文章编号:1000-5862(2019)06-0643-06

基于数据增强及领域适应的神经机器翻译技术

谷舒豪[1,2],单 勇[1,2],谢婉莹[3],郭登级[1,2],王树根[1,2],
邵晨泽[1,2],薛海洋[1,2],张 良[4],冯 洋[1,2]
(1. 中国科学院计算技术研究所,北京 100190;2. 中国科学院大学,北京 100049;
3. 北京语言大学信息科学学院,北京 100083;4. 中国矿业大学(北京)机电与信息工程学院,北京 100080)

摘要:近年来,基于深度学习的神经机器翻译已经成为机器翻译的主流方法. 神经机器翻译模型比统计机器翻译模型更依赖于大规模的标注数据. 因此,当训练语料稀缺或语料领域不一致时,翻译质量会显著下降. 在藏汉翻译中,训练语料大多为政府文献领域且数据稀缺;在汉英语音翻译中,训练语料大多为书面语领域且噪音语料稀缺. 为了提高神经机器翻译模型在这 2 个任务上的表现,该文提出了一种噪音数据增强方法和 2 种通用的领域自适应方法,并验证了其有效性.

关键词:神经机器翻译;藏汉翻译;语音翻译
中图分类号:TP 302.1 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2019.06.14

0 引言

近年来,基于深度学习的神经机器翻译模型已经成为机器翻译领域的主流模型. I. Vaswani 等[1]提出了完全基于自注意力机制的 Transformer 模型,它是目前翻译性能最好的模型架构. 在此之前,D. Bahdanau等[2]提出了基于循环神经网络(RNN)的神经机器翻译模型;Shao Chenze 等[3]提出了基于注意力机制的神经机器翻译模型. 然而,神经机器翻译模型极其依赖于大规模的标注数据,其性能在语料稀缺或语料领域不一致时会显著下降. 在藏汉翻译中,训练语料受限于政府文献领域,数据规模较小;在语音翻译中,大部分训练语料来自书面语领域,而且含噪音的口语语料十分匮乏. 这 2 个问题很难通过传统的端到端方法[1]得到解决. 因此,本文通过充分利用领域外语料来增强模型的翻译能力,并尝试了混合训练、领域精调等多种策略,提出了 2 种领域自适应方法;同时,本文针对噪音语料的语言学特点在领域外语料上进行数据增强,提出了一种噪音数据增强方法. 在 CCMT2019 机器翻译评测上的实验结果表明,本文所提出的方法均能有效提高神经机器翻译模型的性能.

1 系统

1.1 Transformer 模型

近年来,神经机器翻译相较于传统的方法(如统计机器翻译、规则翻译方法)取得了巨大的性能提升,越来越多的研究者开始进行神经机器翻译的研究. 鉴于神经机器翻译取得巨大成功,在本次评测中只使用了神经机器翻译系统. 在本次评测中,主要使用了完全基于注意力机制的 Transformer 神经网络机器翻译模型[1]. Transformer 与其它神经网络翻译模型[2-5]相比,其不依赖于卷积神经网络或者循环神经网络来提取序列特征,而完全依赖于自注意力机制[3],具有较快的训练速度以及出色的性能. 下面将对该系统进行简要的介绍.

与端到端的模型一样,Transformer 模型也采用了编码器-解码器的架构,如图 1 所示,其中左侧接受输入序列,经过映射并结合位置向量后变为输入向量序列;右侧为解码器,负责解码编码器所接收的信息并生成预测概率词表. Transformer 架构使用了多层多头注意力机制以及层级归一化,编码器与解码器都用了全连接层和残差连接. 其中多头自注意力机制接受 3 个输入,分别为 Value、Key 和 Query,

收稿日期:2019-08-31
基金项目:国家自然科学基金(61876174,61662077)和国家重点研发计划(2017YFE9132900)资助项目.
作者简介:谷舒豪(1994-),男,河北保定人,博士生,主要从事机器翻译、自然语言处理研究. E-mail:gushuhao17g@ict.ac.cn

当 3 个输入均为输入序列时,该机制为自注意力机制,如图 2 所示.

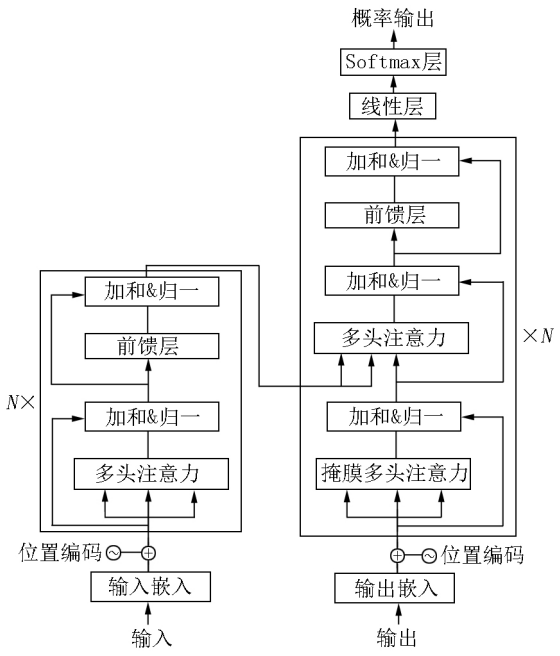

图 1 Transformer 模型框架

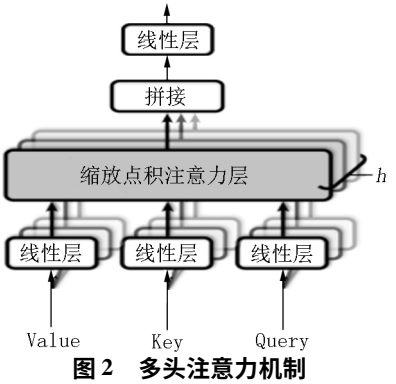

图 2 多头注意力机制

编码器接受输入序列,输入序列与位置编码向量相加作为多头注意力模块的输入,该模块的输出再同输入相加后送入层级归一化函数,并送入全连接层,最终得到编码器的输出. 编码器框架由 N 个相同的编码器模块构成. 编码器框架的输出将作为解码器的一部分输入,参与解码器的运算.

解码器模块在编码器模块的基础上增加了中间一层(多头自注意力层),这一层将编码器的输出作为该层的输入 Key 和 Value,采用解码器的第 1 个子层的输出作为其输入 Query. 解码器框架同样由 N 个相同的解码器模块构成.

1.2 技术说明

1.2.1 **藏汉政府文献机器翻译** 在藏汉政府文献机器翻译中,主要是结合课程学习[4]以及领域适应[5-6]的方法,尝试了不同策略的使用伪语料的方法. 其中包括只使用真实的平行语料、使用包含全部的伪语料以及平行语料的混合语料进行训练、根据语言模型评分使用部分得分较高的伪语料和真实语料的混合语料进行训练、全部真实语料进行评分、先使用混合语料然后每轮逐渐递减伪语料的使用进行训练的策略和先使用真实语料再逐渐增加伪语料进行训练的方法. 将会在下一节实验过程中对这些方法的细节以及实验结果进行详细介绍.

1.2.2 **汉英语音机器翻译** 在汉英语音机器翻译任务中,尝试了 2 种技术方案.

(i) Pipeline 技术方案. 使用百度公司提供的 ASR 识别结果作为输入,解码生成测试集译文. 因为本任务可使用汉英机器翻译任务中的汉英双语语料和汉语单语语料,在汉英双语语料、汉语单语语料、语音翻译双语语料的基础上训练机器翻译模型,将 ASR 识别结果输入机器翻译模型并得出译文.

(ii) 引入 Audio Encoder 的多模态机器翻译方案. 本文参照了文献[7-8]中引入篇章信息的设计,模型在方案 1 所得到的 Transformer 模型基础上加入了一个 Audio Encoder 用于提取音频特征,并在 Audio Encoder 与 Transformer 本身的 Text Encoder 和 Decoder 之间分别加入了 Multi-head Attention 用于音频特征与源端、目标端之间的交互. 因为本任务的 ASR 模块可以使用开源数据集,所以使用了 Common Voice(https://voice.mozilla.org/zh-CN)提供的汉语语音识别数据集训练了 ASR 模型,并将 ASR 模型的 Encoder 部分用于初始化 Audio Encoder 的参数,将方案 1 所得到的 Transformer 模型用于初始化 Text Encoder 与 Decoder 部分,Audio-Text Attention 与 Audio-Decoder Attention 采用随机初始化的方式初始化,在训练时使用较小的学习率进行 Fine-tune,但因为时间有限而未能得到明显的提升.

在方案 1 中,一共训练了 6 个完整的机器翻译模型,并尝试了模型平均和集成解码 2 种融合策略. 在模型平均策略中,将每个模型训练当收敛时最近的 5 个 checkpoint 的参数求平均值,从而得到一个新的 checkpoint 用来解码,模型平均带来的提升约 0.3 BLEU_SBP. 在集成解码策略中,在解码时将 6 个模型中最优的 checkpoint 加载进来,并在解码过程中的每一步根据所有模型的概率分布预测出当前要生成的词语,依次得到完整的译文,这样提升约 1.0 BLEU_SBP. 因为集成解码比模型平均效果更好,在最后提交系统输出时使用集成解码的方式进行融合.

2　数据处理

2.1　分词方法

对于汉语分词，本文使用清华大学开源的分词工具 thulac（http://thulac.thunlp.org/）；对于英语的 tokenize，使用 moses（https://github.com/moses-smt/mosesdecoder）中的分词脚本；对于藏语，使用的是中科院计算所自然语言处理研究组开发的藏语分词工具，该工具使用的是判别式的分词模型. 为了减少未登录词的影响并缩小词表的规模，使用基于子词的 Byte Pair Encoding（BPE）[9] 算法对语料进行处理，采用的是开源工具 subword-nmt（https://github.com/rsennrich/subword-nmt）.

2.2　藏汉政府文献机器翻译

在藏汉政府文献翻译任务中，对官方提供的单语以及双语语料分别做了细致的处理. 对于双语语料，主要处理步骤如下：（i）去除双语语料中的空行；（ii）Unicode 文本正则表示化；（iii）标点符号全角半角转换；（iv）藏文以及中文的分词. 对于汉语单语语料，主要处理步骤如下：（i）按照标点符号对长句进行切分并去重；（ii）去除单句中特殊符号、字母或者数字占比过大的句子；（iii）Unicode 文本正则表示化与标点符号全角半角转换；（iv）分词并去除过短及过长句子；（v）分别训练语言模型并评分，筛选 5.5 M句子[10]. 对于预处理得到的汉语单语语料，主要使用了 Back Translation[11] 方法以及 Fine Tune 方法. 在处理完单语语料之后，首先训练了中文到藏文的 Transformer 翻译模型，考虑到训练与解码的效率，使用了 Small（https://github.com/tensorflow/tensor2tensor）参数，接着利用 fast_align（https://github.com/clab/fast_align）技术对构造的伪平行语料进行评分，将分数过低的句对删除，最终剩余约 4.5 M 句对.

2.3　汉英语音机器翻译

在汉英语音机器翻译任务中，除了使用百度公司提供的语音翻译语料之外，还使用了在汉英新闻领域机器翻译任务中的双语语料和汉语单语语料，模型整体依然为受限训练.

对于汉英双语语料，主要处理步骤如下：（i）去除重复的句对；（ii）替换控制字符和转义字符，Unicode 文本正则表示化，标点符号全角半角转换；（iii）汉语语料分词，英文语料 tokenize；（iv）去除句长过长、过短或者双语句长比例过大、过小的句对；（v）使用 fast_align 工具对平行语料进行评分并选取合适阈值筛选，最终剩余约 5.4 M 句对；（vi）对英文语料进行 truecase 处理；（vii）对汉语语料、英语语料分别进行合并次数为 42 000、32 000 的 BPE 处理.

对于语音翻译双语语料的处理，与汉英双语语料类似. 语料中共含有约 2.8 万个训练样本，每个样本包含 5 个最好的 ASR 识别结果以及 1 个人工转写的文本. 由于语音翻译语料比较小，把训练样本中的每个 ASR 识别结果与人工转写的文本分别与该训练样本的译文组成双语句对，接着去除少数 ASR 识别结果为空的句对，处理之后一共得到约 15.5 万条双语句对.

对于汉语单语语料，主要处理步骤如下：（i）去除重复的句子；（ii）替换控制字符和转义字符，Unicode 文本正则表示化，标点符号全角半角转换；（iii）去除仅含数字、特殊符号的句子；（iv）汉语分词并应用相应的 BPE 处理.

随后，使用了伪语料数据扩充的方法. 在预处理完单语语料之后，首先在汉英平行语料上训练了汉语到英语的 Transformer Big 模型，并用它来翻译单语语料得到伪语料，接着对伪语料进行了如下处理：（i）去除句长过长、过短或者双语句长比例过大、过小的句对；（ii）使用 fast_align 工具对伪语料进行评分，使用不同的阈值进行过滤并将得到的伪语料与汉英平行语料混合起来训练 Transformer Small 模型，根据在语音翻译验证集上的 BLEU_SBP 分数（实验结果此处略去），以 -6 为筛选阈值得到的 3.3 M伪语料效果最好.

所以，在翻译模型的训练中，使用筛选得到的 5.4 M 平行语料、筛选得到的 3.3 M 伪语料、15.5 万条语音翻译平行语料共约 8.7 M 双语句对作为训练集.

3　实验

3.1　实验环境

本次评测使用的硬件环境为：32 颗 Intel（R）Xeon（R）E5-2620 v4@2.10GHz 处理器，4 块 Nvidia Titan XP GPU，128G 内存；软件环境为：Ubuntu 16.04 LTS 64 位操作系统，CUDA9.0，Python 3.6，PyTorch 1.0.1.

本次评测使用的训练代码主要是脸书公司开源的 Fairseq（https://github.com/pytorch/fairseq）工具.

3.2 藏汉政府文献机器翻译

按照上述的方法,构建了伪平行语料并进行了筛选,最终用于训练的语料包括 4.5 M 伪语料以及 157 000 条真实训练语料. 之后尝试了使用不同的伪语料训练方法,测试其在验证集上的表现:(i) Baseline,只使用真实的平行训练语料;(ii) + all,使用平行语料以及全部的伪语料均匀混合进行训练;(iii) + top $5n$,使用平行语料以及语言模型评分最高的 5 倍于平行语料的伪语料;(iv) + top $4n$,使用平行语料以及语言模型评分最高的 4 倍于平行语料的伪语料;(v) + decrease,第 1 轮使用全部的混合语料,以后每一轮减少分数最低的 5% 伪语料,第 20 轮之后只使用平行语料进行训练;(vi) + increase,第 1 轮使用真实平行语料,以后每一轮按照分数由高到低增加 5% 的伪语料.

以上各个系统均使用了 Base 参数,各个系统在验证集(1 000 句)上的得分如为表 1 所示.

表 1 藏汉政府文献翻译不同伪语料使用策略在验证集上的得分

系统	BLEU5_SBP	BLEU5	NIST	GTM	mWER	mPER	ICT
Baseline	0.368 1	0.397 5	6.936 6	0.605 3	0.542 1	0.475 2	0.436 3
+ all	0.248 5	0.271 4	6.160 2	0.566 1	0.608 5	0.529 3	0.381 8
+ top $4n$	0.338 7	0.362 8	6.706 2	0.604 2	0.545 4	0.481 2	0.415 1
+ top $5n$	0.326 6	0.351 0	6.685 3	0.599 4	0.550 8	0.486 4	0.431 8
+ decrease	0.400 5	0.430 5	7.441 0	0.642 4	0.497 2	0.436 8	0.464 1
+ increase	0.247 1	0.269 0	6.118 3	0.566 0	0.610 0	0.530 4	0.372 5

由表 1 可以看到,使用伪语料递减策略的训练方法得到的最终模型的表现最好. 之后使用该策略训练 Transformer-Big 模型,训练参数可以参照 awesome-transformer (https://github.com/ictnlp/awesome-transformer). 在解码时,根据模型在验证集上的表现,尝试不同 beam size 和不同 length penalty [12] 的组合,选取最好的一组作为解码参数. 根据单模型在验证集上的表现,选取排名前 5 的单模型尝试 model average 和 model ensemble 技术进行解码.

最终选定系统在验证集 CWMT500tizh(http://114.212.189.224:8080/mtweb/) 上的结果如表 2 所示,其中 Primary-a 和 Contrast-c 为最后得到的在验证集上表现最好的 2 个单模型,Contrast-b 和 Contrast-d 分别为使用 4 个和 5 个最好的单模型进行 ensemble 解码得到的最终结果. 还尝试了进行参数 average 的策略,但是均未带来提升. 根据验证集上的结果以及人工粗略评估测试集翻译质量确定了最终的主系统. 解码参数 beam size 设置为 11,length penalty 设置为 0.6.

表 2 藏汉政府文献机器翻译最终系统在 CWMT500tizh 验证集上的结果

系统	BLEU_SBP	BLEU_NIST	TER	METEOR	NIST	GTM	mWER	mPER	ICT
Primary-a	0.571 7	0.563 3	0.280 1	0.753 1	9.241 0	0.805 5	0.299 2	0.241 4	0.615 7
Contrast-b	0.570 3	0.563 2	0.281 7	0.751 2	9.246 4	0.804 0	0.301 0	0.243 7	0.614 6
Contrast-c	0.571 7	0.564 2	0.280 6	0.752 4	9.261 8	0.804 6	0.298 8	0.243 1	0.614 8
Contrast-d	0.571 6	0.565 2	0.281 0	0.751 9	9.259 5	0.804 5	0.299 6	0.243 7	0.615 3

3.3 汉英语音机器翻译

由于在语音翻译语料中的领域偏口语化,ASR 识别结果通常会存在一些语气词、重复、句子成分不完整的现象,而且受 ASR 识别精度的影响,识别结果会包含一些标点符号错误和同音异形词错误,这些问题可以认为是机器翻译模型面临的噪音. 若直接在训练集上训练机器翻译模型,则在测试集上翻译时会面临 2 个问题:(i) 模型对于 ASR 识别结果中噪音的处理能力较差;(ii) 因为训练集中大多数语料属于新闻/书面语领域,模型在翻译口语化语料时存在领域不适应的问题. 针对这 2 个问题,本文在模型的训练上应用了噪音数据增强和领域精调 2 种技术.

在噪音数据增强中,在训练集的汉语端句子上进行对应于上述噪声的数据增强操作,通过这样来模拟 ASR 的识别结果,增强模型对于 ASR 识别结果中噪音的处理能力. 噪音数据增强的具体操作为:(i) 随机插入语气词;(ii) 在词附近随机重复该词;(iii) 随机丢弃词;(iv) 随机替换标点符号;(v) 随机将词替换成与之同音异形的词. 虽然训练集中包含有少量语音翻译语料,但由于数据占比很小,在操作时仍然对整个训练集进行数据增强.

对训练集做完噪音数据增强之后,用 Fairseq 在

训练集上训练 Transformer Big 模型，模型参数与 tensor2tensor 中保持一致，详细训练配置参见 awesome-transformer. 在训练时模型大约每轮保存 3 个 checkpoint，大约需要 8 d 训完 30 轮. 随后选取在验证集上 BLEU_SBP 最高的 checkpoint 用于做领域精调.

在领域精调时，只在语音翻译平行语料上进行训练，且不需要进行噪音数据增强，学习率设为 5×10^5，使模型慢慢适应口语领域的翻译特点，精调 10 轮之后停止，选取在验证集上 BLEU_SBP 最高的 checkpoint. 本文也尝试了 + decrease 方法，即递减使用领域外语料（汉英平行语料和伪语料），但是效果不如仅在语音领域的语料上精调.

在解码时，根据模型在验证集上的表现，尝试不同 beam size 和不同 length penalty 的组合，选取最好的一组作为解码参数. 最后得出当 beam size 设为 5、length penalty 设为 1.0 时，效果最好. 对于解码得到的译文，先去掉 BPE 操作引入的特殊标记，再进行 detruecase 处理，最后进行 detokenize 处理.

模型在验证集上的实验结果如表 3 所示，其中 baseline 是仅用汉英平行语料、伪语料、语音翻译平行语料训练得到的单模型的结果，+ noise 是在 baseline 基础上使用噪音数据增强得到的单模型的结果，+ finetune 是在 + noise 基础上使用领域精调得到的单模型的结果，+ ensemble 是使用 6 个分别训练得到的 + finetune 模型集成解码的结果. 从上述实验结果可以看出，噪音数据增强、领域精调、集成解码均能大幅提升翻译质量，验证了本文所采用的策略的有效性.

表 3　汉英语音机器翻译任务中模型在验证集上的实验结果

System	BLEU4_SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
Baseline	0.156 6	0.168 4	4.854 7	0.480 7	0.711 3	0.598 5	0.284 5
+ noise	0.162 5	0.179 2	4.947 2	0.488 7	0.727 1	0.602 7	0.271 8
+ finetune	0.169 6	0.186 6	5.023 0	0.492 4	0.719 2	0.594 4	0.274 0
+ ensemble	0.179 7	0.198 4	5.12 2	0.500 3	0.710 5	0.591 1	0.277 2

最后，使用 + ensemble 实验中的方法作为主系统. 为了避免翻译长句带来的性能下降，对测试集中的长句单独处理：按照标点符号分割成若干个长度不超过 50 的短句，生成对应的译文之后再还原为长句.

4　总结

本文主要介绍了使用数据增强技术以及领域适应技术来解决神经机器翻译技术中数据稀疏以及领域不匹配的问题. 使用语言模型从大量的多领域的含噪音的数据集中进行筛选来扩充训练语料，并通过逐渐减少的方式来进行精细的微调，以此来提高翻译质量. 除此之外，针对口语翻译中噪音特点，对于口语翻译语料进行加噪处理，以提高模型的鲁棒性，提高翻译质量. 在 CCMT2019 机器翻译评测上的实验结果表明，本文所提出的方法均能有效地提高神经机器翻译模型的性能.

致谢：在本次实验中，中国科学院计算技术研究所自然语言处理组的老师们和同学们都付出了很多的努力. 参加本次研究的成员在实际工作中，也收获了许多经验、吸取了不少教训，在此对他们致以衷心的感谢与祝贺，并特别感谢实验室冯洋老师给予的大力支持和帮助.

5　参考文献

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1409.3215.

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1409.0473v2.

[3] Shao Chenze, Feng Yang, Zhang Jinchao, et al. Retrieving sequential information for non-autoregressive neural machine translation [EB/OL]. [2019-03-16]. https://www.aclweb.org/anthology/P19-1288/.

[4] Bengio Y, Louradour J, Collobert R B, et al. Curriculum learning [EB/OL]. [2019-03-16]. http://dx.doi.org/10.1145/1553374.1553380.

[5] Luong M, Manning C D, et al. Stanford neural machine translation systems for spoken language domains [EB/OL]. [2019-03-16]. https://nlp.stanford.edu/pubs/luong-manning-iwslt15.pdf.

[6] Gu Shuhao, Feng Yang, Liu Qun. Improving domain adaptation translation with domain invariant and specific information [EB/OL]. [2019-03-16]. https://arxiv.org/pdf/1904.03879.pdf.

[7] Zhang Jiacheng, Luan Huanbo, Sun Maosong, et al. Impro-

ving the transformer translation model with document-level context [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1810.03581.

[8] Yang Zhengxin, Zhang Jinchao, Meng Fandong, et al. Enhancing context modeling with a query-guided capsule network for document - level NMT [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1909.00564.

[9] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1508.07909.

[10] Axelrod A, He Xiaodong, Gao Jianfeng. Domain adaptation via pseudo in-domain data selection [EB/OL]. [2019-03-16]. https://core.ac.uk/display/21859466.

[11] Rico S, Barry H, Alexandra B. Improving neural machine translation models with monolingual data [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1511.06709.

[12] Zhang Wen, Feng Yang, Meng Fandong, et al. Bridging the gap between training and inference for neural machine translation [EB/OL]. [2019-03-16]. https://arxiv.org/abs/1906.02448.

The Neural Machine Translation Based on Data Augmentation and Domain Adaptation Technology

GU Shuhao[1,2], SHAN Yong[1,2], XIE Wanying[3], GUO Dengji[1,2], WANG Shugen[1,2], SHAO Chenze[1,2], XUE Haiyang[1,2], ZHANG Liang[4], FENG Yang[1,2]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. School of Information Science, Beijing Language and Culture University, Beijing 100083, China;
4. School of Mechanical Electronic and Information Engineering, China University of Mining and Technology(Beijing), Beijing 100080, China)

Abstract: In recent years, neural machine translation based on deep learning has become the mainstream method in machine translation. The neural machine translation model relies more on large-scale annotation data than the statistical machine translation model, so its translation quality will be significantly reduced when the training corpus is scarce or the domain is inconsistent. In Tibetan-Chinese translation, the training corpus is mostly in the government literature domain and the data is scarce. In speech translation, the training corpus is mostly in the written language domain and the noise corpus is scarce. In order to improve the performance of the neural machine translation model on these two tasks, this report proposes a noise data enhancement method and two general domain adaptive methods, and verifies their effectiveness.

Key words: neural machine translation; Tibetan-Chinese translation; speech translation

(责任编辑:冉小晓)