

机器翻译和大语言模型研究进展

朱文昊 周昊 高长江 刘斯哲 黄书剑[†]

计算机软件新技术国家重点实验室, 南京大学

{zhuwh,zhouh,gaocj,liusz}@smail.nju.edu.cn, huangs@nju.edu.cn

摘要

机器翻译旨在通过计算机自动将一种自然语言翻译成另一种自然语言，这个过程对于机器翻译模型的语言理解、语言生成能力有着极高的要求。因此机器翻译一直以来都是一项极具研究价值和研究难度的自然语言处理任务。近期研究表明，大语言模型能够根据人类指令完成包括翻译在内的许多任务，在这一过程中展现出强大的语言理解和生成能力，为自然语言处理范式革新提供了新的可能。为了在大语言模型支持下更好地完成机器翻译任务，研究人员对大语言模型的机器翻译和多语言能力进行了大量的研究和分析。本文从以下三方面介绍相关研究热点和最新进展，包括：大语言模型翻译能力评估、大语言模型翻译能力激发、大语言模型在不同语言上的能力展现。

关键词： 机器翻译；大语言模型；情景学习；指令微调；多语言

Research Development of Machine translation and Large Language Model

Wenhao Zhu, Hao Zhou, Changjiang Gao, Sizhe Liu, Shujian Huang[†]

National Key Laboratory for Novel Software Technology, Nanjing University

{zhuwh,zhouh,gaocj,liusz}@smail.nju.edu.cn, huangs@nju.edu.cn

Abstract

Machine translation aims to automatically translate one natural language into another. This process requires great ability of language understanding and language generation, making machine translation a challenging task. Recent studies have shown that large language models (LLMs) are capable of performing various tasks, including machine translation, based on human instructions. The powerful ability of LLM provides new possibilities for the innovation of natural language processing paradigms. To better accomplish machine translation tasks with the support of LLMs, researchers have conducted extensive research and analysis on the translation and multilingual capabilities of these models. This paper introduces the latest developments in this field from the following three aspects: evaluating translation capabilities of large language models; eliciting translation capabilities of large language models; language ability of large language models in different languages.

Keywords: Machine Translation , Large Language Model , In-Context Learning , Instruction-Tuning , Multilinguality

1 引言

机器翻译（Machine Translation, MT）是利用计算机把一种自然语言自动地转换为另一种自然语言的过程。相较于人工翻译，机器翻译这种快速便捷的翻译方式可以更好地满足人们的基础翻译需求，对促进信息传播和社会经济发展有着重要的实际意义。机器翻译任务的完成依赖于对源语言的准确理解和对目标语言的准确生成，对机器翻译模型的语言能力有着极高的要求 (Nirenburg, 1989; Och and Ney, 2002; Vaswani et al., 2017)。

近年来，在大规模语料上训练的具有大规模参数的大语言模型（Large Language Model, LLM）展现出了极强的语言能力。大语言模型能够理解人类指令，并根据指令完成包括翻译在内的各类任务；还具备情景学习（In-Context Learning, ICL）(Brown et al., 2020),思维链（Chain-of-Thought, CoT）(Wei et al., 2023)等涌现能力,能够利用上下文中的额外信息对自身生成预测结果进行优化调整。大语言模型的强大能力为机器翻译范式革新提供了可能。

目前，针对大语言模型在机器翻译方面的分析和应用已有大量研究。本文整理和综述这些工作，从三个方面介绍大语言模型在机器翻译方面的最新进展：大语言模型翻译能力评估、大语言模型翻译能力激发、大语言模型在不同语言上的能力展现。

通过对这些研究工作进行整理和总结，我们可以得出以下结论：(1) 先进的大语言模型（如ChatGPT）已经可以在部分语言对上超过传统的有监督神经机器翻译模型，但是在低资源语言上仍然存在较大差距；(2) 情景学习与指令微调是两种最常见的翻译能力激发方式，情景学习可以以较小的代价让模型进行机器翻译，而指令微调能够更好地激发模型的翻译能力；(3) 大语言模型在不同语言上的语言能力高度不平衡，但是通过平行数据可以帮助大语言模型建立不同语言之间的对应关系，帮助模型在非英语语言上也展现出不错的语言能力。总体来说，大语言模型的出现为机器翻译研究带来了新的契机和挑战，基于大语言模型建立新的机器翻译范式展现出巨大的潜力，而提升大语言模型翻译能力也可以帮助大语言模型在更多的语言上展现其强大的能力。

本文的后续内容安排如下：第2节将介绍机器翻译和大语言模型的相关背景，第3、4、5节分别介绍大语言模型的翻译能力评估、翻译能力激发和跨语言能力展现相关研究进展，第6节将对整体研究进展和研究趋势进行总结和展望。

2 背景

2.1 机器翻译

从基于规则的机器翻译(Nirenburg, 1989)到统计机器翻译(Och and Ney, 2002)，再到神经机器翻译(Vaswani et al., 2017)，机器翻译范式不断转变，机器翻译效果不断提升。目前，最好的神经机器翻译模型已经可以在少部分高资源语言对(如德语-英语)上超过人类专家翻译水平(Ng et al., 2019)。但是，只构建支持单一翻译方向的机器翻译模型还无法充分满足实际需求。当机器翻译系统需要支持的语言数量增多时，为每一个翻译方向单独部署机器翻译模型代价巨大。于是,构建同时支持多个翻译方向的多语言机器翻译系统逐渐成为近年来的热点研究内容(Johnson et al., 2017; Costa-jussà et al., 2022; Yuan et al., 2022)。此前，多语言机器翻译模型基本均采用编码器-解码器架构。而掌握多种语言的大语言模型为多语言机器翻译系统构建提供了新的可能(Garcia et al., 2023; Zhu et al., 2023)。

2.2 大语言模型

大语言模型的基本架构是Transformer(Vaswani et al., 2017)，基本训练任务是语言建模任务(Bengio et al., 2000)，训练数据基本是以英语为主的多语言单语语料(Zhang et al., 2022; Lin et al., 2022)。其中，语言建模任务要求模型根据前缀序列，准确预测下个词语。在大规模语料上使用语言建模任务进行训练可以使模型掌握语料中蕴含的海量知识，包括事实性知识(Petroni et al., 2019)、语言学知识(Tenney et al., 2019)等，并且具备极强的语言理解和语言生成能力(Pavlick, 2022)。这种强大语言能力也使大语言模型其能够根据人类指令完成各类下游任务。

由于语言模型最初的训练目标仅为预测后续可能的符号，与特定任务并不存在明确的关联。研究人员提出了两种方法教会模型理解某个给定的人类指令，并遵照指令进行对应任务的

执行，这两种方法分别是情景学习和指令微调。情景学习(Brown et al., 2020)利用上下文情景中包含的描述和示例进行学习，仅作用于推断阶段。以翻译任务为例，根据提供任务描述 \mathcal{T} 和示例 $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^k$ ，构造上下文提示 $\mathcal{P} = \mathcal{T}(\mathcal{X}_1, \mathcal{Y}_1) \oplus \mathcal{T}(\mathcal{X}_2, \mathcal{Y}_2) \oplus \dots \oplus \mathcal{T}(\mathcal{X}_k, \mathcal{Y}_k)$ 和待翻译源语言句子 \mathcal{X} ，并将其输入模型。则模型根据示例理解指定的任务，并生成翻译结果 \mathcal{Y} 。翻译结果 \mathcal{Y} 一般通过采样算法获得： $\arg \max_{\mathcal{Y}} p(\mathcal{P} \oplus \mathcal{T}(\mathcal{X}, \mathcal{Y}))$ 。情景学习能够让模型在不更新参数的情况下理解和完成指定任务。Figure 1展示了大语言模型利用情景学习进行机器翻译的例子。

指令微调(Wei et al., 2021; Ouyang et al., 2022)则作用于训练阶段。通过包含特定指令的样本来训练模型，通过调整模型参数，使模型能够更加准确地完成指定任务。相对而言，指令微调方案由于需要改变大语言模型参数，所以对计算资源的要求比较高。

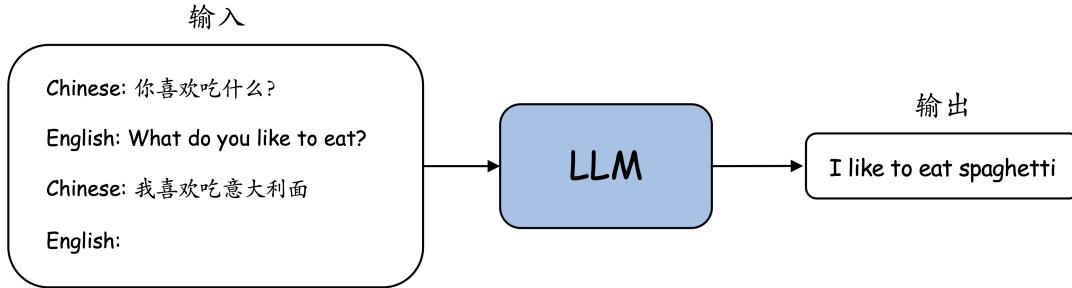


Figure 1: 大语言模型通过情景学习进行机器翻译的示意图

3 大语言模型翻译能力评估

训练大语言模型所使用的海量语料往往以单语数据为主，且其中英文语料占主导地位，而其他语言的语料往往只有很小的比例。大语言模型在这样的数据分布上能否建模好不同语言之间的对应关系，并进一步学习得到可靠的翻译知识，是研究者非常关心的一个问题。因此，研究人员对大语言模型的多语言翻译能力进行了考察和评估(Lin et al., 2022; Moslem et al., 2023; Jiao et al., 2023b; Bang et al., 2023; Hendy et al., 2023; Garcia et al., 2023; Zhu et al., 2023)。

这些研究工作采用情景学习的方式，考察了众多流行的大语言模型在多个翻译方向上的翻译能力。Table 1中列举了这些研究工作的基本情况。其中Zhu et al. (2023)的评测工作是相对最为全面的，他们基于Flores-101多语言机器翻译数据集，在102个语言，202个方向上对XGLM、BLOOMZ、OPT和ChatGPT这四个流行的大语言模型的多语言翻译能力进行了评估，并与现有的强大的监督学习基线模型NLLB-1.3B(Costa-jussà et al., 2022)、M2M-12B(Fan et al., 2021)进行了对比。他们的研究结果表明：在众多被评测的大语言模型中，ChatGPT目前的翻译表现最好。相比于此前的大语言模型，ChatGPT在不同语言间的表现更加平衡，并且在20%左右以英语为核心的翻译方向已经可以超过强大的有监督基线模型NLLB。但与此同时，在大部分翻译方向上，尤其是低资源语言翻译上，ChatGPT仍然落后于有监督模型和商用机器翻译系统（如Figure 2所示）。

评估工作	语言数量	语言对数量	大语言模型
Lin et al. (2022)	13	182	GPT-3, XGLM
Moslem et al. (2023)	6	5	GPT-3, BLOOMZ
Jiao et al. (2023b)	5	8	ChatGPT, GPT4
Bang et al. (2023)	13	24	ChatGPT
Hendy et al. (2023)	18	10	ChatGPT
Zhu et al. (2023)	102	202	XGLM, BLOOMZ, OPT, ChatGPT

Table 1: 翻译能力评估工作概览

值得注意的是 Zhu et al. (2023)发现在使用公开测试集评测大语言模型能力时容易出现数据泄漏问题。由于大语言模型训练数据往往覆盖范围较广且透明度较差，在利用现有数据进行评测时，很容易发生测试数据被包含在训练数据中的情况，导致对应模型的测试表现高

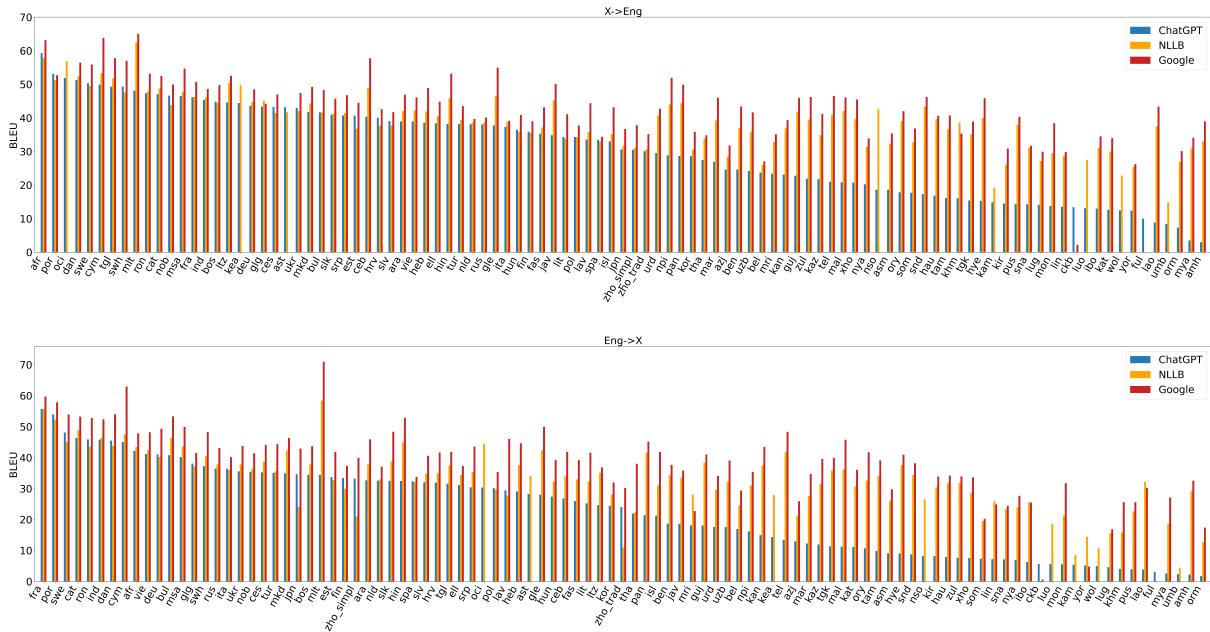


Figure 2: 大语言模型ChatGPT与有监督机器翻译模型NLLB，商用机器翻译系统Google Translate的翻译表现对比(结果摘自(Zhu et al., 2023))

于实际翻译水平。例如，由于BLOOMZ (Scao et al., 2022)利用了Flores-200作为训练数据，在Flores-101数据集上评测BLOOMZ的翻译表现时，就存在数据泄漏的问题，导致评测结果无法准确反映模型的翻译能力 (Zhu et al., 2023)。考虑到不同模型的数据使用各不相同，如何更加公正合理地评估大语言模型的翻译能力，仍然是一个值得关注的问题。

综合而言，大语言模型的翻译能力评估简单有效，展现了大语言模型在翻译上的强大能力，也体现了这种翻译范式的潜在能力。但是，该类研究仅通过情景学习进行翻译，有可能仅发挥了大模型的部分翻译能力。如何进一步激发大语言模型的翻译能力，提升大语言模型的翻译质量，仍然是一个有待解决的开放性问题。

4 大语言模型翻译能力激发方式研究

采取不同的方式激发大语言模型的翻译能力可能会得到不同的翻译表现，研究人员研究了情景学习和指令微调等不同激发方式对翻译表现的影响(Table 2)。

激发方式	影响因素	研究工作
情景学习	模版内容	Zhu et al. (2023)
	模版语言	Zhang et al. (2023a)
	示例来源	Vilar et al. (2022)
	示例挑选	Agrawal et al. (2022), Zhang et al. (2023a), Moslem et al. (2023), Zhu et al. (2023)
	示例个数	Moslem et al. (2023), Agrawal et al. (2022), Zhang et al. (2023a), Zhu et al. (2023)
	示例语言	Zhu et al. (2023)
	示例粒度	Zhu et al. (2023)
指令微调	数据规模	Li et al. (2023), Yang et al. (2023)
	数据质量	Li et al. (2023)
	数据来源	Jiao et al. (2023a), Zhang et al. (2023b)

Table 2: 翻译能力激发研究工作概览

4.1 利用情景学习激发大语言模型翻译能力

情景学习中是用任务描述和示例来描述特定任务，其中，示例往往以某个指定的模板形式给出。在利用情景学习激发大语言模型的翻译能力时，模版的选择、示例的选择等许多因素都

对最终的翻译表现有影响。为了找到更好的情景学习方案，研究人员对这些因素进行了全面的分析研究(Lin et al., 2022; Vilar et al., 2022; Chowdhery et al., 2022; Agrawal et al., 2022; Zhang et al., 2023a; Moslem et al., 2023; Zhu et al., 2023)。

4.1.1 设计合适的情景学习模版

情景学习模版内容是对任务的具体描述，情景学习模板会直接影响翻译能力激发效果。大语言模型在不同模版下的翻译表现有着很大的差距(Zhu et al., 2023; Zhang et al., 2023a)。因此如何为翻译任务设计适合机器翻译任务的模板便成为了一个重要的研究问题。

然而，模板的设计面临着许多困难。首先，模板有效程度不一定符合人类直觉。Zhu et al. (2023)指出，在激发大语言模型XGLM的实验中，“`<X>\n can be summarized as \n <Y>`”这种不合理模版甚至比“`<X>\n can be translated to \n <Y>`”这种合理模版更能激发大语言模型的翻译表现 (Table 3)。模版有效性与人类直觉之间的冲突不仅对模板设计工作提出了巨大的挑战，也促使人们对情景学习的工作原理进行更加深入的思考。

其次，最优模板难以通用，需要按照模型和任务单独定制。现有工作发现，同一模版在不同大语言模型上的使用效果是差别极大的。例如，一种经典的翻译任务模板是“[SRC]: <X>\n [TGT]: <Y>”，其中 “[SRC]”与 “[TGT]”分别为源语言和目标语言的名称，`<X>`与`<Y>`则是源端与目标端句子。这种模版对于PaLM(Vilar et al., 2022)，GLM(Zhang et al., 2023a)模型效果很好，但是对于XGLM模型却效果很差(Zhu et al., 2023)。另外，即使是同样的大模型在不同语言方向上进行翻译时，最优的模板也不同(Lin et al., 2022; Zhang et al., 2023a; Zhu et al., 2023)。这些发现都说明想要为翻译任务设计通用有效的情景学习模版是非常有挑战的。

In-context Template	Deu-Eng	Eng-Deu	Rus-Eng	Eng-Rus	Rus-Deu	Deu-Rus	Average
reasonable instructions:							
<code><X>=<Y></code>	37.37	26.49	29.66	22.25	17.66	17.31	25.12
<code><X> \n Translate from [SRC] to [TGT]: \n <Y></code>	37.95	26.29	29.83	20.61	17.56	15.93	24.70
<code><X> \n Translate to [TGT]: \n <Y></code>	37.69	25.84	29.96	19.61	17.44	16.48	24.50
<code><X> \n [TGT]: <Y></code>	29.94	17.99	25.22	16.29	12.28	11.71	18.91
<code><X> is equivalent to <Y></code>	23.00	4.21	17.76	9.44	8.14	9.84	12.07
<code><X> \n can be translated to \n <Y></code>	37.55	26.49	29.82	22.14	17.48	16.40	24.98
<code>[SRC]: <X> \n [TGT]: <Y></code>	16.95	8.90	14.48	6.88	7.86	4.01	9.85
unreasonable instructions:							
<code><X>\$<Y></code>	37.77	26.43	29.53	20.99	17.72	17.27	24.95
<code><X> \n Translate from [TGT] to [SRC]: \n <Y></code>	38.18	26.21	29.85	20.35	17.75	16.63	24.83
<code><X> \n Compile to [TGT]: \n <Y></code>	37.39	26.35	29.68	19.91	17.52	16.15	24.50
<code><X> \n [SRC]: <Y></code>	27.86	16.69	24.41	18.16	11.98	12.60	18.62
<code><X> is not equivalent to <Y></code>	23.50	3.92	16.90	7.80	8.06	9.23	11.57
<code><X> \n can be summarized as \n <Y></code>	37.46	26.24	29.42	22.62	17.68	17.15	25.10
<code>[SRC]: <X> \n [SRC]: <Y></code>	19.03	8.21	15.96	6.37	7.57	4.40	10.26

Table 3: 不同情景学习模版对翻译表现的影响(该结果摘自(Zhu et al., 2023))

4.1.2 选择合适的情景学习示例

情景学习效果的另一个重要影响因素是情景学习示例。如何为翻译任务提供合适的情景学习示例同样是研究者们关注的问题。情景学习示例一般从有监督数据如训练集、验证集中挑选而来，Vilar et al. (2022)发现从高质量的验证集中挑选情景学习示例比从训练集中挑选效果更好。而为了从候选集合中挑选出最有效的示例，研究人员也尝试了包括稀疏检索、稠密检索、混合检索等多种示例挑选方案(Agrawal et al., 2022; Zhang et al., 2023a; Moslem et al., 2023)，但是相比于随机检索取得收益都比较有限。Zhu et al. (2023)进一步发现，即使根据给定源句的参考译文进行检索，也很难带来进一步的增益。

增加情景学习示例个数是一种简单有效提升翻译表现的途径(Moslem et al., 2023; Agrawal et al., 2022; Zhang et al., 2023a; Zhu et al., 2023)。但是Zhang et al. (2023a)和Zhu et al. (2023)都发现随着示例个数的增加，大语言模型的翻译性能提升幅度会不断放缓。当示例个数在10个以上时，再增加示例个数则很难带来进一步的增益。

此外，情景学习示例中的具体内容会对翻译表现有很大的影响。Zhu et al. (2023)发现使用与测试样例翻译方向不同的跨语言翻译数据作为示例时，能够在某些语言对（如中文-英文）上带来大幅的翻译性能提升，这是一种非常有趣的现象。而如果使用不匹配的源端与目标端句子作为样例，则大语言模型将无法进行翻译任务，这说明大语言模型从示例中了解到需要保持源

句与目标句之间的语义一致性。如果使用词级别与文档级别的翻译作为样例，则会使大语言模型进行句子级别翻译的性能下降，这说明大语言模型需要根据示例中确定翻译任务的粒度。如果使用重复的句子作为样例时，翻译性能同样会下降，这说明保持情景学习示例的多样性是很重要的。

4.2 利用指令微调激发大语言模型翻译能力

另一种激发大语言模型翻译能力的方式是指令微调，通过让大语言模型学习包含指令的有监督数据，可以促使模型更加准确地遵循指令，完成下游任务(Wei et al., 2022; Muennighoff et al., 2022; Chung et al., 2022)。近期研究者开始尝试对大语言模型进行翻译指令微调，针对性激发大语言模型的翻译能力(Li et al., 2023; Jiao et al., 2023a; Yang et al., 2023; Zhang et al., 2023b)。

已有研究发现在特定翻译方向上，仅通过小规模翻译数据（千条至万条数据）微调大语言模型就可以大幅提升大语言模型的翻译能力(Jiao et al., 2023a; Li et al., 2023; Yang et al., 2023; Zhang et al., 2023b)。具体来说，Jiao et al. (2023a)使用翻译数据和通用任务数据对LLaMA、BLOOMZ等大语言模型进行指令微调，发现模型不仅可以完成简单的翻译任务还可以根据人类的特殊需求调整翻译内容。Li et al. (2023)专注于使用单纯的翻译数据微调XGLM模型，其发现随着数据规模增大，以及数据质量提高，模型的翻译性能可以不断提高，这显示了大模型的翻译能力仍存在巨大的提升空间。Yang et al. (2023)使用了102种语言的平行数据对LLaMA模型进行了指令微调，发现利用多语言翻译数据可以同时提升模型在多种语言上的翻译能力，尤其是增强了模型在维语、藏语、蒙古语等低资源语言上的翻译水平。Zhang et al. (2023b)使用了多轮交互式机器翻译数据进行指令微调，发现可以同时提升模型的翻译能力和语言理解能力，让模型能够更好地完成词约束翻译等有特殊需求的翻译任务。

指令微调的研究表明，大模型的潜在翻译能力比使用情景学习展现出来的要高得多。相比于情景学习的激发方式，使用指令微调激发大语言模型的翻译能力存在以下四点优势(Li et al., 2023):(1)激发效果更好，可以取得更强的翻译表现，尤其是在中低资源语言上的翻译效果更好（如Figure 3所示）；(2)泛化性能更好，在未见语言对上，指令微调的翻译表现比情景学习更好；(3)对于指令理解程度更好，在不同的翻译相关指令下，模型的翻译性能稳定，不会出现情景学习中对翻译相关指令非常敏感的情况；(4)推断时不再依赖翻译任务示例，这可以大大减少上下文长度，减少解码计算开销。而相比于情景学习，指令微调的主要劣势在于需要训练大规模参数，对计算资源要求更高。总体来说，指令微调是一种激发大语言模型翻译能力的有效方案，并且随着LoRA(Hu et al., 2023)、QLoRA(Dettmers et al., 2023)等高效微调方案的出现，指令微调的计算代价不断降低，这种方案是非常值得研究人员进一步研究和关注的。

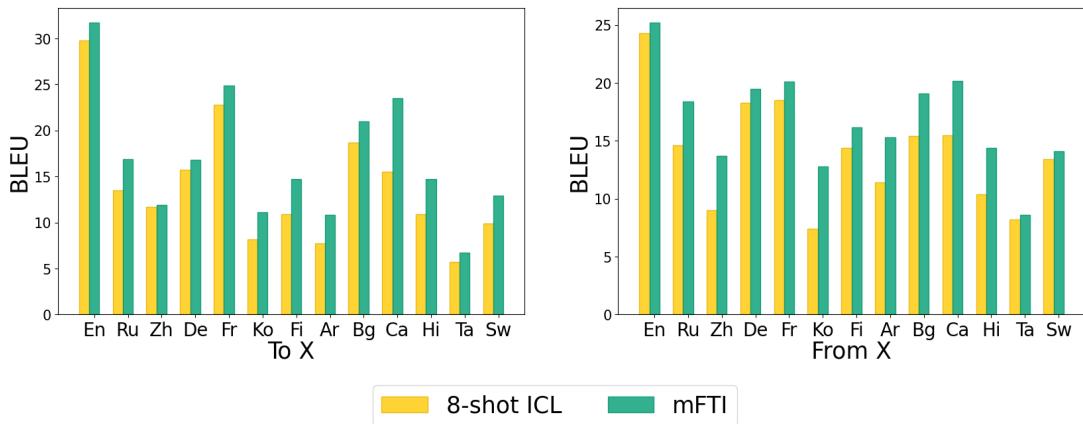


Figure 3: 使用情景学习和指令微调时大语言模型翻译效果对比(结果摘自(Li et al., 2023))

5 大语言模型在不同语言上的能力展现

大语言模型的预训练语料以英文为主，且指令微调时使用的通用任务数据，如alpaca数据(Taori et al., 2023)，也以英文为主。这一方面让模型有着极强的英语语言能力，另一方面

也导致其非英语语言能力较弱。如果大模型能够学会翻译，是否其语言能力也可以在不同语言之间进行迁移呢？对于这一问题，研究者展开了初步研究(Cui et al., 2023; Yang et al., 2023; Zhang et al., 2023b)。

Cui et al. (2023)和Yang et al. (2023)关注于在预训练阶段提升大语言模型在中文上的能力，他们提出可以通过在词表中添加中文字符以及利用中文单语数据进行继续预训练的方式提升模型的中文能力。

Zhang et al. (2023b)则关注于在指令微调阶段提升大语言模型在中文、德语等语言上的能力。他们发现在指令微调阶段，增强大语言模型在英语与非英语之间的翻译能力，是帮助提升模型非英语语言能力的一种有效手段。该方法还可以避免继续预训练大语言模型和收集大规模数据带来的巨大开销。在训练数据设计上，Zhang et al. (2023b)提出使用多轮交互式翻译数据来进行指令微调。从其实验结果来看(Table 4)，相比于没有进行语言能力迁移的Alpaca模型(Taori et al., 2023)、Vicuna模型(Chiang et al., 2023)，使用多轮交互式翻译数据微调得到的Bayling模型(Zhang et al., 2023b)在中文能力评估测试集GaoKao上取得了一定的性能提升。这一结果表明增强大语言模型的翻译能力是帮助模型展现多语言能力的有效手段。

Systems	Avg.	GaoKao(%)									
		chinese	english	mathqa	physics	chemistry	biology	history	geography	mathcloze	
GPT-3.5-turbo	43.87	42.68	86.27	30.48	21.00	44.44	46.19	59.57	63.32	0.85	
BayLing-13B	32.13	29.27	69.28	29.34	21.50	36.71	30.00	34.04	38.19	0.85	
BayLing-7B	28.20	27.64	55.56	26.78	24.50	29.95	29.05	33.19	27.14	0.00	
ChatGLM-6B	31.83	31.71	52.29	26.50	16.00	27.54	28.10	54.04	47.74	2.54	
Vicuna-13B	29.36	21.14	71.24	21.94	23.00	31.88	27.14	33.19	34.67	0.00	
Alpaca-7B	20.03	24.80	36.27	17.95	6.00	20.77	20.95	24.68	27.14	1.69	

Table 4: 不同大语言模型在中文能力评估数据集GaoKao上的表现对比(结果摘自(Zhang et al., 2023b))

6 总结

近期，大语言模型迅猛发展，并凭借其惊人的语言能力在各项自然语言处理任务上都展现了巨大的潜力。本文聚焦于机器翻译任务，对大语言模型在机器翻译方面的相关进展进行了综述，具体介绍了以下三个方面的内容，包括：1) 大语言模型翻译能力评估；2) 大语言模型机器翻译能力激发；3) 大语言模型在不同语言上的能力展现。总体来说，大语言模型已经展现出较强的多语言机器翻译能力，且仍有进一步提升的空间；但其在不同语言上的能力非常不平衡，在大部分中低资源语言上仍然与有监督基线模型有着较大的差距。在未来，如何更好地激发大语言模型的翻译能力，尤其是低资源语言上的翻译能力仍然有待解决。此外，为了让大语言模型在更多语言上发挥其强大的语言能力，多语言翻译的研究和探索可能有着重要的价值。

致谢

本文研究受到国家自然科学基金(No. 62176120)、辽宁省自然科学基金(No. 2022-KF-26-02)等项目资助。黄书剑是本文通讯作者。

参考文献

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research (JMLR)*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2023. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Chen, and Jiajun Chen. 2023. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *arXiv preprint arXiv:2305.15083*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR’s WMT19 news translation task submission. In *Proceedings of the Conference on Machine Translation (WMT)*.

Sergei Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8(1).

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Fei Yuan, Yinquan Lu, WenHao Zhu, Lingpeng Kong, Lei Li, and Jingjing Xu. 2022. Lego-mt: Towards detachable models in massively multilingual machine translation. *arXiv preprint arXiv:2212.10551*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.