Letz Translate：面向卢森堡语的低资源机器翻译

Yewei Song*、Saad Ezzini*、Jacques Klein*、Tegawende Bissyande*、Cl´ement Lefebvre†和 Anne Goujon‡
*安全、可靠性与信任跨学科中心（*SnT*）、卢森堡大学邮箱： *FirstName.LastName@uni.lu,
†‡Banque BGL BNP Paribas，50，J.F Kennedy 大道，L-2951 卢森堡 邮箱： †clement.c.lefebvre@bgl.lu， ‡anne.goujon@bgl.lu

arXiv:2303.01347v1 [cs.CL] 2 Mar 2023

摘要—**低资源语言（LRL）的自然语言处理常常受制于数据匮乏。因此，在低资源环境中实现准确的机器翻译（MT）是一个亟待实用解决方案的问题。多语言模型的研究表明，这类模型能够处理某些低资源语言。然而，这些模型体量庞大、计算需求高，因此难以在受限环境中（例如移动/物联网设备或资源有限的旧服务器）实际使用。本文通过知识蒸馏，利用大型多语言机器翻译模型的能力来解决这一问题。知识蒸馏能够将大型复杂教师模型的知识迁移到更简单、更小的学生模型中，同时基本不损失性能。我们还利用与目标低资源语言相关或具有相同语言根源的高资源语言。在评估中，我们将卢森堡语视为一种与德语具有共同根源和部分相似特性的低资源语言。我们基于德语、从多语言不让任何语言掉队（NLLB）模型进行的知识蒸馏，以及伪翻译，构建了多个资源高效的模型。结果发现，与大型先进的NLLB模型相比，我们的高效模型速度提升了30%以上，性能仅低4%。**

索引词——**神经机器翻译、低资源语言、低资源翻译、知识蒸馏、卢森堡语**

一、引言

自然语言是人类宝贵的遗产，需要加以保护、发展，并与其他语言建立联系。自然语言处理（NLP）已被证明会直接影响自然语言的发展与维护，例如使无文字的闽南语能够通过神经机器翻译译成英语 [1]。机器翻译（MT）是自然语言处理最实用的日常应用之一，能够促进跨文化人际互动。

近年来，神经机器翻译（NMT）彻底改变了机器翻译（MT）任务。语言模型和迁移学习的最新进展，使机器翻译任务能够充分利用现有的大型平行语料库，大幅扩展规模。然而，低资源语言（LRL）缺乏此类大型语料库。因此，无法为低资源语言构建经典的神经机器翻译模型。尽管 Meta 近期推出的「不让任何语言掉队」（NLLB）[2] 多语言翻译模型能够处理 200 种语言，但其庞大的规模和计算需求，使其难以在低资源环境（例如在线服务、移动/物联网设备）中实际运行。

为解决上述局限，我们提出采用知识蒸馏和语言趋近技术，为低资源语言构建资源高效翻译模型。我们提出的技术旨在以最少的计算能力实现接近最先进的性能。知识蒸馏是指利用大型多语言教师模型（即 NLLB）向学生模型传授相关任务，使其以较低的推理成本达到近似的性能。语言趋近技术则通过替换词元和表达，利用与目标低资源语言相关的高资源语言（HRL）的资源。我们的方法旨在以较低的计算成本、较少的平行数据和较快的推理时间，提供准确的翻译模型。在我们的案例研究中，目标是高效地将卢森堡语（LB）翻译成英语（EN）。因此，我们将卢森堡语视为与德语（DE）这一高资源语言相关的低资源语言。在实验中，我们发现，与使用更大的模型（无论是采用更大的平行数据集进行训练，还是利用从德语伪翻译得到的数据扩充现有卢森堡语数据）相比，我们这种数据高效的蒸馏方法取得了更好的结果。

二、相关工作

本节将介绍我们的工作在现有低资源神经机器翻译文献中的定位；这些研究使用了其他高资源语言的模型或数据。

在其文章中， [3], Ko 等人提出了一种适配方法，将大量平行的高资源语言（HRL）数据与低资源语言（LRL）单语数据相结合，创建出能够跨语言群组泛化的无监督翻译模型。作者发现，在英语与葡萄牙语（一种 LRL）之间进行翻译时，他们的方法能够达到最高 36% 的双语评估替补（BLEU）分数（与使用西班牙语模型相比提高了 24%）。他们的方法在最终的微调步骤之前，依赖于回译和对抗式领域适配技术。该论文未讨论所创建模型的计算效率。与我们的文章类似，已有研究利用与目标 LRL 相似的 HRL（例如具有相同词根的语言），来改进 LRL 与英语之间的神经机器翻译（NMT）。Neubig 等人 [4]以及其他研究者 [5]¸[6] 采用了将 LRL 和 HRL 的平行数据混合起来训练并改进相关 LRL 的 NMT 的技术。该技术基于以下两种做法之一：使用 HRL 数据进行预训练、使用 LRL 数据进行微调；或者在对少数语言数据进行平衡和扩充后，将两个数据集混合，形成一个训练集。前一种方法表现更佳。

训练，或在对少数语言数据进行平衡和扩充后，将两个数据集混合，形成一个训练集。前一种方法表现更佳。

与现有文献不同的是，在我们的工作中，我们利用大型多语言模型进行知识蒸馏，生成资源和数据高效的模型。我们还将知识蒸馏与以下方法进行了比较：使用相关的高资源语言模型，在低资源语言平行语料上对其进行微调，以及在通过对大型高资源语言平行数据集进行伪翻译转换而创建的合成语料集上对其进行微调。

III. 方法

本节介绍了本工作中使用的主要流程和各种技术。

A. 问题定义

我们的目标是获得一个**快速**且**高性能的**翻译模型 ① ，仅使用来自低资源语言的少量数据，并在 ② 计算能力有限的环境中运行。

如果模型每个实例的推理时间显著短于 NLLB 等最先进方法所需的推理时间，我们将其视为**快速**翻译模型。例如，大型 NLLB 模型（具有33亿个参数）在一台48核CPU服务器上处理每个实例需要17秒的推理时间，因此相对较慢且不实用。如果其翻译得分接近最先进的大型模型的水平，则可将该翻译模型视为**高性能的**模型。

首先，我们会使用有限的低资源语言数据来模拟低资源场景，例如使用规模较小的卢森堡语（LB）**单语语料库**。其次，我们会在计算资源非常受限的场景中使用模型，例如移动设备、没有 GPU服务器的内联网等。我们的目标是在仅使用 CPU 的服务器上，将每句的推理时间控制在 1 秒以内。

B. 获取平行数据的两种方法

本文比较了以下两种技术在低资源语言到英语的翻译模型中的表现： ① 伪翻译和 ②知识蒸馏。我们还使用现有的平行语料库进行了对比实验。

伪翻译： 与一些关于低资源语言模型的先前研究（如 LuxemBERT [7],和 Translating Translationese [8], ）类似，我们使用伪翻译，从与低资源语言在语言或语义上关系密切的另一种高资源语言中获取质量合格的句子。在我们的案例中，卢森堡语在语法上与德语十分相似，在部分词汇上则与法语（FR）相似。

知识蒸馏： 在一篇 NISP’2014 论文中， [9], Hinton 等人提出了知识蒸馏的概念，旨在将从一个大型模型或多个模型中学到的知识迁移到另一个轻量级单一模型中，以便于部署。在某些情况下，使用软目标计算损失函数，无需真实值也能训练出一个完整的模型。知识蒸馏的概念是这项工作的最初灵感之一。

C. 流程

图 1 概述了生成我们的**伪模型**的流程。为了利用伪平行数据创建模型，我们首先需要现有的高资源语言到英语平行数据，并配合合适的词典执行伪翻译任务，将高资源语言句子转换为伪低资源语言句子。获得伪平行数据后，我们使用这些数据，在伪低资源语言到英语数据集上微调现有的预训练神经机器翻译模型。高资源语言到英语模型和多语言到英语模型均可用于此任务。

图 2 展示了创建我们的**蒸馏模型**的流程。进行知识蒸馏时，我们首先需要一个低资源语言语料库（非平行语料），将其输入大型多语言 NLLB 模型，以便为参数更少的高资源语言到英语模型或多语言到英语模型生成软目标。蒸馏过程是将知识从大型、复杂的教师模型传递给更小、更简单的学生模型，同时尽可能减少性能损失。软目标用于尽量缩小两个模型之间的误差差异。

我们的模型准备就绪后，可以使用少量但高质量的平行数据（低资源语言到英语）进行额外的微调。我们将比较两个维度：使用不同的第一轮微调数据集，以及是否进行第二轮微调。我们还将真实数据集微调集作为基线。

最后，我们希望获得一个计算速度快、计算资源消耗低，且性能接近超大型多语言模型的卢森堡语到英语翻译模型。最理想的结论是，我们的方法能够扩展到所有低资源语言到高资源语言翻译。我们将在下一节评估所创建的模型。

IV. 评估

A. 实验设计

我们的实验依赖于两个数据集。第一组是从德语到英语平行数据中伪翻译得到的卢森堡语数据，使用图1中的第一个流程获得；第二组是使用 NLLB-蒸馏模型（6亿参数）将卢森堡语语料库转换为生成的软目标英语。我们需要比较几个不同模型的翻译性能。

多对一机器翻译模型近年来逐渐受到青睐，通常是将罗曼语族语言译为英语，其中一些模型也涵盖卢森堡语以及其他类似的法兰克语支语言。我们希望测试我们的数据在不同预训练版本的同类机器学习模型上的表现。

B. 研究问题

RQ1. 对于低资源机器翻译，最准确的解决方案是什么?

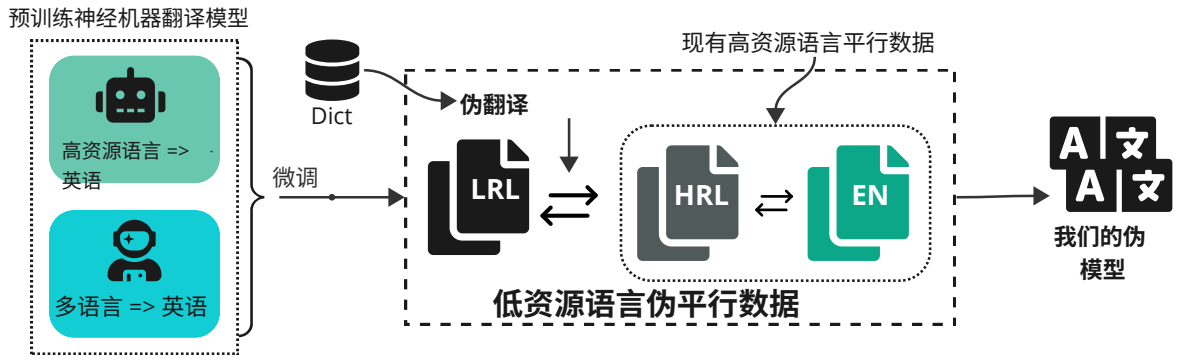

图 1. 获取我们的伪模型的流程

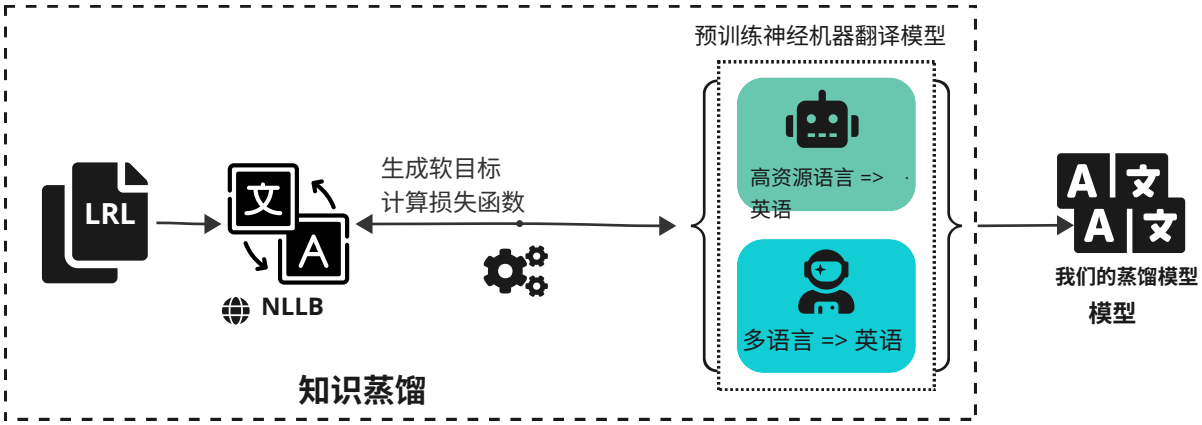

图 2. 获取蒸馏模型的流程

针对 RQ1，我们希望评估基于伪翻译和知识蒸馏的模型在卢森堡语应用场景中的表现，并将其与我们的基线进行比较。

RQ2. 就计算效率和推理时间而言，最适合低资源语言的神经机器翻译模型是什么?

针对 RQ2，我们在相同的硬件上评估所有模型，以确定模型大小、推理时间和翻译准确率之间的最佳权衡。

C. 实现细节与可用性

我们用于创建模型的主要库是 Transformers、PyTorch 和 Redis。我们使用了以下预训练模型：OPUS-MT-DE-EN、OPUS-MT-MUL-EN、NLLB-200-3.3B（NLLB-large）以及 NLLB-200-DISTILLED-6亿参数（NLLB-Distilled）。我们的训练任务在一台配备 Intel Xeon(R) CPU E5-2698 的 Nvidia TESLA V100 32GB 上完成，并在 AMDEPYC7552 上运行推理。创建的模型（经过蒸馏训练、两轮微调，基于 OPUS-MT-DE-EN）可在 HuggingFace 平台上获取[1]。

[1]https://huggingface.co/etamin/OPUS-TruX-Ltz-EN

D. 数据收集

在我们的评估中，我们针对不同目的使用了多个数据集，例如微调、伪翻译和软目标计算。

我们将 Gierschek 在 [10] 提供的数据视为卢森堡语-德语-英语句子的平行数据集。她从 RTL Luxembourg[2],这个卢森堡语新闻网站收集了这些句子。我们删除了长度少于 50 个字母或超过 500 个字母的句子。最终剩下 110,720 组三语平行句子。我们将此数据集称为**真实数据集**。

我们将收集到的双语词典应用于 Gierschek 在 [10]提供的真实数据集中的德语-英语数据集，从而构建**伪翻译**数据集。我们收集了一份包含 51,617 个词条的词典[3]，，并将艾费尔规则（一项针对词尾为 'n' 和 'nn' 的单词的特殊规则）应用于德语句子 [11]。

用于蒸馏微调的**软目标** 由 NLLB-Distilled 模型 [12] 生成。为此，我们使用真实值中的卢森堡语句子（即 [10] 中 Gierschek 的数据集），并且混合损失不使用英语句子。生成软目标耗时约 157 小时。在我们的实验中，我们使用了规模最小的 NLLB

[2]www.rtl.lu[3] github.com/Etamin/Ltz 词典

模型（NLLB-Distilled）。请注意，如果我们使用规模最大的版本（NLLB-large），实验将需要额外 350 小时的处理时间。

我们使用 *Flores-200*[4] 训练集进行微调，并使用 *Flores-200* 测试集和 *Tatoeba*[5] 数据集进行测试。Flores-200 是一个广泛使用的英语与低资源语言之间的多语言翻译数据集，是 Meta 发布的 Flores101 的扩展版本。Flores-200 的训练集（用于微调）包含 997 对句子，测试集包含 1017 对句子，用于测试 [12]。Tatoeba 是一个由社区共同维护的开放式平行句子集合，涵盖 420 种语言 [13]。然而，Tatoeba 中只有 306 对卢森堡语到英语的句子。

E. 预训练模型与微调

为了构建计算效率高的模型，考虑到训练成本低廉，我们选择了两个不同的 OPUS-MT 预训练版本：**OPUS-MT-DE-EN** 和 **OPUS-MT-MUL-EN** [14]。

对于所有微调任务，我们均采用以下相同的训练设置：最大步数为 250,000，AdamW 的学习率设为 2e-5，权重衰减为 0.01，批次大小为 16。此外，我们还使用交叉熵损失函数。

为验证伪翻译或知识蒸馏对模型泛化能力的影响，我们使用高质量 Flores-200 的一个子集，在不到 500 个步骤内进行**第二轮**微调。此过程旨在探究我们的数据是否为模型从卢森堡语映射到英语的能力提供了正确的梯度。

F. 评估指标

在我们的评估流程中，我们使用*SacreBLEU* 和 *ChrF++* 来衡量我们模型的性能，以及所得翻译相较于真实值的质量。

SacreBLEU 是 BLEU（双语评估替补）的一种实现变体，其目标相同，都是通过衡量译文句子与真实数据集句子之间的距离来评估文本翻译质量。为简便起见，我们将使用“BLEU”这一记法，而不是 SacreBLEU。

ChrF++ 是 ChrF 的改进版本，它使用 F-score 统计量来评估字符 n 元语法匹配。ChrF++ 通过将词 n 元语法加入计算，进一步改进了 ChrF。

G. 实验

本节概述我们为回答研究问题而开展的实验。

1）基线模型：我们选取三种支持卢森堡语到英语翻译的多语言机器翻译模型作为基线：NLLB-large（3.3B）、NLLB-蒸馏（6亿参数）和 M2M-100（1.2M）。除此之外，我们还考虑了 OPUS-MT 的两个版本：一个是同样支持卢森堡语的多语言版本（OPUS-MT-MUL-EN），另一个是德语到英语的 OPUS-MT-DE-EN 模型。对于每个版本，我们都评估了其在小型 Flores-200训练集上微调前后的表现。表 I 列出了所选基线模型；在 Flores 训练集上微调的模型带有“*FT*”后缀。

从表 I 可以看出，NLLB-large 翻译一个句子大约需要 17 秒。它的参数数量最多（33亿），因此需要最多的计算能力。显然，与其他基线相比，这个最大的模型具有最佳性能。M2M-100 [15]是第二大的模型，单句执行时间约为 7 秒，规模为 12亿参数。然而，就准确率而言，它仅排名第五，这意味着模型规模并不总是与性能相关。另一方面，尽管 NLLB-Distilled 是 NLLB 的轻量级版本，且在我们的列表中准确率排名第二，但它每个句子仍需要超过 3 秒，并拥有 6亿参数，这使其成为一个很好的对比基线。基于 OPUS-MT 的模型每秒可翻译两到三个句子，且规模最小（比 NLLB-Distilled 小 8 倍，比 NLLB-large 小 44 倍）。这些轻量级模型很适合我们的用例（需要在低计算资源下实现快速推理时间的在线服务）。

2）我们的模型：如前所述，在评估中，我们考虑两种主要方法（即伪翻译和知识蒸馏），并将其应用于两个基于 OPUS-MT 的模型 MUL-EN 和 DE-EN，如表 II 所示。在这项实验中，我们比较了经过伪翻译数据集、蒸馏数据集（通过蒸馏获得）以及全部真实数据集微调后的 OPUS-MT 模型的性能。我们还评估了这些模型在 Flores-200训练集上进行第二轮微调后的表现。

如表 II 所示，经过第二轮微调后，伪翻译模型的表现优于 M2M-100 和 OPUS-MT-DE-EN 模型，但总体结果较差，因为伪卢森堡语句子与实际真实值之间存在较大差异，BLEU 分数低于 35 分。在表 I 的所有数据集上，这些模型的表现都不如 OPUS-MT-MUL-EN 和 NLLB 模型。

通过蒸馏学习获得的模型，其性能与使用“真实”平行数据进行微调的模型非常接近，平均 BLEU 分数差异不到 2 个百分点。使用 Flores-200 进行第二轮微调后，我们的模型在某些场景中有所提升。

在表 II 中进行跨列比较可见，德语-英语模型在将知识迁移至卢森堡语后取得了更好的结果，平均 BLEU 分数优势约为 1%。这可能是因为卢森堡语中的某些句子与德语非常相似，也可能是因为微调过程中保留了部分德语知识。

基于德语-英语模型优于 MUL-

[4]https://huggingface.co/datasets/facebook/flores [5] https://huggingface.co/datasets/tatoeba

表 I 基线模型性能

模型（规模）	Tatoeba LB-EN		Flores 200 测试集		推理时间
指标（参数数量）	BLEU	ChrF++	BLEU	ChrF++	每句秒数
NLLB-large（3.3B）	**56.019**	**71.315**	**44.634**	**68.981**	16.99
NLLB-Distilled（6亿参数）	54.470	70.395	38.589	64.424	3.26
M2M-100（1.2B）	20.679	35.054	30.686	57.029	6.88
OPUS-MT-MUL-EN（77M）	32.186	48.047	19.819	46.89	0.51
OPUS-MT-DE-EN（74M）	11.591	21.111	5.370	24.591	0.28
OPUS-MT-MUL-EN(FT)（77M）	36.763	55.044	24.509	52.003	0.51
OPUS-MT-DE-EN(FT)（74M）	23.941	39.215	19.672	46.259	0.28

所有翻译组的测试结果（表 II）

预训练			OPUS-MT-DE-EN		OPUS-MT-MUL-EN	
第二次微调			Flores	None	Flores	None
Tatoeba	BLEU	伪	26.460	9.697	37.553	14.480
	ChrF++		42.504	17.001	55.697	26.401
Flores	BLEU		20.495	3.911	22.351	5.301
	ChrF++		47.843	21.671	49.823	26.478
Tatoeba	BLEU	蒸馏	47.666	**48.361**	46.051	**47.906**
	ChrF++		63.733	**66.245**	**65.535**	65.508
Flores	BLEU		**29.790**	28.510	**27.988**	25.521
	ChrF++		**57.618**	56.284	**56.054**	53.897
Tatoeba	BLEU	真实值 真值	48.939	50.658	47.813	48.444
	ChrF++		67.960	68.953	67.057	67.824
Flores	BLEU		31.253	30.5207	30.253	29.351
	ChrF++		58.899	58.032	57.602	56.751

英语模型这一优势，我们可以推测，在知识迁移方面，语言上相近的双语模型比多语言模型更有用。不过，为了证明这一点，我们还需要进行更多实验，例如测试以荷兰语（高资源语言）为目标语言的林堡语（来自荷兰的低资源语言）模型，以及以意大利语（高资源语言）为目标语言的罗曼什语（来自瑞士的低资源语言）模型。

在时间性能方面，由于微调不会影响推理时间，这意味着知识蒸馏在提升性能的同时，也使模型具备了 OPUS-MT 模型轻量且快速推理的优势。根据表 I，我们模型的推理速度（即 OPUS-MT-MUL-EN 的推理时间）比 NLLB-large 快 30 倍，比 NLLB-Distilled 快 6 倍。

H. 对研究问题的回答

RQ1. 低资源机器翻译的最佳解决方案是什么?

NLLB-large 模型在翻译准确率方面明显优于其他模型或基线。在我们的解决方案中，蒸馏模型的表现远胜于伪模型和非 NLLB 基线（即表 I 中的 OPUS-MT 和 M2M）。此外，它们的表现与使用大量平行数据进行微调的相同基础模型相近。因此，要回答这个研究问题，使用轻量级模型实现准确的低资源语言翻译，最佳技术是**知识蒸馏**。

RQ2. 从计算效率和推理时间来看，面向低资源语言的最佳神经机器翻译模型是什么?

鉴于通过知识蒸馏获得的**OPUS-MT-DE-EN**模型推理时间和模型规模最小，且性能略优于*OPUS-MT-MUL-EN*，我们可以有把握地说，**OPUS-MT-DE-EN** 是本文中的最佳解决方案。

I. 讨论

首先，我们的研究表明，借助超大规模多语言模型进行知识蒸馏，即使只有单侧卢森堡语语料库（RTL 新闻或维基百科），也能训练出高性能的双语机器翻译模型。其次，我们发现，即使使用从德语到卢森堡语的伪翻译（这两种语言在语言学上相近），在应用词典和语法调整的情况下，也很难训练出有效的无监督翻译模型。

我们还发现，相关语言的模型作为迁移学习的基础，比多语言模型更具优势。为验证这一观察结果，还需要针对其他语言进行更多对比实验，例如林堡语、立陶宛语、芬兰语等。

此外，经一位卢森堡语母语者人工检查后，我们发现 Tatoeba 数据集过于简单，缺乏复杂性。另一方面，Flores-200 缺乏名词多样性。Tatoeba 质量较低并不意外，因为它是一个社区构建的数据集。若能有一个更好的评估数据集，将会很受欢迎。

最后需要指出，直到本文写作的后期，我们才发现原始 OPUS-MT-MUL-EN 模型在预预训练中使用了一些 Tatoeba 数据集，但我们不确定是否使用了卢森堡语数据集。不过，在该数据集上进行评估时，我们没有观察到该模型相比 OPUS-MT-DE-EN 有任何优势。

V. 结论

本文提出了两种技术，用于为低资源语言（LRL）翻译构建轻量级模型。我们的研究表明，基于大型模型进行知识蒸馏，可以获得性能优异的低资源语言小型模型。我们的模型体积更小、速度更快，且性能几乎与大型多语言 NLLB 模型相当。

未来，我们计划改进伪翻译技术，并在其他平行数据稀缺的低资源语言中测试知识蒸馏。我们还希望构建并评估英语到低资源语言的翻译模型。

参考文献

[1] “人工智能首次为闽南语这一没有书写系统的语言提供翻译服务”，https://tech.fb.com/artificial-intelligence/2022/10/ai-translation-unwritten-language，访问日期：2022-11-05。 [2] M. R. Costa-juss`a、J. Cross、O. C.elebi、M. Elbayad、K. Heafield、K. Heffernan、E. Kalbassi、J. Lam、D. Licht、J. Maillard 等，“不让任何语言掉队：扩大以人为中心的机器翻译规模”，*arXiv* 预印本 *arXiv:2207.04672*，2022。 [3] W.-J. Ko、A. El-Kishky、A. Renduchintala、V. Chaudhary、N. Goyal、F. Guzm´an、P. Fung、P. Koehn 和 M. Diab，“在没有平行数据的情况下，让高资源神经机器翻译模型适应低资源相关语言”，*arXiv* 预印本 *arXiv:2105.15071*，2021。 [4] G. Neubig 和 J. Hu，“神经机器翻译向新语言的快速适应”，*arXiv* 预印本 *arXiv:1808.04189*，2018。 [5] M. Johnson、M. Schuster、Q. V. Le、M. Krikun、Y. Wu、Z. Chen、N. Tho- rat、F. Vi´egas、M. Wattenberg、G. Corrado 等，“Google 的多语言神经机器翻译系统：实现零样本翻译”，计算语言学协会会刊，第 5 卷，第 339–351 页，2017。 [6] B. Zoph、D. Yuret、J. May 和 K. Knight，“面向低资源神经机器翻译的迁移学习”，*arXiv* 预印本 *arXiv:1604 02201*，2016。。[7] C. Lothritz、B. Lebichot、K. Allix、L. Veiber、T. F. D. A. Bissyande、J. Klein、A. Boytsov、A. Goujon 和 C. Lefebvre，“Luxembert：在卢森堡语语言模型预训练中简单实用的数据增强”，载于*2022* 年语言资源与评估会议论文集，2022，第 5080–5089 页。 [8] N. Pourdamghani、N. Aldarrab、M. Ghazvininejad、K. Knight 和 J. May，“翻译翻译腔：无监督机器翻译的两步法”，*arXiv* 预印本 *arXiv:1906.05683*，2019。 [9] G. Hinton、O. Vinyals 和 J. Dean，“提炼神经网络中的知识”，*stat*，第 1050 卷，第 9 页，2015。 [10] D. Gierschek，“卢森堡语用户评论中的情感检测”，博士论文，卢森堡大学，卢森堡埃施叙尔阿尔泽特，2022。 [11] P. Gilles，摩泽尔法兰克语中 *n* 的音系删除¨（“艾费尔规则”）。方言学韵律研究的一项贡献，2006 年 1 月，第 29–68 页。

[12] M. R. Costa-juss`a、J. Cross、O. C.elebi、M. Elbayad、K. Heafield、K. Heffernan、E. Kalbassi、J. Lam、D. Licht、J. Maillard 等，“不让任何语言掉队：扩大以人为中心的机器翻译规模”，*arXiv* 预印本 *arXiv:2207.04672*，2022。 [13] M. Artetxe 和 H. Schwenk，“用于零样本跨语言迁移及其他任务的大规模多语言句子嵌入”，计算语言学协会会刊，第 7 卷，第 597–610 页，2019 年 9 月。 [在线]。可访问：https://doi.org/10.1162/tacl a 00288[14] J. Tiedemann 和 S. Thottingal，“OPUS-MT——为世界构建开放式翻译服务”，载于第 *22* 届欧洲机器翻译协会 *(EAMT)* 年会论文集，葡萄牙里斯本，2020。 [15] A. Fan、S. Bhosale、H. Schwenk、Z. Ma、A. El-Kishky、S. Goyal、M. Baines、O. Celebi、G. Wenzek、V. Chaudhary、N. Goyal、T. Birch、V. Liptchinsky、S. Edunov、E. Grave、M. Auli 和 A. Joulin，“超越以英语为中心的多语言机器翻译”，*arXiv* 预印本，2020。