计算机科学
COMPUTER SCIENCE

融入新能源领域术语知识的机器翻译方法

董振恒，任维平，游新冬，吕学强

引用本文

董振恒，任维平，游新冬，吕学强. 融入新能源领域术语知识的机器翻译方法[J]. 计算机科学, 2022, 49(6): 305-312.

DONG Zhen-heng, REN Wei-ping, YOU Xin-dong, LYU Xue-qiang. Machine Translation Method Integrating New Energy Terminology Knowledge[J]. Computer Science, 2022, 49(6): 305-312.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500117

融入新能源领域术语知识的机器翻译方法

董振恒[1]　任维平[2]　游新冬[1]　吕学强[1]
1 北京信息科技大学网络文化与数字传播北京市重点实验室　北京 100101
2 北京信息科技大学外国语学院　北京 100192
(dongzhenheng1@163.com)

摘　要　在领域机器翻译中,领域术语能否被正确翻译对翻译质量起着决定性作用,有效地将领域术语融入到神经机器翻译模型中,提升领域术语的翻译质量具有实际意义。文中提出了一种将新能源领域术语信息作为先验知识融入神经机器翻译中的方法,以新能源领域双语术语知识库构建的术语字典为媒介,提出并比较了两种不同的知识融入方式:1)术语替换,即在源语言端使用目标端术语替换源端术语;2)术语添加,即在源语言端将源端术语与目标端术语拼接,并在源语言端与目标语言端均使用作为特殊外部知识的标识符来标识目标端术语的开头与结尾。以新能源领域中英文双语对齐语料以及构建的中英文对齐术语库为数据基础进行了实验,结果表明,在测试集上,所提方法的BLEU值比基线实验分别高出6.38与6.55,证明了所提方法能有效地将领域术语知识融入到翻译模型中,提升了领域术语的翻译质量。

关键词:领域机器翻译;领域术语;特殊标识;先验知识;术语替换;术语添加
中图法分类号　TP391

Machine Translation Method Integrating New Energy Terminology Knowledge

DONG Zhen-heng[1],REN Wei-ping[2],YOU Xin-dong[1] and LYU Xue-qiang[1]
1 Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,Beijing Information Science and Technology University,Beijing 100101,China
2 School of Foreign Languages,Beijing Information Science and Technology University,Beijing 100192,China

Abstract　In domain machine translation,whether domain terms can be translated correctly plays a decisive role in translation quality. It is of practical significance to effectively integrate domain terms into neural machine translation model and improve the translation quality of domain terms. This paper proposes a method to integrate the term information in the field of new energy into neural machine translation as a priori knowledge. Taking the term dictionary constructed by the bilingual term knowledge base in the field of new energy as the medium,this paper puts forward and compares two different ways of knowledge integration:1)term replacement,that is,replacing the source term with the target term at the source language end;2)term addition refers to the splicing of source side terms and target side terms at the source language side,the identifier as special external knowledge is used to identify the beginning and end of the target term at both the source language end and the target language end. Experiments are carried out based on the Chinese and English bilingual alignment corpus in the field of new energy and the constructed Chinese and English alignment corpus. The results show that on the test set,the Bleu value of the proposed method is 6.38 and 6.55 higher than that of the baseline experiment respectively,which proves that the proposed method can effectively integrate the domain term knowledge into the translation model and improve the translation quality of domain terms.

Keywords　Domain machine translation,Domain terms,Special identification,Prior knowledge,Term replacement,Term append

到稿日期:2021-05-17　返修日期:2021-12-10
基金项目:北京市自然科学基金(4212020);国家自然科学基金(61671070);北京信息科技大学"勤信人才"培育计划项目(QXTCPB201908);北京市市教委科研计划(KM202111232001)
This work was supported by the Natural Science Foundation of Beijing,China(4212020),National Natural Science Foundation of China(61671070),Qin Xin Talents Cultivation Program of Beijing Information Science & Technology University(QXTCPB201908) and Research Planning of Beijing Municipal Commission of Education(KM202111232001).
通信作者:任维平(renweiping@bistu.edu.cn)

1 引言

机器翻译于20世纪30年代被提出，经过70多年的发展，如今取得了突破性进展。从最初基于规则的机器翻译，再到基于统计的机器翻译，在深度学习技术不断发展的推动下，机器翻译取得了长足进步，神经机器翻译逐渐成为翻译领域的主力军[1-3]。神经机器翻译模型一般包括编码器和解码器两部分，编码器负责将源语言映射成向量表示[4]，解码器则根据向量表示生成译文。google团队[5]于2017年提出了仅通过自注意力机制进行编码和解码的Transformer模型，该模型取得了机器翻译的最佳效果。已有的神经机器翻译模型在通用领域取得了较高的翻译质量，但是对于涉及到专业术语的特定领域，其翻译结果存在着大量术语漏翻、错翻的现象[6]，术语翻译的效果仍存在较大的提升空间。表1列出了新能源领域源句子与其翻译结果的示例，其中目前流行的神经机器翻译模型Transformer将源端术语“充电桩”错误地翻译为“charge bank”，而正确的术语翻译为“charging pile”。

表1 翻译示例

Table 1 Translation examples

Source	一种新型新能源充电桩
Target	A novel new energy **charging pile**
Transformer	A novel new energy **charge bank**

对神经机器翻译而言，如何准确翻译特定领域的专业术语，仍是一项具有挑战性的任务。在翻译时将特定领域的术语信息作为先验知识融入到神经机器翻译模型中，提升领域术语的翻译效果一直是神经机器翻译研究的热点问题。

基于此，本文提出了一种在源语言端通过术语替换与术语添加两种不同的方式，将术语信息作为先验知识融入到神经机器翻译中的方法，该方法将目标端术语作为源语言端额外的先验知识输入，让翻译模型学习到源端术语与目标端术语之间的语义关系，在一定程度上减少了领域术语翻译错误的现象。尤其是外部知识的引入，不仅带来了翻译性能的提升，也增加了神经机器翻译的可解释性[7]。

本文以目前流行的Transformer为翻译模型，以新能源领域中英文专利对齐语料与自建中英文术语知识库为数据支撑，对领域机器翻译展开了研究，本文的主要贡献如下：

(1)提出将新能源领域术语作为先验知识的信息输入，在新能源领域机器翻译中融入领域术语知识。

(2)通过自定义双语术语知识库来构建领域术语字典，在源语言端实现目标端术语的替换与添加，并使用作为特殊外部知识的标识符“〈S〉”“〈E〉”来标识目标端术语的开头和结尾，从而扩充源端术语信息。

(3)在新能源领域的专利数据上进行了实验。结果表明，本文方法可以使翻译模型充分学习领域术语信息，有效提升了特定领域译文的翻译质量。

本文第1节为引言部分；第2节介绍了相关工作，包括机器翻译概述与外部知识融入神经机器翻译中的方法研究；第3节介绍了目前神经机器的主流模型Transformer；第4节详细阐述了融入新能源领域术语知识的领域机器翻译研究，介绍了实验过程，阐述了实验结果，并对翻译效果展开了实验分析；最后总结全文并展望未来。

2 相关工作

机器翻译指通过计算机将源语言句子翻译成与之语义等价的目标语言句子的过程，是自然语言处理领域的一个重要研究方向[8]。在领域机器翻译中，将领域先验知识引入到神经机器翻译中，可以指导神经机器翻译的学习过程，使翻译模型学到更多的领域知识，从而有效提升其翻译性能，得到更高质量的翻译结果。

先验知识包括标注数据、翻译规则、双语词典等，是非常重要的翻译知识，也是一种有效处理未登录词、命名实体翻译、术语翻译的方法[9]。

Tang等[10]于2016年在神经机器翻译中融合双语短语信息，使得模型在解码时不仅可以生成目标单词，还可以生成目标短语，改变了以往翻译模型只能生成单词的翻译现象，提升了术语以及实体的翻译效果。Arthur等[11]将双语词典间接融入神经机器翻译中，结合注意力机制将从双语词典中获取到的词语翻译概率转换为目标词语的预测概率，间接地将双语词典作为外部先验知识融入到翻译中，提高了对实词的翻译质量。

Wang等[12]于2017年将短语记忆作为知识集成到神经机器翻译中，指导翻译模型对短语的解码，提高了神经机器翻译对短语的翻译质量。Zhang等[13]于2018年提出使用对数线性框架将双语词典、短语表等外部知识集成到神经机器翻译中，在源端注入了更多言语特征，提升了目标语言的翻译质量。

Han等[14]于2019年提出了一种将单词翻译作为先验知识融入神经机器翻译的方法，使用不同的编码器将单词翻译融入到源语言中，以增强源端信息，从而提升翻译效果。Dinu等[15]使用集成的方法在源端进行术语知识扩充，在模型训练层融入先验知识，提升了术语的翻译质量。

Qiao等[16]于2020年提出将语义角色作为先验知识融入神经机器翻译中，在源端横向编码器和纵向编码器中利用额外的语义角色信息改进传统的神经机器翻译模型，解决了语义漏译、不连续翻译等问题。Cao等[17]于2020年提出将翻译记忆作为先验知识融入神经机器翻译中的方法，该方法将目标端翻译记忆拼接到源语句中，不仅实现了数据规模的扩充，还将翻译记忆传入神经机器翻译模型中，提高了模型的翻译性能。同年，Qin等[7]提出将规则信息转化为近似等价的序列信息，以此将规则等外部知识融入到神经机器翻译模型中，使用规则约束来提升译文的翻译质量。

Zhang等[18]于2021年提出将句法信息作为知识融入到神经机器翻译中，通过自监督双语句法对齐的方式，利用源语言与目标语言单词之间的隐藏状态来建模双语句法关系，将源语言与目标语言的句法结构精确对齐，将句法、语义等先验知识有效地融入模型中，利用对齐后的句法结构信息之间的相互依赖性来提高翻译的性能。Chen等[19]于2021年将编解码器中的注意力权重作为知识，提出了基于词对齐指导的词汇约束神经机器翻译方法，该方法在外部目标约束神经机器翻译的基础上使用对齐词汇约束来指导模型解码，通过

更精准的词汇对齐约束，提升对齐单词的翻译准确性，从而进一步提升翻译性能。

基于以上工作，本文以将领域术语信息作为先验知识为出发点，借助领域术语知识库，通过在源语言端使用目标端术语替换源端术语以及将源端术语与目标端术语进行拼接的方式，将目标端术语作为先验知识融入到源语言中，并使用特殊符号“〈S〉”“〈E〉”来标识替换、添加在源端的目标端术语（“〈S〉”表示术语开头，“〈E〉”表示术语结尾），将目标端术语与源端成分分割，以防止神经机器翻译模型在训练时造成先验知识混淆。

融入先验知识的源语言，在包含领域术语信息输入的同时，标识符也作为特殊的外部知识，标识源语言中的术语信息，使翻译模型能更好地学习目标端术语与源语言之间以及源端术语与目标端术语之间的语义对应关系。作为先验知识的目标端术语，在模型训练时指导术语解码过程，在生成译文时充分考虑术语信息，以达到对术语进行正确翻译的效果。此外，由于本文方法主要涉及数据层面，因此使用目前主流的Transformer模型就能取得很好的术语翻译效果。

3 基于Transformer的神经机器翻译系统

Transformer是Google团队于2017年提出的一种基于注意力机制的经典模型[5]，该模型摒弃了传统循环神经网络的顺序结构，使用自注意力机制并且在拥有全局信息的基础上进行并行化训练，翻译质量和翻译速度均最佳。因此，本文选取Transformer对翻译模型展开实验。

Transformer模型采用序列到序列（Seq2Seq）的模型架构，整体结构由编码器（Encoder Block）和解码器（Decoder Block）两部分组成。模型的编码器将源语言进行编码后转换为向量表示，解码器接收来自编码器的编码信息，然后进行解码生成译文。Transformer的整体结构如图1所示。

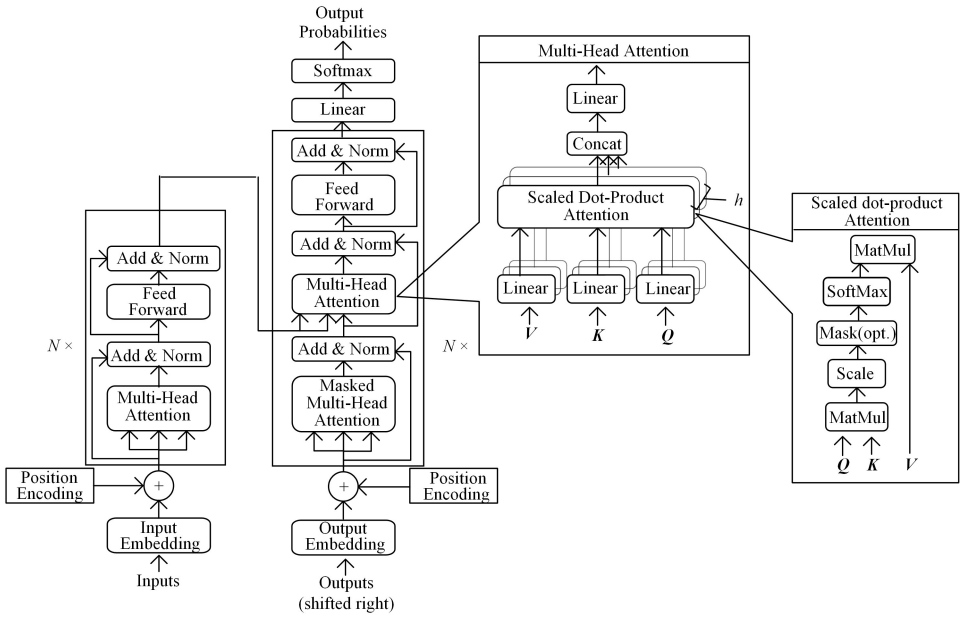

图1 Transformer的结构

Fig. 1 Transformer structure

编码器由6个相同的层堆叠组成，每层内部包含多头注意力子层（Multi-HeadAttention）和前馈神经网络子层（Feed Forward），这两个子层通过残差连接（Residual Connection）和层标准化（Layer Normalization）连接。Transformer中使用位置编码（Positional Encoding）保存单词在序列中的位置，以识别语言中的顺序关系，其中位置编码PE的计算式如式（1）、式（2）所示：

$$PE_{(pos,2i)}=\sin(pos/10\,000^{2i/d}) \tag{1}$$

$$PE_{(pos,2i+1)}=\cos(pos/10\,000^{2i/d}) \tag{2}$$

其中，pos表示单词在句子中的位置，d表示位置嵌入的维度，i表示字向量的维度序号。对于输入的源句子$X=(x_1,x_2,\cdots,x_n)$，通过词嵌入向量与位置编码相加之后得到句子中所有单词的向量表示，如式（3）所示：

$$X_{\text{Embedding}}=WordEmbedding(X)+PositionalEncoding(X) \tag{3}$$

注意力子层中存在的3个线性变换矩阵$\boldsymbol{W}_Q$，$\boldsymbol{W}_K$，$\boldsymbol{W}_V$分别对输入的向量表示进行3次线性变换，衍生出查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$以及值矩阵$\boldsymbol{V}$，并进一步得到X的注意力输出，如式（4）－式（7）所示：

$$\boldsymbol{Q}=Linear(X_{\text{Embedding}})=X_{Embedding}\cdot\boldsymbol{W}_{\boldsymbol{Q}} \tag{4}$$

$$\boldsymbol{K}=Linear(X_{\text{Embedding}})=X_{\text{Embedding}}\cdot\boldsymbol{W}_{\boldsymbol{K}} \tag{5}$$

$$\boldsymbol{V}=Linear(X_{\text{Embedding}})=X_{Embedding}\cdot\boldsymbol{W}_{\boldsymbol{V}} \tag{6}$$

$$X_{\text{attention}}=Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

将自注意力层中的线性变换矩阵$(\boldsymbol{W}^Q,\boldsymbol{W}^K,\boldsymbol{W}^V)$定义为多组，即$(\boldsymbol{W}_0^Q,\boldsymbol{W}_0^K,\boldsymbol{W}_0^V)$，$(\boldsymbol{W}_1^Q,\boldsymbol{W}_1^k,\boldsymbol{W}_1^V)$，计算后得到多组注意力向量，将多组注意力向量进行拼接，经过残差连接与层标准化后输入前馈神经网络中，具体计算式如式（8）、式（9）所示：

$$Z_i=head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V),i=1,\cdots,8 \tag{8}$$

$$Z=MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(Z_1,Z_2,\cdots,Z_8)W^O \tag{9}$$

前馈神经网络由两层全连接网络子层组成，第一层使用

RELU 函数作为激活函数，第二层是线性激活函数，如式(10)所示：

$$FFN(Z)=\mathrm{Max}(0,ZW_1+b_1)W_2+b_2 \tag{10}$$

其中，$\boldsymbol{b}_1$，$\boldsymbol{b}_2$，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$为模型参数。

与编码器类似，解码器同样由 6 个相同的层堆叠组成，解码器每层由两个多头注意力子层和一个前馈神经网络子层组成。第一个多头注意力子层采用了 Masked 操作，被称为 Masked Multi-Head Attention，在翻译当前顺序时用于隐藏下一顺序的信息；第二个多头注意力子层的 $\boldsymbol{K}$，$\boldsymbol{V}$ 矩阵来自编码器的编码信息矩阵，而 $\boldsymbol{Q}$ 矩阵则使用解码器中上一个多头注意力层的输出。在每一子层都进行了残差连接和层归一化处理。

4 融入术语知识的领域机器翻译

4.1 实验数据

本文以新能源领域双语对齐语料库以及构建的双语术语知识库为数据基础对领域机器翻译展开了研究。双语对齐语料库 Pat_Corpus 是根据专利数据搜索引擎 Actionable patent 网站而构建的，中文语料库的大小为 17.5 MB，英文语料库的大小为 26.1MB，两个语料库共包含中英文双语对齐语句 116095 对[1)]。表 2 列出了新能源领域中英文专利语句对实例。

表 2 中英文语句对实例

Table 2 Examples of Chinese and English sentences

中文	英文
新能源侧包括太阳能集热子系统、光伏发电子系统、风力发电子系统和高温燃料电池堆	the energy side comprises a solar heat collection subsystem, a photovoltaic power generation system, a wind power generation system and a high-temperature fuel cell stack
新能源汽车智能语音提示系统	intelligent voice prompting system for new energy automobile
本发明公开了一种传动平稳的新能源汽车高效驱动装置，包括驱动电机和变速箱	the invention discloses an efficient driving device for a new energy automobile with stable transmission. the efficient driving device comprises a driving motor and a gearbox

术语知识库则使用深度学习技术来构建神经网络以对中文专利文本进行术语抽取[20]，将中文术语经过 WIPO 翻译得到对应的英文术语，中文术语库的大小为 713 kB，英文术语库的大小为1 071 kB，两个语料库共包含中英文术语对 39 861 对，表 3 列出了新能源领域中英文专利术语对实例。

表 3 中英文术语对实例

Table 3 Examples of Chinese and English terms

中文术语	英文术语
混合动力电动车	hybrid electric vehicle
太阳能电池模块	solar cell module
开关电路	switch circuit

将 Pat_Corpus 按照 8∶1∶1 的比例划分为训练集、验证集、测试集，数据具体划分情况如表 4 所列，语料库、术语知识库的具体信息如表 5 所列。

表 4 数据集划分的详细信息

Table 4 Detailed information of data set division

数据集	训练集	验证集	测试集
Pat_Corpus	9.2876×10^5	11609	11610

表 5 语料库、术语知识库的统计信息

Table 5 Statistical information of corpus and terminological database

语料库	规模/条	大小
中文语料库	116095	17.5 MB
英文语料库	116095	26.1 MB
中文术语库	39861	713 kB
英文术语库	39861	1071 kB

4.2 实验过程

4.2.1 新能源领域术语知识融合

在本文实验中将中文数据作为源端语言，将英文数据作为目标端语言。对于中文数据，采用术语替换与术语添加两种不同的方法将术语信息作为先验知识融入到神经机器翻译中。使用 jieba 分词工具并借助中文专利术语库来指导分词，有效地保留了中文专利数据中的术语信息，避免了分词不准确造成的术语信息丢失，从而影响对术语的翻译效果，为术语替换以及术语添加提供了基础数据。

(1)术语替换：在源语言端使用目标端术语替换源端术语，并用符号“〈S〉”“〈E〉”对目标端术语进行标识。本文将此方法记为 SE-Replace。

(2)术语添加：在源语言端的源端术语后面添加目标端术语，并用符号“〈S〉”“〈E〉”对目标端术语进行标识。本文将此方法记为 SE-Append。

对于英文数据，使用 NLTK 分词工具进行分词，对其中的术语也进行标识，将分完词后的英文数据进行标准化操作。融合术语知识的翻译实例如表 6 所列。

表 6 融合术语知识的翻译实例

Table 6 Examples of translation integrating terminology knowledge

Source	Zh	一种新型新能源充电桩
Segmentation	Zh	一种新型新能源充电桩
SE-Replace	Zh	一种新型〈S〉new energy〈E〉〈S〉charging pile〈E〉
SE-Append	Zh	一种新型新能源〈S〉new energy〈E〉充电桩〈S〉charging pile〈E〉
Target	En	A novel〈S〉new energy〈E〉〈S〉charging pile〈E〉

表 6 中，Source 表示源语言，Segmentation 表示术语字典指导分词后的结果，SE-Replace 表示本文提出的术语替换方法，SE-Append 表示本文提出的术语添加方法。通过术语替换与术语添加的方式将领域术语信息作为先验知识融入到神经机器翻译中，并使用符号“〈S〉”“〈E〉”分别标识替换、添加在源语言中的目标端术语的开头和结尾。使用术语替换的方法使翻译模型在训练时学习目标端术语与源语句的语义

1) https://github.com/ZhenHengDong/Patent_Data

关系；使用术语添加的方法使模型在训练时能更充分地学习源端术语与目标端术语之间的对应关系，以进一步提高术语翻译结果的准确性。以融入新能源领域先验术语知识的翻译实例为例，翻译模型的训练过程如图 2 所示。

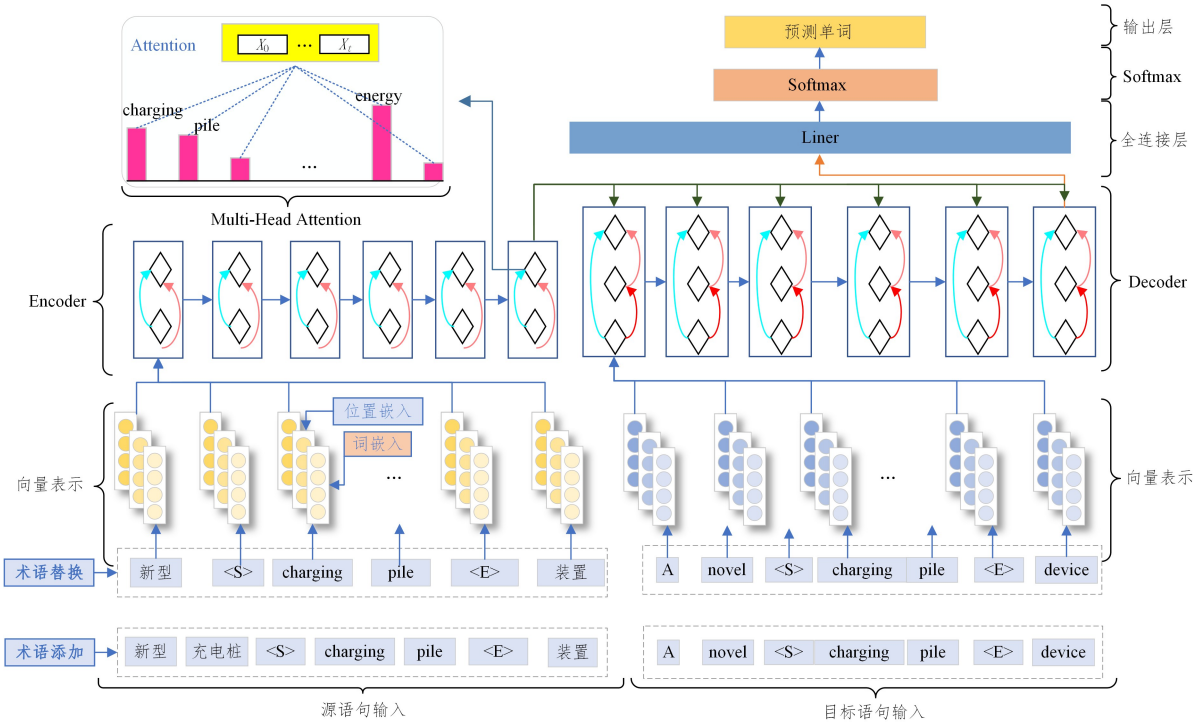

图 2　模型的训练过程

Fig. 2　Model training process

4.2.2　对比实验

为了验证所提方法的有效性，本文设计了以下几种对比实验。

(1)Baseline：源语言中未携带任何作为先验知识的术语信息，将仅经过新能源领域术语指导分词后的数据作为基线实验数据。

(2)Replace：仅替换，使用目标端术语替换源端术语，源语言中携带目标端术语知识信息，未使用任何标识符对源语言成分进行标识。

(3)Append：仅添加，在源端术语后面拼接目标端术语。源语言中携带源端术语、目标端术语知识信息，未使用任何标识符对源语言中的语言成分进行标识。

(4)Sub-Replace：使用目标端术语替换源端术语，源语言中携带目标端术语知识信息，借鉴 Dinu 等的思想，使用标识符对源语言成分进行标识，其中下标 0 标识源语言部分，下标 1 标识源端术语，下标 2 标识目标端术语。

(5)Sub-Append：在源端术语后面拼接目标端术语。源语言中携带源端术语、目标端术语知识信息，使用标识符对源语言中的成分进行标识，其中下标 0 标识源语言部分，下标 1 标识源端术语，下标 2 标识目标端术语。

(6)SE-Replace：本文提出的术语替换方法。在源语言端使用目标端术语替换源端术语，并用符号“〈S〉”“〈E〉”对目标端术语进行标识。

(7)SE-Append：本文提出的术语添加方法。在源语言端与源端术语后面添加目标端术语，并用符号“〈S〉”“〈E〉”对目标端术语进行标识。

对比实验的翻译数据实例如表 7 所列。

表 7　对比实验翻译数据实例

Table 7　Examples of translation data in comparative experiments

Source	Zh	一种新型新能源充电桩
Target	En	A novel new energy charging pile
Replace	Zh	一种新型 new energy charging pile
Append	Zh	一种新型新能源 new energy 充电桩 charging pile
Sub-Replace	Zh	一种$_0$新型$_0$new energy$_2$ charging pile$_2$
Sub-Append	Zh	一种$_0$新型$_0$新能源$_1$ new energy$_2$ 充电桩$_1$ charging pile$_2$
SE-Replace	Zh	一种新型〈S〉new energy〈E〉〈S〉charging pile〈E〉
SE-Append	Zh	一种新型新能源〈S〉new energy〈E〉充电桩〈S〉charging pile〈E〉

4.3　实验设置

本文将 Facebook 团队开源的 fairseq[21] 作为基础工具，在数据处理阶段，根据训练语料生成词表以及数据二值化，中文词表的覆盖率为 97%，英文词表的覆盖率为 98.6%，未登录词用〈UNK〉表示；在模型训练阶段，将 Transformer 模型作为实验模型。模型中编码器与解码器的层数设置为 6，多头注意力数量为 8，其中词向量的维度大小设置为 512，源端的隐层大小和词表大小与目标端相同，分别设置为 38 016 和 16 656。实验采用 adam 优化器进行更新参数，其中$\beta_1=0.9$，$\beta_2=0.98$，初始学习率设置为 5×10^{-4}，为防止训练时过拟合，将神经元随机失活率(dropout)设置为 0.3，并且按照词的数量划分批次，将 max_tokens 设置为 4 096，评价指标使用 BLEU[22] 值，保存 BLEU 值最高的模型。

4.4 实验结果

为了更好地体现所提方法的有效性，在划分训练集、验证集和测试集时设定了不同的随机种子，按划分比例多次进行实验，训练轮次设定为 160，模型训练的超参数设置相同，记录每次实验在验证集和测试集上的 BLEU 得分，求其平均值，并将其作为最终的分数。

为了更加直观地展现不同方法的性能表现，图 3 与图 4 以 BLEU 值为纵坐标，展现了不同方法在验证集与测试集上的 BLEU 值及变化趋势。

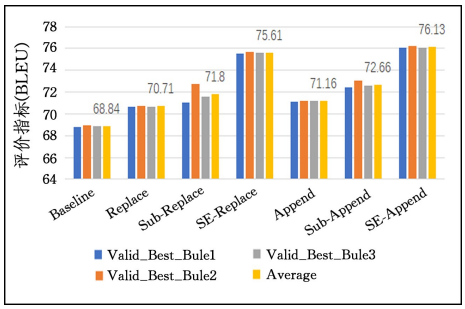

图 3 验证集翻译表现

Fig. 3 Validation sets translation performance

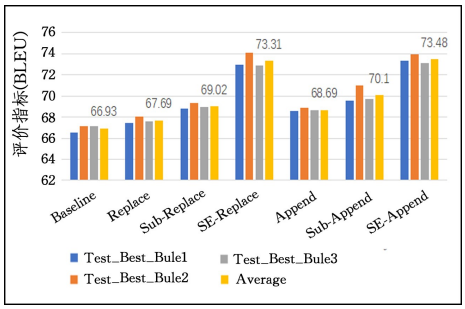

图 4 测试集翻译表现

Fig. 4 Test sets translation performance

实验结果表明，与 Baseline 相比，携带先验术语知识的 Replace 和 Append 的方法在翻译性能上均有提升；使用 Replace 的方法进行 3 次实验后的均值在验证集和测试集上分别获得了 1.87 和 0.76 个 BLEU 点的提升；采用 Append 的方法进行 3 次实验后的均值在验证集和测试集上分别获得了 2.92 和 1.76 个 BLEU 点的提升。Replace 和 Append 方法验证了在神经机器翻译中融入先验知识的有效性，在源端中融入先验术语知识可以有效提升神经机器翻译的翻译效果。

与 Replace 和 Append 方法相比，Sub-Replace 和 Sub-Append 方法在翻译性能上有了进一步的提升。与 Replace 相比，Sub-Replace 在验证集和测试集上分别提升了 1.09 和 1.33 个 BLEU 点；与 Append 相比，Sub-Append 在验证集和测试集上分别提升了 1.5 和 1.41 个 BLEU 点。Sub-Replace 与 Sub-Append 方法验证了在将术语信息作为先验知识融入到神经机器翻译的基础上，使用标识符标识源语言中的句子成分，以区分源端术语与目标端术语的有效性。

与 Sub-Replace 和 Sub-Append 方法相比，本文提出的 SE-Replace 和 SE-Append 方法在翻译性能上再次得到了提升。与 Sub-Replace 相比，SE-Replace 在验证集和测试集上分别再次提升了 3.81 和 4.28 个 BLEU 点；与 Sub-Append 相比，SE-Append 在验证集和测试集上分别再次提升了 3.47 和 3.38 个 BLEU 点。这验证了本文提出的 SE-Replace 和 SE-Append 方法在源语言与目标语言中均使用更为简洁的标识符“〈S〉”“〈E〉”来标识目标端术语，将目标端术语与源语言和目标语言成分分割，并将特殊标识符作为额外信息输入的有效性。

与 Baseline 相比，本文提出的 SE-Replace 和 SE-Append 方法在翻译性能上均有明显提升。与 Baseline 相比，使用 SE-Replace 进行 3 次实验后的均值在验证集和测试集上分别获得了 6.77 和 6.38 个 BLEU 点的提升，采用 SE-Append 方法进行 3 次实验后的均值在验证集和测试集上分别获得了 7.29 和 6.55 个 BLEU 点的提升，并且与 SE-Replace 方法相比，其在验证集和测试集上进一步提升了 0.52 和 0.17 个 BLEU 点。实验结果进一步表明了所提方法在将术语信息作为先验知识融入到神经机器翻译的基础上，使用简单的标识来标识目标端术语，并将标识符作为额外信息输入的有效性。

4.5 实验分析

本文方法将源端术语直接替换为目标端术语，将目标端术语信息拼接在源端术语后面，并使用标识符号“〈S〉”“〈E〉”分别标识目标端术语的开头和结尾，以此将术语信息作为先验知识以及将标识符作为额外信息融入到神经机器翻译模型中。为了进一步比较在源端进行术语替换以及术语添加的差异性，我们对 Baseline，Append，Replace，Sub-Replace，Sub-Append 以及本文提出的 SE-Replace 和 SE-Append 方法展开了实验分析。图 5、图 6 分别给出了是否在源语言端融入术语知识以及融入术语知识后如何对其标识等方法在训练过程中的模型迭代次数以及对 BLEU 取均值之后的折线图。表 8 列出了不同方法针对同一译文的翻译结果。

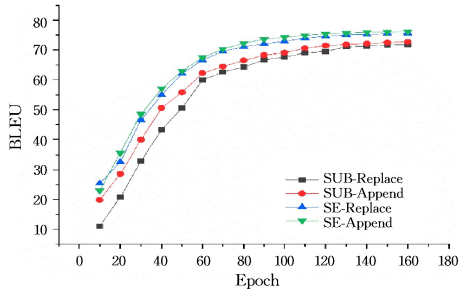

图 5 Epoch-BLEU 变化图 1

Fig. 5 Epoch-BLEU change chart 1

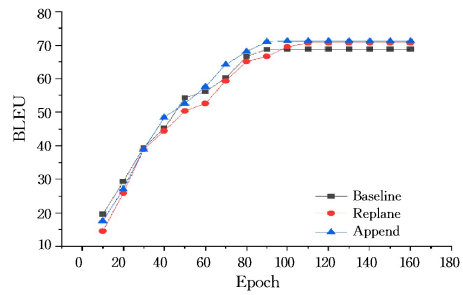

图 6 Epoch-BLEU 变化图 2

Fig. 6 Epoch-BLEU change chart 2

表 8 翻译示例

Table 8 Translation examples

	Source	所述微电机通过转动齿轮与插条活动连接
	Target	The **micromotor** is movably connected with the **insertion strip** through a **rotating gear**
	Terms	微电机转动齿轮插条
实验对比	Baseline	所述微电机通过转动齿轮与插条活动连接
	Replace	所述 micromotor 通过 rotating gear 与 insertion strip 活动连接
	Replace	所述微电机 micromotor 通过传动齿轮 rotating gear 与插条 insertion strip 活动连接
	Sub-Replace	所述$_0$micromotor$_2$通过$_0$rotating$_2$ gear$_2$与$_0$insertion$_2$strip$_2$活动连接$_0$
	Sub-Replace	所述$_0$微电机$_1$ micromotor$_2$通过$_0$传动齿轮$_1$ rotating$_2$ gear$_2$与$_0$插条$_1$ insertion$_2$ strip$_2$活动连接$_0$
	SE-Replace	所述〈S〉micromotor〈E〉通过〈S〉rotating gear〈E〉与〈S〉insertion strip〈E〉活动连接
	SE-Append	所述微电机〈S〉micromotor〈E〉通过传动齿轮〈S〉rotating gear〈E〉与插条〈S〉insertion strip〈E〉活动连接
译文结果	Baseline	The **micromotor** are movably connected with the **patch** through a **rotate rods**
	Replace	The **micromotor** is movably connected with the **insert strip** through a **connecting rods**
	Append	The **micromotor** is movably connected with the **insert strip** through a **rotating rods**
	Sub-Replace	The **micromotor** are movably connected with the **insertion gears** through **connecting rods**
	Sub-Append	The **micromotor** is movably connected with the **inserting strip** through **connecting gear**
	SE-Replace	The 〈S〉**micromotor** 〈E〉 is movably connected with the 〈S〉 **inserting strip** 〈E〉 through 〈S〉**connecting gear** 〈E〉
	SE-Append	The〈S〉**micromotor**〈E〉 is movably connected with the〈S〉**insertion strip**〈E〉through 〈S〉**rotating gear** 〈E〉

如图 5、图 6 所示，Transformer 模型在迭代训练中逐渐趋于收敛。仅进行过普通分词处理的 Baseline 收敛较快，在训练 80 个轮次之后开始趋于平稳，其原因可能是新能源领域中英文专利文本表达的固定性以及 Transformer 模型较强的学习能力，使得 Transformer 模型在短时间内就学会了中英文专利文本的对应表达，从而在较短时间内就趋于稳定。从实验效果来看，由于源语言中没有作为先验知识的术语信息输入，翻译模型在训练时不能充分学习源端术语与目标端术语的对应关系，仅能将简单的术语翻译正确，因此对于较长的术语翻译效果不佳。

使用 Replace 和 Append 的方法分别训练了 100 与 90 个轮次之后趋于稳定，其原因可能是源语言中包含了目标端术语信息，与 Baseline 相比，模型需要进行较长时间的训练才能从中学会替换与添加后的表达。从实验效果看，与 Baseline 相比融入先验知识的源语言携带了目标端术语信息，翻译时提升了对部分术语的翻译效果。但是未有任何标识符将目标端术语与句子成分分割开来，在训练时不能充分地将术语知识融入翻译模型中。

Sub-Replace 与 Sub-Append 方法分别在训练了 140 与 130 个轮次之后才趋于稳定，其原因可能是中英文表达的差异性以及使用了多个数字下标标识源语言中的句子成分，造成了知识混淆，加大了翻译模型的学习难度，从而给神经网络的学习造成了困难。从实验效果来看，源语言中虽携带了术语信息，但是由于源语句中的标识符过多而将句子中所有成分都进行标识，并且把目标端术语作为单个词语而不是整体进行标识，造成了模型对源端术语与目标端术语学习不充分，模型仅能将部分术语翻译正确，对于复杂的术语翻译效果不佳。

SE-Replace 在训练至 120 个轮次时逐渐趋于平稳，其原因可能是将目标端术语直接替换为源端术语，并将其作为新的语料，造成了源语句中的中英文表达较为复杂，但是由于仅使用了简单标识符对目标端术语进行标识，因此其训练时间长于未携带任何标识符的 Replace，短于携带了多个复杂标识符的 Sub-Replace。从实验效果来看，与 Sub-Replace 相比，使用简单标识符将目标端术语作为整体进行标识，在训练时能较好地将术语知识融入翻译模型中，使得模型能够较好地学习目标端术语与源语句的语义关系。

SE-Append 在训练了 110 个轮次时就已经趋于平稳，其原因可能是将目标端术语添加在源端术语后面，在源语言中融入了更多的术语知识。与 SE-Replace 相比，SE-Append 不会对中文的表达产生影响，并且标识符简单，因此模型的收敛时间长于 Append，短于 SE-Replace。而从实验效果来看，SE-Append 的效果在 SE-Replace 的基础上得到了进一步的提升，其原因可能是术语添加可以保留中文表达的完整信息与表达习惯，丰富了源端术语的表达信息，并且在源语言与目标语言中通过标识符仅对目标端术语进行标识，避免因标识的成分过多而造成知识的混淆。翻译模型能更充分地学习中英文术语之间的语义对应关系，从而获得最好的实验结果。

结束语 针对面向领域的机器翻译任务，本文以新能源领域为例展开了领域机器翻译研究，提出了一种通过术语替换与添加的方式将新能源领域术语信息作为先验知识融入到神经机器翻译中的方法，在此基础上，又使用标识符“〈S〉”“〈E〉”对目标端术语进行标识，将目标端术语与源语言、目标语言的语句成分分割，将标识目标端术语开头和结尾的标识符“〈S〉”“〈E〉”作为额外知识输入，指导翻译过程。实验证明，使用目标端术语替换源端术语以及通过拼接目标端术语到源端术语之后可以有效地将领域术语作为先验知识融入神经机器翻译中，提升了术语翻译效果；标识符作为特殊的外部知识会给予神经网络额外的学习指导，神经机器翻译模型学习源语言与目标语言之间的对应关系时，会对特殊符号标识过的目标端术语给予重点关注，从而提升被标记术语的翻译质量，在保证领域术语被正确翻译的基础上提升了整体翻译质量。

本文方法的主要改动在数据层面，相比模型修改，本文方法简单、实用且易于实现，并且在实验中取得了明显的效果。此外，本文方法在确保领域对齐语料与领域术语知识库的前提下在提升领域术语翻译质量方面具有通用性。在下一步的工作中，我们将探索在领域机器翻译中如何使用多个编码器对术语词典、句法信息等外部先验知识进行建模，在模型层面将领域特性更深层次地融入到神经机器翻译模型中。

参 考 文 献

[1] JUNCZYS-DOWMUNT M, DWOJAK T, HOANG H. Is neural machine translation ready for deployment? A case study on 30 translation directions[J]. arXiv:1610.01108, 2016.

[2] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv:1609.08144, 2016.

[3] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]// International Conference on Machine Learning. PMLR, 2017:1243-1252.

[4] BRITZ D, GOLDIE A, LUONG M T, et al. Massive exploration of neural machine translation architectures[J]. arXiv: 1703.03906, 2017.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv:1706.03762, 2017.

[6] LIU F, LU H, NEUBIG G. Handling homographs in neural machine translation[J]. arXiv:1708.06510, 2017.

[7] QIN W J, XIONG D Y. Neural machine translation with rule information[J]. Journal of Xiamen University(Natural Science), 2020, 59(2):185-191.

[8] FENG Y, SHAOCH Z. Review on the frontier of neural machine translation[J]. Journal of Chinese Information Processing, 2020, 34(7):1-18.

[9] LI Y, XIONG D, ZHANG M. review of neural machine translation[J]. Chinese Journal of Computers, 2018, 41(12): 2734-2755.

[10] TANG Y, MENG F, LU Z, et al. Neural machine translation with external phrase memory[J]. arXiv:1606.01792, 2016.

[11] ARTHUR P, NEUBIG G, NAKAMURA S. Incorporating discrete-translation lexicons into neural machine translation[J]. arXiv:1606.02006, 2016.

[12] WANG X, TU Z, XIONG D, et al. Translating phrases in neural machine translation[J]. arXiv:1708.01980, 2017.

[13] ZHANG J, LIU Y, LUAN H, et al. Prior Knowledge In-tegration for Neural Machine Translation using Posterior Regularization[J]. arXiv:1811.01100, 2018.

[14] HAN D, LI J H, ZHOU G D. Neural machine translation based on word translation[J]. Journal of Chinese Information Processing, 2019, 33(7):40-45.

[15] DINU G, MATHUR P, FEDERICO M, et al. Training neural machine translation to apply terminology constraints[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:3063-3068.

[16] QIAO B W, LI J H. Neural machine translation with semantic roles[J]. Computer Science, 2020, 47(2):163-168.

[17] CAO Q, XIONG D Y. Fusion method of translation Memory and neural machine translation based on data expansion[J]. Journal of Chinese Information Processing, 2020, 34(5):36-43.

[18] ZHANG T, HUANG H, FENG C, et al. Self-supervised bilingual syntactic alignment for neural machine translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(16):14454-14462.

[19] CHEN G, CHEN Y, LI V O K. Lexically constrained neural machine translation with explicit alignment guidance[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 12630-12638.

[20] SUN T, CHEN H T, LV X Q, et al. Research on Term Extraction of New Energy Patent Text[J/OL]. Journal of Chinese Computer Systems. [2021-07-16]. http://kns.cnki.net/kcms/detail/21.1106.TP.20210511.1556.002.html.

[21] OTT M, EDUNOV S, BAEVSKI A, et al. fairseq: A fast, extensible toolkit for sequence modeling[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:48-53.

[22] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002:311-318.

DONG Zhen-heng, born in 1995, postgraduate. His main research interests include natural language processing and machine translation.

REN Wei-ping, born in 1962, professor. Her main research interests include applied linguistics and so on.

(责任编辑:何杨)