doi:10.6043/j.issn.0438-0479.202011025

面向低资源神经机器翻译的回译方法

张文博[1,2,3],张新路[1,2,3],杨雅婷[1,2,3],董 瑞[1,2,3],李 晓[1,2,3]*

(1.中国科学院新疆理化技术研究所,新疆 乌鲁木齐 830011;2.中国科学院大学计算机科学与技术学院,北京 100049;3.新疆民族语音语言信息处理实验室,新疆 乌鲁木齐 830011)

摘要:神经机器翻译在高资源情况下已经获得了巨大的成功,但是对低资源情况翻译效果还有待提高.目前,维吾尔语-汉语(维汉)翻译和蒙古语-汉语(蒙汉)翻译都属于低资源情况下的翻译任务.本文提出将汉语单语数据按照领域相似性划分成多份单语数据,并通过回译方法分段利用不同的单语数据训练翻译模型,然后借助模型平均和模型集成等方法进一步提升维汉和蒙汉翻译质量.使用第 16 届全国机器翻译大会(CCMT 2020)的评测数据进行实验,结果表明该方法可以有效地提升维汉和蒙汉翻译的翻译质量.

关键词:神经机器翻译;低资源语言;回译;领域相似性;预训练

中图分类号:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2021)04-0675-05

端到端的神经机器翻译方法已经成为目前主流的机器翻译方法[1-3].在高资源语言对之间的翻译任务上,神经机器翻译已经取得令人满意的效果;但对大多数领域或语言对来说,足够数量的高质量平行语料往往难以获取.相对于平行语料而言,单语数据广泛地存在于互联网上,往往更容易获取,因此利用单语数据提高低资源神经机器翻译质量成为一种常用的手段[4-8].

下一个目标语言单词可通过平行语料训练得到的神经机器翻译模型,根据源语言句子和之前的目标语言单词预测得到;而语言模型可以根据之前的目标语言单词给出下一个目标语言单词的概率分布,并且其只需要单语数据训练得到.因此,通过语言模型来利用大规模的单语数据是一个自然的方式.Gulcehre 等[4]提出通过浅融合和深融合的方式将语言模型整合到神经机器翻译模型中,在预测下一个目标语言单词时,该模型可以综合语言模型和翻译模型的打分;Skorokhodov 等[5]利用大量的源语言和目标语言单语数据分别训练源语言和目标语言的语言模型,然后利用语言模型初始化翻译模型的参数,在预测下一个目标语言单词时,合并目标语言模型的打分.虽然语言模型可以提高低资源神经机器翻译的质量,但是通常需要对翻译模型进行大量的修改来整合语言模型和翻译模型,这使神经机器翻译模型变得更加复杂.Sennrich 等[7]提出一个基于回译的数据增强方法.该方法通过一个反向的翻译模型将目标语言单语数据翻译成源语言数据,并和原始目标语言单语数据联合形成伪平行语料;合并伪平行语料和真实的平行语料可以提升语料规模,从而提升翻译效果.Hoang 等[8]提出通过迭代回译,同时利用源语言单语数据和目标语言单语数据进一步提升翻译效果.

基于回译的方法可以不用修改神经机器翻译模型就能有效地利用单语数据;但是当平行语料规模较小时,通过回译生成的伪平行语料质量往往较差,而且大规模的伪数据和较小规模的平行语料混合使得真实的平行语料难以被有效地利用.针对这些问题,本文提出一种针对较小规模平行语料下的解决方案.该方法通过调节字节对编码(BPE)[9]融合数以及失活率(dropout)[10]参数来训练一个尽可能好的低资源神经机器翻译系统,从而提升伪平行语料的质量;将汉语单语按照词覆盖率划分成不同领域相似的单语数据,从而通过回译利用汉语单语数据生成伪平行语料

收稿日期:2020-11-15 **录用日期**:2021-04-28

基金项目:国家自然科学基金(U1703133);新疆自治区高层次人才引进工程项目(Y839031201);新疆维吾尔自治区重点实验室开放课题(2018D04018)

* **通信作者**:xiaoli@ms.xjb.ac.cn

引文格式:张文博,张新路,杨雅婷,等.面向低资源神经机器翻译的回译方法[J].厦门大学学报(自然科学版),2021,60(4):675-679.

Citation:ZHANG W B,ZHANG X L,YANG Y T,et al. Back translation for low resources neural machine translation[J]. J Xiamen Univ Nat Sci,2021,60(4):675-679.(in Chinese)

和高领域相似的伪平行语料;通过分段训练分别利用伪平行语料和领域相似的伪平行语料进行预训练和微调;使用模型平均和模型集成提升系统的鲁棒性,进一步提升翻译质量.

1 数据及预处理

本文使用第 16 届全国机器翻译大会(CCMT 2020)提供的维吾尔语-汉语(维汉)、蒙古语-汉语(蒙汉)平行语料以及汉语单语语料搭建翻译系统,并进行对比实验. 其中维汉平行语料是新闻领域数据,蒙汉平行语料是日常用语数据,汉语单语语料是新闻领域数据.

1.1 预处理

数据预处理阶段,过滤平行语料中重复的句对,将语料中的字母和数字的全角形式转换成半角形式. 在维汉翻译和蒙汉翻译两个任务中,分别使用 moses 脚本(https://github.com/moses-smt/mosesdecoder)处理维吾尔语和蒙古语,使用 Jieba 分词工具(https://github.com/fxsjy/jieba)对汉语分词.

1.2 子词化处理

BPE[9]是目前缓解机器翻译任务中未登录词问题[11]的一种普遍方法. 本文使用 fastBPE 工具(https://github.com/glample/fastBPE)处理维汉翻译和蒙汉翻译这两个任务,融合数分别为 1 000 和 5 000. 并联合源语言和目标语言进行 BPE 切分处理,而不是分别处理源语言和目标语言. 相应地,神经机器翻译模型中源语言和目标语言也共享同一个嵌入(embedding)矩阵. 同时,本文都只使用平行语料作为 BPE 的词频统计语料,因此对汉语单语语料,分别采用维汉平行语料和蒙汉平行语料学习得到的模型进行汉语单语切分.

2 神经机器翻译模型的训练方法

回译[7]是一种能够有效地利用目标端单语数据的数据增强方法. 针对回译的研究有很多,如 Graça 等[12]和 Edunov 等[13]通过研究生成伪平行语料的方式对回译进行改进,Hoang 等[8]提出同时利用源语言单语数据和目标语言单语数据的回译方法. 基于回译的方法中一般都使用和平行语料数量相差不大的单语数据,本文也通过回译利用汉语单语数据提升维汉和蒙汉翻译质量;但为了能在低资源情况下利用尽可能多的汉语单语数据,本文采用首先在大规模伪平行语料上预训练,再在高领域相似的伪平行语料上微调这样分段式的训练方法来训练翻译模型. 除此之外,本文在生成伪平行语料时,还采用随机采样(sampling)[13]和过滤质量较差的伪平行句对来提升翻译质量. 回译所用模型为 Transformer_big.

2.1 伪平行语料生成

首先训练一个目标端到源端的翻译模型,利用翻译模型将筛选到的目标端单语语料翻译成源端单语语料,翻译得到的源端单语语料和原来的目标端单语语料联合构成伪平行语料. 在解码过程中,为了增强数据多样性,本文设集束大小(beam size)为 1,并使用随机采样[13]的方式生成源端数据. 迭代回译[8]可以同时利用源端单数数据和目标端单语数据获得更好的效果,但是本次评测只提供汉语单语数据,因此本文中的伪平行语料由在平行语料训练得到的反向翻译模型和汉语单语数据一次生成. 为了过滤噪声数据,本文删除伪平行语料中长度(BPE 之后)小于 5 或大于 250 以及长度比大于 2 的句对.

2.2 高领域相似伪平行语料

与 Zhang 等[14]的研究类似,本文根据单语句子中所有词在平行语料词典中出现的比例挑选和平行语料领域接近的单语数据. 为了降低平行语料中低频词的干扰,在统计完平行语料的词典(BPE 之前)之后,删除词典中频率小于 3 的词. 对维汉翻译和蒙汉翻译,本文均挑选比例大于 0.9 的汉语单语句子用来生成伪平行语料. 同时本文还从比例等于 1 的汉语单语数据中额外提取一份作为高领域相似的单语数据,并通过回译生成高领域相似伪平行语料,即高领域相似伪平行语料的汉语部分句子中所有词都在平行语料词典中出现过. 生成的伪平行语料和高领域相似伪平行语料主要用于下面的分段式训练方法.

2.3 分段式训练方法

回译常见的做法是将伪平行语料和平行语料合并作为新的平行语料,用来直接训练翻译模型或者微调[7]. 由于单语语料的规模要远大于平行语料,所以直接合并训练并不能有效地利用平行语料. 过采样虽然可以使平行语料和伪平行语料保持接近的比例,但是对平行语料采样次数过多也容易使得模型对平行语料过拟合. 因此本文使用两段式的训练方法:第一阶段只使用所有伪平行语料训练翻译模型;第二阶段将高领域相似的伪平行语料(由经过筛选得到的高领域相似的汉语单语生成)和平行语料结合,继续训练

翻译模型,并且在训练过程中使用过采样使伪平行语料和平行语料保持 1∶1 的比例. 对于维汉翻译任务,本文在第二阶段训练完成之后,进一步使用平行语料微调.

3　模型平均和集成

模型平均和集成都可以提升模型的鲁棒性,有助于进一步提升翻译质量. 前者将同一个模型在训练阶段不同时刻保存的参数平均作为最后的模型参数;后者使用多个模型同时解码,在生成候选词概率表时,将多个模型生成的概率词表平均作为用于生成下一个词的概率表. 本文对翻译模型进行多次训练和微调,在每个训练阶段之后,根据在验证集上的表现,选择平均最后 5 或 10 个模型或最佳模型作为该阶段的输出模型. 对维汉和蒙汉本文分别使用 3 个不同的随机种子训练 3 个模型(这 3 个模型中的每个模型都是平均之后的模型或 best 模型),最后集成这 3 个模型对测试集进行解码.

4　实　验

4.1　参数设置

使用开源工具 fairseq(https://github.com/pytorch/fairseq)在两块 32 G 英伟达 V100 显卡上进行机器翻译模型的训练. 本文采用 transformer_big 模型[3]作为翻译模型,并且使用 GELU 激活函数. 在训练中,使用 fairseq 的 update-freq 将每个批次(batch)的最大标记(token)设置为 64 000. 为了节省时间,在 CCMT 2020 评测时使用 Transformer_base 测试 dropout 和 BPE 参数,最终将维汉翻译任务的 dropout(d_1)设置为 0.3,activation-dropout(d_2)设置为 0.3,attention-dropout(d_3)设置为 0.2,BPE 融合数设为 1 000;蒙汉翻译任务的 d_1 设置为 0.3,d_2 和 d_3 都设置为 0,BPE 融合数设为 5 000. 所有模型都采用 Adam 优化器,在训练过程中,维汉第一阶段使用 0.000 5 的学习率,蒙汉第一阶段使用 0.000 7 的学习率,warmup 设置为 4 000;第二阶段,warmup 都设置为 1 000,学习率为 0.000 3. 对维汉翻译,最后使用固定学习率 0.000 1 在平行语料进一步微调. 本文所有系统在解码时,beam size 均为 12.

4.2　评测结果

对维汉翻译和蒙汉翻译,都分别提交了 3 个结果,其中主系统 primary-a 为通过 3 个随机种子利用汉语单语料分段式训练得到的模型进行集成的结果. 对比系统 contrast-c 为使用单语语料但没有进行集成的单个模型的结果,contrast-b 为只使用平行语料训练得到的单个模型的结果. 表 1 为这 3 个系统在 CCMT 2020 评测结果中的双语互译评估(BLEU5-SBP)分数.

表 1　不同系统在测试集的测试结果

Tab. 1　Test results of different systems on test sets

系统	BLEU5-SBP/%	
	维汉	蒙汉
contrast-b	46.02	43.41
contrast-c	52.14	50.76
primary-a	53.05	52.99

由表 1 可以看出使用汉语单语语料可以显著提升翻译质量,在维汉翻译和蒙汉翻译上分别提升了 6.12 和 7.35 个百分点. 最后使用模型集成可以进一步分别提升约 0.91 和 2.23 个百分点.

4.3　重要参数对比

低资源神经机器翻译往往具有容易过拟合以及存在较多集外词[11]的问题,因此本文使用 transformer_big 模型在平行语料上,通过对 dropout[10,15] 和 BPE[9] 融合数这两个参数的调节来缓解这两个问题. 考虑到 Transformer_big 与 Transformer_base 的最佳参数可能不一样,故对 Transformer_big 的 dropout 和 BPE 参数进行验证. 表 2 是在维汉翻译和蒙汉翻译上使用 30 000 融合数时不同 dropout 的测试结果. 表 3 是 dropout 为表 2 中的最佳取值时,在维汉翻译和蒙汉翻译任务中,不同 BPE 融合数的测试结果.

表 2　不同 dropout 参数在验证集的测试结果

Tab. 2　Test results of different dropout parameters on validation sets

验证集	BLEU/%			
	$d_1=0.1$, $d_2=d_3=0$	$d_1=0.3$		
		$d_2=d_3=0$	$d_2=d_3=0.1$	$d_2=0.3$, $d_3=0.2$
维汉	30.19	37.91	38.19	39.09
蒙汉	52.83	63.07	64.40	62.73

表 3 不同 BPE 融合数参数在验证集的测试结果

Tab. 3 Test results of different BPE parameters on validation sets

BPE 融合数	BLEU/%	
	维汉	蒙汉
1 000	40.10	64.77
5 000	39.97	65.37
10 000	39.92	64.07
30 000	39.09	64.40

从表 2 和 3 中可以看出，Transformer_big 与 Transformer_base 的最佳 dropout 参数并不同，但最佳 BPE 融合数一致. 该结果还表明 dropout 和 BPE 融合数这两个超参数对维汉翻译和蒙汉翻译在低资源神经机器翻译中均有不同程度的影响. 对低资源神经机器翻译，合适的 dropout 可以显著提升翻译质量，合适的融合数也有利于提高翻译质量. 表明通过调整 dropout 和 BPE 参数可以进一步提高系统的翻译性能.

4.4 其他实验结果与分析

为了验证分段式训练方法的有效性，本文使用 multi-bleu. perl（https：// github. com/moses-smt/mosesdecoder/scripts/generic/multi-bleu. perl）在验证集上计算不同系统基于字的 BLEU[16] 值. 表 4 展现本文在本次评测中筛选之后得到的数据规模. 表 5 对比了不同训练方式在验证集上的实验结果.

表 4 不同类型语料的规模

Tab. 4 The scales of different types of corpus

语料类型	维汉/万对	蒙汉/万对
平行语料	16.8	26.4
伪平行语料	954.2	772.9
高领域相似伪平行语料	370.8	282.5

表 5 不同训练策略在验证集的结果

Tab. 5 The results of different training strategies on validation sets

训练策略	BLEU/%	
	维汉	蒙汉
只使用平行语料	40.10	65.37
只使用伪平行语料	44.04	66.01
伪平行语料混合平行语料	45.39	67.59
伪平行语料预训练＋高领域相似伪平行语料混合经过上采样的平行语料微调	46.60	70.79

由表 5 可以看出语料规模对翻译结果有着重要的影响：只使用大量的伪平行语料就可以获得比只使用平行语料更好的性能；在伪平行语料的基础上加上平行语料虽然可以获得进一步的提升，但是提升幅度并不大；而分段式地训练可以在只使用伪平行语料的基础上获得更显著的提升.

5 结 论

本文通过调节 dropout 和 BPE 融合数两个参数，缓解了低资源神经机器翻译易过拟合以及存在较多低频词和集外词的问题；并借助回译，通过分段式训练同时有效地利用单语语料和平行语料资源，较大地提升了低资源情况下维汉翻译和蒙汉翻译的质量.

参考文献：

[1] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-11-15]. https：// arxiv. org/pdf/1409. 0473. pdf.

[2] LUONG M, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [C] // Conference on Empirical Methods in Natural Language Processing. Lisbon：ACL，2015：1412-1421.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Conference on Neural Information Processing Systems. Long Beach：NIPS，2017：1706. 03762.

[4] GULCEHRE C, FIRAT O, XU K, et al. On using monolingual corpora in neural machine translation[EB/OL]. [2020-11-15]. https：// arxiv. org/1503. 03535. pdf.

[5] SKOROKHODOV I, RYKACHEVSKIY A, EMELYANENKO D, et al. A semi-supervised neural machine translation with language models [C] // Workshop on Technologies for MT of Low Resource Languages. Boston：Association for Machine Translation in the Americas，2018：37-44.

[6] IMAMURA K，FUJITA A，SUMITA E. Enhancement of encoder and attention using target monolingual corpora in neural machine translation[C] // Workshop on Neural Machine Translation and Generation. Melbourne：ACL，2018：55-63.

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

[8] HOANG V C D, KOEHN P, HAFFARI G, et al. Iterative back-translation for neural machine translation [C] // Workshop on Neural Machine Translation and Generation. Melbourne: ACL, 2018: 18-24.

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

[10] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors [EB/OL]. [2020-11-15]. http: // export. arxiv. org/pdf/1207. 0580. pdf.

[11] LUONG M T, MANNING C D. Achieving open vocabulary neural machine translation with hybrid word-character models [EB/OL]. [2020-11-15]. https: // arxiv. org/pdf/1604. 00788. pdf.

[12] GRAÇA M, KIM Y, SCHAMPER J, et al. Generalizing back-translation in neural machine translation [C] // Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 45-52.

[13] EDUNOV S, OTT M, AULI M, et al. Understanding back-translation at scale [C] // Annual Meeting of the Association for Computational Linguistics. Brussels: ACL, 2018: 489-500.

[14] ZHANG J J, ZONG C G. Exploiting source-side monolingual data in neural machine translation [C] // Conference on Empirical Methods in Natural Language Processing. Texas: ACL, 2016: 1535-1545.

[15] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[16] KISHORE P, SALIM R, TODD W, et al. BLEU: a method for automatic eval-uation of machine translation [C] // Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 311-318 .

Back translation for low resources neural machine translation

ZHANG Wenbo[1,2,3], ZHANG Xinlu[1,2,3], YANG Yating[1,2,3], DONG Rui[1,2,3], LI Xiao[1,2,3]*

(1. The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China;
2. School of Computer and Technology, University of Chinese Academy of Sciences, Beijing 100049, China;
3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China)

Abstract: Neural machine translation has achieved great success in high-resource situations, but the translation effect in low-resource situations needs to be improved. At present, both Uyghur-Chinese and Mongolian-Chinese translation are low resource translation tasks. This paper proposes to divide Chinese monolingual data into multiple monolingual data according to domain similarity, and to train a translation model on different monolingual data by pre-training and fine-tuning. Then, the translation quality of Uyghur-Chinese and Mongolian-Chinese is further improved by model averaging and model ensemble. Using the evaluation data of the 16th China Conference on Machine Translation (CCMT 2020) for experimental comparison, the results show that this method can effectively improve the translation quality of Uyghur-Chinese and Mongolian-Chinese translation.

Keywords: neural machine translation; low-resource language; back translation; domain similarity; pre-trainning