

# 面向维汉神经机器翻译的双向重排序模型分析

张新路<sup>1,2,3</sup> 李晓<sup>1,2,3,†</sup> 杨雅婷<sup>1,2,3</sup> 王磊<sup>1,2,3</sup> 董瑞<sup>1,2,3</sup>

1. 中国科学院新疆理化技术研究所, 乌鲁木齐 830011; 2. 中国科学院大学, 北京 100049; 3. 新疆民族语言  
语言信息处理实验室, 乌鲁木齐 830011; † 通信作者, E-mail: xiaoli@ms.xjb.ac.cn

**摘要** 在维吾尔语到汉语等低资源语料库上, 神经机器翻译的拟合训练容易陷入局部最优解, 导致单一模型的翻译结果可能不是全局最优解。针对此问题, 通过集成策略, 有效整合多个模型预测的概率分布, 将多个翻译模型作为一个整体; 同时采用基于交叉熵的重排序方法, 将具有相反解码方向的翻译模型相结合, 最终选出综合得分最高的候选翻译作为输出。在 CWMT2015 维汉平行语料库上的实验结果表明, 与单一的 Transformer 模型相比, 改进后的方法提升 4.82 个 BLEU 值。

**关键词** 神经机器翻译; 集成学习; 双向重排序; 维吾尔语

## Analysis of Bi-directional Reranking Model for Uyghur-Chinese Neural Machine Translation

ZHANG Xinlu<sup>1,2,3</sup>, LI Xiao<sup>1,2,3,†</sup>, YANG Yating<sup>1,2,3</sup>, WANG Lei<sup>1,2,3</sup>, DONG Rui<sup>1,2,3</sup>

1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011; 2. University of Chinese Academy of Sciences, Beijing 100049; 3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011;  
† Corresponding author, E-mail: xiaoli@ms.xjb.ac.cn

**Abstract** The fitting training of neural machine translation is easy to fall into a local optimal solution on a low-resource corpus such as Uyghur to Chinese, resulting in the translation result of a single model may not be a global optimal solution. In order to solve this problem, the probability distribution predicted by multiple models is effectively integrated through the ensemble strategy, and multiple translation models are taken as a whole. At the same time, the translation models with opposite decoding directions are integrated by the reordering method based on cross entropy, and the candidate translation with the highest comprehensive score is selected as the output. The experiment on CWMT2015 Uyghur-Chinese parallel corpus shows that proposed method has 4.82 BLEU values improvement compared with a single transformer model.

**Key words** neural machine translation; ensemble learning; bi-directional reranking; Uyghur

神经机器翻译(neural machine translation, NMT)是一种基于深度学习的机器翻译方法, 在获得大规模平行语料的同时, 可以得到较好的翻译性能。近年来, 随着深度学习技术的发展, 神经机器翻译极大地提高了机器翻译的质量, 在许多语言对上显示出最好的效果<sup>[1]</sup>。神经机器翻译的核心思想是建立一个基于神经网络的 Encoder-Decoder 模型, 通过将源语言句子编码为一个稠密向量, 然后从该向量

解码出目标语言句子, 从而建立源语言与目标语言的映射关系<sup>[2]</sup>。传统的神经机器翻译是将循环神经网络<sup>[3]</sup>(RNN)或卷积神经网络<sup>[4]</sup>(CNN)作为 Encoder-Decoder 模型的基础。最近, Vaswani 等<sup>[5]</sup>提出一种完全基于注意力机制的神经机器翻译模型 Transformer, 该模型在 WMT2018 测评任务中取得单一模型最好的结果。

尽管 Transformer 模型在很多语言对上取得最

好的翻译效果,但在维吾尔语到汉语这种低资源语料库上,神经网络的拟合训练很容易陷入局部最优解,最终单一模型的翻译结果可能不是全局最优解。因此,研究人员在 WMT 和 CWMT 等机器翻译评测任务中,通过整合不同模型的预测结果来提升机器翻译的性能<sup>[6-7]</sup>。Vaswani 等<sup>[5]</sup>提出使用检查点平均法来获得更低的方差和更稳定的翻译结果。Sennrich 等<sup>[8]</sup>首先在 WMT16 测评任务中应用检查点集成的方法,然后在 WMT17<sup>[6]</sup>测评任务中尝试独立集成的方法,并取得显著的改善。李北等<sup>[9]</sup>分析了集成策略对于神经机器翻译的影响。以上研究在使用集成策略解码时,大多使用同一个方向的翻译模型,通过束搜索算法得到概率最大的翻译候选,不能充分地使用每一个翻译模型内部的信息。

为了获得更好的翻译输出,研究人员尝试了各种方法。Shen 等<sup>[10]</sup>提出两种新的基于感知器的重排序算法,提高了机器翻译的质量。Kumar 等<sup>[11]</sup>提出一种用于统计机器翻译的最小贝叶斯风险的解码方法,将预期的翻译错误损失降到最低,从而提升了翻译质量。这些方法主要应用于统计机器翻译。Zhou 等<sup>[12]</sup>使用不同的系统,得到有差异性的翻译结果,通过系统融合的方式,利用混淆网络来重构翻译结果。

本文基于 Transformer 结构<sup>[5]</sup>,通过设置不同的随机初始化种子,生成多个正向翻译模型(从左到

右)和逆向翻译模型(从右到左)。图 1 展示基本的双向重排序模型结构。对于输入句子  $X$ , 为了得到更好的翻译列表,我们使用集成策略对多个正向模型集成翻译产生 N-best 列表。计算候选翻译项对应于集成的每一个翻译模型的交叉熵以及每一个反向翻译模型对于 N-best 列表的交叉熵,将不同的模型对当前翻译候选的交叉熵得分求和。由于神经机器翻译在解码时生成的序列概率是多个条件概率的乘积,所以生成的句子序列越长,概率值相乘的结果就越小,更倾向于生成短序列。因此,需对短句子进行基于句子长度的惩罚,降低生成短序列的倾向。这种方式既可以充分利用集成的优势,又可以有效地整合不同方向的翻译模型对候选翻译的评估。

## 1 神经机器翻译

神经机器翻译是一种使用深度神经网络获取自然语言之间的映射关系的方法,通常采用 Encoder-Decoder 架构,通过编码器,将给定的源语句句子  $X=(x^1, x^2, \dots, x^n)$  编码为中间向量  $Z$ , 解码器根据中间向量产生目标语言句子  $Y=(y^1, y^2, \dots, y^n)$ 。对编码器和解码器进行联合训练,使给定源序列的目标序列的条件概率最大化:

$$P(Y|X; \theta) = \prod_{j=1}^N P(y_j | y_{<j}, x; \theta). \quad (1)$$

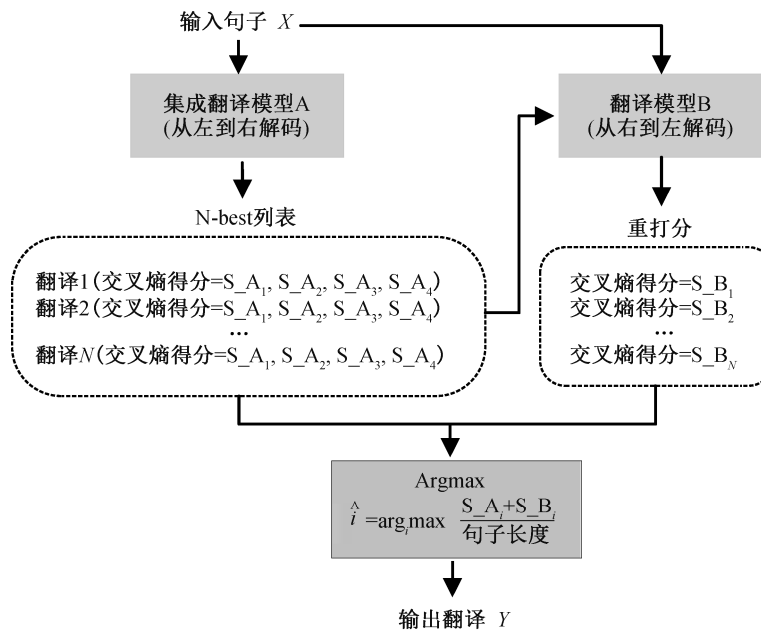


图 1 双向重排序模型示意图

Fig. 1 Schematic diagram of bidirectional reranking model

在 Transformer 模型提出之前,大多数神经机器翻译模型都采用基于注意力机制<sup>[1]</sup>的循环神经网络(RNN)。这种方式在一些任务上取得不错的效果,但由于 RNN 的序列特性,导致在训练的过程中难以并行化,对于长距离和层次化的依赖关系难以建立,使得训练过程十分漫长,从而影响翻译质量。

Transformer 摒弃了传统的循环神经网络的序列结构,采用完全基于注意力机制的结构。在保证模型并行化的同时,也改善了模型的表示能力,准确性也有一定程度的提升<sup>[5]</sup>。

Transformer 同样采用 Encoder-Decoder 的架构,它由  $N$  个堆叠的编码器和解码器层组成。采用全新的注意力机制: Encoder 端的自注意力机制、Decoder 端的自注意力机制以及编码端-解码端的自注意力机制。如图 2 所示,Transformer 模型没有使用任何循环神经网络,为了获得输入序列的顺序特征,需要在词向量的基础上加入位置编码信息。

编码器由  $N$  个相同的层组成,图 2 展示两个 Encoder 层的堆叠。每一层包含两个子层:多头自注意力机制和传统的前馈神经网络。框内的虚线表示直连接网络,是为了在较深的神经网络中减少信

息的损失以及加速模型的收敛。使用残差链接和层正则化<sup>[13]</sup>来保证梯度的传递的稳定。解码器与编码器类似,解码的过程中只能看到已经输出的信息,需要将未输出部分进行遮掩,因此使用带遮掩的多头注意力机制,同时增加一个负责处理编码器输出的多头注意力机制。多头注意力机制用于从不同位置的不同表示子空间获取信息,其基本单元是缩放的点积注意力模型。每一个头对应一个点积注意力模型,计算公式如下:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (2)$$

其中,  $\mathbf{Q}$  表示查询矩阵,  $\mathbf{K}$  表示键值矩阵,  $d_k$  表示  $\mathbf{K}$  的维度,  $\mathbf{V}$  表示权重矩阵。多头自注意力机制就是采用多组  $\mathbf{Q}$ ,  $\mathbf{K}$  和  $\mathbf{V}$  得到不同的点积自注意力输出,最后将这些输出连接起来作为最终的输出结果。

## 2 方法

### 2.1 集成学习策略

集成学习是一种联合多个学习器进行协同决策的机器学习方法<sup>[14]</sup>,通过整合多个学习器的决策结果,有效地减小预测结果的方差和偏差<sup>[15]</sup>,可以显

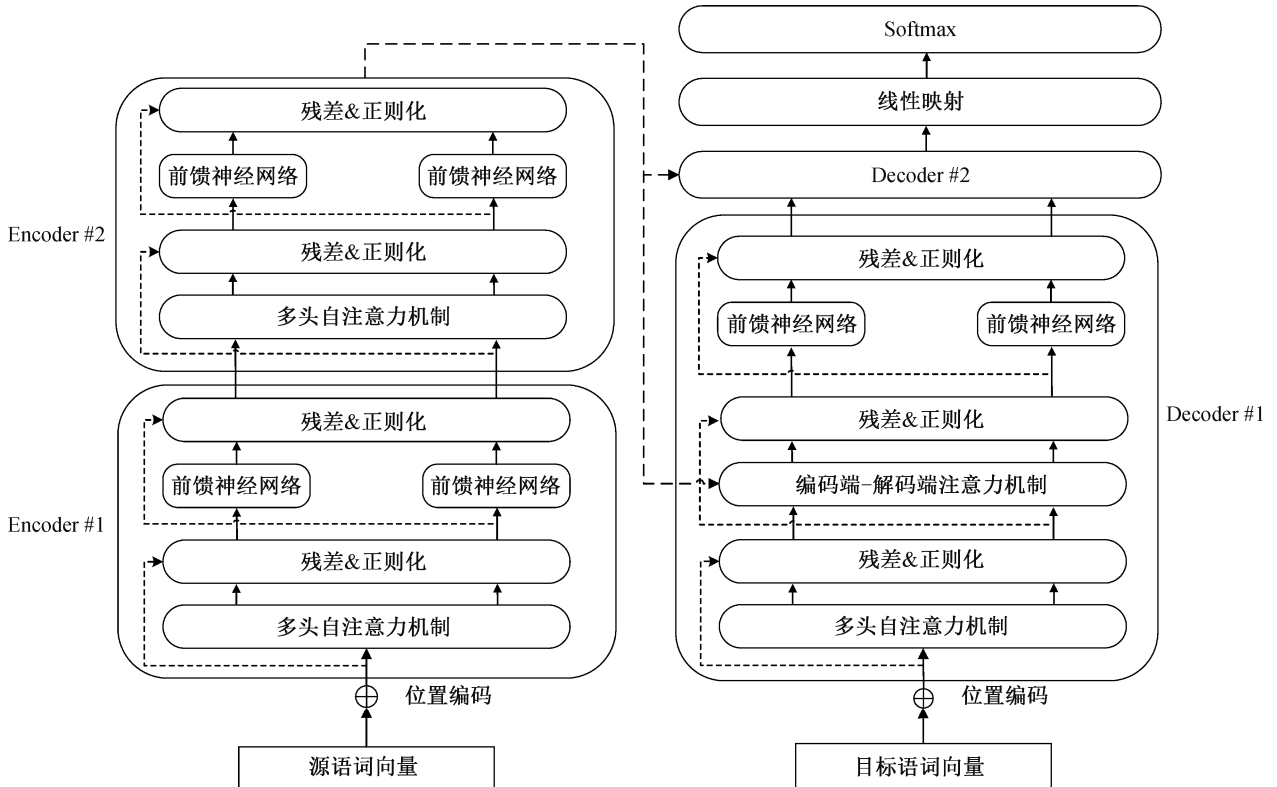


图 2 Transformer 模型示意图  
Fig. 2 Transformer model diagram

著地提升模型的泛化能力<sup>[16]</sup>, 获得比单学习器更好的效果。近年来, 集成学习方法在机器翻译领域取得了较好的效果<sup>[17-19]</sup>。

机器翻译是一种序列生成任务, 在解码时, 每一个时序的输出都依赖于前一个时序的输出。模型会根据当前的语义信息, 计算得到概率分布向量(维度为词表大小), 经过 Softmax 操作, 得到归一化的向量表示, 向量中的每一个元素指代预测下一个词的概率。如图 3 所示, 机器翻译的集成学习策略是在解码的过程中通过整合不同模型得到的概率分布, 从而获得新的解, 进而预测下一个目标端词语。单一模型生成目标端单词的选择如式(3)所示, 集成学习方法生成单词的选择如式(4)所示。

$$\hat{y}_t = \operatorname{argmax} \log P(y_t | y_{1:t-1}, x; M), \quad (3)$$

$$\hat{y}_t = \operatorname{argmax} \frac{1}{J} \sum_{j=1}^J \log P(y_t | y_{1:t-1}, x; M_j), \quad (4)$$

其中,  $y_t$  表示在第  $t$  个位置将要输出的单词,  $y_{1:t-1}$  表示从句首到  $(t-1)$  位置输出的单词序列。 $x$  表示输入单词序列。 $M$  表示翻译中的解码模型( $M_j$  表示第  $j$  个翻译模型),  $J$  表示集成的翻译模型数量。

集成策略也存在一些限制。首先, 所有的翻译模型都必须使用同一个目标词汇表, 因为每一个单词的输出概率都是在同一个词表规模上不同模型的平均值。其次, 所有模型的解码方向必须一致, 这是由于在得到当前时刻的输出单词后, 需采用束搜索的方式生成候选翻译, 只有模型的解码方向一致, 才能生成更好的候选翻译。

## 2.2 重排序策略

在解码的过程中, 神经机器翻译会根据束搜索

算法产生 N-best 的翻译列表, 但有时概率最大的翻译结果不一定是最佳翻译。如何通过重排序的方式在规模为  $N$  的候选翻译中找到最优翻译结果十分重要<sup>[20]</sup>。

本文通过计算交叉熵的方式来评估翻译列表中译文的质量。交叉熵<sup>[21]</sup>的定义如下:

$$H(p, q) = -\sum_{i=1}^n p(x_i) \log(q(x_i)), \quad (5)$$

其中,  $x_1^n = x_1, x_2, \dots, x_n$  为源语言的句子,  $p(x_i)$  为源语言词  $x_i$  对应的候选翻译词的概率分布,  $q(x_i)$  为翻译模型  $q$  翻译得到的候选翻译列表中对应该词的概率。由于交叉熵越小模型的表现越好, 为了用分数来衡量模型的质量, 我们使用交叉熵的相反数作为候选翻译的最终得分。

重排序策略可以应用于相同目标语言的任意模型, 因此通过集成策略改善 N-best 翻译列表的质量, 使用重排序方法选出更优的翻译候选。

## 2.3 双向重排序模型

为了更好地利用集成策略和重排序策略的优势<sup>[22]</sup>, 本文将两种方式相结合, 称为双向重排序模型。通过维汉神经机器翻译中的实例, 说明双向重排序模型的有效性。

给定以下源语言和目标语言。

源语言: چوڭ ھالقىغان چېگرا دىرياسى يالوچىياڭ  
قەدەم نەھمىيەتلىك قۇرۇلۇشلاردا قاتارلىق كۆۋرۈكى  
بىيىسىلدى.

参考译文: 鸭绿江跨境大桥等建设迈出了有意义的步伐。

对于以上实例, 我们通过 4 个正向模型的集成翻译, 根据对数似然概率产生 12 个翻译候选, 对数

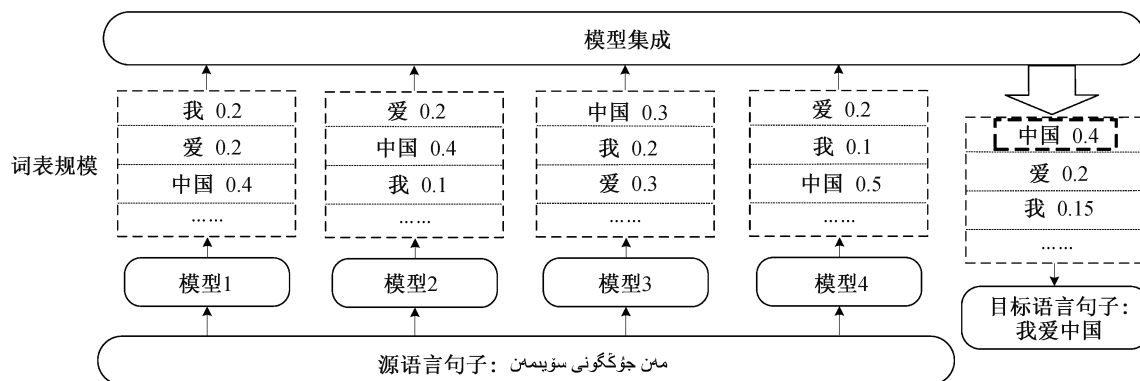


图 3 模型集成示意图

Fig. 3 Schematic diagram of model ensemble

似然概率对应图4中最后一列。计算列表中每一个候选翻译对应的集成翻译模型的交叉熵得分F0~F3, 然后依次使用4个逆向翻译模型, 计算列表中的交叉熵得分R2L0~R2L3。将所有交叉熵得分的和与基于翻译长度的惩罚因子的比值作为最终的结果, 将得分最高的翻译候选作为最终的翻译输出。通过这种方法, 选出与参考译文完全一致的候选翻译6, 而不是对数似然最大的候选翻译1。

### 3 实验

#### 3.1 数据集

我们在CWMT2015提供的维汉双语平行语料上进行实验, 验证基于Transformer的双向重排序模型的有效性。首先, 对语料进行乱码过滤, 剔除混有乱码的语料。然后, 在训练集中过滤掉与开发集测试集相同的句对, 并对汉语进行基于字级别的处理。最终训练集包含维-汉平行句对336224句, 开发集包含700句, 测试集包含1000句。

我们从训练集中学习得到字节对编码(BPE)的规则<sup>[23]</sup>, 其中设置BPE词表的大小为32000。将学习到的规则应用于所有训练集、开发集和测试集, 这样可以有效地减少未登录词出现的频次, 提升翻

译的效果。根据经过BPE处理的源语言和目标语言的训练集, 生成联合词汇表, 将词汇表规模设置为36 K, 将训练集中的句子长度限制为100。采用基于Moses的multi-bleu-detok.perl的脚本<sup>[24]</sup>来计算BLEU值, 通过BLEU值来衡量翻译的质量。

#### 3.2 实验设置

我们使用Marian作为实验的模型框架<sup>[25]</sup>。使用Transformer模型作为实验的基线模型, 参数采用Transformer\_base的设置。编码端和解码端的层数都是6层, 词向量的维度为512, 前馈神经网络为2048维, 学习率为0.0003, 采用5个K80GPU进行训练。采用Adam算法作为优化算法, 通过AAN网络<sup>[26]</sup>来加速Transformer的解码过程, 将Swish作为激活函数。在训练中使用早停机制, 参数设置为5, 在整个训练语料上进行108轮。解码的过程中, 我们使用Beam-search策略, beam-size的大小设置为12。基于该参数的单模型, 在测试集上的BLEU值为50.17。

#### 3.3 N-best 的最优候选

我们采用束搜索的方法生成N-best翻译候选。通常, 束搜索的大小会大于或者等于候选翻译的规

1 鸭绿江跨境大桥等建设也迈出了意义。	F0=-9.58721, F1=-8.36216, F2=-5.59128, F3=-7.11541, R2L0=-25.8942, R2L1=-40.0716, R2L2=-28.2416, R2L3=-27.9918     -5.41194
2 鸭绿江跨境大桥等建设迈出了意义。	F0=-7.21009, F1=-3.73177, F2=-8.79401, F3=-10.2059, R2L0=-20.446, R2L1=-33.5099, R2L2=-24.0751, R2L3=-22.4915     -5.47026
3 鸭绿江跨境大桥等工程迈出了意义。	F0=-8.74809, F1=-6.28332, F2=-6.7897, F3=-11.0161, R2L0=-21.5619, R2L1=-30.7331, R2L2=-24.1634, R2L3=-23.1263     -5.99926
4 鸭绿江跨境大桥等建设也迈出了有意义的步伐。	F0=-12.4652, F1=-11.9089, F2=-7.39339, F3=-7.23471, R2L0=-10.6006, R2L1=-19.6397, R2L2=-11.893, R2L3=-10.527     -6.10431
5 鸭绿江跨境大桥等建设都迈出了意义。	F0=-7.53811, F1=-7.5062, F2=-8.72646, F3=-11.0794, R2L0=-23.5588, R2L1=-38.4265, R2L2=-26.7558, R2L3=-26.9881     -6.15236
6 鸭绿江跨境大桥等建设迈出了有意义的步伐。	F0=-9.98647, F1=-7.03542, F2=-10.6757, F3=-10.5992, R2L0=-5.74414, R2L1=-13.9064, R2L2=-7.13545, R2L3=-6.42047     -6.16356
7 鸭绿江跨境大桥等建设步伐有意义。	F0=-7.95083, F1=-4.76442, F2=-15.6098, F3=-6.14119, R2L0=-18.9959, R2L1=-29.2258, R2L2=-20.7198, R2L3=-20.2328     -6.29686
8 鸭绿江跨境大桥等建设工程迈出了意义。	F0=-6.72428, F1=-8.14605, F2=-11.02, F3=-12.5719, R2L0=-23.7961, R2L1=-31.7659, R2L2=-27.447, R2L3=-24.5967     -6.57328
9 鸭绿江跨境大桥等建设工程迈出了有意义的步伐。	F0=-8.84395, F1=-11.6735, F2=-12.026, F3=-11.3724, R2L0=-9.17272, R2L1=-11.8658, R2L2=-11.3608, R2L3=-7.81557     -6.69245
10 鸭绿江跨境大桥等建设项目迈出了意义。	F0=-10.0848, F1=-9.83906, F2=-12.2144, F3=-10.6601, R2L0=-26.4235, R2L1=-36.3011, R2L2=-31.7313, R2L3=-27.369     -7.31432
11 鸭绿江跨境大桥等建设也迈出了意义的步伐。	F0=-12.3255, F1=-14.0528, F2=-8.69434, F3=-10.5707, R2L0=-18.4994, R2L1=-21.2049, R2L2=-19.1643, R2L3=-18.9075     -7.34592
12 鸭绿江跨境大桥等建设迈出了有意义的步骤。	F0=-14.7599, F1=-12.0839, F2=-10.7454, F3=-15.0582, R2L0=-12.1697, R2L1=-22.3832, R2L2=-12.4494, R2L3=-15.2642     -8.47315

图4 多模型集成的双向重排序实例

Fig. 4 An example of bi-directional reranking for multi-model ensemble

模  $N$ 。本文设置束搜索的大小与  $N$ -best 相等。在生成翻译时,对  $N$ -best 中每一个候选翻译进行交叉熵打分,选择得分最高的候选翻译作为最终的翻译输出。图 5 展示不同的翻译模型在不同  $N$ -best 规模下 BLEU 值的变化,包含从左到右的正向翻译模型、从右到左的逆向翻译模型以及双向重排序模型。本节使用的模型都是单一的翻译模型,没有使用集成的方法来生成翻译。

从图 5 可以看出,从左到右的正向翻译模型,在  $N$ -best 为 9 时, BLEU 值最高(50.49)。当  $N$ -best 的规模达到 12 时, BLEU 值趋于平稳。从右到左的逆向翻译模型,在  $N$ -best 为 8 时, BLEU 值最高(50.74)。当  $N$ -best 的规模达到 11 时, BLEU 值趋于稳定。双向重排序模型,随着  $N$ -best 的增大, BLEU 值也逐渐增大,在  $N$ -best 为 11~15 时, BLEU 值趋于稳定,在  $N$ -best 为 15 时最高(52.03),表明随着翻译候选列表的增大,双向重排序模型能够较好地选择最佳翻译候选。基于以上的实验结果,下面的实验中将  $N$ -best 设置为 12。

### 3.4 多模型集成的双向重排序模型

为了验证重排序策略的有效性,我们在从左到右的 4 个模型(L2R)和从右到左的 4 个模型(R2L)中分别对比  $N$  为 12 时重排序策略和最大概率的翻译质量。从表 1 可以看出,除第 3 个正向翻译模型中最大概率的策略高于重排序策略外,其余重排序模型的翻译质量都高于最大概率, BLEU 值平均提升 0.33。

图 6 展示从左到右、从右到左的解码以及双向重排序多模型集成的 BLEU 值与集成模型数量的关系。在双向重排序模型中,我们进行以下两组实验。

1) 将从左到右的正向模型作为翻译模型,将从

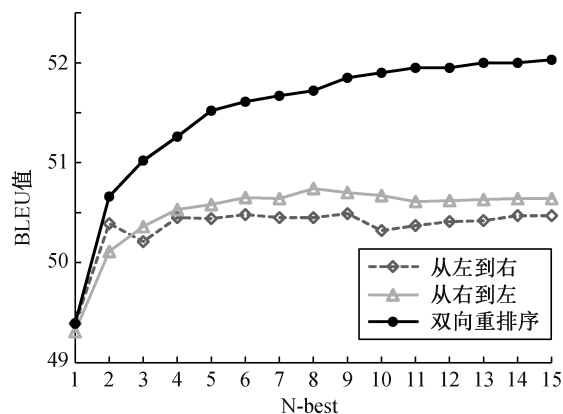


图 5 BLEU 值与  $N$ -best 的关系

Fig. 5 Relationship between BLEU value and  $N$ -best

表 1 重排序策略对 BLEU 值的影响  
Table 1 Effect of reranking strategy on BLEU value

模型名称	最大概率	重排序
L2R_1	50.17	50.41
L2R_2	50.92	51.44
L2R_3	51.04	50.99
L2R_4	50.34	50.84
R2L_1	50.33	50.62
R2L_2	49.69	50.20
R2L_3	50.67	51.13
R2L_4	49.85	50.10

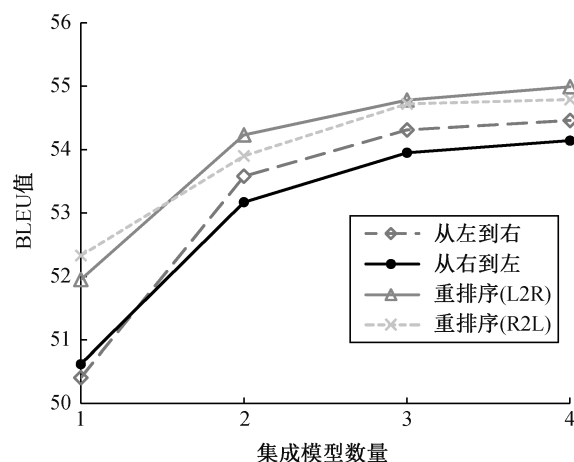


图 6 BLEU 值与集成模型数量的关系

Fig. 6 Relationship between BLEU value and number of ensemble models

右到左的逆向模型作为重排序模型(重排序(L2R))。

2) 将从右到左的正向模型作为翻译模型,将从左到右的逆向模型作为重排序模型(重排序(R2L))。

从图 6 可以得到如下结论。

1) 4 个翻译模型的 BLEU 值都随模型数量的增加而增加,但增长速度变慢。集成 4 个翻译模型时,翻译质量达到最优, BLEU 值分别为 54.99, 54.79, 54.46 和 54.14。双向重排序多模型集成的最好的结果比从左到右的正向单一翻译模型(L2R\_1)的 BLEU 值提升 4.58, 比从右到左的逆向翻译模型(R2L\_1)提升 4.37。

2) 从左到右的集成翻译效果比从右到左的集成翻译效果稍有提高, BLEU 值平均高 0.22。说明翻译质量不仅与数据集和语言对相关,也随解码方向的不同而变化。

3) 重排序后模型的翻译效果显著高于单向的多模型集成,说明重排序策略可以更好地选择最佳

候选翻译。为了进一步验证重排序方法对集成翻译模型提升的效果, 我们进行如下对比实验。

集成了从左到右的4个正向翻译的模型后, 我们依次加入逆向的重排序策略, 结果如表2所示。可以看出, 随着重排序模型数量的增加, BLEU值也随之提高, 与没有逆向重排序策略的集成模型相比, BLEU值提高0.53, 说明逆向重排序策略可以有效地提升集成翻译的质量。

表2 重排序策略对集成翻译的影响

Table 2 Impact of reranking strategies on ensembled translation

系统名称	系统设置	模型数量	BLEU
Ensem	L2R_1, L2R_2, L2R_3, L2R_4	4	54.46
Com-1	Ensem + R2L_1	5	54.85
Com-2	Com-1 + R2L_2	6	54.98
Com-3	Com-2 + R2L_3	7	54.94
Com-4	Com-3 + R2L_4	8	54.99

4) 两个双向重排序模型的翻译效果差别不大, 正向翻译模型比逆向翻译模型的 BLEU 值高 0.2。我们认为这是由于4个正向翻译模型的 BLEU 值的均值比逆向翻译模型略高。这也从另一个角度说明提升单一模型的翻译效果能够在集成策略中更好地获得翻译候选, 从而提升翻译质量。

## 4 总结

本文通过集成策略和重排序策略, 提升维吾尔语到汉语等低资源语言对上的翻译质量。由于重排序方法可以将具有与集成不同属性的模型组合在一起, 所以我们将正向模型与反向模型结合起来进行重排序。通过集成和双向重排序, 获得比单独集成更高的翻译质量。在 CWMT2015 维汉测试集上的实验表明, 当 N-best 的规模为 12 时, 经过 4 个正向模型的集成翻译和 4 个逆向模型的重排序, 得到的 BLEU 值比基线系统提升 4.82。此外, 通过对比不同方向的翻译效果, 验证了解码的方向对翻译质量也有影响。

如果能进一步提高单个模型的质量, 则集成和重排序方法都可以进一步提高翻译质量。因此, 我们未来将采用更多更具有差异性的模型来提升集成翻译的效果, 同时将更多的重排序方法进行融合,

得到更优的翻译候选。

## 参考文献

- [1] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2019-02-01]. <https://arxiv.org/abs/1409.0473v4>
- [2] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks // Advances in Neural Information Processing Systems 27. Montreal, 2014: 3104-3112
- [3] Kalchbrenner N, Blunsom P. Recurrent continuous translation models // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, 2013: 1700-1709
- [4] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning // Proceedings of the 34th International Conference on Machine Learning-Volume 70. Sydney, 2017: 1243-1252
- [5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // Advances in Neural Information Processing Systems 30. Long Beach, 2017: 5998-6008
- [6] Sennrich R, Birch A, Currey A, et al. The University of Edinburgh's neural MT systems for WMT17 // Proceedings of the Second Conference on Machine Translation. Copenhagen, 2017: 389-399
- [7] Tan Z, Wang B, Hu J, et al. XMU neural machine translation systems for WMT 17 // Proceedings of the Second Conference on Machine Translation. Copenhagen, 2017: 400-404
- [8] Sennrich R, Haddow B, Birch A. Edinburgh neural machine translation systems for WMT 16 [EB/OL]. (2016-06-27) [2019-02-01]. <https://arxiv.org/abs/1606.02891>
- [9] 李北, 王强, 肖桐, 等. 面向神经机器翻译的集成学习方法分析. 中文信息学报, 33(3): 42-51
- [10] Shen L, Sarkar A, Och F J. Discriminative reranking for machine translation // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. Boston, 2004: 177-184
- [11] Kumar S, Byrne W. Minimum Bayes-risk decoding for statistical machine translation [R]. Baltimore: Johns Hopkins Univ Baltimore MD Center for Language and Speech Processing (CLSP), 2004

- [12] Zhou L, Hu W, Zhang J, et al. Neural system combination for machine translation [EB/OL]. (2017-04-21) [2019-02-01]. <https://arxiv.org/abs/1704.06393>
- [13] Lei Ba J, Kiros J R, Hinton G E. Layer normalization [EB/OL]. (2016-07-21) [2019-02-01]. <https://arxiv.org/abs/1607.06450>
- [14] Hansen L K, Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1990, 12(10): 993-1001
- [15] Dietterich T G. Ensemble methods in machine learning // *International workshop on multiple classifier systems*. Berlin: Springer, 2000: 1-15
- [16] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, 1999, 36(1/2): 105-139
- [17] Opitz D, Maclin R. Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 1999, 11: 169-198
- [18] Xiao T, Zhu J, Liu T. Bagging and boosting statistical machine translation systems. *Artificial Intelligence*, 2013, 195: 496-527
- [19] Liu Y, Zhou L, Wang Y, et al. A comparable study on model averaging, ensembling and reranking in NMT // *CCF International Conference on Natural Language Processing and Chinese Computing*. Hohhot, 2018: 299-308
- [20] Och F J, Gildea D, Khudanpur S, et al. A smorgasbord of features for statistical machine translation // *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Boston, 2004: 161-168
- [21] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2013
- [22] Imamura K, Sumita E. Ensemble and reranking: using multiple models in the NICT-2 neural machine translation system at WAT2017 // *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, 2017: 127-134
- [23] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [EB/OL]. (2016-06-10) [2019-02-01]. <https://arxiv.org/abs/1508.07909>
- [24] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, 2007: 177-180
- [25] Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, et al. Marian: fast neural machine translation in C++ [EB/OL]. (2018-04-04) [2019-02-01]. <https://arxiv.org/abs/1804.00344>
- [26] Zhang B, Xiong D, Su J. Accelerating neural transformer via an average attention network [EB/OL]. (2018-05-07) [2019-02-01]. <https://arxiv.org/abs/1805.00631>