软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(12):3797−3807 [doi: 10.13328/j.cnki.jos.005889]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合句法解析树的汉-越卷积神经机器翻译*

王振晗[1,2], 何建雅琳[1,2], 余正涛[1,2], 文永华[2], 郭军军[1,2], 高盛祥[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](云南省人工智能重点实验室(昆明理工大学),云南 昆明 650500)

通讯作者: 余正涛, E-mail: ztyu@hotmail.com

摘 要: 神经机器翻译是目前应用最广泛的机器翻译方法,在语料资源丰富的语种上取得了良好的效果.但是在汉语-越南语这类缺乏双语数据的语种上表现不佳.考虑汉语和越南语在语法结构上的差异性,提出一种融合源语言句法解析树的汉越神经机器翻译方法,利用深度优先遍历得到源语言的句法解析树的向量化表示,将句法向量与源语言词嵌入相加作为输入,训练翻译模型.在汉-越语言对上进行了实验,相较于基准系统,获得了 0.6 个 *BLUE* 值的提高.实验结果表明,融合句法解析树可以有效提高在资源稀缺情况下机器翻译模型的性能.

关键词: 神经机器翻译;资源稀缺;句法解析树

中图法分类号: TP18

中文引用格式: 王振晗,何建雅琳,余正涛,文永华,郭军军,高盛祥.融合句法解析树的汉-越卷积神经机器翻译.软件学报,2020, 31(12):3797−3807. http://www.jos.org.cn/1000-9825/5889.htm

英文引用格式: Wang ZH, He JYL, Yu ZT, Wen YH, Guo JJ, Gao SX. Chinese-Vietnamese convolutional neural machine translation with incorporating syntactic parsing tree. Ruan Jian Xue Bao/Journal of Software, 2020,31(12):3797−3807 (in Chinese). http://www.jos.org.cn/1000-9825/5889.htm

Chinese-Vietnamese Convolutional Neural Machine Translation with Incorporating Syntactic Parsing Tree

WANG Zhen-Han[1,2], HE Jian-Ya-Lin[1,2], YU Zheng-Tao[1,2], WEN Yong-Hua[2], GUO Jun-Jun[1,2], GAO Sheng-Xiang[1,2]

[1](School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

[2](Yunnan Key Laboratory of Artificial Intelligence (Kunming University of Science and Technology), Kunming 650500, China)

Abstract: Neural machine translation is the most widely used machine translation method at present, and has sound performance in languages with rich corpus resources. However, it does not work well in languages that lack of bilingual data, such as Chinese-Vietnamese. Taking the difference in grammatical structure between different languages into consideration, this study proposes a neural machine translation method that incorporates syntactic parse tree. In this method, a depth-first search is used to obtain the vectorized representation of the syntactic parse tree of the source language, and the translation model is trained by embedding the obtained vectors and the source language embedding as inputs. This method is implemented on Chinese-Vietnamese, language pair and achieves 0.6 *BLUE* values improvement compared to the baseline system. This experiment shows that the incorporating syntax parse tree can effectively improve the performance of the machine translation model under the resource scarcity.

Key words: neural machine translation; low-resource; syntactic parse tree

* 基金项目: 国家自然科学基金(61732005, 61672271, 61761026, 61866020); 云南省自然科学基金(2018FB04); 云南省省级人才培养计划项目(KKSY201703005, KKSY201703015)

Foundation item: National Natural Science Foundation of China (61732005, 61672271, 61761026, 61866020); National Natural Science Foundation of Yunnan Province (2018FB04); Personal Training Project of the Yunnan Science and Technology Department (KKSY201703005, KKSY201703015)

收稿时间: 2019-04-24; 修改时间: 2019-06-05, 2019-07-20; 采用时间: 2019-09-09

神经机器翻译是 Sutskever 等人[1]在 2014 年提出的一种机器翻译方法,目前主流的神经机器翻译模型都采用编码器-解码器的架构.首先,利用双语平行语料分别生成源语言与目标语言的词表,根据双语词表生成双语数据的向量化表示.通过编码器将代表源语言的向量编码成隐藏向量表示,再利用解码器将该隐藏向量信息解码成目标语言,编码器和解码器之间一般通过注意力机制(attention mechanism)连接,通过不断训练神经网络从而得到源语言映射到目标语言的翻译模型.

目前,神经机器翻译主要有基于循环神经网络(recurrent neural network,简称 RNN)的神经机器翻译模型[2−5]与基于卷积神经网络(convolutional neural network,简称 CNN)的神经机器翻译模型[6−11].在双语数据资源丰富的条件下,通过以上方法训练所得到的神经机器翻译模型均能获得很好的效果.但是针对汉语-越南语这类双语数据较少的资源稀缺型语言来说,翻译效果并不理想.为解决以上问题,本文提出融合源语言句法解析树的神经机器翻译方法.该方法首先对源语言进行句法解析,得到源语言的句法解析树;然后利用深度优先遍历,获得源语言句子中每个单词对应的句法标签序列;在神经网络的编码器端,再将以上得到的标签序列与源语言词嵌入向量及位置嵌入向量拼接;最后,通过全连接网络将拼接后的向量转化为固定长度的向量,作为训练神经网络的输入.考虑到句法解析树所具有的层次化结构特征,在模型的选择上,我们采用多层卷积神经网络作为编码器,这样更容易使模型学习获得源语言句法树的注意力信息.该方法有利于捕捉编码过程中自然语言与语法的依赖关系.实验结果表明:相比基准系统,本文所提方法能有效提高机器翻译模型的质量.

1 相关工作

近年来,国内外研究学者针对资源稀缺的汉语-越南语机器翻译方法开展了许多研究,并取得了一定进展.在汉-越双语平行语料获取方面,Trinh 等人[12]研究通过汉-越双语网站收集双语文本的方法,采用该方法,可以从双语网站中获得大量的汉-越可比语料,并提供了 JSOUP 开源库.我们可以利用得到的汉-越可比语料抽取汉-越平行语料,为汉-越机器翻译研究工作提供基础.Tran 等人[13]对汉-越双语分词方法进行了研究,基于命名实体、共享词汇、词级别对齐结果和字符级别对齐这 4 个因素进行汉语和越南语的分词,以加强汉语和越南语词语之间一对一的对齐,并限制了未登录词的数量,提升了汉-越机器翻译的性能.Huu 等人[14]提出了融合发音特征的汉-越统计机器翻译方法,借助汉语与越南语拼音的相似性,将双语数据转化成声母、韵母、声调的表示形式,以此粒度训练翻译模型,并对解码结果进行还原,从而使译文获得更好的效果.Phuoc 等人[15]通过分析字符级翻译和词级翻译的优点,在词级别的翻译中使用统计与规则的方法,缓解了汉越机器翻译中数据稀疏的问题.针对汉语-越南语机器翻译中未登录词的翻译问题,Tran 等人[16]提出了基于汉语和越南语语义关系的命名实体的翻译方法.针对越南语修饰语后置的特点,He 等人[17]提出一种融合词根位置特征的汉-越机器翻译方法.根据定语位置、状语位置和修饰语排序信息定义排序块,然后与基于短语的统计机器翻译模型融合,使用排序块对模型解码结果进行重排序,从而得到越南语语法结构的译文.

以上工作均能够提升汉-越机器翻译的性能,但由于越南语具有资源稀缺的特点,在汉-越机器翻译中效果提升仍然十分有限.考虑到汉语与越南语之间存在语法差异,本文从句法知识的利用角度研究汉-越机器翻译方法.在句法知识的应用方面,Wu 等人[18]提出了将目标语言句法知识融入神经机器翻译模型的方法,使用两个 RNN 网络分别进行词语生成模型和句法结构模型的构建,通过依存上下文指导译文的生成.Chen 等人[19]使用 LSTM 网络对源语言的输入序列和句法树进行双向编码,使源语言句法信息融入编码过程,有效提高了模型的性能.Zhang 等人[20]将源语言的依存句法信息编码,然后使编码信息与源语言的词嵌入融合,并送入双向 RNN 编码,使模型能够有效学习到源语言与目标语言的词对齐关系.Li 等人[21]将句法树转化为句法标签序列,编码过程中,使用两个 RNN 网络对输入序列和句法标签序列同时编码,使源语言句法信息融入机器翻译模型,显著提高了翻译效果.相比以上方法,本文编码器使用卷积神经网络,将每个单词对应的句法标签向量与词向量相加作为编码器输入,以此提高模型性能.在汉-越机器翻译中,由于汉语和越南语的句法结构存在差异性,即汉语和越南语主语、谓语、宾语的顺序不同,使用依存句法树并不能充分地表示出语法特征.相比依存句法树,短语句法树能够表征出更深层次的词法句法信息,句法解析树结构如图 1 所示.为提高汉-越机器翻译的性能,本文提出了融

合源语言句法解析树的汉-越神经机器翻译方法.

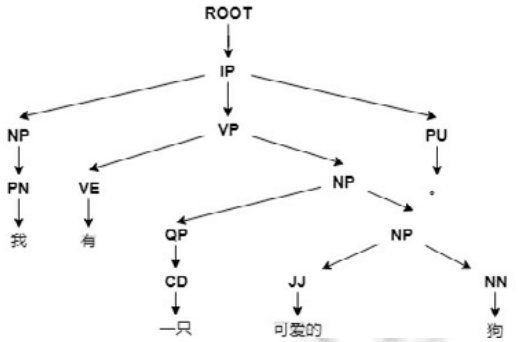


Fig.1 Structure diagram of Chinese syntactic parsing tree

图 1 汉语句法解析树结构图

2 融合源语言句法解析树的卷积神经机器翻译

在本节中,我们对汉语、越南语句法解析树的表征及融合方法进行了详细说明.由于基于 CNN 的神经机器翻译模型与基于 RNN 的神经机器翻译模型相比模型性能更好,因此在基准模型的选择上采用了具有多层卷积神经网络的编码器,解码器使用 LSTM 网络.在本文中,我们对神经网络结构不做修改,而是改变编码器的输入信息,融入了源语言的句法树信息.以下将从汉越句法解析树的获取、汉越句法解析树的向量化、基于 CNN 的神经机器翻译模型及汉越句法解析树的融合这 4 个方面,对融合汉越句法解析树的卷积神经机器翻译方法进行说明.

2.1 汉、越句法解析树的获取

获取汉越句法解析树,是为了得到汉语、越南语句子的语法结构及句子中单词之间的依赖关系,也就是为了得到汉语、越南语的语法信息为神经机器翻译模型训练提供支持.句法解析的准确率对神经机器翻译模型的性能有直接影响.获取高质量的汉语、越南语句法解析树是实验的关键.

目前,汉语的句法解析工具较多,同时准确率高.其中,有代表性的开源中文句法解析工是斯坦福的句法解析模型.本文利用斯坦福的汉语句法解析模型(ChinesePCFG)[22]对汉语进行句法解析,得到了汉语句法解析树,汉语句法解析结果如图 1 所示.

由于越南语的句法解析开源工具较少,在越南语句法解析树获取方面,我们采用李英等人[23]的越南语短语句法解析工具对越南语进行句法解析,得到越南语句法解析树.由于句法树是在单词的粒度上实现,但是越南语以音节为单位,使用越南语句法解析工具前,需要对越南语进行分词及词性标注.因此,首先利用实验室研发的语言信息处理工具对越南语进行预处理,所得到的越南语分词和词性标记结果如图 2 所示.

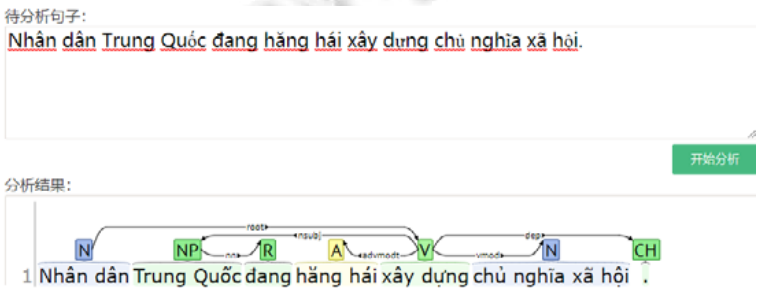


Fig.2 Vietnamese preprocessing

图 2 越南语预处理

然后,使用越南语句法解析工具对经过分词及词性标注的越南语句子进行短语句法解析,得到的越南语句法解析树如图 3 所示.

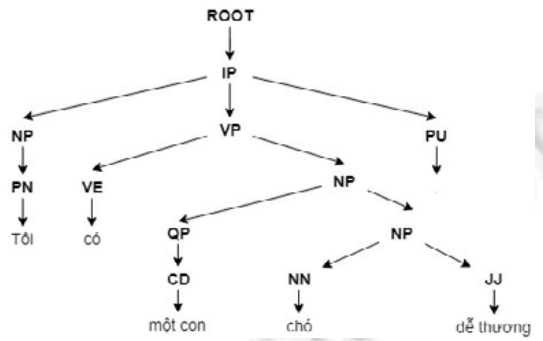


Fig.3 Structure diagram of Vietnamese syntactic parsing tree

图 3 越南语句法解析树结构图

从以上得到的汉语、越南语句法树可以看出:越南语具有修饰语后置的特点,中文短语“可爱的狗”在越南语中应该翻译为“chó(狗) dễ thương(可爱的)”,其中的修饰语“dễ thương(可爱的)”位于名词“chó(狗)”之后.由于得到的越南语、汉语句法树结构并不相同,同时,在汉语与越南语之间这种差异性比较明显,因此,本文将源语言句法信息融合到汉-越神经机器翻译模型中,以此提高模型性能.

2.2 汉、越句法解析树的向量化

在神经机器翻译模型训练中,需要将自然语言表征为特征向量的形式作为模型的输入.为了将句法解析树融入神经网络的编码过程,需要将句法解析树的树状结构转化为特征向量的形式.首先对以上得到的汉语和越南语的句法解析树进行深度优先遍历,对于每个叶子,都存在从根节点到该叶子节点的一条路径.采用这种方法分别得到汉语和越南语句子中每单词的句法标签序列,具体过程如图 4 所示.

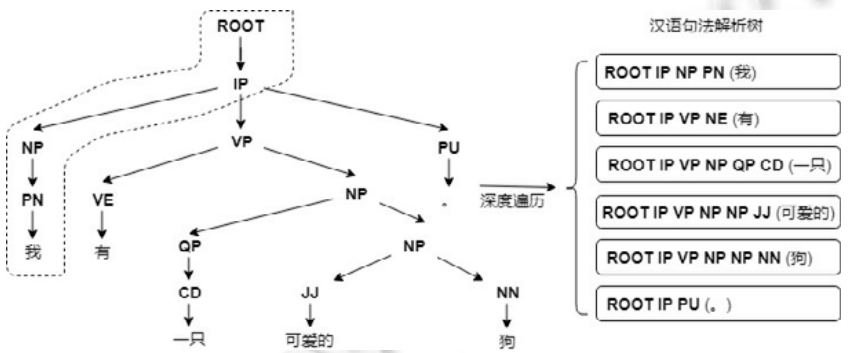


Fig.4 Syntactic tag sequence generation diagram

图 4 句法标签序列生成图

上图汉语句子中,每个单词都对应一个句法标签序列.为了对以上句法标签进行向量化表示,我们对每种标签定义固定的特征编码,如{(*ROOT*,1),(*IP*,2),(*VP*,3),…}.根据定义的标签编码,可以将每个单词对应的句法标签序列表示为以下形式:

$$\boldsymbol{g}_i=((w_1l_1+b_1),(w_2l_2+b_2),\ldots,(w_tl_t+b_t)).$$

其中,$\boldsymbol{g}_i$表示原句中第 i 个单词的句法标签向量,$l=\{l_1,l_2,\ldots,l_t\}$为句法标签序列中预定义的每个标签编码,t 为每个词对应的句法标签数量,$b=\{b_1,b_2,\ldots,b_t\}$表示偏置项初始值为 0.对于每个句法标签对应的权重 w_t,根据标签所在句法解析树的层次对权重进行初始化,越靠近叶子节点的标签对当前节点的影响越大,因此将权值 $w=\{w_1,$

$w_2,\ldots,w_t\}$初始化为$\{0.1,0.2,\ldots,0.1\times t\}$.在本文中,我们将句法标签向量定义为 64 维的向量表征,采用自左向右的填充方式,空白处用 0 表示.生成的句法标签向量如图 5 所示.

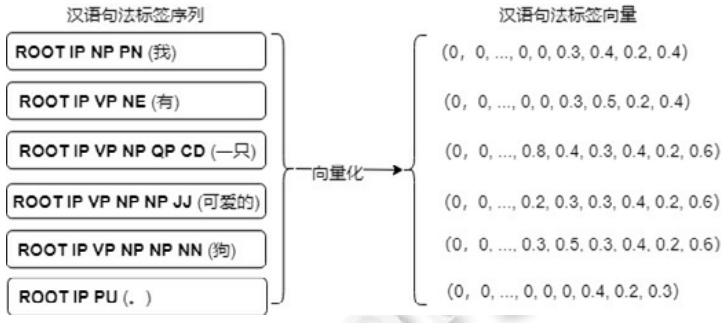


Fig.5 Syntactic tag sequence generation diagram

图 5 句法标签向量化

对于单词“可爱的”有句法标签对应的编码$\{(ROOT,1),(IP,2),(VP,3),(NP,7),(NP,7),(JJ,12)\}$,则有

$$g_{\text{可爱的}}=\{(0.1\times1),(0.2\times2),(0.3\times3),(0.4\times7),(0.5\times7),(0.6\times12)\}.$$

最终得到单词“可爱的”句法标签向量化表示为$g_{\text{可爱的}}=(0,0,0,\ldots,0.1,0.4,0.9,2.8,3.5,7.2)$.

2.3 基于CNN的神经机器翻译模型

我们在实验过程中发现:由于卷积神经网络权值共享及可并行的特点,基于 CNN 的神经机器翻译模型与基于 RNN 的神经机器翻译模型相比参数更少、收敛速度更快.同时,考虑到句法解析树是具有多层的树状结构与卷积神经网络的结构相似,所以选择 Gehring 等人[11]提出的基于卷积神经网络编码器的神经机器翻译模型作为基准模型开展实验.

基于 CNN 的神经机器翻译模型,模型结构如下图 6 所示.在输入层,将源语言的词嵌入和位置嵌入向量相加作为模型的输入,利用多层卷积神经网络对输入向量进行编码,提取源语言中的特征信息.编码过程中,首先选取输入序列$\{1,2,\ldots,k,\ldots\}$中前 K 个单词作为输入,这里将其定义为一个卷积窗口,将一个卷积窗口内所有单词的 embedding 向量相加,然后卷积核对平均之后的 embedding 向量进行卷积,得到当前窗口的编码向量.为了使深层的卷积网络能够获得原始输入信息,在每层卷积神经网络中使用残差连接,将原始输入通过非线性激活函数后直接转递到输出端.通过多层卷积编码最终得到输入序列上下文向量 h.在解码过程中,使用 LSTM 网络根据输入序列的注意力信息以及上下文向量 h 对输出序列进行还原.

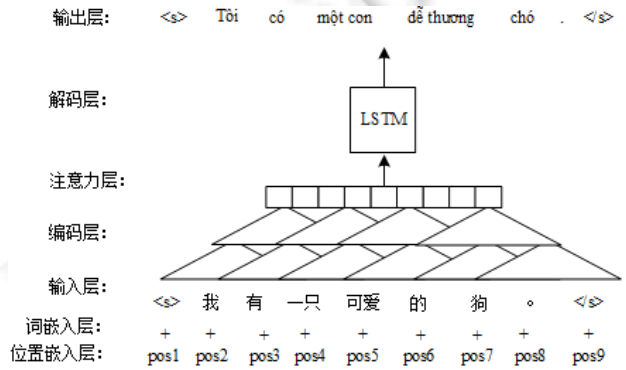


Fig.6 Model encoder-decoder architecture

图 6 基于 CNN 的神经机器翻译模型结构

由于卷积神经网络具有可并行的特点,编码过程中可以多个窗口同时进行卷积编码,相比基于循环神经网络的编码器具有更快的速度.但是在基于 RNN 编码器的模型中,输入序列依次编码,但由于该方法不包含输入序列的位置信息,并且输入序列中的 embedding 在一定程度上很接近,因此在基于卷积编码器的翻译模型中将单词对应的位置信息经过编码融入到源语言的输入过程中,通过向量相加的方式将源语言词嵌入和位置嵌入相加得到的新向量作为编码器输入.使词嵌入过程与上下文相关联,模型学习到输入序列中各单词之间的相对位置关系,提高模型的准确率.

2.4 源语言句法解析树的融入

本节介绍汉语、越南语句法解析树与卷积神经机器翻译模型的融合方法.基于以上汉语、越南语句法解析树,得到了汉语、越南语句子中每个单词对应句法标签序列的向量化表示.为了将句法表征向量融入翻译模型中,将源语言词嵌入向量、位置嵌入向量及句法标签向量进行拼接得到的向量作为模型的输入.由于本文对基准模型的结构不作修改,因此利用全连接网络,将输入向量的维度压缩为与模型编码器相同的维度,并且使全连接网络参与训练过程.最后训练神经网络,得到汉语-越南语的神经机器翻译模型.编码器的输入过程如图 7 所示.

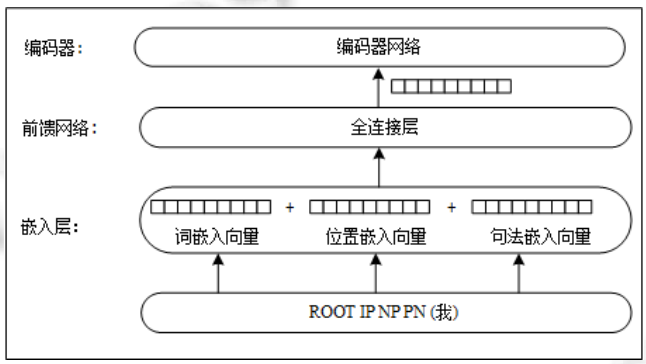


Fig.7 Syntactic parse tree information integration

图 7 句法解析树信息融入

在编码端,首先需要对输入序列进行词嵌入.对于输入序列 $x=(x_1,x_2,\ldots,x_m)$,利用门控线性单元(GLU)将其嵌入到分布空间 e 中,得到源语言的词嵌入向量为 $e_1,\ldots,e_m$,其中,$\boldsymbol{e}_i\in\mathbb{R}^d$ 是嵌入矩阵 $\mathcal{D}\in\mathbb{R}^{m\times d}$ 的列,计算方法如公式(1)所示:

$$e_i=(e_i\times W_1+b_1)\otimes\sigma(e_i\times V_1+c_1) \tag{1}$$

其中,$W_1,V_1\in\mathbb{R}^{k\times m\times n}$ 为权重,$b_1,c_1\in\mathbb{R}^n$ 为偏置项,m 为输入序列长度,σ为 sigmoid 函数,⊗是点乘.

位置嵌入过程与词嵌入类似,对于输入序列 $x=(x_1,x_2,\ldots,x_m)$所对应的绝对位置序列 $p=(p_1,p_2,\ldots,p_m)$,采用同样的方法嵌入到分布空间 e 中,其中,$\boldsymbol{p}_i\in\mathbb{R}^d$,其位置嵌入向量的维度大小与词向量维度大小相同.其计算方法如公式(2)所示.

$$p_i=(p_i\times W_2+b_2)\otimes\sigma(p_i\times V_2+c_2) \tag{2}$$

对于输入序列中各单词对应的句法标签序列 $s=(s_1,s_2,\ldots,s_m)$,同样利用 GLU 方法进行嵌入,其中,$\boldsymbol{s}_i\in\mathbb{R}^d$.计算方法如公式(3)所示.

$$s_i=(s_i\times W_3+b_3)\otimes\sigma(s_i\times V_3+c_3) \tag{3}$$

在基准模型中,采用向量加法将词向量与位置向量相加作为模型的输入.考虑到不同位置向量与词向量相加之后的结果可能相同,会影响模型的性能,因此在编码端,我们将源语言词嵌入向量、位置嵌入向量及句法标签向量进行拼接,然后采用全连接网络对拼接后的向量进行压缩作为编码器的输入,输入向量 I 的表示方法如公式(4)所示:

$$I=[(e_1+p_1+s_1),\ldots,(e_m+p_m+s_m)] \tag{4}$$

由于原有 CNN 编码器在嵌入层中对于单词的 embedding 向量是上下文无关的,对于给定单词在不同句子中均由相同的 embedding 表示,无法表示词语的语义信息.通过以上方法得到的词向量是上下文相关的,相同单词在不同上下文环境最终会得到不同的 embedding 表示,这样将位置信息、句法信息有效地融入到神经机器翻译模型,得到的译文更符合语法规则.

3 实验与分析

3.1 实验设置

为验证本文提出的融合句法解析树的汉-越神经机器翻译方法,我们分别在汉-越、英-越两种语言对上进行了实验,其中,汉-越语料是由实验室通过互联网爬取、人工翻译等方式收集得到的 136K 平行句对,从中随机抽取训练集、开发集与测试集;英-越训练数据是由 IWSLT 提供的并从中抽取 132K 平行句对作为训练集,使用其提供的开发集、测试集.

在实验数据预处理中,首先利用 JIEBA(https://github.com/jieba)中文分词工具对汉语进行分词,然后使用 MOSES 对全部训练数据进行 tokenization,lowercase 以及 clean 最终保留长度在 80 个词以内的句对.本实验使用的基准系统是 FaceBook 的开源神经机器翻译模型 fairseq,采用卷积神经网络进行编码、LSTM 进行解码.以及 Google 开源的 nmt 模型,采用基于 RNN 的编码器与解码器.所有实验均使用 132K 的双语平行语料作为训练集.所使用的实验数据见表 1.

Table 1 Experimental data settings

表 1 实验数据设置

	汉-越	英-越
训练集	132 000	132 000
测试集	2 000	1 930
开发集	2 000	1 925

本文采用单张 Tesla K40m GPU 进行实验.基准模型采用 BPE 词表,词表大小均为 40K.在本文所提方法中,由于句法标签是以词为单位,需要对源语言进行分词,生成词表.在 RNNsearchs 实验中,所使用的词嵌入维度为 512 维,编码器与解码器网络均为 6 层,每层隐含单元数为 256 个,并用 1.0 初始化 LSTM 的遗忘门偏置项,dropout 值为 0.2.在 Transformer 实验中,Transformer Base 模型使用 6 层编码器与解码器网络,每层单元数量为 512,批次大小 4 096,*dropout* 值为 0.1.在此基础上,Transformer Big 模型将隐层大小调整为 1 024,*dropout* 值为 0.3.在基于卷积神经网络的模型中,编码器与解码器的嵌入维度设置为 768 维.编码器设置为 15 层的卷积神经网络,解码器采用 LSTM 网络.编码器中,前 9 层的隐含单元数为 512 个,后 6 层的隐含单元数为 1 024 个,批次大小为 64 以及 *dropout* 值为 0.1,卷积核大小为 k.基准实验中,k 的取值为 5,在对比不同卷积核大小时,k 的取值分别为 k={3,5,7}.

3.2 实验结果与分析

本次实验主要研究融入句法解析树对神经机器翻译性能的影响,实验中对比了 RNNsearchs 模型、Transformer Base 模型、Transformer Big 模型、不带有任何外部信息的 CNN 模型、融入位置信息的 CNN 模型、带有句法树信息的 CNN 模型以及同时带有位置及句法解析树信息的 CNN 模型.在此基础上,为进一步探究影响机器翻译模型的因素,又对比了不同卷积核大小以及不同深度的神经网络对实验结果的影响.每组实验重复进行 3 次,将每次实验最后保存的模型进行评测,并取 3 次平均值作为最终实验结果的 BLEU 值.

(1) 融合句法解析树信息

本文首先将 RNNsearchs,Transformer Base,Transformer Big,CNN,CNN+P(位置信息),CNN+S(句法解析树信息)与 CNN+P+S(位置信息和句法解析树信息)进行对比实验与分析,使用 BLEU 值作为评价标准.汉-越、汉-英、英-越这 3 种语言对的实验结果见表 2.

Table 2 Experimental results of different model settings

表 2 不同模型设置的实验结果

	汉-越	越-汉	英-越	越-英
RNNsearch	17.31	13.92	18.67	18.34
Transformer Base	21.13	20.68	21.43	22.21
Transformer Big	21.65	21.11	21.71	22.43
CNN	18.81	18.07	20.12	20.86
CNN+P	21.52	21.02	21.97	22.12
CNN+S	21.61	21.09	21.83	22.32
CNN+P+S	**22.36**	**21.65**	**22.32**	**22.80**

如表 2 所示,基于 CNN 的神经机器翻译模型性能优于基于 RNN 的神经机器翻译模型以及基于 Transformer 的神经机器翻译模型.实验对比表明:通过融入源语言句法解析树,可使机器翻译性能提升.具体原因分析如下.

通过以上实验可以看出:当目标语言为汉语时,译文的 BLEU 值低于源语言为汉语的翻译效果.例如,在 CNN+P+S 实验中,汉语-越南语、越南语-汉语的翻译中,汉语-越南语的翻译相比越南语-汉语翻译高出 0.71 个 BLEU 值.主要原因是:越南语由音节构成,与汉语拼音类似,每个音节又由声母、韵母及音调组成,与汉语相比构词方法相对简单.相比在汉语中,汉字的构词非常丰富,但是在硬件资源、计算能力有限的情况下,在训练过程中使用有限大小的词表.在资源稀缺情况下,所得到的词表表征能力十分有限,OOV 问题相对严重,导致译文的 BLEU 值较低.

基准实验中,RNNsearchs 模型效果较差.这是因为基于 RNN 的模型在训练过程中存在不足.在训练过程中,编码器依次编码源语言句子中每个单词,产生固定长度的源语言上下文向量;然后,解码器通过这个上下文向量还原目标语言.采用这种编码-解码方式,模型无法充分学习到源语言中某个单词与其他单词的关联关系,也就是单词在一个句子中的上下文环境信息,导致解码器生成的单词脱离原文语境译文质量不佳.并且在基于 RNN 的模型中,未能够将源语言的句法知识融入到翻译模型中,因此译文句法结构与源语言句法结构不符,得到的翻译译文质量较差.同时,在单 GPU 的下,模型编码器解码器的层数及隐含层单元的大小受限,也是影响模型性能的因素.

基于 Transformer 的神经机器翻译模型性能优于 RNNsearchs,主要原因是 Transformer 模型采用多头注意力机制(multi-head attention),同时,在编码过程中,将源语言的词嵌入向量与位置向量相加作为模型输入,使词序信息有效融合到神经机器翻译模型的训练过程,提高了模型的性能.在此基础上,增加了网络宽度的 Transformer Big 模型相比 Transformer Base 模型性能上获得了进一步的提升.

基准实验中,基于 CNN 的神经机器翻译模型在未融入位置及句法信息时,相比 RNNsearch 模型效果有所提升,但是效果弱于融入位置信息或句法信息的神经机器翻译模型.原因是,基于 CNN 的编码器未能获取到源语言中词语的位置信息、词序关系与句法信息.在融入位置信息及句法信息后,相同单词在不同位置或上下文环境中得到不同的 embedding 表示,使编码器能够学习到更充分的语义信息,提高了模型的性能.

通过对比汉-越、英-越两组语言对的实验结果可以看出,融合句法信息在汉-越机器翻译上的作用更加明显.主要原因是:相比英语-越南语,汉语-越南语之间存在的语法差异较大.对于语法结构相似的语种,该方法获得的效果并不明显.因此,融合源语言句法信息能够有效提升汉-越机器翻译的性能.

(2) 不同卷积核大小

在卷积编码器模型中,随着编码器中卷积核大小的改变及编解码器层数的变化,所训练出模型的效果也会产生变化.因此,本文以汉语-越南语、越南语-汉语翻译为例,基于以上提出的 CNN+P+S 模型探讨了编码器层数以及卷积核大小对模型性能的影响.

为研究不同大小的卷积核对模型性能的影响,将编码器层数固定为 15 层,分别选取卷积核大小为 3,5,7 进行实验,结果见表 3.

从表 3 的实验结果中可以看出:当编码器网络层数不变时,卷积核大小变大,译文的 BLEU 值下降.卷积核最小时,模型获得的性能最好.

Table 3 Experimental analysis table of different convolution kernel sizes

表 3 不同卷积核大小下模型的效果

卷积核大小	汉-越	越-汉
3	22.82	22.86
5	22.36	21.65
7	21.98	20.43

(3) 不同编码器网络层数

为研究不同层数的卷积网络对模型性能的影响,下面将卷积核大小固定为 5,分别选取卷积网络层数大小为 5,9,15 进行实验,结果见表 4.

Table 4 Experimental analysis table of different network depths

表 4 不同网络深度下的模型效果

卷积网络层数	汉-越	越-汉
5	21.19	20.86
9	21.52	21.12
15	22.36	21.65

从以上结果可以看出:编码器网络的层数越多,所得到的模型效果越好.在编码器中采用更多层的卷积神经网络,能够更加充分地获得源语言的语义表征,提高模型的性能.在模型训练时,将编码器层数设置为 15 层,卷积核大小设为 3,能够得出较优的模型训练结果.

3.3 汉语-越南语译文分析

下面以汉语-越南语翻译为例分析融入句法解析树对译文的影响,将汉语“中国人民正在积极建设社会主义”以及“大家热泪盈眶,满怀高兴.”作为源语言,使用以上的 CNN+P+S 模型翻译成越南语.翻译结果见表 5.

Table 5 Comparison of example experiments in different groups

表 5 各组实例实验对比

中文	模型	越南语
中国人民正在积极建设社会主义.	参考译文	Nhân dân(人民) Trung Quốc(中国) đang(正在) hang(积极) hái xây dựng(建设) chủ nghĩa(主义) xã hội(社会)
	RNNsearchs	Trung Quốc(中国) Người dân(人民) đang(正在) tích(积极) cực xây dựng(建设) xã hội(社会) chủ nghĩa(主义)
	Transformer Base	Nhân dân(人民) Trung Quốc(中国) đang(正在) căng(积极) cực xây dựng(建设) chủ nghĩa(主义) xã hội(社会)
	Transformer Big	Nhân dân(人民) Trung Quốc(中国) đang(正在) tích(积极) cực xây dựng(建设) chủ nghĩa(主义) xã hội(社会)
	CNN	Người dân(人民) Trung Quốc(中国) đang(正在) căng(积极) cực xâydựng(建设) xã hội(社会) chủ nghĩa(主义)
	CNN+P	Người dân(人民) Trung Quốc(中国) đang(正在) tích(积极) hái xây dựng(建设) xã hội(社会) chủ nghĩa(主义)
	CNN+S	Nhân dân(人民) Trung Quốc(中国) đang(正在) tích(积极) cực xây dựng(建设) chủ nghĩa(主义) xã hội(社会)
	CNN+P+S	Nhân dân(人民) Trung Quốc(中国) đang(正在) hăng(积极) hái xây dựng(建设) chủ nghĩa(主义) xã hội(社会)

对比以上译文可以看出,基于 CNN 模型输出的译文质量高于 RNNsearchs 模型.主要原因是:通过 RNNsearchs 模型输出的译文句法结构与目标语言句法结构不符,如“中国人民”“社会主义”的译文为“Trung Quốc(中国) Người dân(人民)”与“xã hội(社会) chủ nghĩa(主义)”,其都按照汉语中名词作定语的句法结构进行翻译,而非根据越南语的句法结构.对于 CNN 模型的译文,同样存在译文句法结构与目标语言句法结构不相符的问题,如“热泪盈眶”的译文为“đều đầy(满是) nước mắt(泪水)”,译文中未对状语部分进行翻译.在融入位置、句法解析树的信息后,能够对译文的顺序进行调整.因此可以看出:融入句法解析树,能够使神经机器翻译模型学习获得语言的语言信息,对生成译文的词序和句法结构具有约束作用,缓解汉-越神经机器翻译中源语言与目标语言句法结构不相符的问题.

4 总 结

本文针对汉-越神经机器翻译面临的训练语料不足问题,提出了融合汉、越句法解析树的神经机器翻译方

法.该方法首先对汉语和越南语进行句法解析,得到句法解析树;然后,将句法解析树信息转化为向量表示;最后,将得到的句法向量融入到神经机器翻译模型编码器的输入中.本文在汉语-越南语、英语-越南语上进行了实验,同时又对比了不同深度的卷积神经网络、不同大小卷积核对实验结果的影响.结果表明:融入源语言句法解析树信息,能够有效提高汉-越神经机器翻译模型的性能.当然,该方法也存在一些不足,由于越南语在分词、词性标记以及句法解析的准确率不足,导致训练过程中的特征提中存在错误,影响最终神经机器翻译模型的性能.在未来的工作中,我们会探索从神经机器翻译模型的方面进行改进,进一步提升汉-越神经机器翻译模型的性能.

References:

[1] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the Advances in Neural Information Processing Systems 27 (NIPS 2014). 2014. 3104−3112.

[2] Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-Sequence attentional neural machine translation. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. 823−833.

[3] Eriguchi A, Tsuruoka Y, Cho K. Learning to parse and translate improves neural machine translation. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 72−78.

[4] Aharoni R, Goldberg Y. Towards string-to-tree neural machine translation. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 132−140.

[5] Pust M, Hermjakob U, Knight K, Marcu D, May J. Using syntax-based machine translation to parse English into abstract meaning representation. Computer Science, 2015, 482−489.

[6] Wu NR, Su YL, Liu WW, Ren QDEJ. Mongolian-Chinese machine translation base on CNN etyma morphological selection model. Journal of Chinese Information Processing, 2018,32(5):42−48 (in Chinese with English abstract).

[7] Bao WGDL, Zhao XB. Mongolian-Chinese neural machine translation base on RNN and CNN. Journal of Chinese Information Processing, 2018,32(8):60−67 (in Chinese with English abstract).

[8] Gehring J, Auli M, Grangier D, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proc. of the 34th Int'l Conf. on Machine Learning (ICML 2017), Vol.70. 2017. 1243−1252.

[9] Meng FD, Lu ZD, Wang MX, Li H, Jiang WB, Liu Q. Encoding source language with convolutional neural network for machine translation. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. 2015. 20−30.

[10] Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 1506−1515.

[11] Gehring J, Auli M, Grangier D, Dauphin Y. A convolutional encoder model for neural machine translation. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 123−135

[12] Trinh M, Tran P, Tran N. Collecting Chinese-Vietnamese texts from bilingual websites. In: Proc. of the 5th NAFOSTED Conf. on Information and Computer Science (NICS). 2018. 260−264.

[13] Tran P, Dinh D, Nguyen LHB. Word re-segmentation in Chinese-Vietnamese machine translation. ACM Trans. on Asian and Low-Resource Language Information Processing, 2016,16(2):1−22.

[14] Huu AT, Huang HY, Guo Y, Shi SM, Jian P. Integrating pronunciation into Chinese-Vietnamese statistical machine translation. Tsinghua Science and Technology, 2018,23(6):83−91.

[15] Phuoc T, Dien D, Nguyen HT. A character level based and word level based approach for Chinese-Vietnamese machine translation. Computational Intelligence and Neuroscience, 2016,2016(2):1−11.

[16] Tran P, Le T, Dinh D, *et al*. Handling organization name unknown word in Chinese-Vietnamese machine translation. In: Proc. of the 2013 RIVF Int'l Conf. on Computing & Communication Technologies—Research, Innovation, and Vision for Future (RIVF). 2013. 242−247.

[17] He YJL, Yu ZT, Lv CT, Lai H, Gao SX, Zhang Y. Language post positioned characteristic based Chinese-Vietnamese statistical machine translation method. In: Proc. of the 21st Int'l Conf. on Asian Language Processing (IALP). 2017.

[18] Wu SZ, Zhang DD, Yang N, Li M, Zhou M. Sequence-to-dependency neural machine translation. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 698−707.

[19] Chen HD, Huang SJ, Chiang D, Chen JJ. Improved neural machine translation with a syntax-aware encoder and decoder. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 1936−1945.

[20] Zhang MS, Li ZH, Fu GH, Zhang M. Syntax-Enhanced neural machine translation with syntax-aware word representations. In: Proc. of the NAACL 2019. 2019. 1151−1161.

[21] Li JH, Xiong DY, Tu ZP, Zhu MH, Zhang M, Zhou GD. Modeling source syntax for neural machine translation. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. 688−697.

[22] Levy R, Manning C. Is it harder to parse Chinese, or the Chinese treebank? In: Proc. of the 41st Annual Meeting on Association for Computational Linguistics. 2003. 439−446.

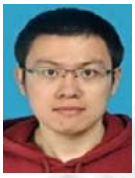
[23] Li Y, Guo JY, Yu ZT, XianYT, Chen W. Construction the Vietnamese phrase treebank by fusion of cietnamese grammatical features and improved PCFG. Journal of Nanjing University (Natural Sciences), 2017,(2):155−165 (in Chinese with English abstract).

附中文参考文献:

[6] 乌尼尔,苏依拉,刘婉婉,仁庆道尔吉.基于 CNN 词根形态选择模型的改进蒙汉机器翻译研究.中文信息学报,2018,32(5):42−48.

[7] 包乌格德勒,赵小兵.基于 RNN 和 CNN 的蒙汉神经机器翻译研究.中文信息学报,2018,32(8):60−67.

[23] 李英,郭剑毅,余正涛,等.融合越南语语言特征与改进 PCFG 的越南语短语树库构建.南京大学学报(自然科学),2017,(2):155−165.



王振晗(1993−),男,学士,CCF 学生会员,主要研究领域为自然语言处理,机器翻译.

文永华(1979−),男,讲师,CCF 学生会员,主要研究领域为自然语言处理,机器翻译.

何建雅琳(1993−),女,硕士,主要研究领域为自然语言处理,机器翻译.

郭军军(1987−),男,博士,讲师,CCF 专业会员,主要研究领域为自然语言处理,神经机器翻译,信息检索.

余正涛(1970−),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,信息检索.

高盛祥(1977−),女,博士,副教授,CCF 专业会员,主要研究领域为自然语言处理,机器翻译,信息检索.