Teaching Large Language Models to Translate on Low-resource Languages with Textbook Prompting

Ping Guo[1,2,‡], Yubing Ren[1,2,‡], Yue Hu[1,2*], Yunpeng Li[1,2], Jiarui Zhang[1,2], Xingsheng Zhang[1,2], Heyan Huang[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Beijing Institute of Technology, Beijing, China
guoping@iie.ac.cn

Abstract

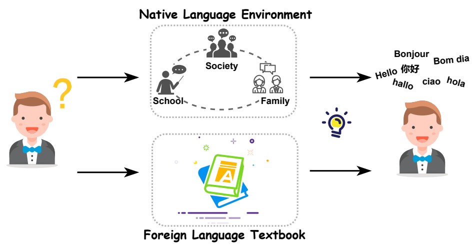
Large Language Models (LLMs) have achieved impressive results in Machine Translation by simply following instructions, even without training on parallel data. However, LLMs still face challenges on low-resource languages due to the lack of pre-training data. In real-world situations, humans can become proficient in their native languages through abundant and meaningful social interactions and can also learn foreign languages effectively using well-organized textbooks. Drawing inspiration from human learning patterns, we introduce the Translate After LEarNing Textbook (TALENT) approach, which aims to enhance LLMs' ability to translate low-resource languages by learning from a textbook. TALENT follows a step-by-step process: (1) Creating a Textbook for low-resource languages. (2) Guiding LLMs to absorb the Textbook's content for Syntax Patterns. (3) Enhancing translation by utilizing the Textbook and Syntax Patterns. We thoroughly assess TALENT's performance using 112 low-resource languages from FLORES-200 with two LLMs: ChatGPT and BLOOMZ. Evaluation across three different metrics reveals that TALENT consistently enhances translation performance by 14.8% compared to zero-shot baselines. Further analysis demonstrates that TALENT not only improves LLMs' comprehension of low-resource languages but also equips them with the knowledge needed to generate accurate and fluent sentences in these languages.

Keywords: Large Language Models, Multilingual Machine Translation, Low-resource Language Evaluation.

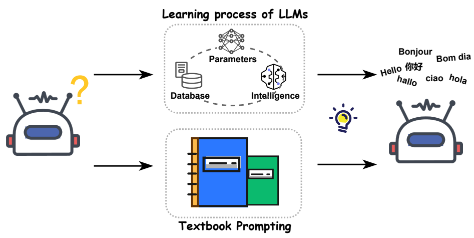
1. Introduction

Large Language Models (LLMs), typically characterized by the large scale of parameters and training corpora, have achieved dominant performance on a wide range of natural language understanding and generation tasks. Through instruction prompting, where certain words or sentences are provided as prompts alongside the base input, LLMs can simulate human-like intelligence to some extent by processing and generating coherent and contextually relevant responses that align with human intentions. Besides their impressive potential across various tasks, LLMs have particularly excelled in the field of Machine Translation (MT), showcasing surprising performance on high-resource languages (Hendy et al., 2023). However, the translation ability of LLMs on low-resource languages is questionable and various works have proven their weak performance at generalizing to low-resource languages (Hangya et al., 2022; Lai et al., 2023; Bang et al., 2023). One main reason is the shortage of pre-training data on such languages. Since the pre-training data for high-resource languages is often orders of magnitude larger than the low-resource ones, it is hard for LLMs to focus on learning low-resource language-specific knowledge.

Unlike the learning process of LLMs, humans possess the remarkable ability to master languages through a diverse range of methods, as shown in Figure 1a. In the context of acquiring our native language, we absorb it from our linguistic surroundings, which immerse us in abundant and meaningful communication in our daily interactions. As for



(a) Human language learning methods for native language (above) and foreign languages (below).



(b) Diverse language learning approach for LLMs: Pre-training (above) and ours (below).

Figure 1: Diverse methods for mastering languages for humans (a) and LLMs (b).

‡ Equal Contribution.
* Corresponding Author. Email: huyue@iie.ac.cn

LREC-COLING 2024, pages 15685–15697
20-25 May, 2024.

foreign languages, humans can achieve proficiency without the requirement of overwhelming data. Human language acquisition for foreign tongues often involves engaging with textbooks, which typically offer a well-organized and systematic approach to the new language. These instructional materials serve as valuable guides, assisting individuals on their journey to fluency, eliminating the need to awkwardly deduce language-specific nuances from extensive language examples. This leads to a question that hasn't been thoroughly explored: Can LLMs effectively translate a low-resource language by adopting a textbook-based learning approach, as depicted in Figure 1b?

In pursuit of this goal, we introduce the Translate After LEarNing Textbook (TALENT) method, aimed at employing LLMs to translate low-resource languages using a textbook-based approach. The TALENT methodology comprises three key stages: **(1) Creating a Textbook for Low-resource Languages.** To mitigate the reliance on extensive data, we develop a tailored Textbook for each low-resource language based on the source sentence. This Textbook consists of two sections: Language Examples, containing valuable examples to grasp language-specific usage patterns, and a Vocabulary List, providing definitions for less familiar words. **(2) Guiding LLMs to Absorb Textbook Content for Syntax Patterns.** Given the intricate nature of syntax and its complex variations across languages, we incorporate an intermediate Absorption Stage: Instruct LLMs to absorb the language knowledge from the Textbook and parse Language Examples to obtain Syntax Patterns. **(3) Enhancing Translation by Utilizing the Textbook and Syntax Patterns.** To seamlessly integrate the Textbook and Syntax Patterns into LLMs, we restructure them into natural language form and embed them within the context of the standard translation prompt. This empowers LLMs to effectively apply this knowledge, resulting in more precise and accurate translations.

To comprehensively assess how effectively TALENT can perform translations in low-resource languages, we selected 112 diverse low-resource languages from the FLORES-200 dataset. Our evaluation focuses on the translation performance between English and these 112 languages. We present our findings based on two LLMs: BLOOMZ-7.1B (Muennighoff et al., 2023) and ChatGPT [1]. We gauge performance using three distinct metrics: COMET (Rei et al., 2022), BLEURT (Sellam et al., 2020), and ChrF++ (Popović, 2017). The outcomes reveal that TALENT yields a 9.2% enhancement on ChatGPT, which further rises to 14.8% when applied to BLOOMZ-7.1B.

We delve deeper into how each stage in TALENT impacts the translation ability of LLMs on low-resource languages. Through our experiments on the Textbook, we observe that the Vocabulary List yields substantial enhancement in low-resource language comprehension. This is attributed to the robust lexical alignment signal it provides between these languages and English. Conversely, the Language Examples bring more obvious improvements when generating low-resource languages. We reveal that Language Examples can get LLMs to recognize unfamiliar language tags and subsequently generate tokens in the correct language. Furthermore, the introduction of a dedicated Absorption Stage enables LLMs to effectively analyze and parse low-resource languages. The acquired syntax insights for low-resource languages can significantly enhance both comprehension and generation abilities. In summary, this work offers the following contributions:

- Inspired by human language learning, we propose the Translate After LEarNing Textbook (TALENT) method, which guides LLMs to absorb a low-resource Textbook for Syntax Patterns and then translate source sentences to improve translation on low-resource languages.
- We comprehensively evaluate TALENT using 112 low-resource languages on both BLOOMZ-7.1B and ChatGPT with three distinct metrics. TALENT consistently delivers improvements across all LLMs and metrics.
- We analyze how the Absorption Stage in TALENT can influence translation performance. The Absorption Stage can benefit the translation by 3.3 COMET score on average and even improves the translation quality on Cyrillic script languages by 13.8 COMET score.
- Examination of the Textbook reveals that with the lexical alignment cues in Vocabulary List, LLMs can better understand low-resource languages, while Language Examples can get LLMs to better recognize language tags and generate tokens in the appropriate languages.

2. Related Work

2.1. Retrieval-Augmented LLMs

Due to inherent limitations in accessing specialized knowledge by LLMs (Ram et al., 2023), promising approaches (Guu et al., 2020; Borgeaud et al., 2022; Jiang et al., 2023; Luo et al., 2023a,b; Shi et al., 2023) involve retrieving relevant information from an external database using similarity-based retrievers. Retrieval-augmented methods have been successfully applied to empower LLMs with domain-specific specialized knowledge for various tasks (Luo et al., 2023c), such as code completion (Zhang

[1] https://chat.openai.com/

et al., 2023), information retrieval (Wang et al., 2023a; Ai et al., 2023), image captioning (Ramos et al., 2023), and biomedical applications (Soong et al., 2023), among others. Our focus is on retrieving language-specific knowledge in the form of low-resource language textbooks to enhance LLMs for low-resource language translation tasks.

2.2. Guiding LLMs for Neural Machine Translation

Numerous studies have focused on evaluating the translation capabilities of LLMs (Zhu et al., 2023a; Wang et al., 2023b; Kocmi and Federmann, 2023; Raunak et al., 2023; Karpinska and Iyyer, 2023; Lu et al., 2023b; Kadaoui et al., 2023; Etxaniz et al., 2023; Yamada, 2023) or enhancing their translation proficiency (Li et al., 2023; Mu et al., 2023; Zeng et al., 2023; AI4Bharat et al., 2023; Chen et al., 2023; Hao et al., 2023; Xu et al., 2023; Ebadulla et al., 2023; Schioppa et al., 2023; Jiao et al., 2023). Some (Puduppully et al., 2023; Gao et al., 2023; Peng et al., 2023; Moslem et al., 2023; Nagy et al., 2023; Jon et al., 2023; Fernandes et al., 2023; Sia and Duh, 2023) have employed straightforward prompts to explore LLMs' translation capabilities, while others (Agrawal et al., 2022; Vilar et al., 2023; Jones et al., 2023; Nguyen et al., 2023; Bhandari and Chen, 2023) have investigated prompts' impact on formality or specific dialects. In addition to traditional prompt methods, certain studies (Cahyawijaya et al., 2023; Tanwar et al., 2023; Kim et al., 2023; Gitau and Marivate, 2023; Yang and Nicolai, 2023; Liu and Hou, 2023) have sought effective in-context examples to enhance translation outcomes. Others (Huang et al., 2023; Nicholas and Bhatia, 2023; Zhu et al., 2023b; Kumar et al., 2023; Araabi et al., 2023; Oh et al., 2023) have explored techniques like Chain-of-Thought to structure translation processes for LLMs. Recently, the integration of dictionaries into LLMs (Lu et al., 2023a) has substantially enhanced the translation ability in LLMs.

3. Translate After Learning Textbook

We present the Translate After LEarNing Textbook (TALENT) framework, which initially constructs a Textbook for the low-resource languages. Following that, a dedicated Absorption Stage for LLMs is employed to extract Syntax Patterns. Finally, TALENT integrates all acquired knowledge as prompts' context for generating translations. The overall TALENT framework is illustrated in Figure 2. Formally, for a translation task from sentence $\mathbf{x}$ in language $l_{\mathbf{x}}$ to sentence $\mathbf{y}$ in low-resource language $l_{\mathbf{y}}$ [2], TALENT operates as follows:

3.1. Creating a Low-resource Language Textbook

In order to enhance LLMs with linguistic ability in the low-resource language $l_{\mathbf{y}}$, we generate a low-resource language Textbook $T_{l_{\mathbf{x}} \to l_{\mathbf{y}}}(\mathbf{x})$. This Textbook encompasses two crucial aspects of low-resource language knowledge, often advantageous in human language acquisition:

Vocabulary List A Vocabulary List is a common feature found in almost all textbooks. It offers specific word-level translations or equivalents between languages. What's more, Vocabulary List is easy to obtain through dictionaries, making them appealing candidates for external resources of translation. In TALENT, we focus on building a Vocabulary List for keywords $\mathbf{x}$ and we employ a statistic method called TF-IDF to select these keywords. Formally, we represent the monolingual corpus for source language $l_{\mathbf{x}}$ as $D_{\mathbf{x}}$, the TF-IDF score for each word $w_{\mathbf{x}}^{(i)}$ in source sentence $\mathbf{x}$ is computed as:

$$f(w_{\mathbf{x}}^{(i)}) = \frac{\sum_{w \in \mathbf{x}} \mathbb{1}(w = w_{\mathbf{x}}^{(i)})}{|\mathbf{x}|} \log \frac{|D_{\mathbf{x}}|}{1 + \sum_{s \in D_{\mathbf{x}}} \mathbb{1}(w_{\mathbf{x}}^{(i)} \in s)}, \tag{1}$$

where $\mathbb{1}(\cdot)$ returns 1 if the statement is true, and 0 otherwise. Also, $|\mathbf{x}|$ represents the length of $\mathbf{x}$. We choose the highest N percentage of words in TF-IDF score as keywords and translate them into the low-resource language $l_{\mathbf{y}}$ using a Dictionary $D_{l_{\mathbf{x}} \to l_{\mathbf{y}}}$. We utilize BabelNet (Navigli and Ponzetto, 2010) as Dictionary $D_{l_{\mathbf{x}} \to l_{\mathbf{y}}}$. Formally, the Vocabulary List $V_{l_{\mathbf{x}} \to l_{\mathbf{y}}}$ is outlined as follows:

$$\begin{aligned} V_{l_{\mathbf{x}} \to l_{\mathbf{y}}}(\mathbf{x}) &= \left\{ \left(w_{\mathbf{x}}^{(i)}, w_{\mathbf{y}}^{(i)}\right) \middle| w_{\mathbf{x}}^{(i)} \in \mathbf{x}, \text{Top-}N(f(w_{\mathbf{x}}^{(i)})) \right\}, \\ w_{\mathbf{y}}^{(i)} &= D_{l_{\mathbf{x}} \to l_{\mathbf{y}}}(w_{\mathbf{x}}^{(i)}). \end{aligned} \tag{2}$$

Language Examples Language Examples are a common feature in many language textbooks. They offer learners insights into the practical usage of words or phrases within specific contexts. In the TALENT framework, we retrieve potentially beneficial sentences $s_{\mathbf{y}}$ from a monolingual corpus $D_{\mathbf{y}}$ in the low-resource language $l_{\mathbf{y}}$. This process is modeled as a selection from the distribution $p(s_{\mathbf{y}}|\mathbf{x})$ using a neural language-agnostic retriever:

$$\begin{aligned} p(s_{\mathbf{y}}|\mathbf{x}) &= \frac{\exp g(\mathbf{x}, s_{\mathbf{y}})}{\sum_{s \in D_{\mathbf{y}}} \exp g(\mathbf{x}, s)}, \\ g(\mathbf{x}, s_{\mathbf{y}}) &= \text{EMBED}(\mathbf{x})^{\top} \text{EMBED}(s_{\mathbf{y}}), \end{aligned} \tag{3}$$

where EMBED refers to an embedding function, implemented using LaBSE (Feng et al., 2022a).

[2] Note that we only introduce the cases of translating from other languages to low-resource ones; The same applies to translating from low-resource languages to the others.

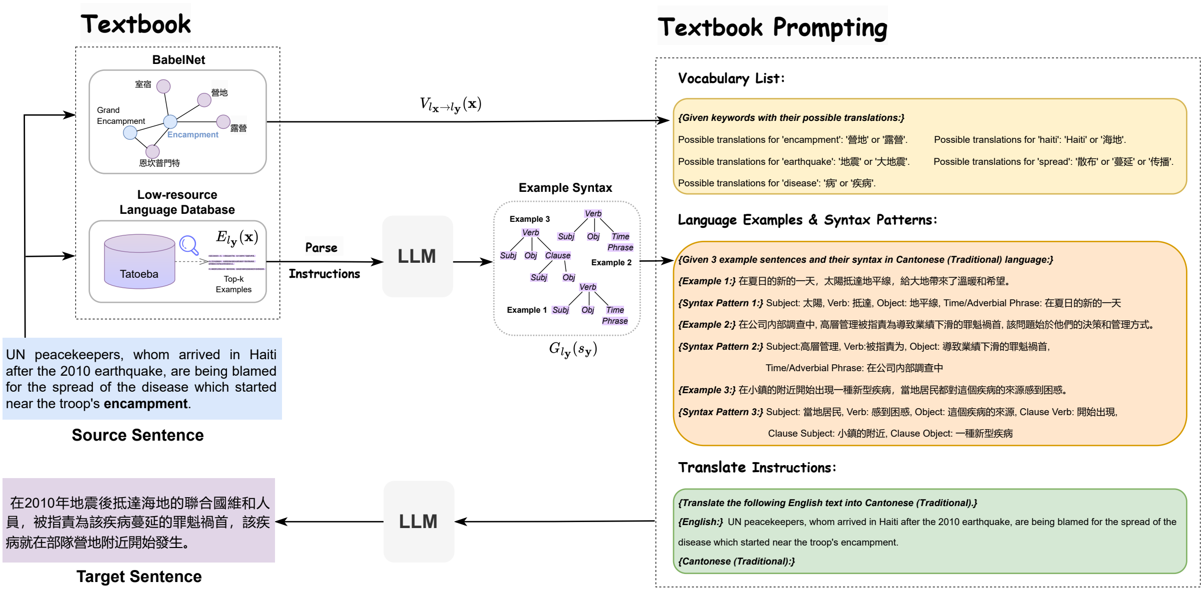


Figure 2: An illustration of TALENT. TALENT first creates a Textbook for low-resource languages. LLMs are then asked to extract Syntax Patterns before finally translating the source sentence.

Within the TALENT framework, we employ Tatoeba [3] as our monolingual database $D_\mathbf{y}$. By averaging the representations from the last hidden layer, we derive sentence representations and subsequently draw K sentences from the monolingual database $D_\mathbf{y}$ with the top probability score $p(s_\mathbf{y}|\mathbf{x})$. These collected sentences form the Language Examples denoted as $E_{l_\mathbf{y}}(\mathbf{x})$.

3.2. Guiding LLMs to Absorb Textbook Content for Syntax Patterns

Human translators often systematically learn the Syntax Patterns of a foreign language before undertaking translation. Drawing inspiration from this human translation process, we introduce a preparatory stage: familiarizing LLMs with the Textbook's content to extract Syntax Patterns. Illustrated in Figure 2, we instruct LLMs to parse the Language Examples within the Textbook, enabling them to capture Syntax Patterns through the following steps:

$$G_{l_\mathbf{y}}(s_\mathbf{y}) = \textsc{Llm}(T_{l_\mathbf{x} \to l_\mathbf{y}}(\mathbf{x}), \textsc{Parse}[s_\mathbf{y}]), \tag{4}$$

where PARSE$[\cdot]$ represents the Parse Instructions: `Given an English-Cantonese (Traditional) Textbook, parse the 3 Language Examples into Syntax Patterns.` $G_{l_\mathbf{y}}(s_\mathbf{y})$ pertains to the Syntax Patterns extracted by LLMs for each Language Example within the Textbook. An illustration of $G_{l_\mathbf{y}}(s_\mathbf{y})$ can be found in Figure 2 for reference.

3.3. Enhancing Translation by Utilizing the Textbook and Syntax Patterns

Once we have obtained the low-resource language Textbook and derived Syntax Patterns, we integrate them into the prompts' context within LLMs, as illustrated in Figure 2. This incorporation allows LLMs to effectively leverage the additional retrieved information alongside the Syntax Patterns they have generated, facilitating the generation of translation outputs. TALENT provides supplementary evidence for the model to effectively incorporate and pinpoint the intended knowledge tailored to the specific translation task. Formally, the translation outcome under the guidance of TALENT is as follows:

$$\begin{aligned} \mathbf{y} = \textsc{Llm}(&\textsc{Textbook}[T_{l_\mathbf{x} \to l_\mathbf{y}}(\mathbf{x}), G_{l_\mathbf{y}}(s_\mathbf{y})], \\ &\textsc{Translate}[\mathbf{x}]), \end{aligned} \tag{5}$$

where TEXTBOOK$[\cdot]$ means our Textbook Prompting and TRANSLATE$[\cdot]$ refers to Translate Instructions.

4. Experiments

4.1. Experiment Settings

Models. We evaluate TALENT using two prominent LLMs: ChatGPT (GPT-3.5-TURBO) and BLOOMZ (7.1B).

- **GPT-3.5-TURBO.** GPT-3.5-TURBO is among the most renowned and powerful LLMs, which is proprietary and utilizes Reinforcement Learning with Human Feedback in conjunction with instruction fine-tuning. Building upon previous studies, we access GPT-3.5-TURBO-0301 through its official Python API.

[3] https://tatoeba.org

Language Family	Direction	Eng-Low						Low-Eng					
	Model	BLOOMZ			ChatGPT			BLOOMZ			ChatGPT		
	Metric.	COMET	BLEURT	ChrF++	COMET	BLEURT	ChrF++	COMET	BLEURT	ChrF++	COMET	BLEURT	ChrF++
Afro-Asiatic (14)	Zero-Shot	58.38	47.71	21.62	71.24	56.73	32.27	66.27	58.10	44.48	73.86	60.08	46.12
	+TALENT	**60.68**	**49.19**	**23.57**	**73.10**	**58.90**	**32.41**	**68.22**	**59.74**	**48.20**	**74.96**	**61.40**	**48.02**
	Few-Shot	65.28	47.34	25.22	73.50	58.32	34.30	69.68	53.25	46.55	75.38	61.76	47.48
	+TALENT	**67.94**	**50.23**	**27.64**	**73.80**	**58.68**	**34.48**	**72.68**	**55.83**	**47.92**	**77.04**	**64.33**	**49.96**
Indo-European (45)	Zero-Shot	48.49	**31.84**	22.20	67.34	48.38	38.03	62.67	51.33	39.98	78.56	**66.83**	53.05
	+TALENT	**50.92**	31.80	**24.62**	**69.92**	**49.88**	**39.79**	**64.24**	**51.62**	**42.70**	**80.26**	66.64	**54.24**
	Few-Shot	50.96	26.87	**25.64**	68.52	48.98	39.59	64.98	48.90	40.05	80.36	69.06	54.83
	+TALENT	**51.35**	**30.82**	25.11	**73.00**	**50.17**	**40.53**	**65.48**	**50.21**	**42.89**	**81.46**	**70.97**	**56.70**
Turkic (6)	Zero-Shot	31.02	12.90	4.47	64.81	36.44	27.03	43.89	29.52	15.76	72.57	56.36	39.34
	+TALENT	**50.75**	**14.23**	**12.27**	**68.35**	**36.75**	**27.41**	**60.47**	**38.85**	**24.10**	**75.41**	**57.62**	**41.53**
	Few-Shot	36.97	9.52	9.90	66.02	36.34	28.20	49.45	23.19	19.49	74.91	59.47	40.92
	+TALENT	**41.04**	**10.35**	**11.69**	**69.73**	**38.42**	**29.53**	**54.54**	**36.81**	**28.83**	**76.72**	**62.64**	**43.18**
Sino-Tibetan (3)	Zero-Shot	42.67	39.11	9.16	74.65	45.41	18.60	58.78	**43.32**	27.43	66.48	45.15	33.44
	+TALENT	**66.92**	**40.41**	**12.20**	**75.23**	**47.62**	**21.29**	**60.47**	42.85	**34.10**	**69.74**	**50.27**	**38.11**
	Few-Shot	78.96	38.64	17.82	75.46	47.86	20.88	62.10	38.80	34.31	67.91	44.76	34.55
	+TALENT	**80.48**	**41.73**	**19.68**	**76.16**	**49.56**	**22.39**	**62.99**	**39.00**	**38.12**	**69.55**	**48.24**	**36.48**
Atlantic-Congo (16)	Zero-Shot	46.98	32.68	**19.73**	52.10	30.54	12.84	**61.54**	**55.44**	**39.80**	57.51	44.50	31.50
	+TALENT	**48.48**	**35.40**	19.02	**52.79**	**32.92**	**14.08**	57.75	48.20	33.63	**59.74**	**46.30**	**32.75**
	Few-Shot	50.46	26.99	19.75	53.78	30.66	12.59	58.21	44.68	33.03	61.30	48.79	33.68
	+TALENT	**51.23**	**29.79**	**23.24**	**53.93**	**32.94**	**14.86**	**58.62**	**46.11**	**33.39**	**62.33**	**50.54**	**34.68**
Dravidian (4)	Zero-Shot	**91.75**	**86.50**	82.69	68.82	57.49	32.72	**90.10**	**80.86**	**71.89**	81.26	63.52	46.46
	+TALENT	90.18	84.97	**84.98**	**69.18**	**58.95**	**33.69**	87.77	74.08	67.54	**81.91**	**64.76**	**49.15**
	Few-Shot	**90.13**	**84.27**	**76.96**	69.57	58.39	32.95	**88.28**	**77.93**	**66.44**	81.85	64.42	46.91
	+TALENT	86.27	79.62	76.45	**69.96**	**59.31**	**34.57**	88.07	77.31	65.09	**84.35**	**68.15**	**51.22**
Austroasiatic (2)	Zero-Shot	29.82	29.80	1.88	57.59	35.88	11.84	40.75	31.85	10.46	61.54	42.55	29.52
	+TALENT	**47.63**	**31.40**	**6.46**	**61.81**	**36.31**	**15.32**	**57.48**	**36.23**	**20.55**	**64.54**	**45.61**	**33.20**
	Few-Shot	64.24	31.25	13.51	61.28	35.32	13.85	46.40	21.97	17.05	62.69	41.05	29.96
	+TALENT	**65.09**	**32.24**	**14.08**	**64.80**	**36.40**	**17.67**	**53.88**	**36.09**	**18.61**	**65.09**	**44.15**	**32.38**
Austronesian (13)	Zero-Shot	48.56	43.25	20.16	69.29	51.78	38.98	52.58	**45.43**	29.01	73.77	64.61	49.55
	+TALENT	**50.99**	**45.05**	**22.22**	**70.51**	**53.63**	**42.93**	**56.60**	45.23	**30.39**	**76.20**	**68.64**	**52.77**
	Few-Shot	52.12	36.45	22.75	69.71	52.65	40.53	58.21	40.82	30.70	75.85	67.07	51.01
	+TALENT	**52.43**	**38.50**	**25.57**	**74.37**	**56.34**	**43.47**	**60.08**	**42.69**	**36.00**	**79.41**	**69.80**	**55.05**
Others (9)	Zero-Shot	33.02	32.18	8.35	55.57	48.43	20.50	46.81	37.18	21.66	63.21	47.10	34.76
	+TALENT	**48.49**	**35.49**	**11.08**	**56.57**	**50.03**	**22.72**	**57.02**	**38.02**	**23.17**	**64.88**	**49.74**	**37.62**
	Few-Shot	46.57	28.03	15.23	54.02	48.58	19.01	52.04	29.72	23.61	66.79	50.23	36.45
	+TALENT	**48.40**	**30.95**	**16.13**	**56.68**	**51.08**	**21.71**	**58.00**	**34.48**	**25.98**	**68.53**	**53.17**	**38.45**
Average	Zero-Shot	47.85	39.55	21.14	64.60	45.68	25.87	58.15	48.11	33.39	69.86	54.52	40.41
	+TALENT	**57.23**	**40.88**	**24.05**	**66.38**	**47.22**	**27.74**	**63.34**	**48.31**	**36.04**	**71.96**	**56.78**	**43.04**
	Few-Shot	59.52	36.59	25.20	65.76	46.34	26.88	61.04	42.14	34.58	71.89	56.29	41.76
	+TALENT	**60.47**	**38.25**	**26.62**	**68.05**	**48.10**	**28.80**	**63.82**	**46.50**	**37.42**	**73.83**	**59.11**	**44.23**

Table 1: Average Results on 9 different language families in COMET, BLEURT, and ChrF++. We report on translating from English into low-resource languages (Eng-Low) and from low-resource languages into English (Low-Eng). Bold text denotes better results between TALENT and its corresponding baseline. We also highlight the greatly improved results with underline.

- **BLOOMZ** (Muennighoff et al., 2023). We employ the publicly available BLOOMZ to gauge the effectiveness of TALENT. BLOOMZ is a multitask model instruction fine-tuned based on BLOOM (Workshop, 2023), which ranks as one of the most multilingual LLMs and has been trained in 46 languages. For our experiments, we utilize its 7.1B model.

Baselines and Hyper-parameters. We report the translation results of TALENT on two baseline settings: zero-shot and few-shot. For few-shot baselines, we randomly choose 3 sentence pairs from the corresponding test set of FLORES-200 as demonstrations. For the "few-shot + TALENT" setting, we opt for a single sentence pair. Empirically, we set distinct hyper-parameters for two LLMs. For BLOOMZ, we set $N = 0.1$ and $K = 2$, while $N = 0.1$ and $K = 3$ are applied for ChatGPT.

Datasets and Evaluation Metrics. To thoroughly evaluate the impact of TALENT on low-resource languages, we report translation results encompassing English and 112 low-resource languages from the FLORES-200 benchmark, which spans diverse domains and topics. We use the dev-test partition of FLORES-200, containing 1012 sentences for each language. Appendix A provides further data statistics. As for evaluation metrics, we report three widely-utilized metrics following prior baselines (He et al., 2023; Ghazvininejad et al., 2023):

- **COMET.** COMET is a neural framework to evaluate machine translation models with a high correlation with human judgments. Among different model settings in COMET, we take the newest "Unbabel/wmt22-comet-da" model as the scorer following baselines (He et al., 2023).
- **BLEURT.** BLEURT is another model-based metrics widely-used in machine translation re-

searches (Fan et al., 2020; Li and Liang, 2021; He et al., 2023). BLEURT indicates to what extent machine output is fluent and conveys the meaning of the reference based on the contextual embeddings from language models.

- **ChrF++.** ChrF++ measures the quality of a translation through a character N-gram F-score by unigrams and bigrams. ChrF++ has been the second most popular metric and is highly recommended (Kocmi et al., 2021).

4.2. Main Results

The empirical outcomes for both English-to-low-resource (Eng-Low) and low-resource-to-English (Low-Eng) translation directions are presented in Table 1. The results have been averaged across language families. Key observations from Table 1 are as follows:

TALENT demonstrates consistent improvements across different settings and LLMS. From Table 1, we observe that TALENT shows an averaged improvement of 6.8% and 5.2% on zero-shot and few-shot settings, respectively. The further improvement on Few-shot settings suggests that TALENT can provide supplementary language-specific insights beyond demonstrations. What's more, TALENT outperforms the BLOOMZ and ChatGPT baselines by margins of 14.8% and 9.2%. The consistent improvement across both LLMS shows the versatility of the language-specific knowledge encapsulated within TALENT.

TALENT enhances performance across metrics and language families. As shown in Table 1, TALENT brings an increase of 17.5% in BLEURT. The improvement further rises to 23.3% in COMET and 31.2% in ChrF++. These results underscore the dual advantage of TALENT's translations, being both fluent (as indicated by COMET and BLEURT) and accurate (as indicated by ChrF++). Notably, the translation performance varies significantly across different language families, ranging from an average of 34.25 (Austroasiatic) to 70.51 (Dravidian) across both LLMS. However, TALENT consistently enhances translation across these diverse language families, averaging a 5.7% improvement and achieving an even more significant 13.9% boost in the Turkic family. We also find some abnormal results in BLOOMZ in two language families: Atlantic-Congo and Dravidian, where the zero-shot performance is higher than the few-shot ones. This suggests that BLOOMZ may have inflated performance due to the data leakage issue (Zhu et al., 2023a; Workshop, 2023). Further analysis on Table 1 shows that after applying TALENT, the standard deviation of the performances across different language families decreases from 10.6 to 9.4. This further proves that TALENT can mitigate the translation performance discrepancy across different language families.

4.3. Ablation Study: Exploring TALENT's Impact

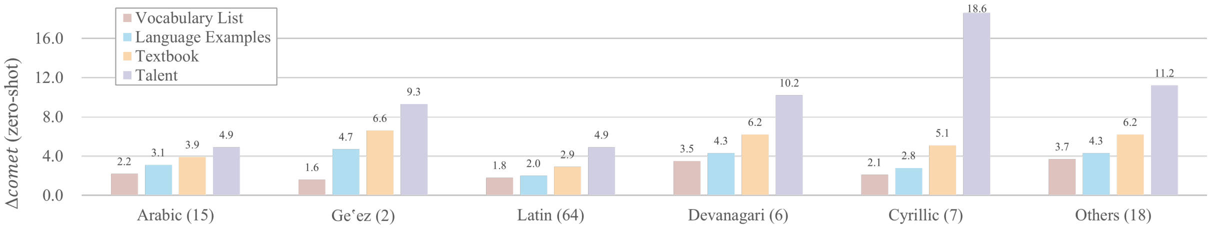
Given the substantial difference in pre-training data between high-resource and low-resource languages (Nguyen et al., 2023), the symmetry between English-to-low-resource (Eng-Low) and low-resource-to-English (Low-Eng) translations is disrupted. Eng-Low can measure LLMS' ability to generate low-resource languages (Natural Language Generation), while Low-Eng gauges their comprehension of sentence meanings in low-resource languages (Natural Language Understanding). To delve into these dynamics, we perform targeted experiments in both translation directions, probing the influence of different TALENT components. Figure 3 presents the outcomes of these investigations.

Performance of TALENT on Different Language Scripts. Observing Figure 3, TALENT delivers a commendable 3.1 COMET score improvement across all 6 language scripts, showcasing the robustness of TALENT. This finding implies the efficacy of TALENT for diverse language scripts, enhancing the generative capacity of LLMS in low-resource contexts. In Figures 3a and 3b, the top 2 improvements occur in the Cyrillic and Ge'ez scripts, with an increase of 6.9 and 4.4 COMET scores, respectively.

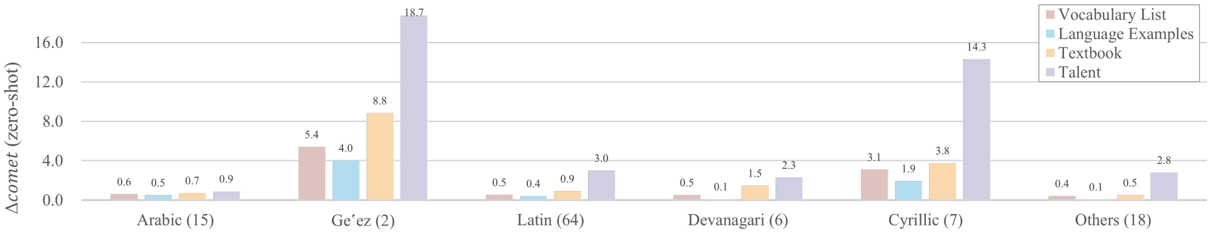
4.3.1. Influence of Textbook.

To thoroughly analyze Textbook's impact, we individually assess the contribution of each section and report the COMET results on Eng-Low and Low-Eng translation directions in Figure 3a and 3b. Drawing from the outcomes, we deduce the following insights regarding the two sections:

(a) Language Examples Improve Generation Capability: The translation direction Eng-Low requires LLMS to generate a coherent sentence in low-resource languages. From Figure 3a, we observe that Language Examples can improve the performance on Eng-Low direction by 3.5 COMET score, which is 42.3% higher than the improvement made by Vocabulary List. As depicted in Figure 3a, the application of Language Examples yields a substantial 3.5 COMET score improvement in the Eng-Low context, which outpaces that achieved by Vocabulary List by a noteworthy 42.3%. As further illustrated in Table 3, incorporating sentences from low-resource languages can also help LLMS to align



(a) Ablation Study for Talent in Eng-Low direction.



(b) Ablation Study for Talent in Low-Eng direction.

Figure 3: Ablation study of Talent. ΔCOMET quantifies how much better the performance is compared to the performance achieved with zero-shot baselines. We report the averaged results on 6 language scripts in two directions.

unfamiliar language tags with their corresponding language tokens.

(b) Vocabulary List Enhances Language Comprehension: In contrast, the application of Vocabulary List emerges as a catalyst for improving LLMs' comprehension of low-resource languages. The results in Figure 3b affirm that a mere integration of Vocabulary List contributes an average improvement of 1.8 COMET score. This achievement surpasses that of Language Examples by a substantial 46.6%. Vocabulary List can furnish lexical alignment insights between English and low-resource languages. Hence, English entries for chosen keywords in Vocabulary List can convey adequate semantic information, which enables precise disambiguation and comprehension of the source sentence.

Impact of Absorption Stage. We report the results of the Absorption Stage with the performance gap between "Talent" and "Textbook" in Figure 3a and 3b. The results highlight that acquiring syntactic insights for low-resource sentences yields substantial enhancements of 3.3 COMET score across diverse language scripts and both translation directions. We conjecture that due to the complexity of syntax, LLMs require a separate Absorption Stage to extract Syntax Patterns, which can improve translation performance. The quality of the Syntax Patterns is listed in Table 3.

4.4. Translation Performance on Non-English-Centric Directions

We expand our evaluation to assess how Talent enhances the translation capabilities of LLMs in scenarios where no high-resource languages are involved. We randomly select 10 translation directions from the pool of 112 low-resource languages and compare the performance of Talent against zero-shot and few-shot baselines. The outcomes are presented in Table 2. Talent surpasses the "pipeline" translation method by an additional 3.1%, indicating its capacity to offer lexical and syntactic component alignment between the source and target languages. With Talent, language-specific knowledge is distorted into basic elements, which alleviates the need for an intermediary pivot language.

4.5. Retrieval Utility and Quality

We present the Retrieval Utility ($\mathbf{RU}$) (Guu et al., 2020) to show how LLMs utilize Textbook and Retrieval Quality ($\mathbf{RQ}$) to measure the accuracy of each component in Talent in Table 3.

Retrieval Utility Following REALM (Guu et al., 2020), we report the retrieval utility to measure the usefulness of retrieved knowledge. We define the retrieval utility ($\mathbf{RU}$) of retrieval knowledge $\mathbf{z}$ (Vocabulary List, examples, or Syntax Patterns) for the given source sentence $\mathbf{x}$ as the difference between the log-likelihood of the knowledge-augmented results and the basic results:

$$\mathbf{RU}(\mathbf{z}) = \frac{1}{|\mathbf{y}|}\sum \log(\mathbf{y}|\mathbf{z},\mathbf{x}) - \frac{1}{|\mathbf{y}|}\sum \log(\mathbf{y}|\mathbf{x}), \tag{6}$$

where $|\mathbf{y}|$ denotes the length of the target sentence $\mathbf{y}$. A negative RU means that $\mathbf{z}$ is useless for predicting $\mathbf{y}$. The $\mathbf{RU}$ results are consistent with the

Method	src	amh_Ethi	bak_Cyrl	ibo_Latn	lao_Laoo	nya_Latn	sag_Latn	smo_Latn	tat_Cyrl	tgk_Cyrl	uig_Arab	Avg.
	tgt	lao_Laoo	amh_Ethi	hye_Armn	snd_Arab	sag_Latn	lin_Latn	lao_Laoo	hye_Armn	amh_Ethi	tgk_Cyrl	
Zero-Shot		9.76	6.74	17.45	8.17	4.95	18.05	17.13	24.06	7.80	16.33	13.04
Few-Shot		12.10	7.42	20.38	8.18	11.03	20.21	17.46	24.91	7.93	19.75	14.94
Pipeline		16.45	9.30	**22.14**	**16.27**	5.00	16.92	18.84	25.11	9.63	**23.62**	16.33
TALENT		**16.82**	**10.41**	20.63	10.97	**13.69**	**22.15**	**19.92**	**25.87**	**11.46**	20.99	**16.83**

Table 2: Translation Performance on Non-English-Centric Directions. "Pipeline" means we translate the source sentence into English and then translate the English sentence into the target language. We only utilize the Textbook for languages on the target side. Results are shown in ChrF++ for translation on 10 Low-Low directions. The best results are bolded.

Method/Metric	Retrieval Utility(↑)		Retrieval Quality(↑)		Off-Target-Rate(↓)	
	Eng-Low	Low-Eng	Low	Eng (ref)	Low-Low	Eng-Low
Zero-shot	-	-	-	-	0.29	0.21
Vocabulary List	0.89	1.48	0.61	0.67	0.12	0.06
Language Examples	1.19	1.14	0.57	0.83	0.07	0.03
Syntax Patterns	2.13	1.82	0.74	0.98	0.06	0.03

Table 3: Retrieval Utility $\mathbf{RU}$, Retrieval Quality $\mathbf{RQ}$, and Off-target Analysis on LLMS. We further report $\mathbf{RQ}$ score for English as a reference.

observations in Figure 3. In the Eng-Low direction, Language Examples are notably more advantageous, with a $\mathbf{RU}$ score 4.4% higher than that of the Vocabulary List. While we note contrasting results in the Low-Eng direction.

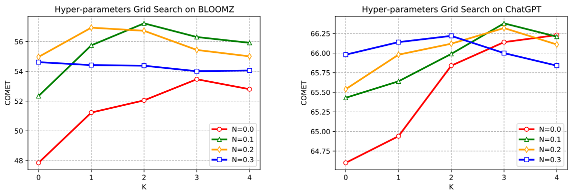
Retrieval Quality Since the Vocabulary List, examples, and Syntax Patterns are different kinds of knowledge, we define different metrics to measure their qualities. For the Vocabulary List, we use the proportion of the words in the Vocabulary List that do exist in the target sentence to reflect the quality of the Vocabulary List. Formally, for retrieved target words,

$$\mathbf{RQ}(\mathbf{VL}) = \frac{\sum_{\mathbf{y}} \sum_i \mathbb{1}(w_{\mathbf{y}}^{(i)} \subseteq \mathbf{y})}{\sum_{\mathbf{y}} \sum_i w_{\mathbf{y}}^{(i)}} \tag{7}$$

As for Retrieval Quality, we observe that the retrieval quality for Syntax Patterns is 0.74, which is relatively lower than that for English. However, LLMS can still gain improvements after applying Syntax Patterns. Consequently, even if the quality of Syntax Patterns and Language Examples is not exceptionally high, LLMS can still glean valuable insights from them, thereby enhancing their translation capabilities.

4.6. Off-Target Analysis

When translating to low-resource languages, the target-side results can contain multiple languages, for LLMS struggle to recognize unfamiliar language tags (off-target problem). We randomly select 10 languages and calculate the off-target rate as shown in Table 3. The results show that direct target information in the context (simply as words in Vocabulary List) can alleviate the off-target problem. Meanwhile, supplying sentences in target languages proves more crucial in aiding off-target problem than language tags.



(a) BLOOMZ (b) ChatGPT

Figure 4: Hyper-parameters Grid Search on BLOOMZ and ChatGPT, respectively. We randomly select 30 languages from 112 low-resource languages and use the average COMET score on Eng-Low directions to investigate the influence of these two hyper-parameters.

4.7. Hyper-parameters Grid Search

Hyper-parameter K determines how many Language Examples we choose in Textbook and N defines how strictly we select the keywords. Since both the number of Language Examples and keywords can affect the length of input prompt in LLMS, we jointly evaluate the influence of these hyper-parameters, as shown in Figure 4a and 4b. When no Language Examples are selected ($K = 0$), $N = 0.2$ achieves the best performance on BLOOMZ. However, when applying both Vocabulary List and Language Examples, the limitation of the total input length may restrict the translation performance. From Figure 4a, we conjecture that ChatGPT can accommodate a larger length of input tokens than BLOOMZ. Heuristically, we set $K = 2, N = 0.1$ for BLOOMZ and $K = 3, N = 0.1$ for ChatGPT to get the best performance.

5. Conclusion

Motivated by the foreign language learning paradigm of humans, we propose the Translate After LEarNing Textbook (TALENT) method, which applies a separate Absorption Stage for LLMs with a retrieved target Textbook before translation. Improved results on 112 low-resource languages show that TALENT can enhance the ability of LLMs to comprehend low-resource languages and provide sufficient language knowledge to generate accurate and fluent sentences.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.U2336202). This work is also supported by the National Natural Science Foundation of China (Grant No.U21B2009).

6. Bibliographical References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, Shen Gao, Jiafeng Guo, Xiangnan He, Yanyan Lan, Chenliang Li, Yiqun Liu, Ziyu Lyu, Weizhi Ma, Jun Ma, Zhaochun Ren, Pengjie Ren, Zhiqiang Wang, Mingwen Wang, Ji-Rong Wen, Le Wu, Xin Xin, Jun Xu, Dawei Yin, Peng Zhang, Fan Zhang, Weinan Zhang, Min Zhang, and Xiaofei Zhu. 2023. Information retrieval meets large language models: A strategic report from chinese ir community.

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Ali Araabi, Vlad Niculae, and Christof Monz. 2023. Joint dropout: Improving generalizability in low-resource neural machine translation through phrase pair variables.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Neel Bhandari and Pin-Yu Chen. 2023. Lost in translation: Generating adversarial examples robust to round-trip translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Ellen Bialystok, Fergus IM Craik, and Gigi Luk. 2012. Bilingualism: consequences for mind and brain. *Trends in cognitive sciences*, 16(4):240–250.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instruct-align: Teaching novel languages with to llms through alignment-based cross-lingual instruction.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

G. De Angelis. 2007. *Third Or Additional Language Acquisition*. G - Reference,Information and Interdisciplinary Subjects Series. Multilingual Matters.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Danish Ebadulla, Rahul Raman, S. Natarajan, Hridhay Kiran Shetty, and Ashish Harish Shenoy. 2023. Exploring linguistic similarity and zero-shot learning for multilingual translation of dravidian languages.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english?

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *ArXiv*, abs/2010.11125.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022a. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022b. Language-agnostic bert sentence embedding.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation.

Catherine Gitau and VUkosi Marivate. 2023. Textual augmentation techniques applied to low resource machine translation: Case of swahili.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. 2023. Rethinking translation memory augmented neural machine translation.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models.

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.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation.

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

Josef Jon, Dušan Variš, Michal Novák, João Paulo Aires, and Ondřej Bojar. 2023. Negative lexical constraints in neural machine translation.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2023. Boosting cross-lingual transferability in multilingual models via in-context learning.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.

Aswanth Kumar, Anoop Kunchukuttan, Ratish Puduppully, and Raj Dabre. 2023. In-context example selection for machine translation using multiple features.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *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 4582–4597, Online. Association for Computational Linguistics.

Yang Liu and Yuexian Hou. 2023. Syntax-aware complex-valued neural machine translation.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023a. Chain-of-dictionary prompting elicits translation in large language models.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023b. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023a. Sail: Search-augmented instruction learning.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023b. Dr.icl: Demonstration-retrieved in-context learning.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023c. Augmented large language models with parametric knowledge guiding.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Attila Nagy, Dorina Petra Lakatos, Botond Barta, Patrick Nanys, and Judit Ács. 2023. Data augmentation for machine translation via dependency subtree swapping.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis.

T. Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge Applied Linguistics. Cambridge University Press.

Seokjin Oh, Su ah Lee, and Woohwan Jung. 2023. Data augmentation for neural machine translation using generative language model.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed prompting for machine translation between related languages using large language models.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. 2023. Cross-lingual supervision improves large language models pre-training.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models.

David Soong, Sriram Sridhar, Han Si, Jan-Samuel Wagner, Ana Caroline Costa Sá, Christina Y Yu, Kubra Karagoz, Meijian Guan, Hisham Hamadeh, and Brandon W Higgs. 2023. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Learning to retrieve in-context examples for large language models.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. 2022. Preserving in-context learning ability in large language model fine-tuning.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yuzhuang Xu, Shuo Wang, Peng Li, Xuebo Liu, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Pluggable neural machine translation models via memory-augmented adapters.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into chatgpt's customizability.

Wayne Yang and Garrett Nicolai. 2023. Neural machine translation data generation and augmentation using chatgpt.

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

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

Fengji Zhang, Bei Chen, Yue Zhang, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. Repocoder: Repository-level code completion through iterative retrieval and generation.

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

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Extrapolating large language models to non-english by aligning languages.