文章编号：1003-0077（2011）00-0000-00

基于深度学习的维汉口语机器翻译研究*

孔金英[1,2,3]，杨雅婷[1,2]，董瑞[1,2]，王磊[1,2]，袁扬[1,2,3]，李晓[1,2]

（1. 中国科学院 新疆理化技术研究所, 新疆 乌鲁木齐 830011；2. 新疆民族语音语言信息处理重点实验室, 新疆 乌鲁木齐 830011；3. 中国科学院大学, 北京 100049）

摘要：深度神经网络是深度学习技术的代表，是一种模拟人脑的计算模型，已经在人工智能的很多领域取得了突破性的成果。基于神经网络的机器翻译是利用深度神经网络进行源语言到目标语言转换的技术，与传统的统计机器翻译技术有着很大的差别。虽然基于神经网络的机器翻译结构十分简单，但是其达到的翻译效果已经接近或者超过传统的基于短语的机器翻译技术。本文提出了一种基于神经网络的维吾尔语到汉语的机器翻译方法。通过实验证明，在特定的短文本口语翻译领域，本文提出的神经网络机器翻译方法的结果优于传统的基于短语的机器翻译技术。

关键词：神经网络；机器翻译；维吾尔语

中图分类号：TP391 **文献标识码：**A

Research of Deep learning for Uyghur-Chinese oral Machine Translation

Kong Jinying[1,2,3], Yang Yating[1,2] ,Dong Rui[1,2], Wang Lei[1,2], Yuan Yang[1,2,3],Li Xiao[1,2]

(1.Unit ,city, province zip code, China ; 2. Unit, city, province zip code ,China)

Abstract: Deep neural network, which is the representative of deep learning technology has made breakthrough in many fields of artificial intelligence, is a computational model that simulates human brain. The Machine Translation technology based on neural network transfer source language to target language by using deep neural network, has big difference to traditional statistical Machine Translation technology. Although structure of neural Machine Translation model is quite simple, the performance of it is close to or better than the traditional phrase based Machine Translation technology. This paper presents an approach to Machine Translation based neural network in the Uyghur-Chinese translation. Through experiments, it is proved that the neural network machine translation presented by this paper is better than the traditional phrase based Machine Translation technology in the specific field of short spoken translation.

Key words: neural network; machine translation; Uyghur

1 引言

近年来，随着”一带一路”战略的施行，新疆作为该战略的桥头堡功能愈发突显。新疆世代是多民族聚居的地方，该地区目前生活了 1500 万以上的维吾尔族同胞，这使得维吾尔语到汉语的机器翻译需求尤为迫切。维吾尔语与汉语在语法上有着巨大的差异，同时维吾尔语语料资源较为匮乏，导致通用的统计机器翻译的方法很难取得令人满意的结果[1]。基于以上原因，维汉机器翻译逐渐成为了国内自然语言处理领域的一个研究热点。

* **收稿日期：** **定稿日期：**

基金项目：新疆自治区青年科技创新人才培养工程项目（2014721032，2015211B034）；新疆自治区自然科学基金（2015211B034）；新疆维吾尔自治区重点实验室资金（2015KL031）；中科院战略性先导科技专项-新疆少数民族信息处理（XDA06030400）

作者简介：孔金英（1988——），男，博士研究生，主要研究自然语言处理和机器学习；杨雅婷（1985——），女，副研究员，主要研究自然语言处理；董瑞（1985——），男，博士研究生，主要研究自然语言处理；王磊（1975——），男，研究员，主要研究自然语言处理，袁扬（1986——），男，博士研究生，主要研究自然语言处理；李晓（1957——），男，研究员，主要研究自然语言处理。。

机器翻译是利用计算机将源语言自动转化成目标语言的一种技术，目前使用的比较多的方法是基于统计学的方法。其中，基于短语的机器翻译系统是现在性能最好的机器翻译系统之一。维汉机器翻译作为机器翻译领域中的一个分支，有着显著的自身特点。维吾尔语具有复杂的语言形态，其形态变化包括曲折（infection）、派生（derivation）、复合（composition）等。相对而言，汉语几乎没有形态变化。当前，国际上机器翻译的研究对象比如英语、德语、汉语等都是形态变化相对比较简单的语言，可以不用考虑词语的形态变化，直接把不同词形的词都当成独立的词语进行翻译。但对于维吾尔语而言，这样会导致译文中出现大量的未登录词(out of vocabulary, OOV)。另一方面，维吾尔语是主宾谓语言，汉语是主谓宾的语言，两种语言在这方面的差异导致了维汉机器翻译的调序问题尤为突出。

维汉口语机器翻译指的是利用计算机对口语化的维吾尔语进行自动翻译。口语一般指的是短信、微博等文本。维汉口语机器翻译相比较新闻政务等书面化语的机器翻译有着很大的不同。首先，维汉口语平行语料较为难获得，比如新闻政务领域的平行语料可以从网站上直接爬取，但是网络上的口语平行语料资源却极度匮乏，质量也很低。此外，口语中的语言随意性很强，日常的口语中时常夹杂着各种语法的谬误和拼写的错误。还需要强调的是，口语中的新词很多，可能会参杂有其他的语言。最后，口语一般都是短文本，一句话的单词数量很少会超过20个。

本文提出的维汉口语机器翻译方法的工作流程大致如图1所示。首先，我们对口语语料中的维吾尔语部分进行词干词缀的切分，这样可以有效的减少未登录词。然后，我们对切分后的维吾尔语部分进行拼写检查，得到基本拼写正确的维吾尔语单词。接下来，我们利用切分过后的维吾尔语和汉语训练一个神经网络机器翻译模型。输入一个经过了词干词缀切割和拼写检查后维吾尔语口语句子，我们就可以利用该翻译模型解码出相应的汉语 n-best 译文。最后综合语言模型得分和翻译模型得分输出最优的结果。

本文是首次将维汉口语机器翻译作为一个单独的方向进行阐述并提出解决方案的文献。另外，本项研究工作也是利用神经网络机器翻译技术解决维吾尔语到汉语机器翻译的初次探索。

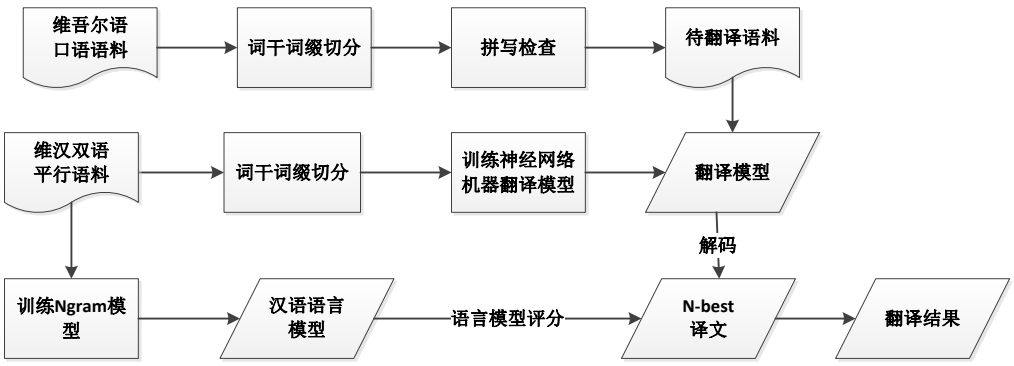

图 1 基于神经网络的维汉口语翻译模型

2 相关工作

目前，对于维汉机器翻译的研究都针对的是新闻政务领域。新闻政务的语言符合语言学规范，句式较为固定，未登录词比较少，能够达到不错的翻译效果。根据CWMT2015年的非受限维汉新闻政务领域的机器翻译评测，最好成绩的BLEU-5值达到了46.93[2]。

BLEU(Bilingual Evaluation Understudy)值是目前采用最广的机器翻译领域中用来自动评价机器翻译准确度的标准，2002年由 IBM 的研究人员提出，是一种基于 N-Gram 匹配的自动评测方法[3]。该方法认为，

机器翻译译文的质量可以通过译文与专业人工翻译结果的相似度进行评价，相似度越高翻译质量越好。N 元匹配是指，机器翻译译文与人工翻译参考译文（数量可以不唯一）之间任意连续 N 个单词完全相同。统计译文之间 N 元词组匹配的数量，并除以机器翻译译文的 N 元词汇总数，得到 N 元匹配的精确度 P_n ，并针对译文中大量重复词条进行剪切修正。由于 N 元统计精度随着阶数的提高呈现指数递减，所以最后的评价公式采用几何平均形式求平均值后加权，再乘以长度惩罚因子 $BP=\begin{cases}1 & ;\ if\ c\geq r\\ e^{1-\frac{r}{c}} & ;\ if\ c<r\end{cases}$，最后的评价如式 1 所示。

$$Score=BP\times\exp\left(\sum_{n=1}^{N}\frac{1}{N}\log P_n\right) \quad (1)$$

当 N 取 1 时，着眼于译文的忠实度，当 N 取大于 1 的整数时，着眼于译文的流畅度，因此 BLEU 方法很好的考虑到译文的忠实度和流畅度。

中科院新疆理化技术研究所采用的维汉机器翻译技术主要是以层次短语翻译模型为基础[4]。顾名思义，层次短语是指短语中的短语，相对短语模型来看层次短语模型具有更好的泛化和远距离调序能力。层次短语翻译模型将短语（连续的单词，非语言学）作为机器翻译的最小粒度，在规则的抽取过程中除了普通短语规则之外，还抽取了变量规则用来处理长距离调序。新疆大学在基于实例的维汉机器翻译领域的耕耘也是颇有建树的。此外上海交通大学的卡哈尔江尝试过构建基于规则的维汉机器翻译系统[5]。

深度神经网络是一种模仿人脑的计算模型，是深度学习的主要方法，由最少三层神经网络组成。因为其在图像识别和语音识别等领域有着突破性的进展，深度神经网络已经成为了目前人工智能领域最为热门的研究方向之一。在自然语言处理领域，也有不少对深度神经网络应用的研究取得了很好的效果。

神经网络机器翻译(NMT)是由 cho 在 2014[6]年提出的，其结构十分简单，由一个深度神经网络组成的编码器（encoder）和一个深度神经网络组成的解码器（decoder）组成。Encoder 接收输入的句子，decoder 则产生输出的译文。本文采用的是Bahdanau于2015[7]年提出的基于注意力（attention）的神经网络机器翻译模型，该模型是目前性能最好的神经网络机器翻译模型之一。Bahdanau 模型中的注意力是基于同目标单词有关的源语言单词。

此外，还有很多 NMT 上的工作。比如将基于注意力的NMT模型当成短语模型中的特征[8]，与其他特征一起经过调参训练后协同进行解码，使用最小风险（MRT）对神经网络进行调参[9]，这些工作得到的译文质量比传统的短语模型有很大的提高。

3 神经网络维汉口语机器翻译模型

神经网络机器翻译模型是一个（序列到序列）sequence-to-sequence 模型，其利用一个 encoder 接收输入的 sequence，然后利用 decoder 产生输出的 sequence。根据 Luong 的研究表明[10]，神经网络机器翻译模型的学习能力与传统的统计翻译模型相比，因为没有严格的词对齐模型，所以神经网络翻译模型的泛化能力更强。不过需要注意的是，虽然加入了注意力后的神经网络机器翻译模型对长文本翻译准确度有了提升，但是性能跟目前最好的短语翻译模型相比还是有差距，弱点主要体现在词对齐上。不过，口语翻译是短文本翻译，口语中较少出现超过二十个单词的句子。使用神经网络模型解决长句子翻译问题不仅结果不令人满意且训练的时间也难以让人接受，但在解决短文本翻译时却得心应手。综上所述，神经网络翻译模型适合使用在某些特定的场合，比如口语的翻译。所以本文采用神经网络机器翻译模型解决维汉口语翻译问题。

本节首先对前处理工作进行描述。口语语料中充斥的各种错误，所以必须进行前处理。接下来，本节具体阐述如何利用标准的新闻政务领域语料训练一个基于注意力的神经网络机器翻译模型。最后，本节对使用训练好的翻译模型进行解码时的后处理工作进行描述。

3.1 前处理

维吾尔语是一种黏着性语言，词干和词缀组合后形成的新词层出不穷，对维吾尔语进行机器翻译时，词干词缀一般都是必须的工作。本文在前处理中利用中科院新疆理化技术研究所杨雅婷提出的方法对维吾尔语进行词干和词缀的切割[11]。接下来，本文使用标准单词表和编辑距离对各个维吾尔语单词进行匹配，替换拼写错误的单词，进行简单的拼写检查。

本文使用杨雅婷 2014 年提出的方法，将词干词缀的识别看成是序列标签问题，用条件随机场模型进行标签的切割。有效地词干词缀的切割可以显著的减小维汉机器翻译中的数据稀疏问题，进一步而言，可以改善翻译模型的质量[12]。比如图 2 中的例子，‘你们不能进行标准化吗’。如果不进行切分，那么这个双语句对的知识将很难在翻译过程中使用。而在对维吾尔语单词进行了词干词缀切割后进行机器翻译，我们可以利用更小粒度的对齐，更准确的翻译知识。

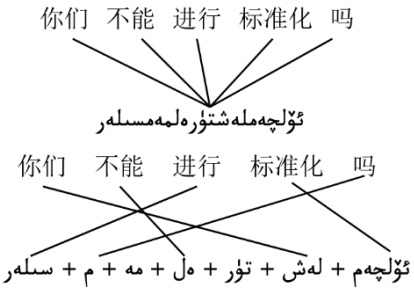

图 2 词干词缀切割例子

另一方面，维吾尔语口语中的拼写错误十分常见，这是由于维吾尔语单词是由多个字母组成的，人们在使用输入法的时候难免出现拼写错误的情况。汉语中的拼写错误，则多是由同音字造成。由于本文研究的对象是维汉机器翻译，所以我们的拼写检查对象是维吾尔语。

本文使用简单的Levenshtein距离和维吾尔语标准词典的方法对维吾尔语单词进行拼写检查。Levenshtein 距离又名最小编辑距离，是一个衡量两个单词相似度的标准，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。Levenshtein 距离在字母构成单词的拼写检查中很有效果。算法 1 是利用Levenshtein距离进行拼写检查的具体思想，输入的标准词典我们利用其他领域（比如新闻政务领域）的标准语料进行提取。该算法保留 10 个与错误单词最近的单词，最后使用 N-gram 模型得分择优选取拼写正确的单词。

算法 1 使用 Levenshtein 距离进行拼写检查

```
Input: sentence={word_1, word_2, word_3, ···, word_i},
dic={word_1, word_2, ···, word_k}
Output: sentence
1:Wordlist=Map[10]
2:for word in sentence :
3:   if dic.has(word) :
4:      continue;
5:   else :
6:      for comparison in dic :
7:         distance=
8:          Levenshtein.distance(compasison, word)
9:         if Wordlist.notFull():
10:            Wordlist[compasison]=distance
11:        else:
12:            key = findMaxValue(Wordlist)
13:            if  Wordlist.getValue(key)>distance:
14:                  delete(Wordlist[key])
15:                  Wordlist[comparison]=distance
16:            else: continue;
17:   for checker in Wordlist:
18:      word=maxNgramCount(checker, sentence)
19:return sentence
```

3.2 神经网络机器翻译模型

传统的短语翻译模型由很多不同的模块构成，与之出入的是，神经网络机器翻译（NMT）模型尝试直接用深度神经网络解决机器翻译问题。神经网络翻译模型由一个编码器 encoder 和一个解码器 decoder 组成。编码器和解码器一般都是由循环神经网络（RNN）组成的，RNN 是 Hopfield 在 1982 年提出的一种能够处理时序问题的循环神经网络[13]。图 3 是一个 RNN 网络按时间步骤展开的示意图。由图 3 可见，按时间序列展开的 RNN 就是一个（MLP）前向神经网络。换言之，RNN 就是一个隐藏层带有回路的前向神经网络。一个典型的 RNN 网络将输入层的信息传入由记忆单元组成的隐藏层，隐藏层根据上一次的输入产生的隐藏层状态和这次输入产生输出，同时隐藏层记录下这一次的隐藏层状态。RNN 在时刻 t 依照式 2 和式 3 计算得到输出。

$$h^{(t)} = \sigma(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h) \quad (2)$$

$$y^{(t)} = soft\max(W^{hx}h^{(t)} + b_y) \quad (3)$$

式 2 中的$\sigma()$指的是激活函数，$h^{(t)}$指的是在时间 t 的隐藏层状态，$x^{(t)}$是从输入层输入的词向量，W^{hx}是输入层与隐藏层连接的参数矩阵，W^{hh}则是 t-1 时的隐藏层与 t 时刻的隐藏层连接的参数矩阵，b_h是容许每个结点学习一个偏移量的偏置参数。式 3 中的$y^{(t)}$是输出的各个类别的概率得分向量，softmax（）是一个用来多分类的函数，W^{yh}指的是时刻 t 的隐藏层状态与输出层连接的参数矩阵，b_y也是一个偏置参数。

从式 2 和式 3 的定义我们可以看出，一个 RNN 在 t 时刻的输出是由 t-1 时刻的隐藏层状态和当前的输入共同决定。换言之，RNN 可以记录之前的序列信息，并根据之前的序列信息和当前的输入产生更加合理的输出。RNN 的这个特性正好可以用来解决机器翻译问题

根据 NMT 中所采用的 RNN 中的每个神经单元的不同，NMT 的性能也会大相径庭，目前在 NMT 中应用的比较普遍的 RNN 网络有 LSTM 和 BiRNN。本文采用的 NMT 模型是 Bahdanau 在 2015 的工作。

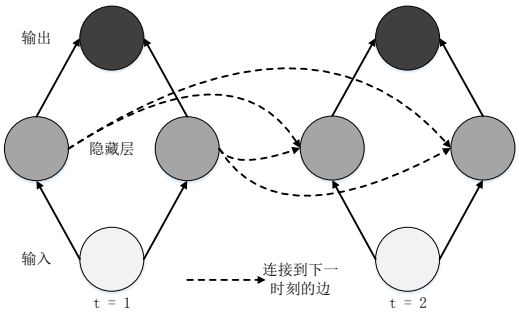

图 3. RNN 按照时间步骤展开的示意图

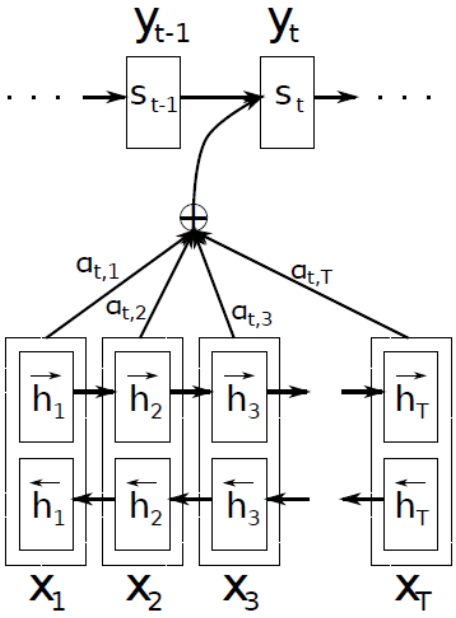

图 4. 基于注意力的 NMT 的工作流程

本文所采用的神经网络机器翻译模型的工作流程如图 4 所示：编码器在输入层按照源语言句子的顺序依次输入源语言中的各个经过 Embedding 的单词向量。当读到句子结束时，解码器中的隐藏层按照源语言句子和编码器隐藏层的状态产生输出并记录下这时的隐藏层状态。解码器中的输出层根据隐藏层状态的输出和源语言的相关信息产生出各个目标单词的概率。最后，基于各个时序产生的目标单词概率，使用柱搜索算法找出概率得分最大的译文。

接下来，本文将在接下来的两节里分别阐述解码器和编码器的构成。

3.2.1 编码器

研究表明，编码器以逆向的顺序读取单词的能够得到更好的翻译结果[14]。因此本文使用（双向循环神经网络）BiRNN 网络作为编码器。BiRNN 工作流程如下：

a. 编码器顺序读取词向量，利用式 4 计算并记录下各个单词在 时刻的隐藏层状态$\overrightarrow{h_t}$。

b. 编码器以逆序的方式读取词向量，利用式 4 计算并记录下各个单词在 时刻的隐藏层状态$\overleftarrow{h_t}$，计算方式与顺序的相同。

$$\overrightarrow{h_t} = f(x^{(t)}, \overrightarrow{h_{(t-1)}}) \quad (4)$$

c. 计算 BiRNN 编码器 t 时刻的状态。BiRNN 在 t 时刻的隐藏层状态的计算方法如式 5 所示。

$$h_t = \begin{bmatrix} \overrightarrow{h_t} \\ \overleftarrow{h_t} \end{bmatrix} \quad (5)$$

d. 依次执行 a, b, c 步，直到输入结束，保存每个时刻的 h_t 。

当编码器读完所有的源语言句子后，我们可以得到一系列编码器状态 $\{h_1, h_2, \ldots, h_{tx}\}$。一般的神经网络翻译系统会将最后得到的定长的向量 h_{tx} 当作源语言的全部信息，并将其输入到解码器中。而基于注意力的神经网络翻译会保存 $\{h_1, h_2, \ldots, h_{tx}\}$，并利用其计算出软对齐模型，进而选取与目标单词 i 相关性较大的那些源语言单词 j 进行输入。

3.2.2 解码器

本文使用的神经网络机器翻译模型是基于注意力的，因此解码器的输入并不是固定的将整个源语言句子的信息压缩到一起的句子向量，而是动态的输入对应输出单词相关性较大的那些源语言单词向量 c_i。在基于注意力的神经网络机器翻译中，目标句子中的单词与源语言句子中的哪些单词相关性较大是由对齐模型决定的。

神经网络机器翻译中的对齐模型有别于传统的统计机器翻译中的对齐模型，是一种软对齐模型。式 6 表示的是对齐模型的定义。

$$e_{ij} = a(s_{i-1}, h_j) \quad (6)$$

e_{ij} 表示的是目标语言中的 i 单词与源语言中的 j 单词的匹配度。$a()$ 表示的是一个非线性的函数（也可以是一个前馈神经网络）。$a()$ 由 s_{i-1} （$i-1$时刻 RNN 的隐藏层状态)和 h_j（源语言句子中的第 j 个词向量)共同产生输出，可以和机器翻译系统中其他的部分一起训练得到的。（该问题可以看成是已知 s_{i-1} 求解类别是 h_i 的概率）。

有了对齐模型，我们可以得到上下文有关向量 c_i。c_i 的计算方式可以由式 7 和式 8 共同算出。

$$c_i = \sum_{j=1}^{T_X} a_{ij} h_j \quad (7)$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (8)$$

除了上下文有关向量的计算方式与其他的神经网络相比有很大的区别，目标单词向量 y_i 的概率得分的计算方式则大同小异。

$$s_i = \mathrm{f}(s_{i-1}, y_{i-1}, c_i) \quad (9)$$

根据式 9，我们可以由 i-1 时刻的隐藏层状态 s_{i-1} 和上一个目标词向量以及上下文有关向量 c_i 计算得到 i 时刻的隐藏层状态 s_i， f 是一个非线性的函数。

最后，我们可以由 i 时刻的隐藏层状态 s_i，和上一个单词以及上下文有关向量 c_i 得到 y_i 单词的概率得分。

$$p(y_i \mid y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i) \quad (10)$$

式 10 中的 $g()$ 一般是非线性函数，也可以是多层神经网络。值得留意的是，在一般

NMT 中解码器中的隐藏状态的初始值 s_0 是编码器中的隐藏状态的最终值。

直观上可见，引入软词对齐模型作为注意力，同将源语言句子压缩成一个定长向量的一般神经网络翻译方法相比，大大减少了将源语言句子中所有信息同时编码的负担。Bahdanau 的实验也证明,这种方法也可以有效的提升最终的翻译结果。

3.3 后处理

上一节，本文描述了 NMT 中编码器和解码器的工作原理，在解码器输出各种目标词向量的相应概率后，我们可以得到一个有限的解空间(充满了各种可能的翻译)。如果我们利用柱搜索（beam search）保留每次译文概率得分的前 n 项，我们就可以得到相应的 n-best 译文。众多研究表明，对 n-best 译文进行有效的重排序可以提升最后的译文质量，而这也是本节所要阐述的内容。

神经网络翻译模型只是利用深度神经网络记录所有的翻译知识，没有使用附加的模型辅助产生译文。哈工大的梁华参[15]曾提出用语言模型困惑度来过滤语料，并且证实了这是一种行之有效的办法。在他的启发下，本文尝试使用目标语言的语言模型对 n-best 译文进行困惑度打分,按照困惑度得分的高低对 n-best 译文进行重排序。使用这种方法添加语言模型具有更好的鲁棒性和可操作性。

困惑度是衡量句子流畅性的指标。困惑度 pp 原本是衡量一个语言模型优劣的准则，是建立在交叉熵的基础上的。反过来，如果我们用相同的语言模型对不同的句子进行困惑度打分，这样就可以衡量各个句子的流畅程度。交叉熵的数学定义如式 11 所示，困惑度的定义如式 12 所示。:

$$H(p_{LM}) = -\frac{1}{n}\sum_{i=1}^{n}\log p_{LM}(w_i \mid w_1,\ldots,w_{i-1}) \tag{11}$$

$$pp = 2^{H\left(P_{LM}\right)} \tag{12}$$

需要留意的是，式 11 和 12 中的 P_{LM} 是语言模型 N-gram 概率得分。最后，按照式 13，我们综合各个候选译文的翻译概率得分和困惑度得分对 n-best 译文进行重排序，选取最优的译文。

$$Score_i = \lambda_m \log p_i + \lambda_n \log pp_i \tag{13}$$

图 5 是利用语言模型困惑度进行 n-best 候选译文重排序的流程图。

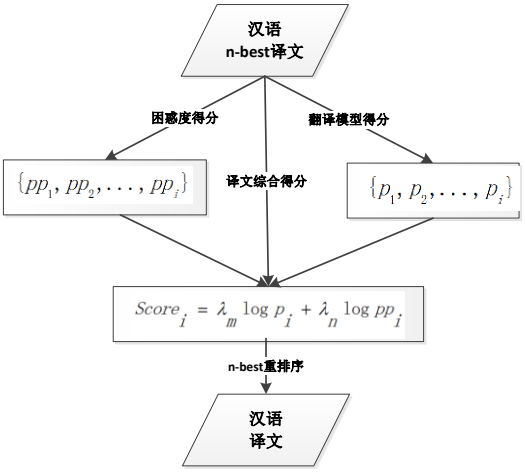

图 5 利用困惑度进行译文重排序

3.4 实验

3.4.1 实验设置

针对本文提出的维汉口语神经网络机器翻译方法，我们进行了实验的论证。本实验使用的语料分成两部分，一部分是新闻政务领域的双语语料，另一部分是少量的短信口语语料，具体情况如表 1 所示。因为短信口语的平行语料是由专门的翻译人员进行翻译完成的，所以数量较少，不足以作为训练集。我们的训练集是从新闻政务领域的网站天山网爬取而来。

表 1 语料大小情况

类别	训练集	开发集	测试集一	测试集二
新闻政务	500，000	------	----------	----------
短信口语	----------	1,000	1,000	1,000

本文使用中国科学院新疆理化技术研究所基于杨雅婷的方法实现的词干词缀切

割工具对维吾尔语进行切分。另外，我们使用开源的python-Levenshtein 0.10.2对词进行拼写检查。本文的神经网络翻译模型是用基于 Python 的 Theano 库开发的groundhog实现的。实验中encoder的BiRNN隐藏层的神经单元个数是 1000,decoder 中LSTM的隐藏层神经单元个数也是1000个。为了减少未登录词，本文使用的汉语词汇表大小是300000,维吾尔语词汇表是450000。此外，本文选用 adadelta 算法作为优化参数的算法，随机正则化噪音项设置为0.01，实验选择的 batch 的大小设置为 60。我们使用 early stop 算法对整个神经网络进行训练。

关于对比的短语翻译系统，本文使用的是MOSES 2.1 ，操作系统是ubuntu12.04。本文使用GIZA++ 开源工具包作为词对齐工具，然后采用“grow-diag-final-and”策略获得多对多的词语对齐。本文的短语抽取限制的长度是7。在调参过程，本文使用的是最小错误训练方法[16]优化模型的参数。另外，本文使用 SRILM 工具分别对训练集里的汉语语料进行五元语言模型的训练，并用Kneser-Ney平滑估计参数。

本文采用了对大小写不敏感的 BLEU-4作为机器翻译最终结果的评测指标。本文的实验分成八个小组，分别是：

Baseline:使用传统的短语翻译模型分别对测试集进行翻译，没有做任何变动。

PRE+PMT:在训练集和测试集进行前处理，然后使用传统的短语翻译模型对分别对测试集进行翻译。

PMT+POST:使用传统的短语翻译模型和后处理方法相结合，分别对测试集进行翻译。

PRE+PMT+POST:在训练集和测试集进行前处理，然后使用传统的短语翻译模型和后处理方法相结合，分别对测试集进行翻译。

NMT：使用基于注意力的神经网络机器翻译模型对测试集进行翻译。

PRE+NMT: 在训练集和测试集进行前处理， 使用基于注意力的神经网络机器翻译模型对测试集进行翻译。

NMT+POST: 使用基于注意力的神经网络机器翻译模型和后处理方法相结合，分别对测试集进行翻译。

PRE+NMT+POST: 在训练集和测试集进行前处理， 使用基于注意力的神经网络机器翻译模型和后处理方法相结合，分别对测试集进行翻译。

3.4.2 实验结果

表2是本文的实验结果，表2中的数值的单位都是BLEU-4。

表2 维汉机器翻译系统的实验结果

序号	组别	开发集	测试集1	测试集2	平均
1	Baseline	20.02	18.46	17.36	18.61
2	PRE+PMT	21.24	19.56	18.43	19.74
3	PMT+POST	20.12	18.45	17.46	18.68
4	PRE+PMT+POST	21.37	19.89	18.63	19.96
5	NMT	21.43	20.03	19.22	20.23
6	PRE+NMT	23.56	23.05	22.21	22.94
7	NMT+POST	21.56	20.45	19.38	20.46
8	PRE+NMT+POST	**23.75**	**23.23**	**22.40**	**23.13**

由于训练集是新闻政务领域的语料，而开发集和测试集是短信口语领域的语料，两个领域的相差性较大，所以本文实验的总体BLEU值都不高。根据表3的结果，我们可以得到以下结论。

单纯比较第1组和第5组，我们发现第5组的 BELU 值都要略高于第一组。这说明在跨领域的机器翻译中，使用神经网络翻译

模型的译文结果要优于短语翻译模型。这是因为神经网络翻译模型有着更强的领域泛化性。分别对比 1，2，3，4 组和 5，6，7，8 组，我们可以看到 BLEU 值是依次缓慢增加的。这也反应了本文提出的前处理和后处理方法是行之有效的。分别对比 1，2 组和 5，6 组，我们发现前处理的效果很明显，可见维汉机器翻译中的词干词缀切割工作十分有必要。因为维吾尔语是黏着语，词汇量十分巨大，有效的词干词缀切割可以减少神经网络翻译模型中的未登录词数量。此外，5，6 组 BLEU 值的提升要大于 1，2 组，这也佐证了神经网络模型中的未登录词要多于短语翻译模型。对比 5，7 组，因为后处理中，困惑度和翻译模型得分的权重确定带有偶然性，所以后处理方法对译文质量的提升并不十分明显，后续可考虑使用启发式的方法来确定权重。最后，我们综合对比 1，8 组，发现 BLEU 值的提升是十分明显的，这也证明了本文的工作是有意义的。

通过观察具体的译文，我们发现神经网络翻译系统的译文风格和短语翻译系统有着很大的区别。神经网络翻译模型得到的译文的流畅性要明显好于短语翻译模型，但是未登录词比短语翻译模型要多，而且短语翻译的准确性要略差于短语翻译模型。

4 结论和展望

口语机器翻译属于机器翻译中的一个分支。本文有别于传统的统计机器翻译方法，使用了一种基于注意力的神经网络机器翻译模型对维吾尔语口语进行了汉语的翻译，并取得了较为理想的效果。本文是使用神经网络模型解决维吾尔语口语翻译的初次探索，可以提供给后续研究者不错的参考价值。如图 6 是使用本文提出的神经网络机器翻译方法进行维吾尔语短文本翻译的例子。我们可以看到，虽然输出的译文与参考译文不是完全一样，但是流畅性较高，且翻译出来的译文意思完全相同。这是因为神经网络翻译模型没有刻意去追求源语言和目标语言中词或短语在翻译时的一一对应，而更注重语言意义的整体相似性。

需要注意的是，本文在训练神经网络机器翻译模型耗费了两周左右的时间。可见，神经网络机器翻译对计算资源要求比较高，这也使得对神经网络机器翻译的研究设置了较高的门槛。如何更高效的使用计算资源进行维汉神经网络机器翻译，将会是我们接下来的研究方向。另外，我们也会尝试将维吾尔语特性添加到基于注意力的神经网络机器翻译的研究。

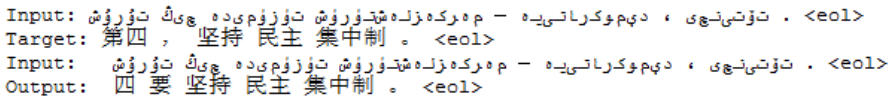

图 6 神经网络机器翻译的例子

参考文献

[1]李晓，蒋同海，周喜,等. 维汉机器翻译关键技术研究概述[J]. 网络新媒体技术，2016,5(1):19-25.

[2]汪昆，姜文斌，杨海彤，向露，赵红梅，第十一届全国机器翻译研讨会（CWMT 2015）评测报告,CWMT2015,2015 年 9 月 24-25 日，合肥

[3] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002:311-318

[4]David Chiang. 2007. 'Hierarchical phrase-based translation'.J. Computational Linguistics(2007), 33(2):201-228.

[5]卡哈尔江·阿比的热西提. 基于实例的汉维—维汉双向机器翻译系统的研究[D]. 上海交通大学,2012.

[6] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[7] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[8]Junczys-Dowmunt M, Dwojak T, Sennrich R. The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT[J]. 2016.

[9] Shen S, Cheng Y, He Z, et al. Minimum Risk Training for Neural Machine Translation[J]. Computer Science, 2015.

[10] Minh-Thang Luong, Ilya Sutskever,Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba.

Addressing the rare word problem in neural machine translation.
In Proceedings of ACL, 2015.
[11]Yang Y, Mi C, Ma B, et al. Character Tagging-Based Word Segmentation for Uyghur[M]// Machine Translation. Springer Berlin Heidelberg, 2014:61-69.
[12] 米莉万•雪合来提, 刘凯, 吐尔根•依布拉音. 基于维吾尔语词干词缀粒度的汉维机器翻译[J]. 中文信息学报, 2015, 29(3):201-206.
[13]Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the National Academy of Sciences, 1982, 79(8):2554-8.
[14]Lipton Z C, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning[J]. Computer Science, 2015
[15] 梁华参.基于短语的统计机器翻译模型训练中若干关键问题的研究[D].博士.哈尔滨：哈尔滨工业大学.2013.1-2
[16] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation [A]. Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics[C]. July 2003. 160-167.

作者联系方式：姓名　地址 邮编　电话（最好手机） 电子邮箱