Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(2), 69-76
Published Online May 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.122010

Hans汉斯

基于汉字简繁转换的汉日神经机器翻译数据增强研究

张津一*，高忠辉，郭　聪

沈阳理工大学，信息科学与工程学院，辽宁 沈阳

收稿日期：2023年4月10日；录用日期：2023年5月9日；发布日期：2023年5月29日

摘　要

本文提出了一种基于汉字简繁转换的神经机器翻译(Neural Machine Translation, NMT)数据增强方法，旨在通过利用简繁转换表将源端文字替换为目标端文字，从而融合汉字简繁转换信息，并提高翻译质量。本文将此方法应用于汉日机器翻译任务，实验结果表明此方法是一种有效的数据增强方法，可以显著地提高汉日机器翻译质量。

关键词

神经机器翻译，简繁汉字转换，数据增强，汉日翻译

Research on Data Augmentation for Chinese-Japanese Neural Machine Translation Based on Conversions between Traditional Chinese and Simplified Chinese

Jinyi Zhang*, Zhonghui Gao, Cong Guo

School of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning

Received: Apr. 10th, 2023; accepted: May 9th, 2023; published: May 29th, 2023

Abstract

This paper proposed a neural machine translation (NMT) data augmentation method based on

*通讯作者。

文章引用: 张津一, 高忠辉, 郭聪. 基于汉字简繁转换的汉日神经机器翻译数据增强研究[J]. 人工智能与机器人研究, 2023, 12(2): 69-76. DOI: 10.12677/airr.2023.122010

conversions between Traditional Chinese and Simplified Chinese. The method aimed to integrate the information of conversions between Traditional Chinese and Simplified Chinese by replacing the source text with target text according to the Chinese characters mapping table, thereby improving the translation quality. The method was applied to the Chinese-Japanese machine translation task, and the experimental results demonstrated that this approach was an effective data augmentation method and could significantly improve the translation quality of Chinese-Japanese machine translation.

Keywords

Neural Machine Translation, Conversions between Traditional Chinese and Simplified Chinese, Data Augmentation, Chinese-Japanese Translation

1. 引言

汉日机器翻译是指用计算机自动将汉语文本翻译成日语文本，或将日语文本翻译成汉语文本的过程。汉日机器翻译既涉及汉语与日语这两种语言，也涉及人工智能与计算机科学这两个领域。

中国和日本是东亚地区最大的两个国家，也是东亚地区最重要的经济体。中日两国在技术、文化、经济等方面的交流与合作已有几千年的历史，但随着时代的发展，中日两国的交流和合作变得越来越紧密。近几十年来，随着中国的经济腾飞，中国成为世界第二大经济体。中国的经济增长也带动了中日两国的贸易增长，中国成为日本的最大贸易伙伴，日本也成为中国的第二大贸易伙伴。随着中日两国经济关系的加强，中日两国的文化交流和科技合作也变得更加密切。由于中日两国语言的不同，语言障碍成为了阻碍中日两国更紧密合作的一个重要因素。因此，汉日机器翻译在这种背景下应运而生，成为了解决中日两国语言障碍的有效手段。

随着人工智能技术的飞速发展，神经机器翻译(Neural Machine Translation, NMT)已经取代了过去的翻译方式，成为了当下最好的翻译应用框架[1] [2] [3]。比起其他常用语种的翻译，汉语与日语这两个相近的语言具有很多共同的特点，但相关机器翻译研究的翻译精度仍未达到实用要求，仍需引导和关注。

想要得到高质量翻译结果的话，NMT 对汉日对译语料库的数据量的需求很大。但是，相比于汉英的公开平行对译语料的数百万至数千万句对，汉日的公开平行对译语料的规模只有几十万句对。如 ASPEC-JC 与 WCC-JC 汉日对译语料库，仅有约 67 万与 215 万对译句子，比起其他语言对的千万级别的量，具有较大差距[4] [5] [6] [7]。

数据增强(Data Augmentation, DA)技术缓解了深度学习中数据不足的场景，在图像领域首先得到广泛使用，进而延伸到自然语言处理领域，并在许多任务上取得效果。主要的方向是增加训练数据的多样性，从而提高模型泛化能力。在 NMT 中，经常使用的方法是回翻译：将原始数据从目标语言 A 翻译到中间语言 B，然后再翻译回目标语言 A，从而实现对原始数据的改写[8]。另一种是单向翻译，将原始数据从语言 A 翻译到语言 B，从而扩充语言 B 的数据，此种方法多出现在多语言场景中。优势是容易使用、使用范围广、保证句法跟语义不变。劣势是不可控且多样性受限，会受限于固定的翻译模型。归根结底，回翻译旨在生成跟原始数据语义尽可能相似的新数据。还有的方法则是在原始数据上添加微弱的噪声，不至于严重影响文本原来的语义，但又能跟原始数据存在一定的差异性。这种方式不仅可以实现训练数

据的扩充，还能增加模型训练的难度，提高模型的鲁棒性。

日语和汉语的读法和语法构造完全不同，但是却有很多相同的特征。比如在汉字方面上，有相当一部分汉字是日语和汉语都通用的(简繁转换后)。而且很多词语的意义也是相同的，这在世界范围内都是很罕见的特征。虽然汉日机器翻译现在仍未达到高品质的地步，但是汉日两种语言在本质上天然就具有互译优势。中澤敏明等人的研究使用了日中语言中共同汉字的信息[9]。还有一些简单利用汉字分解、拼音、五笔和部首等汉日语言信息的研究。但在机器翻译中深度融合更多汉日语言外部知识仍是汉日翻译的难题。

对于汉语和日语这两个相近的语言而言，它们的最大的共同之处就是都具有汉字。日本在 1946 年的时候，用内阁公告的方式来推动了繁体字的简化。中国在 20 世纪 60 年代之后进行了简体字的改革，而韩国则进行了废除汉字运动。可以说，汉字的发展贯穿着整个汉语与日语的历史，也是汉语与日语的最主要的共通信息。那么，考虑到通过简繁转换可以将日语和汉语中的汉字进行无损意义上的变换，我们希望在机器翻译中通过数据增强来深度融合汉字简繁转换的信息。本研究致力于通过数据增强来深度融合汉字简繁转换的信息，提高汉日机器翻译的翻译精度。

2. 相关研究

很多研究都在尝试通过改进翻译模型来为低资源语言对提供支持，例如 Mao 等人针对日语提出了一种特定的联合预训练方法来优化翻译模型[10]。Xu 等人则是通过一种动态学习策略对数据进行重新排序以改进翻译效果[11]。Dou 等人提出了一种权重策略来提升回译模型的性能[12]。Araabi 等人通过优化翻译模型，提高了在稀有资源语言上的翻译效果[13]。Ngo 等人则是针对稀有资源语言对中的稀有词汇翻译问题，提出了相应的解决策略[14]。

智能筛选训练数据被证实是一种成功的技术，可以同时提升机器翻译的训练效率和翻译性能。Amittai 等人提出的数据筛选技术在基于短语的机器翻译中表现出色，取得了良好的效果[15]。Marlies 等人探讨了数据筛选在 NMT 上的应用，主要利用动态数据筛选方法，有效地提升了模型的性能[16]。Zhang 等人也对 NMT 的数据选择进行了研究，并在主动学习框架下选择信息量大的源语句子构建平行语料库，尽量降低人工翻译的成本[17]。Wang 等人针对特定领域的翻译进行了数据筛选研究[18]，为 NMT 提出了一种多域平衡方法，以平衡训练数据的领域分布，从而显著提高 NMT 的性能。此外，Song 等人通过在目标语言中替换目标词汇来加强数据，进而提高翻译准确性[19]。李毅鹏为中日双语平行语料库提出了一种采用 XML 标记的方法[20]。

另一方面，有一些研究专注于通过“回翻译”进行语料库的扩展和增强[6]。Caswell 等人在回翻译中提出了一种更简单的添加噪声方法[21]。Khatri 等人将回翻译应用于无监督 NMT 中[22]，Wei 等人提出了一种反向翻译迭代框架[23]，Abdulmumin 等人提出了一种迭代式自训练方法以改善回翻译[24]，Hieu 等人针对回翻译提出了一种专门的元学习框架[25]，均取得了良好的翻译成果。尤丛丛等人提出了一种基于低频词的同义词替换数据增强方法[26]。贾承勋等人研究了基于更大粒度的替换，提出了一种基于短语替换的汉越伪平行句对生成方法[27]。赵志耘等人介绍了中日两国近年来在机器翻译合作项目中的情况，包括合作背景与基础、知识产权、具体合作内容与成果以及对机器翻译实用化方面的一些思考[28]。Zhuang 等人以及 Hagiwara 等人在 IWSLT 2020 open domain translation task 上针对汉日翻译集成了语料数据增强以及选用最新模型等方法，分别建立了相关的汉日神经网络机器翻译系统[29] [30]。

Zhang 等人提出了一种语料数据增强的方法，该方法有两种变化，一种是针对所有语言对，另一种是针对汉日语言对。该方法利用现有平行语料库的源句和目标句，对包含标点符号的长句对进行切割并通过回翻译生成多个伪句对。本方法使用了词对齐信息，用于确定分割点。针对汉日语言对，使用了“相

同汉字率”来修改句对的分段，有效提升了汉日长句分割的精度。提出的方法和过去的只使用反向翻译的方法相比，取得了更显著的翻译效果[31]。

Zhang 等人曾尝试在字级别的 NMT 上加入额外特征“部首”。因为汉语和日语中的汉字无法拆成 Subword，所以基于字符级别选择了汉字的部首当作外部特征来融合到字向量(Embedding Vector)中。翻译结果展示了把部首当作特征能有效提升翻译效能，甚至可以翻译出参考译文没有翻译成功的单词[32]。

对于汉日神经网络机器翻译中的低频词语的翻译问题，Zhang 等人选择利用汉字分解后产生的部首偏旁来进行低频词语和高频词语的转换。通过提出的词典编码方案，可以把高频词语用低频词语加上特定标注的伪词来表示。这样消除了低频词语之后，学习模型变得更加轻量化，最后再转换回低频词语。最终使翻译效果得到了一定的提升[33]。

3. 基于汉字简繁转换的神经机器翻译数据增强

Table 1. Similarities and differences between Japanese Kanji and Chinese simplified characters
表 1. 日语汉字与汉语简体字的异同点

	一致	大部分一致	小部分一致	和制汉字
日语汉字	雪、安	愛、詞	発、広	雫、躾
简体字	雪、安	爱、词	发、广	无

日语中含有大量汉字，但这些日语汉字与汉语汉字(简体字)还有一些不同。如表 1 中“雪”“安”这样的汉字是完全一致的，日语“愛”“詞”和汉语“爱”“词”大部分一致，日语“発”“広”和汉语“发”“广”小部分一致，和制汉字就是日语独创的汉字，汉语中并不包含和制汉字。但和制汉字并不经常使用。

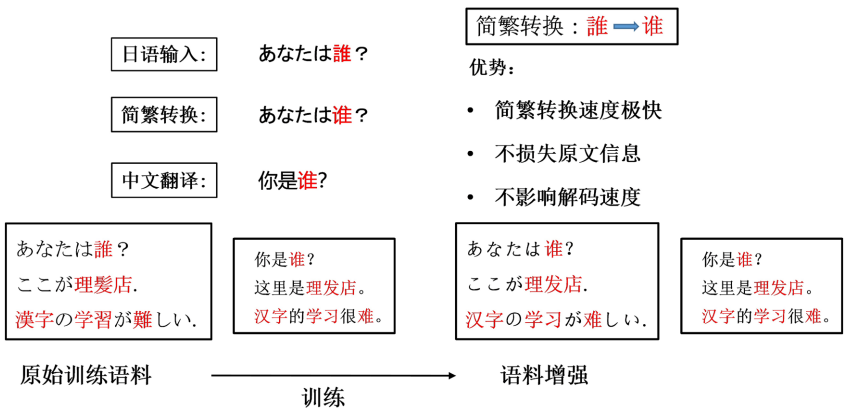

Figure 1. The process of converting traditional Chinese and simplified Chinese
图 1. 融合汉字简繁转换信息的过程

通过汉字简繁转换，我们可以将日语常用汉字和汉语简体字互相转换。那么，为了融合汉字简繁转换信息，可以根据汉字简繁转换表替换源端文字为目标端文字，翻译模型将学会将源端句子中的目标端译文片段复制到译文端，这样，指定简繁转换的原文信息会得到完整保留(以译文的形式出现在源端)。具体步骤如图 1 所示。这种方法中充分利用了汉字简繁转换信息，简繁转换速度极快(Word 软件都可以进行)，而且不损失源端信息，也不影响解码速度。最后，使用和回翻译类似的方法，将增强后的语料与原始语料合并，形成原始语料两倍的新训练语料。

4. 实验与分析

我们在 ASPEC-JC 日中语料库上进行了实验，该语料库是近年来 Workshop on Asian Translation 的日中翻译子任务中共享的语料库。该语料库由手动将日语科学论文翻译成中文的对译语料构建而成[4]。这些日语科学论文属于日本科学技术振兴机构或存储在日本最大的学术期刊电子平台 J-STAGE 上。ASPEC-JC 由 4 个部分组成：训练数据(672,315 个句对)、开发数据(2,090 个句对)、开发测试数据(2,148 个句对)和测试数据(2,107 个句对)。ASPEC-JC 包含论文摘要和部分正文。由于难以包含所有科学领域，ASPEC-JC 仅包括“医学”、“信息”、“生物”、“环境学”、“化学”、“材料”、“农业”和“能源”领域。这些领域是通过调查重要科学领域以及日本研究人员和工程师使用文献数据库的趋势而选择的。

Meng 等人的研究表明汉语的自然语言处理相关任务的最佳选择是使用字符级别与 LSTM 训练模型[34]，因此我们选择使用字符级别与 LSTM 来训练翻译模型。使用的 LSTM 模型有 1 层，每层有 512 个单元，词向量大小为 512。参数在[−0.1, 0.1]范围内初始化，使用 SGD 进行优化，起始学习率为 1.0。Batch size 设定为 10，梯度的范数超过 1 时，将其进行归一化处理。解码时的 Beam Size 为 5。为了计算机器翻译通用评价指标 BLEU 值[35]，我们使用了 Jieba [36]和 Mecab [37]分词工具对输出的汉语和日语句子进行分词处理，最后在词语级别计算 BLEU 值。

在实施所提出的简繁转换方法时，我们采用了 Chu 等人公开发布的汉字简繁对照表[38]。最后，在神经机器翻译框架 OpenNMT 上进行了实验[39]，实验结果如表 2 和表 3 所展示。

Table 2. Japanese → Chinese experiment results
表 2. 日语→汉语实验结果

System	Japanese → Chinese	
	PPL ↓	BLEU ↑
Baseline	3.60	38.71
Double	**3.40**	39.43
Aug	3.41	39.48
Double + Radical	3.44	39.45
Aug + Radical	3.41	**39.65**

Table 3. Chinese → Japanese experiment results
表 3. 汉语→日语实验结果

System	Chinese → Japanese	
	PPL ↓	BLEU ↑
Baseline	2.71	38.58
Double	2.55	40.00
Aug	2.56	39.69
Double + Radical	2.54	39.72
Aug + Radical	**2.54**	**40.16**

表 2 和表 3 中的 Baseline 表示未使用数据增强的字符级别方法；Double 指为与数据增强后的语料库作对比的纯复制方法，使句子对数量增加至 2 倍，与增强后的语料库一致；Aug 是本文提出的简繁转换后的数据增强方法，句子对同样增加至原语料库的 2 倍；Double + Radical 是在纯复制基础上结合额外特征“部首”的方法[32]；Aug + Radical 则是结合本文的方法与额外特征“部首”的方法。PPL (Perplexity) 表示模型复杂度，BLEU 是机器翻译通用评价指标。

从表中的实验结果来看，在日语→汉语和汉语→日语两个翻译方向上，如果不使用加入额外特征“部首”(Radical)的方法，那么使用数据增强(Aug)的方法相较于单纯复制数据(Double)的方法，虽然在汉语→日语的翻译任务中表现略差，但在日语→汉语的任务中表现相近。这说明数据增强方法在一定程度上有助于提高翻译质量。在将额外特征“部首”(Radical)融入方法之后，结合数据增强和部首特征(Aug + Radical)方法在日语→汉语和汉语→日语两个翻译方向上都取得了最佳的翻译品质，相较于基线，分别提高了约 0.9 和 1.6 个 BLEU 值。由此可以得知，结合数据增强和部首特征可以有效提高神经机器翻译模型在日语→汉语和汉语→日语任务中的翻译质量。这两种方法的结合可以互相弥补对方的不足，数据增强方法为模型提供更多学习样本，部首特征有助于模型理解汉字和日本汉字的结构和语义关系，共同提高翻译质量。尽管在日语→汉语的 PPL 方面并未获得最低值，但与最低值的差距仅为 0.01，再结合 BLEU 值评估，证实结合数据增强和部首特征的方法对翻译质量的提升具有重要意义。

5. 结语

在本文中，我们对基于汉字简繁转换的汉日神经机器翻译数据增强研究进行了深入探讨。通过分析不同的数据增强技术和方法，我们发现利用汉字简繁转换可以有效地提高汉日神经机器翻译的性能。我们综合了现有的研究成果，提出了一种新的数据增强策略，通过将简体字与繁体字进行转换，成功增加了语料库的句对数量，提升了最终的翻译结果。

在实验部分，我们展示了所提出的数据增强方法对汉日神经机器翻译的有效性。实验结果表明，该方法在提高翻译准确性方面取得了显著的进展。然而，研究还有很多值得进一步探索的方向。未来，我们将继续研究其他可能的数据增强技术，如结合多模态信息、利用迁移学习等，以提高汉日神经机器翻译的性能。我们希望这项研究为提升汉日翻译的质量提供新的视角，并为未来的相关研究奠定了基础。

基金项目

辽宁省教育厅高等学校基本科研项目(面上青年人才项目) (Grant No.LJKZ0267)，沈阳理工大学引进高层次人才科研支持计划(Grant No.1010147001004)，沈阳理工大学科研创新团队建设计划资助项目(Grant No.SYLUTD202105)。

参考文献

[1] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. *The International Conference on Learning Representations* (*ICLR*), Banff, 14-16 April 2014, 1-15.

[2] Luong, M.T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1412-1421. https://doi.org/10.18653/v1/D15-1166

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 5998-6008.

[4] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H. (2016) ASPEC: Asian Scientific Paper Excerpt Corpus. *Proceedings of the* 10*th International Conference on Language Resources and Evaluation* (*LREC* 2016), Portorož, 23-28 May 2016, 2204-2208.

[5] 徐一平, 曹大峰. 汉日对译语料库的研制与应用研究: 论文集[M]. 北京: 外语教学与研究出版社, 2002.

[6] Zhang, J., Tian, Y., Han, M., Mao, J. and Matsumoto, T. (2022) WCC-JC: A Web-Crawled Corpus for Japanese-Chinese Neural Machine Translation. *Applied Sciences*, **12**, Article No. 6002. https://doi.org/10.3390/app12126002

[7] Zhang, J., Tian, Y., Han, M., Mao, J., Wen, F., Guo, C., Gao, Z. and Matsumoto, T. (2023) WCC-JC 2.0: A Web-Crawled and Manually Aligned Parallel Corpus for Japanese-Chinese Neural Machine Translation. *Electronics*, **12**, Article No. 1140. https://doi.org/10.3390/electronics12051140

[8] Sennrich, R., Haddow, B. and Birch, A. (2016) Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 86-96. https://doi.org/10.18653/v1/P16-1009

[9] 中澤敏明, C. Chu, 黒橋禎夫. 日中共通漢字の整理とこれを利用した日中機械翻訳の高度化[EB/OL]. Japio Year Book: 258-261. https://cir.nii.ac.jp/crid/1523669555917032960, 2023-05-24.

[10] Mao, Z., Cromieres, F., Dabre, R., Song, H. and Kurohashi, S. (2020) JASS: Japanese-Specific Sequence to Sequence Pre-Training for Neural Machine Translation. *LREC* 2020 12*th International Conference on Language Resources and Evaluation*, Marseille, 11-16 May 2020, 3683-3691.

[11] Xu, C., Hu, B., Jiang, Y., Feng, K., Wang, Z., Huang, S. and Zhu, J. (2020) Dynamic Curriculum Learning for Low-Resource Neural Machine Translation. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, 8-13 December 2020, 3977-3989. https://doi.org/10.18653/v1/2020.coling-main.352

[12] Dou, Z.Y., Anastasopoulos, A. and Neubig, G. (2020) Dynamic Data Selection and Weighting for Iterative Back-Translation. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 16-20 November 2020, 5894-5904. https://doi.org/10.18653/v1/2020.emnlp-main.475

[13] Araabi, A. and Monz, C. (2020) Optimizing Transformer for Low-Resource Neural Machine Translation. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, 8-13 December 2020, 3429-3435. https://doi.org/10.18653/v1/2020.coling-main.304

[14] Ngo, T., Nguyen, P., Ha, T., Dinh, K. and Nguyen, L. (2020) Improving Multilingual Neural Machine Translation for Low-Resource Languages: French, English—Vietnamese. *The* 3*rd Workshop on Technologies for MT of Low Resource Languages*, 4-7 December 2020, 55-61.

[15] Amittai, A., He, X. and Gao, J. (2011) Domain Adaptation via Pseudo In-Domain Data Selection. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Association for Computational Linguistics, Cedarville, 355-362.

[16] Marlies, W., Bisazza, A. and Monz, C. (2017) Dynamic Data Selection for Neural Machine Translation. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Copenhagen, 7-11 September 2017, 1400-1410.

[17] Zhang, P., Xu, X. and Xiong, D. (2018) Active Learning for Neural Machine Translation. 2018 *International Conference on Asian Language Processing* (*IALP*), Indonesia, 15-17 November 2018, 153-158. https://doi.org/10.1109/IALP.2018.8629116

[18] Wang, R., Utiyama, M., Finch, A.M., Liu, L., Chen, K. and Sumita, E. (2018) Sentence Selection and Weighting for Neural Machine Translation Domain Adaptation. *IEEE*/*ACM Transactions on Audio*, *Speech*, *and Language Processing*, **26**, 1727-1741. https://doi.org/10.1109/TASLP.2018.2837223

[19] Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K. and Zhang, M. (2019) Code-Switching for Enhancing NMT with Pre-Specified Translation. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 449-459.

[20] 李毅鹏. 中日双语平行语料库之日语科技语标注技术[J]. 企业导报, 2015(2): 175-176.

[21] Caswell, I., Chelba, C. and Grangier, D. (2019) Tagged Back-Translation. *Proceedings of the Fourth Conference on Machine Translation* (*WMT*), Volume 1, 53-63. https://doi.org/10.18653/v1/W19-5206

[22] Khatri, J. and Bhattacharyya, P. (2020) Filtering Back-Translated Data in Unsupervised Neural Machine Translation. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, 8-13 December 2020, 4334-4339. https://doi.org/10.18653/v1/2020.coling-main.383

[23] Wei, H., Zhang, Z., Chen, B. and Luo, W. (2020) Iterative Domain-Repaired Back-Translation. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing*, 16-20 November 2020, 5884-5893. https://doi.org/10.18653/v1/2020.emnlp-main.474

[24] Abdulmumin, I., Galadanci, B.S. and Isa, A. (2021). Enhanced Back-Translation for Low Resource Neural Machine Translation Using Self-training. In: Misra, S. and Muhammad-Bello, B., eds., *ICTA* 2020: *Communications in Computer and Information Science*, vol 1350, Springer, Cham. https://doi.org/10.1007/978-3-030-69143-1_28

[25] Pham, H., Wang, X., Yang, Y. and Neubig, G. (2021) Meta Back-Translation. https://doi.org/10.48550/arXiv.2102.07847

[26] 尤丛丛, 高盛祥, 余正涛, 毛存礼, 潘润海. 基于同义词数据增强的汉越神经机器翻译方法[J]. 计算机工程与科学, 2021, 43(8): 1497-1502.

[27] 贾承勋, 赖华, 余正涛, 文永华, 于志强. 基于短语替换的汉越伪平行句对生成[J]. 中文信息学报, 2021, 35(8): 47-55.

[28] 赵志耘, 石崇德, 何彦青, 高影繁, 姚长青. 面向科技文献的中日机器翻译合作研究[J]. 情报工程, 2017, 3(3): 4-9.

[29] Zhuang, Y., Zhang, Y. and Wang, L. (2020) LIT Team's System Description for Japanese-Chinese Machine Translation Task in IWSLT 2020. *Proceedings of the* 17*th International Conference on Spoken Language Translation*, 9-10 July 2020, 109-113. https://doi.org/10.18653/v1/2020.iwslt-1.12

[30] Hagiwara, M. (2020) Octanove Labs' Japanese-Chinese Open Domain Translation System. *Proceedings of the* 17*th International Conference on Spoken Language Translation*, 9-10 July 2020, 166-171. https://doi.org/10.18653/v1/2020.iwslt-1.20

[31] Zhang, J. and Matsumoto, T. (2019) Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora. *Applied Sciences*, **9**, Article No. 2036. https://doi.org/10.3390/app9102036

[32] Zhang, J. and Matsumoto, T. (2017) Improving Character Level Japanese-Chinese Neural Machine Translation with Radicals as an Additional Input Feature. *Proceedings of the* 2017 *International Conference on Asian Language Processing* (*IALP*), Singapore, 5-7 December 2017, 172-175. https://doi.org/10.1109/IALP.2017.8300572

[33] Zhang, J. and Matsumoto, T. (2019) Character Decomposition for Japanese-Chinese Character-Level Neural Machine Translation. *Proceedings of the* 2019 *International Conference on Asian Language Processing* (*IALP*), Shanghai, 15-17 November 2019, 35-40. https://doi.org/10.1109/IALP48816.2019.9037677

[34] Meng, Y., Li, X., Sun, X., Han, Q., Yuan, A. and Li, J. (2019) Is Word Segmentation Necessary for Deep Learning of Chinese Representations? *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 3242-3252.

[35] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002) Bleu: A Method for Automatic Evaluation of Machine Translation. *Annual Meeting of the Association for Computational Linguistics*, Philadephia, 6-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[36] "结巴"中文分词[EB/OL]. http://github.com/fxsjy/jieba, 2020-01-20.

[37] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab

[38] Chu, C., Nakazawa, T. and Kurohashi, S. (2012) Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese. *Proceedings* 8*th Conference on International Language Resources and Evaluation* (*LREC*'12), Istanbul, 21-27 May 2012, 2149-2152.

[39] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A. (2017) OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL* 2017, *System Demonstrations*, Vancouver, 30 July-4 August 2017, 67-72. https://doi.org/10.18653/v1/P17-4012