

基于 mRASP 的藏汉双向神经机器翻译研究

杨 丹^{1,2,3}, 拥 措^{1,2,3*}, 仁青卓玛^{1,2,3}, 唐超超^{1,2,3}

(1. 西藏大学 信息科学技术学院, 西藏 拉萨 850000;

2. 西藏自治区藏文信息技术人工智能重点实验室, 西藏 拉萨 850000;

3. 藏文信息技术教育部工程研究中心, 西藏 拉萨 850000)

摘 要:藏汉机器翻译技术的研究对于弘扬和传承优秀民族文化,推进藏族地区经济、教育和文化的发展有着十分重要的现实意义。该文立足于藏汉平行语料匮乏而导致的藏汉神经机器翻译效果欠佳的问题,对跨语言预训练模型进行了研究。使用第十八届全国机器翻译大会(CCMT 2022)的藏汉数据集构建藏汉双语的跨语言预训练模型(mRASP),采用谷歌的Transformer神经网络机器翻译架构作为基线模型,主要利用数据增强的方式对藏汉平行语料进行扩充、优化藏汉机器翻译所用到的词表,并探索跨语言预训练模型中的联合词表对翻译性能的影响,最终提出了一种融合跨语言预训练模型(mRASP)与改进后的绿色联合词表的藏汉双向神经机器翻译。经过上述策略,藏汉翻译任务上的BLEU值达到了55.69,汉藏翻译任务上的BLEU值达到了29.57。与传统的基于预训练模型的藏汉双向神经机器翻译相比,在稀缺资源条件下有效地提升了藏汉双向机器翻译的性能。

关键词:跨语言预训练模型;藏汉双向神经机器翻译;mRASP;数据增强;词表

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2023)12-0200-07

doi:10.3969/j.issn.1673-629X.2023.12.028

Research on Tibetan-Chinese Bidirectional Neural Machine Translation Based on mRASP

YANG Dan^{1,2,3}, YONG Cuo^{1,2,3*}, RENQING Zhuo-ma^{1,2,3}, TANG Chao-chao^{1,2,3}

(1. School of Information Science and Technology, Tibet University, Lhasa 850000, China;

2. State Key Laboratory of Artificial Intelligence for Tibetan Information Technology in

Tibet Autonomous Region, Lhasa 850000, China;

3. Ministry of Education Engineering Research Center for Tibetan Information Technology, Lhasa 850000, China)

Abstract: The study of Tibetan-Chinese machine translation technology is of great practical significance to promote and inherit excellent national culture and advance the development of economy, education and culture in Tibetan areas. Based on the problem of poor Tibetan-Chinese neural machine translation caused by the lack of Tibetan-Chinese parallel corpus, we investigate the cross-linguistic pre-training model. We use the Tibetan-Chinese dataset from the 18th National Conference on Machine Translation (CCMT 2022) to construct the cross-lingual pre-training model (mRASP) for Tibetan-Chinese bilingualism, and adopt Google's Transformer neural network machine translation architecture as the baseline model, and mainly use data augmentation to expand the Tibetan-Chinese parallel corpus and optimize the vocabulary used in Tibetan-Chinese machine translation, and explore the influence of the joint vocabulary in the cross-language pre-training model on the translation performance. Finally, a Tibetan-Chinese bidirectional neural machine translation that integrates the cross-language pre-training model (mRASP) and the improved green joint vocabulary is proposed. Through the above strategies, the BLEU value on the Tibetan-Chinese translation task reached 55.69, and the BLEU value on the Chinese-Tibetan translation task reached 29.57. Compared with the traditional Tibetan-Chinese bidirectional neural machine translation based on pre-trained model, it effectively improves the performance of Tibetan-Chinese bidirectional machine translation under the condition of scarce resources.

Key words: cross - language pre - training model; Tibetan - Chinese bidirectional neural machine translation; mRASP; data

收稿日期:2023-02-02

修回日期:2023-06-07

基金项目:国家重点研发计划项目(2017YFB1402202);西藏自治区科技创新基地自主研究项目(XZ2021HR002G);西藏大学珠峰学科建设计划项目(zf22002001)

作者简介:杨 丹(1997-),女,硕士研究生,研究方向为自然语言处理;通信作者:拥 措(1974-),女,博士,教授,CCF会员(74089M),研究方向为藏语自然语言处理、藏文文献智能处理。

augmentation; vocabulary

0 引言

随着网络的快速发展,全球各国之间联系日益紧密、各民族交流日趋频繁。语言作为交流的基础,对实现不同种语言之间的翻译显得尤为重要。由于人工翻译代价很高,为了满足人们的翻译需求,机器翻译凭借翻译速度快,低成本等优点受到了人们的青睐^[1]。在自然语言处理(Natural Language Processing, NLP)任务中,机器翻译作为其重要分支和人们的日常生活息息相关。机器翻译实现了计算机在不同种语言之间的自动转换,纵观机器翻译的发展史,它经历了基于规则的机器翻译(Rule-Based Machine Translation, RBMT)、统计机器翻译^[2](Statistical Machine Translation, SMT)以及神经机器翻译^[3](Neural Machine Translation, NMT)三个主要阶段。

近年来,跨语言预训练语言模型在 NLP 任务上受到普遍关注,比如 mBERT^[4], MASS^[5], XLM^[6], XLM-R^[7], mBART^[8]等。它们在大量语料上进行预训练,然后在下游任务中按照其特点对模型微调。这种预训练加微调的方式在一系列 NLP 任务中取得了很好的效果。2020 年陆金梁、张家俊提出了一种基于 Multi-BERT 跨语言联合编码预训练的语言模型的译文质量估计(Quality Estimation, QE)方法,使用不同神经网络对预训练语言模型进行微调^[9]。2021 年满志博等人针对汉语、英语以及缅甸语三种语言结构差异较大而导致的共享词表大小受限的问题,提出进行联合语义表征来提升缅汉英机器翻译模型的性能^[10]。翁荣祥等人提出 APT 框架,从预训练模型中获取知识到神经机器翻译,在跨语言机器翻译任务上的试验结果表明,该模型优于强基线和微调模型^[11]。黄昊阳等人介绍了 Unicoder。给定一个任意的 NLP 任务,可以在 Unicoder 基础上使用一种语言的训练数据对模型进行训练,并直接应用于其他语言相同任务的输入。同时对多种语言微调可以进一步提升效果^[12]。

在多语言机器翻译中,林泽辉等人提出 mRASP 模型^[13](multilingual Random Aligned Substitution Pre-training, mRASP),其关键思想是随机对齐替换技术(Random Aligned Substitution, RAS)。可以在预训练后,在下游语言对模型微调。首次验证使用多个语言对的少量语料数据可以提高资源丰富的机器翻译,并且可以提高预训练语料库中未曾出现过的其他语言的翻译质量,可以在不同语言中构建语义空间的桥梁,从而有效提高翻译性能。

随着藏汉机器翻译的兴起,很多高校和机构开始研究藏汉统计机器翻译。比如:1998 年,陈玉忠等人

顺利研发出班智达汉藏科技机器翻译系统。2003 年在此基础上研发了基于规则的实用化汉藏机器翻译系统^[14],为基于规则的汉藏机器翻译奠定了坚实的理论基础。2013 年,周毛先提出了基于混合策略的汉藏机器翻译系统^[15]。随后,中科院计算所、中科院软件所、厦门大学、西藏大学、青海师范大学等单位开展了藏汉统计机器翻译的研究工作。2014 年,华却才让提出基于树到串的藏语机器翻译^[16],这是中国第一个基于藏文句法信息的统计机器翻译系统;2015 年,位素东提出基于短语的藏汉统计翻译^[17];2016 年,西藏大学尼玛扎西教授的团队研发完成“阳光藏汉双向机器翻译系统”,并面向社会提供翻译服务,系统在汉藏现代公文领域的翻译平均准确率达到 70%,速度也较高^[18]。近几年,研究人员开始研究藏汉神经机器翻译。比如,2017 年,李亚超等人通过迁移学习方法进行了藏汉神经机器翻译的实验^[19];2018 年,蔡子龙等人利用数据增强技术对语料扩充,增强了藏汉机器翻译的泛化能力^[20];2019 年,慈祯嘉措等人将藏语单语模型融合到神经机器翻译中^[21];2021 年,头旦才让等人改进了字节对编码算法,优化了汉藏神经机器翻译^[22];同年,该学者融入了藏文命名实体识别技术,提出了藏文长句分割方法^[23];2022 年,周毛先为了提高翻译的质量,提出一种融合先验知识的方法^[24];同年,孙义栋等学者对机器翻译的词表进行了优化,显著提升了翻译性能^[25];杨丹等学者经过对数据增强策略的深入研究,有效缓解了因平行语料匮乏而导致的翻译性能较差的问题^[26]。

以上学者提出的方法有效改善了藏汉双向机器翻译的性能,但是由于藏汉平行语料匮乏、语料的质量以及现有语料的领域限制,藏汉机器翻译的性能相较于其他大语种的翻译性能来说效果较差。而 mRASP 是针对机器翻译任务而提出的多语言预训练模型,其翻译效果已经超过 mBART。因此,该文使用一种融合跨语言预训练模型(mRASP)与改进后的联合词表的藏汉双向机器翻译,从而进一步提高藏汉双向机器翻译的质量。相比基线系统来说,在藏汉/汉藏翻译上提高了 3.43/1.27 个 BLEU 值。

1 mRASP 多语言神经机器翻译

mRASP 的关键思想是随机对齐替换技术(RAS),该技术使多种语言中具有相似含义的单词和短语在表示空间中更接近。它利用多个语言对的少量平行语料训练模型,然后在下游语言对微调。

RAS 技术采用了无监督词对齐方法(MUSE)。

(1) 同义词替换。在训练集中随机抽取 15 万条数据采用同义词替换的方式扩充语料。在进行藏语的同义词替换时, 使用 50 万条藏语单语语料训练 word2vec 模型, 从句子中根据替换率分别为 0.08, 0.15 的频率随机选择非停用词进行替换。汉语语料借助中文近义词工具包 Synonyms, 从句子中根据替换率分别为 0.08, 0.15 的频率随机选择非停用词进行替换。

把“འཕེལ་ཁུར་གྱི་སྒྲུབ་པ་ལོ་རྒྱུས་མཐུན་པའི་”翻译成“发展势头”,这3种译文相较于实例来说,mRASP+基线的翻译更接近它所表达的含义。另外,实例中包含了国家名,3个译文均将“ཀྲུང་གཞི་རྒྱུ་”译成“中塞”,结果得当。

表7 藏汉翻译结果对比

[illegible]

表 8 汉藏翻译结果对比

模型

“十二五”规划圆满完成，“十三五”规划顺利实施，经济社会发展取得历史性成就、发生历史性变革

基线

“十四五”规划圆满完成，“十五五”规划顺利实施，经济社会发展取得历史性成就、发生历史性变革

ALBERT
+基线

“十四五”规划圆满完成，“十五五”规划顺利实施，经济社会发展取得历史性成就、发生历史性变革

mRASP+
基线

“十四五”规划圆满完成，“十五五”规划顺利实施，经济社会发展取得历史性成就、发生历史性变革

参考
译文

“十四五”规划圆满完成，“十五五”规划顺利实施，经济社会发展取得历史性成就、发生历史性变革

在翻译实例中,包含了“十二五”“十三五”两个专有名词,ALBERT+基线模型将“十二五”错译为“十三

五”,基线、mRASP+基线翻译准确;基线、ALBERT+基线把“圆满完成”翻译成“བུ་གསུམ་གྲུ་ལོ་གསུམ་འགྲུ་གྲུ་”它偏向于书面语,mRASP+基线中的“圆满完成”翻译为“ཕུན་སུམ་ཚུགས་པ་འོང་ངང་ལོ་གསུམ་གྲུ་བ་བྲུ་ང་བ”,该表达在日常生活和书面语中比较常用。总而言之,mRASP+基线的翻译结果更准确。

上述译文的对比证实了融合 mRASP 模型的有效性,它提高了藏汉/汉藏的翻译效果。

4 结束语

针对 CCMT 2022 提供的藏汉综合领域的平行语料,在 transformer-big 框架下,通过 VOLT 改进词表、探索联合词表对翻译性能的影响,并在 mRASP 跨语言预训练模型上进行融合。通过实验表明,利用 VOLT 改进词表可以对藏汉机器翻译的性能有一定提升;虽然在预训练时没有加入藏语语种进行训练,但是 mRASP 跨语言预训练模型仍然可以有效提高藏汉双向机器翻译的性能。

在下一步研究中,将计划收集更高质量、领域覆盖面更广的藏汉平行语料,也将探索更好的跨语言预训练模型来进一步提高藏汉双向机器翻译的性能。

参考文献:

- [1] 肖 桐,朱靖波. 机器翻译:基础与模型[M]. 北京:电子工业出版社,2021;23-26.
- [2] BROWN P F, COCKE J, DELLA PIETRA S A, et al. A statistical approach to machine translation [J]. Computational Linguistics, 1990, 16(2): 79-85.
- [3] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proc of NIPS. Bangkok: NIPS Foundation, 2014; 3104-3112.
- [4] KENTON J D M W C, TOUTANOVA L K. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proc of NAACL - HLT. Minneapolis: ACL, 2019; 4171-4186.
- [5] SONG K, TAN X, QIN T, et al. Mass; masked sequence to sequence pre-training for language generation[C]//Proc of ICML. California: IMLS, 2019; 5926-5936.
- [6] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining [C]//Proc of NIPS. Vancouver: NIPS Foundation, 2019; 7059-7069.
- [7] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale [C]//Proc of ACL. Washington: ACL, 2020; 8440-8451.
- [8] LIU Y, GU J, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation [J]. Transactions of the Association for Computational Linguistics, 2020, 8(1): 726-742.

- [9] 陆金梁,张家俊. 基于多语言预训练语言模型的译文质量估计方法[J]. 厦门大学学报:自然科学版,2020,59(2):151-158.
- [10] 满志博,毛存礼,余正涛,等. 基于多语言联合训练的汉-英-缅神经机器翻译方法[J]. 清华大学学报:自然科学版,2021,61(9):927-935.
- [11] WENG R, YU H, HUANG S, et al. Acquiring knowledge from pre-trained model to neural machine translation[C]//Proc of AAAI. New York: AAAI, 2020: 9266-9273.
- [12] HUANG H, LIANG Y, DUAN N, et al. Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks[C]//Proc of EMNLP. Hong Kong: SIGDAT, 2019: 2485-2494.
- [13] LIN Z, PAN X, WANG M, et al. Pre-training multilingual neural machine translation by leveraging alignment information[C]//Proc of EMNLP. Punta Cana: SIGDAT, 2020: 2649-2663.
- [14] 陈玉忠,俞士汶. 藏文信息处理技术的研究现状与展望[J]. 中国藏学,2003(4):97-107.
- [15] 周毛先. 基于混合策略的汉藏机器翻译系统的构建方法研究[D]. 西宁:青海师范大学,2013.
- [16] 华却才让. 基于树到串藏语机器翻译若干关键技术研究[D]. 西安:陕西师范大学,2014.
- [17] 位素东. 基于短语的藏汉在线翻译系统研究[D]. 兰州:西北民族大学,2015.
- [18] 仁青东主,头旦才让,尼玛扎西. 汉藏机器翻译研究综述[J]. 中国藏学,2019(4):222-226.
- [19] 李亚超,熊德意,张民,等. 藏汉神经网络机器翻译研究[J]. 中文信息学报,2017,31(6):103-109.
- [20] 蔡子龙,杨明明,熊德意. 基于数据增强技术的神经机器翻译[J]. 中文信息学报,2018,32(7):30-36.
- [21] 慈祯嘉措,桑杰端珠,孙茂松,等. 融合单语语言模型的藏汉机器翻译方法研究[J]. 中文信息学报,2019,33(12):61-66.
- [22] 头旦才让,仁青东主,尼玛扎西,等. 基于改进字节对编码的汉藏机器翻译研究[J]. 电子科技大学学报,2021,50(2):249-255.
- [23] 头旦才让. 汉藏神经机器翻译关键技术研究[D]. 拉萨:西藏大学,2021.
- [24] 周毛先. 融合先验知识的藏汉神经机器翻译研究[D]. 西宁:青海师范大学,2022.
- [25] 孙义栋,拥措,杨丹. 基于 VOLT 的藏汉双向机器翻译[J]. 计算机与现代化,2022(5):28-32.
- [26] 杨丹,孙义栋,拥措. 基于数据增强的藏汉神经机器翻译研究[J]. 计算机与数字工程,2022,50(11):2473-2477.
- [27] XU J, ZHOU H, GAN C, et al. Vocabulary learning via optimal transport for neural machine translation[C]//Proc of ACL (volume 1: long papers). Bangkok: ACL, 2021: 7361-7373.
- [28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30(2):6000-6010.
- [29] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proc of ACL. PA: ACL, 2002: 311-318.
- [30] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self-supervised learning of language representations[C]//Proc of ICLR. Addis Ababa: ICLR, 2020: 1057-1073.