面向维汉神经机器翻译的双向重排序模型分析

张新路[1, 2, 3] 李晓[1, 2, 3, †] 杨雅婷[1, 2, 3] 王磊[1, 2, 3] 董瑞[1, 2, 3]

1. 中国科学院新疆理化技术研究所，新疆 乌鲁木齐 830011； 2. 中国科学院大学，北京 100049； 3. 新疆民族语音语言信息处理实验室，新疆 乌鲁木齐 830011；

† 通信作者，E-mail: xiaoli@ms.xjb.ac.cn

摘要 神经机器翻译已经成为机器翻译的主流方法，但神经机器翻译的拟合训练在维吾尔语到汉语这种低资源语料库上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。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整合起来，选出综合得分最高的候选翻译作为输出。最终在 CWMT2015 维汉平行语料上，该方法相比于单一的 Transformer 模型有 4.82 个 BLEU 值的提升。

关键词 神经机器翻译；集成学习；双向重排序；维吾尔语

中图分类号 H085

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

ZHANG Xinlu[1,2,3], LI Xiao[1,2,3,†], YANG Yating[1,2,3], WANG Lei[1,2,3], DONG Rui[1,2,3]

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

Abstract Neural machine translation has become the mainstream method of machine translation, but 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 this paper, 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. Finally, on CWMT2015 Uighur-Chinese parallel corpus, this 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，在可获得大规模平行语料的情况下可以得到较好的翻译性能。近年来，随着深度学习技

新疆维吾尔自治区重点实验室开放课题（2018D04018）、国家自然科学基金（U1703133）、中国科学院青年创新促进会项目(2017472) 、新疆维吾尔自治区高层次人才引进工程项目(Y839031201)

术的发展，神经机器翻译也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[1]。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，从而建立源语言和目标语言的映射关系[2]。传统的神经机器翻译都是利用循环神经网络[3]（RNN）或者卷积神经网络[4]（CNN）作为 Encoder-Decoder 模型的基础。最近 Vaswani 等人[5]提出了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 中取得了单一模型最好的结果。

尽管 Transformer 模型在很多语言对上取得了最好的翻译效果，但是由于神经网络的拟合训练很容易在维吾尔语到汉语这种低资源语料库上陷入局部最优解，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，CWMT 等机器翻译评测任务中，通过整合不同模型的预测结果用来提升机器翻译的性能[6][7]。但大部分研究人员在使用集成策略解码时都是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作为输出，不能够较好的使用每一个翻译模型内部的信息和反向翻译模型的信息，从而影响翻译的准确性。

针对这个问题，本文基于 Transformer 结构[5]，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）展示了基本的双向重排序模型结构图，对于输入句子 **x**，通过正向翻译模型产生 N-best 翻译列表，使用逆向翻译模型对列表进行重打分。通过综合打分机制得到更好的翻译候选。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。并计算候选翻译项对应于集成的每一个翻译模型的交叉熵，同时计算每一个反向翻译模型对于 N-best 列表的交叉熵。这种方式既可以充分利用集成的优势，又可以有效的整合不同方向的翻译模型对候选翻译的评估。通过综合打分机制进而得到更好的翻译输出。

本文通过在 CWMT2015 的维汉双语平行数据集上进行实验，我们发现随着集成模型数量的增加翻译质量也会随之提高，基于交叉熵的重排序策略能够较好的选出候选翻译，提升翻译质量。多模型集成的双向重排序方法在该语料上取得的最好结果较 Transformer 的单模型有 4.82 个 BLEU 值的提升，显著的改善了维汉机器翻译的质量。

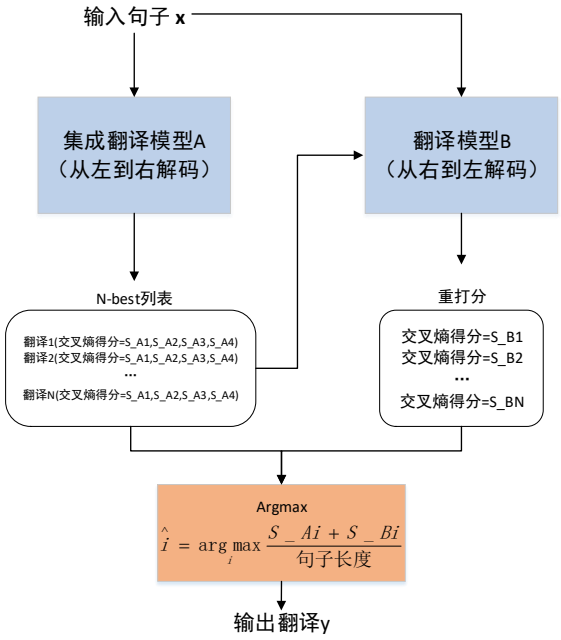

图 1 双向重排序模型示意图

Fig.1 Schematic diagram of bidirectional reranking model

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。神经机器翻译系统通常采用 Encoder-Decoder 架构，将给定的源语句子$\mathrm{X}=（x^1,x^2,\ldots,x^n）$通过编码器编码为中间向量Z，解码器根据中间向量产生目标语言句子$Y=（y^1,y^2,\ldots y^n）$。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：

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

在 Transformer 模型提出之前，大多数神经机器翻译模型都是采用基于注意力机制[1]的循环神经网络（RNN）。这种方式虽然在一些任务上取得了不错的效果，但是由于 RNN 的序列特性导致其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，会导致训练过程十分漫长，从而影响翻译质量。

随着 Transformer 的提出，它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。在提升模型并行化的同时，也提升了模型的表示能力。在翻译结果的准确性上也有一定程度的提升[5]。我们采用 Transformer 作为我们的基线模型，下面简要介绍一下该模型的结构。

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。采用了全新的注意力机制，主要包括 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，由于 Transformer 模型没有使用任何循环神经网络，因此为了能够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 个相同的层堆叠而成，图中展示了两个 Encoder 层的堆叠。每一层都包含两个子层，第一个子层是多头自注意力机制，第二个子层是传统的前馈神经网络。图中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。同时也使用了残差链接和层正则化[8]的方式来保证梯度的传递的稳定。解码器与编码器类似，但是由于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。同时多了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。它的基本单元是缩放的点积注意力模型，每一个头对应于一个点积注意力模型，计算方式如公式（2）所示：

$$\mathrm{Attention(Q,K,V)}=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V} \tag{2}$$

其中 Q 表示查询矩阵，K 表示键值矩阵，d_k表示 K 的维度，V 表示权重矩阵。多头自注意力机制就是采用多组 Q，K，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。

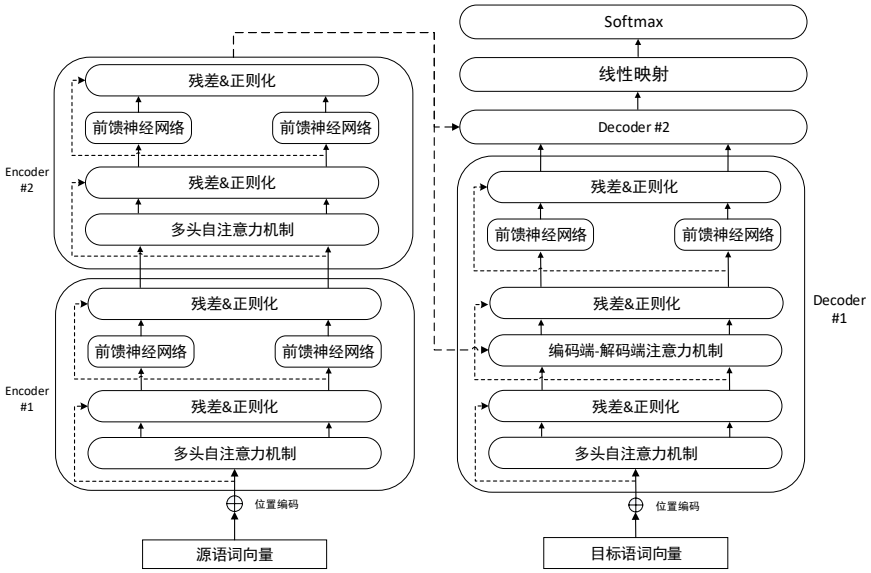

图 2 Transformer 模型示意图

Fig.2 Transformer model diagram

2. 方法

研究人员提出了许多方法用于改善解码过程中的翻译效果。本章节首先介绍如何将不同的翻译模型通过集成策略进行整合，从而得到泛化性能更好的翻译模型。之后介绍重排序策略在在机器翻译中的应用。最后在集成和重排序策略的基础上，介绍如何通过双向重排序的方法提升维汉机器翻译的效果。

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[9]。集成学习方法通过整合多个学习器的决策结果可以有效地减小预测结果的方差与偏置[10]，显著地提升了模型的泛化能力[11]，达到比单学习器更好的效果。因此集成学习方法受到了研究人员的广泛认可，被应用于各种实际任务中[12]。近年来集成学习方法在机器翻译领域，也取得了较好的效果[13][14]。

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输出的结果。模型会根据当前的语义信息计算出一个维度大小是词表大小的概率分布向量，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指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。对于单一模型生成目标端单词的选择如公式（3）所示，与之相对应的集成学习方法生成单词的选择如公式（4）所示。

$$\hat{y}_t = argmax \log P(y_t|y_1^{t-1}, \boldsymbol{x}; M) \tag{3}$$

$$\hat{y}_t = argmax \frac{1}{J} \sum_{j=1}^{J} \log P(y_t|y_1^{t-1}, \boldsymbol{x}; M_j) \tag{4}$$

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$(t-1)$个位置所输出的单词序列。$\boldsymbol{x}$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 表示所集成的翻译模型数量。

然而集成策略仍然存在着一些限制。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。第二，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，要采用束搜索的方式生成候选翻译，模型的解码方向一致时，才能生成更好的候选翻译。

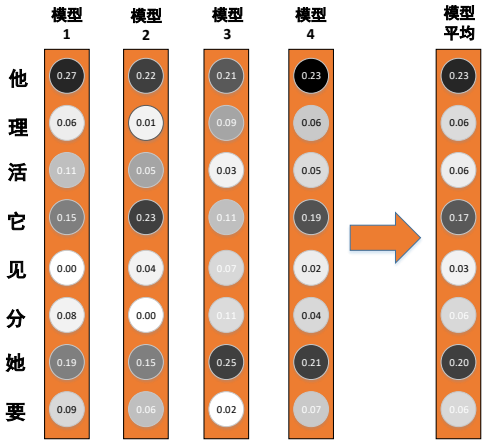

图 3 模型融合示意图

Fig.3 Schematic diagram of model ensemble

2.2 重排序策略

神经机器翻译在解码的过程中，会根据束搜索算法产生 N-best 的翻译列表。有时候概率最大的翻译结果并不一定是最佳翻译。因此如何在规模为 N 的候选翻译中通过重排序的方式找到最优翻译结果就变得十分重要[15]。

在本文中，我们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[16]的定义如下：

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

其中$x_1^n = x_1, x_2, \ldots, x_n$为源语言的句子，$p(x_i)$为源语言词$x_i$所对应的候选翻译词的概率分布表示。$q(x_i)$ 为翻译模型q对x_i所生成目标词的概率估计，即翻译模型q将x_i翻译得到候选翻译列表中对应词的概率。由于模型的交叉熵越小模型的表现越好，为了通过分数的高低来衡量模型的好坏，我们在计算的过程中使用交叉熵的相反数作为该候选翻译最终的得分。

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对于提升机器翻译的效果就变得十分有意义。

2.3 双向重排序模型

集成策略可以生成更好的 N-best 列表，重排序方法将不同解码方向的翻译模型组合起来。为了更好的利用集成策略和重排序策略的优势[17]，我们将这两种方式结合起来，在本文中称之为双向重排序模型。我们通过维汉神经机器翻译中的实例来说明双向重排序模型的有效性。对于给定的源语言和目标语言：

源语言：يالۇجياك دەرياسى چېگرا ھالقىغان چوك كۆۋرۈكى قاتارلىق قۇرۇلۇشلاردا ئەھمىيەتلىك قەدەم بېسىلدى .

参考译文：鸭绿江跨境大桥等建设迈出了有意义的步伐。

我们通过四个正向模型的集成翻译根据对数似然概率产生了 12 个翻译候选，对数似然概率对应图（4）中最后一列。同时计算列表每一个候选翻译对应于集成翻译模型的交叉熵得分 F0-F3，然后依次使用四个逆向翻译模型计算列表中的交叉熵得分 R2L0-R2L3。将所有交叉熵得分进行求和，求和的值和基于翻译长度的惩罚因子相除，得到最终的打分结果。将得分最高的翻译候选作为最终的翻译输出。在这个例子中通过这种方式可以选择出跟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。从而有效改善了译文生成的

质量。

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

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，用以验证基于 Transformer 的双向重排序模型的有效性。我们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和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[18]，其中我们设置 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。这种方式可以有效的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，我们将词汇表规模设置为 36K。同时我们将训练集中的句子长度限制为 100。我们通过 BLEU 值来衡量翻译的质量，采用基于 Moses 的 multi-bleu-detok.perl 的脚本[19]来计算 BLEU 值。

3.2 实验设置

我们使用 Marian 作为我们实验的模型框架[20]。使用 Transformer 模型作为实验的基线模型，参数设置采用了 Transformer_base 参数设置。编码端与解码端的层数分别是 6 层，词向量的维度为 512 维，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 算法作为我们的优化算法，通过采用 AAN 网络[21]来加速 Transformer 的解码过程，采用 Swish 函数作为激活函数。我们在训练中使用了早停机制，参数设置为 5，在整个训练语料上进行了 108 轮。在解码的过程中，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。通常来说，束搜索的大小会大于或者等于候选翻译的规模 N。在本文中，我们设置束搜索的大小与 N-best 大小相等。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。在图（5）中展示了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。在本节中所使用的模型都是单一的翻译模型并没有使用集成的方法来生成翻译。

从图中可以看出从左到右的正向翻译模型，在 N-best 大小为 9 时，BLEU 值最高为 50.49。当 N-best 的规模达到 12 时，BLEU 值趋于平稳。而从右到左的逆向翻译模型，在 N-best 大小为 8 时 BLEU 值最高为 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 值也在逐渐增大，在 N-best 大小为 11-15 时，BLEU 值趋于稳定，在 N-best 为 15 时最高为 52.03.这说明了随着翻译候选列表的增大，双向重排序模型能够较好的选择出最佳翻译候选，从而提升翻译质量。基于以上的实验现象，我们在接下来的实验中将 N-best 的大小设置为 12.

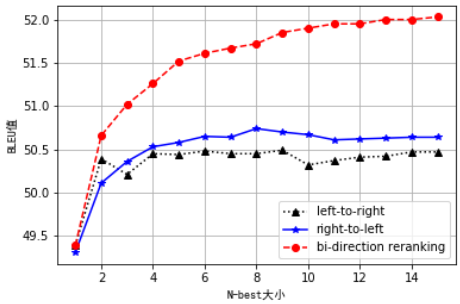

图 5 BLEU 值与 N-best 的关系

Fig.5 Relationship between BLEU value and N-best

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，我们在从左到右的四个模型（L2R）和从右到左的四个模型（R2L）上分别对比了在 N 为 12 的情况下，重排序策略与最大概率的翻译质量。如表（1）所示，从表中可以看出除了第三个正向翻译模型中最大概率的策略要高于重排序策略外，其余模型的翻译质量都有显著提升，BLEU 值平均提升了 0.33.

表 1 重排序策略对 BLEU 值的影响

Table 1 The 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）展示了从左到右，从右到左的解码以及双向重排序多模型集成学习随着集成模型数量的增加翻译结果 BLEU 值的变化。双向重排序模型中我们进行了两组实验

（1）使用从左到右的正向模型作为翻译模型，从右到左的逆向模型作为重排序模型，在图中表示为 bi-directional reranking(l2r)。

（2）使用从右到左的正向模型作为翻译模型，从左到右的逆向模型作为重排序模型，在图中表示为 bi-directional reranking(r2l)。

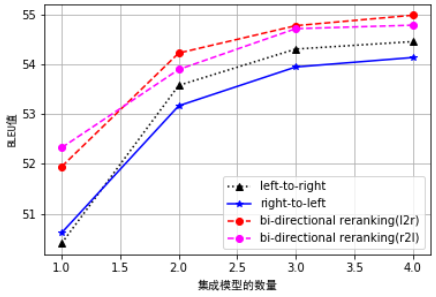

图 6 BLEU 值与集成模型数量的关系

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

1. 对于模型的数量与翻译的质量之间的关系。我们发现对于图中所有的方法，BLEU 值都随着模型数量的增加而增加。然而，随着模型数量的增加，增长速度变慢。在集成了四个翻译模型时，翻译质量达到最优 BLEU 值分别为 54.99，54.79，54.46，54.14.其中双向重排序多模型集成的最好的结果相比于从左到右的正向单一翻译模型（L2R_1）的 BLEU 值提升了 4.58，比从右到左的逆向翻译模型（R2L_1）提升了 4.37.

2. 对于从左到右以及从右到左不同模型集成翻译的结果。我们发现在该语料库上，从左到右的集成翻译效果比从右到左的集成翻译效果有些许提高，BLEU 值平均高了 0.22。通常来说翻译质量与数据集和语言对相关，然而，这些结果表明，翻译质量也会随着解码方向的变化而变化。

3. 对于双向重排序和单向翻译模型集成的翻译效果，从图中可以看出重排序后的翻译模型要显著高于单向的多模型集成。这说明了重排序策略可以更好的选择出最佳候选翻译。为了进一步验证重排序方法对集成翻译模型提升的效果，我们进行了如下对比实验:

在集成了从左到右的四个正向翻译的模型后,我们依次加入逆向的重排序策略,从表(2)中的结果可以看出，随着重排序模型数的增加，BLEU 值也随之提高，相比于没有逆向重排序策略的集成模型，BLEU 值提高了 0.53。这说明了逆向重排序策略可以有效的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 The impact of reranking strategies on ensembled translation

系统名称	系统设置	模型数量	BLEU
Ensem	L2R1, L2R2, L2R3, L2R4	4	54.46
Com-1	Ensem + R2L1	5	54.85
Com-2	Com-1 + R2L2	6	54.98
Com-3	Com-2 + R2L3	7	54.94
Com-4	Com-3 + R2L4	8	**54.99**

4. 通过对比双向重排序模型中的两个实验，我们发现两个实验翻译效果变化不大，正向翻译的模型比逆向翻译的模型高 0.2 个 BLEU 值。我们认为这是由于 4 个正向翻译模型的 BLEU 值的均值要比逆向翻译模型略高。这也从另一个方面说明了，提升单一模型的翻译效果能够在集成策略中更好的获得翻译候选从而提升翻译质量。

4. 相关工作

随着神经机器翻译性能的提升，神经机器翻译越来越受到外界的关注。现有的工作大多都关注于如何改进模型性能，加快模型的运算速度。同时也出现很多在解码端融入不同手段来启发式的改进翻译性能的工作。

模型集成是指在预测下一个目标词之前，对多个模型的概率分布进行综合的一种方法。Vaswani 等人[5]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 等人[22]首先在 WMT16 中应用了检查点集成的方法，然后在 WMT17[6]中进一步尝试了独立集成的方法，与之前的方法相比，取得了显著的改进。李北等人[23]详细分析了集成策略对于神经机器翻译的影响。

为了获得更好的翻译输出，研究人员也尝试了各种方法。Shen 等人[24]提出了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 等人[25]提出了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函数降到最低，从而提升翻译质量。然而，这些方法主要应用于统计机器翻译。L Zhou 等人[26]使用不同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。

在本文中，我们将集成策略应用于维汉神经机器翻译中，使用基于交叉熵的重排序方法充分利用单个模型翻译的信息从中选出较好的候选翻译。经过大量的实验，我们发现双向重

排序模型可以有效提升维汉机器翻译的质量。

5. 总结

在本文中主要介绍了如何通过集成策略和重排序策略来提升维吾尔语到汉语这种低资源语言对上的翻译质量。通过大量的实验，我们发现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。利用这一特性，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，我们获得了比单独集成更高的翻译质量。本文在 CWMT2015 维汉测试集上，当 N-best 的规模为 12 时，经过四个正向模型的集成翻译和四个逆向模型的重排序，实验结果对比于基线系统提升了 4.82 个 BLEU 值。此外我们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。

如果能够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，我们未来计划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，这将是我们下一步工作的重点。

参考文献

1. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
2. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
3. Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.
4. Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1243-1252.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
6. Sennrich R, Birch A, Currey A, et al. The university of edinburgh's neural MT systems for WMT17[J]. arXiv preprint arXiv:1708.00726, 2017.
7. Tan Z, Wang B, Hu J, et al. XMU neural machine translation systems for WMT 17[C]//Proceedings of the Second Conference on Machine Translation. 2017: 400-404.
8. Lei Ba J, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
9. Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990 (10): 993-1001.
10. Dietterich T G. Ensemble methods in machine learning[C]//International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000: 1-15.
11. Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants[J]. Machine learning, 1999, 36(1-2): 105-139.
12. Opitz D, Maclin R. Popular ensemble methods: An empirical study[J]. Journal of artificial intelligence research, 1999, 11: 169-198.
13. Xiao T, Zhu J, Liu T. Bagging and boosting statistical machine translation systems[J]. Artificial Intelligence, 2013, 195: 496-527.
14. Liu Y, Zhou L, Wang Y, et al. A comparable study on model averaging, ensembling and reranking in nmt[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018: 299-308.

15. Och F J, Gildea D, Khudanpur S, et al. A smorgasbord of features for statistical machine translation[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004.
16. 宗成庆. 统计自然语言处理[M]. 清华大学出版社, 2013.
17. Imamura K, Sumita E. Ensemble and Reranking: Using Multiple Models in the NICT-2 Neural Machine Translation System at WAT2017[C]//Proceedings of the 4th Workshop on Asian Translation (WAT2017). 2017: 127-134.
18. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.
19. Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. 2007: 177-180.
20. Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, et al. Marian: Fast neural machine translation in C++[J]. arXiv preprint arXiv:1804.00344, 2018.
21. Zhang B, Xiong D, Su J. Accelerating neural transformer via an average attention network[J]. arXiv preprint arXiv:1805.00631, 2018.
22. Sennrich R, Haddow B, Birch A. Edinburgh neural machine translation systems for wmt 16[J]. arXiv preprint arXiv:1606.02891, 2016.
23. 李北, 王强, 肖桐, 等. 面向神经机器翻译的集成学习方法分析[J]. 中文信息学报, 33(3): 42-51.
24. Shen L, Sarkar A, Och F J. Discriminative reranking for machine translation[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004.
25. Kumar S, Byrne W. Minimum bayes-risk decoding for statistical machine translation[R]. JOHNS HOPKINS UNIV BALTIMORE MD CENTER FOR LANGUAGE AND SPEECH PROCESSING (CLSP), 2004.
26. Zhou L, Hu W, Zhang J, et al. Neural system combination for machine translation[J]. arXiv preprint arXiv:1704.06393, 2017.