文章编号：1003-0077(2019)12-0067-09

融入分类词典的汉越混合网络神经机器翻译集外词处理方法

车万金[1,2]，余正涛[1,2]，郭军军[1,2]，文永华[1,2]，于志强[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500)

摘 要：在神经机器翻译中，因词表受限导致的集外词问题很大程度上影响了翻译系统的准确性。对于训练语料较少的资源稀缺型语言的神经机器翻译，这种问题表现得更为严重。近几年，受到外部知识融入的启发，该文在RNNSearch模型基础上，提出了一种融入分类词典的汉越混合网络神经机器翻译集外词处理方法。对于给定的源语言句子，扫描分类词典以确定候选短语句对并标签标记，解码端利用词级组件和短语组件的混合解码网络，很好地生成单词集外词和短语集外词的翻译，从而改善汉越神经机器翻译的性能。在汉越、英越和蒙汉翻译实验上表明，该方法显著提高了准确率，对于资源稀缺型语言的神经机器翻译性能有一定的提升。

关键词：神经机器翻译；分类词典；资源稀缺；集外词

中图分类号：TP391 **文献标识码**：A

Unknown Words Processing Method for Chinese-Vietnamese Neural Machine Translation Based on Hybrid Network Integrating Classification Dictionaries

CHE Wanjin[1,2], YU Zhengtao[1,2], GUO Junjun[1,2], WEN Yonghua[1,2], YU Zhiqiang[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: In neural machine translation, the problem of unknown words caused by limited vocabulary significantly affects the translation quality. Inspired by the integration of external knowledge, this paper investigates to improve the RNNSearch NMT by incorporating the classification dictionary, and proposes a new hybrid network to deal with the unknown words problem in the Chinese-Vietnamese neural machine translation. For source language sentence, the model scans classification dictionary to determine candidate phrase pairs and tags, the decoder uses hybrid network with both word and phrase level components to generate the translations. Experiments on Chinese-Vietnamese, English-Vietnamese and Mongolian-Chinese NMT show that this method significantly improves the translation performance.

Keywords: neural machine translation, classification dictionaries, resource-scarce, unknown words

0 引言

伴随着我国“一带一路”倡议的提出，面向东南亚国家的汉语-越南语等资源稀缺型语言的翻译需求不断增加。但因为资源受限，目前汉语—越南语的机器翻译模型效果不理想，因此提升汉语—越南语机器翻译系统性能，不仅对于推动两国之间的交流具有十分重要的作用，而且对于资源稀缺型语言的机器翻译研究也具有一定的启发作用。神经机器

收稿日期：2019-04-26 **定稿日期**：2019-09-09

基金项目：国家重点研发计划(2018YFC0830105，2018YFC0830100)；国家自然科学基金(61732005，61672271，61761026，61762056，61866020)；云南省高新技术产业专项(201606)；云南省自然科学基金(2018FB104)；云南省科技人才培养项目(KKSY201703015)

翻译是近年来提出的基于深度神经网络的机器翻译方法，在英—法、英—德等资源丰富语言上取得了很好的翻译效果[1-4]。但其依赖于大规模的平行句对语料，而汉语—越南语属于资源稀缺型语言，没有大规模的双语平行句对资源，因此研究如何在神经机器翻译框架下，提升汉语—越南语神经机器翻译性能是一项具有挑战性的工作。

在传统的神经机器翻译中，为了控制与目标词汇量大小成比例增长的计算复杂性，大多数系统将词表限制为只包含源语言和目标语言中的3万[2]到8万[3]个常见单词，除此以外的词称为集外词，使用UNK进行替换。这种方法虽然能够降低神经机器翻译模型的计算复杂性，但是同时会带来无法翻译集外词以及因UNK导致句子含义模糊等问题。例如，在汉越翻译“阮富仲衷心祝贺中共19大取得圆满成功。双方同意落实好共建‘一带一路’合作文件”中，“阮富仲”和“一带一路”两个词不在词表中，在翻译时导致这两个词无法正常翻译，从而得出“<unk>chân thành chúc mừng Đại hội toàn quôc lần thứ 19 của Đảng Cộng sản Trung Quốc vì đã thành công hoàn toàn.Hai bên đã đồng ý triển khai tài liệu hợp tác '<unk>'”，造成翻译结果语义的不完整。对于资源稀缺型语言的神经机器翻译，由于本身语料规模较小，因而词语的覆盖程度较少，甚至有许多常见的词都不会出现在词表中，成为集外词，导致模型翻译效果不理想。

受到Tang等人[5]和Zhang等人[6]将双语词典引入到神经机器翻译的启发，本文提出一种在RNNSearch模型基础上融入外部分类词典来缓解集外词的混合翻译模型。分类词典包括双语稀有词词典、实体词典和规则词典，分别通过词对齐、维基百科抽取和规则方法生成。在编码端通过查询分类词典，将输入序列标记为词或短语标签，同时在解码端加入选择门控，构造词、短语混合解码结构，结合分类词典得到目标语言翻译结果。本方法通过融入分类词典，一方面通过短语模式可以很好地翻译短语集外词，另一方面还可以通过词级模式翻译词表外的集外词，可以很好地利用词表，从而提升翻译系统的性能和效果。本文首先在汉越神经机器翻译上进行实验，后期又扩展到英越神经机器翻译和蒙汉神经机器翻译，均取得了较好的翻译效果，系统性能有所提升。

1 相关工作

在神经机器翻译中，如何有效地处理集外词问题是近年来的研究热点，但是在资源稀缺型语言的神经机器翻译中，开展集外词问题的研究还相对较少。具体来讲，目前针对集外词问题有以下几类处理方法：

第一类方法是通过指针网络或拷贝机制从源序列中拷贝单词进行翻译。Caglar Gulcehre等人[7]提出Pointer Softmax (PS)使用两个softmax层，预测原输入语句中某个词的位置和预测在预定词表中的单词。Gulcehre等人[7]在神经机器翻译模型上嵌入一种拷贝模式，解码器自动选择是从词表中选择词语进行生成还是从源语言句子中选择词语进行拷贝。

第二类方法将输入序列切分为更小粒度的子词序列来缩小词表规模。Sennrich等人[8]提出使用BPE算法对子词建模。Costa-jussa等人[9]提出将基于字符来产生词嵌入的方法应用于机器翻译。Ling等人[10]的工作中使用双向RNN来编码字符序列。Wu等人[11]提出一种混合字符—词语模型，使用字符序列替代集外词。Chung等人[12]提出了一种新的RNN结构，可以对字符和词进行处理，输出字符序列不需要进行分词。虽然这些方法能够较好地处理罕见/未知单词问题，但因为序列长度的增加而导致训练变得更加困难。

第三类方法为构建大规模词典集和替换技术。Li等人[13]提出集外词“替换—翻译—恢复”的方法。Luong等人[14]提出了在目标语言句子中插入定位符号，以后备词典的方式处理UNK。Jean等人[15]使用大字典并在softmax时进行采样，提出了一种基于重要性抽样的近似训练算法，可以训练一个具有更大目标词汇的NMT模型。所提出的算法在仅使用全部词汇的小子集的水平上可有效地保持训练期间的计算复杂度。

虽然以上工作在处理集外词上有一定的作用，但是均未涉及双语词典等外部知识的融入。最近Arthur等人[16]提出了一种利用NMT模型的注意力向量来选择该模型应该关注的源词词法概率计算候选译文中下一个词的词法概率的方法。Zhang等人[17]提出了新颖的模型，将双语词典转换成足够的句子对，利用混合词/字符模型和合成平行句子保证大量翻译词汇的出现。Tang等人[5]提出了一种存

储词组对的翻译器确定候选短语对标记信息，采用设计的策略一次性生成多个单词序列。但这些方法所使用的都是通用词典，没有研究集外词本身的特点，且多集中在资源丰富语言，没有涉及资源稀缺型语言。本文在 RNNSearch 模型基础上提出一种融合分类词典的混合网络模型结构，在预处理阶段融入分类词典标签标记，在解码端通过不同模式查找词表和分类词典，进而缓解集外词问题。

2　研究背景

本文工作建立在基于注意力机制的神经机器翻译模型上。对于给定一个源语言序列 $x=(x_1,x_2,\cdots,x_{T_x})$，RNNsearch 模型使用双向 RNN 对源语言句子进行编码。它由两个独立的神经网络组成，其中前向 RNN 顺序读入源语言序列 x 后产生源语言前向隐式状态序列 $\overrightarrow{h}=(\overrightarrow{h}_1,\cdots,\overrightarrow{h}_{T_x})$，后向 RNN 逆序读入源语言序列 x 后产生源语言后向隐式状态序列 $\overleftarrow{h}=(\overleftarrow{h}_1,\cdots,\overleftarrow{h}_{T_x})$。前向和后向隐式状态序列中对应位置拼接形成该位置单词的隐式状态 $h=\overrightarrow{h};\overleftarrow{h}$。

在解码时刻 t，解码器分别产生该时刻的目标语言隐式状态和目标语言单词。t 时刻目标语言隐式状态 s_{t-1} 由 $t-1$ 时刻目标语言隐式状态 s_{t-1}，$t-1$ 时刻解码器所生成的目标语言单词 y_{t-1} 和 t 时刻上下文向量 c_t 所决定，如式(1)所示。

$$s_t=f(s_{t-1},c_t,y_{t-1}) \tag{1}$$

其中 t 时刻上下文向量 c_{t-1} 由源语言隐式状态序列 h 和注意力模型所产生的权重加权所得，如式(2)所示。

$$c_t=\sum_{j=1}^{T_x}\alpha_{t,j}h_j \tag{2}$$

$f(\cdot)$ 为 GRU[18-19]。其中注意力模型的权重 $\alpha_{t,j}$ 由 $t-1$ 时刻目标语言隐式状态 s_{t-1} 与源语言隐式状态序列 h 产生，如式(3)～式(4)所示。

$$\alpha_{t,j}=\frac{\exp(e_{t,j})}{\sum_{k=1}^{T_x}\exp(e_{t,k})} \tag{3}$$

$$e_{t,j}=\sigma(s_{t-1},h_j) \tag{4}$$

其中 σ 为非线性函数。权重 $a(t,j)$ 可以解释为源语言词语 x_j 与 t 时刻解码器所产生词语的相关程度。

在生成目标语言隐式状态 s_t 后，为了预测目标词，解码器结合 s_t，c_t 和 y_{t-1} 通过 softmax 函数估计 t 时刻目标语言单词的概率分布，如式(5)所示。

$$p(y_t=y_{<t},x)=\mathrm{softmax}(f(Bs_t+Cc_t+Dy_{t-1})) \tag{5}$$

其中，B、C 和 D 是权重矩阵。

3　模型

3.1　模型总体概述

集外词主要包括稀有词、实体、数字、日期等。本文构建的分类词典将集外词分为三类：一是稀有词；二是实体，包括人名、地名、组织机构名和专有名词；三是数字、日期、符号和时间等。实验发现，如人名“山本五十六”这样的词，在切词后变成“山本”和“五十六”。而在模型的词表中“山本”因为出现的次数少，则被替换为“UNK”，而“五十六”可翻译为“Ngũ Thập Lục”。但实际“山本五十六”正确的翻译为“Yamamoto Isoroku”，二者相差甚大，造成翻译句子质量差，模型性能不理想。

本文通过在端到端的神经机器翻译模型中引入分类词典来解决集外词问题，结构模型如图 1 所示。在预处理时，对切分后的源语言句子进行处理，找到类似于切分成“山本”和“五十六”这样的词，通过扫描查找分类词典进行合并恢复为短语，使用 RNN 编码器将该语句编码为短语表示形式，在编码的时候进行标签化标记。解码端分为短语模式和词级模式，通过混合 RNN 解码器网络生成单词和短语。短语模式为通过分类词典进行翻译的短语，这类短语大多为前面提到的三类集外词，因为在预处理阶段对短语进行标签化标记，解码时模型可以区分这些集外词，然后通过双语分类词典整体进行翻译。词级模式主要分为两种情况：一种情况是翻译的词本身在模型的词表中，对于这类集内词可以直接通过模型的词表翻译生成；另一种情况是这些词不在词表中，即为集外词，对于这样的词，我们同样通过查找分类词典进行翻译。

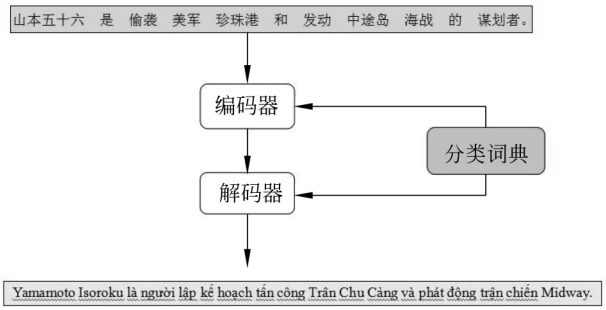

图 1　模型结构

3.2 分类词典

如表1所示，分类词典主要包括稀有词双语词典、实体词典和规则词典。稀有词双语词典的构建包括两个方面。一方面，使用GIZA++词对齐工具对语料进行对齐处理，得到对齐结果，排除词表内的词来构建双语词典，对于一对多的情况，我们只保留对齐概率最大的记录。另一方面，我们在词典中还加入了部分人工整理添加的双语词典。最终本文构建了规模为8 735对的稀有词双语词典。实体词典的构建主要基于维基百科进行词条抽取。页面的词条多为人名、地名等实体词，在左下角会有对应的“Languages”可以链接到越南语的翻译，该链接的HTML信息中包括了翻译后词汇。最终通过维基百科抽取构建的实体词典，包括人名实体6 418对、地名实体2 934对、组织机构名实体5 026对、专有名词实体4 363对，共计18 741对，如表1所示。规则词典的构建采用基于规则方法，对语料进行正则化处理，构建时间、数字、日期等特定标记的词典。

表1 分类词典的类型和内容

分类词典类型		源语言(汉语)	目标语言(越南语)
双语词典		西游记	Tây du ký
		国际歌	Quốc tế ca
		牡丹	Chi Mẫu đơn Trung Quốc
实体词典	人名	习近平	Tập Cận Bình
	地名	台湾	Đài Loan
	组织机构名	全国人民代表大会	Đại hội đại biểu Nhân dân toàn quốc
	专有名词	兵马俑	Đội quân đất nung
规则词典	数字	两千	2000
	日期	2016年2月25日	Ngày 25 tháng 2 năm 2016
	时间	12点25分	12：25
	符号	Ω	Ω

与基于短语的统计机器翻译(SMT)中的规则不同，在SMT中，一个短语可以有多个翻译，并且有一个概率分布。我们分类词典的列表仅仅为一对一的关系，这样可以简化模型设计和训练。同时对于分类词典(Θ)中存在的翻译规则(P,Q)，其中P存在于待翻译的源句子中，Q为P翻译后的词。解码器保持P和Q之间的对应关系，如图2所示。

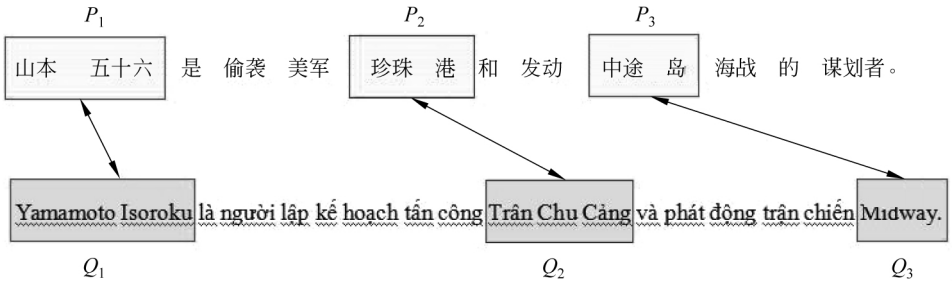

图2 源语言和目标语言句子中的短语对应关系

3.3 编码

分类词典Θ用于在编码之前对句子对进行预处理。为了标记一个源句$x=(x_1,x_2,\ldots,x_{T_x})$，我们需要找到它包含的短语。

我们找出源句x里存在Θ中的规则短语，并将这些规则表示为P_x。同时需要找到P_x在目标句子y中对应词记作Q_x。P_x和Q_x将源句和目标句的单词分成组，如图2所示。源句x中的单词分为两组：短语和单词，而目标句y中的单词分为两组：短语和单词。

我们通过将切分后的词进行处理，查找分类词典对源语言句子中错切的短语进行合并处理。这样的词主要在我们的分类词典Θ中，我们使用RNNSearch中的编码器对短语进行标签标记，标签用于帮助模型定位和区分短语和单词。如图3所示，在句子x中我们将合并后的短语标记为1，其余单词标记为0，在解码过程中通过识别1或者0，从而选择短语模式还是单词模式。

3.4 解码

在RNNSearch解码器中仅包含单词模式，本文为解码器增加短语模式。对于有两个或两个以上的词构成的目标短语$p_t=(y_t,y_{t+1})$，按照短语模式

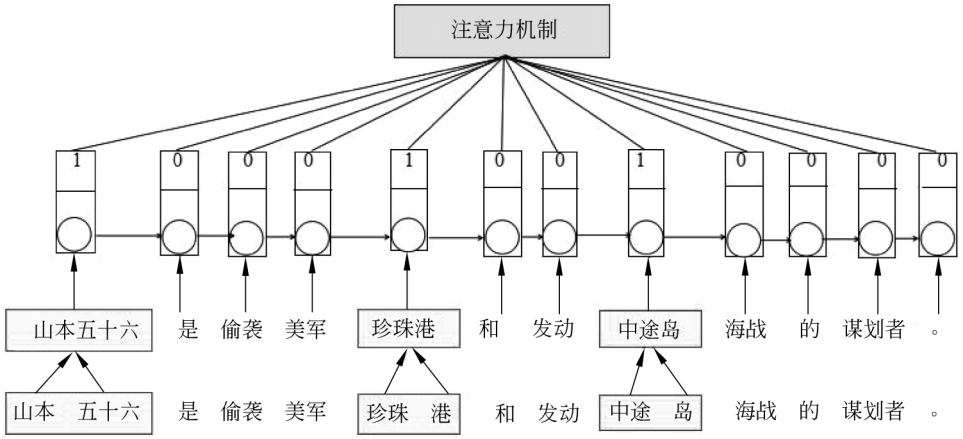

图 3　编码器结构(包括标签)

(整体)生成。在本文的模型中,对于短语的翻译,我们可以通过分类词典进行翻译;对于单词的翻译,如果这个词在模型的词表中,则可直接进行翻译,如果不在词表中则为集外词,可通过查找分类词典进行翻译。

模型的解码器结构如图 4 所示。模型中通过门控单元来决定在 t 时刻使用哪个模式,其中门控单元是二进制指示符变量($\xi \in \{0,1\}$),0 代表词级模式,1 代表短语模式。

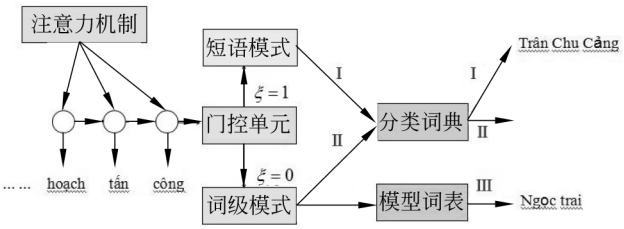

图 4　解码器结构

对于模型,我们简单分为三种情况:

① 当通过门控单元确定为 1 时,则通过短语模式进行翻译。此时需要查找分类词典 Θ 进行翻译,结合图 2 可知,在翻译"công"后面的词时,源语言中为"珍珠港",通过标记后由查找分类词典 Θ 翻译得到"Trân Chu Cảng"。

② 当通过门控单元确定为 0 时,则通过词级模式进行翻译,对应翻译的词不在模型的词表 V 中的情况,为集外词,则通过查找分类词典 Θ 进行翻译。由于正常翻译的时候下一个词为"珍珠",这个词在词表 V 中,所以通过Ⅲ进行翻译。

③ 当通过门控单元确定为 0 时,则通过词级模式进行翻译,对应翻译的词在模型的词表 V 中的情况,可通过查找模型词表进行翻译。结合图 2 和图 4,抛开短语模式来看,正常翻译的时候下一个词为"珍珠",而且"珍珠"这个词在词表 V 中,得到最后的翻译为"Ngọc trai"。

由此我们可以看出,短语模式的翻译和词级模式的翻译得到的翻译结果不同。通过词级模式得到的翻译结果为"Ngọc trai",这里仅仅单指"珍珠"这个词的翻译,放在"珍珠港"这个语境下完全不正确,通过短语模式得到的翻译结果"Trân Chu Cảng"为正确的"珍珠港"的翻译结果。解码过程在 $t=1$ 时刻,编码器通过单词模型生成了单词 y_1,当前的状态为 s_1,然后到下一个时刻 $t=2$。在 s_1 当前状态,注意力机制模型首先生成上下文向量 c_2 作为 h 的加权和(从编码器中)。在有 s_1,c_2 和 y_1 的情况下,解码器更新到下一个状态 s_2,并且生成下一个单词或短语。

4　实验

4.1　实验数据及设置

本文构建 20 万汉越双语平行语料库,为了验证模型在不同语料规模,特别是低资源情况下的性能,将其分为 10 万、20 万两组训练集分别进行实验。另外,构建测试集和验证集,规模均为 5 000 对。在融入本文方法之前对语料做了清洗和 Tokenize 处理。实验中使用的词表为 32 000,句子最大长度为 50,dropout 设置为 0.2,词嵌入维数为 620 维,训练步数为 300 000,hidden_size 为 1 000。实验中使用 BLEU 值作为评测指标。

4.2　实验结果

实验设计包括五个部分,分别是 Moses、

RNNSearch(语料规模为10万)、RNNSearch(语料规模为20万)、本文方法(语料规模为10万)和本文方法(语料规模为20万)。每部分中都包含汉语-越南语、越南语-汉语双向翻译,共计10组实验结果。为了直观地观察和对比,保证实验结果的可靠性,每组的实验结果的BLEU值都采用相同的测试集。表2中列举了汉语-越南语和越南语-汉语两个翻译方向的实验结果。

表2 汉语-越南语和越南语-汉语两个翻译方向的实验结果

模型	BLEU	
	汉语-越南语	越南语-汉语
Moses	19.56	20.12
RNNSearch(10万)	21.74	22.45
RNNSearch(20万)	23.59	24.25
本文方法(10万)	23.42	24.87
本文方法(20万)	25.16	26.07

4.3 实验对比与分析

从实验结果对比看,本文方法通过混合网络融入分类词典后BLEU值有所提升。对于相同规模的10万训练语料,在汉语-越南语翻译方向下,本文方法比RNNSearch有1.68个BLEU值提升;在越南语-汉语翻译方向下,本文方法比RNNSearch有2.42个BLEU值提升。对于相同规模的20万训练语料,在汉语-越南语翻译方向下,本文方法比RNNSearch有1.57个BLEU值提升;在越南语—汉语翻译方向下,本文方法比RNNSearch有1.82个BLEU值提升。

为了进一步验证翻译结果的准确性,我们给出三组汉语的源语言句子和越南语译文,来比较RNNSearch模型和本文方法翻译结果的质量,如表3所示。在第一组中,我们发现源语言句子中存在包括时间"2017年11月12日"、人名"阮氏金银"和专有名词"一带一路"等词语,RNNSearch模型可以将时间翻译为"Ngày 12 tháng 11, 2017",但与正确的译文有差别,没有将"年"-"năm"正确翻译出来。对于人名"阮氏金银",因为在训练语料中出现的次数很少,甚至没有在训练语料中出现过,因此词表中不包含这样的人名,从而不能正常地翻译,最后通过UNK代替。专有名词"一带一路"翻译为"vành đai, đường",其为"皮带,道路"意思,与"一带一路"不一致。本文中的方法因为融入集外词词典,很好地将源语言句子中的日期、人名和专有名词翻译出来。在第二组中,源语言句子中包括地名"富茨克雷"和名词"灵枢"等词,同样由于RNNSearch模型中的词表不包含地名"富茨克雷",所以无法正常翻译,最后用UNK代替,名词"灵枢"最后翻译为"quan tài"。对于"quan tài"这个词,查找翻译其意为"棺材",与原文名词相似,可以认为翻译正确。地名"富茨克雷"的翻译,英语与越南语相同,都为"Footscray"。在第三组中,源语言句子中"人民日报""中共中央"和"联合国教科文组织"都为组织机构名。通过与译文对比,RNNSearch模型对于"人民日报"翻译有误,"中共中央"的翻译"Ủy ban trung ương CPC"中的"CPC"为"the Communist Party of China"的简写,意为"中国共产党",可理解为翻译正确。"联合国教科文组织"翻译正确。本文的方法关于"人民日报"和"中共中央"翻译与译文相同,"联合国教科文组织"翻译为"Tổ chức Giáo dục, Khoa học và Văn hóa Liên Hiệp Quốc",是按照字面直接翻译,实际上没有错误,"联合国教科文组织"英文翻译的简写为" UNESCO",在越南语中同样适用。

表3 不同模型生成摘要的比对结果

1. 源语言句子	2017年11月12日,国家主席习近平在河内会见越南国会主席阮氏金银,双方同意实施"一带一路"合作文件。
译文	**Ngày 12 tháng 11 năm 2017** gặp Chủ tịch Tập Cận Bình ở Hà Nội, quốc gia Việt Nam Chủ tịch Quốc hội **Nguyễn Thị Kim Ngân**, hai bên đồng ý thực hiện "**Một vành đai một con đường**".
RNNSearch	**Ngày 12 tháng 11, 2017**, Chủ tịch Trung Quốc Tập Cận Bình đã gặp Chủ tịch Quốc hội Việt Nam tại Hà Nội **<UNK>,**Hai bên đã đồng ý triển khai tài liệu hợp tác **vành đai, đường**.
本文方法	**Ngày 12 tháng 11 năm 2017**, Chủ tịch Quốc hội Tập Cận Bình đã gặp Chủ tịch Quốc hội Việt Nam tại Hà Nộ **Nguyễn Thị Kim Ngân,** Hai bên đã đồng ý triển khai tài liệu hợp tác "**Một vành đai một con đường**".

2. 源语言句子	几天后，我们去了富茨克雷的一座佛教寺庙，坐在她的灵柩旁。
译文	Vài ngày sau , chúng tôi đến một ngôi chùa đạo Phật ở **Footscray** và ngồi quanh **quan tài** của bà .
RNNSearch	Vài ngày sau, chúng ta đi. là một ngôi chùa Phật giáo **<UNK>**, ngồi bên **linh cữu** của cô ấy.
本文方法	Vài ngày sau, chúng tôi đến một ngôi chùa Phật giáo ở **Footscray**, ngồi cạnh **quan tài** của cô ấy.
3. 源语言句子	人民日报是中共中央机关报，被联合国教科文组织评为世界十大报纸之一。
译文	**Nhân Dân nhật báo** là cơ quan báo **Ủy ban Trung ương Đảng Cộng sản Trung Quốc** , Được **UNESCO** vinh danh là một trong mười tờ báo hàng đầu thế giới.
RNNSearch	**Báo Nhân dân** là cơ quan báo **Ủy ban trung ương CPC**, được **UNESCO** bình cho thế giới. Mười một trong những tờ báo lớn.
本文方法	**Nhân Dân nhật báo** là cơ quan báo **Ủy ban Trung ương Đảng Cộng sản Trung Quốc**, được **Tổ chức Giáo dục, Khoa học và Văn hóa Liên Hiệp Quốc** bình cho thế giới. Mười một trong những tờ báo lớn.

4.4　实验扩展

对于提出的改善集外词方法，本文进一步扩展到其他语种上进行实验，验证其他资源稀缺型语言上的翻译效果。我们选取了英越和蒙汉进行实验，语料规模为英越 20 万语料和蒙汉 26 万语料(CWMT 2018 蒙汉语料)。分类词典的构建和处理方式及方法与汉越翻译基本相同。关于双语词典，英越双语字典构建的规模为 7 642 对，蒙汉双语字典构建的规模为 10 231 对。关于实体词典，英越同样通过维基百科进行抽取。英越实体词典中人名实体数量为 6 416 对，地名实体数量为 2 873 对，组织机构名实体数量为 5 012 对，专有名词实体数量为 4 351 对，共计 18 652 对。对于蒙汉翻译，相关的工具还不够完善，我们通过人工方式对汉语语料中的实体进行查找，通过 NiuTrans[20] 中的蒙汉翻译系统来找到对应的蒙文实体翻译结果，再对蒙文训练语料进行实体词识别。蒙汉实体词典中人名实体数量为 2 857 对，地名实体数量为 2 513 对，组织机构名实体数量为 1 754 对，专有名词实体数量为 2 013 对，共计 9 137 对。关于规则词典则与汉越规则词典构建方式及方法相同。

在融入本文方法之前，对英语和越南语语料做了 Tokenize 处理。实验同样包括双向翻译，分为英语—越南语、越南语—英语、蒙语—汉语和汉语—蒙语，总共 20 组实验数据，结果如表 4 所示。

表 4　英越和蒙汉实验结果

模型	BLEU	
	英语—越南语	越南语—英语
Moses	20.72	21.43
RNNSearch(10 万)	22.43	23.54
RNNSearch(20 万)	24.71	25.63
本文方法(10 万)	25.24	26.49
本文方法(20 万)	26.72	27.84
模型	**BLEU**	
	蒙语—汉语	汉语—蒙语
Moses	9.12	9.06
RNNSearch(13 万)	10.43	10.24
RNNSearch(26 万)	12.16	12.12
本文方法(13 万)	12.97	11.68
本文方法(26 万)	14.06	13.77

从实验结果对比看，本文方法通过混合网络融入分类词典后 BLEU 值有所提升。对于相同规模的 10 万训练语料，在英语—越南语翻译方向下，本文方法比 RNNSearch 有 2.81 个 BLEU 值提升；在越南语—英语翻译方向下，本文方法比 RNNSearch 有 2.95 个 BLEU 值提升。对于相同规模的 20 万训练语料，在英语—越南语翻译方向下，本文方法比 RNNSearch 有 2.01 个 BLEU 值提升；在越南语—英语翻译方向下，本文方法比 RNNSearch 有 2.21 个 BLEU 值提升。对于相同规模的 13 万训练语料，在蒙语—汉语翻译方向下，本文方法比 RNNSearch 有 2.54 个 BLEU 值提升；在汉语-蒙语翻译方向下，本文方法比 RNNSearch 有 1.44 个 BLEU 值提升。

对于相同规模的26万训练语料，在蒙语-汉语翻译方向下，本文方法比RNNSearch有1.90个BLEU值提升；在汉语-蒙语翻译方向下，本文方法比RNNSearch有1.65个BLEU值提升。对于蒙汉的翻译结果，BLEU值整体比较低，我们进行了分析。在蒙语里，尽管有些词是通过空格进行分割的，但有很大一部分词语和句子存在整体性，彼此之间是相连的。对于这样的少数民族语言，没有对应的分词工具，一定程度上会导致BLEU值偏低。

通过以上实验结果可以看出，本文方法不仅在汉越神经机器翻译上表现出优势，在其他资源稀缺型语言上（如英越）和少数民族语言（如蒙汉）上同样提高了神经机器翻译的准确率，对于资源稀缺型语言的神经机器翻译集外词问题的处理具有可行性。

5 结论

本文对RNNSearch模型进行改进，提出一种融入分类词典的汉越混合网络神经机器翻译模型，很好地处理了集外词问题。对于给定的源语言句子，扫描分类词典以确定候选短语句对，并对其标签做标记，解码器端利用单词组件和短语组件混合网络，生成单个集外词和短语集外词，从而改善汉越神经机器翻译的性能。通过对汉越、英越和蒙汉实验结果分析，证明了该方法的有效性。下一步研究中，拟继续探索利用本体库、知识图谱等外部资源来改善低资源神经机器翻译性能的方法。

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. Seattle, 2013: 1700-1709.

[2] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 2015 International Conference on Learning Representations. USA, Diego, 2015: 865-881.

[3] Ilya Sutskever, Oriol Vinyals, Quoc V Le. Sequence to sequence learning with neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Quebec, 2014: 3104-3112.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734.

[5] Tang Y, Meng F, Lu Z, et al. Neural machine translation with external phrase memory[J]. arXiv preprint arXiv: 1606.01792, 2016.

[6] Zhang J, Zong C. Bridging neural machine translation and bilingual lexicon[J]. arXiv: 1610.01272, 2016.

[7] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, et al. Pointing the unknown words[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 140-149.

[8] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural machine translation of rare words with subword units[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1715-1725.

[9] Marta R Costa-jussa, Jose A R Fonollosa. Character-based neural machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 357-361.

[10] Wang Ling, Isabel Trancoso, Chris Dyer, et al. Character-based neural machine translation[C]//Proceedings of the 2016 International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 785-796.

[11] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv: 1609.08144, 2016.

[12] Junyoung Chung, Kyunghyun Cho, Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1693-1703.

[13] Xiaoqing Li, Jiajun Zhang, Chengqing Zong. Towards zero unknown word in neural machine translation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. New York, USA, 2016.

[14] Minh-Thang Luong, Ilya Sutskever, Quoc V Le. Addressing the rare word problem in neural machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 11-19.

[15] Sébastien Jean, Kyunghyun Cho, Roland Memisevic. On using very large target vocabulary for neural machine translation [C]//Proceedings of the Annual Meeting of the Association for Computational Lin-

guistics. Beijing, China, 2015: 1-10.

[16] Arthur P, Neubig G, Nakamura S. Incorporating discrete translation lexicons into neural machine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA, 2016: 1557-1567.

[17] Zhang J, Zong C. Bridging neural machine translation and bilingual dictionaries[J]. arXiv preprint arXiv: 1610.07272, 2016.

[18] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Washington, USA, 2014: 1724-1734.

[19] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]//Processing of the NIPS 2014 Workshop on Deep Learning. Montreal, Canada, 2014.

[20] Tong Xiao, Jingbo Zhu, Hao Zhang, et al. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation[C]//Proceedings of the Association for Computational Linguistics, Jeju Island, Korea, 2012: 19-24.

车万金(1993—)，**硕士研究生，主要研究领域为机器翻译、自然语言处理。**
E-mail: 809501313@qq.com

郭军军(1987—)，**博士，主要研究领域为自然语言处理、机器翻译、机器学习。**
E-mail: guojjgb@163.com

余正涛(1970—)，**通信作者，博士，博士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。**
E-mail: ztyu@hotmail.com