新疆大学维汉翻译评测技术报告

宜年[1,2]，艾山·吾买尔*[1,2]，买合木提·买买提[1,2]，吐尔根·依布拉音[1,2]

(1.新疆大学信息科学与工程学院，新疆 乌鲁木齐 830046;

2.新疆多语种信息技术重点实验室，新疆 乌鲁木齐 830046)

摘要:本文主要介绍新疆大学在第十七届全国机器翻译大会机器翻译评测项目中参赛的基本情况。在本次机器翻译评测中，参加了低资源形态丰富黏着语言维吾尔语-汉语新闻领域的机器翻译评测项目。本报告主要介绍参赛的维汉神经网络机器翻译系统以及该系统所采用的方法及该系统在开发集和测试集上的性能。

关键词: 维汉翻译；自注意力机制；低资源翻译

中图分类号: TP302.1 **文献标志码：**A

1. 引言

本文介绍了新疆大学新疆多语种信息技术重点实验室所参加第十七届全国机器翻译大会（CCMT 2021）（China Conference on Machine Translation，简称 CCMT）的维汉机器翻译技术评测的主要情况。本次评测采用的系统为 Google 提出的 Transformer 神经机器翻译模型架构[4]。为了提高模型的效果，得到质量更好的生成数据，本次测评采用数据蒸馏的方式去提升基线模型的翻译性能。除此之外，考虑到本次测评所使用数据量，本文对不同迭代次数的亚词切分进行了实验，以减少数据处理过程中未登录词的覆盖率得到一个性能更好的模型。同时考虑到时间消耗，本次测评对反向翻译生成的数据进行了过滤，得到高质量的生成数据。之后使用模型平均、模型集成以及重排序的方法来进一步提升翻译结果的质量。同时，基于对模型生成结果的分析，本次测评对生成的结果进一步进行了后处理。

在本次 CCMT2021 双语翻译评测任务中，新疆大学自然语言处理团队提交了维汉新闻领域机器翻译的评测系统。该测试报告中详细描述了新疆大学维汉机器翻译系统的网络架构、数据预处理、亚词单元的选择、单语数据的使用、知识蒸馏、微调策略、模型集成等相关技术，并进行对比分析。

2. 系统

2.1 系统介绍

基金项目：国家自然科学基金资助项目(No.61662077)；国家重点研发计划项目（2017YFB1002103）；国家语委重点科研项目(ZDI135-54)；

* **通讯作者:** Email：hasan1479@xju.edu.cn；

本次评测中，使用 Transformer 模型，主要采用了 FACEBOOK 开源的 FairSeq[1]系统的 PyTorch 版本。本次评测提交的系统，使用了反向翻译、数据过滤、知识蒸馏、微调等方法，系统的流程如图所示。

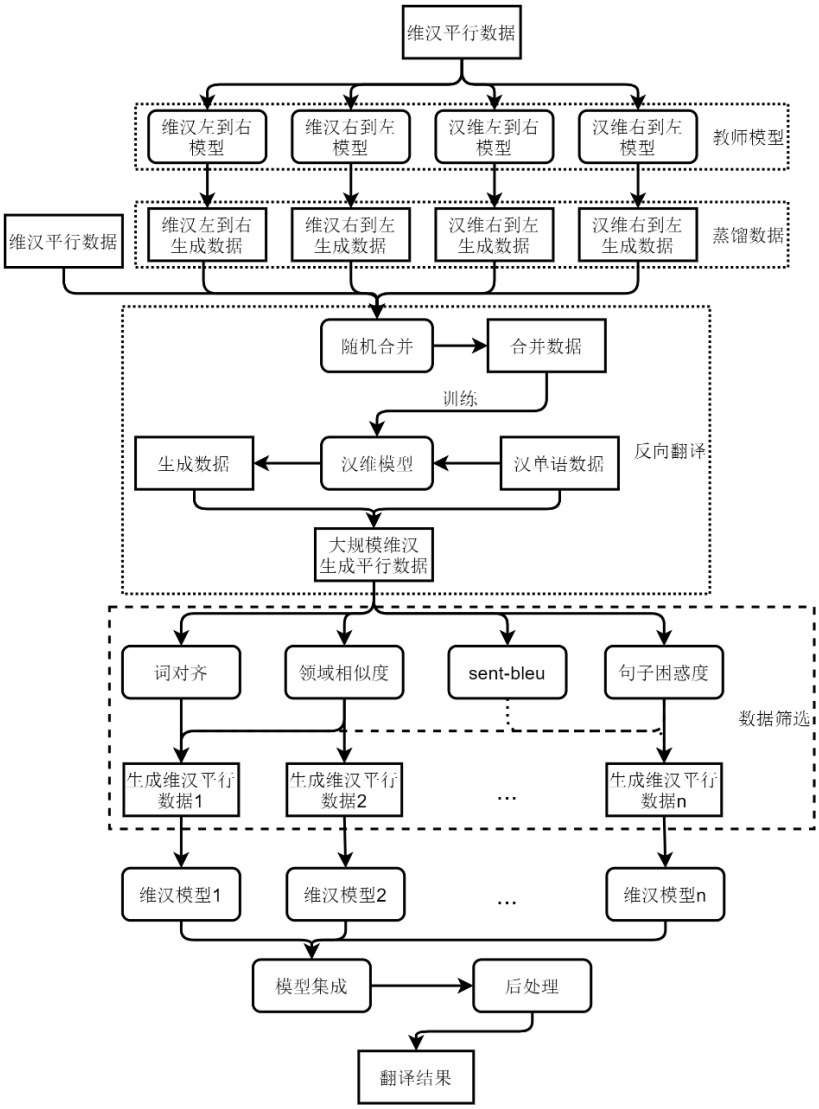

图 1 系统训练流程

Fig.1 Process of training system

2.2 模型结构

神经网络机器翻译(Neural Machine Translation, NMT)[1][2][3]是最近几年提出来的一种机器翻译方法，该方法在多种语言翻译任务对上超过传统的统计机器翻译方法，并逐渐成为目前最主流的机器翻译方法。2017 年 Google 提出一种简单网络架构 Transformer[4]，使用基于自注意力机制，而没有采用传统的循环和卷积神经网络，该模型相比循环神经网络进

[1] https://github.com/pytorch/fairseq.

一步缩短了训练时间，并提升了机器翻译的性能。在本次测评中，我们使用 Transformer 模型实现了维汉机器翻译模型，主要选用 FACEBOOK 团队研发的 FAIRSEQ 开源系统。

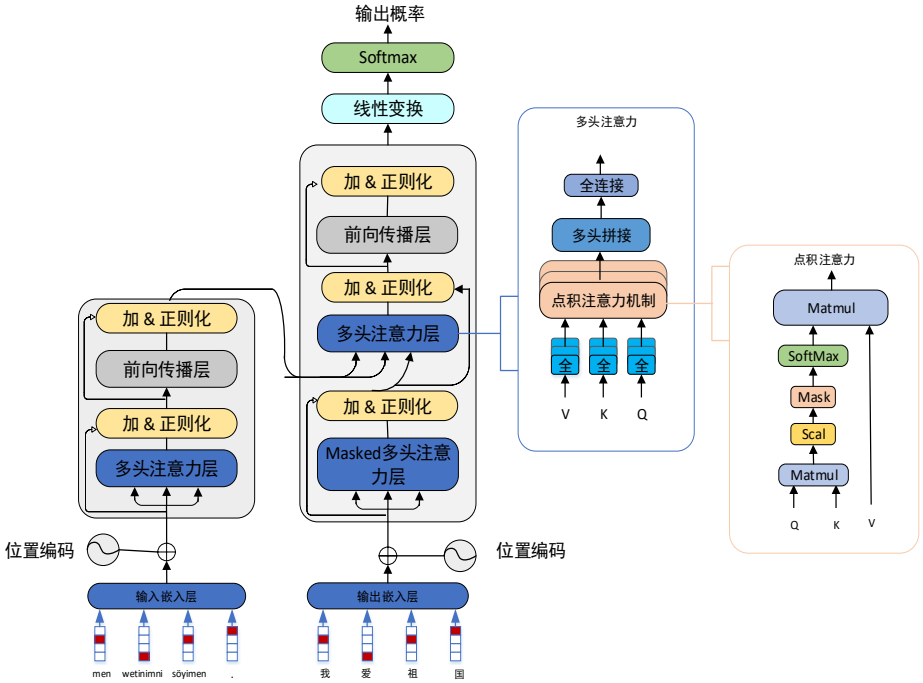

图 2 Transformer 模型架构图

Fig.2 The architecture of Transformer

自注意力机制的神经机器翻译模型 Transformer也是由编码器和解码器组成（如图 2）。编码器由多个子层组成，每个子层由一个多头注意力网络和前馈神经网络组成。同时，在这两个网络之间连接了一个残差连接，并在经过残差之后会对整个输出进行层标准化操作。解码器与编码器相同，也是有多个子层构造。不同的是解码器的子层是由一个遮掩多头注意力网络、多头注意力网络和前馈网络组成，同样在每个网络中间也使用了残差连接及层标准化操作，从而防止因模型结构过于复杂而引起的退化。

自注意力机制的公式如式（1）所示，自注意机制通过 Query-Key 的句子乘积来得到两个向量不同维度的相似度，并利用计算得到的相似度、SoftMax 函数得到对应维度的权重，除以$\sqrt{d_k}$是为了防止相似度结果过大或者过小，利用权重与V的乘积结果，作为当前层的上下文向量。

$$\text{Attention}(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

其中，$\sqrt{d_k}$是缩放因子。常见于加法 Attention 和点积 Attention 中，Q、K、V 为输入向量通过不同线性函数后映射在不同空间的结果。d_k为 Q、K、V 的维度。

多头注意力机制主要通过将输入向量分为维度相同的多个向量，之后会将这些向量输入到多个自注意机制中，最后将自注意力机制输出的结果进行拼接已得到结果向量。整个过程的公式如 2、3 所示。

$$\text{MultiHead}(Q,K,V) = Concat(head_1,\dots,head_n)W^o \quad (2)$$

$$head_1 = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

其中，W_i^Q, W_i^K, W_i^V为第i个头的权重，W^o为拼接之后生成最终上下文的线性映射参数。

给定训练数据$\mathrm{D} = \{\mathrm{X}^{(\mathrm{d})}, \mathrm{Y}^{(\mathrm{d})}\}_1^{|D|}$，其中 X 为源语言，Y 为目标语言，$|D|$为为语料中双语句子的个数。神经机器翻译模型通过优化以下目标函数来得到翻译模型θ。

$$L(\theta, D) = \frac{1}{D}\sum_{n=1}^{|D|}(\log P(Y^n|X^n);\theta) \quad (6)$$

2.3 知识蒸馏与微调

知识蒸馏(Knowledge Distillation, KD)［9］是一种知识迁移的方法，是一种基于“教师-学生网络思想”的训练方法。针对 NMT 的特点，本次评测采用了句子级的知识蒸馏方法，具体的做法如图 1 系统训练流程中教师模型和数据蒸馏阶段。本文通过对句子序列是否为逆向，将句子序列的方向分为 Left2Right(L2R)、Right2Left(R2L)、Right2Right(R2R)，并根据这些情况训练了 6 个 Teacher 模型，根据 Teacher 模型翻译训练数据得到的六个蒸馏文件，与真实数据进行不同程度的合并，训练得到三个学生模型。

2.4 反向翻译与伪造数据选择

[6]提出了反向翻译（Back translation，BT）的方法，该方法通过翻译模型翻译单语言数据来生成大量的数据来提高模型性能的方法。本文利用测评提供的汉语单语言文本进行反向翻译。本文中，使用 CCMT2021 提供的维汉数据，根据知识蒸馏方法得到的汉维翻译模型对全部单语数据进行循环翻译构建{真实汉语句子、伪造维语句子、伪造汉语句子}语料库$\mathrm{D}_{Pseudo} = \{Y, \mathrm{X}_{Pseudo}, \mathrm{Y}_{Pseudo}\}_1^D$。利用语言模型的困惑度、$\mathrm{sim}(\mathrm{Y}, \mathrm{Y}_{Pseudo})$、领域相似度、$Y, \mathrm{X}_{Pseudo}$的词语对齐值等方式对$\mathrm{D}_{Pseudo}$进行排序，然后与真实数据合并进行使用。根据[5]提出了添加<BT>标签的方法，本系统所使用的全部伪造数据均添加<BT>标签。

伪造数据不一定全部是质量高的数据，如何从大量的伪造数据中选出质量高，而且对模型性能提升有帮助的数据是充分挖掘反向翻译的关键问题之一[7]。本评测系统研发中，尝试利用句子困惑度、句子对齐度、句子级 BLEU 值、领域相似度计算等方法来评价数据质量。具体的做法为：

1）句子困惑度：该方法利用 KenLM[1]工具得到维吾尔语的语言模型，通过语言模型对生成的维语句子进行评价，通过排序得到评价值较高的前 3 百万数据。

[1] https://github.com/kpu/kenlm

2）根据 fast_align[1]计算生成的维汉句得到的对齐率来得到前 3 百万数据。

3）利用循环翻译的结果，对真实的汉语句子和生成的汉语句子计算句子 BLEU，已得到句子级 BLEU 较高的 3 百万数据。

4）领域想相似度是基于预训练模型的基础上，2019 年腾讯 CCMT 评测报告[10]中的两层全连接网络的方法计算，模型直接加载预训练语言模型，获得句子向量，将句子向量输入到两层的全连接层中，以实现领域内外作为分类器目标进行建模，具体公式为：

$$P(x) = soft\max(\tanh(W1x + b1)^T W2 + b2)$$

领域内数据为正例，通用领域数据为负例，训练 BERT 领域二分类器，使用 BERT 领域二分类器对伪造数据中的汉语数据使用分类器进行分类，得到分类为正例和负例的概率，其中分类为正例的概率为领域相似度值。

2.5 后处理

在通过模型集成的方式得到数据结果之后，我们对验证集、测试集进行了分析，发现模型对于其中一些数据翻译存在翻译结果中只存在数字等非汉语字符的句子。因此本文对于这种翻译的结果进行了处理。具体流程为：1. 搜索 NMT 翻译结果中未包含汉语字符的句子但是对应的维语句子中有维语字符的结果；2. 将这些维语句子中的非维语数据用一个特殊字符代替，之后将句子重新输入到模型中进行翻译。3. 将非维语的字符替换到翻译的结果中。处理结果前后对比如表 1 所示。由于测试集中只有少部分翻译结果存在该问题，因此处理对于结果的提升并不明显。

表 1 句子与处理前后的翻译结果

Tab.1 The results of sentence before and after processing

维语句子:	66 نى ئىگىلىگەن .	3 ئاشقان .
未处理翻译结果:	66%。	3% 。
非汉语字符替代后翻译结果：	占据 了 num%。	增长 num% 。
非汉语字符还原后结果：	占据 了 66% 。	增长 3% 。

3. 数据

3.1 语料预处理

对于本次维汉评测任务，本文对双语数据和单语数据进行了预处理，主要操作全角字符

[1] https://github.com/clab/fast_align

转半角字符、处理转义字符、控制字符等特殊字符过滤、分词及 token、等，维吾尔语分词使用本课题组自己开发的工具，汉语分词使用清华大学分词系统，过滤掉句子长度比例 3 以上的句子，最终保留维汉句子 164315 条。对汉语单语数据，除了采用以上的预处理方法之外，去掉了“网址，email”或句子中英文或数字的单词比例超过句子的 25%的句子以及长度大于 100 汉语单词的句子，最终保留了 6881603 条句子。

4. 实验

4.1 实验环境

在本次的机器翻译评测中，我们使用的操作系统版本为 CentOS7.2，CPU 为 Intel(R) Xeon(R) CPU E5-2640，内存 251G，显卡 NVIDIA Tesla V100(4 块)，显存 16G。

4.2 模型参数设置

本次评测采用 FaieSeq 系统 Transformer Big Model，每个模型使用 1 块 GPU 核进行训练，每个 batch 大约含有 4096 token，模型训练 60 epoch，每 epoch 保存一次模型用于之后的模型平均。词向量的维度为 1024，隐层状态维度为 4096，编码器与解码器均为 6 层，多头自注意力机制使用 16 个头。本次评测采用了 dropout 机制，dropout 设为 0.3。使用 Adam 梯度优化算法来训练得到最终的模型参数，其中 $\beta 1 = 0.90$，$\beta 2 = 0.98$。初始学习率为 0.001，warmup 步数设定为 4000，beam size=24，BPE[8]迭代次数为 8K。翻译结果使用 Moses 中提供的评价脚本 multi-bleu.perl 计算得到 word 级别的 bleu 值。

4.3 提交结果

对于维汉翻译测评任务，本次提交了 3 个系统的结果，结果的评测指标均为大小写敏感。其实验结果如表 2 所示。其中主系统为使用多个模型集成之后采用后处理的结果，系统 1 则为使用重排序之后的结果，系统 2 则是使用真实数据微调模型得到的结果。

表 2 在开发集和测试集上的实验结果

Tab.2 Experimental results on development valid and test set

	验证集	测试集
主系统	36.15	42.31
系统 1	35.50	40.12
系统 2	34.86	40.29

4.4 数据蒸馏

为了提高系统的适应性，提高仅有少量数据的潜力，采用知识蒸馏的方法开展了实验，本文根据句子序列方向的不同，将 teacher 模型分为以下集中种:L2L、L2R、R2L、R2R。将不同 tearcher 生成的数据合并之后得到数据蒸馏模型的结果如表 3 所示。

表 3 Student 模型的实验结果

Tab.3 Experimental results of the Student model

	维汉		汉维	
	验证集	测试集	验证集	测试集
基线模型	28.95	34.54	24.07	27.21
+数据蒸馏 1	30.15	35.30	25.18	28.36
+数据蒸馏 2	30.72	36.08	25.50	29.02

从表 3，可以看出知识蒸馏对维汉模型和汉维模型的性能有较大幅度的提升。对于不同的数据蒸馏方式，本文添加了 R2R 的 Teacher 模型生成的数据。其中数据蒸馏 1 为采用 L2L、L2R 方式得到的生成数据合并之后的模型结果。数据蒸馏 2 则是在其基础之上添加 R2L，R2R 方式得到的生成数据合并之后的模型结果。实验结果表明，通过添加不同 Teacher 模型的生成结果，可以有效的提升基线模型的结果。这可能与增加了平行数据的数量和数据的分布有关。

4.5 数据增强

表 4 维汉翻译方向上不同数据筛选方式的实验结果

Tab.4 Experimental results of different data filtering methods in Uyghur-Chinese translation

	验证集	测试集
基线模型	28.95	34.54
+数据蒸馏 2	30.72	36.08
+词对齐	34.91	40.60
+句子级 BLEU	34.76	40.50
+领域相似度	34.63	40.81
+1:4:1:4	34.20	39.99
+4:1:4:1	34.34	40.69

为了充分利用评测提供的汉语单语数据，对伪造平行语料库进行数据规模和选择方法方面的实验，使用 10 个模型进行平均得到模型的平均值。本实验中，使用不同的筛选方式对生成数据进行了筛选。其实验结果如表 4 所示。

表 4 中 1:4:1:4 和 4:1:4:1 为语言模型计算伪造句子的困惑度、领域相似度、循环翻译的原句子和最终伪造句子的句子 BLEU 值分数、词对齐分数的不同插值化。从实验结果可以看出，通过循环翻译把汉语句子翻译成维语，然后把伪造维语句子再一次使用机器翻译翻译成汉语，对最初的真实汉语句子和最后的伪造汉语句子进行相似度计算的方法评价伪造维语句子的方法是最有效的数据选择方法之一，其次领域相似度也是相对有一定效果。插值化的方式合并使用词对齐分数、领域相似度等也能接近句子 BLEU 的水平。

4.6 模型集成和微调

表 5 模型微调和集成后的实验结果

Tab.5 Experimental results after model fine-tuning and ensemble

	验证集	测试集
基线模型	28.95	34.54
+数据蒸馏 2	30.72	36.08
+词对齐	34.91	40.60
+真实微调	34.54	40.59
+模型集成	36.01	42.13

本文通过真实的平行数据对数据增强方式得到的模型进行了微调和集成，其实验结果如表 5 所示。从实验结果可以看出，真实的平行数据对于模型进行微调之后，反而模型的结果降低了。而使用模型集成的方法，将多个模型集成起来得到的结果远比单个的模型结果要好，表明不同数据选择方法还是有一定的作用。

4.7 模型的深度与广度

表 6 模型广度和深度实验结果

Tab.6 Experimental results of deeper and wider model

翻译方向	维汉		汉维	
数据集	验证集	测试集	验证集	测试集
基础模型	27.28	32.53	22.55	26.48
深度模型	27.41	33.29	22.92	26.3
广度模型	28.95	34.54	25.18	28.36

对于模型深度和广度的实验结果如表 6 所示，针对 transformer 模型，本文进行了模型上的深度与广度对比。其中基础模型采用 transformer base model 参数。针对广度模型，本文采用 transformer big model 参数。而针对深度模型，本文将 transformer base model 中编码器

和解码器层数由 6 层变为 8 层。实验结果表明，广度模型的结果要好于基础模型和深度模型的结果。

4.8 不同模型结构的集成

本文针对不同模型结构的集成，做了对应的实验，由于时间、设备等问题，实验仅仅开展在 dynamic 模型和基础模型上用于对不同模型架构集成的验证，如果集成方式可以应用在当前这两种模型结构上，则说明该方式同样可以应用在本文的基线模型和 dynamic 模型上。遗憾的是，该集成方式并没有应用在本次测评中。其实验结果如表 7 所示。

表 7 dynamic 模型与 transformer 模型集成实验结果

Tab.7 Ensemble results of dynamic and transformer model

翻译方向	维汉		汉维	
数据集	验证集	测试集	验证集	测试集
基础模型	27.28	32.53	22.55	26.48
dynamic 模型	28.49	34.91	24.68	28.12
模型集成	29.73	35.66	24.84	28.48

5. 总结

在本次评测系统的研究中，对机器翻译系统性能有影响的开源系统、亚词词表大小、反向翻译和数据评价、知识蒸馏和微调等进行比较详细的对比，使用模型平均和集成等有关技术。但是，由于时间、人力、设备等因素，原本计划的句子权重化、双语数据增强方法、迭代式使用伪造数据等工作未能开展。可能经验不足，时间有限等原因，在实验过程中开展的正则化 BPE、Dropout BPE 以及 Morpheme 的亚词化等未取得良好的结果，也没有在 transformer 模型上进行更深层模型的尝试以及在模型集成时没有考虑到不同结构模型之间的集成。今后继续开展以上未完成或者做得不尽人意的工作。在本次评测的准备过程中，对机器翻译系统相关技术的认识进一步加深，对锻炼能力起到了很好的促进作用。这里也感谢评审老师提出的宝贵意见。

参考文献：

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[2] Bahdanau D, Cho K H, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//3rd International Conference on Learning Representations, ICLR 2015. 2015.

[3] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//International Conference on Machine Learning. PMLR, 2017: 1243-1252.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[5] Caswell I, Chelba C, Grangier D. Tagged Back-Translation[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). 2019: 53-63.
[6] Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL), 2016: 86-96.
[7] Edunov S, Ott M, Auli M, et al. Understanding Back-Translation at Scale[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 489-500.
[8] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1715-1725.
[9] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. stat, 2015, 1050: 9.
[10] Hu B, Han A, Zhang Z, et al. Tencent minority-mandarin translation system[C]//China Conference on Machine Translation. Springer, Singapore, 2019: 93-104.

Xinjiang University Uyghur-Chinese Translation Evaluation Technical Report

YI Nian[1,2], AISHAN Wumaier[1,2], Maihemuti Maimaiti[1,2], Turgun Ibrayim[1,2]
（1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046;
2. Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi 830046）

Abstract: This report mainly introduces the basic situation of Xinjiang University in the machine translation evaluation project of the 17th national machine translation conference. In this machine translation evaluation, we participated in the machine translation evaluation project in the field of Uyghur Chinese News with low resources and rich adhesive language. This report mainly introduces the Uyghur Chinese neural network machine translation system, the methods used in the system, and the performance of the system in the development set and test set.

Keyword: Uyghur-to-Chinese Translation; Self Attention Mechanism; Low Resource Translation