基于混合模型的机器翻译优化方法

张 乐[1] 荆晓远[2] 孙其航[1]
ZHANG Le JING Xiaoyuan SUN Qihang

摘 要 预训练语言模型通过学习大量文本数据中的语言模式和结构，展现出对自然语言任务的通用处理能力。通常，为了机器翻译任务而专门训练预训练模型，需要耗费大量的计算资源。为了解决这一问题，文章提出了一种新的混合模型，该模型将 BERT、ROBERTA、ELECTRA 和 LUKE 等预训练模型与 Marian 神经机器翻译模型相结合。该方法的目的是利用其他领域的预训练模型与机器翻译模型相结合，以低成本构建新的预训练机器翻译模型。实验结果表明，混合模型在多个语言数据集上比直接使用 Marian 模型进行微调 BLEU 值平均提升 5.53，最高提升了 16.05。结果显示，结合预训练模型和 Marian，以低成本构建混合模型，可以有效提升机器翻译性能。

关键词 自然语言处理；预训练模型；混合模型；机器翻译

doi：10.3969/j.issn.1672-9528.2025.02.011

0 引言

机器翻译要求理解和转换不同语言之间的复杂语义和信息。随着自然语言处理技术的进步，特别是预训练语言模型的出现 [1]，机器翻译技术取得了突破性进展。

机器翻译技术历经从基于规则的机器翻译到统计机器翻译，再到神经机器翻译 [2] 的演变。随着机器翻译范式不断转变，机器翻译效果不断提升。目前神经机器翻译被广泛应用，例如 Marian NMT 因果语言模型，采用编码器 - 解码器架构，通过编码器将源语言映射为连续的向量表示，然后通过解码器生成目标语言的翻译结果。然而，尽管 Marian NMT 模型在机器翻译任务上取得了较好的性能，但在处理长句子和低频词汇翻译等方面仍存在一定的问题。为了解决这些问题，Transformer 语言模型的发展和出现在捕捉复杂的语言细微差别方面表现出了卓越的能力。因此，在机器翻译的发展历程中，预训练语言模型的出现标志着一种新的转变。

训练一个新的预训练模型需要巨大开销，特别是针对每一对翻译语言重新训练预训练模型，这无疑加剧了计算资源的消耗。本文利用大语言模型在处理自然语言时具有较好的通用性，提出通过融合预训练语言模型如 BERT[3]、ROBERTA[4]、ELECTRA[5] 和 LUKE 等，基本架构为 Transformer 的预训练语言模型 [6] 结合 Marian 机器翻译模型组成新的机器翻译混合模型。

本文的主要贡献有：（1）提出了一种新的混合机器翻译模型，通过融合预训练的语言模型和 Marian 神经机器翻译模型，实现低成本构建新的机器翻译模型；（2）该混合模型在 Wmt14 英 - 译德任务上取得了显著的性能提升，BLEU 得分最高达到了 30.14；（3）在多语言数据集上，混合模型相较于直接使用原本的 Marian 模型进行微调，BLEU 得分平均提升 5.53，最高提升了 16.05。实验表明，这种结合预训练语言模型和 Marian 模型的方法可以以较低的成本构建新的预训练机器翻译混合模型，并且能够有效提升机器翻译任务的性能。

1 机器翻译混合模型设计

机器翻译是一种序列到序列的任务，对源语言 $x=(x_1,x_2,\cdots,x_n)$ 到目标语言 $y=(y_1,y_2,\cdots,y_m)$ 的映射关系进行建模。当前主流的神经机器翻译模型架构图如图 1 所示 [7]。

源语言单词首先通过编码器转换为隐藏表示 Z，其公式为：

$$Z=\text{Encoder}(x)=(z_1,z_2,...,z_n) \quad (1)$$

接着，解码器则利用 Z 和此前已经生成的目标单词序列。$y_{<t}=(y_1,y_2,\cdots,y_{t-1})$ 作为输入，生成第 t 个目标词的表示 h_t，其公式为：

$$h_t=\text{Decoder}(Z,y_{<t}) \quad (2)$$

1. 吉林化工学院信息与控制工程学院 吉林吉林 132022
2. 广东石油化工学院计算机学院 广东茂名 525000

[基金项目] 国家自然科学基金项目“基于类不平衡深度特征学习的石化动设备故障信号分类研究”（62176069）

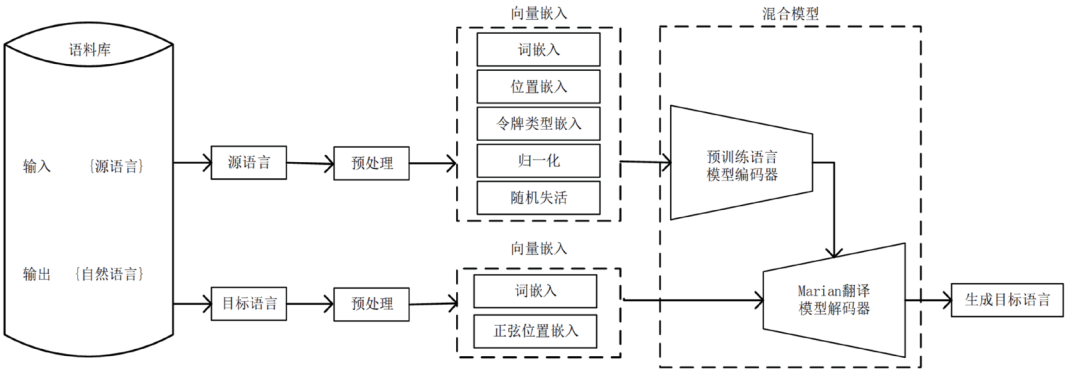

图 1 机器翻译混合模型模型架构图

随后，使用线性投影将 h_t 映射到目标词汇表的维度 $|V|$，并使用 Softmax 函数计算 t 时刻生成词 y_t 的概率分布，其公式为：

$$p_q(y_t \mid Z, y_{<t}; q) = \text{soft max}(\text{Linear}(h_t)) \quad (3)$$

最终，取概率值最大的词为 t 时刻预测的目标词 y_t。整个句子的概率则表示为目标语言各个单词的联合概率分布，其公式为：

$$p_q(y \mid x) = \sum_{t=1}^{m} p_q(y_t \mid Z, y < t; q) \quad (4)$$

为了构建这种机器翻译混合模型，选用了 4 种不同的预训练模型：luke-base、distilroberta-base、distilbert-base-uncased 和 electra-small-discriminator。这些模型的编码器均是基于 transform 架构的编码器解码器模型，其能够有效处理句子中的长距离依赖关系，尤其在较长的句子中有优秀的表现。将 4 种模型的编码器分别与 Marian 模型的解码器相结合，生成了 4 个不同的 Seq2Seq 混合模型。

这些混合模型的原理是基于 Transformer 架构，在神经机器翻译中，将每一个输入的词元使用词嵌入算法转换为词向量，词嵌入过程只发生在最底层的编码器中。输入序列进行词嵌入之后，源语言中每一个词元都会流经编码器中的多头自注意力层 [8] 和前馈网络层。在 Transformer 中的多头自注意力子层使用“自注意力”机制，这样便可以充分地表示词元与词元之间联系的密切程度。自注意力机制允许模型根据输入序列的内在相关性来动态分配权重，其计算公式为：

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{soft max}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\boldsymbol{d}_k}})\boldsymbol{V} \quad (5)$$

式中：$\boldsymbol{Q}$ 表示查询向量；$\boldsymbol{K}$ 表示链向量；$\boldsymbol{V}$ 表示值向量；d 表示 Transformer 的维度，其中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 计算公式为：

$$\boldsymbol{Q}_h = \boldsymbol{W}_h^{\boldsymbol{Q}} \cdot \boldsymbol{X} \quad \boldsymbol{K}_h = \boldsymbol{W}_h^{\boldsymbol{K}} \cdot \boldsymbol{X} \quad \boldsymbol{V}_h = \boldsymbol{W}_h^{\boldsymbol{V}} \cdot \boldsymbol{X} \quad (6)$$

式中：$\boldsymbol{W}_h^{\boldsymbol{Q}}$、$\boldsymbol{W}_h^{\boldsymbol{K}}$ 和 $\boldsymbol{W}_h^{\boldsymbol{V}}$ 是每个注意力头的权重矩阵；$\boldsymbol{X}$ 为输入向量。

混合模型通过添加位置编码 [9] 来注入顺序信息，通过将正弦和余弦波的不同频率与输入向量的每个位置结合。这些波形的频率随位置增加，为每个位置提供独特的编码。位置编码的计算过程为：

$$P(\text{pos}) = \sin(\frac{\text{pos}}{10\,000^2}) + \cos(\frac{\text{pos}}{10\,000^2}) \quad (7)$$

式中：pos 是指序列中的位置。

通过这种混合模型的构建，本文提出的 4 个混合模型，能够利用预训练模型学习到的丰富语言知识，并能够结合不同预训练编码器的优势，进一步提升机器翻译的性能。这种混合模型的设计理念是利用预训练编码器在语言理解方面的能力，结合 Marian 解码器在翻译生成方面的优势，从而实现更高质量的翻译结果。

2 实验与分析

2.1 数据集

为评估提出的机器翻译混合模型，实验选取了 Wmt14 和 Kde4 中的两个数据集进行测试。Wmt14 英 - 德数据集是机器翻译领域广泛使用的一个标准数据集，包含大量的英德语料对；Kde4 数据集则是一个小型但多样的语料库，包含英 - 法、英 - 德两种语言对。两个数据集均被划分为训练集、验证集和测试集，表 1 给出了详细的数据分布。通过在两个数据集上进行测试，可以全面评估文中提出的机器翻译混合模型的性能，包括其在标准数据集上的表现以及在小型但多样的数据集上的泛用性。

表 1 数据集和数据分布

数据集	语料 / 对	数据集划		
		训练集	验证集	测试集
Wmt14 英 - 德	4 514 788	4 508 785	3000	3003
Kde4 英 - 法	210 169	203 863	3153	3153
Kde4 英 - 德	224 035	217 315	3360	3360

2.2 对比实验

本文在 Wmt14 英 - 德数据集上进行了对比实验，表 2 展示了本文提出的 4 个混合模型与多种基线模型在该数据集上的性能对比。

表 2 Wmt14 英 - 德数据集上机器翻译模型性能对比

模型	BLEU	参数量 /MB	年
Marian-Luke	30.14	288	2024
Marian-Distilroberta	29.87	142	2024
Marian-Distilbert	29.65	126	2024
Marian-Electra	29.12	73	2024
PartialFormer	29.56	68	2023
Mega	29.01	67	2022
Mask Attention Network	29.10	63	2021
Lite Transformer	26.50	17	2020
Hardware Aware Transformer	24.40	48	2020
LightConv	28.90	202	2019
DynamicConv	29.70	213	2019

基线模型选取了最近 5 年内发表且参数量相当的模型，并使用 BLEU[10] 值作为评价指标，用以评估模型性能，其中前 4 个加黑的模型为混合模型。

结果显示，混合模型在 BLEU 得分上普遍优于基线模型，例如 Marian-Luke 模型取得了最高的 30.14 的 BLEU 得分。同时部分混合模型在参数小于基线模型的同时，性能也有所提升。例如 Marian-Electra 混合模型与 LightConv[11] 基线模型，虽然性能只提升了 0.22，但是参数量减少了 140 MB，说明混合模型能够在保证性能的同时，降低模型复杂度，加快训练和推理速度。DynamicConv[12] 模型在基线模型中取得了最高为 29.70 的 BLEU 得分，但其他基线模型的性能相对较低。混合模型在 BLEU 得分上普遍有所提升。这些结果表明，本文提出的混合模型能够在 Wmt14 英 - 德这一大数据集上有效提升机器翻译任务的性能，展现出其在处理大规模翻译任务方面的潜力。

为了进一步验证混合模型的有效性，并探究其在多语言翻译任务上的表现，本文选取了 KDE4 数据集进行实验。KDE4 数据集包含了英 - 法、法 - 英、英 - 德、德 - 英 4 种语言对的翻译数据，能够更全面地评估混合模型在多语言翻译任务上的性能，证明模型的通用性。

表 3 表明，混合模型在 KDE4 数据集上的英 - 法、法 - 英、英 - 德、德 - 英多种语言的数据集上均取得了优于 Marian 全量微调模型的 BLEU 得分。

表 3 kde4 多语言数据集混合模型 Belu 得分

模型	英 - 法	法 - 英	英 - 德	德 - 英
Marian-Luke	62.64	58.56	45.11	40.30
Marian-Distilroberta	61.25	45.79	49.81	37.09
Marian-Distilbert	58.64	47.22	45.02	41.37
Marian-Electra	55.27	46.28	45.16	36.39
marian-finetuned	52.94	42.51	43.71	32.69

Marian-Luke 在 kde4 数据集上的 BLEU 得分为 61.5，相较于使用 Marian 模型进行微调，BLEU 得分提高 9.15。证明混合模型在多语言机器翻译任务上的通用性。

2.3 消融实验

本文使用了 kde4 英 - 法在混合模型上进行了消融实验。表 4 中展现了 Marian-Luke、Marian-Distilroberta、Marian-Distilbert 和 Marian-Electra 混合模型在英 - 法翻译任务上的 BLEU 得分、SacreBLEU 得分和 RougeScore 得分。

表 4 kde4 英 - 法数据集混合模型和全量微调对比结果

模型	评价指标		
	BLEU	SacreBLEU	RougeScore
Marian-Luke	62.64	76.78	67.34
Marian-Distilroberta	61.25	76.47	66.50
Marian-Distilbert	58.64	75.76	64.51
Marian-Electra	55.27	74.75	62.97
marian-finetuned	52.94	74.07	61.52

消融实验结果表明，混合模型在所有评价指标上均优于 Marian 全量微调模型，证明了混合模型的有效性。其中，Marian-Luke 模型表现最佳，BLEU 得分达到 62.64，SacreBLEU 得分达到 76.78，RougeScore 得分达到 67.34。混合模型能够有效提升机器翻译性能，相较于 Marian 全量微调模型，BLEU 得分最高提升了 9.7。实验结果表明，混合模型在机器翻译任务中展现出强大的潜力，证明了以低成本构建新的预训练机器翻译模型这一方案的可行性。

3 结论

大语言模型使用机器学习和自然语言处理技术实现自动翻译，这让翻译更加节省了时间和经济成本；本文结合预训练模型和 Transformer 架构的混合模型，能够利用预训练模型学习到的丰富语言知识，以及 Transformer 架构在处理序列数据方面的优势，从而在机器翻译任务中取得优异的性能。为机器翻译领域的发展提供了新的方向。

参考文献：

[1] 罗锦钊，孙玉龙，钱增志，等 . 人工智能大模型综述及展望 [J]. 无线电工程，2023,53 (11):2461-2472.

[2] 贺承浩，王泽辉，滕俊哲，等 . 机器翻译综述 [J]. 电脑知识与技术，2023,19 (21):31-34.

[3] 岳增营，叶霞，刘睿珩 . 基于语言模型的预训练技术研究综述 [J]. 中文信息学报，2021,35 (9):15-29.

[4] 朱嘉辉，韩韧，张生，等．利用压缩多语言 BERT 知识增强的低资源神经机器翻译 [J/OL]. 计算机工程与应用，1-12[2024-04-17].http://kns.cnki.net/kcms/detail/11.2127.TP.20240416.1707.018.html.

[5] 谢思静，文鼎柱．基于联邦分割学习与低秩适应的 RoBERTa 预训练模型微调方法 [J]. 数据采集与处理，2024, 39(3): 577-587.

[6] 尹宝生，孔维一．基于 Electra 预训练模型并融合依存关系的中文事件检测模型 [J]. 计算机科学，2024, 51(S1):235-240.

[7] 薛擎天，李军辉，贡正仙，等．基于预训练的无监督神经机器翻译模型研究 [J]. 计算机工程与科学，2022,44 (4):730-736.

[8] 石磊，王毅，成颖，等．自然语言处理中的注意力机制研究综述 [J]. 数据分析与知识发现，2020, 4(5):1-14.

[9] 亢晓勉，宗成庆．融合篇章结构位置编码的神经机器翻译 [J]. 智能科学与技术学报，2020, 2(2):144-152.

[10]EVTIKHIEV M, BOGOMOLOV E, SOKOLOV Y, et al. Out of the BLEU: how should we assess quality of the code generation models?[J]. Journal of systems and software, 2023, 203(9): 111741.

[11]MENG L X, TAN W J, MA J G, et al. Enhancing dynamic ECG heartbeat classification with lightweight transformer model[J]. Artificial intelligence in medicine, 2022, 124: 102236.

[12]ZHU L J, PENG L, DING S C, et al. An encoder‐decoder framework with dynamic convolution for weakly supervised instance segmentation[J]. IET computer vision, 2023, 17(8): 883-894.

【作者简介】

张乐（1999—），女，陕西西安人，硕士研究生，研究方向：自然语言处理、大语言模型研究、故障诊断。

荆晓远（1971—），通信作者（email: jingxy_2000@162.com），男，江苏南京人，博士，教授、博士生导师，研究方向：模式识别、计算机视觉、故障诊断。

孙其航（1999—），男，山东泰安人，硕士研究生，研究方向：自然语言处理、大语言模型研究、故障诊断。

（收稿日期：2024-10-15）

（上接第 45 页）

参考文献：

[1] LINDBLAD T, KINSER J M. Image processing using pulse-coupled neural networks[M/OL].2nd. Berlin: Springer,2005[2024-06-13].https://link.springer.com/book/10.1007/3-540-28293-9.

[2] ECKHORN R, REITBOECK H J, ARNDT M, et al. A neural network for feature linking via synchronous activity: results from cat visual cortex and from simulations[M]//COTTERILL R. Models of Brain Function, Cambridge: Cambridge University Press, 1989: 255-272.

[3] WANG Z B, MA Y D. Medical image fusion using m-PCNN[J]. Information fusion, 2008,9(2): 176-185.

[4] 苗启广，王宝树．基于局部对比度的自适应 PCNN 图像融合 [J]. 计算机学报，2008(5): 875-880.

[5] 杨艳春，党建武，王阳萍．基于提升小波变换与自适应 PCNN 的医学图像融合方法 [J]. 计算机辅助设计与图形学学报，2012, 24(4): 494-499.

[6] 严春满，郭宝龙，易盟．基于改进 LP 变换及自适应 PCNN 的多聚焦图像融合方法 [J]. 控制与决策，2012, 27(5): 703-708.

[7] 李美丽．基于多尺度变换的 PCNN 和 FOA 图像融合 [J]. 光电子·激光，2016, 27(7): 767-772.

[8] 李奕，吴小俊．粒子群进化学习自适应双通道脉冲耦合神经网络图像融合方法研究 [J]. 电子学报，2014, 42(2):217-222.

[9] 石美红，张军英，张晓滨，等．基于改进型脉冲耦合神经网络的图像二值分割 [J]. 计算机仿真，2002, 19(4):42-46.

[10] 马义德，戴若兰，李廉．一种基于脉冲耦合神经网络和图像熵的自动图像分割方法 [J]. 通信学报，2002(1):46-51.

[11] 高超．图像融合评价方法的研究 [J]. 电子测试，2011(7): 30-33.

[12] 王文文，王惠群，陆惠玲，等．基于压缩感知和 NSCT-PCNN 的 PET/CT 医学图像融合算法 [J]. 重庆理工大学学报 (自然科学版), 2016, 30(2):101-108.

【作者简介】

王观英（1989—），女，江苏徐州人，硕士，讲师，研究方向：模式识别与人工智能、机器学习。

（收稿日期：2024-10-30）