基于知识蒸馏的低频词翻译优化策略

郭逸帆，眷红英✉，阎子悦，许鸿飞

[1]郑州大学计算机与人工智能学院，河南郑州
570714651@qq.com; iehyzan@zzu.edu.cn; 1716958920@qq.com; hfxunlp@foxmail.com

摘要

神经机器翻译通常需要大量的平行语料库才能达到良好的翻译效果。而在不同的平行语料库中，均存在词频分布不平衡的问题，这可能导致模型在学习过程中表现出不同的偏差。这些模型倾向于学习高频词汇，而忽略了低频词汇所携带的关键语义信息。忽略的这些低频词汇也包含重要的翻译信息，可能会对翻译质量产生不利影响。目前的方法通常是训练一个双语模型，然后根据频率为词汇分配不同的权重，通过增加低频词的权重来提高低频词的翻译效果。在本文中，我们的目标是提高那些有意义但频率相对较低的词汇的翻译效果。本文提出使用知识蒸馏的方法来提高低频词的翻译效果，训练在低频词上翻译效果更好的模型，将其作为教师模型指导学生模型学习低频词翻译。进而提出一个更加稳定的双教师蒸馏模型，进一步保证高频的性能，使得模型在多个任务上均获得了稳定的提升。本文的单教师蒸馏模型在英语→ 德语任务上相较于SOTA进一步取得了0.64的BLEU提升，双教师蒸馏模型在汉语→ 英语任务上相较于SOTA进一步取得了0.31的BLEU提升，在英语→ 德语、英语→ 捷克语和英语→ 法语的翻译任务上相较于基线低频词翻译效果，在保证高频词翻译效果不变化的前提下,分别取得了1.24、0.47、0.87的BLEU提升。

关键词： 知识蒸馏 ； 机器翻译 ； 低频词翻译 ； 双教师模型

Knowledge Distillation-Based Optimization Strategy for Low-Frequency Word Translation in Neural Machine Translation

Yifan Guo, Hongying Zan✉, Ziyue Yan, Hongfei Xu

[1]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan
570714651@qq.com; iehyzan@zzu.edu.cn; 1716958920@qq.com; hfxunlp@foxmail.com

Abstract

Neural machine translation typically requires a large number of parallel corpora to achieve good translation results. In different parallel corpora, there is an imbalance in word frequency distribution, which may lead to different biases in the learning process of the model. These models tend to learn high-frequency vocabulary while ignoring the key semantic information carried by low-frequency vocabulary. The overlooked low-frequency vocabulary also contains important translation information, which may have an adverse impact on translation quality. The current method is usually to train a bilingual model and assign different weights to vocabulary based on frequency. By increasing the weight of low-frequency words, the translation effect of low-frequency words can be improved. In this article, our goal is to improve the translation performance of meaningful but relatively low-frequency vocabulary. We propose the use of

第二十三届中国计算语言学大会论文集，第206页-第216页，太原，中国，2024年7月25日至28日。
卷1：主会论文

knowledge distillation to improve the translation performance of low-frequency words, training model with better translation performance on low-frequency words, and using it as teacher model to guide student in learning low-frequency word translation. Furthermore, a more stable dual teacher distillation model is proposed to ensure high-frequency performance and achieve stable improvements in multiple tasks. The monolingual teacher distillation model in this study achieved a further improvement of 0.64 BLEU on the English-to-German translation task compared to the state-of-the-art (SOTA) model. The bilingual teacher distillation model, on the other hand, achieved a 0.31 BLEU improvement on the Chinese-to-English translation task compared to the SOTA model. In the translation tasks of English-to-German, English-to-Czech, and English-to-French, compared to the baseline low-frequency word translation approach, the proposed models achieved BLEU improvements of 1.24, 0.47, and 0.87, respectively, while ensuring that the translation performance of high-frequency words remained unchanged.

Keywords: Knowledge Distillation , Neural Machine Translation , Translation of low-frequency words , Dual Teacher Model

1 引言

神经机器翻译模型通常需要大量的平行语料(Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017)，而在机器翻译任务中，翻译模型通常倾向于选择高频词而忽略低频词。在自然语言中，存在着词汇不平衡的情况，不同的单词可以表示相同的含义，更存在着熟词僻意的情况。例如：以汉语中的“然而”举例，在汉英翻译模型中，模型会更倾向于将其翻译为相对高频的“however”，而忽略“nevertheless”。Gu et al. (2020)指出高频词和低频词之间存在严重的不平衡现象，翻译模型很少有机会在训练过程中去学习低频的真实标签，而这些低频词往往携带着重要的语义信息，通常是一些特定领域、文学作品或方言中的专有词汇，能够准确地表达出某些特定的概念或情感。

在以往的工作中(Luong et al., 2015; Jean et al., 2015; Li et al., 2016; Pham et al., 2018)尝试通过维护短语表或低频词表来改进稀有词汇的翻译效果，一些工作在模型中增加了额外的组件(Gulçehre et al., 2016; Zhao et al., 2018)，但是这些带来了额外的训练复杂性和计算成本。一些基于BPE(Sennrich et al., 2016)或词段(Wu et al., 2016)的模型在一定程度上有效的缓解了标签分布不平衡的现象，并且成为多数翻译模型(Vaswani et al., 2017)的标准流程。但是无论多大的细分力度，它们仍然面临着标签分布不均匀的现象。现在多数的翻译模型大多为每一个词分配以相同的权重，并没有考虑到词频不同的情况。在这种情况下，低频词的训练机会相较于高频词小很多，模型在训练中会逐渐忽略低频词带来的损失，导致与低频词相关的参数无法得到有效的训练。

受益于Li et al. (2021)在图像分类中使用知识蒸馏解决长尾数据问题的启发，本文提出使用知识蒸馏的方法去优化低频词翻译效果。本文提出了一种仅使用现有数据的简单有效的方法用于提升低频词翻译效果，不需要引入额外的数据、人工成本。使用低频数据集训练教师模型，提升教师模型的低频词翻译效果，使用低频词翻译效果强的教师模型指导学生模型，学生模型在保留本身学习高频词的倾向的同时，由教师模型指导学生模型的低频词翻译效果。在此基础之上，提出使用双教师模型分别指导学生模型高频和低频词的学习，进一步保证学生模型中高频词的翻译效果，使得模型更具有稳定性。

本文任务的主要目标是：提升翻译模型中低频词的翻译效果，同时保证高频词的翻译效果不被破坏。本文的主要贡献如下：

- 本文使用知识蒸馏的方法解决低频词翻译效果不理想的问题，提出基于知识蒸馏的低频词翻译优化模型，减缓其倾向于高频的趋势，显著提升了低频词的翻译结果，同时保证了高频词翻译效果不被破坏。

- 本文在基于知识蒸馏的低频词翻译模型的基础上，提出双教师知识蒸馏模型，使用两个教师模型同时指导高、低频词的学习，保证了高频词的翻译性能，进一步提升了模型的稳定性。

- 基于知识蒸馏的低频词翻译优化模型在英语→德语任务SOTA的基础上进一步获得了0.64的BLEU提升；双教师蒸馏模型在汉语→英语任务上相较于SOTA取得了0.31的BLEU提升；相较于基线模型，在英语→德语、英语→法语任务上取得了1.24和0.87的BLEU提升，在英语→捷克语任务上获得了0.47的BLEU提升。

2 相关工作

2.1 机器翻译中的低频词翻译优化

在翻译任务中，常见的低频词类型有生僻词翻译、特殊俚语、专业术语等。低频词增加模型的词表多样性，会给模型带来很大的计算负担，翻译模型在处理较大词汇量时具有局限性。一些工作尝试维护短语表或回退词(Luong et al., 2015; Jean et al., 2015; Li et al., 2016)来解决大词汇问题。而现在的主流技术是使用基于子词的方法(Sennrich et al., 2016; Luong and Manning, 2016; Wu et al., 2016)，这种方法大大减少了词汇量并很好的解决了生僻词表示困难的问题。机器翻译的本质是一个分类任务，类别不平衡问题的主要解决方法有两种：基于数据(Baloch and Rafi, 2015; Sutskever et al., 2014)和基于算法(Zhou and Liu, 2005; Lin et al., 2017)的方法。其中基于数据的方法主要采用过采样和欠采样的方法，通过采样的方法解决类别不平衡的问题；基于算法的方法主要为不同的词分配以不同的训练策略。Jiang et al. (2019)提出一种线性加权的方式，根据词频为翻译任务中的词语分配以不同的权重，以此来解决低频词翻译不充分的问题。Gu et al. (2020)在此基础之上，进一步提出了两种权重分配的方法，通过卡方分布函数和幂函数的方式进一步优化的低频词的翻译质量，达到了近年来的SOTA水平。

2.2 知识蒸馏在机器翻译中的应用

知识蒸馏是近几年来解决各类迁移学习任务的热门方法，学生模型可以从教师模型中快速学习到知识的特性赋予了蒸馏方法很多的可能性。Zhuang and Tu (2023)使用知识蒸馏技术，通过对预训练模型的掩码语言模型进行微调，显著的提高了翻译模型的性能。Zhang et al. (2023)验证了通过知识蒸馏的方法可以从已经训练好的翻译模型中获取到知识并能将特定领域的知识转移到学生模型中。而在知识蒸馏中，并不是教师的知识越多越好，教师与学生模型之间的差异过大可能会损坏整体的性能(Wang et al., 2021)。为了解决这种问题，Huang et al. (2022)提出了一种新的蒸馏损失函数，使学生模型的学习目标由原先需要与教师模型完全一致，变为学习教师模型的软标签之间的相对关系，大大降低了蒸馏学习的难度。类别不平衡问题在各种任务中都可以观察到(Wei et al., 2013; Johnson and Khoshgoftaar, 2019)，在图像领域中，Li et al. (2021)使用知识蒸馏技术优化了长尾数据分类不均衡的问题，这为本文提出使用知识蒸馏技术优化低频词翻译提供了思路。本文使用知识蒸馏的方法，学习教师模型中特定的“低频（高频）词领域”的知识，用于解决词语分布不均匀的问题。

3 方法

本文的目标是提升模型在低频词上的翻译效果，同时保证高频词的翻译不受影响。使用在低频词上翻译效果更好的模型作为教师模型指导学生模型低频词的学习，同时学生模型保留模型本身学习高频词的倾向，学生模型在训练时就会减缓其倾向于高频的趋势。

在蒸馏模型中，首先需要训练一个在低频词上效果更好的模型作为教师模型。根据实验结果，教师模型要在低频词翻译上有显著提高且高频词不被严重破坏才能起到指导低频词有效学习的作用。基于知识蒸馏的低频词翻译模型即使用一个在低频词上翻译效果更好的教师模型去指导学生模型，但是实验过程中发现，随着数据集规模的增长，蒸馏模型指导学生模型学习低频词的效率会变得不稳定，甚至会破坏高频词的翻译效果。基于此本文提出双教师蒸馏模型，额外使用一个高频词教师模型去指导模型保障高频词的翻译效果，实验结果表明双教师模型在任意规模的数据集上均表现出稳定的增长。在蒸馏过程中，为了进一步保证高频词不被严重破坏，本文在进行蒸馏微调时对所使用的数据集进行了采样处理。

3.1 基于知识蒸馏的低频词翻译模型

基于知识蒸馏的低频词翻译模型的模型图如图 1所示，其中教师和学生模型的均参照Vaswani et al. (2017)中的实验设置。两个模型的输入序列相同，记为$X = (x_1, x_2, ..., x_n)$，真实标签记为$Y^* = (y_1, y_2, ..., y_m)$，输入经过Transformer编码器,输入序列均被转为向量序列$\mathbf{E}_x = [E_x[x_1], E_x[x_2], ..., E_x[x_n]]$,其中$E_x[x_i]$表示输入序列中第$i$个词的词向量编码与位置向量编码之和，经过编码器N层线性层后，将第N层的输出输入到解码器。解码器对编码器的输出执行多头注意力机制计算，解码器第N层的最终输出得到了目标隐藏层状态记为：$\mathbf{Y}_s = [[s_1], [s_2], ..., [s_m]]$，如下公式1-2所示。

$$\mathbf{E}_x = Embedding(X) + Embedding(S_{pos}) \tag{1}$$

$$\mathbf{Y}_s = Decoder(Encoder(\mathbf{E}_x)) \tag{2}$$

模型在得到解码器最后一层的隐藏层状态后，经过Softmax函数做归一化处理得到每一个预测类别的概率。然而，通过Softmax形成的概率分布通常是一个比较分散的软标签向量，最终的预测类别概率集合中的大多数类别的概率占比很低，学生模型很难学到这一部分知识。对此，利用超参数温度T对Softmax函数的结果进行平滑处理。最终模型一共可以得到三个输出，分别为经过温度T蒸馏得到教师模型和学生模型的软标签，分别记为Out^{tea}_{soft}、Out^{stu}_{soft}，以及学生模型对$\mathbf{Out}_T$取最大值得到的模型最终输出结果记为学生模型硬标签Out^{stu}_{hard}。

模型最终的损失函数根据公式4所示，由学生模型硬标签Out^{stu}_{hard}和真实标签Y^*计算模型的学生损失，使用教师模型软软标签Out^{tea}_{soft}和学生模型软标签Out^{stu}_{soft}计算蒸馏学习的损失。最终整个模型的损失由学生损失和蒸馏损失分配以不同的权重求和所得。

$$Out_T = Softmax(\frac{\mathbf{Y}_s}{T}) \tag{3}$$

$$Loss = \alpha * CrossEntropyLoss(Out^{stu}_{hard}, Y^*) + \gamma * DIST\_KD(Out^{stu}_{soft}, Out^{tea}_{soft}) \tag{4}$$

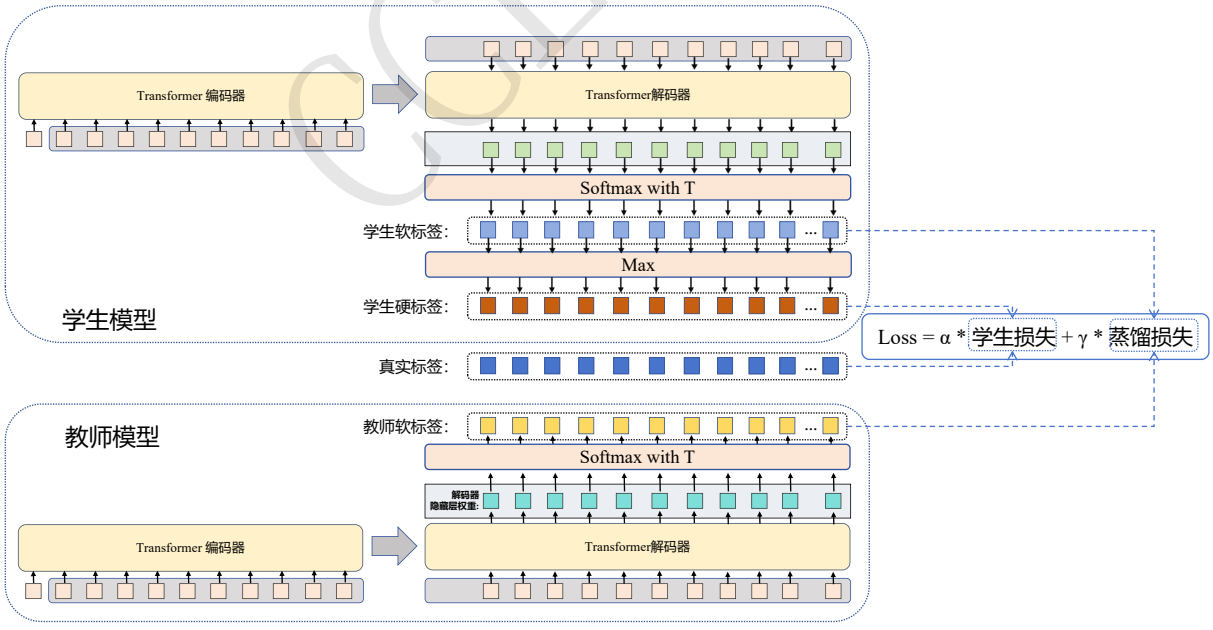

Figure 1: 基于知识蒸馏的低频词翻译优化模型

3.2 双教师知识蒸馏模型

随着数据规模的不断增大，数据集信息量增加而模型容量有限，低频教师模型指导学生模型学习低频词的程度进一步加深，低频蒸馏就会损伤模型的高频知识，从而导致高频词翻译效果不佳。具体而言，只用一个低频教师模型对小数据集的提升是有效的，但是在大数据集任务中表现不稳定。受到使用低频教师模型指导学生模型学习低频词的启发，在通过训练集采样限制模型高频词的学习倾向的同时，设置一个高频教师模型去限制学生模型高频词的大幅降低，将高频词的翻译效果限制在一个可控的范围之内，既不会受低频教师模型影响大幅降低，也不会过于倾向高频词。

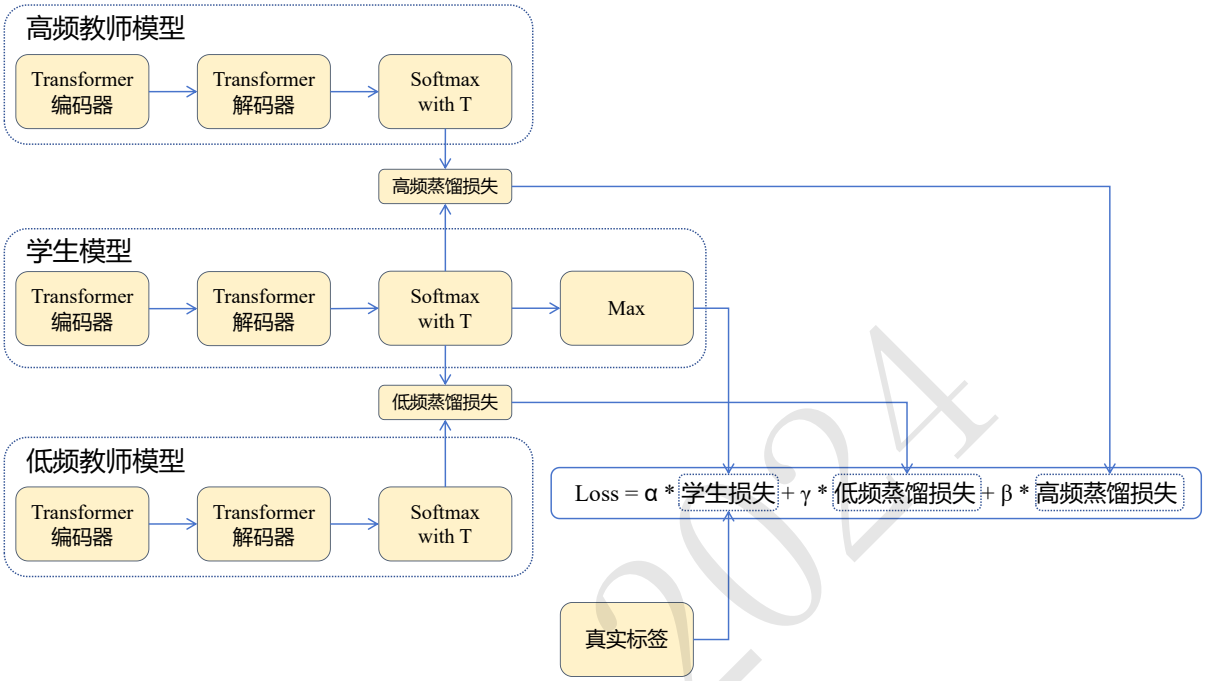

Figure 2: 双教师知识蒸馏教师模型指导学生模型

基于此，本文提出了双教师知识蒸馏模型，通过两个教师模型分别指导学生模型学习高低频词，双教师蒸馏模型的模型图如图2所示。双教师模型在蒸馏模型的基础之上，使用一个在高频词表现优秀的高频教师模型，并使用高频教师模型所得到的软标签$Out_{soft}^{h_tea}$与学生模型的软标签Out_{soft}^{stu}之间形成高频蒸馏损失，最终模型的整体损失如下公式5所示。

$$\begin{aligned} Loss = \alpha * CrossEntropy(Out_{hard}^{stu}, Y^*) + \\ \gamma * DIST\_KD(Out_{soft}^{stu}, Out_{soft}^{tea}) + \\ \beta * DIST\_KD(Out_{soft}^{stu}, Out_{soft}^{h\_tea}) \end{aligned} \tag{5}$$

4 实验

4.1 数据集

受Gu et al. (2020)启发，根据词频按照等式6将训练集和测试集进行打分，其中L表示句长，使用$\frac{1}{L}$消除句长影响；y_i表示句子中的每一个词在数据集中的频率，y_k表示整个词汇表中每一个词的频率。在等式6中，分母表示对整个词汇表中词频求和，在所有句子的得分计算中，分母均相同，句子中含有词汇的词频越小，分子越小，取对数后就会越小，取负号后就会越大，最终一个句子的得分越高表示该句子中包含的词相对低频。将训练集和测试集按照得分进行排序后将二者分别均分为三份数据集记为：$\{Train_{high}, Train_{middle}, Train_{low}\}$,$\{Test_{high}, Test_{middle}, Test_{low}\}$ 分别表示含有更多高频词的数据集、词分布均衡的数据集和含有更多低频词的数据集。蒸馏模型在进行知识蒸馏微调

数据集	训练集	验证集	测试集
英语->德语	4468668	3000	3003
英语->捷克语	10311551	3003	2656
汉语->英语	19489587	2002	2001
英语->法语	35705145	3000	3003

Table 1: 数据集统计表

时，使用采样的方法在三个子训练集中分别取出部分数据作为最终的微调数据集，即微调蒸馏模型的数据集为原数据集大小的1/3。

$$Freq_{sentence} = -\frac{1}{L}\sum_{i=0}^{L} log\frac{Count(y_i)}{\sum_{k=1}^{|V_t|} Count(y_k)} \tag{6}$$

本文主要使用了四个公开数据集，分别为WMT16的英语→德语数据集、WMT14的英语→法语数据集,两个数据集均分别选择newstest-2013和newstest-2014作为验证集和测试集；WMT18的汉语→英语数据集，使用该数据集在WMT官网提供的预处理版本，包含验证集与测试集；以及WMT15的英语→捷克语数据集，分别采用newstest-2013和newstest-2015作为验证集和测试集。按照以下步骤对数据集进行预处理：1）删除有错误编码的句子；2)将数据集中全角字符替换为相应的半角字符，并将所有的命名字符和数字字符（如>, >, &#x3e）转换为对应的Unicode字符；3）使用SentencePiece工具包(Kudo and Richardson, 2018)对数据集进行32000次联合BPE操作，以解决未知单词的问题;

4.2 蒸馏损失函数

Huang et al. (2022)证明在知识蒸馏中，更加强大的教师模型如果与学生模型差异比较大时，学生在普通的知识蒸馏模型上的表现可能会下降，甚至比在没有知识蒸馏的情况下从头开始训练还要糟糕。Huang et al. (2022)提出了一种只关注教师的偏好（即预测的结果相对排名）的方法。相较于传统的知识蒸馏损失关注恢复绝对值的做法，DIST_KD将学生模型从匹配一个强大的教师模型的确切输出的负担中解放出来，只关注于教师模型预测出的不同类别之间的相对关系。在本文中使用该方法，可以有效地减小学生模型对于教师模型的学习难度，模型的结果更加稳定。

4.3 教师模型

教师模型在选择上需优先选择在低频词上表现较好的模型，在Vaswani et al. (2017)的实验设置下，训练步数为100000得到基线模型记为Base，由于本文的方法都是在Base模型上进行蒸馏微调，所以我们额外在Base模型上不增加额外操作，以学习率为$1e-5$微调Base模型训练了一个BaseFT模型作为本文真正的基线模型，本文所有的实验结果均与BaseFT模型做对比。为了得到在低频词上表现更好的教师模型，在Base模型的基础之上，使用$Train_{low}$数据集在Base模型上进行微调，以此得到一个在低频词上表现效果更好的教师模型。在这种情况下得到的微调模型，在高频词的表现上大幅度降低，在$Test_{high}$上的结果低于Base模型。值得注意的是，DIST_KD在指导学生模型时依赖于教师模型预测结果的相对关系，所以低频教师模型在选择上需要保证BLEU结果在$\{Test_{high}, Test_{middle}, Test_{low}\}$上的结果递减。在高频词教师模型的选择上，因为BaseFT模型在微调时,依旧保留了其倾向于学习高频词的能力，所以BaseFT模型本身可作为一个在高频词上表现优异的教师模型。在双教师蒸馏模型中不同损失函数的权重上，为了保证学生损失的稳定，将α置为1，而后以0.1为间隔在验证集上测试得到最优的损失函数权重α、β、γ分别为1、0.6、0.4。

4.4 主要结果

本文的方法在WMT18中英数据集上的结果如表2所示，其中EMNLP_Exp与EMNLP_K2方法来自于(Gu et al., 2020)中的方法，代表当前的SOTA结果。在将测试集按照公式6划分为三个测试集$\{Test_{high}, Test_{middle}, Test_{low}\}$后，本文的方法在保证含有更多高频词的测试集$Test_{high}$的翻译效果不受影响的情况下，小幅度提升了词均衡的测试集$Test_{middle}$的翻译效

果，显著提升了含有更多低频词的$Test_{low}$的测试集的翻译效果。与当前的SOTA模型相比，本文的方法在高频测试集和均衡测试集上的效果基本与其持平或小幅提升。在低频测试集上进一步提升了0.31的BLEU值。相较于基线模型，整体提升了0.86的BLEU值，证明本文的方法十分具有竞争力。

	汉语→英语			
	$Test_{low}$	$Test_{middle}$	$Test_{high}$	Total
Base	20.95	22.29	25.76	23.22
BaseFT	21.51	22.48	26.45	23.88
$Teacher_{low}$	22.50	22.62	23.71	22.92
EMNLP_Exp (Gu et al., 2020)	22.06	22.57	26.54	24.01
EMNLP_K2 (Gu et al., 2020)	22.01	22.81	26.59	23.98
Our_Double	**22.37**	22.90	26.67	24.05

Table 2: 蒸馏模型在WMT18汉语→英语数据集上的结果

4.5 蒸馏模型在多种翻译任务上的结果

本文的两种方法在WMT16英德数据集上的结果如3所示，当前的SOTA模型相比，本文的方法在高频测试集和均衡测试集上的效果基本与其持平或小幅提升。在低频测试集上，单教师蒸馏模型的结果进一步提升了0.64的BLEU值，更加稳定的双教师蒸馏模型进一步提升了0.72的BLEU值。相较于基线模型，整体提升了1.24的BLEU值。

	$Test_{low}$	$Test_{middle}$	$Test_{high}$
BaseFT	25.55	26.97	28.88
EMNLP_Exp (Gu et al., 2020)	26.07	27.33	28.91
EMNLP_K2 (Gu et al., 2020)	25.99	27.28	28.90
Our_Sigle	26.71	27.40	28.97
Our_Double	**26.79**	27.44	28.99

Table 3: 蒸馏模型在WMT16英语→德语数据集上的结果

为了测试方法在更多数据集上的可靠性与稳定性，分别选取了WMT15的英语→捷克语以及WMT14的英语→法语数据集继续验证本文的蒸馏方法。在WMT15的英语→捷克语和WMT14的英语→法语数据集上的实验结果如表4所示。在两个语种上的结果，双教师蒸馏模型相较于基线模型BaseFT,同样做到了保证高频词不被破坏且基本与基线模型持平，小幅度提升了均衡测试集结果，大幅度提升含有更多低频测试集的翻译效果。在英语→捷克语和英语→法语的低频测试集上分别提升了0.47和0.87的BLEU值，双教师蒸馏模型的稳定性得到了验证。

	英语→捷克语				英语→法语			
	$Test_{low}$	$Test_{middle}$	$Test_{high}$	Total	$Test_{low}$	$Test_{middle}$	$Test_{high}$	Total
Base	27.74	27.98	30.39	28.48	37.16	39.58	41.47	39.65
BaseFT	27.80	28.42	31.40	29.07	37.75	40.16	41.93	40.23
$Teacher_{low}$	28.82	28.08	29.09	28.64	38.67	40.21	40.22	39.76
Our_Double	**28.27**	28.64	31.49	29.33	**38.62**	40.45	42.01	40.65

Table 4: 蒸馏模型在英语→捷克语、英语→法语数据集上的结果

4.6 不同数量教师模型对蒸馏结果的影响

为了验证单教师模型和双教师模型在不同数据集上的影响，本文在每个数据集上均测试了单教师蒸馏模型和双教师蒸馏模型的实验效果。为了更好的表现数据集大小对不同方法稳定度的影响，选择了iwslt14的德语→英语的数据集补充为额外的小规模数据集。该数据集在数据量

上仅有174K句句对文件，相较于其它四个数据集，iwslt14数据集属于低资源数据，实验结果如表5所示。

实验结果表明，无论是单教师模型还是双教师模型在$Test_{low}$上的结果均表现出增长，证明使用蒸馏的方法对低频词的翻译效果是有效的。但是随着训练数据集规模的增长，单教师蒸馏模型在$Test_{high}$上逐渐显现出负增长，这是因为数据集信息量增加而模型容量有限，低频教师模型指导学生模型学习低频词的程度进一步加深，低频蒸馏就会损伤模型的高频知识，从而导致高频词翻译效果不佳。双教师蒸馏模型使用一个高频教师模型约束在蒸馏过程中对高频词的破坏，实验结果表明在任何规模的数据集上，双教师蒸馏模型均表现出稳定的增长。

任务	模型	测试集			
		$Test_{low}$	$Test_{middle}$	$Test_{high}$	Total
iwslt14德语→英语 (174K句对)	BaseFT	27.87	30.88	35.70	31.10
	Single_Model	28.73↑	31.54	35.67→	31.56
	Double_Model	28.74↑	31.59	35.77→	31.69
WMT16英语→德语 (4.4M句对)	BaseFT	25.55	26.97	28.88	27.21
	Single_Model	26.71↑	27.40	28.97→	27.88
	Double_Model	26.79↑	27.44	28.99→	27.93
WMT15英语→捷克语 (10M句对)	BaseFT	27.80	28.42	31.40	29.07
	Single_Model	28.58↑	28.06	29.59↓	28.68
	Double_Model	28.27↑	28.64	31.49→	29.33
WMT18汉语→英语 (19M句对)	BaseFT	21.51	22.48	26.45	23.88
	Single_Model	22.28↑	22.32	24.71↓	23.16
	Double_Model	22.37↑	22.90	26.67→	24.05
WMT14英语→法语 (35M句对)	BaseFT	37.75	40.16	41.93	40.23
	Single_Model	38.48↑	39.37	40.74↓	39.26
	Double_Model	38.62↑	40.45	42.01→	40.65

Table 5: 不同数量教师模型在不同规模数据集上的实验

4.7 不同损失函数对蒸馏结果的影响

本文在损失函数的选择上，选择了 Huang et al. (2022)提出的DIST_KD损失函数，而并非传统的KLD蒸馏损失函数。DIST_KD蒸馏损失函数相较于传统的KLD蒸馏损失函数，学生模型更加专注于学习教师模型预测的词之间的相对关系，即教师模型的翻译倾向，缓解因为教师模型与学生模型差距过大而导致的学习不充分的问题，并且在一定程度上降低了学习难度。传统的KLDloss在计算蒸馏损失时，关注于教师模型预测概率的绝对关系。而本文在蒸馏实验中，低频词模型和高频词模型与学生模型之间相差很大，我们希望模型能够学习到低频教师模型对低频词的倾向，而并非要求学生模型与教师模型完全一致，这样会破坏学生模型对于高频词的学习。DIST_KD契合本文使用教师模型指导学生模型的学习倾向的目的。以WMT18中英数据集为例，选用不同蒸馏损失函数的实验结果如表6所示，实验结果表明DIST_KD损失由于学习低频词预测的相对关系，降低了学习难度，不会严重破坏学生模型的高频词。传统的KLDloss由于学习教师模型的绝对关系，低频教师模型对学生模型造成的影响更大，促使学生模型以完全学习低频词为目标，破坏了高频词的翻译效果，从而在$Test_{high}$上表现不佳。使用DIST_KD不仅进一步提升了低频词的翻译效果，同时保证了高频词的翻译稳定性。

	汉语→英语		
	$Test_{low}$	$Test_{middle}$	$Test_{high}$
BaseFT	21.51	22.48	26.45
+KLDloss	21.83	22.85	26.29
+DIST_KD	**22.37**	22.90	26.67

Table 6: 不同蒸馏损失函数对蒸馏结果的影响

4.8 数据采样对蒸馏结果的影响

在进行限制高频词翻译效果下降的措施上，本文同时采用了采样方法，分别在三个训练集{$Train_{high}$,$Train_{middle}$,$Train_{low}$}上按照不同比例进行采样后形成新的训练集，为了排除最终结果的提升是否受采样方法的影响，在保证和蒸馏实验其它步骤一致的情况下，不使用蒸馏方法而选择采样训练集直接微调。以WMT18中英数据集使用采样数据集进行微调的实验结果如表7所示，实验结果表明仅采用采样数据进行微调时，虽然提升了低频词的翻译效果，但是对高频词造成了严重破坏，在$Test_{high}$上降低了0.79的BLEU值。而在采样的基础之上增加单教师模型，中英数据集由于数据规模性较大且低频教师模型进一步指导学生模型的低频词翻译效果，导致高频词效果进一步下降。而双教师模型通过高频教师模型的作用，保证了高频词的翻译效果。

	汉语→英语		
	$Test_{low}$	$Test_{middle}$	$Test_{high}$
BaseFT	21.51	22.48	26.45
+sample data	21.96	22.79	25.66
+single_distill	22.28	22.32	24.71
+double_distill	**22.37**	22.90	26.67

Table 7: 数据采样对蒸馏结果的影响

5 总结

在本文的研究中，我们关注到了机器翻译任务中类别不均衡的问题。类别不均衡会导致翻译模型倾向于翻译高频词，进而导致词汇多样性被破坏。为了提升那些低频但具有关键含义的词，本文研究了利用知识蒸馏技术解决低频词翻译不充分的问题，进而提出了两种基于知识蒸馏的低频词翻译优化模型。基于知识蒸馏的低频词翻译优化模型有效的提升了低频词的翻译质量，双教师蒸馏模型在此基础上进一步保证了高频词的效果，在保证高频词不被破坏的情况下显著提升了低频词的翻译效果，在多项翻译任务中均获得了稳定的增长。最后的结果表明，本文的方法可以实现显著的性能改善，相较于SOTA模型获得了进一步的提升。

致谢

本文受国家自然科学基金联合基金重点支持项目（U23A20316）资助。许鸿飞受国家自然科学基金青年项目（62306284）、中国博士后科学基金面上项目（2023M743189）和河南省自然科学基金青年项目（232300421386）资助，昝红英是本文的通讯作者。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Maher Baloch and Muhammad Rafi. An investigation on topic maps based document classification with unbalance classes. *Journal of Independent Studies and Research (JISR)*, 13(1), 2015.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, 2020.

Çaglar Gulçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2016.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33716–33727. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/da669dfd3c36c93905a17ddba01eef06-Paper-Conference.pdf.

Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL-IJCNLP 2015*, pages 1–10. Association for Computational Linguistics (ACL), 2015.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, pages 2879–2885, 2019.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 630–639, 2021.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. Towards zero unknown word in neural machine translation. In *IJCAI*, pages 2852–2858, 2016.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Minh-Thang Luong and Christopher D Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, 2016.

Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, 2015.

Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. Towards one-shot learning for rare-word translation with external experts. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 100–109, 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL), 2016.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 6456–6466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.504. URL https://aclanthology.org/2021.acl-long.504.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16: 449–475, 2013.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yuanchi Zhang, Peng Li, Maosong Sun, and Yang Liu. Continual knowledge distillation for neural machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7996, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.443. URL https://aclanthology.org/2023.acl-long.443.

Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. Addressing troublesome words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 391–400, 2018.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18 (1):63–77, 2005.

Yimeng Zhuang and Mei Tu. Pretrained bidirectional distillation for machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1132–1145, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.63. URL https://aclanthology.org/2023.acl-long.63.