昆明理工大学 CCMT2022
机器翻译评测报告

王振晗，叶俊杰，陈蕊，朱志国，高盛祥*，毛存礼

（昆明理工大学 信息工程与自动化学院，云南省人工智能重点实验室，云南 昆明 650500）

摘要：本文详细介绍了昆明理工大学云南省人工智能重点实验室参加 2022 年全国机器翻译(CCMT2022) 评测任务的情况。本次评测我们参加泰语-汉语受限域及非受限域两个测评任务。我们的系统采用基于深度神经网络的 Transformer 模型和基于回译的数据增强方法进行，分别在受限及非受限两种方式下进行模型学习和训练，受限方式即训练数据完全来自于评测方提供的训练数据，非受限方式即在评测方提供的训练数据的基础上，增加实验室收集的双语平行语料，双语词典。

关键词：神经机器翻译，回译，数据增强；

中图分类号：TP 391　　文献标志码：A

1 引言

昆明理工大学云南省人工智能重点实验室从事自然语言处理与机器翻译的研究工作已有 10 多年的研究历史。在 2022 年中国机器翻译研讨会(CCMT2022) 机器翻译评测中，我们提交了两个翻译测评任务，分别是泰-中受限和泰-中非受限系统机器翻译评测任务。其中，非受限域翻译结果是先利用实验室积累的泰语-汉语平行语料训练泰-中模型，再将处理过的汉语单语语料使用回译的方式翻译成泰语，与之前的处理好的汉语单语语料构成伪平行语料，然后经过数据筛选，最后利用伪平行语料训练泰-中模型。在实验中，我们采用 Transformer 神经网络机器翻译架构作为基线翻译模型。下面给出系统描述，数据扩充方法，实验参数设置及实验结果。

2 系统描述

2.1 基于深度 Transformer 的泰-中神经机器翻译

Transformer[1] 系统及其变体遵循标准的编码器-解码器范例。 在编码器端，有许多相同的堆叠层。它们中的每一个都由一个自注意力子层和一个前馈子层组成。 Transformer 中使用的注意力模型是多头注意力[2]，其输出被馈送到一个全连接的前馈网络。同样，解码器还有另一个相同层的堆栈。 除了每个编码器层中使用的两个子层之外，它还有一个编码器-

*通信作者：gaoshengxiang.yn@foxmail.com

解码器注意力子层。一般来说，由于编码器和解码器共享相似的架构，我们可以使用相同的方法来改进它们。在本节中，我们将讨论更一般的情况，不限于编码器或解码器。

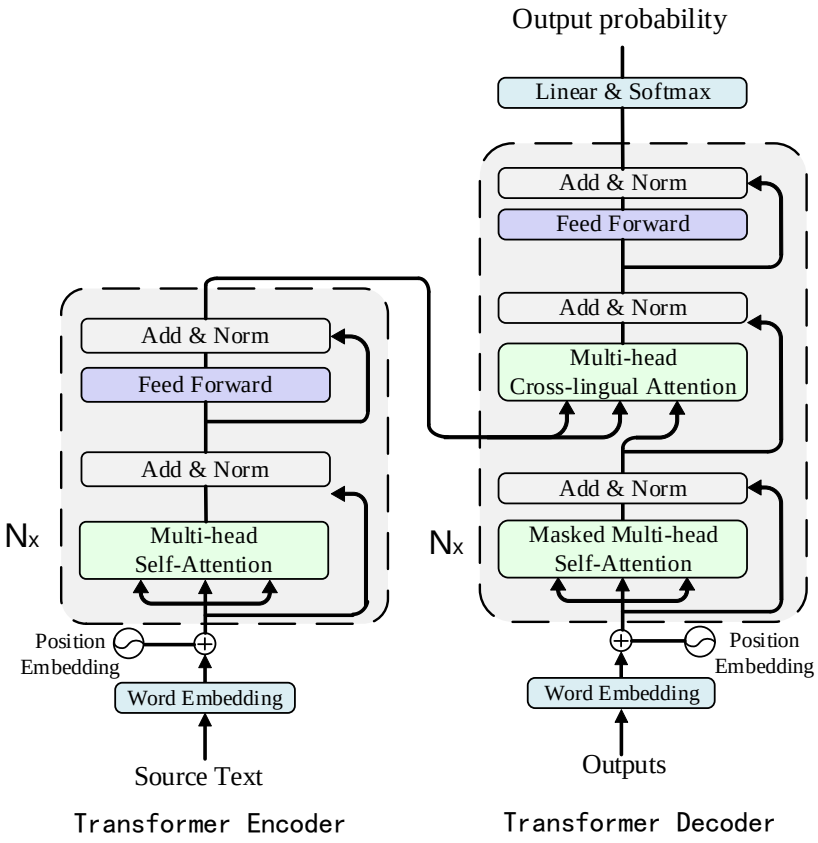

图 1 深度 Transformer 的泰-中神经机器翻译模型

Fig. 1 Deep Transformer's Thai-Chinese neural machine translation model

2.2 模型布局

对于 Transformer 来说，在编码器端和解码器端都不容易训练堆叠的层。这些子层的叠加阻碍了信息在网络中的有效流动，很可能导致训练失败。采用残差连接[3]和层归一化方法求解。设 F 为编码器或解码器中的子层，θ_l为子层的参数。

$$x_{l+1} = f(y_l) \tag{1}$$

$$y_l = x_l + F(x_l\,;\,\theta_l) \tag{2}$$

其中x_l和x_{l+1}为第 l 个子层的输入和输出，y_l为中间输出，后面是后处理函数 f(·)。通过这种方式，x_l显式暴露于y_l。

此外，由于隐藏的状态动态有时会导致收敛训练时间较长，因此采用层归一化来减少子层输出的方差。有两种方法将层归一化纳入残差网络。

在早期版本的 Transformer[1]，层归一化放置在元素的残差连接之后，如式(3)所示，

$$x_{l+1} = LN(x_l + \mathcal{F}(x_l; \theta_l)) \tag{3}$$

式中 LN(·)为层归一化函数，为简单起见，省略了层归一化函数的参数。它可以看作是输出的一个后处理步骤。

在最近的实现中[4-5]，对每个子层的输入进行层归一化处理，将层归一化作为子层的一部分，对残差连接后的数据不做任何处理，如式(4)所示，

$$x_{l+1} = x_l + \mathcal{F}(\mathrm{LN}(x_l); \theta_l) \tag{4}$$

这两种方法都是实现 Transformer 的良好选择。在我们的实验中，对于基于 25 层编码器的系统，它们在 BLEU 中的性能相当。

2.3 层的动态线性组合

残差网络是学习深层网络最常用的方法，在变压器中起着重要作用。原则上，残差网络可以视为常微分方程（ODE）的实例，类似于具有初始值的前向 Euler 离散化[6]。Euler 方法可能是 ODE[9]最流行的一阶解。但这还不够准确。一个可能的原因是，只有前一步用于预测当前值[7]。在机器翻译中，残差网络的单步特性使模型“忘记”遥远的层[8]。因此，如果模型非常深，则无法轻松访问从较低层提取的特征。

在这里，我们描述了一个模型，该模型与之前的所有层直接链接，并提供对深层堆栈中较低级别表示的有效访问。我们称之为层的动态线性组合（DLCL）[17]。该设计受数值常微分方程线性多步法（LMM）的启发。与 Euler 方法不同，LMM 可以通过线性组合有效地重用前面步骤中的信息，以达到更高的阶数。设$\{y_0, \ldots, y_l\}$为 0~l 层的输出, 层 l+1 的输入定义为：

$$x_{l+1} = \mathfrak{b}(y_0, \ldots, y_l) \tag{5}$$

其中，$\mathfrak{b}(.)$是一个线性函数，它将先前生成的值$\{y_0, \ldots, y_l\}$合并为一个新值。对于 Pre-Norm Transformer，我们定义$\mathfrak{b}(.)$为：

$$\mathfrak{b}(y_0, \ldots, y_l) = \sum_{k=0}^{l} W_k^{(l+1)} LN(y_k) \tag{6}$$

其中，$W_k^{(l+1)} \in$是一个可学习的标量，并以线性方式对每个传入层进行加权。等式(6)提供了一种了解堆栈不同级别中的层偏好的方法。即使对于相同的传入层，其对后续层的贡献也可能不同。该方法也适用于后范数变压器模型。对于 Post-Norm，$\mathfrak{b}(.)$可以重新定义为：

$$\mathfrak{b}(y_0, \ldots, y_l) = LN\{\sum_{k=0}^{l} W_k^{(l+1)} y_k\} \tag{7}$$

与 LMM 进行比较。DLCL[17]与 LMM 在两个方面不同，尽管它们的基本模型是相同的。首先，DLCL[17]以端到端的方式学习权重，而不是通过多项式插值等方式确定其值。这提供

了一种更灵活的方法来控制模型行为。其次，DLCL 具有任意大小的过去历史窗口，而 LMM 通常只考虑有限的历史。此外，最近的工作表明 LMM 在计算机视觉中的成功应用，但在类似 LMM[10]的系统中只使用了前两个步骤。

请注意，DLCL[17]是一种非常通用的方法。例如，标准残差网络是 DLCL[17]的特例，我们看到，稠密剩余网络是一个具有统一加权模式的全连通网络[11]。多层表示融合[8]和透明注意方法[12]可以学习加权模型来融合层，但它们仅适用于最顶层。DLCL[17]模型可以涵盖所有这些方法。它提供了在整个堆栈中加权和连接层的方法。我们强调，虽然通过可学习标量对编码器层进行加权的想法类似于透明的注意力，但有两个关键区别：1）我们的方法鼓励在编码过程中早期层之间的交互，而透明的注意力中的编码器层将合并，直到标准编码过程结束；2） 对于编码器层，我们没有学习每个解码器层的唯一权重，而是为每个连续的编码器层创建单独的权重。这样，我们可以在层之间创建更多连接。

3 数据扩充方法

得益于近几年文本翻译领域的显著进展、各种先进翻译模型的开源（包括百度、google 等翻译工具的接口开放），基于回译[13-14]（back translation）方法的文本数据增强成为了质量高又几乎无技术门槛的通用文本增强技术。回译数据增强目前是文本数据增强方面效果较好的增强方法, 将文本数据翻译成另外一种语言(一般选择小语种),之后再翻译回源语言, 即可认为得到与源语料同标签的新语料, 新语料加入到源数据集中即可认为是对源数据集数据增强。当然，很多时候只采用一种中间语种也可以实现很好的增强效果。

回译方法增强数据优势是操作简便获得新语料质量高，但也存在一些问题，主要有：1）在短文本回译数据中, 新语料与源语料很可能存在很高的重复率, 2）并不能有效增大样本的特征空间，语义失真等问题。本文中我们采用回译的方法对泰-中语进行数据增强，进一步提升了泰语机器翻译的性能。

3.1 基于语言模型约束的双语平行语料扩充方法

数据多样化[15]是一种简单有效的数据扩充方法，其避免了开发新架构的开销成本并且增强了现有模型的性能，是一种提高神经机器翻译 (NMT) 性能的简单但有效的策略。它通过使用多个前向和后向模型的预测，然后将它们与训练最终 NMT 模型的原始数据集合并，使训练数据多样化。具体地，我们首先在后向（汉语→泰语）和前向（泰语→汉语）的翻译任务上训练多个模型；然后，我们使用这些模型生成一组不同于原始数据的合成训练数据，，最后把这些合成训练数据和原始数据进行合并，以增强原始数据的多样性，从而提升模型的

性能。

3.2. 基于迭代回译的双语平行语料扩充方法

大规模并行语料库是训练统计机器翻译 (SMT) 和神经机器翻译 (NMT) 系统的重要资源。创建一个高质量的大规模平行语料库需要时间、财力和大量文本的专家翻译。因此，许多现有的大规模并行语料库仅限于特定的语言和领域。相比之下，大型单语语料库更容易获得。现在已经提出了各种方法来从单语语料库创建伪平行语料库，如基于回译的数据扩充方法，回译方法增强数据优势是操作简便获得新语料质量高，但也存在一些问题，主要有：1）在短文本回译数据中, 新语料与源语料很可能存在很高的重复率, 2）并不能有效增大样本的特征空间，语义失真等问题。本文中我们采用迭代回译[16]的方法对泰-中语进行数据增强，提升伪平行语料库的数据质量，进一步提升了泰语机器翻译的性能。具体迭代回译步骤如下：

(1) 使用在汉语到泰语方向上训练的模型，把汉语翻译为合成源语言泰语，构建一个低质量的汉泰伪平行语料库。

(2) 再使用泰语到汉语方向上训练的模型，把伪平行语料库中的泰语翻译为合成目标语言汉语。

(3) 以单语目标语言汉语为参考，合成目标语言汉语为候选，计算二者的句子级相似度。

(4) 将单语目标句子汉语和相应的合成源句子泰语按照句子级相似度度量分数的降序排序，过滤掉分数低的句子，构建高质量的汉泰伪平行语料库。过滤阈值由数据集的翻译质量决定。

(5) 构建的高质量的泰-中伪平行语料库用于模型训练。

4 实验步骤

4.1 实验数据

实验中受限系统采用规模为 20 万句对的泰-中平行句对，非受限系统采用规模为 3000 万句对的泰-中平行句对，测试集和验证集分别包含 10000 和 2000 双语句对，如表 1 所示。

表 1 数据集

Tab.1 Datasets

系统	训练集(句对)	验证集(句对)	测试集(句对)
泰-中受限系统	20 万	2000	10000
泰-中非受限系统	3000 万	2000	10000

4.2 语料筛选

我们通过对模型生成的翻译结果与参考结果进行对比，观察到部分翻译结果存在过翻译的问题，即存在多个重复的无意义字符串。经过数据扩充后的语料还存在着许多的问题，这我们将从下面几个方面对扩充后的语料进行筛选：1）去重，重复的数据会使训练过程有偏。我们先简单的去除扩充数据库里重复的数据，然后计算源文本句和扩充句子之间的局部哈希值，把相似度小于某个阈值的句对都去除，初步过滤掉一些扩充时质量不佳的句子。2）按长度筛选规则，长度太短的句对，对训练没有帮助;长度太长的句对，在送入模型后也会进行截断，所以也没有必要保留。另外，长度比偏离 3 倍标准差的句对，基本上都是有问题的句对，删除不符合标准差的句对。3）我们利用 2 种开源词对齐工具 giza++和 bekerley aligner 来获取所有方向上的语料对齐文件。我们在评测中从对齐文件中提取翻译规则，然后融合在一起构成我们系统的翻译规则。为了保证规则的高效性，我们采用相对熵对规则表进行适当的过滤。4）语言模型筛选:计算语言模型得分，源端和目标端相加，得分太低的句对都可以删除。

4.3 语料预处理

在训练之前首先对双语语料做了人工校对处理，去除语料中存在的重复、空格和不规则符号；然后语料进行分词操作，其中泰语使用了实验室研发的泰语分词模型进行分词，中文使用 jieba 分词进行分词；对于分词之后的语料，因为源句和目标句比例相差过大可能就不平行了，因此需要去掉，我们用了 moses 脚本的 clean 进行过滤，挑选的是长度在 1-80 之间以及长度比在 0.6-1.8 之间的句对，过滤之后受限系统剩余大概 16 万句对，非受限系统剩余大概 2200 万句对。对于过滤后的句子，采用了 BPE 的方法分别将泰语和汉语句子切分成子词，其中 BPE 的操作数设置为 10K，vocabulary-threshold 设置为 1。同时，我们只使用了训练的平行语料作为 BPE 的词频统计语料，因此，验证集和测试集都是根据训练的泰-中平行语料学习到的模型进行切分。最终泰语词表大小为 27964，汉语词表大小为 37549。

4.4 神经机器翻译模型训练

本次评测我们用的是 Facebook 实验室的开源框架 Fairseq 来进行数据预处理，经过数据预处理后构建的词表规模为：受限系统中文词，泰语词；非受限系统中文词，泰语词。

我们采用 Fairseq 内置的深度 Transformer 架构来训练我们的模型，深度 Transformer 采用 25/30 层编码器和解码器，每个 batch 的最大 token 设置为 4096，学习率设置为 0.0002，dropout 设置为 0.2，activation-dropout 以及 attention-dropout 被设置为 0，优化器用的是 Adam

优化器，解码阶段 beam size 设置为 6。所有实验均使用 BLEU 值作为翻译效果的评测指标。

为了体现我们模型的有效性和优异性，我们在受限和非受限的泰-中语数据集上设计了下面这几组实验：Transformer 基准模型，Transformer 25 层编码器模型，Transformer 30 层编码器模型。

4.5 实验结果

以下实验的评估指标均为 BLEU-4，测试集由 ALT 数据集中抽取 2000 句泰语-汉语平行语料进行模型测试，测试结果如表 2 所示。

表 2 泰-中在不同模型上的实验结果对比

Tab.2 Comparison of experimental results of Thai-Chinese on different models

模型	系统	BLEU-4
Transformer Base	受限系统	4.76
	非受限系统	22.96
Deep Transformer-25 Encoder	受限系统	5.54
	非受限系统	23.64
Deep Transformer-30 Encoder	受限系统	5.79
	非受限系统	23.87

受限系统和非受限系统上的实验结果表明，我们的深度 Transformer 模型可以有效的提升基线模型的结果，并且在 BLEU-4 评价指标上获得了大约一个点提升。我们发现，在基线模型的基础上，增加 Transformer 的深度可以有效地提升翻译效果。

为了确定数据增强的有效性，我们在 3 个模型上去除数据增强策略，然后训练泰-中语言模型，实验结果如表 3 所示。

表 3 数据增强策略对翻译性能的影响

Tab.3 Effectiveness of data augmentation strategies on translation performance

模型	系统	BLEU-4
Transformer Base	受限系统	4.16
	非受限系统	22.36
Deep Transformer-25 Encoder	受限系统	5.24
	非受限系统	23.04
Deep Transformer-30 Encoder	受限系统	5.29
	非受限系统	23.07

根据表中的实验结果我们可以发现，BLEU-4 评价指标在没有采用数据增强的模型在上都出现了下降。因此，在数据增强是一种强有效的策略为了提高泰-中机器翻译的性能。

5 结论

在本次评测中我们使用了基于深层 Transformer 的神经机器翻译模型。在泰语-汉语评测

项目中，非受限系统使用实验室收集的约 3000 万规模汉-泰双语平行语料， 通过数据增强方法，扩充语料，从而得到质量较高的泰语-汉语机器翻译模型。

参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Li J, Wang X, Tu Z, et al. On the diversity of multi-head attention[J]. Neurocomputing, 2021, 454: 14-24.

[3] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.

[4] Vaswani A, Bengio S, Brevdo E, et al. Tensor2tensor for neural machine translation[J]. arXiv preprint arXiv:1803.07416, 2018.

[5] Domhan T. How much attention do you need? a granular analysis of neural machine translation architectures[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1799-1808.

[6] Chang B , Meng L , Haber E , et al. Multi-level Residual Networks from Dynamical Systems View[C]// arXiv. arXiv, 2017.

[7] Butcher, J. C . Numerical methods for ordinary differential equations[M]. Clarendon Press, 1976.

[8] Wang Q , Li F , Xiao T , et al. Multi-layer Representation Fusion for Neural Machine Translation[C]// 2020.

[9] UM Ascher, Petzold L R . Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations ||[J]. SIAM Review, 1998, 10.1137/1.9781611971392(2):400-401.

[10] Lu Y . Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations[C]// Thirty-fifth International Conference on Machine Learning (ICML). 2017.

[11] Britz D , Goldie A , Luong M T , et al. Massive Exploration of Neural Machine Translation Architectures[J]. 2017.

[12] Bapna A , Chen M , Firat O , et al. Training Deeper Neural Machine Translation Models with Transparent Attention[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[13] Chen H Y, Boore J R P. Translation and back-translation in qualitative nursing research: methodological review[J]. Journal of clinical nursing, 2010, 19(1-2): 234-239.

[14] Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale[J]. arXiv preprint arXiv:1808.09381, 2018.

[15] Nguyen X P, Joty S, Wu K, et al. Data diversification: A simple strategy for neural machine translation[J]. Advances in Neural Information Processing Systems, 2020, 33: 10018-10029.

[16] Imankulova A, Sato T, Komachi M. Filtered pseudo-parallel corpus improves low-resource neural machine translation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2019, 19(2): 1-16.

[17] Wang Q , Li B , Xiao T , et al. Learning Deep Transformer Models for Machine Translation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

KUST Technical Report of Machine Translation for the 2022 China Conference on Machine Translation

Zhenhan Wang, Junjie Ye, Rui Chen, Zhiguo Zhu, Shengxiang Gao*,

Cunli Mao

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Key Laboratory of Artificial Intelligence of Yunnan Province, Kunming 650500, China)

Abstract: This article details the participation of the Yunnan Provincial Artificial Intelligence Key Laboratory of Kunming University of Science and Technology in the 2022 National Machine Translation (CCMT2022) evaluation task. In this evaluation, we participated in two evaluation tasks of Thai-Chinese restricted domain and unrestricted domain. Our system adopts the Transformer model based on deep neural network and the data enhancement method based on back translation. Model learning and training are carried out in two ways: restricted and unrestricted. The restricted method means that the training data is completely from the evaluation party. The training data provided, in an unrestricted manner, is based on the training data provided by the evaluator, adding bilingual parallel corpora and bilingual dictionaries collected by the laboratory.

Keywords: Thai-Chinese Neural Machine Translation; Back Translation, Data Augmentation;