文章编号：

基于 Bi-LSTM-CRF 模型的蒙古文形态素切分方法

吴都，徐金安，陈钰枫，张颖，陈圣爱，张玉洁

（北京交通大学 计算机与信息技术学院，北京 100044）

摘要：蒙古文形态素切分是蒙古文自然语言处理的核心任务之一。该文针对传统蒙古文的构词特点，提出了一种新的蒙古文形态素标注方法，在蒙古文天然的词边界划分基础上，进一步将形态素进行划分。相比传统的蒙古文词切分方法，本文重点研究构词成分的形态素单元切分，提出的方法在充分学习蒙古文词和字知识的同时，通过自动学习蒙古语构词的形态素成分上的语言学知识，能够更加有效地捕捉形态素单元上的语义信息。该文使用新标注方法并基于 Bi-LSTM-CRF 模型构建了蒙古文形态素切分系统，显著提高了蒙古文形态素切分的精度。

关键词：蒙古文；形态素切分；Bi-LSTM-CRF

中图分类号： **文献标识码：**

Mongolian Morphological Segmentation Method Based on Bi-LSTM-CRF Model

WU Du , XU Jinan , CHEN Yufeng , ZHANG Ying， CHEN Shengai , ZHANG Yujie

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044,China)

Abstract：Mongolian morphological segmentation is one of the core tasks in Mongolian natural language processing. Aiming at the characteristics of traditional Mongolian word formation features, this paper proposes a new Mongolian morphological annotation method, which further divides morphological elements based on Mongolian natural word boundary division. Compared with the traditional Mongolian word segmentation method, this paper focuses on the morphological element segmentation of word formation components. The proposed method is to learn the Mongolian word and word knowledge and capture semantic information on morpheme units more effectively by automatically learning the morphological components of Mongolian word formation and Linguistic knowledge. By using this new annotation method and the Bi-LSTM-CRF model, a Mongolian morphological segmentation system is constructed, which significantly improves the accuracy of the Mongolian morphological segmentation.

Key words: Mongolian；Morphological Segmentation；Bi-LSTM-CRF

0 引言

蒙古文属于黏着语系，蒙古文词由词根、词干和词缀构成。通常把汉语的单词分割称为汉语分词，蒙古文则需要做形态素切分。虽然蒙古文的词和词之间有天然的空格，但是由于蒙古文词的词根和词干后缀接了多种不同的词尾，从形态素粒度出发，需要对蒙古文中构词的成分，即形态素进行切分，识别出每个词的词根、词干和词

收稿收稿日期：；**定稿日期**：

基金项目：本文由国家自然科学基金（61370130,61473294，61876198），中央高校基本科研业务费专项资金(2015JBM033)，科学技术部国际科技合作计划（K11F100010）资助。

缀。形态素切分任务是蒙古文自然语言处理的基础任务。良好的形态素切分结果，能够极大地提高例如蒙中机器翻译，蒙古文阅读理解等下游任务的准确性。

蒙古文自然语言处理研究基础较为薄弱。目前已有的蒙古文词切分方法大多为传统切分方法：1997 年，那顺乌日图提出了一种基于规则的蒙古文词切分方法，实现了词根、词干、词尾的自动切分系统[1]；2005 年，那顺乌日图、雪艳等人提出了 DarhanTaggingSystem 系统，该系统使用基于词典和规则相结合的方法，使用词干、构形附加成分词典和切分规则库来完成蒙古文的词切分工作[2]；2009 年侯宏旭、刘群等人提出了一种基于规则和统计相结合的蒙古文词切分方法[3]，通过对蒙古文语言特点和切分技术的研究，采用规则作为词切分的基础，融入基于统计的语言模型，使蒙古文词切分的准确率达到 93.9%；2010 年赵伟、侯宏旭等人提出了基于条件随机场模型的蒙古文词切分系统[4]提高了词切分准确率。

综上所述，蒙古语分词发展至今，目前仍然存在许多问题亟待解决，诸如没有共享的开源标注语料，没有成熟的分词方法等问题，极大地制约了蒙古文自然语言处理的发展。

本文将蒙古文切分问题转化为序列标注问题，首先针对蒙古文语言特点，提出了一种融合蒙古文语言特点的六字标注方法，该方法能够有效地利用蒙古文本身的构词方法、形态学特征，准确地划分出蒙古文中的词根、词干和词缀。

以新的六字标注方法为立足点，本文使用 BI-LSTM-CRF 深度神经网络模型[5]进行训练，在解决了长距离依赖问题的基础上，综合考虑文本双向的记忆信息，最终获取到全局最优的输出序列。本文通过针对性的实验，验证了这一方法的可行性，形态素切分准确率达到了 96.86%,F1 值达到了 97.08%。

本文其余部分安排如下：第一节介绍传统蒙古文形态素切分方法；第二节阐述本文提出的融合蒙古文语言特点的六字标注方法；第三节介绍基于 BI-LSTM-CRF 模型的蒙古文形态素切分方法；第四节介绍本文对蒙古语语料预处理部分的工作；第五节给出实验数据并进行结果分析；第六节对全文进行总结。

1 传统形态素切分方法

传统的蒙古文形态素切分方法主要包括基于词典的切分方法、基于规则的切分方法、基于词典和规则相结合的切分方法、基于统计的切分方法等。

在蒙古文形态素切分的任务中，基于词典的切分方法是将已经正确切分的蒙古文词统一存放到词典中，通过查找词典的方式得到一个词是由哪些词根、词干、词缀组合而成的。由于词典里列举出了可能出现的所有切分形式，使用这种方法会快速地查到切分结果，能够获得比较高的准确率。这一方法的缺点非常明显，在当今这个数据爆炸的时代，仅仅使用词典无法穷举所有词以及其变化形式。无法解决未登录词以及词的二义性问题。

基于规则的方法就是利用蒙古文词的构成规律进行规则制定，但是，由于蒙古文自身的语言特色，构形规则十分复杂，不仅无法将规则完善编制，而且有一些词在进行切分的时候会发生词形的变化。因此规则的方法经常会导致切分错误。

基于词典和规则相结合的方法结合了基于词典和基于规则两种方法。首先需要构建蒙古文词的词干词典、构形附加成分词典以及一个包含蒙古文词切分规则的切分规则库。在词切分的时候，从前向后在词干词典中进行匹配，匹配出词干部分之后将词的其他部分利用切分规则库将各构形附加成分进行还原，并将其与词典中的附加成分进行比较，如果匹配正确则将其作为正确的结果输出。此方法需要大量的人力资源以人工构建词典库以及规则库，目前已基本不被采用。

基于统计的切分方法主要是利用条件随机场模型将其转化为序列标注任务。条件随机场模型对已标注的数据进行充分学习后能比较精确地输出测试数据相对应的标注序列。该方法节省人力，精确度也比较高，但是该方法无法学习句子中的上下文信息，极大地限制了切分精度的提升。

以上传统方法存在过度依赖于人工构建词

典库、规则库，无法自动学习上下文信息等缺点，本文将进一步考虑训练数据的上下文信息，以进一步提升形态素切分的准确度。

2 蒙古文语言特点的数据标注方式

蒙古文的形态素切分任务即对一个蒙古文词的词根、词干和词缀进行切分，本文将其转化为序列标注任务。对于序列标注任务来说，选择一种合适的数据标注方式有助于对该任务的研究。中文分词一般采用 BMES 四字标注法。因为不同语言存在完全不同的语法特点和标注规则，BMES 标记集并不适用于蒙古文。本文针对蒙古文的语言特色及构词特点，提出了一种融合蒙古文语言特点的六字标注方法，即“B、M、E1、E2、E3、S”，详细内容如表 1 所示。

表 1 融合蒙古文语言特点的六字标注法

标记名称	含义
B	词的首字符
M	词的中间字符
E1	词的末尾字符，后面出现连写附加成分，代表“+”
E2	词的末尾字符，后面出现分写附加成分，代表“-”
E3	词的末尾字符，后面出现一个新的词，代表空格
S	单个字符成词

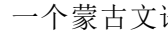
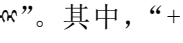
例如，一个蒙古文语句为：“  ”，中文意思为“不停地”。其对应的切分结果为：“  ”。其中，“+”连接的是连写附加成分，“-”连接的是分写附加成分，采用六字标注法进行标注的结果如表 2 所示。其中，第一列为将原始的蒙古文语句转化成单个蒙古文字符的序列，第二列为每一个蒙古文字符所对应的“B、M、E1、E2、E3、S”六字标签序列。

表 2 标签举例

字符序列	标签序列
	B
	M
	M
	M
	E1
	B
	M
	E2
	B
	M
	M
	E3

此种标注方式的优点在于可以很好地对蒙古文的词根、词干和词缀进行划分，且可以通过不同形式的结尾标签“E1、E2、E3”更好地区分蒙古文中的连写附加成分、分写附加成分以及词边界。

3 基于 Bi-LSTM-CRF 模型的蒙古文形态素切分

为了提升蒙古文形态素切分的准确率，本文将使用六字标注法标注的数据作为输入，使用目前主流的深度学习框架 Bi-LSTM-CRF 模型进行训练，该框架通过双向 LSTM 网络[6]提取蒙古文语言的内在特征，并使用 CRF 网络来预测序列标签。

3.1 基于 Bi-LSTM 模型的蒙古文形态素切分

首先介绍融合蒙古文六字标注方法的 Bi-LSTM 网络。网络结构图如图 1 所示：

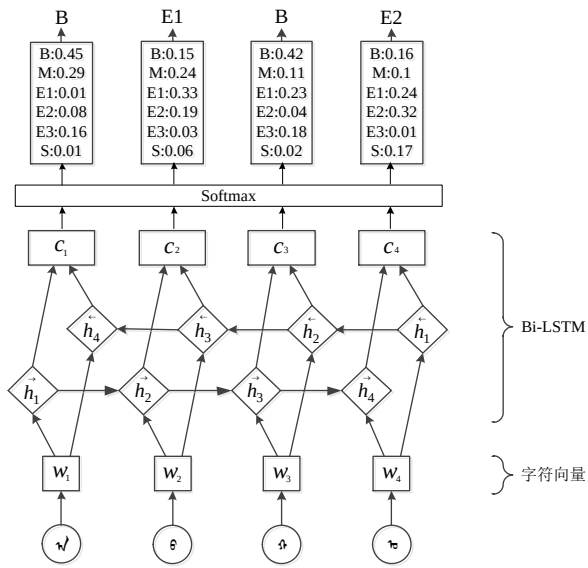


图 1 Bi-LSTM 网络结构图

与中文的字嵌入[7] (word embedding)不同，本文将每一个蒙古文语句转化成单个蒙古文字符的形式来表示蒙古文的字符嵌入。将一个含有

n 个蒙古文字符的序列表示为式(1)：

$$x=(x_1,x_2,...,x_n) \tag{1}$$

模型首先利用随机初始化的蒙古文字符向量矩阵，将一句话中的每一个蒙古文字符由 one-hot 向量映射成稠密的低维字符向量。由于使用神经网络训练，在训练数据较少的情况下容易造成过拟合，因此在进入 Bi-LSTM 网络之前使用 dropout[8][9] 缓解过拟合问题。加入了 dropout 的神经网络在数据训练的过程中，迫使每个隐藏单元必须学会随机选择其他单元样本，这种方法可以使每个隐藏的单元更加健壮，并驱使隐藏单元本身学到有用的特征。假设 dropout 为 0.6，每个神经单元都有 60%的几率暂时被移除，从而达到神经元单元之间不相互依赖的作用。

接下来，将蒙古文的字符嵌入序列信息输入到 Bi-LSTM 中，使用神经网络来自动提取出特征。双向 LSTM 包括一个前向 RNN 和一个后向 RNN。相较于传统的 LSTM 网络[10]，双向 LSTM 网络同时考虑了当前蒙古文字符的前向特征和后向特征。其中，后向特征的提取过程就是将原始的蒙古文字符序列逆向输入到 LSTM 网络中。

假设 $(\overrightarrow{h_1},\overrightarrow{h_2},\overrightarrow{h_3},...,\overrightarrow{h_n})$ 为正向 LSTM 网络输出的隐藏状态序列，$(\overleftarrow{h_1},\overleftarrow{h_2},\overleftarrow{h_3},...,\overleftarrow{h_n})$ 为反向的 LSTM 网络输出的隐藏状态序列，将两者拼接后得到最后的隐藏状态序列，即作为 Bi-LSTM 网络的输出，如式(2)所示：

$$(h_1,h_2,h_3,...,h_n)\in R^{n\times m} \tag{2}$$

在 LSTM 后连接一个 softmax 线性层，将隐藏状态的向量映射成 n 维，将标记集合的标签数设为 k，得到自动提取蒙古文语句的特征，如式(3)所示：

$$P=(p_1,p_2,p_3,...,p_n)\in R^{n\times k} \tag{3}$$

最后得到每个蒙古文字符所对应标签的得分，分值越大表明对应于此标签的概率越大，从而输出每一个蒙古文字符所对应的得分最高的标签。

3.2 基于 BI-LSTM-CRF 模型的蒙古文形态素切分

以上基于 Bi-LSTM 的蒙古文形态素切分方式存在一定的局限性：不能保证整体结果最优，只能保证局部结果最优；打标签的过程是彼此独立的分类，上文预测出来的标签不能直接被利用，只能通过隐藏层状态来获取。因此，可能导致预测出来的标签是不合理或不合法的，如：标签“B”后面接的是标签“S”或者标签“E3”后面接的是标签“M”、“E1”、“E2”而不是标签“B”、“S”。

因此，为了获取全局最优序列，在 Bi-LSTM 网络的基础上以 CRF 网络层[11]替代 softmax 层获得概率输出，将 Bi-LSTM 模型的输出作为 CRF 层的输入。在对当前的蒙古文字符进行标注时，能够利用前面已经标注过的标签信息，考虑整体序列，对全部的标签序列进行打分，从而得到概率最大的全局最优序列。增加 CRF 层后的网络结构如图 2 所示

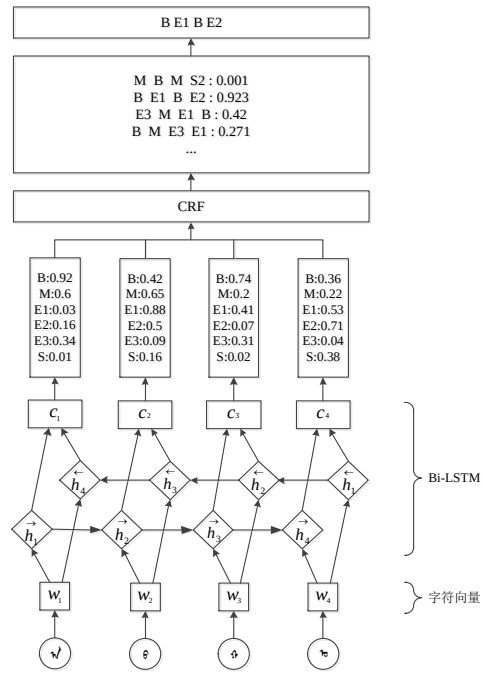


图 2 Bi-LSTM-CRF 网络结构图

CRF 层在训练的时候采用梯度下降的方法学习到最优的参数，如式（4）所示：

$$\theta^*=\arg\max\sum_i\log(P(x^i\mid y^i;\theta)) \tag{4}$$

使用维特比算法预测出最优路径，如式(5)所示：

$$y^* = \arg\max score(x, y^*) \quad (5)$$

CRF 网络层能够对整个序列进行考虑，对全部标签序列打分，并将得分最高的标签序列作为最后的结果输出。

此外，CRF 网络还能够添加一些约束来保证预测出来的标签是合理、合法的，考虑到标签之间是否存在一定程度的依赖关系，例如：

(1) 语句中的第一个字符的标签一定以标签“B”或“S”开始。

(2) 标签序列“BE1E2”是合法序列，而“BE2E1”则是非法序列。因为连写附加成分后面可以出现分写附加成分，而分写附加成分后面不可以直接出现连写附加成分。

以上约束可以降低标签序列中非法序列出现的概率，从而提高蒙古文形态素切分任务的准确率。

4 蒙古文语料预处理

语料处理的好坏程度会直接对后续任务造成影响。因此，本文针对蒙古文自身的语言特点，进行了特殊的数据前处理和后处理。

4.1 数据前处理

4.1.1 蒙古文特殊控制符

蒙古文和中文不同，除了代表词边界的天然空格之外还存在蒙古文特殊控制符，主要包括窄宽度无间断空格、元音间隔符两种。在统一采用的 UTF-8 编码格式下，若删除上述特殊控制符，词的写法将发生一定程度的变化。经统计，在本文所使用的传统蒙古文语料中窄宽度无间断空格共出现 69103 次，元音间隔符共出现 21423 次。由于特殊控制符在语料中出现的频率很高，因此需要提前做特殊的预处理。

窄宽度无间断空格表示为“ᠵ”、“ᠪ”等格附加成分开头部分的特殊空隙。若不进行任何处理，在模型训练或标签还原的时候可能会和正常的词边界空格混淆。因此，统一将此特殊控制符替换为“*”。本文对包含窄宽度无间断空格的格附加成分进行归纳总结，如表 3 所示。其中第一列为含有该控制符的格附加成分名称，第二列为对应的蒙古文格附加成分。

表 3 包含蒙古文窄宽度无间断空格的格附加成分

控制符名称	蒙古文表示
定格	ᠵ/ᠵ/ᠪ
向位格	ᠳᠤ
宾格	ᠶᠢ
凭借格	ᠪᠠᠷ
名词领属	ᠤᠨ
带有领属附加成分的变格	ᠲᠤᠶ
带有领属附加成分的变格	ᠳᠤᠨ

元音间隔符的通常表现形式为“᠊”前面的空隙。由于其固定出现在“᠊”前面，且一般在词的中间出现，因此本文没有对其做额外的处理。

4.1.2 一元模型对词缀词频排序并反切原始语料

蒙古文词的形态变化十分丰富，通过在词根、词干后面连接不同的词缀来实现。因此，词缀在蒙古文的形态素解析中十分重要。本文将转换之后的传统蒙古文语料中的词缀提取到一个词缀词典中，通过一元模型对其进行由高到低的词频排序。经统计，出现频率最高的词缀频次达到了 15125 次。本文将出现频率达到 1000 次以上的词缀默认为是必须进行切分的词缀，通过运用反向最大匹配法，利用出现频率高于 1000 次的词缀重新反切原始语料，进一步完善语料中切分不规范的语句并进行人工校正。

4.1.2 人工校正语料

蒙古文的一部分词干或词缀出现在不同的词中可能代表完全不同的语法意义，从而会产生不同的切分规则。因此，这种人工编制的语料中不可避免地会存在错误的情况。例如，在语料中“ᠮᠣᠩᠭᠣᠯ”这个词的切分结果为“ᠮᠣᠩᠭ+ᠣ+ᠯ”。蒙古文的语言学者很容易地就会判断出该词的切分是错误的，应该修正为“ᠮᠣᠩᠭᠣᠯ+ᠤᠨ+ᠠᠷ”。诸如此种错误如果在前处理部分不做任何修正操作，很可能会在模型训练及测试解码部分产生错误传递，影响后续的实验结果。因此，本文对语料进行了一定程度上的人工校正。

4.2 数据后处理

4.2.1 词边界恢复

蒙古文和英文类似，存在固定的词边界，词与词之间用天然的空格进行断开。根据本文 4.2.1 节提出的六字标注方式可知，标签“E3”和标签“S”代表词的边界。其中，“E3”对应的字符为一个蒙古文词中的最后一个字符，且此字符后应出现空格来对词边界进行识别；“S”对应的是单个字符成词，此字符后面也应出现空格符。但是，在进行蒙古文形态素切分实验的时候，我们发现存在词边界识别错误的现象，例如：原本应该标注“E3”标签，却被错误地标注成了“E1”、“E2”或其他标签。因此，本文在模型测试解码之后，对词边界进行恢复：如果原始标签是“E3”或“S”，则无论模型预测出来的是什么标签，在根据标签进行还原的时候均将其修正为对应的正确标签，从而确保词边界是正确的。

例如，蒙古文语句为：“  ”。

上述语句对应的形态素切分的六字标签如表 4 所示。其中第一列为将蒙古文语句转换成字符的序列、第二列为正确的标签序列、第三列为形态素切分系统可能预测出来的标签序列、第四列为进行词边界还原后的标签序列。

表 4 词边界恢复举例

字符	正确标签	预测标签	还原后的标签
	B	B	B
	M	M	M
	E3	E2	E3
	B	B	B
	M	M	M
	M	M	M
	M	M	M
	E3	E3	E3
··	S	S	S

4.2.2 命名实体恢复

一些人名、地名和组织机构名等命名实体不遵循蒙古文的元音和谐律现象规则[12]。在进行形态素切分的时候，应将其作为一个整体不进行任何切分，因此我们将此类违反元音和谐律的专有名词抽取到命名实体的词典中。在进行形态素切分和词性标注实验的时候，需要进行后处理操作，即对原句中的命名实体进行识别，将误被切开的部分进行恢复，从而提高形态素解析的准确性。

例如，人名：“  ”，译为“雷锋”。

经过形态素切分后的结果可能为：“”。

通过命名实体恢复的后处理将其修正为：“  ”。

4.2.3 词性词典还原

为了提高蒙古文词性标注结果的准确率，本文对字符级别的词性标注和形态素级别的词性标注的实验结果进行词性词典还原的后处理。由于通过模型解码得到的词性标注结果可能是错误的，因此本文从训练语料中构建了一个词性词典，包含每一个蒙古文形态素及其对应的词性信息。词性标注模型解码之后，采用最大字符串匹配的方式去词性词典中查找并修正标注错误的词性，提高词性标注模型的效果。

例如，蒙古文语句：“  ”。

正确的词性标注结果：“  /Ne1  /Ve2+/Fn1/Ne1/Ve2+  /Fs14 ··/Wp1”。

模型得到的标注结果：“/Ne1/Ve2+/Fn3  /Ne1/Nt2+/Fs14 ··/Wp1”。

词性词典还原修正后：“/Ne1/Ve2+/Fn1/Ne1  /Ve2+/Fs14 ··/Wp1”。

5 实验

5.1 实验语料

蒙古文形态素解析实验使用的原始数据是内蒙古大学的那顺乌日图教授和吴金星博士提供的拉丁新蒙文语料，根据传统蒙古文与拉丁新蒙文的转换对照表对其进行转换后得到本文使用的传统蒙古文语料。由于两种语料之间存在非常大的差异性，转换后的结果无法避免地存在一些错误情况，因而对其进行人工修正。最后，得

到了 13755 条语句的蒙古文数据，共包含 446153 个词语。从语料中剔除掉词性标签，得到完整的蒙古文形态素切分语料。

在基于条件随机场的切分实验中，随机抽取 1375 条语句作为实验测试集，其余的 12380 条语句作为训练集。在基于神经网络的切分实验中，随机抽取 877 条语句作为实验的测试集，抽取 878 条语句作为开发集，其余的作为训练集。详细信息如表 5 和表 6 所示：

表 5 基于条件随机场的形态素切分数据集

数据集	句子数	词语数
训练集	12380	401326
测试集	1375	44827

表 6 基于神经网络的形态素切分数据集

数据集	句子数	词语数
训练集	12000	389018
开发集	878	28632
测试集	877	28503

此外，本文使用第十四届全国机器翻译研讨会(CWMT2018)提供的数据集，验证了本文方法对于蒙中机器翻译任务的性能提升的有效性。本文通过对蒙中双语平行语料进行对齐、全半角转换、筛选乱码和特殊字符等预处理操作来进行数据修正，得到最终的实验数据，详细信息如表 7 所示：

表 7 机器翻译数据集

数据集	语言	句对数	语料库词表
训练集	蒙文	258154	150715
	中文	258154	106578
开发集	蒙文	3000	10982
	中文	3000	10660
测试集	蒙文	3000	11305
	中文	3000	11052

5.2 实验设置和评测指标

本文进行蒙古文形态素切分使用的实验设置是根据经验值选定的。经统计，本文所使用的标注数据中语句长度均小于 150，因而将句子的最大长度设置为 150。神经网络的超参设置如下：双向 LSTM 的隐藏层单元个数设为 100；蒙古文字符向量维数为 100；学习率为 0.001，优化方法采用 Adam，这种方法可以自动调整学习率从而找到最小极值点[13]；batch 的大小设置为 128；dropout 设置为 0.5；迭代轮数定为 45 000 轮。

蒙古文语料的编码格式均采用 UTF-8 编码。训练过程在 GPU 服务器上运行。

本文使用的评测指标为分词任务常用的 P(准确率)、R（召回率）、F1 值。此外，由于使用机器翻译效果验证分词的有效性，本文将 BLEU 值与 NIST 值作为一项重要评价指标。

5.3 实验结果与分析

本文将赵伟等人提出的蒙古文数据标注方法“S、A、B、I、E、S”标注法[4]作为对照组，并使用上文提到的表 5 和表 6 数据集进行形态素切分实验。实验结果如下表，表 8 左侧所示。

然后，采用本文第 2 节提出的六字数据标注方法以及第 4 节所述的数据处理方法在相同的数据集上进行形态素切分实验。将第 4 节所述的蒙古文语料的数据前处理(Mongolian data pre-processing)记为“MDP”、词边界恢复(Word boundary recovery)记为“WBR”。实验结果如表 8 右侧所示。

总的来看，使用前人方法得到的最优 F1 值为 88.58%，而使用本文方法所得到的 F1 值达到了 97.08%，增长了近 8.5 个点。

另外，纵向对比本文方法，使用基于 Bi-LSTM-CRF 模型得到的 F1 值比基于 CRF 模型的 F1 值高了 4.22 个点。在 Bi-LSTM-CRF 的对比试验中，相较于单纯地进行本文 4.1 节所述的数据前处理，在此基础上增加本文 4.2 节所述的词边界恢复后 F1 值增长了 1.07 个点。

横向对比前人方法与本文方法可以观察到：在使用相同 CRF 模型的情况下，本文提出的方法比 SABIES 标注方法 PRF 值提高了 2%左右；在 Bi-LSTM-CRF 模型中，本文方法相较 SABIES 方法提高了 30%，由此可见本文提出的方法结合 Bi-LSTM-CRF 模型能更好地提高形态素切分的精度。由表 1 本文的标注方法可以看出，本次使用的六字标注法着重区分了蒙古文词尾缀接不同附加成分的三种情况—连写附加成分、分写附加成分以及空格。在准确对这些关键附加成分区别

表 8 蒙古文形态素切分实验结果

模型	SABIES 方法				本文方法				
	标注方式	准确率	召回率	F1 值	标注方式	语料处理	准确率	召回率	F1 值
CRF++	SABIES	87.96%	89.20%	88.58%	BME1E2E3S	MDP	91.81%	91.76%	91.79%
Bi-LSTM-CNN	SABIES	60.17%	62.86%	61.49%	BME1E2E3S	MDP	95.16%	95.72%	95.44%
Bi-LSTM-CRF	SABIES	62.10%	63.95%	63.01%	BME1E2E3S	MDP	96.07%	95.96%	96.01%
Bi-LSTM-CRF	SABIES	62.10%	63.95%	63.01%	BME1E2E3S	MDP+WBR	**96.86%**	**97.31%**	**97.08%**

标注的情况下，Bi-LSTM-CRF 模型能更准确地获得训练数据的上下文信息以及词干主要信息，实验效果的提升验证了本文方法的有效性。

此外，本文分别将 BPE、CRF 与 BPE 结合、Bi-LSTM-CRF 与 BPE 结合作为蒙中机器翻译的基础对蒙古文进行形态素切分，中文均采用 Jieba 和 BPE 相结合的方式来进行切分，统一使用 Transformer 翻译模型框架进行蒙中机器翻译的对比实验。其中，“SABIES”为前人提出的蒙古文数据标注方法；将本文提出的方法记为“Ours”，即使用第 2 节所述的“B、M、E1、E2、E3、S”六字标注方法和第 4 节所述的蒙古文数据处理方法。二者使用的模型框架相同但对数据的标注方法和语料的前后处理方法不同。对比实验结果如表 9 所示：

表 9 蒙中机器翻译结果

不同翻译系统	BLEU		NIST	
	SABIES	**Ours**	SABIES	**Ours**
BPE + Transformer	0.39	**0.39**	8.46	**8.46**
CRF + BPE + Transformer	0.37	**0.40**	8.33	**8.69**
Bi-LSTM-CRF + BPE + Transformer	0.28	**0.42**	7.69	**8.87**

纵向来看，相较于单纯使用 BPE 子词切分，使用本文提出的六字标注方式的 CRF 形态素切分模型与 BPE 结合后的翻译结果有了近 0.015 个 BLEU 值的提升，NIST 值增长了 0.233 个点。将本文 4.2.2 所述的 Bi-LSTM-CRF 形态素切分模型与 BPE 结合后的翻译结果又增长了 0.022 个 BLEU 值，NIST 值增长了近 0.195 个点，比单纯使用 BPE 算法进行子词切分的翻译结果有了 0.037 个 BLEU 值的提升，NIST 值增长了近 0.428 个点。横向来看，运用本文提出的方法进行蒙中机器翻译后得到的 BLEU 值和 NIST 值均高于前人方法，证明了本文提出的六字标注法以及蒙古文语料预处理的工作对于蒙中机器翻译任务的有效性。

6 总结与展望

本文根据蒙古文本身的语言学特点构造了新的六字标注法，经实验验证，该方法极大地提升了蒙古文形态素切分任务准确度。并且我们将 BI-LSTM-CRF 模型运用到蒙古文形态素切分任务中，充分考虑了蒙古文词的上下文信息，构建了完整的形态素切分模型。最后，我们还使用目前效果最好的开源 Transformer 神经机器翻译系统，验证了本文提出的蒙古文形态素切分方法的有效性。

针对目前蒙古文标注语料极度稀缺的现状，我们下一步将本文提出的标注方法应用于蒙古文形态素词性标注任务，并进一步研究蒙古文的命名实体识别和未登录词识别等相关问题。

参考文献

[1] 那顺乌日图.蒙古文词根、词干、词尾的自动切分系统[J]. 内蒙古大学学报(人文社会科学版),1997(2):53-57.

[2] 那顺乌日图,雪艳,叶嘉明.现代蒙古语语料库加工技术的新进展—新一代蒙古语词语自动切分与标注系统(Darhan Tagging System)[C]//第十届全国少数民族语言文字信息处理学术研讨会, 2005:122-127.

[3] 侯宏旭,刘群,那顺乌日图等.基于统计语言模型的蒙古文词切分[J]. 模式识别与人工智能,2009,22(1):108-112.

[4] 赵伟,侯宏旭,丛伟等.基于条件随机场的蒙古语词切分研究[J].中文信息学报,2010,24(5):31-36.

[5] Huang Z, Xu W, Yu K, et al. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv: Computation and Language, 2015.

[6] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for Chinese word segmentation[C]//Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing. 2015:1197-1206.

[7] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31th International Conference on Machine

Learning. 2014:1188-1196.
[8] Srivastava N. Improving neural networks with dropout[J]. University of Toronto, 2013, 182:566.
[9] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[10] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[11] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[12] 清格尔泰.蒙古语语法[M].呼和浩特:内蒙古人民出版社,1991.
[13] Kingma D P, Ba J. Adam:A Method for Stochastic Optimization[J]. Computer Science, 2014.

作者联系方式：
姓名：吴都　地址：北京市海淀区交大东路北京交通大学 邮编：100089　电话：13858152783
电子邮箱：wd1712@bjtu.edu.cn