doi:10.6043/j.issn.0438-0479.202412009

融合领域双语词典的泰-汉法律领域神经机器翻译方法

李 畅[1,2],高盛祥[1,2*],余正涛[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,云南 昆明 650500)

摘要:[**目的**]对于泰-汉法律领域神经机器翻译,由于训练所需平行语料稀缺,缺乏双语词级别的有效语义监督,难以做到一些实体词以及领域术语等领域关键词的精准翻译.针对该问题,本文提出一种融合领域双语词典的神经机器翻译方法.[**方法**]首先利用法律领域语料库挖掘领域双语词典;然后使用基于不频繁法律领域中文词覆盖的方法筛选额外伪法律领域训练数据;在模型输入端,利用源句匹配的领域双语词典目标词构建额外提示数据;在模型结构中,引入提示浅编码器对提示进行表征,在源句编码器层中加入新的交叉注意力模块强化源句的表征,将源句表征与提示表征拼接作为最终编码器表征,在解码器端对编码器表征的交叉注意力作用下指导产生目标词序列.[**结果**]基于本文筛选方法混合训练的 Transformer 模型相比于仅用领域数据训练的模型,BLEU 值提高了 0.54 个百分点;采用该混合训练集,融入提示信息的模型相比 Transformer 模型 BLEU 值又提升了 0.90 个百分点,并且相比于经典"语码转换"方法训练的模型 BLEU 值提升了 0.61 个百分点.[**结论**]本文数据筛选方法可在解决领域专业术语稀疏问题的同时降低通用高频词对翻译结果的干扰,提升领域翻译基线模型的性能;引入提示信息的模型能够有效地与提示进行交互,实现翻译性能的提升,并使领域关键词翻译更准确.

关键词:法律领域;泰-汉神经机器翻译;领域双语词典;数据筛选;提示浅编码器;交叉注意力
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2025)04-0597-09

Legal field neural machine translation method for Thai-Chinese by integrating domain bilingual dictionaries

LI Chang[1,2],GAO Shengxiang[1,2*],YU Zhengtao[1,2]
(1. Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China;
2. Key Laboratory of Artificial Intelligence of Yunnan Province,Kunming 650500,China)

Abstract:[**Objective**] For Thai-Chinese neural machine translation in the legal field,achieving accurate translation of domain-specific keywords,such as entity names and specialized terms,appears challenging due to the scarcity of parallel corpora for training and the lack of effective semantic supervision at the bilingual word level. To address this issue,we propose a neural machine translation method that incorporates bilingual domain-specific dictionaries. [**Methods**] First,we extract bilingual dictionaries from the legal corpus. Then,a method is employed based on recovering rare Chinese legal terms to filter additional pseudo-legal training data. At the model's input stage,we further construct supplementary prompt data by matching target words from the domain-specific bilingual dictionary with source sentences. Then,within the model architecture,we introduce a prompt-shallow encoder to represent the prompts. Next,a new cross-attention module is added to the source sentence encoder layer to enhance source sentence representation. Finally,the source sentence and prompt representations are concatenated to form the final encoder representation,which guides the

收稿日期:2024-12-11 **录用日期**:2025-06-03
基金项目:国家自然科学基金(U23A20388,U21B2027,62366027);云南省重点研发计划(202303AP140008,202401BC070021,202203AA080004);云南省基础研究项目(202301AT070393);昆明理工大学"双一流"科技专项(202402AG050007);云南省科技人才与平台计划(202105AC160018)
*通信作者:gaoshengxiang.yn@foxmail.com
引文格式:李畅,高盛祥,余正涛.融合领域双语词典的泰-汉法律领域神经机器翻译方法[J].厦门大学学报(自然科学版),2025,64(4):597-605.
Citation:LI C,GAO S X,YU Z T. Legal field neural machine translation method for Thai-Chinese by integrating domain bilingual dictionaries[J]. J Xiamen Univ Nat Sci,2025,64(4):597-605. (in Chinese)

generation of the target word sequence through cross-attention in the decoder. [**Results**] Experimental results show that the Transformer model trained with this filtering method has achieved a 0.54 percentage points improvement in BLEU score compared to the model trained solely on domain data. Using this mixed training set, we find that the model incorporating prompt information witnesses a BLEU score increase of 0.90 percentage points compared to the standard Transformer model, and 0.61 percentage points compared to the model trained with the traditional "code-switching" method. [**Conclusions**] This paper's data filtering approach simultaneously resolves domain term sparsity and suppresses generic word noise, ultimately improving baseline domain translation performance. The model incorporating prompt information effectively interacts with these prompts, thus improving overall translation performance and increasing accuracies of domain-specific keyword translations.

Keywords: legal field; Thai-Chinese neural machine translation; domain bilingual dictionary; data filter; prompt shallow encoder; cross-attention

中国与泰国的民间交流和官方合作日渐频繁,无论是经济合作、政治外交,还是文化互动,都需要充分了解泰国的法律条文,以降低合作风险,提升效率. 但泰语作为小语种,长期以来存在较大的翻译人才缺口,尤其随着近来跨境电商的飞速发展,缺口更加明显. 同时,法律领域的翻译还需具备法律专业知识,对人才的要求更高. 为了缓解这一问题,降低翻译成本,提升工作效率,有必要对提高泰-汉法律领域神经机器翻译(NMT)的质量进行研究.

目前对于特定领域的 NMT 主要采用混合训练或者领域适应方法得到领域翻译基线模型,然后通过融合领域知识进一步提升模型翻译质量. 混合训练一般将数据增强[1]得到的额外数据与原始领域数据混合训练基线模型,进而通过提供额外上下文信息以提升模型性能. 领域适应的方法则通过模型的参数微调或增加模型的额外约束将通用领域预训练的模型迁移到特定领域,使现有模型适应特定领域数据而无需重新混合训练. Luong 和 Manning[2]在大规模数据上预先训练源领域模型,然后用特定领域数据进行参数微调训练基线模型实现领域迁移;Khandelwal 等[3]将特定领域数据输入通用域模型生成翻译上下文,并存储作为领域翻译记忆库,当翻译领域数据时,检索领域翻译记忆库的最近邻生成词汇,对模型训练进行约束而实现领域迁移.

因此,领域适应的本质是在学习通用知识的基础上同时学习特定领域的知识,其性能由大规模通用源域数据和领域数据共同决定. 当源领域和目标领域的数据分布差异较大时,在源领域学到的知识可能无法直接应用到目标领域,甚至会发生负迁移[4]. 考虑到泰-汉这样的低资源语言对难以获取覆盖较广的大规模通用数据,获取的数据分布往往和法律领域差异较大,因此从易获得的外部平行语料中筛选与领域数据相似的伪领域数据,通过混合训练方式得到领域翻译基线模型是一个比较可行的方案. 法律领域文本通常包含大量的实体词、专业术语等领域词,而面向法律领域的泰-汉 NMT 任务平行语料稀缺,有些领域词出现的频率较低,仅利用平行语料训练传统 Transformer 的 NMT 方法[5]缺乏双语词级别的有效语义监督,难以实现这些领域源词到目标词的精准映射,从而在自回归生成目标序列时导致词汇翻译不准确甚至句子语义偏差[6]. 通过词汇约束等方式融合领域双语词典,可以使模型学习利用领域词汇级别的知识辅助目标词汇以及序列的生成.

由于 NMT 模型的端到端特性,直接对其施加约束是具有挑战性的,因此许多研究探索了词汇约束的 NMT 方法,一般将这些研究分为硬约束和软约束. 硬约束通过设计新颖的解码算法强制添加目标约束词出现在生成序列中. 无需更改模型结构和训练过程,但会大大增加解码时间成本,并且强制约束出现在翻译中缺乏灵活性,导致当存在噪声约束时会解码生成错误词汇. 如 Hokamp 等[7]为了实现用户对生成内容的精确控制,提出在解码时加入词汇约束的网格波束搜索算法,其计算复杂度显著高于传统束搜索. 为了降低计算复杂度,Post 等[8]提出动态波束分配算法,Hasler 等[9]使用对齐来获得目标侧约束对应的源词,同时使用有限域机和多堆栈[10]解码来引导波束搜索. 但以上方法均需要较高的概率剪枝阈值,可能丢失优质候选,同时显存占用较大. 软约束不是在解码时强制生成约束词汇,而是在训练时学习如何额外利用约束词汇指导目标词的生成,并在推断中利用约束词汇生成翻译. Dinu 等[11]通过在训练数据内部匹配位置连接或替换术语对的方式将目标约束引入源语句中,对源句原词、目标约束以及对应的源词赋予不同的因子,并与词嵌入拼接共同输入编码器,将术语约束转化为模型内部任务,避免解码时的硬性约束,兼具效率与灵活性. Song 等[12]创建一个源端用目标约束词替换对应源词的合成转换语料库,对 NMT 的训练进行数据增强,并且通过指针网络综合学习从目标约束

中复制约词汇和从目标语言词表中选择词汇的行为. Wang 等[13]使用约束词的派生词、约束词和自由词的表示组成的模板以及自由词的派生词,通过更改模型输入序列,使得模型学习生成目标模板和目标自由词的派生,再据此以及目标约束词的派生生成自然目标语句. 这些方法都只能处理一个源词显式对应一个目标约束词的情况,因此 Chen 等[14]将所有约束词汇连接到源句之后构建增强数据,引入了分段嵌入来增强模型对源句子和约束词汇之间区分的能力,模型通过指针网络学习在没有显式对齐信息的情况下自动利用约束指导词汇生成. 这些软约束方法通过修改训练过程来实现,可以更好地平衡解码质量和速度,使模型学习如何更加灵活地利用约束辅助指导生成过程,而不是强制约束词汇生成.

基于此,本文提出了一种融合领域双语词典的泰-汉法律领域 NMT 方法. 在领域翻译基线模型训练方面,使用领域双语词典通过基于不频繁法律领域中文词覆盖的方法筛选额外伪法律领域训练数据,并通过混合训练 Transformer 模型获得领域翻译基线模型. 在领域知识融合方面,使用该混合训练集,首先在输入端利用领域双语词典构建源端提示数据,然后在 Transformer 模型上引入提示浅编码器对源端提示进行表征,在源句编码器层中加入新的交叉注意力层强化源句的表征,最后将源句表征与提示表征拼接作为最终编码器表征.

1　方法描述

为提升泰-汉法律领域翻译的效果,本文提出了融合领域双语词典的 NMT 方法,总体上可以分为如图 1 所示的 4 个过程. 首先从构建的领域平行语料中挖掘领域双语词典;然后进行伪领域数据筛选来扩充训练集,用于训练领域翻译基线模型;在模型输入端,每个源句匹配领域双语词典构建提示数据;最后在模型方面,将提示数据进行编码,融入进领域翻译基线模型中.

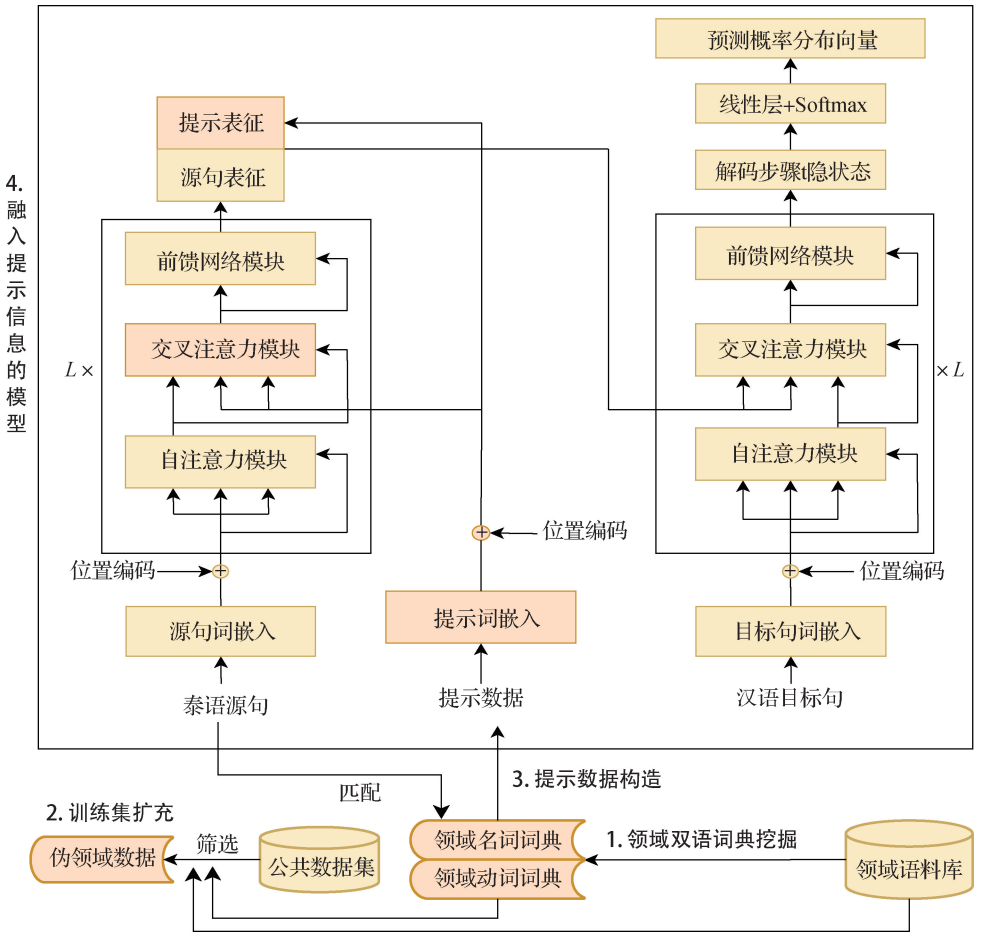

图 1　融合领域双语词典的 NMT 方法的整体框架

Fig. 1　Overall framework of NMT method by integrating domain bilingual dictionaries

1.1 领域双语词典挖掘

首先获取泰-汉词对齐词典,考虑到低资源神经机器翻译模型性能较差,难以通过基于神经机器翻译模型的词对齐模型获取高质量词对齐,因而采用基于统计的 fast_align(https://github.com/clab/fast_align)工具进行词对齐抽取.具体地,获得每个已分词泰-汉句对对应泰-汉词在句子中下标的对齐,从而获得每个句对对应的泰-汉词对齐,然后整合所有句对的词对齐,经过调用翻译接口与人工筛选获得初步的泰-汉词对齐词典.

然后从中筛选泰-汉领域双语词典:对泰语使用 pythainlp(https://github.com/PyThaiNLP/pythainlp)工具进行分词;对中文端使用 jieba(https://github.com/fxsjy/jieba)工具精准模式进行分词,利用 jieba.posseg 进行词性标注,获得中文名词 n、动词 v 2 种词性词,整合去重后构建 nv 词性词典,并且使用 jieba.analyse.textrank 获取中文关键词;中文句子中既属于关键词又属于 nv 词性词典的词定义为中文领域关键词,从泰-汉词对齐词典中筛选含中文词领域关键词的词对构成领域双语词典.

最后通过人工筛选删除与翻译结果对比词义差异较大且具有明显错误的少部分词典噪声.人工筛选的标准是词典中的中文词必须是实体、领域名词和动词术语等领域关键词,以进一步提升词典的质量.领域双语词典挖掘的流程图如图 2 所示.

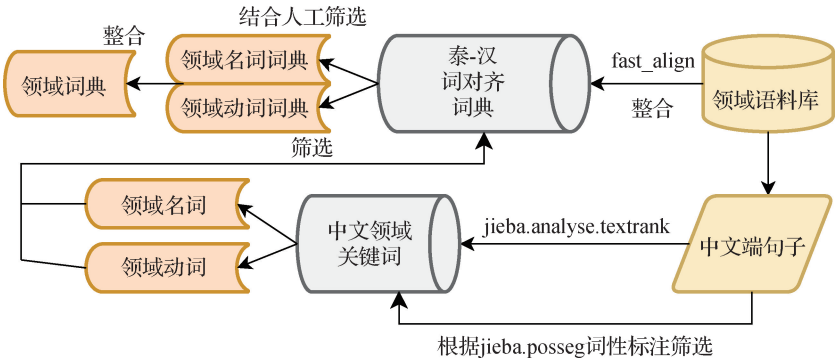

图 2 领域双语词典挖掘流程

Fig. 2 Flowsheet for mining bilingual domain-specific dictionary

1.2 训练集扩充

受 Xu 等[15]基于模糊匹配和 Poncelas 等[16]基于 n-gram 的语料筛选方法的启发,在获得领域双语词典之后,本文利用该词典通过基于模糊匹配和基于不频繁法律领域中文词覆盖的两种筛选算法,从公共数据中筛选固定数量的泰-汉伪领域双语数据,并与原始领域数据共同作为领域翻译训练集.其中将公共数据的中文端含有领域词的句对加入候选池.

对于基于模糊匹配的筛选算法,计算法律领域句子对中文端 t_{in} 与候选池公共数据集中句子对中文端 t_{out} 基于编辑距离的模糊匹配得分:

$$S_{\text{edit}}(t_{\text{in}},t_{\text{out}})=1-\frac{d_{\text{edit}}(t_{\text{in}},t_{\text{out}})}{\max(l(t_{\text{in}}),l(t_{\text{out}}))},\tag{1}$$

其中 l 表示求字符数.选择分数高于阈值的公共数据句对 t_{out} 加入伪领域数据集,并且从候选池中移除,之后的每个法律领域中文句子 t_{in} 对候选池中的句对迭代选择直到选择到指定数据量.

对于基于不频繁法律领域中文词覆盖的筛选方法,计算候选池公共数据集中每个句对的分数,选择其中分数最高的固定数量的句对作为伪领域数据.句对分数 score 的计算方法:统计公共数据集中中文端 t_{out} 中存在领域词 w_{d} 的出现次数 $c(w_{\text{d}})$,设置一个中文领域词次数阈值 C_{T},如果领域词出现次数低于阈值则该领域词在该句对的分数(式(2))设为其出现次数与阈值之间的差距,反之则设为 0,句对分数为中文端存在的领域词分数之和.

$$\text{score}=\sum_{w_{\text{d}}\in t_{\text{out}}}\max(0,C_{\text{T}}-c(w_{\text{d}})).\tag{2}$$

基于不频繁法律领域中文词覆盖的筛选方法可以在解决领域专业术语稀疏问题的同时,降低通用高频词(如"的""规定")对翻译结果的干扰.

1.3 提示数据构造

首先对泰语源文本进行词性标注,本文采用 spaCy-Thai(https://github.com/KoichiYasuoka/spaCy-Thai)工具标注出泰语词汇的名词和动词;然后根据泰语词词性和领域双语词典构建提示数据.例如某泰语词在句子中属于名词词性,且该源词匹配到领域名词词典中对应的目标中文词,则将该词性标签和目标词一起组成一个提示,源句匹配的所有提示一起构成提示数据(构建流程如表 1 所示).

表 1　提示数据构造流程

Tab. 1　Prompt data construction process

伪代码:提示数据构造伪代码

输入:源句 X,名词领域双语词典 D_{N},动词领域双语词典 D_{V}
P_{T} 初始化为空字符串
对于 X 中的每个词 w:
　如果 w 的词性标记是 NOUN 且 w 在 D_{N} 中:
　　对于 w 匹配到每个目标词 C:
　　将标签<NP>和 C 连接到 P_{T}
　如果 w 的词性标记是 VERB 且 w 在 D_{V} 中:
　　对于 w 匹配到每个目标词 C:
　　将标签<VP>和 C 连接到 P_{T}
输出:提示数据 P_{T}

例如,源句“หลักเกณฑ์ และ วิธีการ ที่ กำหนด ไว้ ใน กฎหมาย ระเบียบ มติ คณะรัฐมนตรี และ คำ สั่ง ที่ เกี่ยวข้อง โดย เคร่งครัด”对应目标句为“严格遵守法律、法规、内阁决议和相关命令规定的标准和方法.”构建的提示数据为“<NP>法律<NP> 法规<NP>内阁<NP>决议<VP>规定<NP>标准”,其中“กฎหมาย”对应“法律”,“ระเบียบ”对应“法规”,“คณะรัฐมนตรี”对应“内阁”,“มติ”对应“决议”,“กำหนด” 对应“规定”,“หลักเกณฑ์”对应“标准”,<NP>标签连接的是领域名词,<VP>标签连接的是领域动词.

1.4　融入提示信息的模型

在 Transformer 结构[5]的领域基线模型基础上,模型输入为源句 X、提示数据输入 P_{T} 和目标句 Y 的 (X,P_{T},Y)三元组,将融入领域双语词典提示信息的模型预测概率分布定义为

$$P(Y\mid X,P_{\mathrm{T}};\theta)=\sum_{t=1}^{T}\log\ p(y_t\mid y_{<t},X,P_{\mathrm{T}};\theta). \tag{3}$$

如图 1 所示,提示数据输入的编码器不同于 Transformer 编码器,该编码器对提示数据进行浅层编码(式(4)),在词嵌入层后加上独立位置编码,其中提示词共享目标词嵌入,提示标签作为源语言词表的一部分.

$$\boldsymbol{H}_{P_{\mathrm{T}}}=\mathrm{WordEmbedding}(P_{\mathrm{T}})+\mathrm{PositionEncoding}(P_{\mathrm{T}}). \tag{4}$$

为了充分利用提示数据,在每一层源句编码器层的自注意力模块后新增一个额外的交叉注意力模块,即注意力机制 Prompt_Attention(见式(6),其中 $\boldsymbol{H}^{l}_{\mathrm{self}}$ 为该层自注意力模块输出,$\boldsymbol{H}^{l-1}$ 为该层源句编码器的输入)以及残差连接和后归一化层,该模块输出为 $\boldsymbol{H}^{l}_{\mathrm{cross}}$(见式(7)).

$$\boldsymbol{H}_{\mathrm{self}}{}^{l}=\mathrm{LayerNormalization}(\boldsymbol{H}^{l-1}+\mathrm{Attention}(\boldsymbol{H}^{l-1},\boldsymbol{H}^{l-1},\boldsymbol{H}^{l-1})), \tag{5}$$

$$\mathrm{Attention}(\boldsymbol{H}^{l}_{\mathrm{self}},\boldsymbol{H}_{P_{\mathrm{T}}},\boldsymbol{H}_{P_{\mathrm{T}}})=\mathrm{softmax}\left(\frac{\boldsymbol{H}^{l}_{\mathrm{self}}\cdot\boldsymbol{H}_{P_{\mathrm{T}}}{}^{\mathrm{T}}}{\sqrt{d_{\mathrm{model}}}}\right)\cdot\boldsymbol{H}_{P_{\mathrm{T}}}. \tag{6}$$

每一层源句编码器层额外交叉注意力模块输出通过前馈网络(feed-forward network,FFN)得到该层源句编码器的输出表征 $\boldsymbol{H}^{l}$(见式(8)).

$$\boldsymbol{H}^{l}_{\mathrm{cross}}=\mathrm{LayerNormalization}(\boldsymbol{H}^{l}_{\mathrm{self}}+\mathrm{Attention}(\boldsymbol{H}^{l}_{\mathrm{self}},\boldsymbol{H}_{P_{\mathrm{T}}},\boldsymbol{H}_{P_{\mathrm{T}}})), \tag{7}$$

$$\boldsymbol{H}^{l}=\mathrm{LayerNormalization}(\boldsymbol{H}^{l}_{\mathrm{cross}}+\mathrm{FFN}(\boldsymbol{H}^{l}_{\mathrm{cross}})). \tag{8}$$

将新源句编码器最后一层输出表征 $\boldsymbol{H}^{L}$ 与提示编码器表征 $\boldsymbol{H}_{P_{\mathrm{T}}}$ 进行拼接作为最终编码端表征 $\boldsymbol{H}$,这样当前向传播经过解码器的交叉注意力模块时,解码器除了对源句编码器表征部分计算注意力之外,该交叉注意力还引入了额外对提示信息的注意力.

$$\boldsymbol{H}=\mathrm{Concat}(\boldsymbol{H}^{L},\boldsymbol{H}_{P_{\mathrm{T}}})=[\boldsymbol{H}^{L};\boldsymbol{H}_{P_{\mathrm{T}}}]. \tag{9}$$

由于低资源下模型的脆弱性,干预基于生成概率的目标函数易出现过拟合.本文额外添加提示的方式不会改变源句,不必担心因改变源语句带来的复杂性或语义缺失问题;同时在提示数据中灵活地添加多个目标词,使模型学习目标提示信息强化源端表征,与目标端信息交互,进而使提示知识继续在解码器端交叉注意力模块中传播,指导翻译上下文表示的生成,从而产生更加精准的领域目标词汇.

2　实　验

2.1　数据集构建

首先采样两种方式从网络中获取法律领域泰-汉双语语料:一是通过网络爬取直接获取部分法律领域网站的泰-汉平行语料;二是通过下载网络 PDF 和 DOC 文件直接提取泰语文本,以及使用 OCR 识别工具和语种识别工具识别图片中泰语文本,例如专业网站数据(https://law.m-society.go.th/law2016/law/group_list?page=2),然后去除文本中的非法字符和多余空格,去除含有过多英文字符和数字的文本,去除过短文本和空行,去重后得到高质量泰语文本,并且调用翻译接口得到泰语文本的中文翻译,从而获得泰-汉双语语料.对得到的所有泰-汉法律领域双语数据进行二次清洗,去除中文端中文文字过短的句对,经过人工标注得到最终高质量法律领域泰-汉双语语

料. 对于公开泰-汉平行语料,进行同样清洗,移除空行,去除源或目标文本含非法字符或英文的句对,去除中文端中文文字过短和过长的句对,移除重复的句对,得到最终的公开泰-汉平行语料.

对于法律领域泰-汉数据集,人工选取 2 000 对验证集和 2 000 对测试集,剩余 143 902 对作为训练集. 对于伪领域数据,根据初始领域语料库和领域双语词典(其中名词词典大小为 3 859,动词词典大小为 2 140,总词典大小为 5 768,总词典大小小于名词和动词词典大小的总和是因为有些词既是名词又是动词,本文构造提示数据的方法不会导致匹配重复)在公开泰-汉平行语料中筛选 50 000 对伪法律领域句对,最终得到 193 902 对泰-汉伪领域训练数据集,初始领域语料的 2 000 对验证集和测试集作为最终验证集和测试集. 数据集规模如表 2 所示.

表 2 数据集的划分

Tab. 2 The divided of dataset

数据集	训练集	验证集	测试集
领域数据集	143 902	2 000	2 000
伪领域数据集	193 902	2 000	2 000

2.2 实验参数设置

本文方法在 Fairseq 开源框架基础上实现,所有 NMT 模型上均使用 Transformer 模型作为实验的架构,实验在 NVIDIA GeForce RTX 3090 GPU 上进行. 对于伪领域数据筛选,基于模糊匹配筛选方法的分数阈值设为 0.3,基于不频繁中文法律领域词的方法的频率阈值设为 3. 为避免 NMT 模型改变对实验结果的影响,实验的 NMT 模型参数均相同,设置参考 Transformer-base,使用 6 层编码器和解码器,词嵌入矩阵的向量维度为 512,模型隐藏层的维度也为 512,前馈神经网络中间层单元数为 2 048,多头注意力的头数为 8. 另外对于模型训练,把每次训练的样本批量大小 batch size 设为 512;初始学习率设置为 1×10^{-4},使用 warm-up 策略更新学习率,学习率更新步数为 4 000 步,在训练初期逐渐增加学习率,以平滑地引导模型达到一个更稳定的训练状态,防止梯度爆炸使模型更好地收敛;使用 Adam 作为优化器,为防止模型过拟合,标签平滑值设为 0.1;为防止过拟合和增加模型鲁棒性,随机去除神经元的 Dropout 值设为 0.1,patience 设为 10,即在 10 个 epoch 训练中验证集上的性能不再提升时停止训练,以避免过拟合. 为减少过度局部化和处理不确定性,翻译解码过程使用束搜索算法,搜索束宽设为 5.

为评估翻译模型是否能够有效利用领域双语词典知识,引入复制成功率(copy successr rate,CSR)指标,即源句匹配的双语词典目标词约束中被翻译出的词汇所占百分比.

BLEU(bilingual evaluation understudy)值[17]反映了机器翻译输出与参考翻译之间的重叠程度,主要关注词汇的匹配度,因此使用标准 BLEU-4 作为评价指标衡量模型对句子的翻译性能. BLEU 对于每个 n-gram,计算候选译文中与参考译文匹配的 n-gram 数目占候选译文总 n-gram 数目的比例 p_n,即 n-gram 匹配度. 通过引入短句子惩罚因子 f_{BP}(候选译文长度 c 相对于最短参考译文长度 r 的比例),以解决系统生成短句的 BLEU 得分过高的问题. 为计算精确高效,将 n-gram 匹配度对数化,然后计算几何平均,乘以惩罚因子得到最终 n-gram 精确匹配度的 BLEU 值 V_{BLEU}.

$$f_{\mathrm{BP}}=\begin{cases}\exp(1-r/c), & c\leqslant r,\\ 1, & c>r,\end{cases} \tag{10}$$

$$V_{\mathrm{BLEU}}=f_{\mathrm{BP}}\cdot\exp\Big(\sum_{n=1}^{N}(w_n\cdot\log p_n)\Big). \tag{11}$$

其中,对于标准 BLEU-4,N 取 4,$w_n=1/4$.

2.3 实验结果分析

2.3.1 不同训练方式的基线模型对比实验

为了选择出对领域翻译效果提升更好的训练方式,对不同数据训练的基线 Transformer 模型的泰-汉翻译效果进行对比. 将法律领域数据训练集记为“law”. 根据领域双语数据的中文句子,对中文端含有领域双语词典中文词的外部双语数据进行模糊匹配[15]得到的混合数据记为“law+fuzz”. 根据领域双语词典的中文词,按照法律领域词词频匹配外部数据得到的混合数据记为“law+lfm”;与 Poncelas 等的 INR 方法[16]类似,使用基于不频繁法律领域中文词覆盖的方法对外部双语数据进行筛选得到的混合数据记为“law+ilr”.

表 3 实验结果显示,使用领域内训练集加额外训练集的训练方式相较于仅使用领域内训练集,BLEU 值均得到提升,说明增加额外训练集可以提升模型泰语到汉语的翻译性能. 基于模糊匹配训练集筛选方法的 BLUE 值仅提升了 0.26 个百分点,这是由于外部数据与领域数据的中文端之间编辑距离普遍相对较远,筛选的外部数据蕴含的可利用相似语义知识较少. 基于法律领域中文词频率匹配方法的 BLEU 值提升了 0.35 个

百分点,这是由于筛选的数据提供了目标端更多领域词汇的上下文,提升了模型的生成能力. 而基于不频繁法律领域中文词覆盖的方法 BLEU 值提升了 0.54 个百分点,这是由于筛选的外部数据包含了低频领域词更丰富的上下文,提升了部分低频领域词的翻译性能.

表 3　不同训练集的基线模型翻译性能

Tab. 3　Translation performance of baseline models with different training sets

训练集	BLEU 值/%
law	18.37
law+fuzz	18.63
law+lfm	18.72
law+ilr	18.91

2.3.2　不同融合领域双语词典方法的对比实验

为验证本文融合领域双语词典的 NMT 方法的有效性,将其与几种灵活的经典软词汇约束的 NMT 方法进行对比,结果如表 4 所示. 其中,模型都采用"law+ilr"数据进行训练,本文方法匹配的领域双语词典目标词作为约束词汇. Song 等[12]提出的增加源句内部替换的语码转换(Code-switched)语料库并添加指针网络的语码转换约束方法记为"CS+Ptr";Chen 等[14]提出的源句后部连接约束进行数据增强并添加分段嵌入和指针网络的约束方法记为"LeCA+Ptr".

表 4　不同词汇约束方法对比实验结果

Tab. 4　Comparative experimental results of vocabulary constraint methods

方法	BLEU 值/%	CSR 值/%
CS+Ptr	19.20	84.1
LeCA+Ptr	19.47	85.4
本文	**19.81**	86.3

由表 4 可知,本文提示模型的泰-汉翻译结果相较于"CS+Ptr"方法的 BLEU 值提升了 0.61 个百分点,CSR 值提升 2.2 个百分点. 这是由于"CS"使用目标约束词更改源句内部作为增强数据的方式虽然在训练时保留了初始源句的上下文信息,但在推断时破坏了源句结构,会丢失部分源句语义信息,特别是约束词存在噪声时. 本文模型相对于"LeCA+Ptr"方法的 BLEU 值提升了 0.34 个百分点,CSR 值提升 0.9 个百分点. 这是由于"LeCA+Ptr"方法无法避免一个源词在领域双语词典中存在多个目标词或词典存在噪声的情况,这种复杂性会在源合成数据中引入歧义导致上下文表征的偏差从而在模型结构里错误传播;而本文模型不更改源句,通过与提示信息进行交互的方式可以一定程度上缓解该问题. 相比于两种经典约束方法,本文方法在 BLEU 值和 CSR 上均得到了一定的提升,表明该方法可以有效利用领域双语词典约束,并且在复杂情况下提升领域关键词翻译准确度.

2.3.3　消融实验

为验证本文模型各模块的有效性,进行如表 5 所示的消融实验. 其中:采用"law+ilr"数据训练,将传统源编码器-解码器的 Transformer 基线模型记为"1∶Base";在基线模型 1 的基础上,将源编码器的输出隐状态与提示数据的浅层编码拼接的模型记为"2∶1+Concat";在基线模型的基础上,在每一层编码器自注意力层后添加提示表征交叉注意力模块的模型记为"3∶1+Prompt_Attention";为便于对模型的直观比较,将本文添加交叉注意力模块和拼接提示表征的完整模型记为"4∶1+Concat+Prompt_Attention". 在本文模型基础上,在生成端增加指针网络的模型记为"5∶4+Ptr".

表 5　验证各模块有效性的消融实验结果

Tab. 5　Results of ablation experiments to verify the effectiveness of modules

模型	BLEU 值/%	CSR 值/%
1∶Base	18.91	—
2∶1+Concat	19.14	80.0
3∶1+Prompt_Attention	19.40	83.2
4∶1+Concat+Prompt_Attention	**19.81**	**86.3**
5∶4+Ptr	19.48	84.1

由表 5 可知,基线模型的编码表征引入提示表征后,CSR 值有 80.0%的较高水平,表明融入提示表征能够保持较高水平的领域关键词翻译准确度;但其 BLEU 值相对于基线模型仅提升 0.23 个百分点,表明仅依靠拼接提示表征的方式对模型性能提升有限. 基线模型增加对提示表征的额外交叉注意力模块后,BLEU 值提升 0.49 个百分点,并且 CSR 值相比于引入提示表征也提升了 3.2 个百分点,证明交叉注意力模块对于提升模型性能以及领域关键词翻译准确度

优于引入提示表征.本文完整模型相比于基线模型的BLEU值提升了0.90个百分点,同时CSR值相比于引入提示表征和添加交叉注意力模块分别提升了6.3个百分点和3.1个百分点,这表明提示表征拼接与交叉注意力模块的组合才能发挥最大的作用,更有效提升模型性能与领域关键词翻译准确度.值得注意的是,本文完整模型比仅引入提示表征的模型在BLEU值上提升了0.67个百分点,进一步体现出了交叉注意力的重要性,这可能是因为注意力模块能有效抑制冗余信息进而大大提升提示表征对模型的作用.在本文模型的基础添加指针网络,BLEU值降低0.33个百分点,这可能是由于模型过拟合不够精确的指针网络预测概率,导致模型预测出现偏差;CSR值降低2.2个百分点,表明预测偏差降低了领域关键词的翻译准确度.

2.3.4 翻译任务实例分析

为了直观地显示本文方法的翻译效果,选取法律领域测试集中含有较多领域关键词的典型例句,采用本文方法与基线模型进行泰汉翻译,得到的句子结果如表6所示.

表6 翻译任务实例分析

Tab.6 Case analysis of translation tasks

泰语源句	翻译结果的来源	汉语目标句
หลักเกณฑ์ และ วิธีการ ที่ กำห-นด ไว้	参考译文	严格遵守法律、法规、内阁决议和相关命令规定的标准和方法
ใน กฎหมาย ระเบียบ มติ คณะรัฐมนตรี	基线 Transformer 的翻译结果	严格遵守法律、秩序、内阁决议和相关命令规定的方法和原则
และ คำ สั่ง ที่ เ-กี่ยวข้อง โดย เคร่งครัด	本文方法翻译结果	严格遵守法律、法规、内阁决议和相关命令规定的方法和标准

表6的内容直观地显示了翻译结果与参考译文之间的异同,本文提出的翻译方法能够产生与参考译文匹配度更高的结果.例如,在翻译词汇"ระเบียบ"时,本文方法将其正确地翻译为领域双语词典中对应词汇"法规",而基线Transformer模型则将其错误地翻译为"秩序",造成这种误差可能是因为生成的预测向量中两者的预测概率相同或相近,而本文模型将该目标词的表征作为提示信息融合进来,使其被较准确地翻译出来.实例表明,本文提出的融合领域双语词典的泰-汉神经机器翻译方法能够提升词典中领域关键词的翻译准确度,提升翻译结果与高质量参考译文的词汇匹配程度,从而有助于人们对整个法律译文的理解.

3 总 结

本文提出了一种融合领域双语词典的泰-汉法律领域NMT方法,以缓解该机器翻译任务中领域关键词翻译准确度不足的问题.从语料库挖掘泰-汉法律领域双语词典,以此筛选额外训练数据以引入更多上下文信息,并对源句匹配提示数据,通过新交叉注意力与其进行交互加强编码端表征,并在编码端引入提示信息表征,使得解码端利用强化表征信息指导目标词汇的生成.根据构建的数据集进行的实验结果显示,本文方法能够有效提升该低资源翻译任务性能,证明筛选伪领域数据方法和添加提示模块的有效性.相比于使用纯法律领域数据训练的Transformer模型,本文方法BLEU值提升1.44个百分点;相比于经典"语码转换"方法训练的模型,本文方法BLEU值提升0.61个百分点.通过实例分析发现,本文方法可以准确翻译出作为提示的法律领域中文关键词,翻译准确性得到提升.

参考文献:

[1] LI B H, HOU Y T, CHE W X. Data augmentation approaches in natural language processing: a survey[J]. AI Open, 2022, 3: 71-90.

[2] LUONG M T, MANNING C D. Stanford neural machine translation systems for spoken language domains[C]// International Workshop on Spoken Language Translation: Evaluation Campaign. Stroudsburg: ACL, 2015: 76-79.

[3] KHANDELWAL U, FAN A, JURAFSKY D, et al. Nearest neighbor machine translation[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2010.00710.

[4] SATO S, SAKUMA J, YOSHINAGA N, et al. Vocabulary adaptation for domain adaptation in neural machine translation [C] // Findings of the Association for Computational Linguistics: EMNLP. Stroudsburg: ACL, 2020: 4269-4279.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2024-12-01]. https://arxiv.org/abs/1706.03762.

[6] BENGIO S, VINYALS O, JAITLY N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[J]. Advances in Neural Information Processing Systems, 2015, 1: 1171-1179.

[7] HOKAMP C, LIU Q. Lexically constrained decoding for sequence generation using grid beam search[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1535-1546.

[8] POST M, VILAR D. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation[C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie. Stroudsburg: ACL, 2018: 1314-1324.

[9] HASLER E, GISPERT A, IGLESIAS G, et al. Neural machine translation decoding with terminology constraints[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 506-512.

[10] ANDERSON P, FERNANDO B, JOHNSON M, et al. Guided open vocabulary image captioning with constrained beam search[C]// Conference on Empirical Methods in Natural Language. Stroudsburg: ACL, 2017: 936-945.

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

[12] SONG K, ZHANG Y, YU H, et al. Code-switching for enhancing NMT with pre-specified translation[C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019: 449-459.

[13] WANG S, LI P, TAN Z X, et al. A template-based method for constrained neural machine translation[C]// Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 3665-3679.

[14] CHEN G H, CHEN Y, WANG Y, et al. Lexical-constraint-aware neural machine translation via data augmentation [C]// International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2020: 3587-3593.

[15] XU J T, CREGO J, SENELLART J. Lexical micro-adaptation for neural machine translation[C]// International Workshop on Spoken Language Translation. 2019: 1.27.

[16] PONCELAS A, DE BUY WENNIGER G M, WAY A. Transductive data-selection algorithms for fine-tuning neural machine translation[EB/OL]. [2024-12-01]. https://arxiv.org/abs/1908.09532v3.

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

(责任编辑:任滢滢)