文章编号:1000-5862(2019)06-0638-05

低资源维汉神经机器翻译研究

王 坤[1],殷明明[1],俞鸿飞[1],韩 冬[1],斯拉吉艾合麦提·如则麦麦提[2],西热艾力·海热拉[2],刘文其[2],艾山·吾买尔[2],李军辉[1],段湘煜[1*],张 民[1]
(1. 苏州大学计算机科学与技术学院,江苏 苏州 215000;2. 新疆大学信息科学与工程学院,新疆 乌鲁木齐 830046)

摘要:该文介绍了在第 15 届全国机器翻译大会的机器翻译评测项目中苏州大学的参赛情况,主要介绍参评系统使用的神经机器翻译模型基准结构以及采用的策略、方法,并介绍该系统在评测数据上的实验性能.

关键词:神经机器翻译;维汉翻译;低资源机器翻译
中图分类号:TP 302.1 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2019.06.13

0 引言

第 15 届全国机器翻译大会(China Conference on Machine Translation,简称 CCMT)的机器翻译技术评测主要由 4 部分组成:双语翻译、多语翻译、语音翻译、翻译质量评估.在双语翻译任务中,评测项目由汉英新闻领域机器翻译(CE)、英汉新闻领域机器翻译(EC)、蒙汉日常用语机器翻译(MC)、藏汉政府文献机器翻译(TC)和维汉新闻领域机器翻译(UC)组成.在 CCMT2019 双语翻译评测任务中,苏州大学自然语言处理团队参加了维汉新闻领域机器翻译的评测项目.

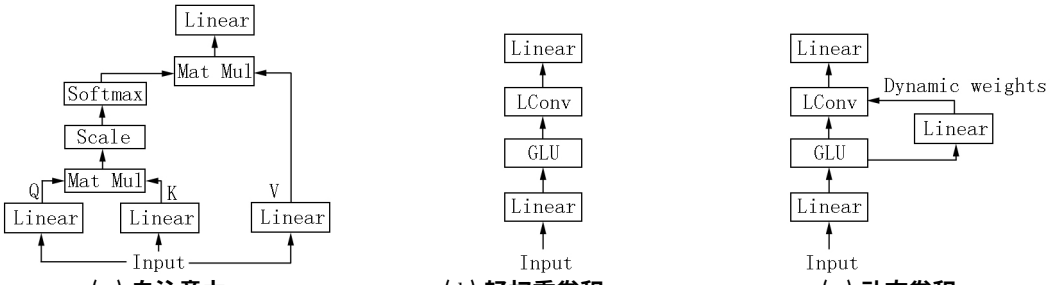
该文详细介绍了在本次参赛中使用的模型结构、数据处理策略、主要技术,以及参赛模型在维汉语言对上的翻译性能.

1 系统

1.1 模型结构

在本次的维汉机器翻译任务中,使用了 Wu Felix 等[1]提出的轻权重卷积(Lightweight Convolution,简称 LightConv)、动态卷积(DynamicConvolution,简称 DynamicConv)的神经机器翻译(Neural Machine Translation,简称 NMT)模型结构,该模型结构的翻译性能优于序列到序列的 NMT 模型[2]、带有注意力机制的 NMT 模型[3-4]、由卷积神经网络组成的 NMT 模型[5].

图 1(a)为自注意力机制图,图 1(b)和图 1(c)展示了由轻权重卷积和动态卷积组成的神经机器翻译系统的模型结构.



(a)自注意力 (b)轻权重卷积 (c)动态卷积

图 1 自注意力、轻量权重卷积及动态卷积图示

收稿日期:2019-08-11
基金项目:国家重点研发计划"政府间国际科技创新合作"重点专项(2016YFE0132100),国家自然科学基金(61673289,61662077)和新疆多语种信息技术实验室开放课题(2016D03023)资助项目.
通信作者:段湘煜(1976-),男,湖南洞口人,教授,博士,主要从事机器翻译、自然语言处理研究. E-mail: xiangyuduan@suda.edu.cn

从文献 [1] 可知: 轻权重卷积的机器翻译模型性能与自注意力机制的神经机器翻译模型[6] 性能相当,动态卷积 NMT 性能优于 Transformer 的性能.

1.2　模型融合

在最终提交的系统中,笔者使用了模型融合的策略. 模型融合的策略分为 2 种: (i) 策略 1,平均在单次训练过程中最好的 n 个模型参数; (ii) 策略 2,融合不同结构或者不同随机数种子的多个模型. 策略 1 在解码过程中输入的模型个数为 1,策略 2 在解码过程中输入的模型个数为 m,m 为不同随机数种子或者不同模型结构训练的 checkpoint 个数. 策略 2 在 Beam Search 阶段中同时需考虑多个模型的候选译文概率,并且平均了多个模型的候选译文概率. 策略 1 可以略微提升机器翻译的性能(+0.1 ~ +0.2 BLEU) ,策略 2 则可以显著提升机器翻译的性能(+1.0 BLEU) [7-8] . 因此,在最终提交的系统中,使用了融合不同结构以及不同随机数种子的多个模型.

2　数据

在本次维汉机器翻译任务中,训练语料由 170 058条维汉双语数据组成,开发集由 1 000 条维汉双语数据组成,测试集 1 和测试集 2 各由 500 条不同的维语数据组成,同时,评测组织方提供了 5 435 155条汉语单语数据. 由于参赛数据规模较小,笔者保留了维汉的所有数据. 下面分别介绍双语的数据处理方法以及回译策略.

2.1　双语数据处理

在双语数据的预处理阶段,采用不同的数据处理策略,并在 Transformer 模型结构上验证不同数据策略的实验性能. 在机器翻译任务中,为了解决数据稀疏以及集外词(Out of vocabulary,简称 OOV) 的问题,R. Sennrich 等[9-10] 提出了双字节编码(Byte pair encoding(https: / /github. com/rsennrich/subword-nmt) ,简称 BPE) 的方式,该方法可以将单词切分成子词单元,有效地缓解了机器翻译中罕见词和未登录词较难翻译的现象. 在本次的维汉机器翻译评测任务中,笔者也对维吾尔语和汉语进行了 BPE 处理,同时也对其他的数据处理方式进行了验证,主要包括: (i) 维吾尔语的拉丁化处理; (ii) 维吾尔语的音节化切分策略; (iii) 汉语的文字分词策略.

维吾尔语拉丁化处理是指将难以被非该语种使用者理解的维语转化为拉丁语形式. 维吾尔语是一种典型的黏着语,形态比较复杂,单词一般由词干和词缀组成,一个词可能会有多种变形,从而造成了语言词汇量比较丰富、低频词较多的情况,导致在训练过程中存在严重的数据稀疏问题和 OOV 问题. 在本次的评测任务中,新疆大学自然语言处理团队开发了维吾尔语音节切分程序,该程序将维吾尔语单词切分为音节,以更小的粒度作为翻译单元.

维吾尔语音节化处理过程为: 在一般情况下,在维吾尔语中音节由 1 个元音或 1 个元音加 N 个辅音结合组成. 音节结构通常是(起音) + 领音 + (收音) ,其中领音必须是元音而且音节中不能没有,起音和收音可以有也可以没有. 用 C 表示辅音,用 V 表示元音,单词音节类型总共有 12 种,可表示为 V、VC、CV、CVC、VCC、$CVCC$、CCV、$CCVC$、$CCVCC$、CVV、$CVVC$、$CCCV$. 其中前 6 个是常见的维吾尔语单词音节类型,后 6 个是外来词的音节类型. 一般在维吾尔语中 CV 与 CVC 类型的音节出现频率最高,外来词音节出现频率较低. 新疆大学团队按照这 12 种音节类型规则来研发音节切分工具,该工具切分正确率为 96.0% ~99.6%.

表 1 为维吾尔语处理的例子.

表 1　维吾尔语切分实例

数据处理方法	维吾尔语数据
–	، پېشىگەنەكشىنلزۇيااقتايىققمر ەتكىلتەپۇسىر ۇقويىنتىششىپە ەتتەنرۇسىر ۇقويىۋگكۇج . ۇدىتاۋىرىكەگرۋەدىگىپىيىداسىتقىنە
BPE 切分	ىشىپە ەتتەنرۇسىر ۇقويى@@ىنلزۇيااقتايىققمر ەتكىلتەپۇسىر ۇقويىنتىش@@نەكشىپىشىگە . رۋەدىگىپىيىداسىتقىنەۋگكۇج ، @@ىرىكەگ@@ۇدىتاۋ
音节化切分	ەكشىنىلزۇيااقتايىققمر ەتكىلتەپۇسىر ۇقويىنتىششىپە ەتتەنرۇسىر ۇقويى . ۇدىتاۋ ىر ىكەگرۋەد ىگىپىيىد اسىتقىنەۋگكۇج ، پىشىگەن

2.2 回译

在机器翻译中,为了更好地利用丰富的、大规模的目标端单语数据,R. Sennrich 等[9-10]提出了回译(Back Translation,简称 BT)的方法. 在回译的方法中,首先需要训练从目标语言到源语言的机器翻译模型,并通过该模型翻译大规模的目标端单语数据,获得伪源端语料;其次将该伪源端语料与原始的双语数据合并,构成新的数据,作为模型的训练数据. 回译的方法能有效地提升低资源机器翻译的译文质量,是目前机器翻译评测的重要方法.

2018 年,S. Edunov 等[11]对回译方法进行了进一步分析. 他们的研究表明:在生成伪源端语料时采用抽样的方法或者取 topk 的方法,回译方法的翻译效果高于采用 beam search 生成伪源端语料的方法;随机打乱生成的伪源端数据,可以进一步提升机器翻译性能. 在本次的维汉机器翻译评测中,对该实验方法进行重现,并将该方法融合到最终提交的模型中.

3 实验

3.1 实验环境

在本次的机器翻译评测中,笔者使用的系统为 Ubuntu 16.04.2,内存大小为 256 GB,显卡型号为 GTX1080Ti,显存大小为 12 GB.

3.2 实验设置

在本次的机器翻译评测中,使用了 Facebook AI Research 开源的神经机器翻译程序 fairseq[12](https://github.com/pytorch/fairseq),该系统集成了 LSTM、Transformer、LightConv 以及 DynamicConv 等机器翻译模型结构. 在本次参赛系统中,使用的 LSTM、Transformer、LightConv 以及 DynamicConv 的参数设置如表 2 所示,其中第 1 行为模型结构,第 1 列为超参数设置.

表 2 模型参数设置

超参数设置	LSTM	Transformer	LightConv	DynamicConv
arch	lstm	transformer	lightconv	lightconv
clip_norm	0	0	0	0
optimizer	adam[13]	adam	adam	adam
lr	0.000 5	0.000 5	0.000 5	0.000 5
min-lr	1e-09	1e-09	1e-09	1e-09
weight-decay	0.000 0	0.000 0	0.000 1	0.000 1
label-smoothing	0.1	0.1	0.1	0.1
ddp-backend	–	–	no_c10d	no_c10d
其他超参数	–	–	–	encoder-glu 1,decoder-glu 1

3.3 实验结果

在本次的机器翻译评测中,评测指标使用自动评价的方式,主要的评价标准为 BLEU_SBP、BLEU_NIST、BLEU 等(http://114.212.189.224:8080/mt-web/about). 在评测任务开启后,首先使用 Transformer 模型结构验证不同的维吾尔语以及汉语的数据处理策略在开发集上的 BLEU 值[14]. 不同数据处理的优劣如表 3 所示.

表 3 不同的数据处理策略测试集译文 BLEU

维语处理	汉语处理	Dev BLEU
+拉丁化	jieba 分词 + BPE	36.31
+ Tokenizer + BPE	+ Tokenizer + jieba 分词 + BPE	37.62
+ Tokenizer + BPE(删除@@)	+ Tokenizer + jieba 分词 + BPE(删除@@)	38.18
+ Tokenizer + BPE(删除@@)	+ Tokenizer + pkuseg 分词 + BPE(删除@@)	38.20
+音节化	+字符切分	38.76

在表 3 中,Tokenizer(https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl)为 Mosesdecoder[15]中的切分脚本;jieba(https://github.com/fxsjy/jieba)分词和 pkuseg[16](https://github.com/lancopku/pkuseg-python)分词为开源的中文分词程序;删除“@@”为在对数据进行 BPE 处理后,将数据中的“@@”删除,这样可以有效减少词典大小. 删除“@@”的例子如表 4 所示.

表 4 BPE 示例

BPE	BPE(删除@@)
害怕批评是进步的绊@@ 脚@@ 石.	害 怕 批 评 是 进 步 的 绊脚石.

在进一步的实验中,从模型层面出发,对不同的模型结构进行验证. 使用了最优的数据处理方法(维吾尔语音节化,汉语字符切分)分别验证了 Transformer、LightConv 以及 DynamicConv 模型结构对实验性能的影响. 实验结果如表 5 所示.

表 5　Transformer 和 Lightconv 翻译性能对比

	Transformer	LightConv	DynamicConv
Dev BLEU	38.76	40.14	40.69

在表 5 的 LightConv 以及 DynamicConv 实验中,将长度惩罚项 lenpen 设置为 2.0,并且译文长度设置为不超过原文长度 +20.

进一步地,在 DynamicConv 模型上,使用回译方式对训练数据进行扩充. 从汉语单语数据中随机挑选了 100 万条数据,并训练了随机数种子为 1 和随机数种子为 2 的汉语-维吾尔语的 DynamicConv 模型,使用模型融合的方式获得反向翻译的译文,增加了回译方法,处理结果如表 6 所示.

表 6　回译方法性能对比

回译方法	Dev BLEU
DynamicConv	40.69
DynamicConv + BT(beam search)	41.73
DynamicConv + BT(topk)	41.45
DynamicConv + BT(sample)	42.11

Beam search、topk、sample 为重现 S. Edunov 等[11] 的实验方法,发现对反向翻译的维吾尔语增加噪声可以进一步提升机器翻译性能. 因此,在最终提交的实验系统中,融合了多种模型结构以及实验方法,实验结果如表 7 所示.

表 7　多个系统模型融合的性能对比

模型	Dev BLEU
LSTM + BT(topk + noise)	40.22
DynamicConv + BT(beam search + noise)	43.11
DynamicConv + BT(topk + noise)	42.13
DynamicConv + BT(sample + noise)	42.94
Ensemble models above	45.15

3.4　提交结果

不同参赛单位在测试集上的实验结果如表 8 所示.

表 8　不同单位测试集上翻译性能(前 10)

参赛单位	BLEU_SBP
腾讯科技(北京) 有限公司	48.08
中国科学院自动化所	46.94
苏州大学	45.67
中国科学技术大学	43.49
北京理工大学	43.05
北京交通大学	40.67
中科软科技股份有限公司	39.55
中央民族大学	38.17
北京航空航天大学	33.36
南京信息技术研究院	30.49

4　总结

本文介绍了苏州大学自然语言处理团队在第 15 届全国机器翻译大会中参加低资源语言对从维吾尔语到汉语进行机器翻译的任务情况.

在本次机器翻译评测中所做的主要工作有: (i) 在数据处理方面,对维吾尔语和汉语的多种处理方法进行了验证,最终选择维吾尔语按音节化处理、汉语按字符切割的方法,并通过生成伪预料的方式对原始训练数据进行扩展; (ii) 在模型结构方面,采用了最新的轻权重卷积以及动态卷积的神经机器翻译模型,并且使用了模型融合的策略,在解码过程中同时考虑多个模型预测的译文概率分布. 以上工作在测试集上取得了较好的效果.

但是由于时间有限,笔者验证过的单语、双语的数据策略并没有融合到最终提交的系统中,包括数据筛选、词典使用以及 re-rank 等方法. 线下的验证结果表明,这些策略可以进一步提升机器翻译的性能,并且具有良好的应用场景.

5　参考文献

[1] Wu Felix, Fan Angela, Baevski A, et al. Pay less attention with lightweight and dynamic convolutions [EB/OL]. [2019-05-11]. https://arxiv.org/abs/1901.10430v1.

[2] 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.

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-05-11]. https://arxiv.org/abs/1409.0473.

[4] Wu Yonghui, Schuster M, Chen Zhifeng, et al. Google's neural machine translation system: bridging the gap between human and machine translation [EB/OL]. [2019-04-10]. http://arxiv.org/pdf/1609.08144v1.

[5] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning [C] // DoinaPrecup, Yee Whye The. Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia: PMLR, 2017: 1243-1252.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Isabelle Guyon, Ulrike von Luxburg, Samy-Bengio, et al. Advances in neural information processing systems. 30: Annual Conference on Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017: 5998-6008.

[7] Wang Yuguang, Cheng Shanbo, Jiang Liyang, et al. Sogou

neural machine translation systems for wmt17 [EB/OL]. [2019-04-18]. https://www.aclweb.org/anthology/W17-4742.pdf.

[8] Deng Yongchao, Cheng Shanbo, Lu Jun, et al. Alibaba's neural machine translation systems for wmt18 [EB/OL]. [2019-04-25]. https://dblp.uni-trier.de/pers/hd/c/Cheng:Shanbo.

[9] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [EB/OL]. [2019-05-11]. https://arxiv.org/abs/1508.07909.

[10] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [EB/OL]. [2019-05-11]. https://arxiv.org/abs/1511.06709.

[11] Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale [EB/OL]. [2019-05-11]. https://arxiv.org/abs/1808.09381.

[12] Ott M, Edunov S, Baevski A, et al. Fairseq: a fast, extensible toolkit for sequence modeling [EB/OL]. [2019-03-25]. https://arxiv.org/abs/1904.01038v1.

[13] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. [2019-05-25]. https://arxiv.org/abs/1412.6980.

[14] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [EB/OL]. [2019-03-22]. https://www.bibsonomy.org/bibtex/20f9b18cd58fb3a930a96fe9fa9f39896/izzy278.

[15] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation [EB/OL]. [2019-03-29]. http://cs.jhu.edu/~ccb/publications/jhu-summer-workshop-final-report.pdf.

[16] Luo Ruixuan, Xu Jingjing, Zhang Yi, et al. PKUSEG: a toolkit for multi-domain Chinese word segmentation [EB/OL]. [2019-03-29]. https://www.researchgate.net/publication/334082086_PKUSEG_A_Toolkit_for_Multi-Domain_Chinese_Word_Segmentation.

The Study on Low-Resource Uygur-Chinese Neural Machine Translation

WANG Kun[1], YIN Mingming[1], YU Hongfei[1], HAN Dong[1], SILAJIAIHEMAITI · Ruzemaimaiti[2], XIREAILI · Hairela[2], LIU Wenqi[2], AISHAN · Wumaier[2], LI Junhui[1], DUAN Xiangyu[1*], ZHANG Min[1]

(1. School of Computer Science and Technology, Soochow University, Suzhou Jiangsu 215000, China;
2. College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046, China)

Abstract: The submission systems of Soochow University for the 15th CCMT on the task of Low Resource Language Pair Translation are mainly introduced in the paper. The report principally describes the benchmark structure of the neural machine translation model used in the task, fundamental strategies and methods adopted, as well as experimental performance on evaluation data.

Key words: neuralmachine translation; Uyghur-to-Chinese translation; low resource machine translation

(责任编辑:冉小晓)