

文章编号: 1003-0077(2018)07-0030-07

基于数据增强技术的神经机器翻译

蔡子龙, 杨明明, 熊德意

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要: 神经机器翻译是目前机器翻译领域最热门的研究方法。和统计机器翻译相比, 神经机器翻译在语料丰富的语种上可以取得非常好的结果, 但是在资源比较稀缺的语种上表现一般。该文利用数据增强技术对资源贫乏语种的训练数据进行扩充, 以此增强神经机器翻译的泛化能力。该文在藏汉、汉英两种语言对上进行了实验, 当训练数据规模只有 10 万平行句对时, 相较于基准系统, 在两种语言对上均获得了 4 个 BLEU 值的提高。实验表明, 数据增强技术可以有效地解决神经机器翻译因为训练数据太少而导致的泛化能力不足问题。

关键词: 神经机器翻译; 数据增强技术; 泛化

中图分类号: TP391

文献标识码: A

Data Augmentation for Neural Machine Translation

CAI Zilong, YANG Mingming, XIONG Deyi

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Neural machine translation performs well in large-scale language pairs, but less satisfactory for low-resource language pairs. This paper employs the data augmentation technology to expand the training data for low-resource language pairs, which can enhance the generalization ability of neural machine translation. Experimented with the language pairs of Tibetan-Chinese and Chinese-English, the translation quality is improved for both tasks significantly, achieving 4.0 BLEU point increase at the training scale of 100,000 pairs of sentences.

Key words: neural machine translation; data augmentation technology; generalization

0 引言

神经机器翻译是 Sutskever 等人^[1]在 2014 年提出的一种基于编码器-解码器模型的机器翻译方法。和传统基于短语的统计机器翻译^[2]不同, 神经机器翻译没有特征工程、隐藏结构设计等方面的困扰, 而是简单地通过训练一个单一、大型的神经网络对输入句子产生合适的翻译。该方法刚被提出来的时候, 效果还不如统计机器翻译。2015 年, Bahdanau 等人^[3]在此工作的基础上通过引入注意力机制使得神经机器翻译在多种语言对上的评测结果超过统计机器翻译, 神经机器翻译因此得到了广泛的关注。

神经机器翻译本质上是训练一个大型的神经网络,

该网络由上万个神经元构成。为了能够充分地学习到网络的权重值, 神经机器翻译需要大量的平行句对作为训练数据, 往往平行句对越多, 训练效果越好。然而, 对于资源贫乏语种来说, 获得充足的训练语料是十分困难的。

为了解决神经机器翻译在资源贫乏语种上因为训练数据太少而导致的泛化能力不足问题, 本文提出了一个简单有效的方法: 数据增强技术。该方法首先对句子进行分块, 然后找到句子中最相似的两个模块, 通过对调它们的位置形成新的句子。利用数据增强技术, 不但可以将训练语料扩充一倍, 而且也会使句子的结构变得多样化。本文在藏汉语种上进行了实验, 较于基准系统, 获得了 4 个 BLEU 值^[4]的提高。实验表明, 本文提出的数据增强方法可以显著提高神经机器翻译对于资源贫乏语种的泛

收稿日期: 2017-09-15 定稿日期: 2017-10-24

基金项目: 国家自然科学基金优秀青年基金(61622209)

化能力。

本文其他部分的组织如下：第一节和第二节分别介绍了神经机器翻译的背景知识和在资源贫乏语种上研究的相关工作；第三节详细说明如何对训练语料进行数据增强；第四节对实验结果进行分析；最后，对本文的工作进行了总结，并对下一步工作进行了展望。

1 背景知识

本节主要介绍基于注意力机制的神经机器翻译。如图 1 所示，神经机器翻译分为两个部分，一个是编码器，另一个是解码器。编码器采用双向循环的神经网络，对源句子 $\mathbf{x} = x_1, x_2, \dots, x_{T_x}$ 进行编码，得到该句子的隐藏层 $\mathbf{h} = h_1, h_2, \dots, h_{T_x}$ 。解码器使用注意力机制，从左往右逐单词地生成目标端句子 $\mathbf{y} = y_1, y_2, \dots, y_{T_y}$ 。

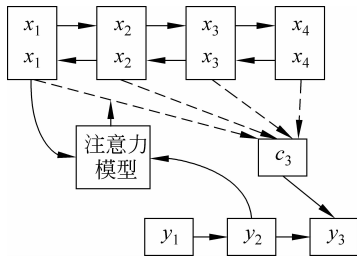


图 1 基于注意力机制的神经机器翻译模型

在训练阶段，神经机器翻译计算一个平行句对 $\langle \mathbf{x}, \mathbf{y} \rangle$ 的条件概率，如式(1)所示。

$$P(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^{T_y} p(y_i | \mathbf{y}_{<i}, \mathbf{x}) \quad (1)$$

其中 y_i 是解码器第 i 时刻生成的目标端单词， $\mathbf{y}_{<i} = y_1, y_2, \dots, y_{i-1}$ 。条件概率 $p(y_i | \mathbf{y}_{<i}, \mathbf{x})$ 计算如式(2)所示。

$$p(y_i | \mathbf{y}_{<i}, \mathbf{x}) = \text{softmax}(f(s_i, y_{i-1}, c_i)) \quad (2)$$

其中 $f(\cdot)$ 是非线性函数， s_i 是解码器第 i 时刻的隐藏状态，如式(3)所示。

$$s_i = g(s_{i-1}, y_{i-1}, c_i) \quad (3)$$

其中， $g(\cdot)$ 是非线性函数。 c_i 是源端句子的内容向量，它是源端隐藏层 \mathbf{h} 的线性和，权重 $a_{i,j}$ 代表解码器解码第 i 个单词时，对源端第 j 个单词所对应的隐藏向量的关注度。计算方式如式(4)所示。

$$c_i = \sum_{j=1}^n a_{i,j} h_j \quad (4)$$

基于注意力机制的神经机器翻译使用最大似然

函数训练整个模型的参数，如式(5)所示。

$$C(\theta) = \sum_{n=1}^N \sum_{i=1}^{T_y} \log(p(y_i^n | \mathbf{y}_{<i}^n, \mathbf{x}^n)) \quad (5)$$

其中， θ 指的是模型的参数， N 指的是语料中的平行句对数。

在测试阶段，如式(6)所示，给定一个源端句子 \mathbf{x} ，用束搜索^[5]方法找到使条件概率 $p(\mathbf{y} | \mathbf{x})$ 最大的句子 $\hat{\mathbf{y}}$ ，并把它作为 \mathbf{x} 对应的翻译。

$$\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}} (p(\mathbf{y} | \mathbf{x})) \quad (6)$$

2 相关工作

本节主要介绍神经机器翻译在资源贫乏语言对上研究的相关情况。2016 年，Zoph 等人^[6]最先提出：在资源贫乏语言对上，神经机器翻译的效果不如统计机器翻译，他们在四组资源贫乏语言对上进行了实验，分别用神经机器翻译和统计机器翻译进行训练，结果发现统计机器翻译在这四组语言对上的评测结果均优于神经机器翻译。

国内外很多研究者对此问题提出了各自的解决方法。这些方法大致可以分为两类，一类是通过补充训练数据，使得神经机器翻译模型得到较为充分的训练；另一类是将文字的语法、语义信息和神经机器翻译模型相融合，从而提高神经机器翻译的性能。

本文主要关注的是第一类解决方法，即通过增加训练数据，改善神经机器翻译在资源贫乏语言对上翻译性能不佳的情况。增加训练数据，并不是去挖掘真实的平行句对，而是通过技术手段，构造伪平行句对。

Sennrich^[7]是第一个提出利用单语语料来构造伪平行句对的研究者。他认为，对于资源缺乏的语言对来说，单语语料的获取往往相对容易，充分地利用单语语料来增强神经机器翻译是十分必要的。在收集到单语语料之后，可以用现有的翻译工具或者在小规模语料上训练好的神经机器翻译模型对单语语料进行翻译，从而获得伪平行句对，之后将伪平行句对和真实的平行句对放到一块进行模型训练。

和 Sennrich 的思路不同，Fadaee^[8]提出了一种新的增加语料的方法。该方法首先在规模较大的单语语料上训练出语言模型，然后用语言模型找到句子中可以被低频词替换的高频词的位置。通过这种简单的单词替换，增加了训练语料中低频词出现的

次数,从而增强神经机器翻译对低频词的理解能力。

本文在 Sennrich 和 Fadaee 等人工作的基础上,提出了自己的数据增强技术,即将句子中最相似的模块进行位置上的对调,以此形成新的语料。与 Sennrich 提出的方法相比较:不同点在于我们利用真实的平行句对而非单语语料进行伪语料的构造,相同点在于构造的伪语料都存在错误。与 Fadaee 提出的方法相比较:不同点在于我们改变的是语料中句子的结构信息而非语料中的词频信息,相同点在于都是对原语料进行扩充。

3 数据增强技术

本文中,我们把神经机器翻译当作一个“黑盒子”,不进行任何修改,而是利用数据增强技术提高神经机器翻译对于资源贫乏语种的泛化能力。本节从三个方面对数据增强技术进行详细的说明。第一,分析数据增强技术在神经机器翻译上面临的难点;第二,提出解决这些难点的方法;第三,介绍数据增强技术具体实现的细节。

3.1 难点分析

虽然数据增强在图像处理任务中已经成为一个标准的技术用于提高神经网络的泛化能力,但是由于语言的特殊性,我们并不能简单地将该技术拓展到机器翻译任务上来。

图 2 是数据增强技术在图像分类任务中的一个典型应用。新图像(b)由原图像(a)翻转 180 度所得,因为是简单的旋转变换,所以图像(b)的内容、标签与图像(a)完全一致。把图像(b)放入训练数据,可以增强图像训练的鲁棒性^[9]。

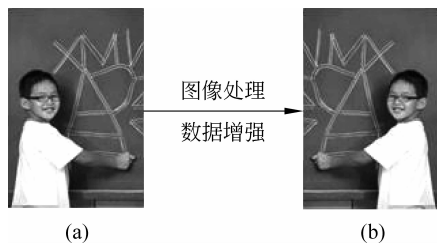


图 2 数据增强技术在图像处理任务中的应用

自然语言处理与图像处理方式大有不同。自然语言有着严格的语法约束,如表 1 第一个例子所示,简单地将一句话从右往左地倒着读既破坏了语法规则,句子本身也失去了语义信息。第二个例子,原句子本身就是一个十分简单的主谓宾结构,将“我”和

“篮球”进行对调,虽然新句子没有破坏语法规则,但是在语义上存在错误。第三个例子,“西电东送”和“西气东输”是对等的两个实体,将它们对调形成的新句子在语义和语法上均保持正确。

根据上面的分析,我们不难发现,如果原句子的长度较短,本身结构简单,那么无论对这个句子进行怎样的变化,新句子都会存在语法或者语义上的错误。对于这类存在一定错误的新句子,我们并不会丢弃,而是将它作为噪声来增强神经机器翻译的泛化能力。

而对于表 1 中第三个例子,我们应尽可能地将句子中最相似的模块进行调换。因此,数据增强技术在神经机器翻译上的难点如下:①如何获得句子的模块;②如何计算模块之间的相似度?

表 1 自然语言变化的示例

序号	原句子	新句子
1	全面 增强 我们 的 科技 创新 能力	能力 创新 科技 的 我们 增强 全面
2	我 喜欢 打 篮球	篮球 喜欢 打 我
3	西电 东送、西气 东输 全面 投产	西气 东输、西电 东送 全面 投产

3.2 解决方法

获得句子的模块是指如何对一个句子进行切分。句子可以分为三个层级,其中单词是构成句子的最小单位,单词组成短语,短语的再上一级是最小翻译单元。以单词为单位对句子进行切分会存在一对多的问题。例如,图 3 是词对齐中常见的一对多问题,源端 A 单词分别和目标端 a, b 对齐,将 a 和 b 的位置进行对调,那么源端 A 的位置并不能唯一地确定下来。

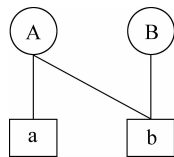


图 3 词对齐中的一对多问题

因此,本文以最小翻译单元为单位对句子进行切分,理由如下:①最小翻译单元在句子的结构中处于最上层,除了涵盖单词和短语的信息外,它还具有一些句子级别的信息;②最小翻译单元具有闭包性,单元与单元之间不存在词对齐关系,这个性质避免了词对齐中的一对多问题。

计算句子模块之间的相似度分两步：第一步，获得模块的向量表示；第二步，对模块进行余弦相似度计算。这里句子模块指的是最小翻译单元，最小翻译单元由句子中连续的若干个单词构成，因此有两种方式获得它的向量表示：第一种，以单词为单位，用 Word2Vec 对原语料进行训练，获得单词的向量表示。然后将构成最小翻译单元的单词向量相加作为其对应的向量表示。第二种，把最小翻译单元当做一个整体，用 Word2Vec 直接获得它所对应的向量表示，得到最小翻译单元 mtu_1 和 mtu_2 的向量表示之后，如式(7)所示，用向量夹角的余弦值来描述它们的相似度，余弦值越大，最小翻译单元就越相似。

$$\text{sim}(mtu_1, mtu_2) = \frac{mtu_1 \cdot mtu_2}{|mtu_1| \times |mtu_2|} \quad (7)$$

3.3 具体实现

本节通过一个例子，具体说明实现数据增强技术的四个步骤。如表 2 所示，首先获取平行语料，然后利用 moses 对平行语料进行训练，获得词对齐信息。接着利用 pbmt 工具得到句子的最小翻译单元。最后通过调换原句子最相似的两个模块得到新的平行句对。

对于第四个步骤，我们要分情况进行讨论。第一种情况是 3.2 节中提到的，最小翻译单元向量的表征方式分为直接和间接两种。第二种情况是原句对中源端最相似的模块不一定和目标端最相似的模块相互对齐。

表 2 数据增强技术的四个步骤

原始句对	ལྷན་སྐྱོད་གན་སྐྱུལ་དང་ལྷན་སྐྱུལ་གན་འདྲིན་ ཕྱགས་ཡོངས་ནས་སྐྱུལ་འདྲིན་བྱེད་འགོ་ཚུགས།
	西电 东送、西气 东输 全面 投产，
词对齐	0-0 1-0 0-1 2-1 3-1 4-2 5-3 7-3 12-3 6-4 8-4 11- 4 9-5 10-5 13-6 14-7
最小翻 译单元	ལྷན་སྐྱོད་_གན་_སྐྱུལ་ དང་ ལྷན་སྐྱུལ་_གན་_འདྲིན་ ཕྱགས་ ཡོངས་_ནས་_སྐྱུལ་འདྲིན་_བྱེད་འགོ་ ཚུགས།
	西电_东送 西气_东输 全面 投产，
新句对	西气 东输、西电 东送 全面 投产，

对于第一种情况，我们做两组对比实验。第一组，以单词为单位，用 Word2Vec 对语料进行词向量训练，然后用单词向量的和对最小翻译单元进行向量表征，我们把通过这种方式得到的向量称为最小翻译单元的间接向量(I-MTU)。第二组，我们把最

小翻译单元当做一个单词，用 Word2Vec 进行训练，获得的词向量，我们称为最小翻译单元的直接向量(D-MTU)。

对于第二种情况，我们分三种方式产生新句对。如表 3 所示， x_i, y_i 是源端和目标端相互对齐的最小翻译单元对， (x_i, y_i) 表示源端第 i 个和第 j 个最小翻译单元的相似度(sim)。

第一种方式是以源端为基准产生新句对(source-based generate, SBG)。该方法首先找到源端最相似的最小翻译单元 x_1 和 x_2 ，并对调它们的位置，然后根据对齐信息，找到目标端对应的最小翻译单元 y_1 和 y_2 ，并对调它们的位置。

第二种方式是以目标端为基准产生新句对(target-based generate, TBG)，与 SBG 类似，这里不再赘述。

第三种方式是将源端和目标端相结合产生新句对(combination-based generate, CBG)。CBG 综合考虑源端和目标端 sim 值排在前 k 个的最小翻译单元对，两者取交集，若该交集非空，则取交集里相似度最高的作为两端最相似的最小翻译单元。若该交集为空，则比较源端和目标端最大的 sim 值，当源端 sim 值高于目标端的时候，我们采用 SBG 产生新句对；否则，我们采用 TBG 产生新句对。实验中我们把 k 值设置为 3。

表 3 各种新句对产生方式

Source-MTU	x_1, x_2, x_3, x_4
Target-MTU	y_1, y_2, y_3, y_4
Rank of Source-MTU	$(x_1, x_3) > (x_2, x_3) >$ $(x_1, x_4) > (x_1, x_2) >$ $(x_2, x_4) > (x_3, x_4)$
Rank of Target-MTU	$(y_1, y_2) > (y_2, y_3) >$ $(y_2, y_4) > (y_1, y_3) >$ $(y_1, y_4) > (y_3, y_4)$
SBG	x_3, x_2, x_1, x_4
	y_3, y_2, y_1, y_4
TBG	x_2, x_1, x_3, x_4
	y_2, y_1, y_3, y_4
CBG	x_1, x_3, x_2, x_4
	y_1, y_3, y_2, y_4

4 实验结果与分析

表 3 中，Source-MTU 指的是句子源端的最

小翻译单元,Traget-MTU 指的是句子目标端的最小翻译单元,Rank of Source-MTU 指的是将源端的最小翻译单元对按照余弦值从大到小进行排序,Rank of Traget-MTU 指的是将目标端的最小翻译单元对按照余弦值从大到小进行排序。

4.1 实验设置

为了验证本文提出的数据增强技术,我们分别在藏汉、汉英这两个语言对上进行实验。其中,藏汉语料是 2011 年全国机器翻译研讨会提供的 10 万平行句对,测试集为 650 句。中英语料是本实验组收集整理的,共 100 万平行句对,测试集为 nist06。

本文用 Word2Vec 获得最小翻译单元的向量表征。Word2Vec 包含两种训练模型,本文用的是 skip gram 模型^[10],其中词向量的维度设置成 30,训练窗口大小设置为 5。

本文用的神经网络机器翻译系统是本课题组基于 Bahdanau 等人的工作开发出来的,用“RNNSearch”表示。其中,对训练语料的句子长度限制在 80 以下,源端和目标端的词向量维度设置为 620,隐藏层维度设置为 1 000,单词表大小设置为 3 万,采用 ADADELTA^[11]方法对参数进行更新,训练中 batch 的大小设置为 80,Dropout^[12]设置为 0.5。

本文还将基于数据增强技术的神经机器翻译与统计机器翻译作对比,实验采用爱丁堡等大学联合开发的 Moses^[13]作为统计机器翻译的基准系统,Moses 采用默认配置,实验以 BLEU-4 作为评测标准。

4.2 最小翻译单元的向量表征

本文针对最小翻译单元的向量表征提出了两种方法,分别是 I-MTU 和 D-MTU。其中,I-MTU 是一种间接获取短语向量表征的方式,D-MTU 把短语作为一个整体,其向量表示由 Word2Vec 训练得到,是一种直接获取短语向量表征的方式。从图 4 中我们可以看出,在三种不同生成句子的策略下,D-MTU 的结果都要比 I-MTU 好,这表明,虽然 Word2Vec 训练的词向量具有良好的语义信息,但是简单地用词向量的和对短语进行表征,还是存在一定问题。

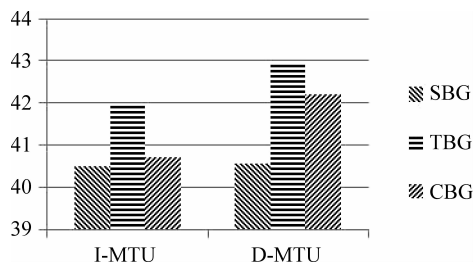


图 4 藏汉六组实验结果

4.3 生成句子的策略选择

本文用三种不同的方式产生新的句对,分别是 SBG、TBG 和 CBG。图 4 的实验结果表明,在藏汉翻译上,TBG 对翻译性能的提升是最显著的。

为了说明 SBG、CBG 和 TBG 三种方法间的差异,我们对产生的新语料进行了统计,结果如表 4 所示。SBG 和 TBG 产生的新语料中有近 60% 的句子是不同的,CBG 和 TBG 产生的新语料中有近 40% 的句子是不同的,这表明翻译源端和目标端语种的不同,对找出相似的最小翻译单元是有影响的。对于藏汉翻译来说,以汉语为基准产生新句对要比以藏语为基准产生新句对的方法好。在其他语言的翻译任务中,我们并不能事先知道 SBG 和 TBG 哪种方法更好,而训练神经机器翻译往往需要大量的时间和资源,这时采用折中的方法 CBG 是一个不错的选择。

表 4 SBG、CBG 与 TBG 不相同句子数所占的百分比

SU	I-MTU	D-MTU
TBG-SBG	59.7%	59.4%
TBG-CBG	38.7%	39.8%

注: SU 是差集的缩写。

4.4 数据增强技术的验证

为了验证本文提出的数据增强技术,我们做了以下几组实验进行对比分析。

从表 5 中,我们可以看出,在藏汉这种小语料上,神经机器翻译的基准系统比统计机器翻译系统低了 3 个点,这验证了 Zoph 等人提出的在资源稀缺的语言对上,神经机器翻译要弱于统计机器翻译。通过使用数据增强技术,神经机器翻译系统的性能得到大幅提升,BLEU 值提高了 4 个点,甚至比统计机器翻译的结果还要高 1 个点,这验证了我们提出的数据增强技术的有效性。

表 5 藏汉测试集上,三种系统结果的比较

SYSTEM	test
Moses	41.71
RNNSearch	38.61
RNNSearch+TDA	42.89

为了进一步分析数据增强技术在不同程度的资源贫乏场景下的效果,如表 6 所示,我们以汉英作为我们的训练语言对,分别在语料规模为 10 万(极度贫乏)、30 万(十分贫乏)、50 万(中度贫乏)、70 万(轻微贫乏)、100 万(不贫乏)上做实验,实验结果表明,当语言对处于极度贫乏时,本文提出的数据增强技术可以有效地提升神经机器翻译的性能,当语言对不是很贫乏时,本文提出的方法也是正向反馈的,也能提高大概 1 个多点的 BLEU 值。

表 6 不同程度的资源贫乏场景下的效果对比

Zh-en	10 万	30 万	50 万	70 万	100 万
Baseline	16.77	25.53	28.11	29.38	30.51
+TDA	20.44	26.72	29.0	30.02	31.59
+fake	24.85	—	—	—	—
+fake +TDA	26.29	—	—	—	—

如表 6 最后两行所示,我们对比了 Sennrich 提出的用伪语料加强神经翻译训练的方法。实验中,我们用谷歌翻译工具对随机抽取的与 Baseline 不相同的英文句子进行中文翻译,从实验结果上来看,这种伪语料技术的效果要比本文提出的数据增强技术要好,但是考虑到谷歌公司可能会把我们的联合国语料放到它自己的模型上进行训练,因此伪语料技术实际上未必能比我们提出的数据增强技术高 4 个 BLEU 值。我们也将本文提出的数据增强技术和伪语料技术相结合,实验结果表明两种方法联合使用可以进一步提升资源贫乏语言对的翻译质量。

4.5 示例分析

表 7 是我们从测试集中挑选的句子,用于说明数据增强技术对神经机器翻译系统的帮助。对比参考译文,Moses 生成的译文丢失了动词“提高”;基准系统 RNNSearch 生成的译文丢失了名词“企业”,且不通顺;RNNSearch+TDA 基本翻译正确,而且通过调换“要”的位置,相比于参考译文,也显得更加通顺。

表 7 译文示例

原文	ཁྱེད་ཀྱི་ལུ་ཁྱེད་ཀྱི་དཔལ་འབྱོར་ ཀྱི་ཁྱེད་ཀྱི་འཕྲན་ཚུགས་ལ་ ཆེད་ཀྱི་དགོས་ན་གནད་འགག་ ནི་ཁྱེད་ཀྱི་འཕྲན་ཚུགས་ ཆེད་ཀྱི་དཔལ་འབྱོར་
参考译文	一个国家的经济要提高 国际竞争力,关键是要提高企 业的竞争力。
Moses	一个国家的经济的国际 竞争能力,关键是企业竞争 力。
RNNSearch	增强具备国家经济国际实 力的能力,关键会提高资源竞 争力。
RNNSearch+TDA	要增强一个国家的经济 国际竞争力,关键是增加企业 的竞争力。

5 总结

本文针对神经机器翻译在资源贫乏语种上面临的训练语料不足问题,提出了数据增强方法。该方法首先对句子进行分块,然后调换最相似的两个模块得到新的句子,最后将新的句子加入到语料中,对翻译模型进行训练。本文在藏汉、汉英语种上的实验结果表明,数据增强技术既能有效提高神经机器翻译对于资源贫乏语种的泛化能力,也能对语料较丰富的语种起到提升翻译质量的作用。当然,该方法也有自己的缺点,通过这种方法产生的新句子往往存在语义或者语法上的错误,这种错误对神经机器翻译产生的影响有待研究。在未来的工作中,我们会考虑将句法知识引入数据增强技术,以此改善生成的句子。

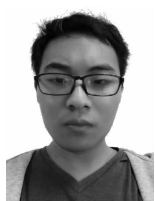
参考文献

[1] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 27(4): 3104-3112.

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

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]// Proceedings of the International Conference on Learn-

- ing Representations. San Diego, CA: International Conference on Learning Representations, 2015.
- [4] Kishore Papineni, et al. Bleu: A method for automatic evaluation of machine translation[C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002.
- [5] Philipp Koehn. A beam search decoder for phrase-based statistical machine translation models[J]. Computer Science, 2004, 3265: 115-124.
- [6] Zoph Barret, et al. Transfer learning for low-resource neural machine translation [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016.
- [7] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016.
- [8] Fadaee Marzieh, et al. Data augmentation for low-resource neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017.
- [9] Ken Chatfield, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//Proceedings of the British Machine Vision Conference. Jubilee Campus, British.
- [10] Tomas Mikolov, Quoc V Le, Ilya Sutskever. Exploiting similarities among languages for machine translation[OL]. <https://arxiv.org/abs/1309.4168>, 2013.
- [11] Matthew D Zeiler. ADADELTA: An adaptive learning rate method[OL]. <https://arxiv.org/abs/1212.5701>, 2012.
- [12] Nitish Srivastava, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014 (15): 1929-1958.
- [13] Koehn Philipp, et al. Moses: open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions, Prague, Czech Republic: Association for Computational Linguistics, 2007, 9(1): 177-180.
- [14] Xing Wang, Zhengdong Lu, Zhaopeng Tu, et al. Neural machine translation advised by statistical machine translation[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017: 3330-3336.
- [15] Xing Wang, Deyi Xiong, et al. Learning Semantic Representations for Nonterminals in Hierarchical Phrase-Based Translation [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015.



蔡子龙(1993—), 硕士研究生, 主要研究领域为自然语言处理、机器翻译。
E-mail: pauseman@yeah.net



熊德意(1979—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理、人工智能。
E-mail: dyxiong@suda.edu.cn



杨明明(1986—), 博士研究生, 主要研究领域为自然语言处理、机器翻译。
E-mail: mmyang@mtlab.hit.edu.cn