# 基于正则泛化的中泰机器翻译系统

林楠铠[1]，林晓钿[2]，黄锦荣[2]，蒋盛益[2,3]*

（1. 广东工业大学计算机学院，广东省 广州市，510006；2. 广东外语外贸大学信息科学与技术学院，广东省 广州市，510006；3. 广州市非通用语种智能处理重点实验室，广东省 广州市，510006）

**摘要：**基于 Transformer 的神经机器翻译模型在高资源语言上取得了成功，然而在低资源语种上则模型效果较差。针对中泰机器翻译任务，本文梳理了该任务的研究现状，并尝试将正则泛化技术应用到该任务上，通过不同的 dropout 方式得到不同的概率标签，通过要求不同的输出之间的差异尽可能小，从而提高模型的泛化能力，减少训练和推理中存在的不一致性。实验结果表明，本文采用技术的有效性，该方法可以有效提升低资源场景下的中-泰与泰-中机器翻译效果。

**关键词：**中泰机器翻译，正则泛化，Transformer

## 1. 引言

近年来，随着神经机器翻译模型结构的改进，模型容量和数据规模的增加，高资源语言对之间的翻译效果已经接近人工水平。然而，在低资源语言上，机器翻译模型的效果仍然不是很理想。现有的低资源语言机器翻译主要研究通过数据增强等手段扩充翻译样本[1][2][3]，然而针对模型进行改进以进一步提高模型在低资源翻译任务的表现能力的工作仍然较少。本文从中泰机器翻译出发，梳理了该任务的研究现状，并尝试将正则泛化技术应用到该任务上。本文的主要贡献有：

（1）对中泰机器翻译的相关研究展开了梳理与分类；

（2）从模型的角度考虑，尝试将正则泛化技术应用到低资源语言机器翻译任务；

（3）验证了正则泛化技术可以有效提升低资源场景下的中-泰与泰-中机器翻译效果。

## 2. 相关研究

目前面向中泰机器翻译任务的研究主要分为以下四类：

（1）模型与算法研究

Luekhong[4]等将基于 3-gram 短语的翻译模型、基于 5-gram 短语的翻译模型和基于 3-gram 层次短语的翻译模型应用于中泰翻译与泰中翻译，实验结果表明，基于 3-gram 层次短语的翻译模型在中泰翻译与泰中翻译任务上更具潜力。Li 和 Lai[5]提出了一种基于双向依赖自注意力机制的依赖句法知识融合方法，实验结果表明，双向依赖的 self-attention 机制为模型提供了更丰富的依赖信息，可以有效提高翻译性能。李自荐[6]将双向长短期记忆神经网络模型应用于分句技术，提出了一种基于 Glove+Bi-LSTM+CRF 架构的泰语句子切分模型，使用该模型能够成功实现对泰语句子的精确切分，对于 NMT 模型翻译精度与训练执行周期均能够带来较大的提升。

（2）资源构建研究

Lin[7]等设计与实现了汉泰英互译有声电子词典，主要功能包括泰语查询、泰中双语翻译、泰人朗读、英文显示等常用功能。还支持词库添加、修改、删除自定义动作，实现了良好的人机交互功能。Chen 和 Kongjit[8]通过文本分析收集和过滤中国游客的评论，对互联网上的泰式菜单翻译进行分类，根据中泰语序规则和术语的不同，分析并制定更准确的菜名翻译规则，建立一个有利于餐饮行业标准化的泰国食品翻译框架。Zhao 和 Guo[9]开发了一个中文-泰语电子词典软件。词典查询软件由即时翻译模块和即时词典模块组成，它还提供了一个维护词库的接口。刘峰[10]设计并实现了在 Android 平台下的汉-英-泰互译有声电子词典软件，系统由泰语语料库创建本地词库，具有对话翻译、拍照翻译的特色, 还实现了汉-英-泰三语查询互译、泰语真人朗读等功能。

（3）评价方法研究

为了解决泰语、越南语的语言特性导致的评估问题，赖华[11]等提出一种基于子词、音节、词组等多粒度特征的文本生成评价方法，在机器翻译等任务上的实验结果表明，该文提出的多粒度特征评价方法相比 ROUGE[12]、BLEU[13]等基于统计的评价方法都取得了更好的性能，与人工评价结果相关性更高。马文倩[14]采用预先对参考译文进行近义词分析提取的方法，扩展生成多参考译文，尽可能覆盖多种翻译的表达。同时设置阈值对提取出的近义词进行筛选，排除低质量近义词，避免因为扩展带来参考译文质量下降，再基于该扩充译文完成非通用语的质量评估。实验结果表明：在面对非通用语等参考语料不足的质量评估时，采用该方法可有效地提高评估的准确度，降低误判率。

（4）众包机制研究

邹一军[15]在对众包结果筛选和质量控制方法研究的前提下，以中泰翻译语料作为研究对象，设计了一个面向机器翻译质量改进的众包质量管理和结果筛选机制。原博洋[16]以泰语与汉语互译为例，基于工作者特征的众包翻译工作者筛选模型构建方法，并进而设计完成一个基于协同计算的众包翻译系统模型和框架。

# 3. 方法

## 3.1 Transformer

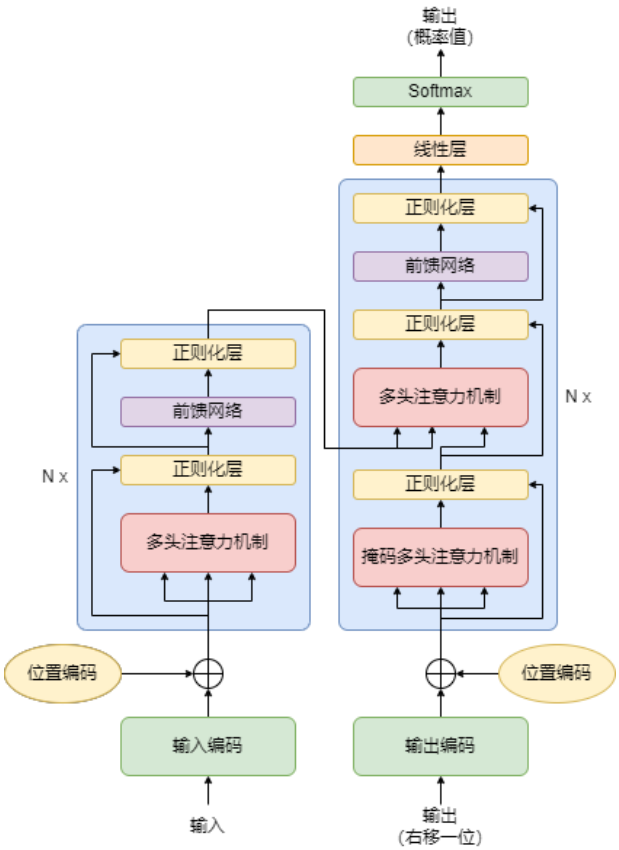


图 1 Transformer 框架图

Fig. 1 Transformer frame diagram

Transformer[17]是一个序列生成模型，其内部由多头注意力机制(Multi-head attention mechanism)组成，包括编码器(Encoder)模块和解码器(Decoder)模块，如图 1 所示。在编码器模块中，编码器将输入序列映射到高维隐式语义表示。在解码器模块中，解码器根据编码器模块输出的高维隐式语义表示，对它们进行解码，得到当前步骤的输出向量，每次生成的输出向量为当前 token 的表示向量。此外，该模型在生成下一个输出的表示向量时会使用先前生成的表示作为附加输入，将所有时刻的

模型输出结果拼接后得到最终的模型输出句子。

**编码器** 编码器由 N 个相同的编码层组成，前一编码层的输出为下一编码层的输入。对于每一个编码层都包含两个模块：一个多头注意力机制和一个由全连接组成的前馈网络。其中，多头注意力机制包含有多个注意力层，注意力层内部采用缩放式点积的方法计算，该层的公式如下所示：

$$Att(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$$Multi-Head = [Att_1; Att_2; \dots; Att_n] \tag{2}$$

其中,$Q$、$K$和$V$分别表示注意力层的三个参数矩阵,它们通过各自独立的线性层从输入向量映射得到。$d_k$指的是查询矩阵和键值矩阵的维度，即调整$Q$和$K^T$的点积大小，避免执行 Softmax 函数过程中向量分布存在较大偏差。$Multi-Head$为$n$个注意力层的堆叠。此外，为了进一步提取与表示句子中的语义信息，前馈层将注意力层的输出进行进一步特征转换，前馈层包含有两个线性层，两个线性层具有独立的参数，前馈层的公式如下所示：

$$FF = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{3}$$

$W_1$，$b_1$，$W_2$，$b_2$分别是两个线性层的权重和偏差。

**解码器** 解码器同样由 N 个完全相同的解码模块组成。除了与编码器中相同的多头注意力层与前馈层，解码器还有第三个模块，即编码-解码多头注意力层，用于计算编码器和解码器两个模块中的向量之间的注意力。与编码器中的多头注意力层不同，在编码-解码多头注意力层中，$Q$和$K$指的是编码器输出的向量，而$V$是在编码器中的掩码多头注意力层的输出。

## 3.2 正则泛化

本文将正则泛化策略 Regularized Dropout[18]应用到中泰机器翻译任务上，以提高模型的语法泛化能力。形式上，给定一个源语言句子S，它由一系列的单词$\{w_1, w_2, w_3, \dots, w_n\}$组成，目标语言$T$同样由一系列目标语言的单词$\{y_1, y_2, y_3, \dots, y_z\}$ 组成，每个词语$w_i$在通过基于 Transformer 的模型时随机丢弃一些隐藏单元，前向传递两次以获得两个不同的标签概率分布$P_1(y_j|w_i)$和$P_2(y_j|w_i)$。然后，为了迫使两个分布$P_1(y_j|w_i)$和$P_2(y_j|w_i)$彼此一致，在使用交叉熵作为损失函数的基础上，最小化两个分布之间的双向 KL 散度。

$$L_{KL}^i = \frac{1}{2}D_{KL}\left(P_1(y_j|w_i)\middle|P_2(y_j|w_i)\right) + \frac{1}{2}D_{KL}\left(P_2(y_j|w_i)\middle|P_1(y_j|w_i)\right) \tag{4}$$

$$L_{CE1}^i = -y_j \log\left(P_1(y_j|w_i)\right) \tag{5}$$

$$L_{CE2}^i = -y_j \log\left(P_2(y_j|w_i)\right) \tag{6}$$

其中$P_1(y_j|w_i)$和$P_2(y_j|w_i)$是词语$w_i$的输出值概率分布，由 Transformer 模型通过两次前向传播获得。在从方程 4-1 中得到 KL 散度后，本文进一步加权 KL 散度和交叉熵损失$L_{CE1}^i$和$L_{CE2}^i$以获得句子$S$中词语$w_i$的最终损失$L^i$：

$$L^i = \alpha * L_{KL}^i + \beta * L_{CE1}^i + \beta * L_{CE2}^i \tag{7}$$

其中$\alpha$和$\beta$是损失权重，$\beta$默认设置为 0.5。将损失权重归一化后，模型的损失值为：

$$L^i = \alpha' * L_{KL}^i + \beta' * L_{CE1}^i + \beta' * L_{CE2}^i \tag{8}$$

$$\alpha' + 2 * \beta' = 1 \tag{9}$$

最终，整个数据集的损失函数计算如下：

$$L = \frac{1}{N}\sum_{n=0}^{N}\frac{1}{M}\sum_{m=0}^{M}L_n^m \tag{10}$$

其中$N$表示数据集中的句子数量，$M$表示句子中的词语数量，$L_n^m$表示第$n$个句子的第 m 个词语的损失。

# 4. 实验

## 4.1 实验数据

本文采用 CCMT 2022 中泰翻译的评测数据进行实验，该数据共包含 20 万个中泰句对，本文将该数据随机划分了 5000 个句对作为线下测试集，剩余的 195000 个句对作为训练集。

## 4.2 实验设置

本文基于 fairseq[1]进行实验。编码器词嵌入矩阵和编码器词嵌入矩阵的词表维度为 512 维，编码器与解码器均包含 3 个编码模块或解码模块，每个编码模块或解码模块中的多头注意力层由 8 个注意力头组成, 全连接层节点数为 2048，标签平滑权重为 0.1。本文训练模型时使用的学习率为 0.0001, dropout 设置为 0.3，一共训练 50 个 epoch。在解码过程中，本文设置 beam search 大小为 5，并取概率最大的结果作为模型最终的输出句子。在预处理阶段，中文文本采用 Jieba 工具[2]进行分词，泰语文本采用 Polyglot[3]进行分词。实验过程中，本文对超参数$\alpha'$与$\beta'$进行了网格搜索，超参数$\alpha'$与$\beta'$的最佳值均为$\frac{1}{3}$。

## 4.3 实验结果

如表 1 所示，本文探究了不同层数的 Transformer 模型的性能，其中 3 层的 Transformer 表现最佳。基于 3 层的 Transformer 模型，本文进一步验证了所采用的正则泛化策略的性能，此外，本文还采用了低资源语言下具有竞争力的 MASS 模型[19]与本文的模型进行对比，实验结果表明，在受限语料的情况下，本文的模型性能最佳，在中-泰翻译上取得了 30.00 的 BLEU 指标，在泰-中翻译上取得了 22.45 的 BLEU 指标。

表 1 主要实验结果

Tab. 1 Main experimental results.

| 模型 | 任务 | BLEU |
|---|---|---|
| Transformer(2-Layer) | 中-泰 | 25.13 |
| | 泰-中 | 16.57 |
| Transformer(3-Layer) | 中-泰 | 28.58 |
| | 泰-中 | 21.45 |
| Transformer(3-Layer) with R-drop | 中-泰 | **30.00** |
| | 泰-中 | **22.45** |
| MASS | 中-泰 | 25.69 |
| | 泰-中 | 18.45 |

此外，我们还对 Transformer(3-Layer) with R-drop 与 Transformer 在训练过程中的性能进行了动态对比，结果如图 2 和图 3 所示，可以看到，Transformer(3-Layer) with R-drop 与 Transformer 的训练趋势大致相同，但是在同一训练时刻的模型的 BLEU 指标更高。

[1] https://github.com/pytorch/fairseq
[2] https://github.com/fxsjy/jieba
[3] https://github.com/aboSamoor/polyglot

本文抽取了线上测试集的结果进行了样例分析（如表 2 所示），将本文提出的系统与现有的机器翻译系统谷歌翻译[4]进行对比，经语言学专家的人工评估，谷歌翻译系统的结果句子不完整，有错译，但用词更好，而本文提出的系统在翻译结果的流畅性与完整度更高，断句基本无误，但仍然存在部分词语不够准确的问题。

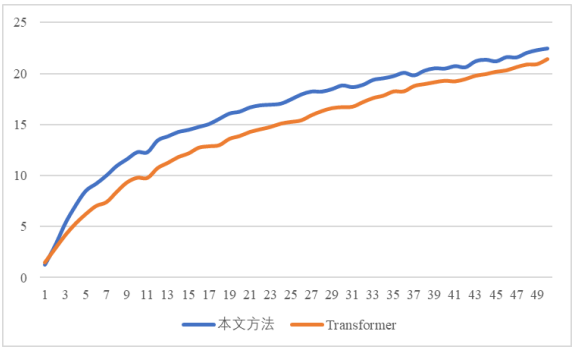


图 2 中-泰机器翻译训练过程

Fig. 2 Chinese-Thai machine translation training process

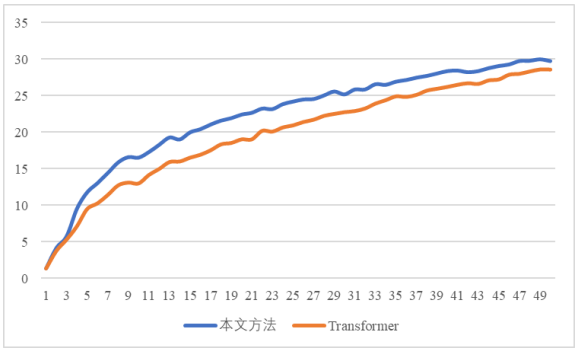


图 3 泰-中机器翻译训练过程

Fig. 3 Thai-Chinese machine translation training process

表 2 样例分析

Tab. 2 Sample Analysis

| | |
|---|---|
| 样例 1 | |
| 原句子 | ไป๋หยุนเฟยถึงกับตะลึงงัน<br>ก่อนจะมองไปที่สุนัขป่าอสนีบาตอย่างสับสนมันมองสุนัขป่าสีเทาขนาดเท่าโคที่ยืนนิ่งตรงหน้าหงยินราวรูปสลักแม้ดวงตาสุนัขป่าตรงหน้าจะเป็นสีแดงฉานแต่ก็ปราศจากชีวิตชีวาหากไม่เห็นว่าตัวของมันยังสั่นระริกอยู่เล็กน้อยทั้งยังมีลมหายใจ<br>ไป๋หยุนเฟยคงสงสัยแล้วว่ามันยังมีชีวิตอยู่หรือไม่ |
| 谷歌翻译结果 | 白云飞愣住了。不解的看向雷狼之前，它又看了看像雕塑一样站在自己面前的牛大小的灰狼，虽然眼前的狼的眼睛是鲜红的，但如果他不看，那就是没有生命了。没看见。它还在微微颤抖和呼吸。白云飞肯定怀疑它是不是还活着。 |
| 本系统翻译结果 | 白云飞愣了愣，不解地看了一眼那雷狼，只见那只灰溜溜的灰色巨狼，就像雕刻在弘胤身前的一样大小，眼前这只狼目虽然是一片泛红，但是却没有一丝生机，要不是他的呼吸还在微微颤抖，白云飞甚至怀疑它到底是否还活着。 |
| 样例 2 | |
| 原句子 | “ ฉัน ไม่สามารถทำได้ ...” หยาง หลิง สายหัว ” มีคน เพียงไม่กี่คน ที่ได้ เข้ารับการ ฝึก เร่งรัด หรือ กลายเป็น ผู้บริหาร ระดับสูง ในเขต ระบบ ของสมาคม นักล่า จะดำเนินการ ตรวจสอบ ใน ทุกๆปี นาย จะถูก เปิดโปง โดย ง่าย ฉัน ขอแนะนำให้ นาย ใช้ ตัวตน ปลอม หลังจากที่ นาย ย้าย เมือง ” |
| 谷歌翻译结果 | “我不可能……”杨凌摇头,“每个获得的人,都是重点培养或者是重点培养的大执事,在猎人协会的系统中，每年都会进行检测，你可以轻易的暴露。我建议,你搬家之后，也不要轻易暴露身份。” |
| 本系统翻译结果 | “我不能……”杨凌摇头，“在掠夺者公会的系统域里，少数经过强化训练或者成为高级管理人员的人，都会在……中进行调查。每年你都很容易暴露。我建议你搬家后使用假身份。” |

[4] https://translate.google.cn/

## 5. 总结

针对中泰机器翻译任务，本文梳理了该任务的研究现状，并尝试将正则泛化技术应用到该任务上，实验结果表明，本文采用的技术的有效性，该方法可以有效提升低资源场景下的中-泰与泰-中机器翻译效果，未来我们将构建中泰相关的语言资源，以进一步提高中泰机器翻译任务的性能。

## 参考文献：


[1] 吴章淋,魏代猛,李宗耀,於正哲,商恒超,陈潇雨,郭嘉鑫,王明涵,雷立志,陶士敏,杨浩,秦璎. 面向神经机器翻译的正向翻译与反向翻译相结合的改进方法[J].厦门大学学报(自然科学版): 1-8.

[2] 刘欢,刘俊鹏,黄锴宇,黄德根. 面向低资源俄汉机器翻译的领域适应方法[J].厦门大学学报(自然科学版): 1-7.

[3] 何乌云,秀芝,包晶晶,陈美兰,王斯日古楞. 基于词切分的蒙汉神经机器翻译中 BERT 数据增强方法[J]. 厦门大学学报(自然科学版): 1-9.

[4] Luekhong P, Sukhauta R, Porkaew P, et al. A Comparative Study on Applying Hierarchical Phrase-based and Phrase-based on Thai-Chinese Translation[C]//Lee V C, Ong K L. International Conference on Knowledge, Information and Creativity Support Systems. Melbourne: IEEE Computer Society, 2012: 126-133.

[5] LI Y J, LAI H, WEN Y H, et al. Neural machine translation integrating bidirectional-dependency self-attention mechanism[J]. Journal of Computer Applications, 2022: 1-8.

[6] 李自荐. 面向机器翻译的数据处理关键技术研究[D].鞍山市：辽宁科技大学,2020: 1-74.

[7] Lin R, Wang J M, Li B Z, et al. Chinese-Thai-English Translation Audible Electronic Dictionary Design and Implementation[C]//Zhu S H. Proceedings of the 2016 4th International Conference on Mechanical Materials and Manufacturing Engineering. Netherlands: Atlantis Press, 2016: 113-119.

[8] Chen Z, Kongjit C. Knowledge Translation Framework for Translating the Name of Thai Dishes from Thai into the Chinese Language[J]. Multicultural Education, 2021, 7(7): 1-9.

[9] Zhao J, Guo H, Zheng Z, et al. The Implementation of Chinese-Tai Lue Electronic Dictionary Based on C#[C]//Tan H. International Conference on Machine Vision and Human-Machine Interface. Kaifeng: IEEE Computer Society, 2010: 300-303.

[10] 刘峰,王嘉梅,林睿等. 基于 Android 平台的汉-英-泰互译有声电子词典[J]. 计算机系统应用, 2018, 27(4): 54-62.

[11] 赖华,高玉梦,黄于欣等. 基于多粒度特征的文本生成评价方法[J]. 中文信息学报, 2022, 36(3): 45-53.

[12] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.

[13] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002; 311–318.

[14] 马文倩,王丽清,王娟等. 基于近义词扩充的非通用语翻译评估[J]. 计算机技术与发展, 2021, 31(8): 125-128.

[15] 邹一军. 面向机器翻译的众包质量管理机制研究[D]. 昆明: 云南大学, 2017: 1-46.

[16] 原博洋. 基于工作者特征模型的众包翻译任务分配方法研究[D]. 昆明: 云南大学, 2019: 1-65.

[17] Vaswani, A, Shazeer, N, Parmar, N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017.

[18] Wu L, Li J, Wang Y et al. R-drop: regularized dropout for neural networks[C]//Advances in Neural Information Processing Systems. 2021: 10890-10905.

[19] Song K, Tan X, Qin T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.


# Chinese-Thai Machine Translation System Based on Regularized Generalization

**Abstract:** Transformer-based neural machine translation models have achieved success in high-resource languages, but the model is less effective in low-resource languages. For the Chinese-Thai machine translation task, this paper sorts out the research status of this task, and tries to apply the regularized generalization technology to this task. Different dropout methods are used to obtain different output probabilities. The difference is as small as possible, thereby improving the generalization ability of the model and reducing the inconsistency in training and inference. The experimental results show the effectiveness of the technology adopted in this paper, and the method can effectively improve the effect of Chinese-Thai and Thai-Chinese machine translation in low-resource scenarios.
**Keywords:** Chinese-Thai Machine Translation, Regularized Generalization, Transformer