文章编号：1003-0077（2011）00-0000-00

子字粒度切分在蒙汉神经机器翻译中的应用

任众[1]，侯宏旭[1]，吉亚图[1]，武子玉[1]，白天罡[1]，雷颖[1]

（1.内蒙古大学 计算机学院，内蒙古自治区 呼和浩特，010021）

摘要：在蒙汉神经机器翻译任务中，由于语料稀少使得数据稀疏问题严重，极大影响了模型的翻译效果。本文对子字粒度切分技术在蒙汉神经机器翻译模型中的应用进行了研究。通过 BPE 算法将切分粒度控制在字符和词之间的子字粒度大小，将低频词切分成相对高频的子字片段，来缓解数据稀疏问题，从而在有限的数据和硬件资源条件下，更高效的提升模型的鲁棒性。实验表明，在两种网络模型中使用子字粒度切分技术，BLEU 值分别提升了 4.81 和 2.96，且随着语料的扩大，训练周期缩短效果也越显著。实验证明了，子字粒度切分技术有助于提高蒙汉神经机器翻译效果。

关键词：蒙汉神经机器翻译；数据稀疏；子字粒度切分

中图分类号：TP391　　**文献标识码：**A

Application of Sub-word Segmentation in Mongolian-Chinese Neural Machine Translation

Zhong Ren[1], Hongxu Hou[1] ,Yatu Ji[1],Ziyu Wu[1],Tiangang Bai[1],Ying Lei[1]

(1.Inner Mongolia University, Hohhot, 010021,China)

Abstract: In the Mongolian-Chinese neural machine translation task, the data sparse problem is serious, which affects the BLEU of the translation model. This paper studies the application of sub-word granularity segmentation in the Mongolian-Chinese neural machine translation model. The Byte Pair Encoding algorithm controls the granularity of the sub-word between the characters and the words and reduces the low-frequency words into relatively high-frequency sub-units to alleviate the data sparseness problem. Thus, under the condition of limited data and hardware resources, improve the robustness of the model more efficiently. Experiments show that using the sub-word granularity segmentation technique in the two models, the BLEU values are increased by 4.81 and 2.96 respectively, and with the corpus is expanded, the effect of shortening the training time is more significant. Experiments have proved that the sub-word granularity segmentation technique can improve the effect of the Mongolian-Chinese neural machine translation.

Key words: Mongolian-Chinese neural machine translation; data sparseness; sub-word segmentation

1 引言

蒙古语是我国少数民族蒙古族的语言，也是蒙古国的官方语言，所以蒙汉机器翻译对于民族团结以及中蒙交流都有着重要意义和价值。近年来，统计机器翻译[1]的发展进入瓶颈期，深度学习[2]成为研究热潮，神经网络机器翻译成为机器翻译研究的重要方向。

蒙古语属于黏着语，其构词规则如图 1 所示，一个蒙古文词由一个蒙古文词根与多个词缀组成。蒙汉机器翻译存在双语对齐语料不足、资源稀少等问题。这使得数据稀疏问题更加凸显。因此本文想要通过对蒙汉双语使用子字粒度的切分，来缓解数据稀疏问题，以提升翻译质量。与此同时，子字粒度切分缩小了双语词典的大小，降低了神经网络模型的计算复杂度，可以大大减少训练周期。

* **收稿日期：**　　**定稿日期：**

基金项目：内蒙古自然科学基金面上项目（2018MS06005）；内蒙古蒙古语言文字信息化专项扶持项目（MW-2018-MGYWXXH-302）；内蒙古自治区研究生科研创新项目（10000-16010109-18）

作者简介：任众（1994——），男，硕士生，自然语言处理；侯宏旭（1972——），男，教授，自然语言处理、信息检索；吉亚图（1990——），男，博士，自然语言处理。

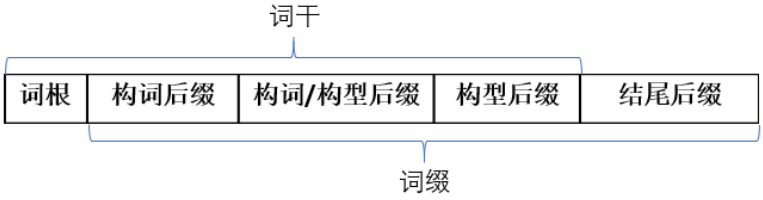


图 1 蒙古文词汇构成

本文使用 CWMT2009[3]和 CWMT2017 蒙汉双语语料，并在基于循环神经网络（Recurrent Neural Network，RNN）和卷积神经网络（Convolution Neural Network，CNN)的蒙汉神经机器翻译系统上进行了子字粒度切分的多方面对比实验。

2 蒙汉神经翻译模型

本节简要介绍本文所使用的神经翻译模型。

2.1 RNN 模型概述

在本文中，RNN 是在基于注意力机制的 RNN 模型[4]上重现的神经网络的编码、解码翻译模型[5]。模型的编码器和解码器是双向的 RNN[6]，其中加入注意力机制的解码器状态更新公式如（1）-（5）所示：首先用解码器 *i*-1 时刻的隐层状态z_{i-1}以及相应编码器的隐层状态h_j，计算得到对齐权重a_{ij}（1）。然后将编码器各隐层状态与对齐权重a_{ij}相乘，并求和得到编码器输出的摘要向量c_i（2）。然后由 i-1 时刻解码器输入的词w_{i-1}，以及 *i*-1 时刻隐层状态z_{i-1}和向量c_i计算更新 i 时刻解码器隐层状态 z_i（3）。每次更新状态，均计算目标词典中的每一个词w_k^T在当前解码器的隐层状态 z_i下的分数(4)。最后利用多分类算法 Softmax[7]将 z_i转换为最终的目标词翻译的概率分布（5）。

$$a_{ij} = \alpha(z_{i-1}, h_j) \tag{1}$$

$$c_i = \sum_{j=1}^{T} a_{ij} h_j \tag{2}$$

$$z_i = \emptyset_\theta(w_{i-1}, z_{i-1}, c_i) \tag{3}$$

$$e(k) = w_k^T z_i + b_k \tag{4}$$

$$\mathrm{p}(w_i = k | w, c_i) = \frac{\exp(e(k))}{\sum_t \exp(e(t))} \tag{5}$$

2.2 CNN 模型概述

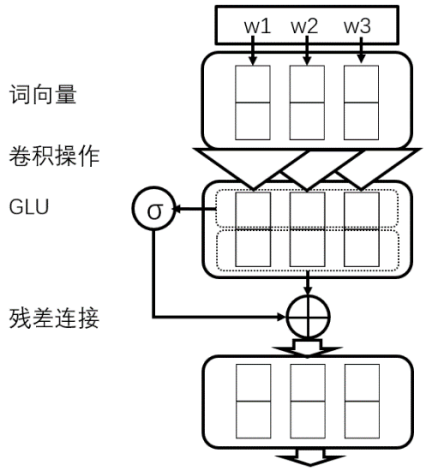


图 2 CNN 编码器结构图

在本文中，CNN 系统采用 Gehring 等提出的架构[8][9]的开源 CNN 模型。该模型依然是由编码器、解码器和注意力机制三个部分组成，如图 2 所示，其中编码器和解码器由多层卷积层组成，且每个卷积层之间加入线性门控单元（GLU）[10]和一个残差连接[11]。该模型利用 CNN 卷积核获取局部信息，并通过增加 CNN 卷积层数来获取长距离依赖信息，因此编码器和解码器均为多层的深层 CNN，每层解码器配备一个注意力机制。

3 子字粒度切分

在机器翻译任务中，对语料的切分是语料预处理过程中非常重要的一步。句子所包含的特征，由多个局部特征共同组成。而在语料预处理阶段，句子切分粒度越大，切分结果越能够保存更完整的局部特征，但是加重了数据稀疏问题；切分粒度越小，其包含局部特征越少，但是数据稀疏问题会得到缓解。尤其在双语资源相对匮乏的蒙汉机器翻译任务当中，切分粒度的把控尤为重要。

子字粒度切分是将句子切分成介于词和字之间的粒度大小，这样的做法可以在一定程度上保留局部特征，同时尽可能减小粒度，从而缓解数据稀疏问题，提高翻译效果，同时缩小词典规模，减少训练周期。

3.1 Byte Pair Encoding（BPE）

Sennrich 和 Haddow[12]在 2016 年提出了一种使用 BPE 编码[13]来处理文本切分粒度。

BPE 算法的基本思路：首先将语料以最小单元（蒙古文是指蒙古文字母，汉文是指汉字）切分；然后统计语料中所有相邻最小单元组合出现的频数；再找出频数最大的组合；将最大组合合并然后替换语料库中原来组合；循环上面操作。

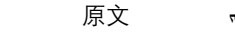
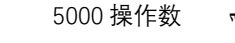
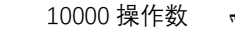
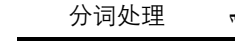
传统的中文分词是利用一个第三方词典切分句子，而 BPE 算法切分子字，是通过统计自身语料中词频，获得自身的词典，再根据词典切分句子。相较于传统分词处理，首先切分使用的词典来源于自身训练集，因此在对测试集的切分粒度也会与训练集保持一致，从而间接减少集外词的数量；更重要的是，BPE 允许出现多粒度切分，即词频高的词会被切分成词，词频低的词则会被切分成字，这样可以在保留一部分局部信息的前提下，缩小词典大小，间接缓解了数据稀疏问题。

BPE 算法有两种应用方法：一是独立 BPE，即构建两个独立词典，一个源语言词典，一个目标语言词典；二是联合 BPE，两个语言放一块，共同生成一个词典。理论上后者效果好一点，可以保证源语言和目标语言分割的一致性。尤其表现在拥有共享字母表的两种语言以及同源词和外来词上。

3.2 蒙汉子字切分

蒙汉机器翻译模型非常适合使用子字切分来提升翻译效果。在子字切分时，我们往往需要在目标端语言中加入标记符号，用来还原词，但是汉文中字、词、短语的界限非常模糊，且最小单位字也能作为独立的切分模块，因此省去了标记符号，降低了噪声；蒙古文是一种纯粹的拼音文字[14]，它是由蒙古文字母组成蒙古文词，再组成蒙古文句子，因此子字粒度可以大大缓解数据稀疏，从而提升整体的翻译效果。

表 1 使用子字粒度切分蒙汉双语对比

	蒙古文	汉文
原文		得到有关单位的允许
5000 操作数		得到 有关 单 位 的 允 许
10000 操作数		得到 有关单位 的 允 许
分词处理		得到 有关 单位 的 允许

如表 1 所示，因为其蒙古文字母种类相对较少，因此在两种切分粒度下（操作数指相邻最小单元组合合并次数，可以间接影响切分粒度，操作数越大，粒度相对越大）与原文比较变化不明显；汉文可以明显看到传统分词与 BPE 切分子字结果不同。子字粒度切分使得高频词得以保留，低频词切分为更小粒度的蒙文字母或高频字母组合，缓解了数据稀疏问题。

同样，如表 2 所示，我们在 CWMT2009 提供的 6 万句对的蒙汉双语语料和 CWMT2017 蒙汉评测提供的 26 万句对的双语语料上分别统计了不同切分粒度的词典大小。蒙古文切词处理，汉文分别切字（*_char）和切词（*_word）处理对比使用 BPE 算法进行蒙汉子字切分（*_BPE）处理，可以明显看出蒙汉两端词典大小均有明显减小；同时，当我们使用较大训练语料时，词典规模也会变大，从而增加了计算量。而子字切分的方式可以大大缩小词典规模，来缓解这个问题。

表 2 使用子字粒度切分蒙汉双语词典大小统计

	蒙古文词典	汉文词典
6w_char	30113	4781
6w_word	30113	34167
6w_BPE	15600	8480
26w_char	83561	8447
26w_word	83561	50656
26w_BPE	32053	15659

3.3 两种 BPE 在蒙汉双语上的应用

本文在 3.1 介绍了使用 BPE 算法实现子字粒度切分有两种方式：独立 BPE 和联合 BPE。我们在 CWMT2009 的蒙汉双语对齐语料上分别进行了两种切分，来分析在蒙汉神经机器翻译任务中，哪种子字切分方式更适合。

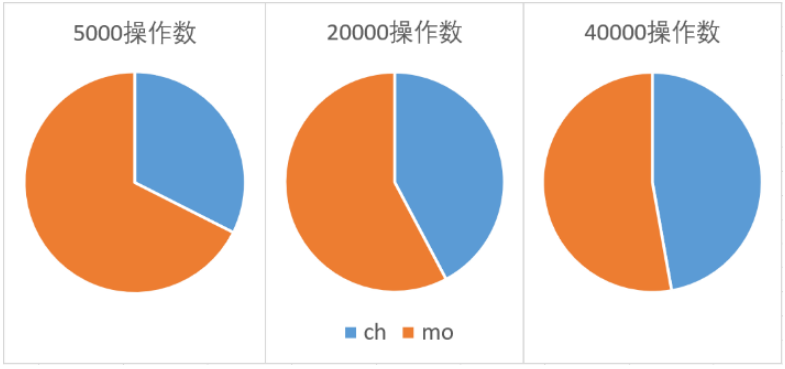


图 3 CWMT2009 不同操组数联合词典占比图

如图 3 所示，我们分别进行了操作数为 5000、20000 和 40000 的联合 BPE 切分实验。理论上联合 BPE 可以尽可能的保持源语言和目标语言分割的一致性。但是我们看到操作数不同的情况下，联合词典中蒙汉的占比差距较大。蒙汉占比不相近，使得蒙汉在切分子字粒度时粒度不一致。我们分析认为这与联合 BPE 的优势并不矛盾，词典占比不均的原因主要是蒙汉语言差异大，蒙古文是典型的拼音文字，而汉文字、词、短语没有明确分割界限，且作为蒙古文最小切分单元的蒙古文字母的数量远远少于汉文字，经统计我们得知，蒙古文单词平均由 5 个蒙古文字母组成，而汉文中绝大多数是二字词。在 BPE 计算共现频数时，蒙古文字母组合频数大，会优先组合蒙古文。因此随着操作数的增加，蒙古文高频组合结合完成后，越来越多的汉文字组合才被加入到词典中，蒙汉占比也越

来越趋近于相等。

表 3 40000 操作数蒙汉子字切分结果

原文	40000 操作数
[illegible] ?	[illegible] ?
巴西人讲葡萄牙语。	巴西 人 讲 葡萄 牙 语。
今天的兑换率是多少?	今天的兑换率是多少?

我们将蒙汉占比较为平衡的40000操作数的切分结果与原文进行了对比如表3所示。我们发现蒙古文中存在大量与原文一致的句子，而汉文中粒度却参差不齐。因此，我们分析认为使用联合BPE进行子字粒度切分，在蒙汉双语语料上无法体现自身的优势，而使用独立BPE进行子字粒度切分可以人为控制蒙汉双语各自的切分粒度，达到最优的切分结果，这样子字粒度切分才会发挥出其应有的效果。

4 实验

4.1 实验数据配置

实验数据均基于CWMT2009的蒙汉双语语料（6万句）和CWMT2017蒙汉评测提供的双语语料（26万句）。汉文语料分词处理采用中科院计算所开发的ICTCLAS[15]中文分词系统进行切分，蒙古文和汉文均使用tokenizer[16]进行词语切分处理。BPE子字切分操作数均为最优结果的经验值。在本文中采用BLEU4[17]作为翻译效果的评测指标。

由于蒙汉机器翻译双语语料有限，词典本身就较小，因此我们的实验不限制词典大小。我们想要通过对比子字粒度切分处理后词典大小变化，并结合最终翻译模型的BLEU值的提升，来验证子字粒度可以通过缓解数据稀疏问题来提高模型翻译质量。

我们使用CWMT2009的蒙汉双语语料分别训练RNN和CNN两个模型，且分别进行了分词和子字粒度切分等对比实验，用来验证子字切分粒度的普适性以及对模型的提升。其次，我们使用CWMT2017的语料训练RNN模型，与使用CWMT2009训练的RNN进行对比，以验证子字粒度切分对数据稀疏的缓解和缩短训练周期的作用。

实验使用的RNN和CNN模型是此前调优的结果，本文所有实验均使用以下参数：

RNN模型，使用的Google 开源的seq2seq系统作为RNN系统。系统参数如下：编码器和解码器均为4层的双向LSTM循环单元[18]，隐藏层节点数512，batchsize 128，dropout 0.3。

CNN模型，使用的是Facebook AI Research 开源系统fairseq作为CNN翻译系统。系统参数如下：编码器不少于5层和解码器不少于9层，解码器的每一层均配备一个注意力机制，编码器和解码器的核宽度不小于 3，词向量维度 500，隐层单元数量不少于500，batchsize 32，训练算法Nesterov’s accelerated gradient (NAG)[19]。

4.2 蒙汉双语子字切分

4.2.1 蒙古文子字粒度切分

由于BPE操作数是个经验值，不同的语言种类、语料大小最优操作数也不同。

我们设置操作数为5000、10000、15000、17500和20000，将蒙古文切分成子字粒度，汉文分词，分别训练RNN模型，测试集BLEU值如图4所示（横坐标为操作数，纵坐标为BLEU值）。最终得出该语料蒙古文的最优操作数的经验值为17500。

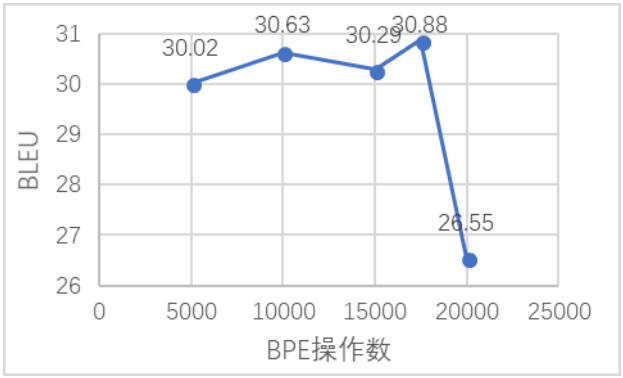


图 4 不同操作数切分语料训练 RNN 表现

如表 4 所示，通过原文与最优切分结果的对比，我们发现，蒙古文子字切分粒度还是比较大的，与分词差别不大。蒙古文词构成通常是由词干和多个词缀构成[20]。结合蒙古文形态分析，我们认为子字粒度切分其实相当于蒙古文词缀的切分，而最优的切分结果能够使得切分的蒙古文子字片段既保留了较为完整的局部信息，又提高出现频率，从而缓解数据稀疏问题。

表 4 17500 操作数蒙汉子字切分结果

原文	17500 操作数
ᠬᠡᠵᠡᠶ᠋ᠢ ᠪᠠ ᠪᠢ ᠰᠠᠶᠢᠨ ᠤᠭ ᠪᠣᠯᠪᠠᠴᠤ ᠤᠨ ᠵᠠᠪ ᠬᠡᠯᠪᠡ ᠂	ᠬᠡᠵᠡᠶ᠋ᠢ ᠪᠠ ᠪᠢ ᠰᠠᠶᠢᠨ ᠤᠭ ᠪᠣᠯᠪᠠᠴᠤ ᠤᠨ ᠵᠠᠪ ᠬᠡᠯᠪᠡ ᠂
ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠪᠣᠯᠤᠮᠵᠢ ?	ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠪᠣᠯᠤᠮᠵᠢ ?
ᠲᠡᠳᠡ ᠭᠠᠵᠠᠷ ᠳᠤ ᠰᠢᠨᠵᠢᠯᠡᠭᠡᠡᠨ ᠦ ᠵᠠᠰᠠᠭ ᠶᠡᠬᠡ ᠂᠂	ᠲᠡᠳᠡ ᠭᠠᠵᠠᠷ ᠳᠤ ᠰᠢᠨᠵᠢᠯᠡᠭᠡᠡᠨ ᠦ ᠵᠠᠰᠠᠭ ᠶᠡᠬᠡ ᠂᠂
ᠲᠡᠷᠡ ᠪᠠ ᠠᠮᠢᠳᠤ ᠪᠠ ᠮᠥᠨᠬᠡ ᠪᠠᠶᠢᠬᠤ ᠭᠡᠵᠦ ᠂	ᠲᠡᠷᠡ ᠪᠠ ᠠᠮᠢᠳᠤ ᠪᠠ ᠮᠥᠨᠬᠡ ᠪᠠᠶᠢᠬᠤ ᠭᠡᠵᠦ ᠂
ᠪᠣᠯᠤᠭᠰᠠᠨ ᠤ ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ ᠠ ᠭᠡᠵᠦ ᠂	ᠪᠣᠯᠤᠭᠰᠠᠨ ᠤ ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ ᠠ ᠭᠡᠵᠦ ᠂

同理，我们也经过大量实验，得到 CWMT2017 蒙古文独立 BPE 的最优操作数的经验值为 35000 左右。

4.2.2 汉文子字粒度切分

汉文与蒙古文差异较大，尤其是蒙古文有明确的分词边界，而汉文字、词、短语界限模糊。因此我们想到两种子字粒度切分方式，如图 5 所示：一是汉文直接使用 BPE 算法切分子字；二是模拟蒙古文子字粒度切分方法，先进行中文分词处理，人为加入词边界，再使用 BPE 算法切分子字。

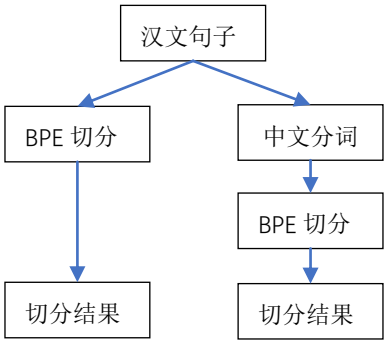


图 5 汉文子字切分方法

两者相比，前者切分结果存在大于分词的粒度，可能包含更多的局部特征，也可能破坏原本词的局部特征，例如：“大学 生活”→“大学生活”或者“大学生”+“活”；后者切分结果的粒度一定小于分词，不会破坏分词前后信息，例如：“大学 生活”→“大学”+“生活”或者“大学”+“生”+“活”。

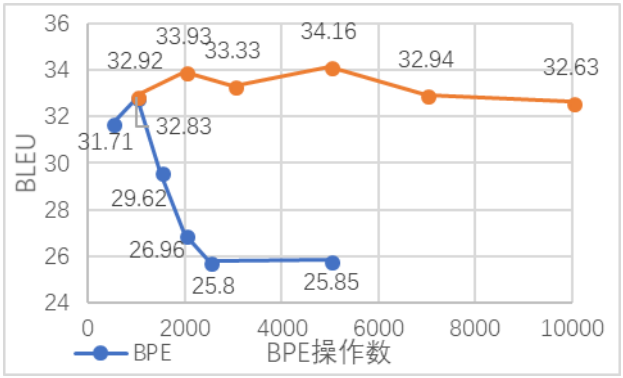


图 6 汉文两种子字切分方法对比

如图 6 所示，我们在保持蒙古文独立 BPE 操作数为最优操作数不变的基础上，汉文使用第一种子字切分方法，操作数设置为 500、1000、1500、2000、2500 和 5000，分别训练 RNN 模型。当操作数为 1000 时，模型效果最佳，BLEU 值达到 32.83。同理，我们使用第二种方法，操作数分别设置为 1000、2000、3000、5000、7000 和 10000，可以看到当操作数为 5000 时模型效果最佳，BLEU 值达到 34.16。

通过对比实验，我们明显可以看出汉文先进行中文分词，再进行子字粒度切分的效果更好。同理，我们得到 CWMT2017 汉文独立 BPE 的最优操作数的经验值为 15000 左右。

4.2.3 词缀切分与 BPE 子字切分对比

我们根据蒙古文形态分析发现，理论上对蒙古文切分后缀也可以减小粒度，缓解数据稀疏问题，因此我们通过规则和字典对蒙古文进行了词缀分析，用来作为子字切分的对比实验。根据蒙古文构词规则，我们知道改变构词后缀一般会改变该词的词义和语义，而改变构形后缀只会改变词性，因此我们通过蒙古文构词规则和后缀字典对构形后缀和结尾后缀进行切分。

如表 5 所示，从后缀切分与子字粒度切分实例对比我们可以发现子字粒度切分结果与两种后缀切分结果相似。

表 5 使用子字粒度切分蒙汉双语词典大小统计

处理方法	示例
蒙古文原文	ᠪᠢ ᠰᠤᠮᠤᠨ ᠤ ᠠᠮᠢᠲᠠᠨ ᠢ ᠬᠠᠷᠠᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠃
子字粒度切分	ᠪᠢ ᠰᠤᠮᠤᠨ ᠤ ᠠᠮᠢᠲᠠᠨ ᠢ ᠬᠠᠷᠠᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠃
结尾后缀切分	ᠪᠢ ᠰᠤᠮᠤᠨ ᠤ ᠠᠮᠢᠲᠠᠨ ᠢ ᠬᠠᠷᠠᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠃
构形后缀切分	ᠪᠢ ᠰᠤᠮᠤᠨ ᠤ ᠠᠮᠢᠲᠠᠨ ᠢ ᠬᠠᠷᠠᠭᠤᠯᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠃

如表 6 所示，我们统计了 CWMT2009 语料上两种后缀切分后词典大小，可以看到与原始语料相比词典大小明显缩小，但是比子字粒度切分的词典大。

表 6 使用子字粒度切分蒙汉双语词典大小统计

	蒙古文词典
CWMT2009 原始语料	30113
CWMT2009 子字粒度切分	15600
CWMT2009 结尾后缀切分	22718
CWMT2009 构形后缀切分	21697

如表 7 所示，因为子字粒度切分和后缀切分结果相近，因此在 CWMT2009 上的表现

两者也互有输赢。我们分析认为两者相比，后缀切分的优势在于切分准确，不会破坏蒙古文词本身。但是子字粒度切分优势在于它完全根据共现频率来切分，不需要外部的规则和字典，且粒度可控，应用更广泛（汉文也可以使用）。

表 7 CWMT2009 上表现对比

模型	蒙古文	汉文	BLEU	训练周期
RNN	词	字	33.03	≈4.8 h
	BPE	字	34.08	≈4 h
	切分结尾后缀	字	32.42	≈4.5h
	切分构形后缀	字	33.85	≈4.5h
	BPE	BPE	34.16	≈3.7 h
CNN	词	字	33.88	≈3 h
	BPE	字	34.08	≈4 h
	切分结尾后缀	字	34.63	≈2.5h
	切分构形后缀	字	35.16	≈2.5h
	BPE	BPE	35.28	≈2 h

4.3 实验结果与分析

表 8 可以看出，同为 RNN 模型，无论小语料库（6w）还是大语料库（26w），汉文分词均不如汉文分字的模型效果好，但是随着语料库的扩大，两种粒度的差距从 3.68 缩小到了 0.39。我们结合表 2 认为，汉文分词效果较差的原因并不是较大粒度本身带来的，而是大粒度切分会放大小语料库数据稀疏问题，致使翻译结果不理想。语料库的扩大，直接缓解数据稀疏问题，因此汉文分词与分字效果差距并不明显。

表 8 RNN 中不同语料大小与不同粒度结果对比

语料大小	蒙古文	汉文	BLEU
6w(CWMT2009)	词	词	29.35
	词	字	33.03
26w(CWMT2017)	词	词	31.36
	词	字	31.75

表 9 CWMT2009 在 RNN 模型上表现

蒙古文	汉文	BLEU	训练周期
词	词	29.35	≈5.5 h
词	字	33.03	≈4.8 h
BPE	词	32.88	≈4.5 h
词	BPE	32.53	≈4 h
BPE	字	34.08	≈4 h
BPE	BPE	**34.16**	**≈3.7 h**

如表 9 所示，蒙汉双语均使用子字粒度切分的 RNN 模型效果最好，BLEU 值达到了 34.16，相比于蒙汉均分词和蒙古文分词，汉文分字的模型分别提升了 4.81 和 1.13。

由于语料较小，训练周期本身不长，因此子字粒度切分虽然使得 RNN 模型训练周期缩短，但效果并不显著。其中蒙汉双语均使用子字粒度切分的模型训练周期最短。

如表 10 所示，相同语料下，CNN 模型对比效果与 RNN 模型相近，依然是蒙汉双语均使用子字粒度切分的模型效果最好，BLEU 值达到了 35.28。因此子字粒度切分在不同模型中都能提高翻译效果，且缩短训练周期，其普适性较高。

表 10 CWMT2009 在 CNN 模型上表现

蒙古文	汉文	BLEU	训练周期
词	词	32.32	≈3 h
词	字	33.88	≈3 h
BPE	词	33.55	≈3.5 h
词	BPE	33.28	≈2 h
BPE	字	34.06	≈2.5 h
BPE	BPE	**35.28**	**≈2 h**

如表 11 所示，RNN 模型在 CWMT2017 的 26 万语料下的表现。由于测试集使用的是日常用语领域的语料，而 CWMT2017 在日常用语的基础上加入了大量政府文献、法律等领域的语料，使得最终翻译效果没有 CWMT2009 上的理想。因此我们只能做同语料下，不同处理方法之间的对比。

因此我们可以看到，蒙汉均使用子字粒度切分的模型翻译效果提升依然显著，且训练周期缩短更为明显，缩短了三分之二。由此可见子字粒度切分随着语料的扩大，缩短训练周期的效果越来越显著。

表 11 CWMT2017 在 RNN 模型上表现

蒙古文	汉文	BLEU	训练周期
词	词	31.36	≈24 h
词	字	31.75	≈23 h
BPE	**BPE**	**33.75**	**≈7.5 h**

5 总结

实验从同语料不同模型和同模型不同语料两个角度进行验证，使用子字粒度切分，将低频词切分成相对高频的子字片段，缓解数据稀疏问题，可以使得多种模型翻译效果得到显著提升，且通过缩小词典大小，显著缩短训练周期。

子字粒度切分技术在语料相对匮乏的前提下，主要是通过切分低频词，来缓解语料匮乏带来的严重数据稀疏问题，从而提高蒙汉神经机器翻译的效果。词典规模和网络模型的训练周期随着语料规模的增大而增加。我们通过限制词典的大小来减少计算复杂度，但这会使得一部分原本在词典中的低频词成为集外词，影响翻译质量。子字粒度切分会将低频词切分成更小单元，这样集外词可以由粒度较小的子字单元拼接而成。因此子字粒度切分在大语料限制词典大小的前提条件下，可以充分体现其减少集外词的能力，来提高模型的翻译效果。

因此，在蒙汉神经机器翻译任务中，子字粒度切分技术无论在现在还是未来都是一个很有应用价值的技术。

参考文献

[1] Luong M T, Sutskever I, Le Q V, et al. Addressing the Rare Word Problem in Neural Machine Translation[J]. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. 2014，27(2):82-86

[2] Hinton G E. To recognize shapes, first learn to generate images [J]. Progress in brain research, 2007, 165: 535-547.

[3] Zhao H, Yajuan L V, Guosheng B, et al. Summary on CWMT2011 MT Translation Evaluation[J]. Journal of Chinese Information Processing, 2012, 26(1):22-30.

[4] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio.Neural machine translation by jointly learning to align and translate. [C]//Proceedings of ACL – IJCNLP, Volume 1: Long Papers. 2015.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems, 2014. 3104-3112.

[6] Mike Schuster and Kuldip K.Paliwal,Bidirectional recurrent neural networks[J].Signal Processing IEEE Transaction, 1997, 45(11):2673-2681

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. 2013, arXiv preprint arXiv: 1301.3781.

[8] Gehring J, Auli M, Grangier D, et al.Convolutional Sequence to Sequence Learning [C]// Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[9] Gehring J, Auli M, Grangier D, et al.A Convolutional Encoder Model for Neural Machine Translation [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 123–135 Vancouver, Canada, July 30 - August 4, 2017

[10] Dauphin, Yann N., Fan, Angela, Auli, Michael, and Grangier, David. Language modeling with gated linear units. arXiv preprint arXiv:1612.08083, 2016.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778

[12] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 1715-1725

[13] Philip Gage .A New Algorithm for Data Compression. C User J.,12(2):23-38,February. 1994

[14] 清格尔泰. 蒙古语语法[M]. 内蒙古人民出版社, 1992.

[15] Papineni K, Roukos S, Ward T, et al. IBM Research Report Bleu: a Method for Automatic Evaluation of Machine Translation[J]. Acl Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002, 30(2):311—318

[16] Zhang R, Yasuda K, Sumita E. Improved statistical machine translation by multiple Chinese word segmentation: Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics[C] Ohio: Association for Computational Linguistics, 2008,216.223.

[17] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation: In Proceedings of the Association for Computational Linguistics, 2007[C]. Prague (Czech Republic): Association for Computational Linguistics, 2007.

[18] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780

[19] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]// International Conference on International Conference on Machine Learning. JMLR.org, 2013:III-1139.

[20] 特格希都楞.蒙古语构词词法研究[M].辽宁民族出版社，2006.

作者联系方式：姓名 任众 地址 内蒙古自治区呼和浩特市赛罕区大学西街 235 号 内蒙古大学 邮编 010021 电话（最好手机） 18586037896 电子邮箱 18586037896@163.com