

基于深度学习和上下文感知算法的 跨语言翻译方法

王丹[†]

(哈尔滨医科大学公共外语教研部,黑龙江 大庆 163000)

摘 要:针对目前跨语言机器翻译存在性能较低的问题,提出了一种基于深度学习和上下文感知算法的跨语言翻译模型。还提出了一个具有 12 层的跨语言 Transformer 编码器-解码器结构,充分学习不同语言表达特征。再提出了一种数据对齐方法,可有效扩充数据样本数量,增加数据多样性和数据量,减少数据和模型之间复杂度的相对差异,缓解过拟合问题。实验阶段,通过有监督、无监督、零样本数据集测试,证明所提模型具备较优性能。实验充分验证了所提模型可有效表达未知语种。与 m-Transformer 和 mRASP 等主流模型相比,所提模型 BLEU 分数较优。实验结果证明了所提出跨语言翻译模型的有效性及其实用性,该模型可为跨语言机器翻译领域的发展提供一定借鉴作用。

关键词:机器翻译;跨语言模型;深度学习;数据对齐;Transformer

中图分类号:TP393

文献标识码:A

Cross Language Translation Method Based on Deep Learning and Context Aware Algorithms

WANG Dan[†]

(Department of Public Foreign Language Teaching and Research, Harbin Medical
University, Daqing, Heilongjiang 163000, China)

Abstract: A cross language translation model based on deep learning and context aware algorithms is proposed to address the issue of low performance in current cross language machine translation. A cross language Transformer encoder-decoder structure with 12 layers was proposed to fully learn different language expression features. A data alignment method has been proposed, which can effectively expand the number of data samples, increase data diversity and volume, reduce the relative differences in complexity between data and models, and alleviate overfitting problems. During the experimental phase, the proposed model demonstrated superior performance through supervised, unsupervised, and zero sample datasets testing. The experiment fully verified that the proposed model can effectively express unknown languages. Compared with mainstream models such as m-Transformer and mRASP, the proposed model has better BLEU scores. The experimental results demonstrate the effectiveness and practicality of the proposed cross language translation model, which can provide some reference for the development of cross language machine translation.

Key words: machine translation; cross language model; deep learning; data alignment; Transformer

Transformer^[1,2]作为深度学习领域^[3]一个重要分支,凭借其丰富的双语并行语料库在机器翻译

方面取得了巨大成就。随着翻译要求的不断提升,机器翻译^[4]已经从单一语言向跨语言发展。跨语言机器翻译的工作旨在创建一个统一的模型来翻译多种语言。跨语言翻译模型之所以有吸引力,有两个原因:其一为模型效率高,能够更容易地进行部署;其二,不同语言之间的参数共享鼓励了知识转移,这有利于资源高效利用,并有可能实现零样本翻译,即对从未训练过的语言进行直接翻译。

目前,许多学者针对机器翻译技术进行研究,取得了丰硕的研究成果。文献[5]提出了一种融合机器翻译方法与复述生成方法的两阶段中文手语文本生成方法。该方法首先在多语言双向和自回归转换器模型中加入注意力机制,从而提升模型对手语文本敏感度。文献[6]提出基于深度可分离卷积的英语神经机器翻译方法。该方法以英语为研究中心,将英语切分为单词、音阶、字符、词等不同粒度,并基于深度可分离卷积和注意力机制提升机器翻译质量。文献[7]提出了一种基于预训练的无监督神经机器翻译模型。该模型基于二次预训练语言模型和自注意力机制层优化机器翻译模型,从而提升跨语言模型翻译性能,有效解决高质量平行语料库难以获取的问题。尽管有众多好处,但跨语言机器翻译问题仍然面临挑战。跨语言机器翻译模型权重必须在多种语言之间分配,模型所容纳的语言数量不断增加,必然导致模型性能的急速下滑。此外,跨语言机器翻译的最佳配置应该对任何语言对都有效,然而目前大多数工作都集中在改进以英语为中心的翻译方向,许多学者通过数据挖掘或反翻译引入更多的非英语语料库训练跨语言机器翻译模型。

为提高跨语言机器翻译性能,本文提出了一种基于深度学习和上下文感知算法的跨语言翻译模型。该模型以 Transformer 为基础,通过对齐扩充、对比学习策略充分挖掘不同语言之间联系,可减少不同语言的表达差距,提升跨语言翻译质量。

1 跨语言翻译模型

1.1 模型框架

所提基于深度学习和上下文感知的跨语言翻译模型总体框架如图 1 所示。模型采用一对并行句子作为输入,并使用跨语言 Transformer 的编码器—解码器计算正态交叉熵损失。所提模型将并行语料库和单语言语料库对比学习,以缩小不同语言之间的表达性差距,从而提升翻译的质量。此

外,本研究还引入了一种有效的对齐扩充技术,通过计算对齐对(正样本)和随机选择的非对齐对(负样本)的表示的对比损失最小化训练差异,提升训练质量。所提模型可以充分利用来自所有监督方向的知识,通过跨语言对比学习以最小化相似句子的表示距离,确保模型在共享空间中充分学习跨语言表示相似的句子。

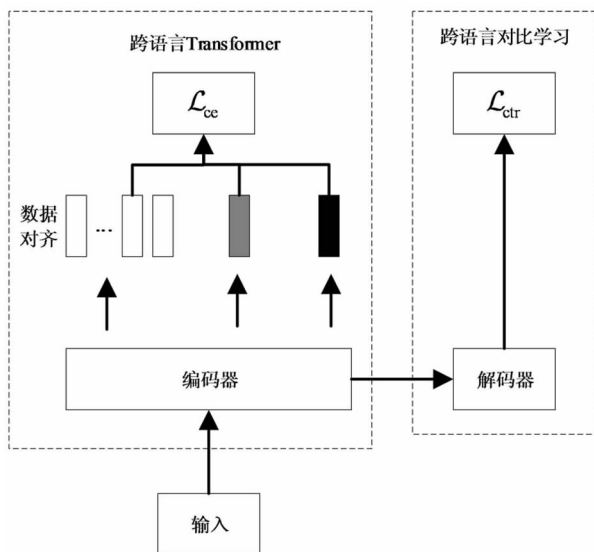


图 1 模型总体框架

1.2 跨语言 Transformer

跨语言翻译模型旨在学习多对多映射函数 f 以从一种语言翻译到另一种语言。为了区分不同的语言,本研究在每个句子前面添加一个额外的语言识别标记,用于区分源端和目标端。Transformer 因其在预测文本中可有效感知上下文信息并较好地捕捉语义关联效果,现已广泛应用于自然语言处理(Natural Language Processing, NLP)领域。为此,本研究提出的基于深度学习和上下文感知的跨语言翻译模型以 Transformer 为基础结构。

跨语言 Transformer 能够隐式学习不同语言的共享表示,并基于上下文感知不同句子之间的复杂关系。本研究提出的跨语言 Transformer 带有 12 层编码器和 12 层解编码器,以增加模型容量,从而更好地学习不同语言之间的语义关系。模型输入设置为 1024。为了简化深度模型的训练过程,对编码器和解码器应用归一化进行词嵌入和预范数残差连接。

令跨语言语料库集合为 $\mathcal{L} = \{L_1, L_2, \dots, L_M\}$, 其中 M 为跨语言的个数。令 D_{ij} 表示由语料库 L_i 和 L_j ($i, j \in M$) 构成的并行语料库数据集, D 表示所有并行语料库数据集。跨语言 Trans-

former 训练损失为交叉熵损失函数,具体定义为:

$$\mathcal{L}_{ce} = \sum_{x^i, x^j \in D} -\log P_{\theta}(x^i | x^j) \quad (1)$$

式中: $\mathcal{L}_{ce}$ 为跨语言 Transformer 训练损失; x^i 为语料库 L_i 中的一个句子; x^j 为语料库 L_j 中的一个句子; θ 为跨语言 Transformer 模型的参数; P_{θ} 为预测概率。

1.3 跨语言对比学习

所提模型中引入了跨语言对比学习损失,可将不同的语言映射到共享的语义空间,从而提升不同语言之间相关性学习。对比学习的核心思想是最小化相似句子的表征差距,最大化无关句子的表征差异。

给定一个双语翻译对 $(x^i, x^j) \in D$, 且 (x^i, x^j) 为一个正样本。接下来从语料库 L_j 中随机选择一个句子 y^j 形成一个负样本 (x^i, y^j) 。需注意,可能存在 $L_j = L_i$ 。跨语言对比学习的目的是尽量减少以下损失函数:

$$\mathcal{L}_{ctr} = - \sum_{x^i, x^j \in D} \log \frac{e^{\sin^+(R(x^i), R(x^j))/\tau}}{\sum_{y^j} e^{\sin^-(R(x^i), R(y^j))/\tau}} \quad (2)$$

式中: $\mathcal{L}_{ctr}$ 为跨语言对比学习损失; $\sin(\cdot)$ 为计算不同句子的相似性函数; $+$ 和 $-$ 分别表示正和负符号; $R(\cdot)$ 为任意句子的平均汇集编码输出函数; τ 为控制参数,用来区分正样本和负样本的难度。本研究中,设置 τ 为 0.1。同时,两个句子的相似性函数 $\sin(\cdot)$ 是用平均合并编码输出的余弦相似性来计算。为了简化实现,从同一训练批次中对负样本进行采样。接着,通过最大化 softmax 项 $\sin^+(R(x^i), R(x^j))$, 并对比损失迫使相似的两个句子的语义投影彼此靠近。同时,softmax 函数还最小化了不匹配对 $\sin^-(R(x^i), R(x^j))$, 即使两个不相似句子的语义投影分离得更远。

所提基于深度学习和上下文感知的跨语言翻译模型训练过程中,可以通过联合最小化对比训练损失和 Transformer 训练损失来优化模型:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda |s| \mathcal{L}_{ctr} \quad (3)$$

式中: λ 为平衡两种训练损失的平衡系数; $|s|$ 为平均序列长度。由于 $\mathcal{L}_{ctr}$ 在句子级别上计算,而 $\mathcal{L}_{ce}$ 是在语料库级别计算,因此 $\mathcal{L}_{ctr}$ 应该乘平均序列长度 $|s|$ 。

1.4 对齐扩充技术

为提升样本多样性,增强模型鲁棒性,本研究使用对齐扩充技术增强样本,具体策略包括为引入有噪声的双语数据和有噪声的单语数据。对齐扩

充技术结构图如图 2 所示。

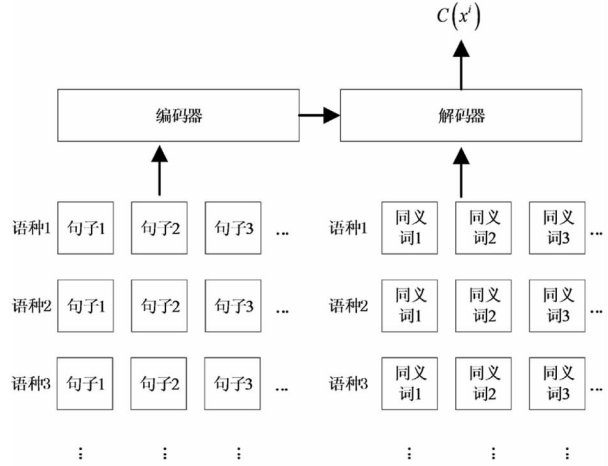


图 2 对齐扩充技术结构图

对于任意双语或单语句子对 (x^i, x^j) , 本研究提出的对齐扩充技术通过替换同义词字典中对齐的词创建扰动句子 $C(x^i)$ 。同时,对于同义词词典中包含的每个词,本研究以 90% 的概率将其随机替换为其同义词。

对于双语句子对 (x^i, x^j) , 可基于对齐扩充技术创建一个伪并行训练样本 $(C(x^i), x^j)$ 。对于单语数据,通过对齐扩充技术提取一个句子 x^i , 并生成其扰动 $C(x^i)$, 以形成并行样本 $(C(x^i), x^i)$ 。接着,在训练中通过 $(C(x^i), x^j)$ 和 $(C(x^i), x^i)$ 计算 Transformer 训练损失和对比学习损失。

2 实验与分析

为验证所提基于深度学习和上下文感知的跨语言翻译模型有效性,本节以案例样本库对模型性能进行交叉验证及分析。

2.1 数据集与实验设置

本研究所用数据集包括 PC32、Newscrawl10-Z、OPUS-100 零样本集三个基础语料库以及一个跨语言词汇语料库。

PC32 为由 32 对以英语为中心的语言对组成的大型公共语料库,其中句子对总数约为 9760 万。在基础 PC32 数据集上应用对齐扩充技术,即将任意双语词典中的同义词随机替换其他源句中的单词。对于字典中的单词,实验时以 90% 的概率将其替换为同义词之一,否则保持不变。

Newscrawl10 为一个多语言大型公共语料库。为简化实验过程,本研究选取 Newscrawl10 中 24

种语言形成一个单语言文本数据集 Newscroll10-Z。该子集只保留与 PC32 中类似的语言,加上 PC32 中没有的另外 3 种语言。为了平衡不同语言之间的数量,实验时在数据集上应用温度采样:

$$\tilde{n}_i = \left[\frac{n_i}{\sum_j n_j} \right]^{\frac{1}{T}} \tag{4}$$

式中: T 为温度系数,实验时设置为 5; n_i 为第 i 种语言中的句子数量。

接着,基于对齐扩充技术扩展单语言文本数据集,方法为用多语言词典中的同义词随机替换源句子中的单词。同时,替换概率也设置为 90%。经扩充后,Newscroll10-Z 中的句子总数约为 10.1 亿。

OPUS-100 零样本测试集由 6 种语言(俄语 Ru、德语 De、法语 Fr、荷兰语 Nl、阿拉伯语 Ar、中文 Zh)组成,共有 15 对语言和 30 个翻译方向。

跨语言词汇语料库包括 59 种语言。该词汇表包含 64808 个标记。在添加 59 个语言标记后,词汇表的总大小为 64867。

实验时使用的 Transformer 模型包括 12 个编码器层和 12 个解码器层,词嵌入大小和 FFN 维度设置为 1024。Dropout 设置为 0.1,学习率设置为 0.0001,学习率衰减函数设置为多项式衰减调度,预热步骤设置为 10000。优化时使用 Adam 优化器,其中维持数值稳定性的参数设置为 1×10^{-8} ,动量参数设置为 0.9 和 0.999。为了稳定训练,将

梯度范数的阈值设置为 5.0,并用较大的范数剪裁所有梯度。

实验环境设置如下:硬件使用 8×4 NVIDIA A100 训练模型,软件采用 pycharm2021 作为算法编译环境,使用 python3.8 和 pytorch1.13.0 编写 Transformer 及其改进模型。

2.2 对比与分析

本节首先进行了有监督消融实验,对比指标选取 BLEU 分数。实验时基于三种基础语料库对模型进行预训练,之后选取跨语言词汇语料库中英语—法语(En-Fr)、英语—土耳其语(En-Tr)、英语—西班牙语(En-Es)、英语—罗马尼亚语(En-Ro)、英语—中文(En-Zh)作为案例进行分析。对比模型包括含有 6 层编码器层和解码器层的 Transformer-6、12 层编码器层和解码器层的 Transformer-12、m-Transformer、跨域预训练语言模型(Cross-lingual Language Model Pretraining, XLM)、掩码序列到序列预训练(Masked Sequence to Sequence Pre-training, MASS)、mRASP 以及所提模型。不同模型有监督消融实验 BLEU 分数对比结果如表 1 所示。其中符号“→”表示从前一语种翻译至后一语种,如 En→Fr 指从 En 翻译至 Fr;符号“—”表示翻译未成功导致 BLEU 分数无效。

表 1 不同模型有监督消融实验 BLEU 分数对比结果/%

模型	En-Fr		En-Tr		En-Es		En-Ro		En-Zh		均值
	→	←	→	←	→	←	→	←	→	←	
Transformer-6	43.23	39.86	—	—	—	—	34.32	34.02	—	—	—
Transformer-12	41.41	—	9.51	12.25	33.20	—	34.31	36.81	20.23	21.85	—
m-Transformer	42.01	38.15	18.82	23.10	32.79	33.68	35.89	37.74	20.04	28.21	31.04
XLM	—	—	—	—	—	—	—	38.54	—	—	—
MASS	—	—	—	—	—	—	—	39.13	—	—	—
mRASP	44.31	39.38	20.01	23.41	34.02	34.28	37.52	38.78	22.01	29.19	32.29
所提模型	43.51	39.29	21.45	25.83	34.47	35.01	38.04	39.11	23.38	30.15	33.02

可以看出,在众多对比模型中,所提基于深度学习和上下文感知算法的跨语言翻译方法具备竞争力,平均 BLEU 分数为 33.02%。与 m-Transformer 和 mRASP 模型相比,所提模型 BLEU 分数分别提高 1.98%和 0.73%。除英语—法语外,所提模型基本具有较高的 BLEU 分数。此外,通过有监督实验对比可以看出,Transformer 层数对 BLEU 分数具有较大影响。如 Transformer-6 提

升为 Transformer-12 时,模型 BLEU 分数提升较大。

接着,本研究还进行了无监督消融实验对比。与有监督对比实验类似,对比指标选取 BLEU 分数。不同之处主要设置如下:其一为无监督消融实验只进行预训练,无微调过程,即预训练后直接基于跨语言词汇语料库进行测试;其二为仅对比 m-Transformer、mRASP 以及所提模型;其三为测试

的翻译对包括英语—荷兰语(En-Nl)、英语-葡萄牙语(En-Pt)、英语-波兰语(En-Pl)、荷兰语—葡萄牙语(Nl-Pt)。不同模型无监督消融实验 BLEU 分数对比结果如表 2 所示。可以看出,所提方法在无监督翻译消融实验时取得了最优的结果。m-

Transformer 和 mRASP 模型无法翻译从未直接训练过的英语—波兰语(En-Pl)语言对。相比之下,所提模型在无监督情况下平均 BLEU 分数为 15.75%。

表 2 不同模型无监督消融实验 BLEU 分数对比结果/%

模型	En-Nl		En-Pt		En-Pl		Nl-Pt		均值
	→	←	→	←	→	←	→	←	
m-Transformer	1.31	7.04	3.69	10.69	0.58	3.28	—	—	—
mRASP	0.73	10.65	3.71	11.64	0.55	5.32	—	—	—
所提模型	10.11	28.54	18.53	30.52	3.48	17.12	9.33	8.35	15.75

接下来进行了非零样本消融实验对比。测试时在 OPUS-100 零样本测试集上评估所提模型性能,目标语种包括:阿拉伯语(Ar)、中文(Zh)、荷兰语(Nl)、法语(Fr)、德语(De)、俄语(Ru)。非零样本 BLEU 分数对比结果如表 3 所示。表 3 中 X 表

示任意语种翻译至目标语种,如 X→Ar 指任意语种翻译至阿拉伯语;“*”表示荷兰语(Nl)不在训练集中。可以看出,所提模型性能优于 m-Transformer,平均 BLEU 分数为 15.49%。实验结果验证了所提模型可更好地表达未知句子。

表 3 非零样本 BLEU 分数对比/%

模型	Ar		Zh		Nl(*)		均值
	X→Ar	X←Ar	X→Zh	X←Zh	X→Nl	X←Nl	
m-Transformer	5.53	17.01	28.49	16.43	2.21	6.03	
mRASP	3.72	5.65	6.73	4.11	2.32	6.30	
所提模型	5.35	17.32	29.03	14.51	5.33	6.10	

模型	Fr		De		Ru		均值
	X→Fr	X←Fr	X→De	X←De	X→Ru	X←Ru	
m-Transformer	16.10	12.31	4.43	14.20	6.60	8.14	11.46
mRASP	7.68	4.82	4.21	4.83	5.76	4.82	5.08
所提模型	23.63	21.74	12.31	15.03	16.43	19.10	15.49

3 结 论

提出了一种基于深度学习和上下文感知算法的跨语言翻译模型。模型以跨语言 Transformer 为基础网络,包含 12 层编码器—解码器结构。同时,该模型将并行语料库和单语言语料库对比学习,以缩小不同语言之间的表达性差距,从而提升翻译的质量。最后,通过试验验证了所提出的方法能够高质量处理跨语言翻译问题,并为跨语言翻译的发展提供了一定的技术基础。

参考文献

[1] 尚春磊,李颖俊,张璐. 基于数据挖掘的自动化翻译系统优化

设计研究与应用[J]. 自动化与仪器仪表, 2023(8): 273—276.

[2] 宫昀. 基于 Transformer 模型的神经机器翻译改进方法研究[J]. 自动化与仪器仪表, 2023(8): 257—261+267.

[3] 齐硕. 基于深度学习的音乐情绪分类研究[J]. 云南师范大学学报(自然科学版), 2023, 43(2): 29—33.

[4] 何娟. 基于深度学习网络的手写英文自动化识别模型在机器英汉互译中的应用研究[J]. 自动化与仪器仪表, 2023(7): 191—195.

[5] 李世伟,侯霞,汪良果. 融合机器翻译与复述生成的手语文本生成方法[J]. 北京信息科技大学学报(自然科学版), 2023, 38(2): 75—81.

[6] 晏芳,罗刚峰,司海峰. 面向平行语料库和多层次语言特征的英语翻译系统研究[J]. 自动化与仪器仪表, 2023(3): 213—217.

[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.