文章编号：1003-0077(2019)12-0054-07

中文字粒度切分在蒙汉机器翻译的应用

苏依拉，高 芬，仁庆道尔吉

（内蒙古工业大学 信息工程学院，内蒙古 呼和浩特 010080）

摘 要：在机器翻译任务中，主流的深度学习算法大多使用词或子词作为基础的语义单元，在词或子词层面学习嵌入表征。然而，词粒度层面存在一系列缺点。该文基于 LSTM 和 Transformer 蒙汉翻译模型，对蒙文进行子词粒度切分，对中文分别进行子词和字粒度切分对比实验。实验结果显示，相比于子词粒度切分，基于 Transformer 的蒙汉翻译模型和基于 LSTM 的蒙汉翻译模型的字粒度切分有极大的 BLEU 值提升，字级别的蒙汉翻译模型在验证集和测试集上都显著优于混合字和词的子词级别的蒙汉翻译模型。其表明，字级别的蒙汉翻译模型更能捕捉单元之间的语义联系，提高蒙汉翻译性能。

关键词：字粒度切分；Transformer；LSTM

中图分类号：TP391 **文献标识码**：A

Application of Chinese Character in Mongolian-Chinese Machine Translation

SU Yila, GAO Fen, RENQING Dao'erji

(College of Information Engineering, Inner Mongolia University of Technology, Hohhot, Inner Mongolia 010080, China)

Abstract: Most current NMT models applies word or sub-word as the unit to learn embedded representations. To deal with the existing errors at the word level, this paper conducts sub-word segmentation for Mongolian, and sub-word and character segmentation for Chinese, respectively, on the translation models of LSTM and Transformer. Experimental results show that transformer and LSTM models with char segmentation both achieve significant improvements in terms of BLEU.

Keywords: char granular segmentation; Transformer; LSTM

0 引言

自 2013 年以来，得益于深度学习方面取得的进展以及一些资源丰富的大规模平行语料库的可用性，神经机器翻译(neural machine translate，NMT)模型获得了迅速发展[1]，与统计机器翻译(statistical machine translation，SMT)模型相比，其翻译质量明显提升。机器翻译是解决不同民族和国家之间信息交流所面临的“语言屏障”问题的关键技术，在加强文化交流、促进民族团结以及推动对外贸易等方面都意义重大[2]。蒙汉机器翻译对于蒙汉两种文化的价值观相互渗透，凝聚民族的核心文化，促进良好民族关系的建立具有重要意义。然而由于内蒙古地区经济发展相对缓慢，蒙汉平行语料收集困难，利用现有的神经网络方法会放大数据稀疏以及训练过拟合等问题，导致翻译质量不高。并且在翻译模型中，编码器和解码器的计算复杂度比较高，由于计算量和 GPU 内存的限制，神经机器翻译模型需要事先确定一个规模受到限制的常用词词表，神经机器翻译系统往往将词汇表限制为高频词，并将其他所有低频词视为未登录词(out-of-vocabulary，OOV)。中文的字是最小的音义结合体，但是在现代汉语中，词才是中文的基本表达单位，并且中文以双字词或者多字词为主，而大部分的词都是由多个字组合而成。中文分词的统计机器学习方法优于传统的规则方法，尤其是在未登录词上具有无可比拟的优势。主流的深度学习算法也大多使用词或子词

收稿日期：2019-06-13 **定稿日期**：2019-08-19

基金项目：国家自然科学基金(61966027，61966028)；内蒙古自治区自然科学基金(2016MS0605)；内蒙古自治区民族事务委员会基金(MW-2017-MGYWXXH-03)

作为基础的语义单元，在词或子词层面学习嵌入表征。

虽然词粒度级别的模型在很多任务上得到普遍应用，但是词粒度级别的输入单元有大量显著不足：(1)词粒度数据会放大稀疏问题，容易导致过拟合，并且过多的未登录词会阻碍模型的学习能力。(2)分词会引入噪声，并且词的多义性会影响分词的方法，同时人对词的颗粒度的认识都会有差别，故分词方法的不同导致分词效果不同，错误的分词将会影响后续翻译模型效果；(3)直观上看，词携带的语义信息比字更丰富，但是对于神经网络机器翻译而言，词粒度模型是否优于字粒度模型还不清楚。

相比于词粒度，字粒度具有以下优点：(1)可以缩小词汇表，以词为单位时，词汇表太大，而以字为单位时，词汇表(即字表)的大小适中。词汇表越大，深度学习所需要的参数就越多，训练起来就越困难。(2)以字粒度为单元，不依赖于分词工具的正确性。(3)可以让模型集中在不同字的交互方面，例如，以“可能”“这个词为例，“可”可能有“可爱”“可能”“可以”等多种意思。“能”也有可能有“能力”“能够”等多种意思。但是因为深度学习模型善于处理远距离依赖关系，模型在后面的层是会去学习“可”和“能”组合起来的意思。

文献[3-4]证明，中文输入粒度是混合字与词粒度的子词粒度的效果比单纯的词粒度的效果更好。中文分词效果差的原因是在较小语料库的情况下，大粒度的切分会放大数据稀疏问题，使得翻译效果不太好，子词粒度切分后，由分词后的低频词切换成高频词，从一定程度上缓解数据稀疏问题，提高了翻译质量。故本文基于 LSTM 和 Transformer 为基本的蒙汉翻译模型，以内蒙古工业大学构建的 126 万蒙汉平行语料库和 3 万蒙汉专有名词为实验数据，对中文输入粒度进行子词粒度和字粒度切分的对比实验。

本文的内容安排如下：第 1 节为神经机器翻译模型概述，主要介绍了目前比较热门的神经机器翻译技术和方法。第 2 节为相关技术介绍，主要介绍了中文的词、子词、字粒度。第 3 节为实验部分，主要包括语料库的划分、实验设置和实验结果。第 4 节为结论部分，对文中所做的工作进行了总结和展望。

1　神经机器翻译模型

NMT 相比于传统的 SMT 而言，是一个能够通过训练把一个序列映射到另一个序列的神经网络，输出的可以是长度变化了的序列，这对实现自动翻译、人机对话以及文字概括等有很大好处。NMT 其实就是一个编码器—解码器结构，编码器把源语言序列进行编码，并提取源语言中的有用信息，再通过解码器将这些信息转换至目标语言中来，以实现对语言的翻译。

1.1　基于 LSTM 的神经网络翻译模型

长短期记忆网络[5](long short term memory，LSTM)是一种特殊的 RNN。LSTM 能够避免长期依赖性问题，其实现建模的公式如式(1)～式(4)所示。

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + bi) \tag{1}$$

$$c_t = (1-i_t)\odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{3}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4}$$

其中，σ 是元素的 Sigmoid 函数。对于给定的一个包含 n 个单词的句子 $(x_1, x_2, \ldots, x_n)$，每个单词表示为 d 维的词向量。

本文的实验使用了哈佛大学开源的神经机器翻译系统 OpenNMT，这是基于注意力机制的 LSTM 翻译模型。注意力机制[6]的结构原理如图 1 所示。

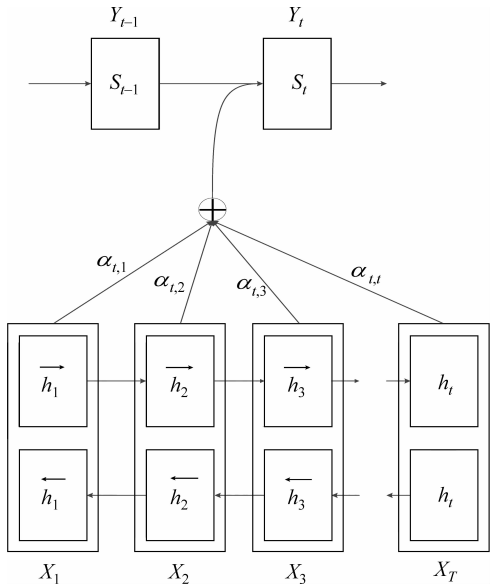

图 1　注意力机制

1.2　基于 Transformer 的神经网络翻译模型

Transformer[7]是完全基于注意力机制的 Seq2Seq[8]

模型，采用多头自注意力[9]来构造编码器和解码器。它的设计思想是把序列中的所有单词并行处理，同时借助自注意力机制将上下文与较远的单词结合起来。在每个步骤中，句子中的每一个词的信息都可以借助自注意力机制与句子中的所有其他的词进行沟通。图 2 为 Transformer 结构图。

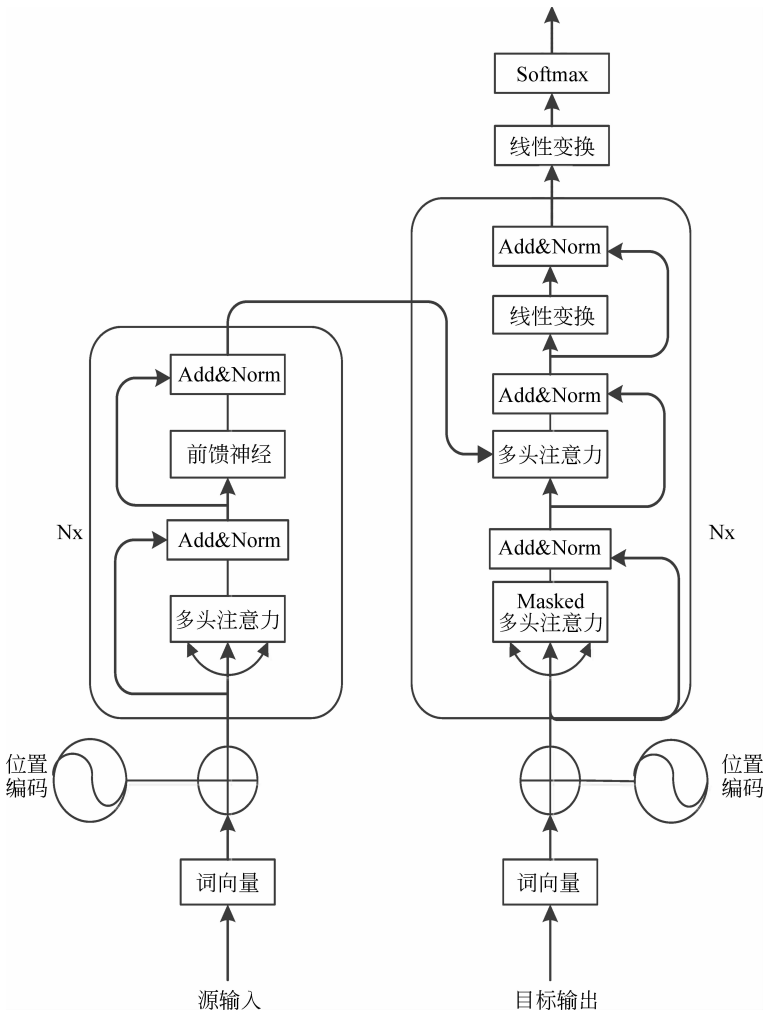

图 2 Transformer 结构

2 相关技术

2.1 蒙古文预处理

传统蒙古文虽然属于阿尔泰语系蒙古语族，但却与印欧语系中的大部分语言一样，属于一种拼音文字，只是它的构成基础是回鹘字母[10]。所以和英语一样，蒙古文句子由字、空格以及标点符号构成，句子本身就已经算是具有词级粒度的形式。蒙古语属于黏着语，黏着语的一个特点是通过在词根的前、中、后位置缀接其他构词成分作为派生新词的手段，因此蒙古文构词及其形态变换非常丰富，导致集外词和未登录词现象频发。文献[3-4]证明，蒙古文进行子词粒度的切分能够减少集外词和未登录词现象，故本文蒙古文的输入粒度是子词。蒙古文的具体构词规则如图 3 所示。

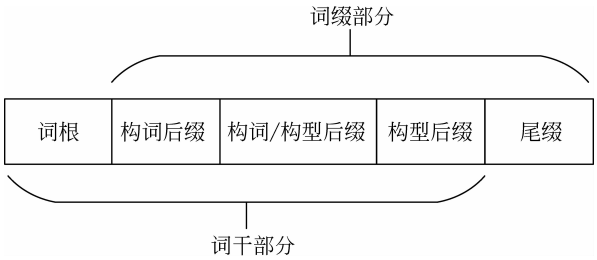

图 3 蒙古文构词规则

对蒙古文进行子词处理，是指把蒙古文切分为比词级粒度更小的粒度。Sennrich 和 Haddow 提出一种子词级粒度处理文本的方式，称为字节对编码(byte pair encoder, BPE)[11]。BPE 处理使用

Github 上的开源系统 fastBPE，实验中将 BPE codes 数设置为 32 000。例如，蒙古文“ᠪᠠᠶᠠᠷᠲᠠᠢ᠃”（中文意思是衷心祝贺！），经过 BPE 处理后，蒙古文变成“ᠪᠠᠶᠠᠷ@@ ᠲᠠᠢ᠃”。将原蒙古文单词切分为两部分，前半部分属于蒙古语词干，后半部分是词缀，其实并不是语言学上真正的词缀，而是剩余蒙古语词缀连接在一起的结果，词缀并没有实际意思，这样切分的好处是增加了低频词子词的共现次数，从而一定程度上减少未登录词的出现。

2.2　中文预处理

2.2.1　词级粒度

汉语属于汉藏语系，每句话只由单个的字和标点符号构成。汉语分词是将连续的中文字符序列按照某种规则分割成词的序列过程[12]。当然，语料的切分也会带来相应的一些问题，比如会出现歧义词现象和破坏语料整体的语义关系等。

本文中文分词使用的分词工具是双向 LSTM 和 CRF（Bi-LSTM-CF）模型[13]。CRF[14]（conditional random fields）中文名称是条件随机场，其主要作用是弥补最大熵马尔科夫模型（maximum entropy markov model，MEMM）分词方法的不足，也就是为了缓解 MEMM 中的标记偏置问题。CRF 是一种判别式模型，CRF 通过定义条件概率 $P(Y|X)$ 来描述模型。

融合双向 LSTM 和 CRF（Bi-LSTM-CRF）模型分词算法结构如图 4 所示。图 4 中，输入层为词嵌入表示层，经过一个双向的 LSTM 网络编码，输出层是一个 CRF 层，经过 LSTM 网络输出的实际上是当前位置对于各词性的得分，CRF 是对词性得分加上前一位置的词性概率转移的约束，其优点在于引入了一些语法规则的先验信息。

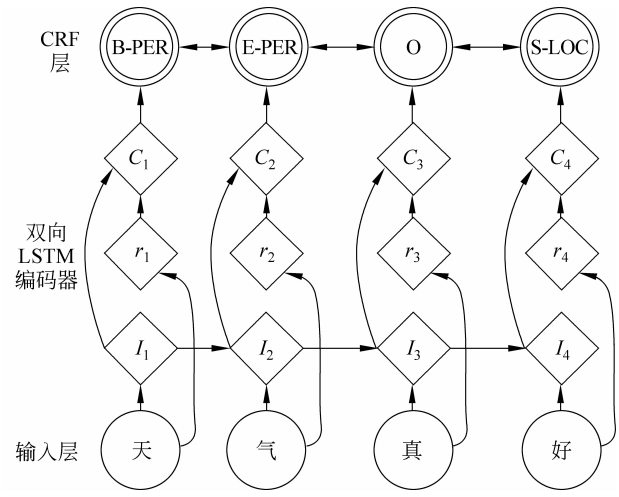

图 4　融合双向 LSTM 和 CRF 模型的网络结构图

2.2.2　子词级粒度

子词级粒度切分指的是将语料切分为比词级粒度更小的单元。把罕见词拆分为子词单元的组合，子词单元的拆分策略，是借鉴了一种数据压缩算法 BPE，其基本原理是：首先将语料以最细粒度为单元进行分切，构造一个初始化子词表，接着统计出语料中所有单元之间组合出现的次数，最后将统计到的次数按从大到小的顺序排列，选择出现次数最多的组合替换初始子词表中的单个单元，在这个过程中最重要的是系统需要自动学习一个词典，从而才能依据学习到的词典对语料进行切分。经过子词切分后，即便是训练语料库里没有见过的新词，也可以通过拼接子词来生成翻译。子词切分可以有效地缓解神经机器翻译中的 OOV 和罕见词翻译。

2.2.3　字级粒度

将中文进行字粒度切分，分字使用自己编写的脚本文件，通过将原语料读入之后，为每个字之间加入空格，然后输出到新的文本保存即可。

3　实验

3.1　语料库划分

本文实验数据来源于内蒙古工业大学构建的 126 万蒙汉平行语料库和 3 万蒙汉专有名词。3 万的蒙汉名词库包含地名、人名、农业名词、医学名词和物理名词等。126 万蒙汉平行语料包括政府新闻、法律公文、日常对话、日常用语、网络对话与聊天等。蒙汉平行语料库数据集的划分如表 1 所示。

表 1　实验数据集划分

	训练集	验证集	测试集	总数据集
数据量	1 290 559	3 000	1 500	1 295 059

对蒙汉平行语料库中的蒙文进行子词粒度切分后，统计出词表，如表 2 所示。

表 2　蒙古文词表统计

	Total words	Unique words	ration
词	14 946 860	323 508	0.022
子词	17 128 867	19 865	0.001

从表 2 可以看出，语料库里蒙古文总共有 14 946 860 个，只出现过一次的词有 323 508 个。经过子词粒

度切分操作后，蒙古文总共有17 128 867个词，只出现过一次的词有19 865个。只出现一次的比率从2.2%下降到0.1%，有效缓解了数据稀疏。

对126万蒙汉平行语料库中的中文分别进行词粒度、子词粒度和字粒度切分后，统计出词表如表3所示。

表3 中文词表统计

	Total words	Unique words	ration
词	11 891 530	320 860	0.027
子词	14 180 095	24 241	0.002
字	21 808 187	7 412	0.000 3

Zipf定律表明，大部分中文词的出现频率都很小，并且在数据集中的占比十分有限，这导致模型不能充分学得数据中的语法、语义知识。如表3所示，经过双向LSTM和CRF(Bi-LSTM-CR)模型分词后，数据集中的中文词总共有11 891 530个，其中有320 860个词仅出现了一次，占数据集总词的2.7%。这表明，词粒度级别的数据放大了数据稀疏，比较容易导致过拟合问题。子词切分粒度下，中文词总共有14 180 095个，并且有24 241个词只出现了1次，比率为0.2%，说明经过子词粒度切分后，减少了数据稀疏。字粒度模型下，中文词(即字)总共有21 808 187个，只出现了一次的词仅有7 142个，占比率0.03%。经过字粒度切分后，只出现过一次词的比率从词粒度的2.7%缩小到0.03%，极大减少了低频词。

3.2 实验设置

基于Transformer的蒙汉翻译模型本文选用清华大学自然语言处理小组开发的机器翻译库THUMT。使用一台搭载NVIDIA Tesla P100 GPU，RAM 16GB；基于LSTM的蒙汉翻译模型本文选用开源库OpenNMT，使用一台搭载NVIDIA 1070 Ti GPU，RAM 8GB。实验环境为Ubuntu 16.04，Linux系统，语言为Python 2.7.0，TensorFlow版本为1.6.0，Anaconda 3-5.2.0。用multi-bleu.per脚本评测翻译性能BLUE值。

LSTM LSTM神经网络的编码器和解码器的隐藏层数设置为4层，词向量维度设置为1 000，解码器中全局注意力机制中输入特征设置为500，输出特征设置为500，激活函数选择tanh()。Torch 0.4.0，TorchText 0.2.3。dropout设置为0.3，train_steps设置为200 000，学习率初始值设置为0.1，学习率衰减速率设置为1。

Transformer Transformer的神经网络层数设置为6层，多头注意力机制设置为8头，激活函数使用GELU，优化函数使用Adam优化算法，学习率初始值设置为0.1，一阶矩估计的指数衰减率设置为0.9，二阶矩估计的指数衰减率设置为0.98，train_steps设置为200 000，batch_size设置为4 096。

3.3 实验结果

如图5、图6所示，本文分别统计出了基于LSTM神经网络蒙汉机器翻译模型和基于Transformer神经网络蒙汉机器翻译模型在200 000 train_steps上对蒙文进行子词粒度切分，中文分别进行子词和字粒度切分的测试集和验证集的BLEU值以及其变化趋势。

表4和表5分别是基于LSTM翻译系统和Transformer翻译模型，对蒙文进行子词粒度切分，

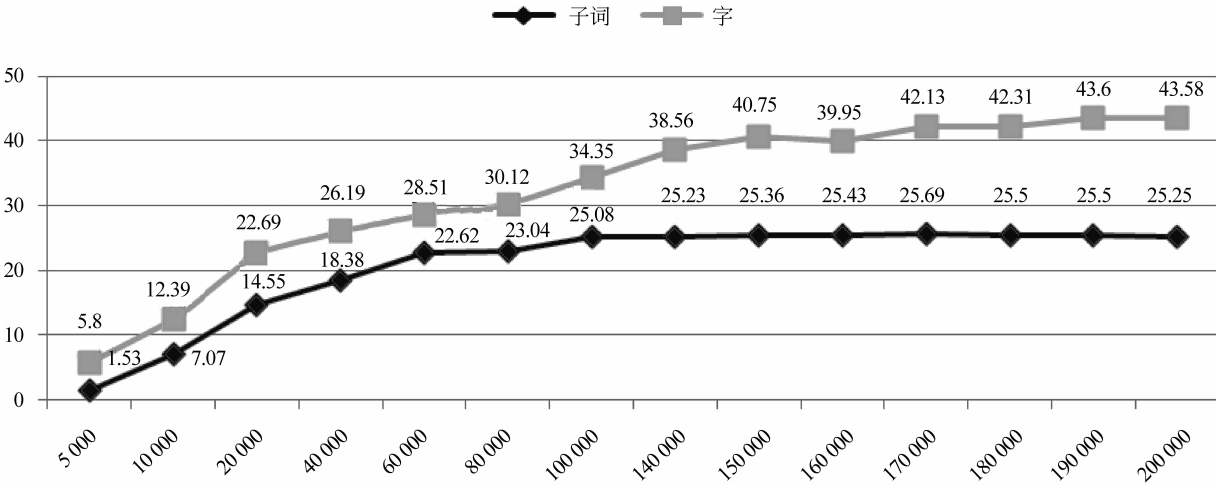

图5 基于LSTM翻译模型测试集的BLEU值变化趋势

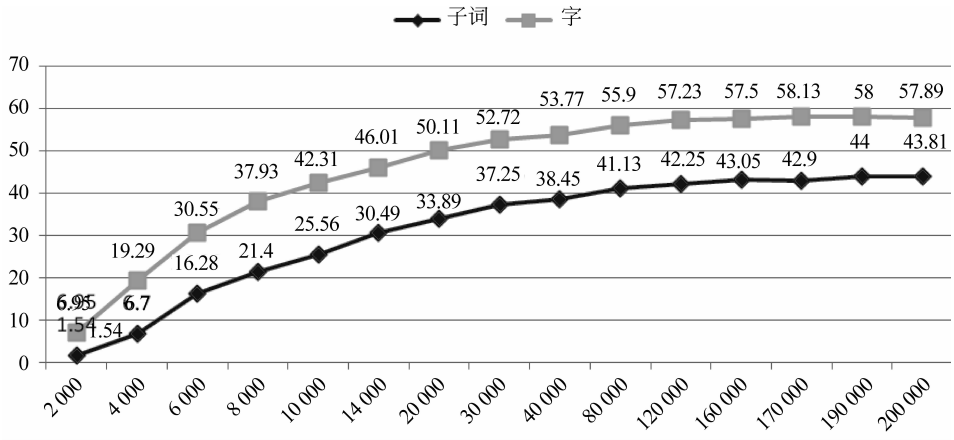

图 6　基于 Transformer 翻译模型验证集的 BLEU 值变化趋势

中文分别进行子词和字粒度切分的测试集的 BLEU 值和 ACC 值。

表 4　基于 LSTM 的翻译模型的 BLEU 值和 ACC 值

	BLEU	ACC
子词	25.25	58.5
字	43.58	72.7

表 5　基于 Transformer 的翻译模型的 BLEU 值和 ACC 值

	BLEU	ACC
子词	44.14	70.2
字	57.90	80.1

从表 4 可以看出，基于 LSTM 翻译模型，对蒙文进行子词粒度切分，对中文进行子词粒度切分后，BLEU 值达到了 25.25，准确率 ACC 达到了 58.5；对中文进行字粒度切分后，BLEU 值达到 43.58，准确率 ACC 达到了 72.7。基于 Transformer 翻译模型中，中文子词粒度翻译模型的 BLEU 值为 44.14，中文字粒度翻译模型的 BLEU 值为 57.90，相对于子词粒度提高了 13.76 个 BLEU 值，准确率大约提高了 10 个值。

图 5、图 6、表 4 和表 5 说明基于字粒度的模型优于混合字和词的子词模型，这表明了基于字粒度的模型已经编码了语言建模任务所必要的语言信息，另外加入词反而会损害其翻译表现。在蒙汉翻译中，解码端的 UNK 对于词影响更大。字级别的模型在验证集和测试集上都显著优于混合词级别和字级别的子词模型。字粒度的模型能集中在不同字的交互方面，更能捕捉单元之间的语义联系，提高翻译质量。

4　总结与未来工作

在蒙汉机器翻译中，翻译单元的大小直接影响着翻译的性能，而翻译单元的大小又是通过语料的切分粒度而体现的，所以本文主要介绍了对中文语料切分的两种方式：字切分粒度和混合字与词的子词切分粒度。对于每一种方法，都通过原理概述和实验来进行说明，最后对各方法的实验结果进行了对比分析。结果表明在蒙汉机器翻译中，对中文进行字级别粒度切分要优于混合词和字的子词粒度切分。本文的结论与 Meng Y[15] 等人的结论也一致。但是字粒度切分也存在问题，一个很关键的问题就是一字多义，而词在一定程度上减轻了这个问题，这也是在统计时代分词存在的必要性。而最新超火的预训练语言模型 bert 就完全舍弃了分词的过程，而是采用字粒度划分。如果一字多义的问题能够通过预训练语言模型来解决，下一步拟将最新的预训练模型应用于蒙汉翻译。本文对中文划分粒度的研究为后续的蒙汉机器翻译的研究做了很好的理论铺垫，也为后续的实验打下了坚实基础，是非常重要的一个环节。

参考文献

[1] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1700-1709.

[2] 刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展, 2017, 54(6): 1144-1149.

[3] 任众, 侯宏旭, 吉亚图, 等. 子字粒度切分在蒙汉神

经机器翻译中的应用[J]. 中文信息学报，2019，33(01)：90-97.
[4] 韩冬，李军辉，熊德意，等. 基于子字单元的神经机器翻译未登录词翻译分析[J]. 中文信息学报，2018，32(4)：74-79,119.
[5] Gers F A, Schmidhuber E. LSTM recurrent networks learn simple context-free and context-sensitive languages[J]. IEEE Transactions on Neural Networks, 2001, 12(6): 1333-1340.
[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473, 2014.
[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems, 2017: 5998-6008.
[8] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems, 2014: 3104-3112.
[9] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[J]. arXiv preprint arXiv: 1703.03130, 2017.
[10] 道布. 蒙古语简志[M]. 北京：民族出版社，1983，137-142.
[11] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of the ACL. Stroudsburg, PA: ACL, 2016: 1715-1725.
[12] Conneau A, Lample G, Ranzato M A, et al. Word translation without parallel data[J]. arXiv preprint arXiv: 1710.04087, 2017.
[13] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.
[14] McCallum A, Freitag D, Pereira F C N. Maximum entropy markov models for information extraction and segmentation[C]//Proceedings of the ICML, Stanford, CA, USA, 2000, 17(2000): 591-598.
[15] Meng Y, Li X, Sun X, et al. Is word segmentation necessary for deep learning of Chinese representations? [J]. arXiv preprint arXiv: 1905.05526, 2019.

苏依拉(1964—)，博士，教授，主要研究领域为人工智能、自然语言处理。
E-mail: suyila@tsinghua.org.cn

仁庆道尔吉(1982—)，博士，教授，主要研究领域为人工智能、云计算与数据挖掘、自然语言处理。
E-mail: 854766886@qq.com

高芬(1994—)，硕士研究生，主要研究领域为人工智能与模式识别。
E-mail: 2191670553@qq.com