蒙古文词切分方法综述

1 词典和规则相结合的方法

--现代蒙古语语料库加工技术的新进展\_省略\_\_Tagging\_System\_那顺乌日图

2 基于规则的方法

3 基于词典的方法

--蒙古文词根\_词干\_词尾的自动切分系统\_那顺乌日图

11 基于词素的蒙汉神经机器翻译

构形附加成分的切分

分写附加成分的切分

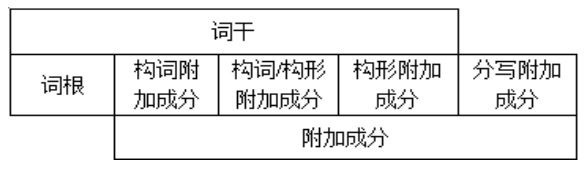
\_蒙古语语法信息词典构形附加成分分库\_的设计与实现\_淑琴

蒙古文词切分方法综述

1 概述

20世纪80年代开始，蒙古文信息处理技术兴起。蒙古文信息处理技术相比于中文信息处理起步较晚，但经过研究者们的努力，现如今它已经取得了一些成绩，被广泛的应用于蒙汉机器翻译,蒙古文语音识别和文字识别等多种领域中。蒙古文信息处理技术不仅对内蒙古自治区的发展起到了重要作用而且还对人们的日常生活带来了很大的便利。蒙古文词切分是蒙古文信息处理技术的基础以及关键步骤之一，词切分的准确性将直接影响蒙古文信息处理的实用性和可靠性。

相比于英文分词和中文分词，蒙古文词切分要复杂的多，因为蒙古文属于黏着语，其词的形态变化特别丰富。蒙古文的词一般可以分解为词干（或词根）+附加成分的形式。蒙古文附加成分从功能角度可以分为构词附加成分和构形附加成分。构词附加成分改变词汇意义，而构形附加成分不改变词汇意义，只增加语法意义。构形附加成分可以分为与词干连写的附加成分和与词干分写的附加成分。（论文：蒙古语语法信息词典构形附加成分分库\_的设计与实现\_淑琴）蒙古文词素（词干,词根,附加成分之间的关系）之间的关系如下表所示：

表1 蒙古文词素之间关系

2 蒙古文词切分方法

在蒙古文词切分方面，研究者们从刚开始的人工切分到基于词典的方法,基于规则的方法,基于统计的方法再到现如今基于神经网络的方法，这当中付出了很多的努力。本文将从蒙古文词切分的技术,方法,算法,系统等方面去进行研究。

2.1基于词典的方法

基于词典的方法是一种比较传统的方法。若想采用基于词典的方法去切分蒙古文，首先得去建立一个蒙古文词切分词典，然后通过去查词典的方法查看这个蒙古文词是不是在词典中，如果在就看这个词是由哪些词干和附加成分构成，最后将相对应的词切分结果输出来。

通过那顺乌日图（蒙古文词根\_词干\_词尾的自动切分系统\_那顺乌日图）在具有20000多词条的《蒙汉词典》和100万词级的《现代蒙古语文数据库》的语料上做的实验结果可以看得出，基于词典的蒙古文词切分方法不仅能覆盖在95%的蒙古文词中，而且准确率也达到了很高的水准。

基于词典的方法的优缺点如下：

优点：准确率较高，跟人工切分比速度快。

缺点：（1）因为很多词都有好几种切分方法，所以会导致存在二义性问题。

（2）蒙古文词复杂多变，词典方式不能给出所有词的切分形式，所以会有未登录词的情况。

（3）想切分准确率高，就得建立大规模词典，但如果词典规模过于大会导致人力物力无法支持。

2.2基于规则的方法

2.3词典与规则结合的方法

2.4 基于SKIP-N语言模型的统计方法

基于SKIP-N语言模型的蒙古文词切分方法是在统计自然语言处理的基础上提出来的方法。SKIP-N的语言模型任务就是根据当前的词去预测其上下文词语。该方法中构建了三种语言模型，分别是基于词的语言模型，基于词干/词缀的语言模型和基于字母的语言模型。该方法的基本步骤为，首先得把语料通过切分规则表切分成所有可能切分的形式（切分中不需要去管切分结果的准确与否只需要给出所有可能性的切分结果），然后通过SKIP-N语言模型以打分形式去评估哪些切分是对的，哪些是错误的，经过筛选有了候选的切分结果后，就可以再利用语言模型去进行评价，然后选出最终的切分结果。

基于SKIP-N语言模型的统计方法的优缺点如下：

优点：准确率高达94%，解决了词切分中存在的二义性问题。

缺点：由于是人工构建的切分规则集，对人工依赖还是很大。

文献[] (--基于统计语言模型的蒙古文词切分\_侯宏旭)

2.5 基于条件随机场的蒙古语词切分

--基于条件随机场的蒙古语词切分研究\_赵伟

2.6 基于短语统计机器翻译模型

--基于短语统计机器翻译模型蒙古文形态切分\_李文

2.7 基于判别式分类的词干词缀切分方法

--蒙古语有向图形态分析器的判别式词干词缀切分\_姜文斌

2.8 基于词典，规则和统计的蒙古文词切分系统

--基于词典\_规则与统计的蒙古文词切分系统的研究\_明玉

2.9 统计与规则相结合的词切分方法

--蒙古文词法分析及其在蒙汉统计机器翻译中的应用\_玉霞

2.10 基于神经网络的方法

--基于双向门限递归单元神经网络的维吾尔语形态切分\_哈里旦木\_阿布都克

3 蒙古文词切分主要问题

3.1 歧义切分问题

3.2 未登录词问题