基于数据增强技术的神经机器翻译

韩东旭*，叶娜，张桂平

(沈阳航空航天大学人机智能研究中心，辽宁 沈阳 110136)

摘要：神经机器翻译自兴起以来，不断取得了更好的翻译效果。但是神经网络有一个瓶颈，就是只有在大规模平行语料的前提下才会取得好的效果，对于低资源领域的效果一般。本文从数据增强的角度出发，利用翻译模板的思想，识别并抽取出句子中的名词或术语，保留句子的主干。然后通过将抽取出的术语集合在句子主干框架上进行重组之后生成伪平行语料。最后通过计算句子的困惑度来对生成的伪语料进行把关，生成质量较好的伪语料。该方法有效的缓解了神经网络因为语料不足而导致模型的泛化能力不足的问题。实验结果表明，该方法获得的译文与基线系统相比，BLEU 值提升了 2.32。

关键词：翻译模板，神经机器翻译，数据增强，伪语料

中图分类号：TP391　　文献标志码：　A

神经机器翻译是通过利用一个非线性的神经网络来进行建模，以达到实现自然语言之间相互转换的目的。神经机器翻译作为一种全新的方法，尽管使机器翻译取得了飞速的发展，但是神经网络只有在大规模语料训练的前提下训练出的模型效果才好，在资源稀缺的领域中的翻译效果还差强人意。在针对低资源的任务中，对翻译模型进行数据增强，是一个非常值得研究的一个方向。

能够进行机器翻译的前提之一是可以从各种各样的语言现象中总结出可以指导翻译过程的知识，让计算机学习到这些知识例如实体和概念以及术语等，并利用这些知识来指导翻译的过程。翻译模板是这一系列知识的合理表示的一种方法。一般认为在一个句子中，句子主干中包含了句子大部分的信息，可以将句子主干抽取出来当作翻译模板，在机器翻译解码的时候通过对翻译模板的匹配从而完成翻译过程。在此前提之下，采用基于句子模板生成的伪语料，即可以在一定程度上保留句子的主干信息，又能对语料库中的术语等进行数据增强，是提升模型性能的一个强而有力的手段。

针对机器翻译任务，考虑到翻译模板可以在一定程度上保留句子的结构信息，本文提出了一种基于模板驱动的伪平行语料的生成方法。通过对语料的分析抽取出句子模板和句子中包含的术语，然后利用术语和翻译模板的重组形式来生成伪平行语料，并通过计算句子的困惑度筛选出质量较优的语料，并通过词对齐结果生成源语言对应的目标语言，最终将生成的伪平行语料用于模型的训练，以达到提升译文整体质量的效果。实验结果表明，该方法获得的译文与基线系统相比，BLEU 值提升了 2.32。

1 相关研究

1.1 翻译模板

在商用的机器翻译中，为了提升译文的质量，使用翻译模板进行翻译是一项不可或缺的工作[1]。其中翻译模板如图 1 所示，即根据原始的英汉平行语料“李久越 坐 飞机 离开 北京 ➞ Li Jiuyue left Beijing by plane” 通过将任务、地点及交通工具的去除，可以得到英汉的翻译模板 “$ 坐 $ 离开 $ ➞ $ left $ by $ ” 获得翻译模板之后，在实际应用时通过对翻译模板的匹配，例如 “ 小王 坐 客车 离开 北京 ” 通过对槽内的内容进行翻译，最终拼接出正确的译文。

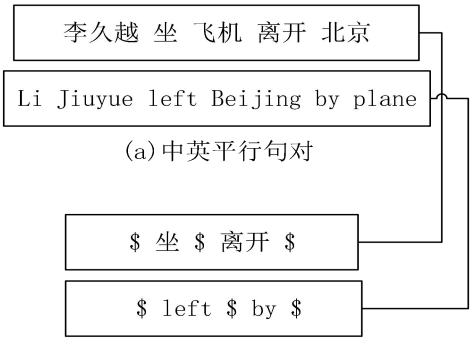

图 1： 翻译模板示例

Fig.1：Example translation template

针对翻译模板可以保留句子当中的主干信息,Chiang 等人提出了基于层次短语的统计翻译模型[2]，其本质是通过对模板的匹配来完成翻译过程。在 Chiang 的工作中，其方法首先是在大规模的训练数据中生成同词对齐结果保持一致的翻译模板和短语对，在翻译的过程当中通过对翻译模板和短语对的重组来生成译文。在基于一定规模的翻译模板的前提下，李强等人提出了引入另外一个编码器来对翻译模板进行建模，该模型通过引入知识阀门和注意力阀门来控制不同来源的信息对当前解码词汇的贡献度的大小。模型中的注意力阀门的作用主要体现在若当前词“left”存在模板编码器中,则更多的使用来自模板编码器的信息，否则更多的使用来自句子编码器的信息。知识阀门则更多的体现在解码器的第一个节点的初始化上，当翻译模板中含有更多的词汇时，则解码器在翻译时更多的使用来自模板的信息，当翻译模板中含有极少的真实词汇或没有真实词汇的时候，解码

器则更多的使用来自源语言句子的编码器信息，不同源的信息由神经网络自动学习。

1.2 数据增强

Zoph 等人在 2016 年通过四组在低资源的领域中的对比试验，率先证明了在低资源的情况下，神经机器翻译的效果不如统计机器翻译的效果[3]。对此很多国内外的专家都给出了自己的解决方案。这些方案主要分为两大类，一类是通过技术手段来扩充训练数据，使模型得到更为充分的训练。另外一类是通过将词的语义、语法等结构信息等同神经网络融合来提升网络的性能。本文主要关注第一类的解决方案，即在低资源的领域下基于已有的少量的训练数据，在保证句子语义结构的情况下来构造伪平行语料，使模型能够得到更加充分的训练。Fadaee 等人提出的对语料库中的罕见词进行建模，在保证句子的语义和语法结构的情况下，使得一个词可以被另一个罕见词进行替换[4]。

蔡子龙在 Fadaee 的工作基础之上提出了一种在保证句子语义和语法结构的情况下，通过对句子中的最小单元建模的方法[5]。该方法是通过对句子进行切分成最小单元建模的方法，利用余弦相似度的计算找到两个相似的模块，并通过对最小单元的对调之后生成了新的句子。

本文针对标准领域内的语料进行实验，在标准领域中，术语翻译不准的现象尤为严重，对于 Fadaee 等人的方法，我们不能只对语料中的低频词进行建模。且在标准领域中的语料有一定的书写规范，对于蔡子龙等人对调句子中的相似模块的的方法并不适用。基于标准领域内语料书写规范且结构性强的特点，本文在 Fadaee 和蔡子龙等人的工作基础之上，对标准翻译领域提出的基于术语的数据增强技术，即将句子的主干信息抽取出作为翻译模板，并将抽取出的术语和翻译模板进行重组，从而生成新的语料对模型进行增强。

2 神经机器翻译

机器翻译，又称自动翻译，是利用计算即将一种语言转换成另外一种自然语言的过程。基于神经网络的翻译是通过让神经网络自动学习语言的特征，找到输入和输出之间的关系，即利用神经网络来实现自然语言之间的映射。其核心问题是对条件概率进行建模：

$$p(y \mid x;\theta) = \prod_{n=1}^{N} p(y_n \mid x, y_{<n};\theta)$$

其中 y_n 是当前预测的目标词，x 是源语言词，$y_{<n}$ 是已生成的目标语言词。即根据源语言词的信息和已生成的目标语言词的信息来预测下一个目标词的过程。对该条件概率进行马尔可夫分解：

$$\begin{aligned} & p(y_n \mid x, y_{<n};\theta) \\ &= \frac{\exp(\varphi(y_n, x, y_{<n}, \theta))}{\sum_{y \in Y} \exp(\varphi(y, x, y_{<n}, \theta))} \\ &= \frac{\exp(\varphi(v_{y_n}, c_s, c_t, \theta))}{\sum_{y \in Y} \exp(\varphi(v_y, c_s, c_t, \theta))} \end{aligned}$$

其中 v_y 表示目标语言词向量，Y 表示目标语言词汇，c_s 表示源语言上下文向量，c_t 表示目标语言词向量。故而神经机器翻译的关键在于定义函数 φ，其技巧在于用向量表示句子，因此其最关键的是如何生成源语言上下文向量 c_s，目标语言上下文向量 c_t 来生成当前译文。

Kyunghyun Cho 和 Stuskever 在 2014 年先后提出了一种端到端的模型，模型直接对输入和输出建立了联系，前者的模型命名为 Encoder-Decoder 模型，后者的模型命名为 Sequence-to-Sequence 模型，其模型都是利用循环神经网络（RNN）来生成源语言和目标语言的上下文信息，并利用长短时记忆法（LSTM）来处理长距离依赖的问题，但是模型的任意长度的句子的向量都被编码成固定维度的向量。于是 Azmitry Bahdanau 等人在 2015 年提出了将注意力机制引入端到端的模型当中，这使得依赖关系的建模不再受到输入或输出序列中距离的制约。Vaswani 在 2017 年提出了一个新的网络架构 Transformer[6]，完全基于注意力机制的机器翻译，这种网络架构仅仅基于注意力机制，完全无需循环和卷积，其模型的性能优越，且更容易并行化处理，需要的训练时间更少。

尽管 Transformer 模型在机器翻译任务上取得了突破性的进展，但是其网络模型在低资源的机器翻译任务的条件下，模型的翻译效果一般。在基于语料库的条件

下，对翻译模型进行数据增强是一个值得研究的课题。一般对于数据增强的方法是对语料中随机位置进行截取之后利用神经网络进行训练或将句子当中的翻译知识加入到神经网络中例如句法知识，短语知识等。本文基于 Google 开源的 Tensor2Tensor 深度学习库中的 Transformer 模型来建模。针对语料进行处理，对平行语料库进行翻译模板和专业术语的抽取之后重组来生成新的语料。

3 整体框架

神经机器翻译在近年来取得了迅速的发展。但是神经机器翻译只有在大规模的语料库的情况下效果才明显，在低资源的领域中的效果一般，特别是在某些特定的领域中例如工业标准的翻译中的效果一般。针对上述问题，本文提出了一种基于语料库的数据增强技术，该技术主要是通过生成伪平行语料的方法来对数据进行扩充。该方法首先对语料进行术语和翻译模板的抽取，然后对抽取出来的术语和翻译模板进行重组，以这种方式来生成伪平行语料，并通过计算句子的困惑度来对质量进行把关，以实现通过扩充数据的方式来对模型的翻译性能的提升的目标。

3.1 数据增强流程图

鉴于术语的边界特征较为明显，有利于句子的整体结构的分析的前提下，本文主要基于“抽取➡重组”的框架来生成伪平行语料，图 2 给出了该方法的完整示例。

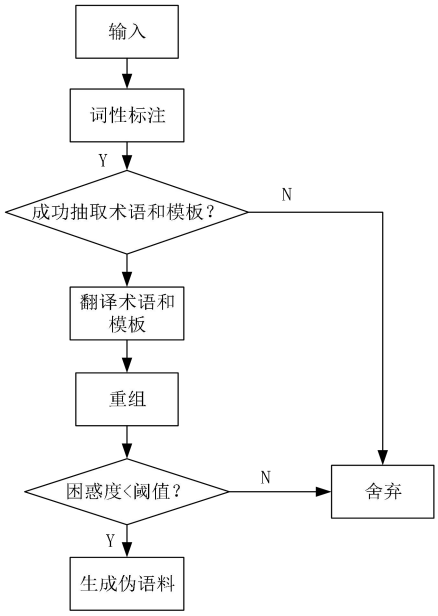

图 2 数据增强模型框架

Fig.2 Data augmentation model

在示例中，翻译模板保留了特殊的标识 ” $ ” ，通过对术语和槽内的内容进行重组，最终拼接出正确的伪平行语料。其具体细节将在 3.2 节和 3.3 节逐一说明。

3.2 抽取

抽取的任务是对平行语料中的翻译模板和领域术语进行抽取，本文处理的语料是工业标准文献中整理的语料。本文所说的术语是工业领域内相关的实体和概念的名称，工业标准领域的语料用词严谨，形式规范，结构性强，词与词之间不存在歧义现象。在一定程度上适合使用规则进行结构成分的识别。本文通过对语料的分析，定义了以下术语抽取规则，抽取规则如表 1 所示：

表 1 术语抽取规则

Tab.1 Term extraction rules

n+n	树脂添加剂
n+n+n	酚醛树脂夹层板
d+v+n+n	非包铝合金板材
b+n	增强型聚酯机
n+n+b+n	氟氯聚乙烯刮型板材料

对模板和术语进行抽取，其核心是对句子进行词性标注和词对齐。抽取任务以中文为基准，其模板抽取操作下表所示：

表 2 模板抽取过程

Tab.2 Template extraction

原句	类别 涉及 材料 的 燃烧 性能 。 Category refers to the combusition performance of materials .
词性标注	（类别（n）涉及（v）材料（n）的（u）燃烧性能（v））
词对齐	（NULL(4,7)类别(1)涉及(2,3)材料(8)燃烧(5)性能(6)。(9)）
模板抽取	$ 涉及 $ 的 燃烧性能。 $ refers to the combusition performance of $.

如上图所示，对源语言进行句法分析的结果为（类别（n） 涉及（v） 材料（n） 的（u）燃烧性能（v）），实验将名词作为核心词进行抽取，抽取结果为“类别、材料”，抽取出的模板为“$ 涉及 $ 的 燃烧性能”，根据词对齐结果，类别➡category、材料➡material，则例句对应目标语言端的翻译模板为“$ refers to the combusition performance of $ ”。

对语料进行术语的抽取过程举例如表 3 所示：

表 3：术语抽取过程

Tab.3 Term extraction process

原句	浸渍 玻璃 纤维 的 非结构性 成型 聚酯 树脂。 Nonstructural molded polyester resin impregnated glass fiber
词性标注	（浸渍（v）玻璃（n）纤维（n）的（v）非结构性（n）成型（v）聚酯（n）树脂（n））
术语抽取	（玻璃纤维（class fiber）、聚酯树脂（polyester resin））

如上图所示，输入的源语言为“浸渍 玻璃 纤维 的 非结构性 成型 聚酯 树脂”，其词性标注的结果为“（浸渍（v） 玻璃（n）纤维（n）的（v）非结构性（n）成型（v）聚酯（n）树脂（n）”，通过上述定义的术语抽取规则，抽取出的术语为“玻璃纤维”，“聚酯树脂”。

在抽取任务中存在两个重要的问题：

（1）词对齐结果中会出现源语言词没有对齐到目标语言词。

（2）词对齐结果中会出现源语言词对其到的目标语言词的位置不连续。

问题 1 会导致在术语和翻译模板重组的时候对应的目标语言端无法重组。问题 2 而导致的结果是在目标语言端进行重组时会导致句子原有的语义结构被破坏。针对以上问题，本文对出现以上问题的术语过滤掉，不进行抽取。

3.3 重组

重组的过程是对抽取出的翻译模板和在语料库中抽取出的术语库进行重新组合，在保证句子语法结构的合理性的同时，将抽取出的术语替换到句子框架中的相应的位置。为了保证生成伪平行语料的合理性，需要为术语找到一个适合的语境。因为在生成翻译模板而抽取核心词的时候，该核心词就包含了当前文本上下文的语义信息，故而将当前的核心词当作查询键，与从语料库中抽取出的术语进行余弦相似度的计算，并选取其 Top-K 作为当前核心词的重组的候选集。

其中术语库中的术语是由多个词组合而成的，我们将多个词组合成的术语作为一个整体，并通过预训练模型来获取其向量表征。重组过程如表 4 所示：

表 4：重组过程

Tab.4 Reorganization process

输入	夹层结构 和 芯材 的 实验方法
生成模板	$ 和 $ 的 实验方法
核心词及其候选集	夹层结构（夹层板，夹层组件）
	芯材（蜂窝芯材，芯材斜面区）
生成伪语料	夹层板 和 蜂窝芯材 的 实验方法
	夹层板 和 芯材斜面区 的 实验方法
	夹层组件 和 蜂窝芯材 的 实验方法
	夹层组件 和 芯材斜面区 的 实验方法

例如当前例句为“夹层结构 和 芯材的实验方法 ”，其抽取出的模板为“$ 和 $ 的实验方法 ”，其核心词为“夹层结构”、“芯材”，并利用核心词作为查询键在术语库中进行语义相似度的计算，并取其最相似的两个术语作为当前核心词的重组候选集，其中“夹层结构”的重组候选集为“夹层板，夹层组件”，“芯材”的候选集为“蜂窝芯材，芯材斜面区”，则模型重组之后会生成新的语料，然后对其对应的目标语言端进行重组。

但是本文实验在重组的过程中会出现以下两个问题。

（1）因为重组过程伪语料的数量是以指数的形式增加，会出现因指数爆炸的情况。

（2）在文本中抽取核心词的时候某些核心词与其前后词为固定搭配，会重组之后句子的合理性遭到破坏。

针对以上问题，本文采取如下方式解决，针对问题一，本文对生成翻译模板时抽取核心词的数量设置阈值，过滤掉核心词的数量超过阈值的语料。针对问题二，本文对核心词的重组候选集设置阈值，并将当前核心词一并加入重组的候选集当中，以保证生成伪语料的合理合法性。最终对生成的伪语料计算其句子的困惑度，对生成质量较差的句子进行过滤。

3.4 过滤

过滤的过程是对重组之后的生成的伪语料进行筛选，本实验通过计算生成伪语料的困惑度来进行质量筛选，本文选用 KenLM 模型对标准领域的语料进行训练，然后选择大小为 3 的窗口对语料进行困惑度的计算，其中困惑度的计算公式为：

$$ppl = P(w_1, w_2, ..., w_N)^{-1/N}$$

对等式量两边取对数后的公式为：

$$\log PPL(s) = \frac{-\sum_{i=1}^{N} \log P(w_1, w_2, ..., w_N)}{N}$$

其中标准领域的语料的困惑度分布情况如下图所示：

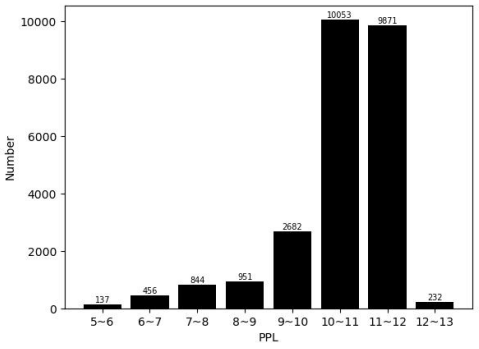

图 3 困惑度分布情况

Fig.3：Perplexity distribution

如图所示，图中 X 轴表示的是标准领域语料的困惑度的分布区间，Y 轴表示的是区间内语料的数量，其中标准领域的语料中的困惑度的阈值为 13，实验选取的困惑度阈值为标准语料中困惑度的阈值。

4 实验

4.1 语料设置

为了验证本文提出的数据增强技术的可行性，我们针对中英的任务，在工业标准的语料上进行实验。其中语料源自本实验组收集的中英工业标准的句对约 3 万句，实验过程中在工业标准的语料中随机抽取开发集、测试集。其语料信息如下表所示：

表 5 语料说明

Tab.5 Corpus description

工业标准语料	31816
训练集	25226
开发集	2851
测试集	3739

本实验利用哈工大平台开源的 LTP 工具(https://github.com/HIT-SCIR/ltp)对语料进行分词和词性标注。选用 Giza++模型(https://github.com/moses-smt/giza-pp)以中文为基准对语料进行词对齐。利用 KenLm 模型(https://kheafield.com/code/kenlm)来训练语言模型并计算句子的困惑度。

4.2 参数设置

本文使用的神经网络模型时 Google 开源的 Tensor2Tensor 深度学习库中的基于Transformer 的机器翻译模型，参数集为Transformer-base，其模型的主要参数如下表所示：

表 6 参数设置

Tab.6 Parameter settings

参数	值
learning_rate	0.2
Batch_size	4096
Adam_beta1	0.9
Adam_beta2	0.997
Train_steps	1200000

其中模型的训练步数实验设定为 120 万步，在生成翻译模板时抽取核心词的阈值为 10，每个核心词的重组候选集的阈值为 1，2。对术语训练的词向量维度为 728 维，并将困惑度（PPL）高于 13 的伪语料进行过滤。实验中利用科大讯飞联合实验室（HFI）开源的 Bert-wwm 预训练模型来获取术语的向量表征，实验以 Moses 工具中 multi-bleu.perl 脚本作为模型 BLEU 值的评测标准。

4.3 结果与分析

本文采用一种基于语料库的数据增强技术，该技术主要以术语和翻译模板的重组的形式生成伪平行语料进行训练，因此，能否准确的抽取出术语和模板直接影响到生成的伪平行语料质量。本文从来自标准领域的训练集中进行模板和术语的抽取，句子的相关信息如表 7 所示：

表 7 句子的相关信息

Tab.7 Sentence related information

句子的平均长度	18.9
翻译模板数	21898
术语数	10120
术语的平均长度	4.6

生成伪语料的数量如表 8 所示，生成伪语料示例如表 9 所示：

表 8 生成伪语料数量

Tab.8 Number of pseudo-corpora

阈值	伪语料数量
1	56949
2	120148

表 9 伪语料示例

Tab.9 Examples of pseudo-corpus

原句	本规范 涉及 利用 端纹轻木 作为 芯材夹板层 的 要求。
伪语料	本规范 涉及 利用 轻木原木 作为 芯材夹板层 的 要求。
	本规范 涉及 利用 端纹轻木 作为 夹层板面板 的 要求。
	本规范 涉及 利用 轻木圆木 作为 芯材夹板层 的 要求。

为了验证本文提出的数据增强实验可以提升模型的翻译质量,本文分别对 Fadaee 和蔡子龙等人的方法和本文提出的数据增强方法进行了对比。其中基线模型是指利用工业标准领域的中英双语语料训练出的翻译模型，数据增强模型是指在基线系统原有的训练集上融入了生成的伪平行伪语料进行训练的模型。其中融入伪语料之后语料的句子的平均长度如下表所示：

表 10 融合伪语料之后句子的平均长度

Tab.10 Average sentence length

阈值	句子的平均长度
1	20.78
2	23.75

本文实验采用 Moses 开源系统中集成的 BLEU 方法对译文的质量进行自动评价，实验结果如下表所示：

表 11 模型结果

Tab.11 Model results

语料	BLEU 值
基线系统	29.71
交换单元	29.85
低频词	29.88
基线系统+伪语料（阈值=1）	30.86
基线系统+伪语料（阈值=2）	30.90

由上表可以看出,Fadaee 和蔡子龙等人提出的数据增强方法在标准领域中模型的 BLEU 值分别有 0.14 和 0.16 的提升,而通过引入本实验提出的数据增强技术，在阈值等于 1 和阈值等于 2 时，相比于基线模型效果分别提升 1.15 和 1.19 个 BLEU 值，证明了本文提出的数据增强方法在标准翻译领域的有效性。

为了验证本实验提出的数据增强技术的有效性，本文对 WMT 官方提供的数据集进行了对比实验，实验选用的是 WMT17 提供的新闻数据集，其中训练集共 227603 句，开发集共 2002 句，测试集共 2002 句，其中模型生成伪语料数量如下表所示：

表 12 生成伪语料数量

Tab.12 Number of pseudo-corpora

阈值	伪语料数量
1	56949
2	120148

融合伪语料之后新闻领域的模型训练效果如下表所示：

表 13 模型结果

Tab.13 Model results

语料	BLEU 值
基线系统	10.32
基线系统+伪语料（阈值=1）	10.38
基线系统+伪语料（阈值=2）	10.57

实验结果表明，本文提出的数据增强技术在新闻领域指提升了 0.25 个百分点，因为在工业标准领域的语料用词严谨，形式规范，在一定程度上可以使用规则进行结构成分的分析，新闻领域内语言现象复杂，并不适用于利用规则进行句子成分分析，在利用本文提出的数据增强方法进行实验时，过滤掉了大部分生成的伪平行语料，所以实验效果有所提升，但提升效果并不明显，证明了本文提出的数据增强的方法在标准领域内的有效性。

为了进一步验证实验在工程领域的实用性，我们进行了如下对比实验，一种是从原有数据集中抽取术语生成伪平行语料，另外一种是利用本实验小组收集的工业领域当中的术语来生成伪平行语料，为了使模型有更好的泛化能力，我们将标准领域的数据集和 WMT 官方提供的联合国中英数据集共 15886041 句对融合之后进行训练。结果显示我们通过使用数据增强技术，其中模型 UN+标准+伪语料（集外词）是本实验利用收集到的工业术语与翻译模板生成的伪语料来训练出的模型，其翻译性能提升了 1.55 个 BLEU 值，模型 UN+标准+伪语料(集内词)从训练集内抽取出的术语和翻译模板进行重组之后生成的伪语料训练出的模型其 BLEU 值提升了 2.32 个点，证明了本文数据增强技术的有效性。

表 14 融合 UN 语料的模型效果

Tab.14 Model effect with UN corpus

En-Zh	BLEU 值
UN+标准	35.22
UN+标准+伪语料(集外词)	36.77
UN+标准+伪语料(集内词)	37.54

表 15 是我们从测试集中挑选的句子，通过示例我们可以看到，在使用了数据增强技术之后，神经机器翻译模型可以准确的预测出术语“粘接层”，而没有使用数据增强技术的模型将“粘接层”预测成了胶接线，原因是因为我们将“粘接层”这一术语跟翻译模板重组之后生成伪语料对模型进行了增强，所以模型效果相对于基线系统效果要好。

表 15 译文示例

Tab.15 Translation example

原文	Any adhesive bondline shall not have voids greater than 0.250 inch in the maximum dimension
参考译文	任何胶粘剂粘接层的气隙的最大尺寸不应大于 6.35 毫米（0.250 英寸）
UN+标准	所有胶粘剂胶接线在最大尺寸内的空隙不应大于 0.25 毫米(0.250 英寸）
UN+标准+伪语料（集内词）	所有胶粘剂粘接层在最大尺寸内的空隙不应大于 6.4 毫米（0.250 英寸）
UN+标准+伪语料(集外词)	任何胶粘剂胶接线的最大尺寸空隙不应大于 12.7（0.250 英寸）

5 总结与展望

本文针对当前神经机器翻译模型在低资源的情况下效果不理想的问题，提出了一种基于语料库的数据增强的技术。依据翻译模板和术语重组的思想，对平行语料进行词性标注和词对齐分析，进一步对翻译模板和术语进行抽取之后，在合理的语境下对术语和翻译模板进行重组之后生成伪平行语料，在一定程度上缓解了因语料不足而导致模型的泛化能力不足的问题。

实验结果表明，该方法通过对伪语料的生成，进而对训练数据进行扩展之后对模型带来了积极的影响，相对基线系统，BLEU 提升了 2.32 个值。然而该方法在模板抽取和术语抽取以及重组等方面都存在一定的错误，并且，例如对句子中的术语进行更加准确的识别等问题尚待解决。下一步的工作重心拟定在两个方面：首先，将本文中的术语抽取方法进行进一步的提升，其次，在翻译模板和术语进行重组的时候虽然设定了阈值，但是时间复杂度是呈指数形式增加，应在翻译模板和术语进行重组的时候对算法进行优化，在降低时间复杂度的时候可以得到质量较高的伪语料。

参考文献

[1] 李强, 黄辉, 周沁, et al. 模板驱动的神经机器翻译[J]. 计算机学报, 2019, 42(03):116-131.

[2] Chiang , David.Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 33(2):201-228.

[3] Zoph Barret,et al. Transfer learning for low-resource neural mathine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.Austin,Texas:Association for Computational Linguistics,2016.

[4] Fadaee M , Bisazza A , Monz C . Data Augmentation for Low-Resource Neural Machine Translation[J]. 2017.

[5] 蔡子龙, 杨明明, 熊德意. 基于数据增强技术的神经机器翻译[J]. 中文信息学报, 2018, 32(7).

[6] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. 2017.

[7] Kalchbrenner N , Blunsom P . Recurrent continuous translation models. 2013.

[8] Sutskever I , Vinyals O , Le Q V . Sequence to Sequence Learning with Neural Networks[J]. 2014.

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

[10] Cho K , Van Merrienboer B , Bahdanau D, et al. On the properties of neural machine translation: encoder=decoder approaches. 2014.

[11] Jean, Sébastien, Cho K , Memisevic R , et al. On Using Very Large Target Vocabulary for Neural Machine Translation[J]. 2014.

[12] Freitag M , Al-Onaizan Y . Fast Domain Adaptation for Neural Machine Translation[J]. 2016.

[13] Chu C , Dabre R , Kurohashi S . An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation[J]. 2017.

[14] Chu C , Wang R . A Survey of Domain Adaptation for Neural Machine Translation[J]. 2018.

[15] Ding L , He Y , Zhou L , et al. Combining

Domain Knowledge and Deep Learning Makes NMT More Adaptive[J]. 2017.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, Neural machine translation by jointly learning to align and translate. CoRR, abs/1490.0473, 2014.

Neural Machine Translation Based on Data Enhancement Technology

Abstract： Since the rise of neural machine translation, it has continuously achieved better translation results. But the neural network has a bottleneck, which is that it can achieve good results only on the premise of large-scale parallel corpora, and the effect on low-resource areas is not good. From the perspective of data enhancement technology, this paper uses the idea of translation templates to identify and extract nouns or noun phrases in sentences, retaining the backbone of sentences. Then, the pseudo-parallel corpus is generated by reorganizing the extracted term set on the main skeleton of the sentence. Finally, the generated pseudo-corpus is checked by calculating the large confusion of the sentence, and a pseudo-corpus of good quality is generated. This method effectively alleviates the problem of insufficient generalization ability of the model due to insufficient corpus. The experimental results show that compared with the baseline system, the BLEU value improved by this method is 2.32.

Key words： Translation template；Neural machine translation；Data augmentation；Pseudo-corpus