$

$$F_1 = \frac{2 \times prec \times recall}{prec + recall},$$

其中切分的单元为词干或词缀.

原始语料

AI ! TERE NIGENTE NIGE VDAG_A SILGALTA-DV HIRI TENGCEGSEN UGEI BOLJAI, BI TEGUN-DU SEREMJI OGBEL SAYIN.

TEDE BOL MINU SAYIN NAYIJI MON, TEGUSUGSEN-U DARAG_A BI MON TEDEN-TEI HARILCAJV BAYIBAL SAYIN.

NAYIJI-YIN GER-TU NIGE HONON_A.

切分后的语料

AI/Is! TERE/Rj NIGENTE/Dc NIGE/Mu VDAG_A/Qn SILGALTA-DV/Fc21 HIRI/Ne TENGCE/Ve +GSEN/Ft11 UGEI/Ve BOL/Ve +JAI/Fs11, BI/Rb TEGUN-DU/Fc21 SEREMJI/Ne OG/Vt +BEL/Fn71 SAYIN/Ac.

TEDE/Rb BOL/Ve MI/Nt +N/Fn3 +U/Zv2 SAYIN/Ac NAYIJI/Nt MON/Sb, TEGUSUGSEN-U/Fc11 DARAG_A/Oa BI/Rb MON/Sb TEDEN-TEI/Fc61 HARILCA/Ve +JV/Fn1 BA/Cw +YI/Fc32 +BAL/Fn71 SAYIN/Ac.

图 1　训练语料的例子

Fig. 1　Example for training corpus

测试集

NAYIJI-YIN JOBALANG-IYAN TOGACIHV-YI SONOSBA.

NAYIJI-YIN-IYAN SANAG_A SEDHIL-I OYILAGABA.

NAYIJI-DV NIGVCA-BAN HELEJU OGBE.

NIGE NIGVCA-YI ILECILEN_E.

参考答案

NAYIJI-YIN/Fc11 JOBALANG-IYAN/Fx11 TOGACIHV-YI/Fc31 SONOS/Vt +BA/Fs14.

NAYIJI-YIN-IYAN/Fx11 SANAG_A/Ne SEDHIL-I/Fc31 OYILA/Vt +G/Fb31 +A/Zv1 +BA/Fs14.

NAYIJI-DV/Fc21 NIGVCA-BAN/Fx11 HELE/Nt +JU/Fn1 OG/Vt +BE/Fs14.

NIGE/Mu NIGVCA-YI/Fc31 ILECILE/Vt +N_E/Fs21.

图 2　词切分的测试集和参考答案

Fig. 2　Test set and keys of word segmentation

实验中我们首先采用规则进行粗切分,采用的规则是在原规则系统上的 400 多条规则中添加了几条不确定的切分规则(见图 3). 原来的规则要求切分必须是正确的,而新加的规则不保证切分得到的结果一定正确,只是给出所有的可能性,其中可能有正确的切分结果,而排歧通过语言模型进行.

```
@@ 'AYI& → 'AI/! +&/!
@@ 'EYI& → 'EI/! +&/!
@@ 'OYI& → 'OI/! +&/!
@@ 'VYI& → 'VI/! +&/!
@@ TVNGG' → TVN/V +G'/'
@@ 'LO→ 'L/' +O/Zv2
```

图 3　部分切分规则

Fig. 3　Some of segmentation rules

表 1　蒙古语词切分实验结果

Table 1　Results of Mongolian word segmentation

	准确率	召回率	F_1 值
参考系统	0.860	—	—
仅规则	0.525	0.666	0.587
规则 + 三元语言模型	0.878	0.852	0.865
规则 + 三元语言模型 + 严格词性的语言模型	0.902	0.851	0.876
规则 + 三元语言模型 + 模糊词性语言模型	0.914	0.865	0.889
规则 + 三元语言模型 + 模糊词性语言模型 + Skip-N 语言模型	0.939	0.867	0.902

表 1 是使用不同方法的实验结果. 在表 1 中,1) 参考系统是文献[11]给出的一个基于规则和词典的系统. 由于文中并未给出召回率的数值,我们只能比较准确率. 需要说明的是,由于我们无法得到文献[11]的测试环境,因此两者的测试环境有一定的差别,数值比较只能作为参考. 2) 三元语言模型采用 SRI 开发的 SRILM(SRI Language Model)语言模型工具包. 我们在训练语料上训练三元语言模型,采用 modified Kneser-Ney 平滑算法. 3) 词性语言模型是从训练语料中提取词性信息,形成词性文件,然后再用 SRILM 训练出一个词性层面的三元语言模型. 4) SKIP-N 语言模型采用我们自行编写的语言模型的训练程序.

表 1 中后 3 个系统分别是加上不同语言模型后得到的结果. 第 3 个系统和第 4 个系统分别采用严格词性的语言模型和模糊词性的语言模型,从结果来看模糊词性的语言模型效果要更好一些. 从实验结果可以看到,利用语言模型可得到较满意的词切分结果.

5 结束语

虽然我们的词切分结果达到 94% 的准确率,但是,这对于一个词切分系统来说还不够. 还需要其它的手段来提高准确率. 这主要通过增加训练集的规模来进行. 目前采用的训练集的大小只有 38 000 句,仍然属于较小的,会在今后不断扩大语料库. 另一个是希望通过结合统计方法和词典来提高准确率. 目前,我们并没有利用词典信息,今后将会在方法上研究其它统计方法,例如利用条件随机场等模型进一步改进算法,提高准确率. 另一方面,考虑到蒙古语的独特特点,我们希望能够通过研究更好的蒙古语语言模型架构来改进蒙古语语言模型的计算,并以此提高蒙古文词切分的准确率.

参 考 文 献

[1] Nasanurtu. A Segmentation System of Mongolian Etyma, Stem and Affix. Journal of Inner Mongolia University: Humanities and Social Sciences, 1997, 29(2): 53 - 57 (in Chinese)
(那顺乌日图. 蒙古文词根、词干、词尾自动切分系统. 内蒙古大学学报:人文社会科学版, 1997, 29(2): 53 - 57)

[2] Hua Shabao. The POS Tagger System for Mongolian Corpus. Journal of Inner Mongolia University: Humanities and Social Sciences, 1999, 31(5): 33 - 37 (in Chinese)
(华沙宝. 对蒙古文语料库的词类标注系统——AYIMAG. 内蒙古大学学报: 人文社会科学版, 1999, 31(5): 33 - 37)

[3] Hou Hongxu, Liu Qun, Zhang Yujie, *et al*. Research and Implement of the 2005 HTRDP(863) Evaluation on Machine Translation. Journal of Chinese Information Processing, 2006, 20(Z1): 7 - 18 (in Chinese)
(侯宏旭,刘 群,张玉洁,等. 2005 年度 863 机器翻译评测方法研究与实施. 中文信息学报, 2006, 20(Z1): 7 - 18)

[4] Badma-Odsar. A Study of Part of Speech Classification of Mongolian Language. Journal of the Central University for Nationalities: Philosophy and Social Sciences Edition, 2004, 31(3): 94 - 100 (in Chinese)
(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, 31(3): 94 - 100)

[5] Nasanurtu. Semantic Research for the Mongolian Language to Be Oriented to Information Processing. Journal of Inner Mongolia University: Humanities and Social Sciences, 2002, 34(5): 43 - 48 (in Chinese)
(那顺乌日图. 关于面向信息处理的蒙古语语义研究. 内蒙古大学学报:人文社会科学版, 2002, 34(5): 43 - 48)

[6] Hua Shabao. The Technological Countermeasure to Deal with the Net Information in Mongolian. Minority Languages of China, 2002, 6: 58 - 60 (in Chinese)
(华沙宝. 蒙古文网络信息技术处理的对策. 民族语文, 2002, 6: 58 - 60)

[7] Hou Hongxu, Deng Dan, Zou Gang, *et al*. An EBMT System Based on Word Alignment // Proc of the 4th International Workshop of Spoken Language Translation. Trento, Italy, 2004: 47 - 49

[8] Zhang Huaping, Yu Hongkui, Xiong Deyi, *et al*. HHMM-Based Chinese Lexical Analyzer ICTCLAS // Proc of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan, 2003: 184 - 187

[9] Ye Jiaming. Research and Implement of Mongolian Lexical Analysis Based on Rules. Master Dissertation. Beijing, China: Peking University. School of Electronics Engineering and Computer Science, 2005 (in Chinese)
(叶嘉明. 基于规则的蒙古语词法分析研究与实现. 硕士学位论文. 北京:北京大学. 信息科学技术学院, 2005)

[10] Liu Qun, Zhan Weidong, Chang Baobao, *et al*. Computing Model and Language Model of Chinese-English Translation System // Proc of the 3rd Intelligent Interface and Intelligent Application. Zhangjiajie, China, 1997: 253 - 258 (in Chinese)
(刘 群,詹卫东,常宝宝,等. 一个汉英机器翻译系统的计算模型与语言模型//第 3 届全国智能接口与智能应用学术会议. 张家界, 1997: 253 - 258)

[11] Hou Hongxu, Liu Qun, Liu Zhiwen. Skip-N Mongolian Statistical Language Model. Journal of Inner Mongolia University: Natural Sciences, 2008, 39(2): 220 - 224 (in Chinese)
(侯宏旭,刘 群,刘志文. Skip-N 蒙古文统计语言模型. 内蒙古大学学报:自然科学版, 2008, 39(2): 220 - 224)

[12] Katz S M. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans on Acoustics, Speech and Signal Processing, 1987, 35(3): 400 - 401

[13] Och F J. Minimum Error Rate Training in Statistical Machine Translation // Proc of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan, 2003: 160 - 167