基于降噪原型序列的汉越神经机器翻译

杨汉清[1,2]，赖华[1,2]，于志强[1,2*]，余正涛[1,2]

（1. 昆明理工大学信息工程与自动化学院，云南 昆明 650500；2. 云南省人工智能重点实验室，云南 昆明 650500）

摘要：原型序列旨在用目标端语言信息指导机器翻译，已有的工作主要是利用相似性翻译作为目标端原型序列,提升神经机器翻译的性能。然而在汉越低资源场景下,平行语料匮乏,原型序列蕴含庞杂的信息,直接使用会增加翻译模型训练的难度，甚至引入噪声。针对此问题，本文提出了一种基于降噪原型序列的汉越神经机器翻译方法。首先，利用跨语言检索得到原型序列；其次，基于实体词典对原型序列中的噪声信息进行掩盖，再综合稀有词词频及语义相似度，得到原型序列的参考价值；最后使用额外的编码器接收原型序列，并允许解码器到两个编码器间建立注意力机制。实验结果表明,相比基线模型,本文所提出的方法能够有效提升汉越神经机器翻译的性能。

关键词：汉越神经机器翻译；低资源；原型序列；降噪

中图分类号：TP 391　　**文献标志码：** A

随着深度学习技术的发展,基于深度学习的神经机器翻译系统[1-3](neural machine translation, NMT)取得了显著的效果，成为机器翻译任务的主流框架。但目前神经机器翻译的性能较大程度上依赖于大规模平行语料， 对于汉语到越南语这类平行语料较少的低资源语言对,神经机器翻译的效果往往不尽如人意。

为了提升神经机器翻译的效果, 许多研究人员进行了大量的研究。一方面，人工翻译源句时会参考相似句子的翻译方法，受其启发，众多学者尝试将目标端语言的语言信息作为指导融入神经机器翻译中：Qian Cao 等[4]利用源端语言的相似性进行模糊匹配,将匹配结果作为翻译记忆融入 NMT 当中；Yiren Wang 等[5]将编码器-解码器框架外用于提升模型性能的序列定义为原型序列（prototype），并把源语言和目标语言之间词语的概率关系作为原型序列，为 NMT 引入了目标语言的全局信息；Cao 等[6]提出一种位于 NMT 的解码端的门控机制，来平衡目标端信息对源端的影响。另一方面，为了防止模型陷入局部优化，使其更快速收敛，Emmanouil 等[7]还提出了基于课程学习的 NMT 训练框架，在模型学习能力和数据难度的基础上调整训练数据。这些方法一定程度上提高了神经机器翻译的性能,但需要依赖较大规模的双语平行语料来训练模型。

在资源匮乏的情况下,目标语言端的单语数据已经被证实能够极大的提升模型的翻译质量，并被广泛利用，最著名的就是回译[8]（back-translation）。越南单语语料相较于平行语料具有数量多、获取成本低等特点。为了在汉越神经机器翻译中利用越南语单语的语言信息,可以通过跨语言相似性检索出越南语的单语语句[9]，将其作为原型序列,并通过编码器结构引入神经机器翻译。然而，由于汉语、越南语分属不同语系，在符号表示、语法等方面存在差异，而且越南语拥有独特的数字表示方式。两种语言多方面的差异导致检索到的越南语原型序列质量不佳。在汉越神经机器翻译中应用越南语作为原型序列时存在两个问题。一方面，基于相似性检索的原型序列中所含的实体、数字多数情况下和源句子中的实体、数字无法对应，这将会在汉越翻译模型的训练过程中引入噪声。另一方面，相似性较低的越南语原型序列会包含稀有词[7]，模型学习这部分原型序列时需要耗费更多时间以及运算成本，稀有词的词嵌入[10]在模型计算损失时也会带来一定误差,这部分的越南语原型序列对模型来说指导作用是偏弱的。此外，越南语是一种拼音文字，书写时以音节作为最小粒度。在传统汉越神经机器翻译中,往往会使用切分后的子词粒度进行模型训练,这虽然可以提升模型的翻译表现,但却给一些需要进行词级粒度处理的工作带来不便。

针对以上问题，本文构建了基于降噪原型序列的汉越神经机器翻译模型。首先将汉语和越南语的句子映射到向量空间，利用跨语言相似性检索出目标端原型序列；接着对子词级原型序列进行粒度还原,再依据构建好的越南语实体词典做噪声掩盖；之后依据越南语原型序列与源端的相似性以及稀有词词频对其进行权重分配，加大原型序列之间的特征差异，赋予其更合理的参考价值判别标准；最后将处理好的

基金项目：国家自然科学基金（61732005，61972186，U21B2027）；云南省重大科技专项（202103AA080015，202002AD080001，202202AD080003）；云南省高新技术产业专项（201606）；云南省教育厅科学研究基金项目（2022J0449）

***通讯作者：**yzqyt@ymu.edu.cn

原型序列作为模板信息，应用在双编码器-单解码器结构，指导翻译任务，从而进一步提高神经机器翻译的效果。本文章节安排如下：第 1 节介绍原型降噪策略；第 2 节介绍本文提出的模型；第 3 节通过实验分析本文所提出方法的有效性；第 4 节对本文进行总结。

1 原型序列的降噪策略

1.1 基于 mask 机制的原型去噪策略

在训练模型过程中，将源语言句子 x 以及候选的目标语言句子 s 通过各自的编码器映射到向量空间之后，得到对应的高维向量E_x和E_s，x 与 s 之间的相关性分数可通过计算得出：

$$r(x_i, s_i) = E_{x_i}^T E_{s_i} \quad (1)$$

我们可以筛选出与源语言句子x_i相关性较高的前 M 个目标语言句子作为原型序列。然而在实践中，我们发现源语言句子与原型序列中的实体、数字多数情况下无法对应，具体示例如表 1 所示。

表 1 检索的原型序列示例

Tab.1 examples of prototype

源语言(汉语)	原型序列(越南语)
我快满 17 岁了	1、 Bây giờ đã có hơn 30000 người （快要 3 万年了） 2、 Mình đã giảm được năm kg đấy （我瘦了将近 5 千克） 3、 Đó à hơn 500000 từ （不止 500 万）
Mike 也听到了歌声	1、 Tommy có thể nghe thấy tiếng vọng （汤米能听到回声） 2、 Nam huyn soo có thể nghe tôi （Nam huyn 能听到你说话） 3、 Jenna rink cũng đang nghe. （珍娜·林克在听）

为了解决这个问题，本文首先针对由越南语构成的候选句子库，构建了一个包含数字、人名、地名的实体词典，在检索出原型序列之后，利用 mask[10]的方法对原型序列的实体进行掩盖，避免引入噪声。给原型序列中实体$S_{illegal}$的位置加上一个无穷大的负偏置，即：

$$S_{illegal} = S_{illegal} + Bias_{illegal} \quad (2)$$

$$Bias_{illegal} \to -\infty \quad (3)$$

处理好的原型序列会经过编码器，会通过自注意力机制[10-11]捕获原型序列中各个单词之间的依赖特征，在这个过程中，这些噪声实体就不会参与运算：

$$\sigma(S_{illegal}) = \frac{e^{S_{illegal}}}{\sum_{j=1}^{K} e^{S_j}} \quad (4)$$

1.2 基于稀有词词频的原型评估策略

检索到的原型序列在翻译模型中起到指导的作用，原型序列中的单词频率会对原型序列的参考价值带来影响。用包含稀有词的原型序列作为指导信息会给模型训练带来困扰：（1）模型在学习包含稀有词[7]的训练样本时，需要不断重新访问样本中的稀有词，这增加了训练成本；（2）稀有词词嵌入的梯度往往具有较高方差，在梯度计算时往往会带来偏差。在相似度较低的原型序列中，会包含部分稀有词。为了解决这个问题，本文结合课程学习[7]中，根据稀有词词频判断样本难度的方法，评估出原型序列中每个句子的难度d_{s_i}，再综合与源语言句子的相似度 r(x_i, s_i)得到每个句子的参考价值：

$$c_{s_i} = r(x_i, s_i) + d_{s_i} \quad (5)$$

这可以将其中对机器翻译有益的知识特征进行增强，给模型更加有效的启发。

2 基于降噪原型序列的汉越神经机器翻译模型

为了提升汉越神经机器翻译的效果，本文提出了基于降噪原型序列的汉越神经机器翻译模型。图 1 展示了模型的整体结构，模型基于 Transformer[10]框架构建，总体上可以分为三个模块：(1)检索模块：利用源语言 x 和目标语言 y 之间的语义相似性，在候选句子库中检索出原型序列 s；(2)原型序列处理模块：对照实体词典，掩盖掉原型序列 s 中的噪声实体，再根据原型序列中稀有词的词频判定其难度分数，结合原型序列与源语言序列的相似度，得出原型序列的参考价值；(3)生成模块：为了适应额外的原型序列，基于 Transformer 模型扩展出一个双编码器-单解码器结构，并允许解码器到两个编码器构建注意力机制。

2.1 检索模块

检索模块用于从一个大规模的目标语言候选句子库当中，检索出和源语言句子最相关的句

子。候选句子库中的句子数量众多，为了更加高效地完成检索，本文采用的是最大内积搜索MIPS[12] (Maximun Inner Product Search)。

给定源语言编码器一个包含l个词的源语言句子$x=\{x_1, x_2,\ldots, x_l\}$，词嵌入层将其转换为词嵌入向量 $E_x=\mathrm{E}(x_1, x_2, \cdots, x_l)$。同理可以得到目标语言 $s=\{s_1, s_2,\ldots, s_l\}$ 的词嵌入向量 $E_s=\mathrm{E}(s_1, s_2, \cdots, s_l)$。源语言句子 x 和供候选的目标语言句子 s 之间的相关性得分可以通过公式(1)计算得到。

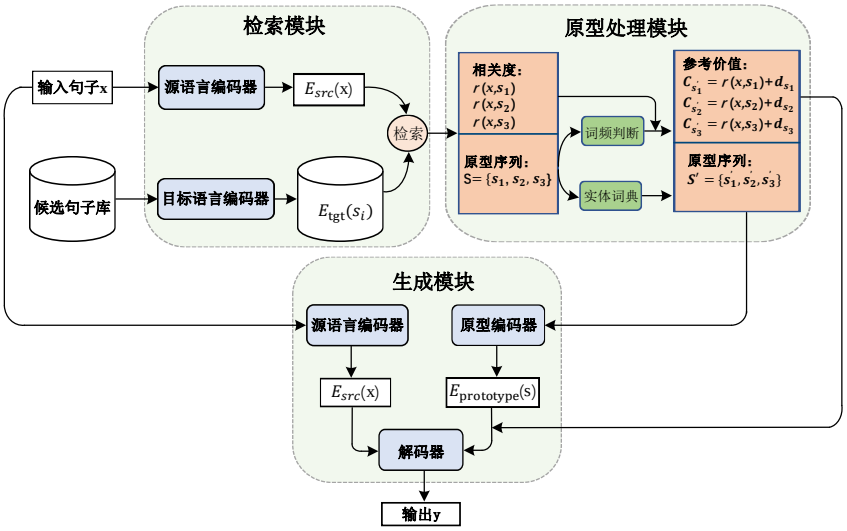

图 1 基于降噪原型序列的神经机器翻译模型

Fig.1 NMT Based on Denoising of Prototype

我们采用了一个开源的检索工具 Faiss 进行检索。给定一个源语言句子 x，计算出其向量表征E_x之后，借助开源的检索工具 Faiss 就能够检索出前 M 个与之最相关的目标语言句子 $S=\{s_i\}_{i=1}^{M}$作为原型序列。为了方便后续原型序列处理，我们分别检索出了基于音节粒度以及基于子词粒度的目标语言句子作为原型序列。

2.2 原型序列处理模块

在将检索到的原型序列用于翻译任务之前，我们需要先对其进行处理。

一方面，根据构造好的实体词典，对原型序列进行词匹配，掩盖不需要的数字以及实体。由于越南语在书写时以音节为最小粒度，而在整个模型训练中，为了缓解词表过大的问题，我们使用的是更小的子词粒度，因此在词匹配之前我们先对子词粒度的原型序列做处理，将被拆分的音节还原，便于之后的词匹配。

首先，对照实体词典，找出音节粒度的原型序列中的数字以及实体，再利用还原后的句子对照出子词粒度原型序列中数字以及实体的位置，最后将子词粒度句子中对应位置的字符用符号 ‘<mask>’替换，流程如图 2 所示。

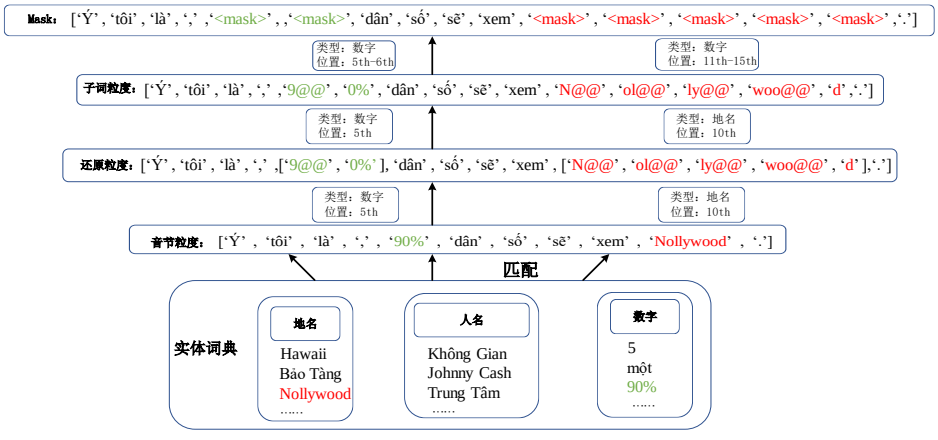

图 2 基于 mask 机制的原型去噪策略

Fig.2 Denoising Strategy Based on Masking

另一方面，原型序列中的单词频率也会对原型序列的参考价值带来影响。根据词表，对候选句子库$\{s_i\}_{i=1}^{m}$的句子进行词频统计，将出现频率低于 10%的词作为稀有词ω_j。在检索模型检索出原型序列以后，计算原型序列稀有词的相对词频：

$$\hat{P}(\omega_j) = \frac{1}{N_{total}} \sum_{k=1}^{N_i} 1_{\omega_i=\omega_j} \quad (6)$$

$1_{\omega_i=\omega_j}$为指示函数，当原型序列中的词ω_i为稀有词时，$1_{\omega_i=\omega_j}$函数值为 1；相反，若不是稀有词时，函数值为 0 。N_{total}为原型序列总的子词数目。接下来，我们将原型序列中所有稀有词的相对词频聚合，判断该原型序列的整体难度分数。

$$d_{s_i} = -\sum_{k=1}^{N_i} log\hat{P}(\omega_j) \quad (7)$$

结合检索模块中的得到相关性得分 r(x_i, s_i)以及当前模块得到的原型序列难度d_{s_i}，根据公式(5)得到原型序列的参考价值c_{s_i}。

2.3 生成模块

给定一个源语言句子 x，经过检索模块及原型处理模块得到的一组原型序列$S'=\{s_i'\}_{i=1}^{M}$，以及原型序列对应的参考价值$\left\{c_{s_i'}\right\}_{i=1}^{M}$，在生成模块中可以计算条件概率 p(y|x, $s_1', c_{s_1'}, \ldots, s_i', c_{s_M'}$)。

生成模块基于标准的编码器-解码器框架，源语言编码器将源语句 x 转换为稠密的向量表示，解码器以自回归的方式生成一个序列 y。在每个时间步 t，解码器根据先前时间步生成的序列$y_{1:t-1}$以及源语言编码器的输出，生成隐藏状态h_t。经过一次线性变换以及 softmax 运算以后，可以得到下一个子词的预测概率值：

$$P_v = softmax(W_v h_t + b_v) \quad (8)$$

对于原型序列$S' = \{s_i'\}_{i=1}^{M}$，需要在编码器-解码器框架之外引入一个额外的原型编码器。原型编码器将原型序列的每个句子s_i'转换为一组词嵌入$\left\{s_{i,k}'\right\}_{k=1}^{L_i}$，$L_i$为原型序列的句子长度，在这个过程中，被掩盖的位置遵照公式(2)，(3)，给其加上一个无穷大的负数偏置，之后计算隐状态h_t与所有原型序列的注意力[10]：

$$\alpha_{i,j} = \frac{e^{(h_t^T W_m s_{i,j} + \beta c_{s_i})}}{\sum_{i=1}^{M}\sum_{k=1}^{L_i} e^{(h_t^T W_m s_{i,k} + \beta c_{s_i})}} \quad (9)$$

$\alpha_{i,j}$计算的是隐状态h_t与第 i 个原型序列句子中第 j 个子词的注意力，同时，还需要考虑到每一个原型序列对模型的参考价值c_{s_i}。W_m是一个维度变换矩阵，β为一个可训练权重参数，用来权衡参考价值c_{s_i}的影响。利用$\alpha_{i,j}$对原型序列s_i的每一个词做加权平均：

$$c_t = W_c \sum_{i=1}^{M}\sum_{j=1}^{L_i} \alpha_{i,j} s_i \quad (10)$$

整个过程中，被掩盖的位置将不会参与运算。用c_t更新隐状态后，综合公式(8)计算出下一个子词的预测概率值：

$$p(y|.) = (1-\lambda_t)P_v(y_t) + \lambda_t \sum_{i=1}^{M}\sum_{j=1}^{L_i} \alpha_{i,j} 1_{s_{ij}=y_t} \quad (11)$$

其中，$1_{s_{ij}}$为指示函数，λ_t是一个由前馈网络构成的门控单元，用以平衡原型序列编码器的影响。

3 实验设计与分析

3.1 实验数据

本实验采用了规模为 12 万句对的汉越平行语料，其中测试语料为 0.2 万句对，验证语料为 0.1 万句对；由于翻译模型是单向的，为了验证模型的泛化能力，本实验还采用了 20.7 万的英越平行语料，测试语料为 0.3 万句对，验证语料为 0.2 万句对；候选句子库为 30 万的越南语单语句子。在训练之前对实验数据进行过滤乱码和分词处理，其中汉语使用 Jieba 分词工具，英语及越南语使用 Subword-nmt 工具，构建词典时使用 Underthesea-Vietnam NLP 工具，词典包含 14000 个实体及数字。

3.2 实验设置

为了评估本文方法的有效性，实验选取了 4 个基线系统：基于 OPENNMT 框架的 Transformer[10]模型、基于源语言相似度检索原型序列[13-14]的 NMT(Source similarity1，Source similarity2)模型、DengcCai[9]等提出的基于跨语言相似度检索原型序列的 NMT(Cross-lingual)模型，我们还将原型处理模块加在 NMT(Source similarity2)模型当中作为对比。

Transformer、NMT(Source similarity1，Source similarity2)、NMT(Cross-lingual)以及本文的模型都使用相同的汉越、英越平行语料作为训练集。NMT(Source similarity1，Source similarity2)中，汉语到越南语使用训练集本身作为候选句子库，英语到越南语使用额外的 30 万英越平行语料作为候选句子库。

本文的模型各个模块使用的是 Transformer 的模块，头数为 8，前馈层维度为 1024，词嵌入的维度为 512，检索模块编码器的隐藏层层数为

3，原型编码器的隐藏层层数为 4，生成模块编码器解码器的隐藏层层数都是 6，学习率设置为 0.1，dropout 设置为 0.3，优化器选用了 Adam，其参数设置为$\beta_1=0.9, \beta_2=0.98$，$\varepsilon=10^{-9}$，检索时相关句子的数目为 3.

3.3 实验结果与分析

3.3.1 对比试验

本文采用双语互译评估(BLEU)值作为评价指标，表 2 给出的是四个基线模型以及本文模型在汉越以及英越数据集上的 BLEU 值对比结果，这里使用“DP”(Denoising of Prototype)表示原型序列的降噪策略。

表 2 在汉越及英越数据集上的实验结果(单位：BLEU)

Tab.2 Comparison of BLEU values of Chinese-Vietnamese and English-Vietnamese corpus

方法类别	模型	BLEU(%)			
		汉语—越南语		英语—越南语	
		dev	test(本文方法Δ)	dev	test(本文方法Δ)
现有方法	Transformer	19.74	20.51 (+1.67)	21.98	22.78 (+2.75)
	NMT(Source similarity1)	18.13	19.02 (+3.16)	21.43	22.97 (+2.56)
	NMT(Source similarity2)	18.07	18.93 (+3.25)	22.71	24.51 (+1.02)
	NMT(Cross-lingual)	20.97	21.71 (+0.47)	23.37	24.94 (+0.59)
本文方法	NMT(Source similarity2 + DP)	18.15	18.34 (+3.84)	22.89	24.83(+0.70)
	NMT(DP)	**21.43**	**22.18**	**24.13**	**25.53**

表 2 为不同模型在汉越、英越数据集上的进行翻译时的 BLEU 值对比，可以看出：

（1）在汉越数据集的验证集以及测试集上，本文方法均取得了最佳效果。在测试集，相较于 Transformer、基于源语言相似度检索的 NMT(Source similarity1)、NMT(Source similarity2)分别提升了 1.67、3.16 以及 3.25 个百分点；相比于基于跨语言相似度检索的 NMT(Cross-lingual)提升了 0.47 个百分点。在验证集，相较于四个基线模型分别提升了 1.69，3.30，3.36 以及 0.46 个百分点。

（2）值得注意的是，汉越翻译任务上，基于源语言相似度检索的 NMT(Source similarity1、Source similarity2)效果不佳，分析可能的原因是该类模型检索依赖较大规模平行句对构成的候选句子库，而由于资源匮乏，本实验将训练数据集本身作为候选句子库，导致检索不到有用的原型序列，甚至引入了噪声，致使实验的结果较差。在此基础上对原型序列做降噪处理，可能无法改善原型序列的质量，使得结合源语言相似度检索方法以及本文方法的 NMT(Source similarity2 + DP)实验效果无法提升。

（3）在英越数据集的验证集以及测试集上，本文方法同样均取得了最佳效果。在英越翻译任务上，我们引入了额外的平行句对扩充了候选句子库，将本文方法结合基于源语言相似度检索的 NMT(Source similarity2 + DP)相较于只基于源语言相似度检索的 NMT(Source similarity2)在测试集上的效果提升了 0.32 个百分点，这表明在资源相对充足的情境下，本文的方法对基于源语言相似度检索的方法也是有效的。

（4）在英越数据集的测试集上，NMT(Cross-lingual)效果优于 NMT(Source similarity2 + DP)有 0.11 个百分点；NMT(DP)效果优于 NMT(Source similarity2 + DP)有 0.7 个百分点，其原因可能是在中小规模的候选句子库中，基于跨语言相似度检索得到原型序列的质量较基于源语言相似度检索得到的更优质。

3.3.2 消融实验

为了验证原型处理模块的有效性，本文在汉越翻译任务上设置了多组消融实验。NMT(none)表示在本文翻译模型基础上去除整个原型处理模块；NMT(Similarity-score)表示在 NMT(none)基础上仅用相似性分数评估原型序列参考价值；NMT(Rare-words)表示在 NMT(none)基础上仅用稀有词词频评估原型序列参考价值，这里用(1- d_{s_i})来表示原型序列的参考价值；NMT(Entitiy-dictionary)表示在 NMT(none)基础上仅基于越南语实体词典对原型序列做降噪处理。

分析表 3 可知，即使未对原型序列做降噪处理，NMT(none)的表现依然要优于 Transformer，分析其原因可能是原型序列本身对翻译模型的指导作用。在经过相似度分数评参考价值、稀有

词词频评估参考价值、实体词典降噪、整个原型序列处理模块降噪之后，相较于 NMT(none)分别提升了 0.24、0.12、0.29 以及 0.69 个百分点，其中， NMT(Rare-words)较 NMT(none)提升 0.12 个百分点，其原因可能是稀有词涵盖面较小，产生的特征差异有限。但是依然可以看出原型序列处理模块在消除噪声实体、通过稀有词及相似性分数赋予原型序列参考价值这两种降噪方式的有效性。

表 3 原型处理模块对模型的影响

Tab.3 The impact of prototype processing module

模型	BLEU(%)
	汉语-越南语
	test
Transformer	20.51
NMT(none)	21.49
NMT(Similarity-score)	21.73
NMT(Rare-words)	21.61
NMT(Entitiy-dictionary)	21.78
NMT(DP)	22.18

3.3.3 实例分析

表 4 给出的是基线系统与 NMT(DP)在汉语到越南语方向的翻译对比示例。

从表 4 的第一组数据可以看出，Transformer 模型的译文出现了不准的现象，将小儿麻痹“bại liệt”错译为“triệu chứng nhỏ”，并且句子语义不连贯。出现此问题的原因是小儿麻痹“bại liệt”这个词在训练语料中出现次数较少，Transformer 模型无法有效学习到该词的语义信息；NMT(Cross-lingual)模型相比于 Transformer 语义更加连贯，但将人名：布鲁斯“Bruce Aylward”漏译。原因可能是模型训练过程中原型序列所包含的人名给 NMT 模型带来了干扰。本文基于降噪原型序列的翻译方法，使得 NMT 模型在解码端能够更好地学习到目标端语言的语义信息，因此 NMT(DP)模型取得了更好的翻译效果。

以上示例说明，本文方法虽然仍存在着翻译不充分的问题，但相较于基线模型能产生更加优质的译文。

表 4 不同模型的译文实例

Tab.4 Translation examples of different models

来源	例句
源语言句子	布鲁斯 • 艾尔沃德 : 我们 怎样才能 把 小儿麻痹症 治 好
参考译文	Bruce Aylward : Chúng ta sẽ ngăn chặn bệnh bai liệt như thế nào
Transformer	Bruce : Ta sẽ chữa triệu chứng nhỏ. (布鲁斯: 我会 对待 症状 小的)
NMT(Cross-lingual)	Làm sao chữa được bai liệt. (如何做 治愈 瘫痪症状)
NMT(DP)	Bruce : Chúng ta sẽ sao chữa bai liệt. (布鲁斯：我们将 如何 对待 瘫痪症状)

4 结论

本文提出了一种基于降噪原型序列的汉越神经机器翻译方法，该方法在利用越南语单语数据缓解双语资源匮乏的同时，将原型序列进行噪声过滤并将其中对机器翻译有益的知识特征进行增强，在汉-越翻译任务上的实验结果表明，论文提出方法有效提高了模型的翻译性能，进一步的英-越翻译任务的实验结果证明了提出方法的鲁棒性。然而，本文使用现有工具构建的实体词典存在少部分识别误差，在未来工作中考虑用更精准的实体词典来提升本文模型的性能。

参考文献：

[1] 刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展，2017：1144-1149.

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

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. Proceedings of the 6th International Conference on Learning Representations. Piscataway:IEEE, 2015:1-15.

[4] Jiatao Gu,Yong Wang,et al.Search engine guided neural machine translation [J].In AAAI, 2018: 5133-5140.

[5] Yiren Wang,Yingce Xia,et al. Neural Machine Translation with Soft Prototype [C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems,2019.

[6] Qian Cao,Deyi Xiong.Encoding Gated Translation Memory into Neural Machine Translation [C]. 2018 Conference on Empirical Methods in Natural Language Processing,2018: 3042-3047.

[7] Platanios E A ,Stretcu O,Neubig G , et al.Competence-based Curriculum Learning for Neural Machine Translation[C].2019 North American Chapter of the Association for Computational Linguistics ,2019.

[8] Xie Q,Dai Z,Hovy E,et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems, 2020: 6256-6268.

[9] Cai D,Wang Y,Li H,et al. Neural Machine Translation with Monolingual Translation Memory[C].Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021:7307-7318.

[10]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Proceedings of the 2017 Neural Information Processing Systems. Cambridge: MIT Press, 2017: 5998-6008.

[11]Menghao Guo,Tianxing Xu,Jiangjiang Liu, et al. Attention Mechanisms in Computer Vision: A Survey [J]. Comp. Visual Media 8, 2022: 331-368.

[12]Shrivastava, A. , and P. Li . Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS).MIT Press MIT Press, 2014:2321-2329.

[13]Jingyi Zhang, Masao Utiyama, Eiichro Sumita, et al. Guiding neural machine translation with retrieved translation pieces. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1325-1335.

[14]Mengzhou Xia, Guoping Huang, Lemao Liu, et al. Graph based translation memory for neural machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 7297-7304.

Chinese-Vietnamese Neural Machine Translation Based on Denoised Prototype

Hanqing Yang[1,2]， Hua Lai[1,2]， Zhiqiang Yu[1,2*]， Zhengtao Yu[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500,China; 2. Key Laboratory of Artificial Intelligence of Yunnan Province, Kunming 650500,China)

Abstract: Prototype is designed to guide neural machine translation with target-side language information. The existing work mainly uses similarity translation as the target prototype to improve the performance of neural machinetranslation in resource rich scenarios. However, in the context of resources like Chinese to Vietnamese, due to the lack of parallel corpus resources, the prototype contains a large amount of information. Direct use will increase the difficulty of translation model training, and even introduce noising information. To solve this problem,a Chinese-Vietnamese Neural Machine Translation Based on Denoised Prototype is proposed. First, prototype are obtained by cross language retrieval; secondly, the noising information in the prototype is masked based on an entity dictionary, and then the reference value of the prototype is obtained by synthesizing the word frequency and semantic similarity of sentences in the prototype; finally, an additional encoder is used to receive the prototype and allow the decoder to establish an attention mechanism with the encoders. The experimental results show that the proposed method can effectively improve the performance of Chinese-Vietnamese neural machine translation compared with the baseline models.

Keywords: Chinese－Vietnamese neural machine translation;low resource;prototype;denoising