Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models

Shuang Li[1], Jiangjie Chen[1]*, Siyu Yuan[2], Xinyi Wu[1], Hao Yang[3], Shimin Tao[3], Yanghua Xiao[1,4]*

[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]School of Data Science, Fudan University
[3]Huawei Translation Services Center
[4]Fudan-Aishu Cognitive Intelligence Joint Research Center
{lishuang18, jjchen19, wuxinyi20, shawyh}@fudan.edu.cn, syyuan21@m.fudan.edu.cn,
{yanghao30, taoshimin}@huawei.com

Abstract

To translate well, machine translation (MT) systems and general-purposed language models (LMs) need a deep understanding of both source and target languages and cultures. Therefore, idioms, with their non-compositional nature, pose particular challenges for Transformer-based systems, as literal translations often miss the intended meaning. Traditional methods, which replace idioms using existing knowledge bases (KBs), often lack scale and context-awareness. Addressing these challenges, our approach prioritizes context-awareness and scalability, allowing for offline storage of idioms in a manageable KB size. This ensures efficient serving with smaller models and provides a more comprehensive understanding of idiomatic expressions. We introduce a multilingual idiom KB (IDIOMKB) developed using large LMs to address this. This KB facilitates better translation by smaller models, such as BLOOMZ (7.1B), Alpaca (7B), and InstructGPT (6.7B), by retrieving idioms' figurative meanings. We present a novel, GPT-4-powered metric for human-aligned evaluation, demonstrating that IDIOMKB considerably boosts model performance. Human evaluations further validate our KB's quality.

Introduction

Idioms are non-compositional expressions whose *figurative meanings* deviate from the meanings of the constituent words (*literal meanings*) (Bobrow and Bell 1973; Swinney and Cutler 1979; Salton, Ross, and Kelleher 2018; Fadaee, Bisazza, and Monz 2018). For example, the figurative meaning of the idiom "bite the bullet" is "to endure a painful situation", deviating from the literal meanings of the constituent words "bite" and "bullet". Given the diverse array of idioms across various cultures and languages, appropriately translating texts that contain idioms (*idiomatic texts*) has become an important research problem (Strakšienė et al. 2009; Tzou, Vaid, and Chen 2017; Shao et al. 2018; Qiang et al. 2023).

However, due to the non-compositionality of idioms, idiomatic translation poses a significant challenge for current machine translation (MT) systems and general-purposed language models (LMs). Traditional phrase-based statistical machine translation systems do not give special consideration to idioms, resulting in low-quality translations (Salton, Ross, and Kelleher 2014a; Manojlovic, Dajak, and Bakaric 2017). Transformer-based MT models (Vaswani et al. 2017; Dankers, Lucas, and Titov 2022) usually treat idioms as compositional expressions, leading to literal translation errors. As shown in Figure 1, they usually translate idiomatic texts based on the literal meaning, failing to convey the intended information. One way to address the idiomatic translation problem would be scaling up the model size and training data, *i.e.*, large language models (LLMs) (Ouyang et al. 2022; OpenAI 2022), where various strong abilities emerge (Wei et al. 2022a). However, deploying models of such sizes for offline scenarios or real-time responses is costly and demanding. Therefore, the research question arises: *Can we enable smaller or specialized models to do idiomatic translation better?*

To tackle this challenge, one intuitive solution is to utilize the figurative meanings of the idioms, which are equivalent to semantically literal expressions, as support for translation. Recent research in linguistics and education has leveraged this insight to develop idiom knowledge bases (KBs), which have been employed to evaluate (Cucchiarini, Hubers, and Strik 2020; Wang 2021) and assist (Jiang et al.

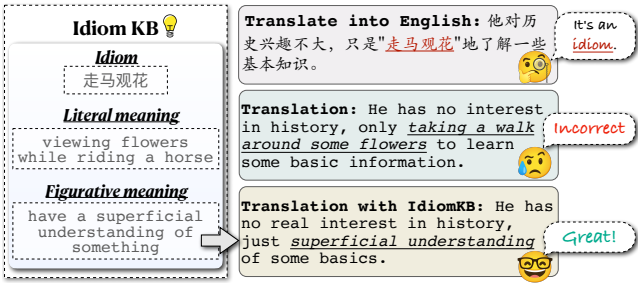

Figure 1: An example of idiomatic translation from Chinese to English. Current machine translation systems incorrectly translate idiomatic texts based on the *literal meaning* of the idiom, resulting in unsatisfactory translation. Incorporating *figurative meaning* from an idiom knowledge base (IDIOMKB) improves the translation performance.

*Corresponding authors.

2018; Tang 2022) second language learners' comprehension of idioms and idiomatic texts. Some studies leverage idiom KBs as valuable transition aids for models to accurately infer idiomatic figurative meanings without dealing with non-compositional expressions (Salton, Ross, and Kelleher 2014b; Modh and Jatinderkumar 2021). However, obtaining these meanings from dictionaries or manual annotation is time-consuming and suffers from idiom coverage issues. As a result, existing idiom KBs are relatively small in scale and notably lack multilingual meanings. This hinders translations since the idiom in the source sentence may not be included in the KB, or the figurative meaning in the target language could be absent.

In this paper, we propose IDIOMKB, a machine-generated KB consisting of idioms and their multilingual meanings, to facilitate the translation of idiomatic texts for LM-based MT systems. To circumvent the need for labor-intensive and expensive human efforts, we adopt the idea of symbolic knowledge distillation (West et al. 2022; Yuan et al. 2023; Bhagavatula et al. 2023) and employ LLMs to distill multilingual figurative meanings of the idioms based on their powerful generation ability. Based on the human evaluation, our IDIOMKB exhibits high quality, with an average score of 2.92 out of 3. Furthermore, we propose to incorporate the figurative meanings from IDIOMKB into idiomatic translation as a ***transition*** between source idiomatic texts and target language, which is similar to Chain-of-Thought (CoT) prompting (Wei et al. 2022b). This method differs from earlier MT systems which directly replace idiomatic expressions with their figurative meanings (Salton, Ross, and Kelleher 2014b; Modh and Jatinderkumar 2021). As demonstrated in Figure 1, our approach effectively incorporates the figurative meaning of the idiom into the current context in the idiomatic text.

Additionally, previous automatic evaluation metrics for MT (Ali and Renals 2018; Wang et al. 2023; Chen et al. 2023) only analyze entire sentences without assessing idiomatic translation quality explicitly. To address this issue, we propose an automatic evaluation metric based on GPT-4 (OpenAI 2023), which analyzes different aspects of idiomatic translation quality more effectively and with a higher correlation to human judgments.

Our contributions are summarized as follows: *1)* To tackle the non-compositional nature of idioms, we propose a multilingual idiom knowledge base, *i.e.*, IDIOMKB, to enhance idiomatic translation, particularly for smaller LMs. *2)* We propose a better method for idiomatic translation with IDIOMKB, utilizing the figurative meanings of idioms in IDIOMKB as a transition for more accurate idiomatic translation. *3)* We design a new metric based on GPT-4 to assess idiomatic translation, which is better aligned with human annotations. This metric demonstrates that our approach improves idiomatic translation quality.

Related Work

Non-compositionality and Idioms in Machine Translation Non-compositional multiword expressions (MWEs), notably idioms, which cannot have their meanings directly derived from their component words, present a significant challenge in various tasks (Lin 1999; Zhu, Guo, and Sobhani 2015; Hwang and Hidey 2019). These expressions further complicate the machine translation process due to their non-compositional nature (Tzou, Vaid, and Chen 2017; Dankers, Lucas, and Titov 2022; Dankers, Bruni, and Hupkes 2022). Previous research has proposed specific strategies such as identifying these MWEs, particularly idioms, learning distinct embeddings for them, or reformulating them into simpler, more understandable phrases (Weller et al. 2014; Ullman and Nivre 2014; Hashimoto and Tsuruoka 2016; Constant et al. 2017). Furthermore, evidence has shown that machine translation of idioms can be improved by incorporating parallel meanings from dictionaries or other external resources (Salton, Ross, and Kelleher 2014b; Zaninello and Birch 2020; Modh and Jatinderkumar 2021). Drawing inspiration from this, we build IDIOMKB to assist smaller models in idiomatic translation.

Resources of Idiom Knowledge To enhance the ability of neural models to comprehend idiomatic text, researchers have developed various resources, which can be categorized into two types: *1) Multilingual Idioms Datasets* (Moussallem et al. 2018; Agrawal et al. 2018; Shao et al. 2018; Qiang et al. 2023; Tang 2022): which consist of parallel translations of idioms in one language to another to improve the idiomatic translation. *2) Monolingual Idioms Datasets* (Jiang et al. 2018; Zheng, Huang, and Sun 2019; Saxena and Paul 2020; Tan and Jiang 2021; Adewumi et al. 2022): which focus on discerning idiomatic expressions within a single language. In this paper, we employ LLMs to construct IDIOMKB, setting itself apart with its large scale and the feature of containing multilingual idiom meanings.

Large Language Models With the recent great success of LLMs (Brown et al. 2020; Ouyang et al. 2022; Leiter et al. 2023; OpenAI 2023), *in-context learning* and *instruction learning* have become prevailing paradigms for deploying LLMs for downstream tasks (Min et al. 2022; Ram et al. 2023). Through these paradigms, LLMs can directly generate high-quality outputs for given tasks without parameter updates (Rubin, Herzig, and Berant 2022; Chung et al. 2022). For dataset construction, LLMs can be a promising alternative to resource-intensive large-scale crowd-sourcing (West et al. 2022; Yuan et al. 2023) to improve the performance of smaller and specialized models, even surpassing teacher models in some settings. Furthermore, compared to existing traditional n-gram-based metrics like BLEU (Papineni et al. 2002) or neural metrics such as COMET (Rei et al. 2022), LLM-based evaluation can offer greater flexibility and better alignment with humans on more challenging tasks, particularly in a reference-free setting (Wang et al. 2023; Chen et al. 2023; Luo, Xie, and Ananiadou 2023). In this paper, we employ them in the construction of IDIOMKB and the assessment of idiomatic translation quality.

IDIOMKB for Idiomatic Translation

In this section, we create an idiom KB, *i.e.*, IDIOMKB, to provide figurative meanings of idioms to smaller LMs to improve their translation quality.

IDIOMKB Construction: Knowledge Distillation from LLMs

Recent research demonstrates that LLMs are promising alternatives to costly large-scale crowd-sourcing for constructing datasets (West et al. 2022; Yuan et al. 2023). Therefore, we leverage LLMs to distill large-scale multilingual figurative meanings of idioms to create IDIOMKB.

Source Data Collection To construct a comprehensive KB, we consider the coverage of idioms in each language and gather idioms across multiple datasets to create multilingual idiom lists. These idioms are collated from three languages, *i.e.*, English (En), Chinese (Zh), and Japanese (Ja):

- **English**: Our English idioms are sourced from the MAGPIE (Haagsma, Bos, and Nissim 2020), IMIL (Agrawal et al. 2018), EPIE (Saxena and Paul 2020), and PIE (Zhou, Gong, and Bhat 2021) datasets. MAGPIE contains Potentially Idiomatic Expressions (PIEs) in context. IMIL includes idiom translations in several Indian languages. EPIE contains idioms categorized as static or formal based on lexical changes and includes their English meanings. PIE provides parallel idiomatic and literal sentences.
- **Chinese**: For Chinese idioms, we use the PETCI (Tang 2022) and CCT (Jiang et al. 2018) datasets. PETCI includes idiom English translations from dictionaries, Google and DeepL. CCT is a cloze test dataset that includes Chengyu, the most prevalent Chinese idioms. The ChID dataset (Zheng, Huang, and Sun 2019), another cloze-style Chinese idiom dataset, also contributes to our Chinese idiom collection.
- **Japanese**: The Japanese segment is built on the OpenMWE (Hashimoto and Kawahara 2008) and ID10M (Tedeschi, Martelli, and Navigli 2022) datasets. OpenMWE is designed for idiom identification and includes many idiomatic and literal sentences per idiom. ID10M collects idioms from several languages but does not include their meanings.

Idiomatic Meanings Distillation from LLMs We inherit the idea of symbolic knowledge distillation from models and use LLMs via in-context learning to generate figurative meanings of idioms for IDIOMKB construction. As shown in Table 1, we first manually design instructions emphasizing the non-compositional nature of idioms. Then, we randomly select several idioms for each language pair and manually annotate their meanings from online dictionaries for reference as examples for in-context learning. For example, we can extract the English meaning of "一气呵成", "to complete a task or work in one go, without stopping or taking a break" from the LLM output to construct IDIOMKB. By relying on the ability of LLMs to comprehend and generate accurate figurative meanings for idioms, our method is both straightforward and computationally efficient, as it avoids processing large amounts of text.

We report the statistics of our KB and other existing idiom corpora for comparison. As shown in Table 2, IDIOMKB boasts a larger number of idioms with multilingual figurative meanings, setting it apart from the existing corpora that provide only monolingual meanings or completely lack figurative meanings. Through the utilization of IDIOMKB, we can retrieve the idiomatic meanings of idioms and effectively enhance smaller models when translating idiomatic text.

KB Meaning Generation

/ Task prompt */*
Given a Chinese idiom, please write the idiom's figurative **English** meaning. Please note: Idiom always expresses figurative meaning which is different from literal meaning of its constituent words.
/ Examples */*
Case 1:
Chinese idiom: 明目张胆
English meaning: straightforwardly, without any concealment
...
/ Test Data */*
Case 5:
Chinese idiom: 一气呵成
English meaning: *to complete a task or work in one go, without stopping or taking a break*

Table 1: Prompt for LLMs to generate IDIOMKB. The prompt contains four examples. Generated texts are *underlined*.

Idiom Lang.	Dataset	Size	Meaning Lang.
English (En)	PIE	1,197	En
	IMIL	2,208	Indian
	EPIE	717	En
	MAGPIE	1,756	-
	IDIOMKB	**3,990**	En, Zh, Ja
Chinese (Zh)	PETCI	4,310	En
	CCT	7,395	Zh
	ChID	3,848	-
	IDIOMKB	**8,643**	En, Zh, Ja
Japanese (Ja)	ID10M	165	-
	OpenMWE	146	-
	IDIOMKB	**270**	En, Zh, Ja

Table 2: Comparison between IDIOMKB with other existing idiom corpora in different languages.

IDIOMKB Application: Improving Idiomatic Translation

To bypass non-compositional expressions in idiomatic translation, we propose to prompt smaller LMs with the figurative meanings of idioms retrieved from IDIOMKB to guide the translation process.

First, we identify the target idiom within the source sentence to correlate it with an idiom in IDIOMKB. We adopt the idiom-sentence pairs in the source datasets rather than retrieved idioms. We think that idiom detection has already been studied, allowing us to focus our efforts on translation tasks without idiom detection as an intermediate step. Once the target idiom is identified, we can retrieve its correspond-

I: Direct Prompt

Translate the following Chinese sentence into English.
Chinese: 为使讨论一气呵成，我们会在本报告第381至396段回应这些关注。
English: *We will answer these questions in the report's 381-396 sections.*

II: KB-CoT

"一气呵成" means ***"to complete a task or work in one go, without stopping or taking a break"***.
Given the above knowledge, translate the following Chinese sentence into English.
Chinese: 为使讨论一气呵成，我们会在本报告第381至396段回应这些关注。
English: *To make the discussion flow smoothly, we will respond to these questions in 381-396 sections of the report.*

III: Self-CoT

"一气呵成" means *"a success that comes as a result of one's without stopping or taking a break"*.
Given the above knowledge, translate the following Chinese sentence into English.
Chinese: 为使讨论一气呵成，我们会在本报告第381至396段回应这些关注。
English: *In order to have a successful discussion, we will respond to these concerns in the 381-396 sections of this report.*

Table 3: Instruction for InstructGPT (6.7B) to translate the idiomatic text with direct (*i.e.*, Direct Prompt) and CoT prompts which utilize meanings retrieved from IDIOMKB (*i.e.*, KB-CoT) or self-generated meanings (*i.e.*, Self-CoT). The meaning retrieved from IDIOMKB is highlighted in *bold*. Texts generated by InstructGPT (6.7B) are *underlined*.

ing multilingual meanings from IDIOMKB.

Then we incorporate idiom meaning in the prompt when doing idiomatic translation. The prompt begins by formulating a short task description introducing the machine translation task. We provide the meaning of the idiom retrieved from IDIOMKB to encourage smaller LMs to focus on comprehending the non-literal meaning of the idiom before attempting a full translation. As shown in Table 3 (II), smaller LMs can successfully translate the Chinese idiomatic text into English by obtaining the correct figurative English meaning of the Chinese idiom "一气呵成" from IDIOMKB as a hint (*i.e.*, KB-CoT prompt). Compared to directly translating in Table 3 (I), the KB-CoT prompt guides LMs to focus on understanding the non-literal meaning of idioms before translating them in context, resulting in more accurate idiomatic translations. Furthermore, we explore the Self-CoT method in Table 3 (III), where smaller LMs generate the meaning without IDIOMKB. However, the inaccurate meaning leads to an incorrect translation, highlighting the benefits of using the correct meaning retrieved from IDIOMKB to enhance the performance of smaller LMs.

Metrics for Idiomatic Translation

In this section, we aim to establish evaluation criteria for evaluating translations of idiomatic text. As mentioned above, idioms are unique linguistic constructs characterized by their figurative meanings which frequently deviate from their literal ones. This difference renders the general metrics unsuitable for idiomatic translation evaluation, as they treat idioms equally with other compositional parts and fail to have a profound understanding of idioms. Due to the lack of aligned data between idiomatic text and its respective translations in other languages, we need to develop a reference-free metric for evaluation. To this end, we draw from linguistic studies to formulate a comprehensive method for determining the quality of idiomatic translations using LLMs.

/ Task prompt */*
Evaluate the idiom translation in the given Chinese translation of an English sentence. Focus on the idiom's figurative meaning.
/ Evaluation Criteria */*
1 point: Ignores, mistranslates, or only translates the literal meaning of the idiom.
2 points: Conveys basic figurative meaning but may lack refinement or have minor imperfections.
3 points: Exceptional translation, accurately conveying figurative meaning, context, and cultural nuances.
/ Test Data */*
Evaluate the following translation:
English sentence: <source>
Idiom in the English sentence: <idiom>
Chinese translation: <translation>
Evaluation (score only): *<score>*

Table 4: The prompt for LLMs to evaluate idiomatic translation quality. This prompt specifies in detail the translation quality and ask the LLM to generate <*score*>.

Evaluation Criteria

As shown in Table 4, we design a prompt based on a 1-3 point evaluation criteria to help LLM focus on idiomatic translation quality while providing detailed guidelines for each point on the scale. In the evaluation criteria, a 1-point score reflects poor idiom translation, 2 points indicate a basic, though imperfect understanding and 3 points represent an exceptional and accurate translation incorporating figurative meaning, context, and cultural nuances. Then LLMs are provided with the test data and asked to generate a score-only evaluation.

Can LLMs Evaluate Idiomatic Translation?

We manually construct a small-scale evaluation set on three language pairs, *i.e.*, Chinese-to-English (Zh→En), English-to-Chinese (En→Zh), and Japanese-to-English (Ja→En), from source language datasets:

- **English**: PIE Corpus (Zhou, Gong, and Bhat 2021), focusing on idiomatic sentence generation and paraphrasing with 1,197 idioms and 5,170 related sentences.
- **Chinese**: We extract idioms from the PETCI dataset (Tang 2022) and identify sentences containing them in the WMT22 dataset (Kocmi et al. 2022).

Pair	Metric	r	ρ	τ
Zh→En	BLEU	0.0936	0.0660	0.0515
	COMET	0.2510	0.2511	0.1984
	GPT-4	**0.6939**	**0.6923**	**0.6375**
En→Zh	BLEU	0.3368	0.3277	0.2484
	COMET	0.5367	0.5186	0.4029
	GPT-4	**0.7891**	**0.7879**	**0.7338**
Ja→En	COMET	0.4174	0.4031	0.3198
	GPT-4	**0.6708**	**0.6718**	**0.6174**

Table 5: Pearson's r, Spearman's ρ and Kendall's τ correlations between human and sacreBLEU, COMET or GPT-4 evaluation in different language pairs. Note that 'BLEU' stands for sacreBLEU and Ja→En translation does not have sacreBLEU results because the Japanese sentences lack corresponding English references.

- **Japanese**: OpenMWE Corpus (Hashimoto and Kawahara 2008), containing 67,575 sentences with 146 idioms for idiom token identification.

We first randomly select 800 data samples from each source language dataset and translate them into the target language utilizing three distinct language models: two smaller LMs, InsructGPT (6.7B) (Ouyang et al. 2022) and BLOOMZ (7.1B) (Muennighoff et al. 2023), and one LLM, namely InstructGPT$_{003}$ (*i.e.*, `text-davinci-003`) (Ouyang et al. 2022), a variant of GPT-3 (Brown et al. 2020) tuned on instructions using reinforcement learning with human feedback (RLHF). We opt for this approach to ensure that translation generated by models of different types and sizes can all be evaluated correctly. Then we annotate 20 sentence and translation pairs for each point across all language pairs. We compare the evaluation results of the LLMs with the human-annotated results to determine their level of consistency by calculating Pearson's r (Pearson 1920), Spearman's ρ (Spearman 1987), and Kendall's τ (Kendall 1948) correlations.

The results in Table 5 demonstrate that LLMs can serve as an evaluator for the translation quality of idiomatic expressions across different language pairs. Conversely, sacreBLEU and CometKiwi (Rei et al. 2022) find it challenging to align with human evaluation. This difficulty arises largely because they lack the capacity to comprehend the nuances and meanings of idioms, thereby failing to accurately reflect human evaluations. We also explore using few-shot prompts and observe performance decline. This decline can potentially be due to the distraction caused by the presence of examples. They may cause LLMs to lose focus on the current test example, thereby leading to biases.

Experiments

Experimental Settings

IDIOMKB Construction For IDIOMKB construction, we choose several LLMs to generate high-quality multilingual idiom meanings via in-context learning: GPT-3.5 series[1], including InstructGPT$_{003}$ (∼175B), ChatGPT (OpenAI 2022) and multilingual LMs, *i.e.*, BLOOM (176B) (Scao et al. 2022) and BLOOMZ (176B). GPT-3 is an auto-regressive LLM with billions of parameters achieving strong performance on NLP tasks. InstructGPT$_{003}$ is a variant of GPT-3 (Brown et al. 2020) fine-tuned on instructions and code via reinforcement learning with human feedback (RLHF). ChatGPT is a dialogue-oriented model that is built on InstructGPT with RLHF. BLOOM is a multilingual language model, and BLOOMZ is built on BLOOM using multitask-prompted fine-tuning.

Idiomatic Translation For idiomatic translation, We choose mBART (680M) (Liu et al. 2020) as a representative of multilingual Transformer models, which is an autoregressive sequence-to-sequence model and has strong performance on machine translation. NLLB model (1.3B, distilled) (Team et al. 2022) is a supervised MT model distilled from a 54.4B Mixture-of-Experts model NLLB-200 to improve performance on low-resource languages. We also choose InstructGPT (6.7B), BLOOMZ (7.1B) and Alpaca (Taori et al. 2023), which are all instruction-finetuned for better performance. We also present the results of ChatGPT and GPT-4 as the upper bound for this task.

The dataset used for idiomatic translation is the same as the one employed for generating the translation evaluation set to ensure that idiomatic texts are of high quality. Due to budget constraints, we randomly select 500 instances from each dataset for evaluation. To ensure a fair comparison between direct prompting and KB-CoT prompting, we employ the same task description presented in Section and set the temperature to 0.7 for all generations.

For evaluation, we set the temperature to 0.1, intending for less randomness. We also adopt two additional metrics, sacreBLEU and CometKiwi, where sacreBLEU represents the n-gram-based evaluation, and CometKiwi represents the reference-free neural-based evaluation. For Zh→En, we directly use the parallel English text in WMT22 as a reference. For En→Zh, we first acquire the parallel literal English text from the PIE dataset and then use ChatGPT to translate the literal text into Chinese. We use the ChatGPT's translation result as a reference. For Ja→En, as there is no parallel text available for Japanese idiomatic text, we only report GPT-4 evaluation score and CometKiwi since they can be conducted in a reference-free setting.

Can IDIOMKB Improve Idiomatic Translation for Small LMs?

Main Results and Analysis The results in Table 6 show that: *1)* KB-CoT prompting with meaning retrieved from IDIOMKB consistently surpasses direct prompting for smaller models, revealing the universality of our KB-CoT method; *2)* For direct prompting, BLOOMZ (7.1B) outperforms InstructGPT (6.7B) and Alpaca while gaining less for KB-CoT prompting. This means that BLOOMZ (7.1B) itself shows relatively strong performance on idiomatic translation. However, its ability to combine instruction with translation is relatively weaker; *3)* For ChatGPT and GPT-4, KB-CoT prompting also improves idiomatic translation quality

[1]Note that OpenAI does not release detailed information about GPT-3.5s.

Model	Size	Setting	Zh→En			En→Zh			Ja→En	
			BLEU	COMET	GPT-4	BLEU	COMET	GPT-4	COMET	GPT-4
mBART	560M	Direct	**30.64**	**82.35**	**2.09**	**43.74**	**75.93**	**1.69**	**74.51**	**1.48**
NLLB	1.3B	Direct	25.03	80.75	1.95	21.64	64.98	1.65	70.62	1.44
InstructGPT	6.7B	Direct	14.26	72.73	1.66	50.20	62.76	1.50	**68.96**	1.34
		KB-CoT	9.64	73.88	2.08	13.92	65.49	1.99	67.50	**1.64**
BLOOMZ	7.1B	Direct	20.60	**79.39**	2.11	49.41	**76.88**	2.08	65.29	1.22
		KB-CoT	15.40	77.26	**2.21**	**50.59**	73.78	**2.17**	63.29	1.54
Alpaca	7B	Direct	24.80	71.87	1.54	21.36	44.28	1.11	65.82	1.23
		KB-CoT	**29.66**	72.74	2.12	5.92	48.09	1.46	65.71	1.57
ChatGPT	?B	Direct	25.90	**82.57**	2.74	**26.89**	**79.87**	2.62	**77.37**	2.52
		KB-CoT	**26.42**	82.01	**2.82**	24.87	77.91	**2.71**	76.32	**2.61**
GPT-4	?B	Direct	21.40	**82.43**	2.73	24.87	**79.74**	2.73	**77.55**	2.63
		KB-CoT	**26.42**	81.38	**2.86**	**32.21**	77.89	**2.83**	76.02	**2.69**

Table 6: The translation performance of LMs in different language pairs. The source language is either directly translated (Direct) or generated via KB-CoT prompting with meaning from IDIOMKB (KB-CoT). The best results are bolded. Note that 'BLEU' stands for sacreBLEU and Ja→En translation does not have sacreBLEU results because the Japanese sentences lack corresponding English references.

Model	Resource	Zh→En	Ja→En
InstructGPT (6.7B)	- (Direct)	1.66	1.34
	Self	1.69	1.35
	BLOOM	1.97	1.36
	BLOOMZ	2.07	1.47
	InstructGPT_{003}	2.07	1.46
	ChatGPT	**2.08**	**1.64**
BLOOMZ (7.1B)	- (Direct)	2.11	1.22
	Self	2.14	1.22
	BLOOM	2.15	1.28
	BLOOMZ	2.20	1.41
	InstructGPT_{003}	2.19	1.38
	ChatGPT	**2.21**	**1.54**

Table 7: Translation results with different sourced meanings retrieved from IDIOMKB constructed by different LLMs.

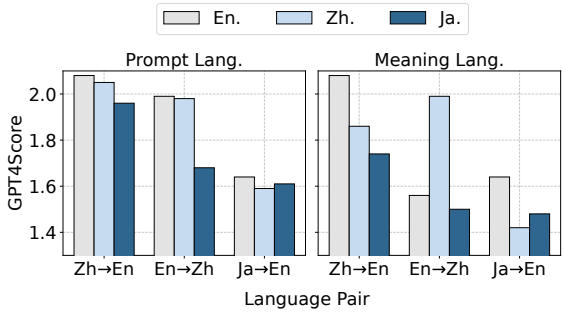

Figure 2: The performance of InstructGPT (6.7B) + KB-CoT, with prompts and meanings of different languages generated by ChatGPT.

compared to direct prompting. This indicates that the use of KB-CoT prompting is not solely beneficial to smaller LMs, but can also enhance the performance of LLMs such as ChatGPT and GPT-4 for idiomatic translation task. However, it should be noted that GPT-4 may exhibit a bias towards the translation generated by GPT models (Liu et al. 2023); *4)* Both pre-trained model mBART and supervised-MT model NLLB encounter difficulties in accurately translating idiomatic text; *5)* scareBLEU does not align well with the idiomatic translation quality score assessed by GPT-4, indicating that n-gram based metrics are unsuitable for idiomatic translation evaluation. CometKiwi, as a sentence-level metric, performs better than sacreBLEU but cannot evaluate idiom translation quality properly.

What language should we use to form prompts? We compare the performance of InstructGPT (6.7B) with KB-CoT prompts of different languages. The results in Figure 2 demonstrate the superiority of utilizing English prompts. A potential reason is that the instruction-tuning datasets predominantly consist of English data. The findings suggest English is ideal for creating KB-CoT prompts across various language pairs, which could stem from the inherent complexity and rich vocabulary of the language, enabling precision in delivering complex instructions (Shi et al. 2023).

What language should we use for idiom meaning? To construct IDIOMKB, we ask LLMs to generate multilingual meanings for idioms. Since this translation task involves multiple languages, deciding which language meaning to provide the model as a reference is critical. We compare InstructGPT (6.7B)'s performance using KB-CoT prompts in English and evaluate its ability to incorporate meanings in different languages on the same idiomatic translation datasets as above. The results in Figure 2 indicate that when performing Zh→En, En→Zh and Ja→En translations, utilizing the meaning in the target language yields better results. This could be due to the inherent structural and semantic differences between these languages. When translating, retaining the meaning in the target language ensures

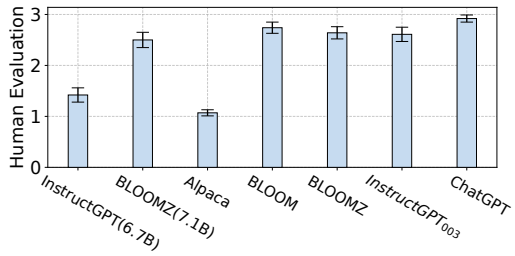

Figure 3: Human evaluation of different meaning sources. We use a wide range of models, including LLMs and relatively smaller models to obtain meanings.

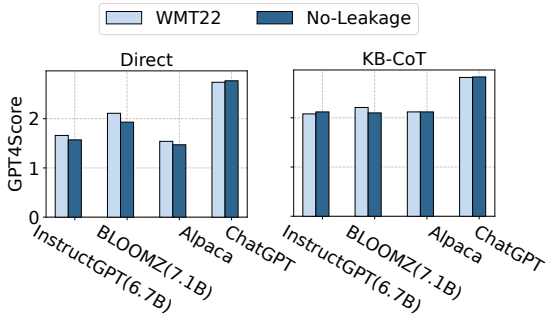

Figure 4: Model performance on idiomatic text extracted from WMT22 and no-leakage dataset (Zh→En).

better context understanding and cultural sensitivity. Moreover, syntax variances, for example, word order in Chinese or Japanese differs greatly from in English, making direct translation complex and often inaccurate. Hence the utilization of meaning in the target language proves more effective.

How Is the Quality of IDIOMKB Under Human Evaluation?

Which model generates the best IDIOMKB? We manually annotate the quality of idiom meaning generated by different LMs utilizing the method in § . Each model generates Chinese meanings of 100 randomly selected Chinese idioms. To evaluate idiom quality, we assign points ranging from 1-3, with 1 indicating a completely inaccurate meaning, 2 indicating the meaning requiring minor refinements, and 3 indicating a perfect capture of nuanced cultural meanings. The results in Figure 3 show that IDIOMKB generated by ChatGPT produces the highest quality results, which are consistent with the translation performance presented in Table 7. In contrast, smaller models struggle to generate high-quality idiom meanings independently.

Will data leakage affect idiomatic translation? Although we strive to select the most current datasets, there remains a possibility that LMs may have encountered the specific sentences we employ. To address this concern, we follow the approach of Zhu et al. (2023) and manually create a no-leakage dataset consisting of 60 recently published news sentences that contain idioms. Then we compare the Zh-En idiomatic translation performance of LMs on our idiomatic translation test set extracted from WMT22 and the no-leakage dataset. The results in Figure 4 indicate that the effect of data leakage on idiomatic translation is negligible.

Source sentence: "即使是发达经济国家也不能永远寅吃卯粮。" **Figurative meaning**: A metaphor for economic hardship and inadequate income, borrowing and misappropriating in advance **Literal meaning**: Eat the grain of the year of the Tiger during the year of the Rabbit. **Reference**: "Even developed economies cannot **[live beyond their means forever]**. "
mBART: "Even the advanced economies will not be able to <*eat their fill*> forever. " ⇒ Literal Translation Error **NLLB**: "Even in developed economies, <*food cannot be eaten*> forever. " ⇒ Literal Translation Error
InstructGPT (6.7B) (Direct): "Even a developed country cannot indefinitely <*subsist on rice porridge*>. "⇒ Literal Translation Error **InstructGPT (6.7B) (KB-CoT)**: "Even a developed economy can't always indefinitely **[live off tomorrow's food]** indefinitely. "⇒ Correct Translation
BLOOMZ (7.1B) (Direct): "Even in developed economies, it is not possible to <*live on borrowed time*> forever. "⇒ Idiom Misunderstanding **BLOOMZ (7.1B) (KB-CoT)**: "Even in developed economies, it is not possible to **[live off tomorrow's resources today]**. "⇒ Correct Translation

Table 8: A comparative case study illustrating the performance of mBART, NLLB, InstructGPT (6.7B), and BLOOMZ (7.1B) on Zh→En idiomatic translation. Each model's translation of the Chinese idiom "寅吃卯粮" is given in <*italics*> (inaccurate) or [bold] (accurate).

Case Study We present the case study in Table 8. "寅吃卯粮" is a Chinese idiom referring to spending resources in advance, which literally translates as eating the food stored up for the next year. "Yin" (寅) and "Mao" (卯), which refer to the year of tiger and rabbit in this context, are the 3rd and 4th terms in the 12 Earthly Branches widely used in traditional Chinese calendars and horoscopic astrology, and thus the model needs to understand the cultural nuances to translate this idiom accurately. Consistent with the main experiment, all four translation models struggle with this idiom to differing extents. mBART and NLLB generate literal translations such as "eat their fill" and "food cannot be eaten", respectively. Similarly, InstructGPT (6.7B) with direct prompting also makes a literal translation error, mistakenly translating the idiom as "subsist on rice porridge" while BLOOMZ (7.1B) misunderstands the idiom as "live on borrowed time". Conversely, when enhanced by IDIOMKB and KB-CoT, the performance of InstructGPT (6.7B) and BLOOMZ (7.1B) significantly improves. Both models successfully capture and convey the figurative meaning of the idiom. Interestingly, there are idioms of similar meanings in different cultures and languages, such as, in this case, the English id-

iom "robbing Peter to pay Paul", which means using next month's (or period's) resources to cover this month's (or period's) expenses. We believe our IDIOMKB would be very useful in studying the cross-culture alignments of idioms, which we leave for future work.

Conclusion

In this paper, we present a solution to tackle the challenges of idiomatic translation, which can be applied to various sizes of models. We develop IDIOMKB, a multilingual idiom knowledge base that leverages the figurative meanings of idioms as a transitional aid to prevent non-compositional expressions. We build IDIOMKB from LLMs, which are finite and can be stored offline, and then retrieve the idiom meanings from the KB and add them in the CoT prompting to improve translation quality. Furthermore, we introduce an automatic evaluation method with GPT-4 to assess the translation quality of idioms, showing the effectiveness of our approach. We believe IDIOMKB will be a valuable resource to advance the research on idiomatic translation and the study of the cross-culture alignment of idiomatic expressions.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work is supported by Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), National Natural Science Foundation of China (No.62102095), and Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). Yanghua Xiao is also a member of Research Group of Computational and AI Communication at Institute for Global Communications and Integrated Media, Fudan University.

References

Adewumi, T.; Vadoodi, R.; Tripathy, A.; Nikolaido, K.; Liwicki, F.; and Liwicki, M. 2022. Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms. In *Proc. of LREC*.

Agrawal, R.; Chenthil Kumar, V.; Muralidharan, V.; and Sharma, D. 2018. No more beating about the bush : A Step towards Idiom Handling for Indian Language NLP. In *Proc. of LREC*.

Ali, A.; and Renals, S. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In *Proc. of ACL*.

Bhagavatula, C.; Hwang, J. D.; Downey, D.; Le Bras, R.; Lu, X.; Qin, L.; Sakaguchi, K.; Swayamdipta, S.; West, P.; and Choi, Y. 2023. I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation. In *Proc. of ACL*.

Bobrow, S. A.; and Bell, S. M. 1973. On catching on to idiomatic expressions. *Memory & cognition*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Proc. of NeurIPS*.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: A Preliminary Empirical Study. arXiv:2304.00723.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Constant, M.; Eryiğit, G.; Monti, J.; van der Plas, L.; Ramisch, C.; Rosner, M.; and Todirascu, A. 2017. Survey: Multiword Expression Processing: A Survey. *CL*.

Cucchiarini, C.; Hubers, F.; and Strik, H. 2020. Learning L2 idioms in a CALL environment: the role of practice intensity, modality, and idiom properties. *CALL*.

Dankers, V.; Bruni, E.; and Hupkes, D. 2022. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proc. of ACL*.

Dankers, V.; Lucas, C.; and Titov, I. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proc. of ACL*.

Fadaee, M.; Bisazza, A.; and Monz, C. 2018. Examining the Tip of the Iceberg: A Data Set for Idiom Translation. In *Proc. of LREC*.

Haagsma, H.; Bos, J.; and Nissim, M. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proc. of LREC*.

Hashimoto, C.; and Kawahara, D. 2008. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-Specific Features. In *Proc. of EMNLP*.

Hashimoto, K.; and Tsuruoka, Y. 2016. Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings. In *Proc. of ACL*.

Hwang, A.; and Hidey, C. 2019. Confirming the Non-compositionality of Idioms for Sentiment Analysis. In *Proc. of MWE-WN 2019*.

Jiang, Z.; Zhang, B.; Huang, L.; and Ji, H. 2018. Chengyu Cloze Test. In *Proc. of BEA*.

Kendall, M. G. 1948. Rank correlation methods.

Kocmi, T.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Gowda, T.; Graham, Y.; Grundkiewicz, R.; Haddow, B.; Knowles, R.; Koehn, P.; Monz, C.; Morishita, M.; Nagata, M.; Nakazawa, T.; Novák, M.; Popel, M.; and Popovic, M. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Conference on Machine Translation*.

Leiter, C.; Zhang, R.; Chen, Y.; Belouadi, J.; Larionov, D.; Fresen, V.; and Eger, S. 2023. ChatGPT: A Meta-Analysis after 2.5 Months. *arXiv preprint arXiv:2302.13795*.

Lin, D. 1999. Automatic Identification of Non-compositional Phrases. In *Proc. of ACL*.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Proc. of TACL*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *In Proc. of EMNLP*.

Luo, Z.; Xie, Q.; and Ananiadou, S. 2023. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization. arXiv:2303.15621.

Manojlovic, M.; Dajak, L.; and Bakaric, M. B. 2017. Idioms in state-of-the-art Croatian-English and English-Croatian SMT systems. *In Proc. of MIPRO*.

Min, S.; Lewis, M.; Zettlemoyer, L.; and Hajishirzi, H. 2022. MetaICL: Learning to Learn In Context. In *Proc. of NAACL*.

Modh, J. C.; and Jatinderkumar, R. 2021. Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms. *ACSA*.

Moussallem, D.; Sherif, M. A.; Esteves, D.; Zampieri, M.; and Ngonga Ngomo, A.-C. 2018. LIdioms: A Multilingual Linked Idioms Data Set. In *Proc. of LREC*.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Le Scao, T.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; and Raffel, C. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proc. of ACL*.

OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-08-15.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Gray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*.

Pearson, K. 1920. Notes on the history of correlation. *Biometrika*.

Qiang, J.; Li, Y.; Zhang, C.; Li, Y.; Zhu, Y.; Yuan, Y.; and Wu, X. 2023. Chinese Idiom Paraphrasing. *Proc. of TACL*, 11.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Rei, R.; Treviso, M.; Guerreiro, N. M.; Zerva, C.; Farinha, A. C.; Maroti, C.; C. de Souza, J. G.; Glushkova, T.; Alves, D.; Coheur, L.; Lavie, A.; and Martins, A. F. T. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proc. of SIGMT*.

Rubin, O.; Herzig, J.; and Berant, J. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proc. of NAACL*.

Salton, G.; Ross, R.; and Kelleher, J. 2014a. An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In *Proc. of HyTra*.

Salton, G.; Ross, R.; and Kelleher, J. 2014b. Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation. In *Proc. of MWE*.

Salton, G. D.; Ross, R. J.; and Kelleher, J. D. 2018. Exploring the Use of Attention within an Neural Machine Translation Decoder States to Translate Idioms. arXiv:1810.06695.

Saxena, P.; and Paul, S. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Proc. of TSD*.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

Shao, Y.; Sennrich, R.; Webber, B.; and Fancellu, F. 2018. Evaluating Machine Translation Performance on Chinese Idioms with a Blacklist Method. In *Proc. of LREC*.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; Das, D.; and Wei, J. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh ICLR*.

Spearman, C. 1987. The proof and measurement of association between two things. *The American journal of psychology*.

Strakšienė, M.; et al. 2009. Analysis of idiom translation strategies from English into Lithuanian. *Kalbų studijos*.

Swinney, D. A.; and Cutler, A. 1979. The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*.

Tan, M.; and Jiang, J. 2021. Learning and Evaluating Chinese Idiom Embeddings. In *Proc. of RANLP 2021*.

Tang, K. 2022. PETCI: A Parallel English Translation Dataset of Chinese Idioms. arXiv:2202.09509.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model.

Team, N.; Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; Sun, A.; Wang, S.; Wenzek, G.; Youngblood, A.; Akula, B.; Barrault, L.; Gonzalez, G. M.; Hansanti, P.; Hoffman, J.; Jarrett, S.; Sadagopan, K. R.; Rowe, D.; Spruit, S.; Tran, C.; Andrews, P.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Goswami, V.; Guzmán, F.; Koehn, P.; Mourachko, A.; Ropers, C.; Saleem, S.; Schwenk, H.; and Wang, J. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.

Tedeschi, S.; Martelli, F.; and Navigli, R. 2022. ID10M: Idiom Identification in 10 Languages. In *Proc. of ACL Findings*.

Tzou, Y.-Z.; Vaid, J.; and Chen, H.-C. 2017. Does formal training in translation/interpreting affect translation strategy? Evidence from idiom translation. *Bilingualism: Language and Cognition*.

Ullman, E.; and Nivre, J. 2014. Paraphrasing Swedish Compound Nouns in Machine Translation. In *Proc. of MWE*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048.

Wang, X. 2021. Applying cognitive linguistics to second language idiom learning.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022a. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022b. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Proc. of NeurIPS*.

Weller, M.; Cap, F.; Müller, S.; Schulte im Walde, S.; and Fraser, A. 2014. Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. In *Proc. of ComAComA 2014*.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J.; Jiang, L.; Le Bras, R.; Lu, X.; Welleck, S.; and Choi, Y. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proc. of NAACL*.

Yuan, S.; Chen, J.; Fu, Z.; Ge, X.; Shah, S.; Jankowski, C.; Xiao, Y.; and Yang, D. 2023. Distilling Script Knowledge from Large Language Models for Constrained Language Planning. In *Proc. of ACL*.

Zaninello, A.; and Birch, A. 2020. Multiword Expression aware Neural Machine Translation. In *Proc. of LREC*.

Zheng, C.; Huang, M.; and Sun, A. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proc. of ACL*.

Zhou, J.; Gong, H.; and Bhat, S. 2021. PIE: A Parallel Idiomatic Expression Corpus for Idiomatic Sentence Generation and Paraphrasing. In *Proc. of MWE*.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. arXiv:2304.04675.

Zhu, X.; Guo, H.; and Sobhani, P. 2015. Neural Networks for Integrating Compositional and Non-compositional Sentiment in Sentiment Composition. In *Proc. of LCS*.