doi:10.6043/j.issn.0438-0479.201811019

基于增量训练的维汉神经机器翻译系统

杨郑鑫[1,2],李京谕[1,2],胡镓伟[1,2],冯 洋[1,2*]

(1.中国科学院计算技术研究所,智能信息处理重点实验室,北京 100190;
2.中国科学院大学计算机科学与技术学院,北京 100049)

摘要:目前,基于深度学习的神经机器翻译已经成为机器翻译领域的主流方法.神经机器翻译模型相较于统计机器翻译模型具有更庞大的参数规模,因此其翻译质量取决于训练数据是否充足.由于与维吾尔语相关的平行语料资源严重匮乏,神经机器翻译模型在维汉翻译任务上表现不佳,为此提出了一种利用伪语料对神经机器翻译模型进行增量训练的方法,可有效提升神经机器翻译在维汉翻译任务上的质量.

关键词:自然语言处理;神经机器翻译;维吾尔语

中图分类号:TP 183 **文献标志码**:A **文章编号**:0438-0479(2019)02-0195-05

目前,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.Cho 等[1]提出了基于循环神经网络(RNN)的神经机器翻译方法;Bahdanau 等[2]提出了基于注意力机制的神经机器翻译方法;Gehring 等[3]提出了基于卷积神经网络(CNN)的神经机器翻译方法;Vaswani 等[4]提出了完全基于自注意力机制的 Transformer 模型架构的神经机器翻译方法,该方法使用的是目前国际上翻译性能最好的模型架构.

神经机器翻译模型相较于传统的基于短语的统计机器翻译模型具有更庞大的参数规模,因此其翻译质量取决于双语平行语料的训练数据是否充足.然而,目前许多小语种的双语语料资源仍匮乏.因此,Sennrich 等[5]针对这一问题提出了利用单语语料构造伪双语平行语料的方法,从而有效地扩充了训练语料,解决了平行语料数据量不足的问题.

目前,由于与维吾尔语相关的平行语料资源严重匮乏,所以神经机器翻译模型在维吾尔语-汉语的翻译任务上表现不佳;但是汉语的单语语料十分充足.因此,本文中提出了一种利用汉语伪语料对神经机器翻译模型进行增量训练的方法,以提升其在维吾尔语-汉语翻译任务上的翻译质量.

1 系统架构

本文中采用 Vaswani 等[4]提出的完全基于自注意力机制的 Transformer 架构的神经机器翻译系统.该系统与基于注意力机制的神经机器翻译系统[2,6-8]不同,其不依赖于 CNN 和 RNN 来抽取序列特征,而是完全依赖于自注意力机制.

1.1 整体架构

与端到端的模型一样,Transformer 架构也采用了编码器-解码器的架构,如图 1 所示:其中左侧为编码器,接收输入序列;右侧为解码器,预测词概率.

Transformer 架构使用多层多头的自注意力机制以及层级归一化,编码器与解码器间都使用全连接层和残差连接.其中多头自注意力机制包含 h 个参数独立的头,每个头接受 3 个输入(Value、Key 和 Query),当 3 个输入均为输入序列时,则表示自注意力机制(图 2).

1.2 编码器

编码器接收输入序列,输入序列与位置编码向量相加作为多头自注意力模块的输入,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

收稿日期:2018-11-11 **录用日期**:2018-12-03
基金项目:国家自然科学基金(61662077)
* **通信作者**:fengyang@ict.ac.cn
引文格式:杨郑鑫,李京谕,胡镓伟,等.基于增量训练的维汉神经机器翻译系统[J].厦门大学学报(自然科学版),2019,58(2):195-199.
Citation:YANG Z X,LI J Y,HU J W,et al. Uyghur-to-Chinese neural machine translation based on incremental training[J]. J Xiamen Univ Nat Sci,2019,58(2):195-199.(in Chinese)

http://jxmu.xmu.edu.cn

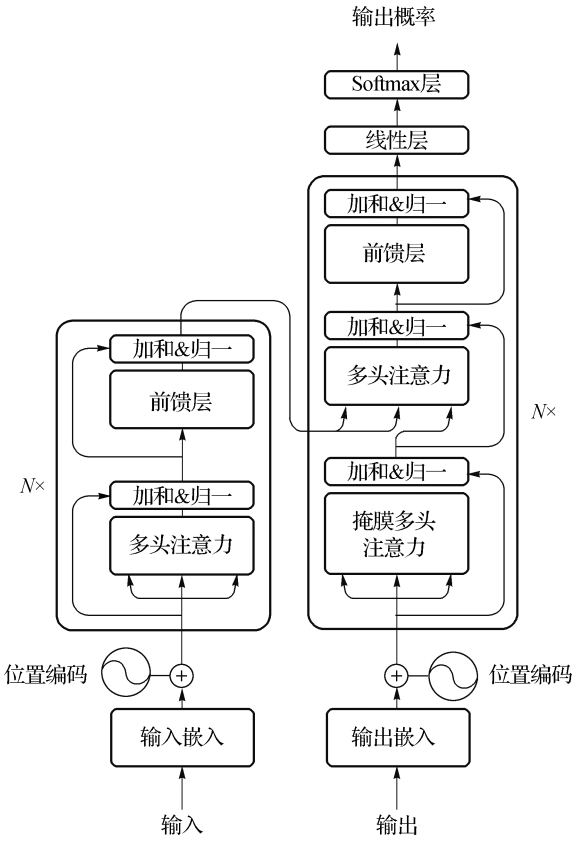

图 1 Transformer 模型架构(修改自文献[4])

Fig. 1 Architecture of the Transformer model (modified from reference[4])

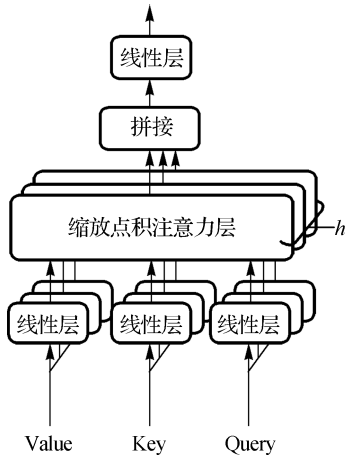

图 2 多头自注意力模块(修改自文献[4])

Fig. 2 Multi-head self-attention module (modified from reference[4])

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,其输出将作为解码器的一部分输入,参与解码器的运算.

1.3 解码器

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,该层将编码器的输出作为该层的输入 Key 和 Value,采用解码器第一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.

1.4 系统融合

利用增量训练的方法训练的 4 个不同汉语单语语料子集的子模型进行系统融合(每个子集的规模为 2×10^5 句,然后对 4 个模型分别进行集束搜索(beam search),生成N-best的目标端候选译文,最后对 $4\times N$ 个候选译文进行句子级别的再排序(re-ranking). 具体地,利用汉语单语语料训练目标端的汉语语言模型,然后对 4 个子模型解码后得到的 $4\times N$ 个候选译文按一定比例对单语汉语语言模型的困惑度和翻译系统解码后目标端句子的困惑度求和,最后对所有候选译文重新排序,并将困惑度最低的句子作为系统的最终输出结果.

2 数 据

使用 2018 年全国机器翻译研讨会(CWMT 2018)发布的维吾尔语-汉语的双语平行语料进行实验. 由于维吾尔语-汉语的双语平行语料规模较小,因此使用一定规模的汉语单语语料来构造伪双语平行语料以扩充训练集数据规模.

2.1 语料预处理

针对语料的数据预处理是机器翻译的重要环节,语料处理的优劣在一定程度上决定了机器翻译系统训练模型时的效果好坏.

主要的预处理步骤如下:

1) 编码转换;

2) 全角字符转半角字符;

3) 处理转义字符;

4) 过滤控制字符等特殊字符;

5) 分词及 token;

6) 筛选单语、双语语料.

2.2 单语语料过滤

为了获得与双语平行语料所属领域一致的单语语料,使用 SRILM 工具(http://www.speech.sri.com/projects/srilm/)对双语平行语料的汉语端构建汉语语言模型;然后利用构建好的汉语语言模型计算预处理好的单语汉语语料句子的困惑度,并按照困惑

http://jxmu.xmu.edu.cn

度由低到高进行排序;最后选取靠前的 8×10^5 个单语句子构建伪双语平行语料. 语料规模为:双语平行语料 4×10^5 句对,汉语单语语料 8×10^5 句.

2.3　单语语料使用

CWMT 2018 在维汉翻译项目中提供了大量的单语语料. Sennrich 等[5]提出了利用单语语料构造伪双语平行语料的反向翻译(back-translation)方法,从而有效地扩充了训练语料,解决了平行语料数据量不足的问题.

首先利用已有的双语平行语料训练一个汉语-维吾尔语神经机器翻译系统,然后使用训练好的系统将汉语单语语料翻译为维吾尔语,从而得到伪平行语料;进而在使用平行语料训练的维吾尔语-汉语神经机器翻译系统的基础上,继续利用伪平行语料进行训练,并得到最终的翻译系统. 构建方法如图 3 所示.

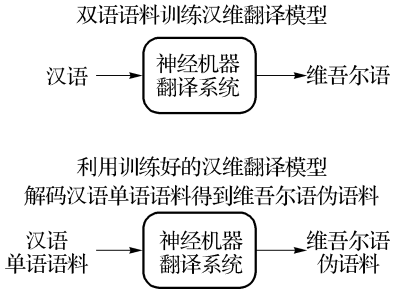

图 3　利用反向翻译方法构建伪语料

Fig. 3　Constructing synthetic corpus using back-translation method

2.4　分词方法

在机器翻译任务中,对未登录词以及罕见词的处理是一大难题. 尤其是维吾尔语作为一种黏着性语言,本身在语素的组合上具有高度灵活性,理论上可以构成无限多的词语;但是绝大多数维吾尔语的词在语料库中只出现一次,测试集中也极易出现在训练集中未出现的词. 因此,如何针对维吾尔语进行分词对神经机器翻译模型的泛化能力显得尤为重要. 近期 Sennrich 等[9]提出了一种基于字节对编码(byte-pair encoding,BPE)的分词算法,该方法通过对子词进行切分从而提高翻译模型对罕见词和未登录词的处理能力. Wu 等[7]又提出了一种混合字词的分词模型,该模型在汉语上可以提供较好的分词效果.

本文中对维吾尔语端采用 BPE 算法进行分词,对汉语端采用混合字词模型的方法进行分词. 维吾尔语分词使用 Sennrich 开源的 Subword 分词工具(https://github.com/rsennrich/subword-nmt);汉语分词采用 Tensor2Tensor 工具(https://github.com/tensorflow/tensor2tensor)内建的 Subword 分词方法.

3　实　验

3.1　实验环境

使用 Ubuntu 15.04 64 位操作系统,2 块 Intel(R) Xeon(R) E5-2609v3 CPU@1.90 GHz 处理器和 64 GB 内存,4 块 GeForce GTX TITAN X GPU.

3.2　实验设置

采用 Google 开源的基于 Tensorflow 的 Tensor2Tensor 深度学习工具包进行模型训练[10],并对模型的分词部分基于 2.4 节的方法进行了重写.

具体地,维吾尔语和汉语的词表大小均为 3.2×10^4,其中维吾尔语 BPE 的迭代轮数为 3.2×10^4. 采用 Transformer 架构中的 Big 模型参数对主系统 primary-a 进行多 GPU 训练,由于 Tensor2Tensor 中数据批大小(batch size)以词为单位,将其调整为 3.4×10^3 以适应实验室 GPU 显存大小. 所有模型均使用 4 块 GeForce GTX TITAN X GPU 进行训练.

主系统 primary-a 由 4 个基于不同单语语料训练的子模型融合而来,每个子模型都选取模型参数收敛后的最近 20 个检验点(checkpoints)进行参数平均. 首先利用双语平行语料训练一个基准模型,记录停止训练时的学习率为 L;然后使用通过 SRILM 筛选后的单语语料,选取困惑度最低的 8×10^5 句单语生成伪训练语料,将其按困惑度由低到高平均分为 4 份,即 D1、D2、D3 和 D4.

3.3　增量训练

采用增量训练的方法分别训练 4 个子模型. 首先在基准模型的基础上,将 D1 伪双语平行语料加入双语平行语料中继续进行训练,直到模型收敛,得到子模型 A;然后将学习率调回至 L,将 D1 伪双语平行语料从双语平行语料中去除,并将 D2 伪双语平行语料加入双语平行语料中继续进行训练,直到模型收敛,得到子模型 B;接着再将学习率调回至 L,将 D2 伪双语平行语料从双语平行语料中去除,并将 D3 伪双语平行语料加入双语平行语料中继续进行训练,直到模型收敛,得到子模型 C;最后将学习率再次调回至 L,将 D3 伪双语平行语料从双语平行语料中去除,并将 D4 伪双语平行语料加入双语平行语料中继续进行训练,直到模型收敛,得到子模型 D. 得到 A、B、C 和 D 4 个子模型后对模型进行系统融合,得到最终的翻译结果. 增量训练的方法如图 4 所示.

http://jxmu.xmu.edu.cn

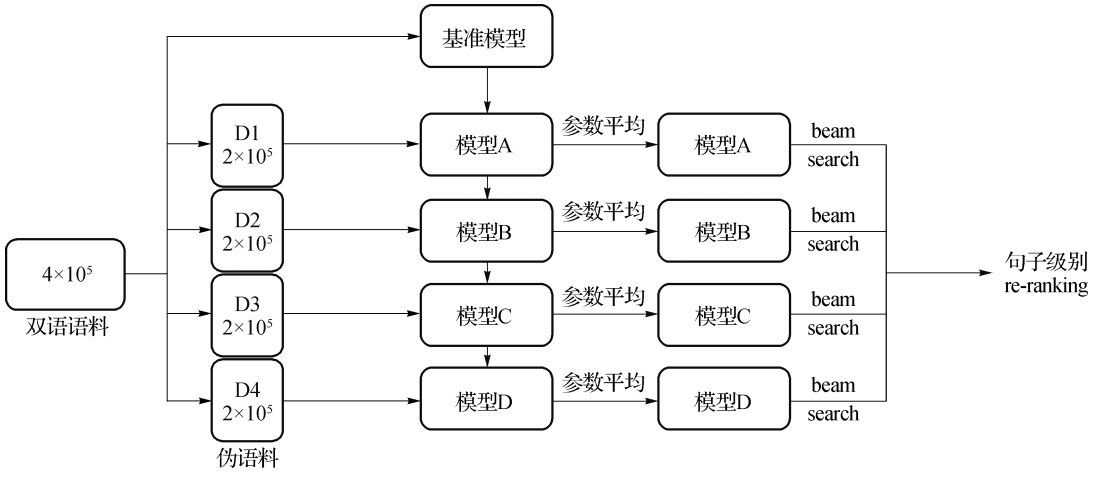

图 4 利用伪平行语料的增量训练方法

Fig. 4 Incremental training method using pseudo-parallel corpus

对比系统 contrast-b 和 constrast-c 采用了与主系统 primary-a 相同的配置,区别是对比系统在每个子模型解码时使用的解码参数 alpha(长度惩罚项)不同,具体地,对比系统 contrast-b 的子模型均采用 alpha=0.6,contrast-c 的子模型均采用 alpha=0.7,而主系统 primary-a 的子模型 A、B 均采用 alpha=0.7,子模型 C、D 均采用 alpha=0.6. 其中,对比系统 contrast-d 即基准模型,基准模型在 beam search 时参数 alpha=0.6,该系统仅训练了一个 Transformer 架构的基础模型,并采用了 Transformer 架构的 Big 参数进行训练.

3.4 实验结果与分析

表 1 为主系统和 3 个对比系统在 CWMT 2018 测试集上的测试结果,其中 primary-a、contrast-b 和 contrast-c 是基于多个子模型进行系统融合得到的,而 contrast-d 是平行双语语料训练得到的单个模型. 由表 1 可以看出系统融合对最终翻译质量的提升帮助较大;从 primary-a、contrast-b 和 constrast-c 的测试结果可知,解码时参数 alpha 的设置对翻译结果也有一定影响,但并不显著.

表 1 基于不同评价指标的测试结果

Tab. 1 Testing results based on different evaluation metrics

评价指标	primary-a	contrast-b	contrast-c	contrast-d
BLEU5-SBP[11]	0.380 2	0.380 4	0.377 6	0.364 6
BLEU5	0.388 7	0.389 2	0.388 4	0.375 2
BLEU6	0.348 8	0.349 2	0.347 1	0.334 8
NIST6[12]	9.220 0	9.235 9	9.260 7	9.062 8
NIST7	9.229 3	9.245 3	9.269 7	9.071 5
GTM[13]	0.718 8	0.719 5	0.719 9	0.709 1
mWER[14]	0.449 5	0.449 8	0.455 7	0.466 7
mPER[15]	0.335 1	0.334 3	0.331 0	0.343 9
ICT[16]	0.380 7	0.380 5	0.372 6	0.363 7
METEOR[17]	0.598 0	0.598 8	0.600 3	0.587 4
TER[18]	0.404 5	0.404 1	0.407 1	0.419 1

注:BLEU-SBP 为使用严格的长度惩罚因子的机器双语互译评估(BLEU)指标,BLEU5 基于 5 元组,BLEU6 基于 6 元组;NIST 为在 BLEU 指标上的一种改进方法,NIST6 基于 6 元组、NIST7 基于 7 元组;GTM 用于计算文本之间的相似度,是一个 *N*-gram 准确率/召回率评价指标;mWER 为基于多参考译文的词错误率评价指标,mPER 与其相似,但不考虑词的位置;ICT 为中科院计算所研制的以熵为基础的自动评测方法,利用匹配片断计算加权熵;METEOR 为基于单精度的加权调和平均数和单字召回率的评价指标;TER 为计算一个机器翻译译文需要经过修改多少次才能与参考译文一致的指标。

4 结 论

本文中主要提出了一种基于增量训练的神经机器翻译系统,增量训练的方法可以使模型在不丢失已经学习到的原有数据信息的基础上,充分学习新的训练数据的信息,从而增强整个模型的翻译性能. 另外,该系统充分利用单语汉语语料来构建伪双语平行语料,从而解决了维吾尔语语料资源匮乏的情况.

致谢:对中国科学院计算技术研究所自然语言处理小组老师和同学的努力付出致以衷心的感谢,并特别感谢实验室赵红梅老师和李响给予的帮助.

参考文献:

[1] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings

http://jxmu.xmu.edu.cn

of EMNLP 2014. Baltimore: Association for Computational Linguistics, 2014: 1724-1734.

[2] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// Proceedings of ICLR 2015. San Diego: International Conference on Learning Representations, 2015: 1409.0473.

[3] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]// Proceedings of ICML 2017. Sydney: International Conference on Machine Learning, 2017: 1705.03122.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of NIPS 2017. Long Beach: Conference on Neural Information Processing Systems, 2017: 1706.03762.

[5] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[C]// Proceedings of ACL 2016. Berlin: Association for Computational Linguistics, 2016: 86-96.

[6] LUONG M, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [C]// Proceedings of EMNLP 2015. Lisbon: Association for Computational Linguistics, 2015: 1412-1421.

[7] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB/OL]. [2018-11-27]. https://arxiv.org/pdf/1609.08144.

[8] SUTSKEVER I, VINYALS V, LE Q V. 2014. Sequence to sequence learning with neural networks [EB/OL]. [2018-11-27]. https://arxiv.org/pdf/1409.3215.

[9] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units [C]// Proceedings of ACL 2016. Berlin: Association for Computational Linguistics, 2016: 1715-1725.

[10] JOHNSON M, SCHUSTER M, LE Q V, et al. 2016 Google's multilingual neural machine translation system: enabling zero-shot translation[EB/OL]. [2018-11-27]. https://arxiv.org/pdf/1611.04558.

[11] CHIANG D, DENEEFE S, CHAN Y S, et al. Decomposability of translation metrics for improved evaluation and efficient algorithms[EB/OL]. [2018-11-27]. https://www3.nd.edu/~dchiang/papers/bleu.pdf.

[12] DODDINGTON G. Automatic evaluation of machine translation quality using *N*-gram co-occurrence statistics [EB/OL]. [2018-11-27]. http://www.mt-archive.info/HLT-2002-Doddington.pdf.

[13] MELAMED I D, GREEN R, TURIAN J P. Precision and recall of machine translation[C]// Proceedings of HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003: 61-63.

[14] KLAKOW D, PETERS J. Testing the correlation of word error rate and perplexity[J]. Speech Communication, 2002, 38(1/2): 19-28.

[15] LEUSCH G, UEFFING N, NEY H. A novel string-to-string distance measure with applications to machine translation evaluation[C]// Proceedings of MT Summit Ⅸ 2003. New Orleans: [s.n.], 2003: 240-247.

[16] 刘群,刘洋. 一种机器翻译自动评测方法及其系统:中国,ZL200410000628.8[P]. 2009-10-28.

[17] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of ACL 2005. Ann Arbor: Association for Computational Linguistics, 2005: 65-72.

[18] SNOVER M, DORR B, SCHWARTZ R, et al. A study of translation edit rate with targeted human annotation[C]// Proceedings of AMTA 2006. Cambridge: Association for Machine Translation in the Americas, 2006: 223-231.

Uyghur-to-Chinese neural machine translation based on incremental training

YANG Zhengxin[1,2], LI Jingyu[1,2], HU Jiawei[1,2], FENG Yang[1,2*]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China; 2. School of Computer Science and Technology, University of Chinese Academy of Science, Beijing 100049, China)

Abstract: At present, the neural machine translation based on deep learning has become the mainstream method in the field of machine translation. The neural machine translation model requires a larger parameter size than the statistical machine translation model does. Therefore, its translation quality depends on the sufficiency of the training data. Due to the serious lack of parallel corpus resources related to Uyghur, the neural machine translation model performs poorly on Uyghur-to-Chinese translation tasks. This paper proposes a method of incremental training of neural machine translation models using pseudo-corpus, which effectively improves the quality of neural machine translation in Uyghur-to-Chinese translation tasks.

Keywords: natural language processing; neural machine translation; Uyghur

http://jxmu.xmu.edu.cn