使用 Transformers 对波斯语—英语机器翻译数据集进行评估

Amir Sartipi伊斯法罕大学
amirsartipi.msc@eng.ui.ac.ir

Meghdad Dehghan伊斯法罕大学
meghdadd78@gmail.com

Afsaneh Fatemi伊斯法罕大学 a_fatemi@eng.ui.ac.ir

arXiv:2302.00321v1 [cs.CL] 1 Feb 2023

摘要

如今，许多研究人员都在关注机器翻译（MT）。然而，尽管人们已对英语等资源丰富的语言开展了大量研究，波斯语机器翻译仍未得到充分探索。此外，尽管针对一些波斯语数据集的统计机器翻译研究已取得相当进展，但目前尚无适用于每个语料库的、针对基于 Transformer 的文本到文本模型的标准基线。本研究收集并分析了用于波斯语-英语翻译的最受欢迎且最有价值的平行语料库。此外，我们分别在每个数据集上对两种最先进的基于注意力机制的序列到序列模型进行了微调和评估（共 48 项结果）。我们希望本文能帮助研究人员将波斯语译为英语及英语译为波斯语的机器翻译结果与标准基线进行比较。

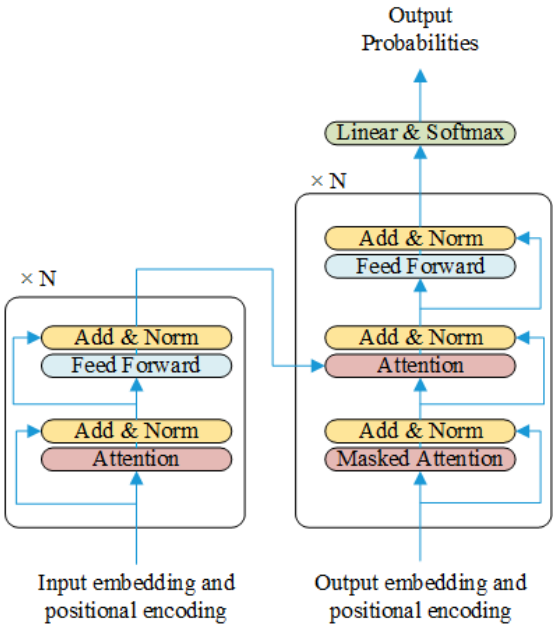


图 1：Transformer 模型架构

1 引言

机器翻译的主要目的是将文本从一种语言翻译成另一种语言。此前，统计语言模型曾被视为这一任务的前沿（Brown 等人，1993年；Koehn，2009年；Lopez，2008年）。然而，由于当前可用数据量庞大，神经机器翻译（Bahdanau 等人，2015年；Kalchbrenner 和 Blunsom，2013年；Wu 等人，2016年；Cho 等人，2014年）如今已超越统计方法。随后，Vaswani 等人（2017年）提出了一种仅基于注意力机制的全新简洁网络架构，用以取代当时占主导地位的、基于循环神经网络和卷积神经网络的序列转换模型。编码器部分

Transformer 架构已被广泛应用于Devlin 等人（2019年）和Liu 等人（2020年（b））提出的模型中，这些模型都在大量未标注文本上进行了预训练。 Raf- fel 等人（2020年）研究了自然语言处理领域的迁移学习策略，促使迁移学习成为自然语言处理中的一种强大技术。该研究提出了一种将所有语言任务转换为文本到文本格式的系统，称为 T5。mT5 是 T5 模型的多语言变体，使用一种基于 Common Crawl 的新数据集进行了预训练，该数据集包含 101 种语言（Xue 等人，2021）。为避免在数千项任务上训练时出现过拟合，Costa-jussà 等人（2022年）提出了多项架构和训练改进。他们使用人工翻译的基准数据集 Flores-200，来评估

超过 40,000 种不同翻译方向上的表现。与此前最先进的序列到序列模型相比，他们的模型 BLEU 分数提升了 44%。这两个模型（谷歌 T5 和 Meta NLLB）都采用了 Transformer 架构，并对编码器或解码器部分进行了一些修改和改进。Transformer 架构如图1所示。

本文的目的可概括如下：

1. 我们回顾统计机器翻译和神经机器翻译系统，以及相关数据集。

2. 我们公开了所有实验结果，包括最后一个模型检查点、最佳模型检查点、模型预测、训练和开发阶段的历史记录以及执行时间，均可在 Hugging Face[1] 上获取；代码也可在 GitHub[2] 代码仓库中获取。

3. 我们为波斯语-英语机器翻译任务建立基线，以供未来研究进行比较。

4. 我们研究实例数量对 BLEU分数的影响。

本文其余部分安排如下。在第2 节中，我们总结波斯语-英语机器翻译的既有方法。第 3 节介绍实验所用的主流语料库。此外，我们还在本节详细分析这些语料库的统计数据。第4节针对每个数据集进行语言模型的大量实验，且实验在两个方向上开展。第5节讨论本研究面临的挑战，并分析模型的预测结果。最后，第6节给出本研究的结论。

2 相关工作

关于以往的研究，已有若干研究探讨了英语到阿拉伯语（Nagoudiet al.，2022年）、法语（Tian 等人，2022年；Dione 等人，2022年；Liu 等人，2020a）和俄语（Yu，2019年；Littell 等人，2019年），这些研究侧重于将 Transformer 作为基础架构，并呈现结果。不过，有一些波斯语-英语数据集并未提供任何关于语言模型的结果。

本节将探讨波斯语-英语机器翻译方面的既有研究。首先，我们介绍机器翻译的两种方法：统计方法和神经网络方法，并概述这两个领域的近期研究。然后，我们回顾波斯语-英语机器翻译平行语料库的相关尝试。

统计机器翻译系统的基线。 波斯语-英语统计机器翻译系统的结果最早见于 PersianSMT（Pilevar 和 Faili，2010年）。他们采用基于短语的统计机器翻译系统，并以电影字幕领域作为平行语料库的主要资源，获得了相应结果。此外，Bakhshaei 等人(2010年)也报告了基于短语的波斯语-英语统计机器翻译系统的结果。他们测试了统计机器翻译系统参数的不同取值，并比较了各参数取值对应的结果。Mohaghegh 和 Sarrafzadeh(2010年)以及 Mohaghegh 等人(2010年)针对不同规模的语言模型语料库，报告了统计机器翻译系统的结果。他们得出结论，使用更大规模的语料库训练统计机器翻译系统能够取得更好的结果。Mohaghegh 等人(2011年)创建了一个名为 NSPEC 的组合平行语料库，其统计机器翻译系统取得的结果优于他们此前的研究。

Pilevar （2011年）创建了一个基于规则的统计机器翻译系统，随后进行了统计编辑，并报告了该系统的结果。他们的新方法优于现有的基于规则的统计机器翻译系统，但统计机器翻译系统仍比他们的方法更有效。 Mo-haghegh （2012）比较了两种统计机器翻译系统：层次型系统（Joshua）和经典系统（Moses）。他们报告了两个翻译方向的结果；然而，仅在英语到波斯语的翻译方向上使用层次型系统，取得了更好的结果。

Jabbari 等人（2012）创建了一个新语料库，其统计机器翻译系统的结果优于此前的结果。Mansouri 和 Faili（2012）比较了多个统计机器翻译系统，并使用最大熵分类器改进现有的最先进的统计机器翻译系统。 Rasooli 等人（2013年）指出，分割波斯语动词是有效的，并能提高 BLEU分数。Passban 等人（2015）改进了现有的 TEP 语料库，并创建了 TEP++。他们还在新语料库上取得了结果，并将其与 TEP 和 Mizan 等其他语料库进行了比较。研究结果超过了此前在 TEP 和 Mizan 语料库上取得的结果。在他们的研究中，

[1] https://huggingface.co/ [2] https://github.com/

Mizan

"A wonderful privilege it was to be thus admitted into the soul of a man of genius, to be allowed to share the ecstasies and the agonies of his inmost life."

چه خوشبختی بزرگی بود که انسان در عمق روح نابغه‌ای نفوذ می‌کرد و در نشئه و شکنجه زندگی درو نیش سهیم می‌گردید.

Bible

Jesus said to them, Can the friends of the bridegroom mourn, as long as the bridegroom is with them? But the days will come when the bridegroom will be taken away from them, and then they will fast.

عیسی بدیشان گفت: آیا پسران خانه عروسی، مادامی که داماد با ایشان است، می‌توانند ماتم کنند؟ و لکن ایامی می‌آید که داماد از ایشان گرفته شود؛ در آن هنگام روزه خواهند داشت.

Quran

The Creator of the heavens and the earth -- how should He have a son, seeing that He has no consort, and He created all things, and He has knowledge of everything?

نوپدیدآرنده آسمانها و زمین است. چگونه او را فرزندی باشد و حال آنکه برای همسری نبوده است، و هر چیزی را آفریده و او به هر چیزی داناست.

PEPC Bidirectional

Spencer was born on 31 October 1929 in Santa Lucia a historical rione of the city of Naples

باد اسپنسر یا کارلو پدرسولی بازیگر ایتالیایی و متولد ۳۱ اکتبر ۱۹۲۹ در ناپل است

PEPC One Directional

The School forms part of the Waterloo campus on the South Bank of the River Thames and is now one of the largest schools in the university

مدرسه به شکل بخشی از دانشگاه واترلو در کرانه جنوبی رودخانه تیمز بوده و در حال حاضر یکی از بزرگترین مدارس در دانشگاه کینگز لندن است

TEP

"from false accusing me of things that you know i did not , could not have done ."

اتهام دروغي كه به من واسه چيزي زدي كه خودت ميدوني من انجام ندادم ، نمي تونستم انجام بدم ، مبرا كني .

TEP ++

my father died to uphold the truce with your world . you must honor his noble intentions .

پدر من به خاطر حمایت از عهدنامه برای دنیایی شما کشته شد . شما باید به خاطر این کارش به ان افتخار کنین

Open Subtitles

"Had We so wished, nothing could have hindered Us from replacing you by others like yourselves, or transforming you into beings you know nothing about."

بر این که امثالتان را به جای شما قرار دهیم و شما را در جهان دیگری که نمی‌دانید آفرینش تازه‌ای بخشیم

图 2：每个数据集中英语（上）和波斯语（下）两侧实例的示例

	波斯语						英语					
	avg	min	max	92%	all	独特	avg	min	max	92%	all	独特
Mizan	13	1	232	26	13,464,236	131,751	13	0	226	26	13,360,397	259,182
《圣经》	28	3	124	48	1,796,084	18,166	23	2	100	38	1,428,716	40,202
《古兰经》	29	1	373	61	30,235,077	28,380	33	1	772	74	34,227,828	92,976
PEPC 双向	20	7	178	35	4,163,011	169,637	21	7	153	36	4,354,619	142,792
PEPC 单向	22	7	178	37	3,539,183	158,707	21	7	153	36	3,359,635	138,489
TEP	8	1	37	14	716,113	22,710	7	1	33	14	684,242	36,634
TEP ++	7	1	34	13	4,445,543	92,037	8	0	32	14	4,720,821	57,753
OPUS-100	10	1	1,487	21	10,284,744	155,874	9	1	839	20	9,524,220	342,979

表1：数据集的一般统计信息

	训练集	dev	test	all
Mizan	1,006,430	5,000	10,166	1,021,596
《圣经》	51,329	5,000	5,704	62,033
《古兰经》	1,013,756	5,000	10,240	1,028,996
PEPC 双向	175,442	5,000	19 ,494	199,936
PEPC 单向	138,005	5,000	15,334	158,339
TEP	72,748	5,000	8,084	85,832
TEP ++	515,925	5,000	57,326	578,251
OPUS-100	1,000,000	2,000	2,000	1,004,000

表 2：训练集\开发集\测试集中的实例数量

Kashefi(2018) 计算了统计机器翻译系统在其所用语料库（Mizan）上的BLEU分数。他们分别在域内和域外测试集上取得了结果。

神经机器翻译系统的基线。 已有多项研究尝试为使用神经机器翻译系统进行波斯语—英语机器翻译提出基线。 Bastan 等人 (2017年) 使用神经机器翻译（NMT）系统研究了翻译和音译两项任务。他们在 NMT 架构中使用了不同层数的循环神经网络。此外，他们通过调整代价函数并对波斯语语料进行预处理，提升了结果。与现有的神经机器翻译系统相比，Zare- moodi 等人 (2018) 和 Zaremoodi和 Haffari(2018) 表明，多任务学习方法能够提升波斯语等低资源语言的机器翻译效果。PasriNLU 首次使用神经语言模型进行波斯语与英语之间的机器翻译（Khashabi 等人，2021）。他们在自己创建的一个基准数据集的部分数据上，对 Google mT5 文本到文本模型的四种变体进行了微调。微调过程中使用的训练数据集整合了四个语料库，以提高泛化能力。

3 个数据集

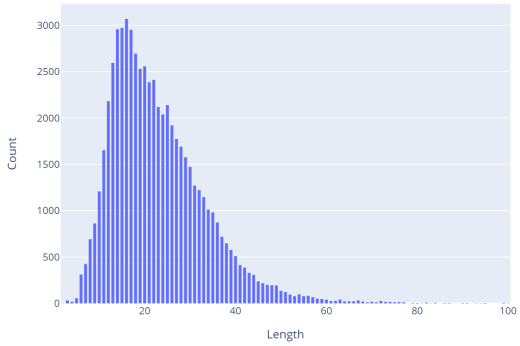
绝大多数关于机器翻译任务的研究和基准测试都是基于 WMT数据集开展的（Bojar等人，2014年）。此外，还有 OPUS-100（Zhang 等人，2020年）和 OpenSubtitles（Lison和 Tiedemann，2016年）等数据集，分别包含 60 种和 100 种语言，并用于其他语言的机器翻译任务。

对于波斯语-英语语言对，我们收集了九个数据集，用于通过神经序列到序列模型进行微调，并获得各数据集的结果。此外，ParsiNLU 是一组针对波斯语的语言理解任务，其中包括机器翻译（Khashabi 等人，2021）。在他们工作的机器翻译部分，他们整合多个语料库，构建了一个大型平行语料库。训练数据集包含四个领域：他们的问句改写任务中的问题、Mizan 语料库、TEP 语料库和 Global Voice 语料库。训练数据集包含近 160 万条记录。评估集由《古兰经》、Mizan、《圣经》和 QPP 数据集组成，包含约 4.7 万个句子。

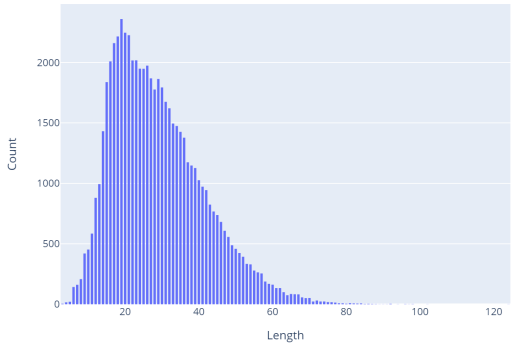
下面将逐一介绍收集到的各个数据集，并考察其主要属性。

《**古兰经**》。《古兰经》主要是一本阿拉伯语书籍，已被翻译成多种语言。 Tiede-mann (2012) 提出了 Tanzil 项目中的 Tanzil 数据集，并将其作为 OPUS 项目的一部分。该数据集包含 42 种语言。该数据集的波斯语-英语语言对包含近 100 万个句对和 5702 万个单词。

《**圣经**》。《圣经》是另一本被译成多种语言的宗教书籍。作为 OPUS 项目的一部分，圣经数据集以 100 种语言发布（Tiedemann，2012）。该数据集中的波斯语—英语语言对包含近 62,000 个句对和 289 万个词。



(a) 英语部分



(b) 波斯语部分

图 3：《圣经》中每个句子的词元分布

TEP. TEP（德黑兰英语-波斯语）是另一个由电影字幕构成的平行语料库。从 OpenSubtitles 收集了近 21000 个字幕文件，去除重复文件后仅保留了 1200 对字幕文件。最终数据集包含超过 550,000 行文本（Pilevar 等人，2011年）。

TEP++。 Passban 等人介绍了 TEP 语料库的一个改进版本，名为 TEP++。Passban 等人 (2015年）。他们指出，TEP 语料库存在噪声，并尝试在新语料库中解决这一问题。他们还通过使用 TEP++ 语料库，改进了统计机器翻译系统的结果。该语料库的波斯语和英语各包含近 570,000 个对齐句子和近 500 万个词元。

OPUS-100。 OPUS-100 是由电影字幕、GNOME 文档和《圣经》数据集拼接而成，包含 100 种语言和 99 个语言对，所有语言对均以英语为源语言或目标语言（Zhang 等人，2020年）。

PEPC。 PEPC 是另一个波斯语-英语平行语料库，取自维基百科文档（Karimi 等人，2018）。他们采用双向和单向方法从维基百科中提取文档，因此根据提取方法提出了两个版本的数据集。双向 PEPC 数据集包含近 200,000 个句对，而单向 PEPC 数据集包含近 160,000 个句对。

Mizan. Mizan 在发布时是最大的波斯语语料库。它由文学名著编纂而成，包含一百多万个句对，波斯语和英语文本各有超过 2300 万个单词（Kashefi，2018）。

我们从每个语料库中随机选取了一个实例，如图 2所示。可以看出，OPUS-100 数据集将首字母大写的“We”和“Us”置于句子中间；波斯语字幕中有一处听写错误；逐词翻译与其含义也并非完全对应。有些句子带有引号，或者在英语中以小写字母开头。数据集的这些特征可能会影响评估结果。

我们使用 SPARK NLP（Kocaman 和 Talby，2021）来提供有关数据集的一般统计信息。根据这些信息，可以更准确地选择序列长度等参数。表1中的 max 列表示句子中允许的词元数量上限。由于每个数据集都包含少量较长的序列，这些序列可视为离群值，并可通过更精确地设定长度加以截断，因此这个上限可能并不合适。因此，我们为每个数据集计算了一个覆盖 92% 数据集的数值。换言之，92% 的句子所含词元数量小于或等于该数值。就每个句子的词元数量而言，该数值远低于最大值。此外，该表还列出了每个数据集的平均词元数和最小词元数，以及波斯语和英语语料库的词元总数和唯一词元总数。

4 实验

为了构建我们的网络，我们使用了 PyTorch（Paszke 等，2019年）和 Hugging Face 的 Transformers 库（Wolf 等，2020年）作为实现工具。

	英语-波斯语			波斯语-英语		
	mt5-small	mt5-base	nllb-distilled	mt5-small	mt5-base	nllb-distilled
Mizan	12.22	12.69	15.00	16.29	16.70	18.05
《圣经》	13.93	22.06	69.78	16.28	18.83	49.93
《古兰经》	4.79	4.97	18.10	10.39	10.04	27.65
PEPC 双向	7.10	7.21	13.13	10.28	10.22	17.01
PEPC 单向	5.37	5.71	13.20	8.82	9.85	16.84
TEP	11.70	14.11	16.06	13.63	23.64	26.74
TEP ++	21.02	23.09	26.44	30.14	31.63	35.98
OPUS-100	10.81	10.46	11.62	20.66	20.91	24.16

表 3：语言模式下英语到波斯语（英语-波斯语）和波斯语到英语（波斯语-英语）的评估 ls

数据集划分。 表2 列出了各数据集的实例总数及训练集/开发集/测试集划分情况。对于 OPUS-100 数据集，我们使用了预先定义的数据划分。对于其他数据集，我们手动将全部数据划分为训练集、开发集和测试集。首先，我们打乱了各数据集中的所有实例，以随机化其顺序。然后，对于实例数超过一百万的数据集，我们将全部实例的 1% 作为测试集，选取 5,000 个实例作为开发集，其余实例作为训练集。

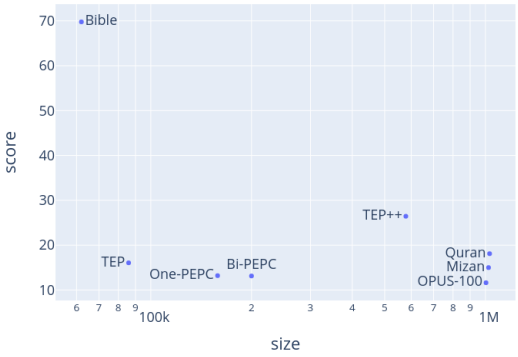
超参数： Khashabi 等人(2021) 在微调阶段使用 1e-3 的学习率（lr）。本研究采用相同的 lr，并使用 ADAMW 优化器对模型进行 7 个训练轮次的微调（Loshchilov 和 Hutter, 2019年）。为了在训练阶段选择序列长度，我们考虑了能涵盖数据集 92% 内容的序列长度。此外，我们绘制了句子数量与每个句子的词元数量之间的关系，以便选择合理的序列长度。图3展示了《圣经》数据集的相关示例。

模型 我们使用的序列到序列模型之一是 mT5，它包含波斯语嵌入。另一种文本到文本模型是 NLLB，其性能优于此前最先进的模型。由于这类模型的参数数量庞大，且需要大量计算资源，我们仅对 2 个谷歌 mT5 变体 { mT5 small, mT5 base} 和一个 Facebook NLLB 模型 {蒸馏版 NLLB} 进行了微调。下面总结这些模型的主要属性

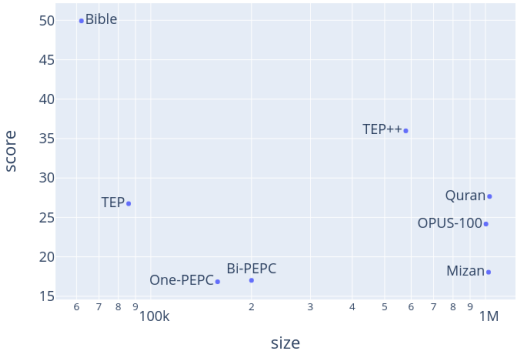
- **Google mT5：** Google T5 模型是一种基于文本到文本 Transformer 的语言模型。这意味着该模型的输入和输出都是文本。该模型可用于不同的任务，例如问答、机器翻译和文本分类。该模型的 mT5 版本在多语言 mC4 数据上进行了预训练，该数据包含 101 种语言，其中包括波斯语。该模型的 mT5-small 版本是最小的版本，只有 3 亿个参数。mT5-base 是第二小的模型，拥有 5.8 亿个参数。该模型最大的版本约有 130 亿个参数。

- **Meta NLLB：** NLLB 模型是当时最先进的文本到文本模型，其提出旨在提升低资源语言的机器翻译性能。它支持近 200 种语言的嵌入。该模型还采用基于 Transformer 的架构，并分为两种类型：Dense 和 MoE。Dense 类型会针对每个输入序列激活模型的所有参数，而 MoE 模型则只会针对每个输入激活一部分参数。NLLB 模型根据参数规模分为 5 个变体。最小的模型只有 6 亿个参数，属于 Dense 模型；最大的模型则属于 MoE 模型，约有 545 亿个参数。

评估指标： BLEU分数（Papineni 等人, 2002 年）多年来一直是评估机器翻译结果最常用的指标。该指标结合了不同 N 元大小的 N 元精确率，并采用句子长度惩罚。由于选择 BLEU 分数参数时存在多种配置，研究人员报告的不同基线结果并不太适合相互比较。例如，许多研究都没有报告最大 N 元的大小和分词方法。该



(a) 英语到波斯语方向



(b) 波斯语到英语方向

图4：根据数据集规模得出的 BLEU 分数最高值

SacreBLEU 指标由 Post (2018) 提出，旨在解决其中一些问题，并建立一种标准指标，以便在不同研究中进行比较。

训练过程： 由于模型可以同时针对波斯语和英语进行微调，因此每项实验只考虑一个方向。在训练阶段，我们在每个训练轮次结束时评估模型。根据模型在开发数据集上的评估指标值选择最优模型。在训练模型前对数据进行预处理十分重要，但由于我们希望为这些数据集建立基线，因此没有进行预处理。对数据集应用数据清洗方法可以改进机器翻译系统。

硬件： 我们的谷歌模型使用 float32 在 TITAN RTX 和 RTX 3090 Ti GPU 上进行了微调。由于 Meta 模型需要更强的计算能力，我们使用 NVIDIA V100 GPU 来运行该模型。后者使用 PyTorch 的一项名为自动混合精度的功能进行了微调，从而降低了 GPU 消耗和执行时间；与使用 float32 相比，该功能采用 float16。

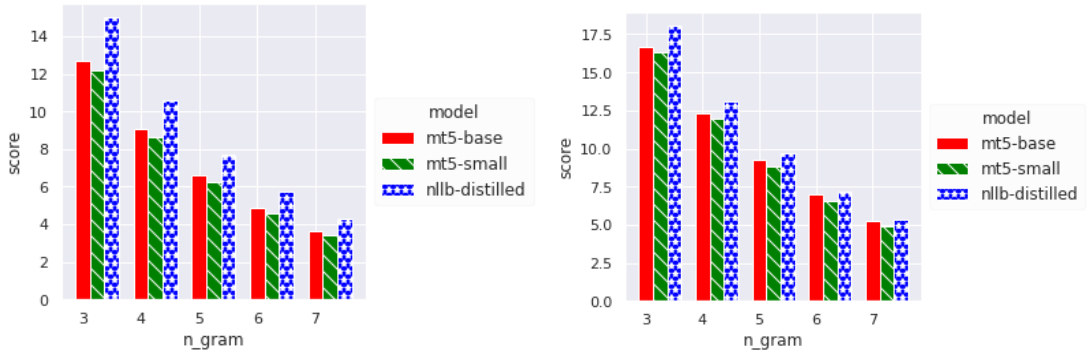
结果 我们使用 SacreBLEU 作为评估指标来评估微调后的模型。由于计算能力有限，针对测试数据生成的句子的最大序列长度低于该值。无法精确比较真实测试数据与预测实例。为解决这一问题，在计算分数之前，我们截断了超过预测句子最大序列长度的测试实例。表 3显示了 SacreBLEU 的值，$N-gram=3$。

N 元组的取值是决定最终 BLEU 分数的重要因素。该指标将 N 元组作为给定文本样本中连续的 {N} 个项目序列。为避免歧义，并使结果便于与未来的研究进行比较，我们报告了 {3, 4, 5, 6, 7} 对应的 BLEU 指标。图 5 展示了三个模型的 N 元组与分数之间的关系，以比较其性能并确定 N 元组对性能的影响。正如预期的那样，N 元组取值较大时，结果低于取值较小时的结果。在所有数据集中，Meta NLLB 模型的表现均优于两个版本的 Google mT5 模型。

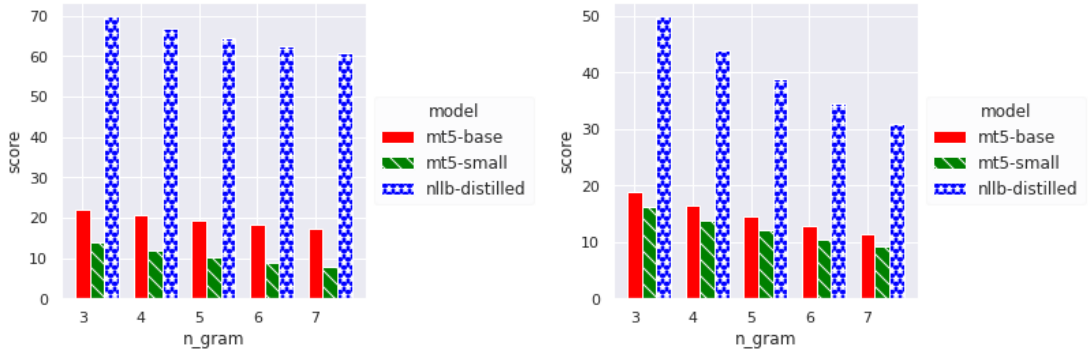
模型评估 7展示了关于训练和验证困惑度以及训练期间开发集 BLEU 分数的实验详细信息。

训练困惑度从第一个训练轮次到第二个训练轮次大幅下降，随后逐渐下降，直至第七个训练轮次。然而，验证困惑度从第一个训练轮次到第二个训练轮次下降得更快，之后则逐渐下降。在某些模型中，该值开始上升，模型因而出现过拟合。在这一阶段，困惑度值起初非常高，但经过一个训练轮次后便会下降。

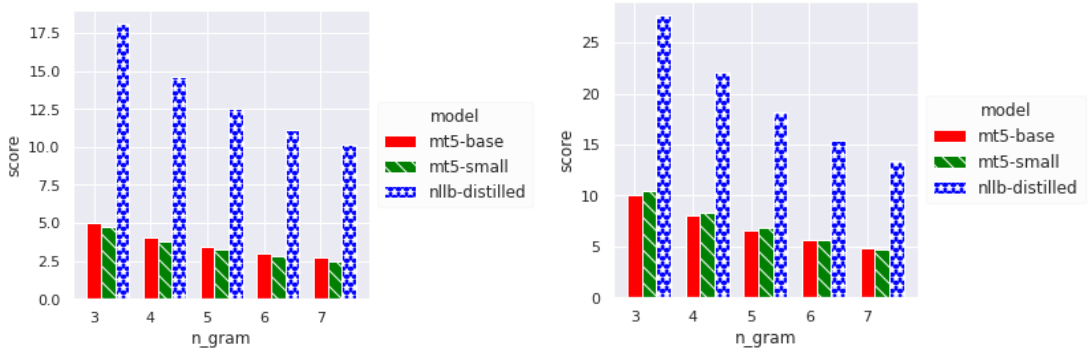
为了展示训练阶段 BLEU 分数的变化，并全面了解模型在各个数据集上的表现，我们按训练轮次计算了各开发集的 BLEU 分数。大多数模型的分数稳步上升，随后趋于下降或保持平稳。不过，在三组实验中，包括 mt5-small-fa-en 和 mt5-base-fa-en 的 PEPC 双向实验，以及一组 mt5- 单向实验，



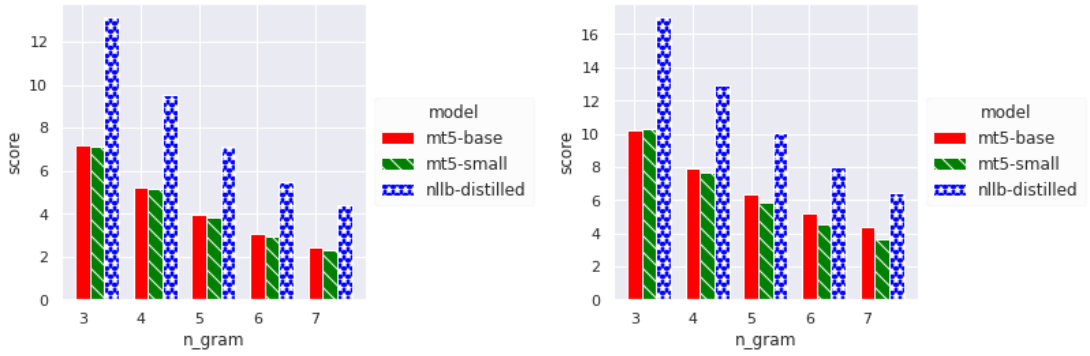
(a) Mizan



(b) 《圣经》

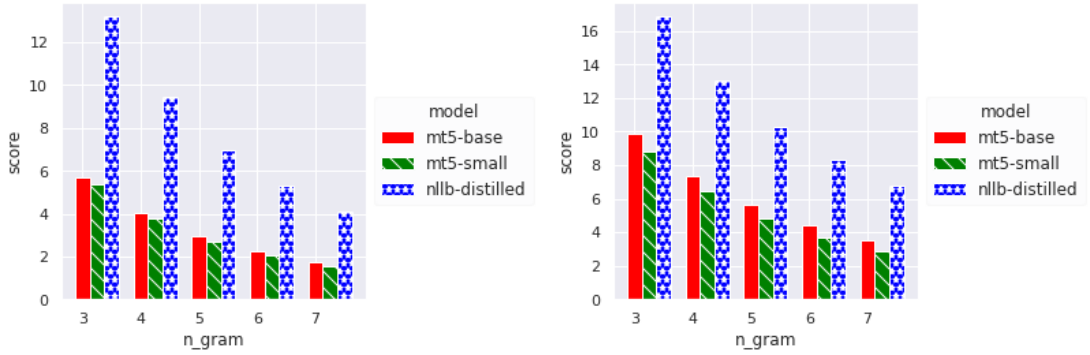


(c) 《古兰经》

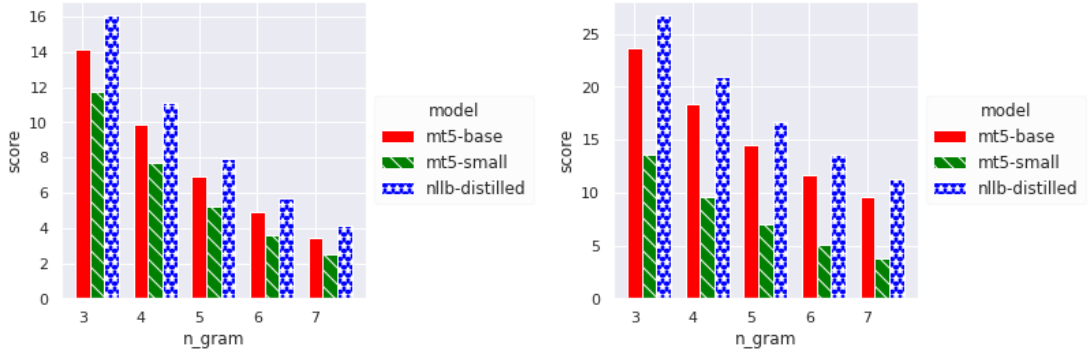


(d) PEPC 双向

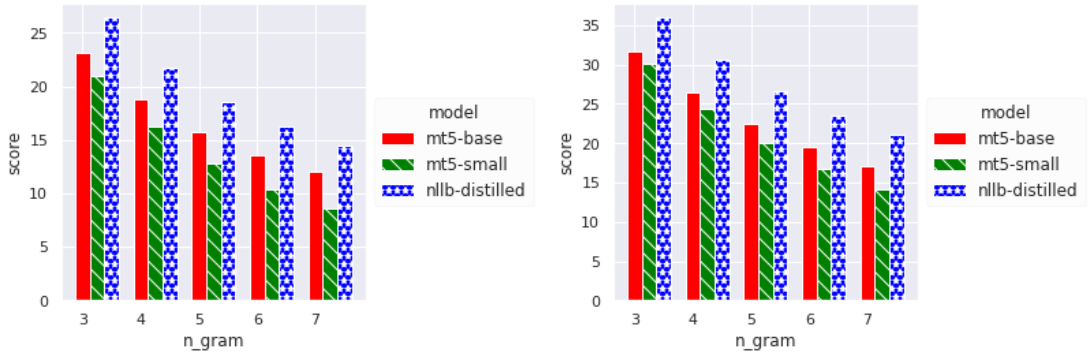
图5：不同 n 元语法下按翻译方向（左侧为英语到波斯语，右侧为波斯语到英语）和模型划分的 BLEU分数结果**第一部分**



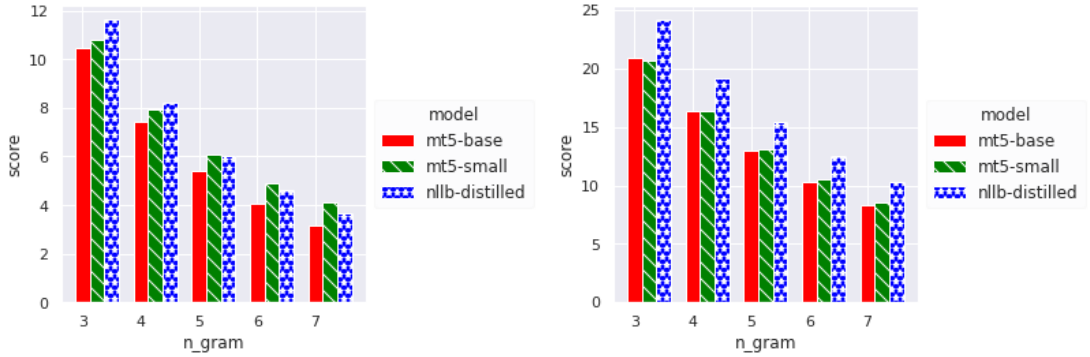
(e) PEPC 单向



(f) TEP



(g) TEP ++



(h) OPUS-100

图5：按翻译方向区分的不同 n 元语法的 BLEU 分数结果（左侧为英语到波斯语，右侧为波斯语到英语）和模型**第二部分**



图 6：训练实例数量对 mT5 small 模型在波斯语到英语翻译任务的评估数据集上表现的影响。

对于 base-en-fa，评估指标在训练轮次 2 时下降，随后迅速恢复。

5 讨论

本节将介绍对实验结果的一些见解。此外，我们还将根据实例数量讨论所用数据集的质量。图 6 展示了各数据集在两个方向上的最高 BLEU 分数与数据集规模之间的关系，从而便于更好地比较结果。

总体而言，像《古兰经》、OPUS-100 和 Mizan 这样拥有超过一百万个实例的数据集，其 BLEU 分数低于或几乎与规模较小的数据集相同，例如《圣经》、TEP、TEP++ 和 PEPC 的变体。

与 TEP 相比，TEP++ 数据集取得了更高的分数，这表明清理噪声实例并增加实例数量对数据集的结果有积极影响。相比之下，PEPC 数据集的不同版本在分数上没有显著差异。

尽管《圣经》是实例总数最少的数据集，但它在两个翻译方向上的得分都高于其他所有数据集。另一个值得一提的是，该数据集中实例的平均序列长度仅次于《古兰经》的平均序列长度，位居第二，但《古兰经》的得分却低得多。

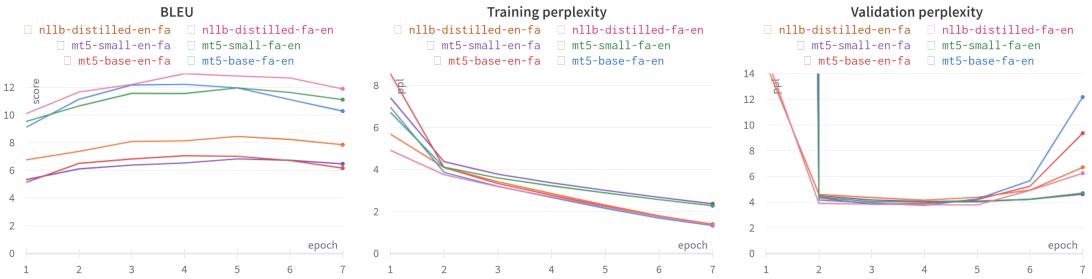
质量还是数量? 按照传统的提升机器翻译结果的方法，增加训练数据的规模预计会提高 BLEU 分数。然而，本研究表明，实例数量较多的数据集往往比实例数量较少的数据集获得更低的 BLEU 分数。因此，微调阶段所用数据的质量可能比实例数量更为重要。就质量而言，听写错误、翻译内容未对齐、标点错误，以及源语言和目标语言方向上的语序错误，都可能改变原意，并对最终评估结果产生负面影响。

我们测试了三个实例数超过一百万的数据集，以展示训练样本数量如何影响评估指标的数值。我们从这些数据集中抽取了 4 万和 8 万个实例，并对 Google mT5 small 模型进行了微调。根据这项实验，图 6 显示，随着实例数量的增加，模型的结果也会更好。

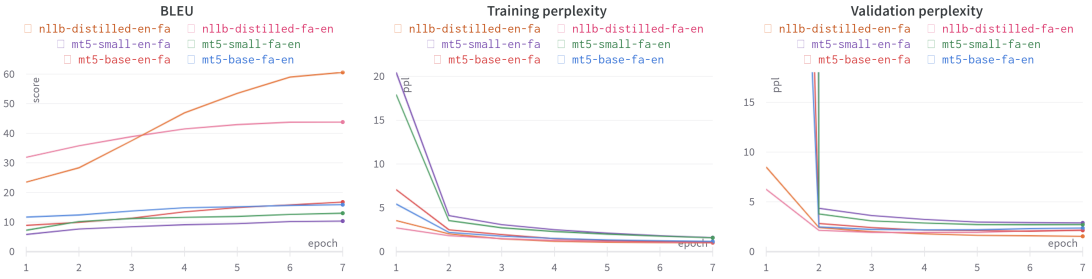
翻译方向 两个方向上获得的 BLEU 分数总体排序几乎完全相同。有几个因素值得考虑。《圣经》数据集在两个方向上都取得了最高的 BLEU 分数。然而，在英语到波斯语方向，OPUS-100 数据集的 BLEU 分数最低；而在波斯语到英语方向，单向的 PEPC 数据集的 BLEU 分数最低。尽管几乎所有数据集在波斯语到英语翻译中的表现都更好，但《圣经》数据集在英语到波斯语翻译中的表现明显更佳，BLEU 分数高出近 20%。在波斯语到英语翻译中，OPUS-100 数据集的表现明显优于 Mizan 数据集；而在相反方向上，Mizan 数据集的表现更好。

6 结论

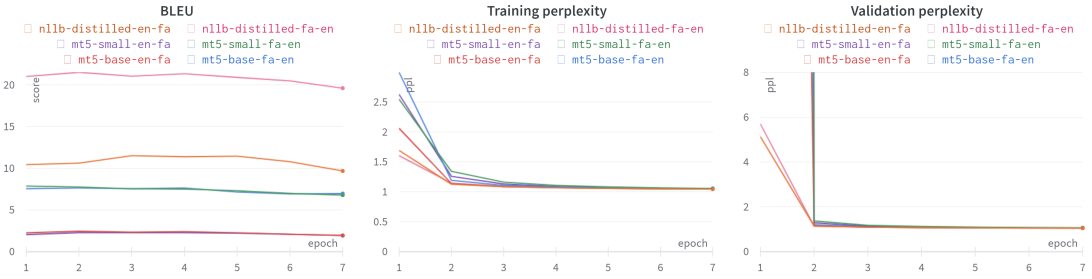
在本研究中，我们回顾了大部分波斯语-英语平行语料库，并为八个数据集建立了标准基线。我们使用基于 Transformer 架构的两种多语言序列到序列模型对这些数据集进行了评估。对 48 项实验的分析表明，《圣经》数据集和 PEPC 数据集的 BLEU 分数分别最高和最低。此外，我们得出结论，Meta 的基础版本显著优于以往基于 Transformer 的方法。研究结果还表明，在大多数实验中，波斯语到英语翻译的评估指标高于



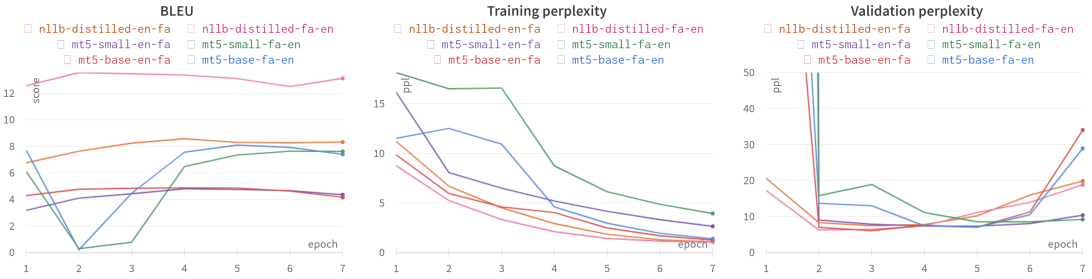
(a) Mizan



(b) 《圣经》

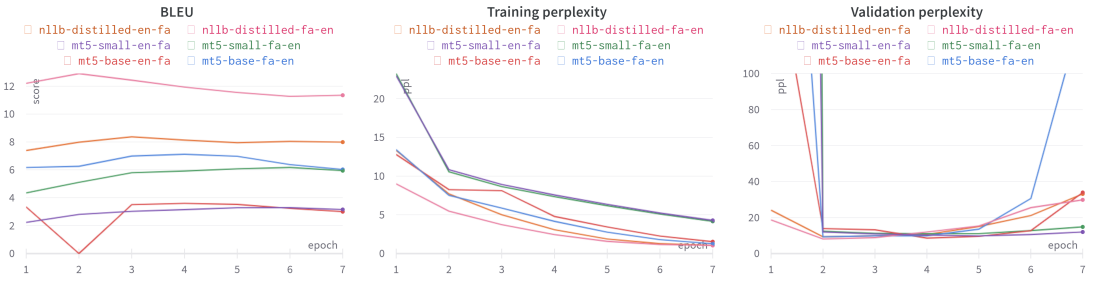


(c) 《古兰经》

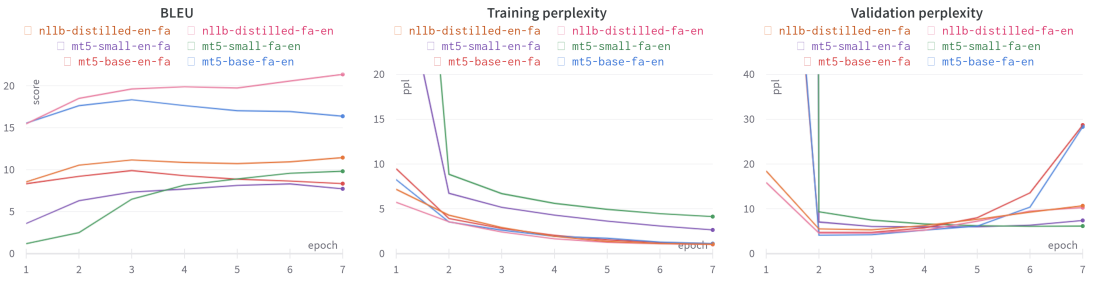


(d) PEPC 双向

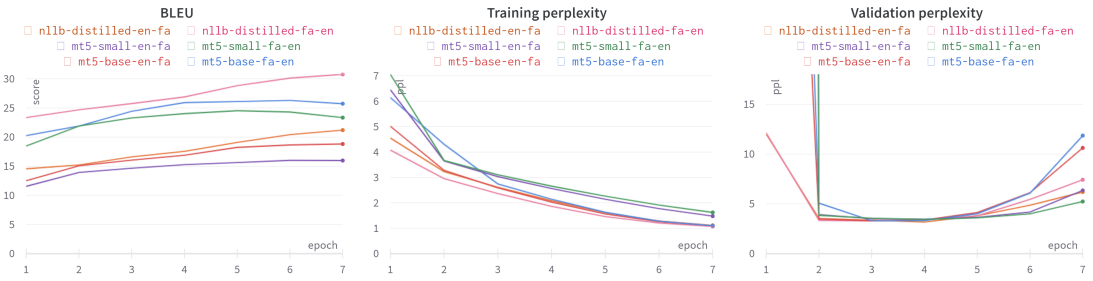
图 7：各数据集的 BLEU 分数、训练困惑度和验证困惑度。**第一部分**



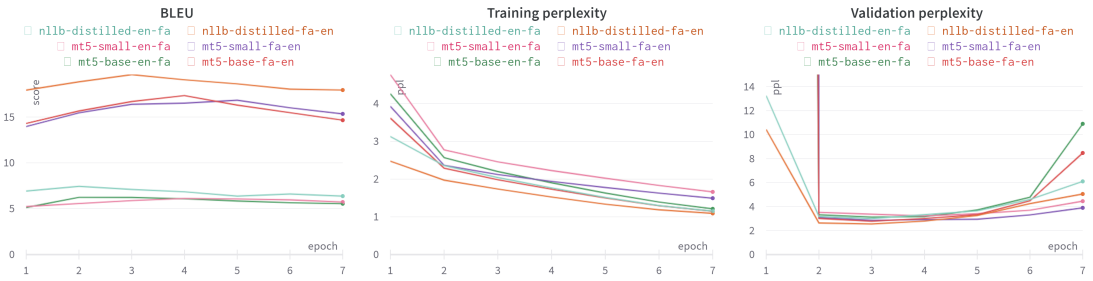
(e) PEPC 单向



(f) TEP



(g) TEP++



(h) OPUS-100

图 7：各数据集的 BLEU 分数、训练困惑度和验证困惑度。**第二部分**

从英语到波斯语的翻译。据我们所知，这是首项使用序列到序列模型分别为每个数据集建立基线的研究。我们希望这项研究能够帮助研究人员将其方法与基线进行比较，并针对波斯语进行具体评估。

致谢

本研究得到西莫尔格超级计算机——阿米尔卡比尔理工大学的资助，合同编号为ISI-DCE-DOD-Cloud-900808-1700。

参考文献

Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。2015年。通过联合学习对齐与翻译实现神经机器翻译。载于第三届表征学习国际会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集。

Somayeh Bakhshaei、Shahram Khadivi、Noushin Riahi 和 Hossein Sameti。2010年。探究影响波斯语-英语统计机器翻译系统的参数。载于*2010*年第五届国际电信研讨会，第985–991页。

Mohaddeseh Bastan、Shahram Khadivi 和 Moham- mad Mehdi Homayounpour。2017年。资源稀缺条件下的神经机器翻译：波斯语-英语案例研究。载于*2017*年伊朗电气工程会议（*ICEE*），第 1485–1490 页。

Ondˇrej Bojar、Christian Buck、Christian Federmann、Barry Haddow、Philipp Koehn、Johannes Leveling、Christof Monz、Pavel Pecina、Matt Post、Herve Saint-Amand、Radu Soricut、Lucia Specia 和 Aleš Tamchyna。2014年。2014年统计机器翻译研讨会论文集。载于第九届统计机器翻译研讨会论文集，第 12–58 页，美国马里兰州巴尔的摩。计算语言学协会。

Peter F. Brown、Stephen A. Della Pietra、Vincent J. Della Pietra 和RobertL.Mercer。1993年。统计机器翻译的数学：参数估计。《计算语言学》，19(2)：263–311。

Kyunghyun Cho、Bart van Merriënboer、Dzmitry Bah- danau 和 Yoshua Bengio。2014年。神经机器翻译的特性：编码器–解码器方法。载于《*SSST-8*：第八届统计机器翻译语法、语义与结构研讨会论文集》，第103–111页，卡塔尔多哈。计算语言学协会。

Marta R Costa-jussà、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Mail- lard 等。2022年。不让任何语言掉队：扩大以人为本的机器翻译规模。*arXiv* 预印本：*2207.04672v3*，页码 arXiv–2207。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于计算语言学协会北美分会*2019*年人类语言技术会议论文集：第*1*卷（长论文和短论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Cheikh M. Bamba Dione、Alla Lo、Elhadji Mamadou Nguer 和 Sileye Ba。2022年。低资源神经机器翻译：沃洛夫语最先进 Transformer 的基准测试<->法语。载于第十三届语言资源与评测会议论文集，第6654–6661页，马赛，法国。欧洲语言资源协会。

Fattaneh Jabbari、Somayeh Bakshaei、Seyyed Mohammad Mohammadzadeh Ziabary 和 Shahram Khadivi。2012。使用 AFEC 开发开放领域英语-波斯语翻译系统：阿米尔卡比尔波斯语-英语双语语料库。载于第四届阿拉伯文字语言计算方法研讨会，第17–23页，美国加利福尼亚州圣迭戈。美洲机器翻译协会。

Nal Kalchbrenner 和 Phil Blunsom。2013年。循环连续翻译模型。载于*2013*年自然语言处理经验方法会议论文集，第 1700–1709 页，美国华盛顿州西雅图。计算语言学协会。

Akbar Karimi、Ebrahim Ansari 和 Bahram Sadeghi Bigham。2018。从可比语料库中提取英语-波斯语平行语料库。载于第十一届国际语言资源与评测会议论文集（*LREC2018*），日本宫崎。欧洲语言资源协会（ELRA）。

Omid Kashefi。2018。Mizan：大型波斯语-英语平行语料库。*arXiv* 预印本 *arXiv:1801.02107v3*。

Daniel Khashabi、Arman Cohan、Siamak Shakeri、Pedra m Hosseini、Pouya Pezeshkpour、Malihe Alikhani、Moin Aminnaseri、Marzieh Bitaab、Faeze Brahman、Sarik Ghazarian、Mozhdeh Gheini、Arman Kabiri、Rabeeh Karimi Mahabagdi、Omid Memarrast、Ahmadreza Mosallanezhad、Erfan Noury、Shahab Raji、Mohammad Sadegh Rasooli、Sepideh Sadeghi、Erfan Sadeqi Azer、Niloofar Safi Samghabadi、Mahsa Shafaei、Saber Sheybani、Ali Tazarv 和 Yadollah Yaghoobzadeh。2021。

ParsiNLU：一套针对波斯语的语言理解挑战。《计算语言学协会汇刊》，9：1147–1162。

Veysel Kocaman 和 David Talby。2021。SPARK NLP：大规模自然语言理解。软件影响，第100058页。

Philipp Koehn。2009年。统计机器翻译。剑桥大学出版社。

Pierre Lison 和 Jörg Tiedemann。2016年。OpenSubtitles2016：从电影和电视字幕中提取大型平行语料库。收录于 第十届国际语言资源与评测会议论文集（*LREC'16*），第923–929页，波尔托罗日，斯洛文尼亚。欧洲语言资源协会（ELRA）。

Patrick Littell、Chi-kiu Lo、Samuel Larkin 和 Darlene Stewart。2019年。用于哈萨克语-俄语-英语神经机器翻译的多源Transformer。收录于 第四届机器翻译会议论文集（第*2*卷：共享任务论文，第*1*天），第267–274页，意大利佛罗伦萨。计算语言学协会。

Xiaodong Liu、Kevin Duh、Liyuan Liu 和 Jianfeng Gao。2020a。《用于神经机器翻译的超深Transformer》。*ArXiv*，abs/2008.07772。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettle moyer 和 Veselin Stoyanov。2020年（b）。Ro{bert}a：一种经过稳健优化的 {bert} 预训练方法。Adam Lopez。2008年。统计机器翻译。*ACM*计算综述 *(CSUR)*，40(3):1–49。Ilya Loshchilov 和 Frank Hutter。2019年。解耦权重衰减正则化。载于国际表征学习会议。Amin Mansouri 和 Heshaam Faili。2012。最先进的英语到波斯语统计机器翻译系统。载于第*16*届 *CSI* 国际人工智能与信号处理研讨会（*AISP 2012*），第174–179页。Mahsa Mohaghegh。2012。英语-波斯语层次统计机器翻译的进展。载于*NZCSRSC* 新西兰计算机科学研究生会议。Mahsa Mohaghegh 和 Abdolhossein Sarrafzadeh。2010年。英语-波斯语统计机器翻译中各种训练数据的性能评估。载于第*10*届文本数据统计分析国际会议（*JADT 2010*），意大利罗马。Mahsa Mohaghegh、Abdolhossein Sarrafzadeh 和 Tom Moir。2010年。改进英语-波斯语统计机器翻译的语言建模。载于第*4*届统计翻译句法与结构研讨会论文集，第75–82页。

Mahsa Mohaghegh、Abdolhossein Sarrafzadeh 和 Tom Moir。2011年。改进波斯语-英语统计机器翻译：领域适配实验。载于第*2*届南亚和东南亚自然语言处理研讨会论文集（*WS-SANLP*），第9–15页，泰国清迈。亚洲自然语言处理联合会。

El Moatez Billah Nagoudi、AbdelRahim Elmadany 和 Muhammad Abdul-Mageed。2022年。TURJU-MAN：用于神经阿拉伯语机器翻译的公共工具包。载于第五届开源阿拉伯语语料库和处理工具研讨会论文集（含《古兰经》问答和细粒度仇恨言论检测共享任务），第 1–11 页，法国马赛。欧洲语言资源协会。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002 年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第四十届年会论文集，第 311–318 页，美国宾夕法尼亚州费城。计算语言学协会。

Peyman Passban、Andy Way 和 Qun Liu。2015年。使用 TEP 基准测试波斯语的统计机器翻译性能++语料库。载于欧洲机器翻译协会第 *18* 届年会论文集，土耳其安塔利亚。欧洲机器翻译协会。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019年。PyTorch：一种命令式风格的高性能深度学习库。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett 编，神经信息处理系统进展第 *32* 卷，第 8024–8035 页。Curra n Associates, Inc. Abdol Hamid Pilevar。2011年。使用统计后编辑来改进基于规则的机器翻译系统的输出。《国际计算机科学与通信杂志》，330：330–000。Mohammad Taher Pilevar 和 Heshaam Faili。2010年。PersianSMT：首次尝试英语到波斯语的统计机器翻译。载于*JADT*，第 2010 卷，第 10th 页。Mohammad Taher Pilevar、Heshaa m Faili 和 Abdol Hamid Pilevar。2011年。TEP：德黑兰英语-波斯语平行语料库。载于计算语言学与智能文本处理，第 68–79 页，柏林、海德堡。施普林格柏林海德堡。Matt Post。2018。呼吁清晰报告 BLEU 分数。载于第三届会议论文集：

机器翻译：研究论文，第 186–191 页，比利时布鲁塞尔。计算语言学协会。

Colin Raffel、Noam Shazeer、Adam Roberts、Kather- ine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索统一文本到文本 Transformer 的迁移学习极限。《机器学习研究杂志》，21(140):1–67。

Mohammad Sadegh Rasooli、Ahmed El Kholy 和 Nizar Habash。2013年。波斯语到英语统计机器翻译的正字法与形态处理。载于第六届自然语言处理国际联合会议论文集，第1047–1051页，日本名古屋。亚洲自然语言处理联合会。

Taoling Tian、Chai Song、Jin Ting 和 Hongyang Huang。2022年。一种使用 Transformer 网络的法语到英语机器翻译模型。《计算机科学*Procedia*》，199:1438–1443。第八届信息技术与定量管理国际会议（ITQM 2020 2021）：新冠肺炎疫情后发展全球数字经济。

Jörg Tiedemann。2012年。OPUS 中的平行数据、工具和接口。载于第八届国际语言资源与评估会议（*LREC'12*）论文集，第 2214–2218 页，土耳其伊斯坦布尔。欧洲语言资源协会（ELRA）。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就够了。载于神经信息处理系统进展，第 30 卷。Curran Associates, Inc.。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。2020年。Transformers：最先进的自然语言处理。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，在线。计算语言学协会。

Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V. Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey、Jeff Klingner、Apurva Shah、Melvin John- son、Xiaobing Liu、Lukasz Kaiser、Stephan Gouws、Yoshikiyo Kato、Taku Kudo、Hideto Kazawa、Keith Stevens、George Kurian、Nishant Patil、Wei Wang、Cliff Young、Jason Smith、Jason Riesa、Alex Rud- nick、Oriol Vinyals、Greg Corrado、Macduff Hughes 和 Jeffrey Dean。2016年。谷歌神经机器翻译系统：弥合人类与机器翻译之间的差距。*CoRR*，abs/1609.08144。

Linting Xue、Noah Constant、Adam Roberts、Mi- hir Kale、Rami Al-Rfou、Aditya Siddhant、Aditya Barua 和 Colin Raffel。2021。mT5：大规模多语言预训练文本到文本转换器。载于《*2021*年北美计算语言学协会分会会议论文集：人类语言技术》，第483–498页，在线。计算语言学协会。

Doron Yu. 2019年。基于 Transformer 的英俄双向集成机器翻译系统。载于《第四届机器翻译大会论文集》（第 *2* 卷：共享任务论文，第 *1* 天），第 434–439 页，意大利佛罗伦萨。计算语言学协会。

Poorya Zaremoodi、Wray Buntine 和 Gholamreza Haffari。2018。多任务学习中的自适应知识共享：提升低资源神经机器翻译。载于计算语言学协会第 *56* 届年会论文集（第 *2* 卷：短论文），第 656–661 页，澳大利亚墨尔本。计算语言学协会。

Poorya Zaremoodi 和 Gholamreza Haffari。2018。双语资源稀缺场景下的神经机器翻译：一种深度多任务学习方法。载于计算语言学协会北美分会 *2018* 年会议论文集：人类语言技术，第 *1* 卷（长论文），第 1356–1365 页，美国路易斯安那州新奥尔良。计算语言学协会。

Biao Zhang、Philip Williams、Ivan Titov 和 Rico Sennrich。2020年。改进大规模多语种神经机器翻译和零样本翻译。载于计算语言学协会第*58*届年会论文集，第1628–1639页，在线。计算语言学协会。