Leveraging Multilingual Knowledge Graph to Boost Domain-specific Entity Translation of ChatGPT

Min Zhang zhangmin186@huawei.com
Limin Liu cecilia.liulimin@huawei.com
Yanqing Zhao zhaoyanqing@huawei.com
Xiaosong Qiao qiaoxiaosong@huawei.com
Chang Su suchang8@huawei.com
Xiaofeng Zhao zhaoxiaofeng14@huawei.com
Junhao Zhu zhujunhao@huawei.com
Ming Zhu zhuming47@huawei.com
Song Peng pengsong2@huawei.com
Yinglu Li liyinglu@huawei.com
Yilun Liu liuyilun3@huawei.com
Wenbing Ma mawenbing@huawei.com
Mengyao Piao piaomengyao1@huawei.com
Shimin Tao taoshimin@huawei.com
Hao Yang yanghao30@huawei.com
Yanfei Jiang jiangyanfei@huawei.com
Huawei Translation Services Center, Beijing, 100095, China

Abstract

Recently, ChatGPT has shown promising results for Machine Translation (MT) in general domains and is becoming a new paradigm for translation. In this paper, we focus on how to apply ChatGPT to domain-specific translation and propose to leverage Multilingual Knowledge Graph (MKG) to help ChatGPT improve the domain entity translation quality. To achieve this, we extract the bilingual entity pairs from MKG for the domain entities that are recognized from source sentences. We then introduce these pairs into translation prompts, instructing ChatGPT to use the correct translations of the domain entities. To evaluate this novel MKG method for ChatGPT, we conduct comparative experiments on three Chinese-English (zh-en) test datasets constructed from three specific domains, of which one domain is from biomedical science, and the other two are from the Information and Communications Technology (ICT) industry — Visible Light Communication (VLC) and wireless domains. Experimental results show that both the overall translation quality of ChatGPT (+6.21, +3.13 and +11.25 in BLEU scores) and the translation accuracy of domain entities (+43.2%, +30.2% and +37.9% absolute points) are significantly improved with MKG on the three test datasets.

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 77–87
September 4–8, 2023, Macau SAR, China.

1 Introduction

Recently, the emergence of ChatGPT[1] has brought remarkable influence on Natural Language Processing (NLP) tasks. It is an intelligent chatbot developed by OpenAI, based on InstructGPT (Ouyang et al., 2022). ChatGPT is built on the top of GPT-3.5 and GPT-4 families of Large Language Models (LLMs) and has been fine-tuned using both supervised and reinforcement learning techniques. It possesses diverse abilities of NLP, including question answering, dialogue generation, code debugging, generation evaluation (Qin et al., 2023; Zhong et al., 2023; Wang et al., 2023; Kocmi and Federmann, 2023; Lu et al., 2023). We are particularly interested in ChatGPT for domain-specific Machine Translation (MT) tasks, especially how domain-specific knowledge can be introduced into ChatGPT to improve the translation accuracy of domain entities.

Source: 甚至光纤上去一照变成黑光纤了。

Prompt for ChatGPT: 请将下面的句子翻译成英文：甚至光纤上去一照变成黑光纤了。

MT by ChatGPT: Even the fiber optic cable turned into a black fiber optic cable when it was lit up.

Reference: Even the optical fiber is irradiated to become a dark fiber.

Figure 1: An MT example by ChatGPT in the VLC domain, where 光纤 and 黑光纤 are two domain entities and the corresponding English translations are *optical fiber* and *dark fiber* respectively, and the prompt "请将下面的句子翻译成英文: ..." means "Please translate the following sentence into English: ...".

Recent studies (Jiao et al., 2023; Hendy et al., 2023; Peng et al., 2023; Zhao et al., 2023) on translation tasks have found that ChatGPT performs competitively with commercial translation products (e.g., Google Translate and Microsoft Translate) in the general domain, but performs poorly in specific domains. According to the analysis by Jiao et al. (2023) and Zhao et al. (2023), it is a challenge for ChatGPT to correctly translate domain entities. Fig. 1 shows a Chinese sentence that contains two domain entities 光纤 and 黑光纤 in the Visible Light Communication (VLC) domain. We use the translation prompt in Fig. 1 for ChatGPT to translate this sentence into English. From Fig. 1, it could be seen that the two entities are wrongly translated as *fiber optic cable* and *black fiber optic cable* by ChatGPT, and the correct translations are *optical fiber* and *dark fiber*. Although ChatGPT shows strong translation capabilities in general domains, it is not so effective in specific domains, especially in domain entity translation.

In this paper, we propose to utilize Multilingual Knowledge Graph (MKG) to help ChatGPT address this issue. After we introduce bilingual entity pairs from MKG into translation prompts, ChatGPT can use the correct translations of domain entities. Experimental results in three specific domains (biomedical science, VLC, and wireless) demonstrate that the overall translation quality of ChatGPT is improved by +6.21, +3.13 and +11.25 respectively in terms of BLEU scores, and the translation accuracy of domain entities is improved by +43.2%, +30.2% and +37.9% absolute points respectively.

The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose to introduce MKG for ChatGPT to improve the translation accuracy of domain entities.

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

- Experimental results in three specific domains demonstrate that the proposed method can significantly improve the translation quality of ChatGPT and the translation accuracy of domain entities.

2 Related Work

2.1 ChatGPT for MT

With ChatGPT showing remarkable capabilities in various NLP tasks, research on ChatGPT for MT has sprung up (Jiao et al., 2023; Hendy et al., 2023; Peng et al., 2023; Zhang et al., 2023).

Jiao et al. (2023) provided a preliminary evaluation of ChatGPT for MT, including translation prompt, multilingual translation, and translation robustness. Hendy et al. (2023) presented a comprehensive evaluation of ChatGPT for MT, covering various aspects such as quality of ChatGPT in comparison with state-of-the-art research and commercial systems, effect of prompting strategies, robustness towards domain shifts and document-level translation. These studies show that ChatGPT does not perform as well as commercial translation products in low-resource languages or specific domains.

Peng et al. (2023) proposed two simple yet effective prompts (task-specific and domain-specific prompts) to mitigate these issues. However, ChatGPT still faces great translation challenges if it is prompted only with the type name of an unfamiliar domain (i.e., ChatGPT knows little about the domain). In such case, domain-specific knowledge needs to be provided to ChatGPT (similar to the way by Zhang et al. (2023) for automatic post-editing). In this paper, we propose to introduce domain-specific MKG to ChatGPT.

2.2 Knowledge Graph for MT

Entity translation plays a particularly important role in determining the translation quality. Therefore, various methods (Shi et al., 2016; Lu et al., 2019; Moussallem et al., 2019; Zhao et al., 2020a,b; Zhang et al., 2022) are proposed to improve the entity translation quality with Knowledge Graph (KG).

With the help of KG, Shi et al. (2016) built and formulated a semantic space to connect the source and target languages, and applied it to the sequence-to-sequence framework to propose a Knowledge-Based Semantic Embedding method. Lu et al. (2019) utilized the entity relations in KG as constraints to enhance the connections between the source words and their translations. Under the hypothesis that KG could enhance the semantic feature extraction of neural models, Moussallem et al. (2019) proposed two strategies for incorporating KG into neural models without modifying the neural network architectures. Zhao et al. (2020a,b) not only proposed a multi-task learning method on sub-entity granularity for MT task and knowledge reasoning task, but also designed a novel KG enhanced Neural Machine Translation (NMT) method (i.e., transforming the source and target KGs into a unified semantic space). To apply the entity pairs of MKG, Zhang et al. (2022) proposed a data augmentation strategy for NMT.

In a word, all the above approaches are about how to introduce KG into neural networks for MT. In this paper, we utilize MKG to generate translation prompts for ChatGPT, which is very simple and effective.

3 Translation Prompt with MKG

The generation of MKG-based translation prompts for ChatGPT is described in Fig. 2, which consists of three steps:

(1) *Named Entity Recognition (NER)*: We extract domain entities from source sentences. In Fig. 2, two VLC domain entities 光纤 and 黑光纤 are extracted from the source sentence "甚至光纤上去一照变成黑光纤了。".

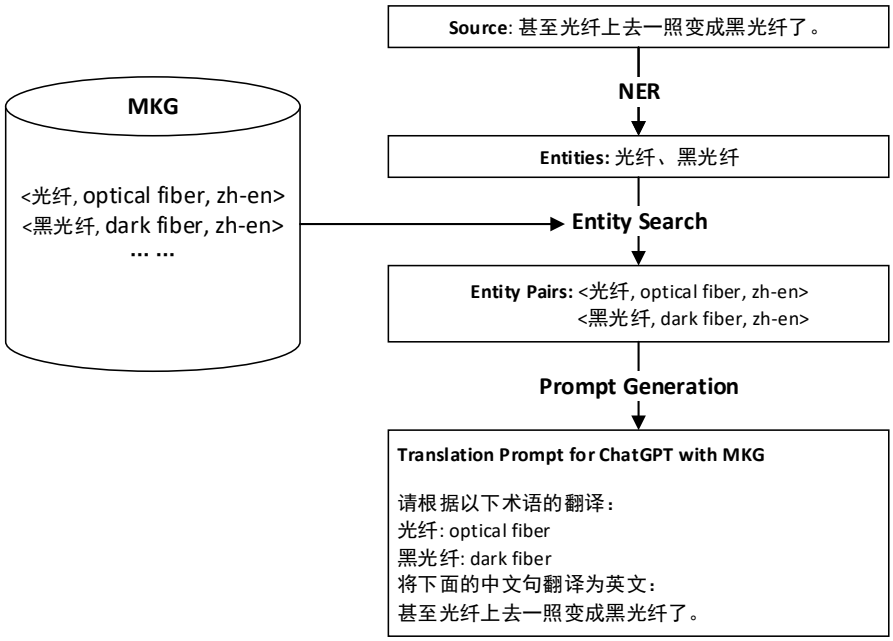

Figure 2: An example of translation prompt generation with MKG for ChatGPT.

(2) *Entity Search*: We search the MKG for the extracted entities to obtain bilingual entity pairs. In this paper, MKG is composed of triplets as <source entity, target entity, language pair>. In Fig. 2, by searching the MKG for 光纤 and 黑光纤, two entity pairs <光纤, optical fiber, zh-en> and <黑光纤, dark fiber, zh-en> are obtained.

(3) *Prompt Generation*: We introduce the obtained entity pairs into the translation prompt for ChatGPT, as the example illustrated in Fig. 2.

Source: 甚至光纤上去一照变成黑光纤了。

MT by ChatGPT: Even the fiber optic cable turned into a black fiber optic cable when it was lit up.

MT by ChatGPT + MKG: Even when light was shone on the optical fiber, it turned into dark fiber.

Reference: Even the optical fiber is irradiated to become a dark fiber.

Figure 3: Translation results of ChatGPT with or without MKG for the sentence in Fig. 2.

Fig. 3 shows the translation results of ChatGPT with or without MKG for a sentence in the VLC domain (i.e., using the prompts in Fig. 2 and Fig. 1 for ChatGPT respectively). It could be seen that ChatGPT correctly translates the two domain entities 光纤 and 黑光纤 with MKG, but incorrectly without MKG. This means that introducing MKG to translation prompts does improve the entity translation accuracy of ChatGPT, thereby contributing to the translation quality of ChatGPT.

Domain	# Sent	# Ent	Precision	Recall	$F1$ Score
Biomedical	264	574	0.933	0.977	0.954
VLC	397	586	0.955	0.952	0.953
Wireless	200	855	0.956	0.988	0.972

Table 1: NER results on the three test datasets from the three specific domains, where "# Sent" and "# Ent" indicate the number of source sentences and the number of entities in these sentences respectively.

System	Biomedical (zh-en)			VLC (zh-en)			Wireless (zh-en)		
	BLEU	ACC	COMET	BLEU	ACC	COMET	BLEU	ACC	COMET
ChatGPT	29.37	0.500	81.2	22.10	0.563	81.6	21.08	0.507	80.5
ChatGPT+MKG	35.58	**0.932**	**82.8**	25.23	**0.865**	**82.6**	**32.33**	**0.886**	**82.6**
Google Translate	**37.06**	0.527	81.6	**29.80**	0.619	81.6	26.13	0.529	80.0

Table 2: BLEU scores, ACC values and COMET scores of ChatGPT with or without MKG and Google Translate on the three test datasets.

4 Experiments

4.1 Experiment Settings

We provide a brief description of the experiment settings, which mainly includes the used models, test datasets, MKG and evaluation metrics.

4.1.1 Models

We compare the translation results of ChatGPT with or without MKG in specific domains. The results in this paper come from the `gpt-3.5-turbo` models, which power the ChatGPT. In addition, the results of Google Translate[2] are used for reference. The NER models are fine-tuned on 1,000 annotated domain sentences with traditional BERT-LSTM-CRF Devlin et al. (2019) architecture for the three domains respectively.

4.1.2 Data

To evaluate the translation quality of ChatGPT with or without MKG, we construct three zh-en test datasets from three specific domains. The datasets consist of 264 parallel sentences from WMT22 biomedical science domain Neves et al. (2022), 397 parallel sentences from the VLC domain, and 200 parallel sentences from the wireless domain. It should be pointed out that data in the biomedical science domain is public, and the data in the other two domains is our own non-public data. And the MKGs for the three domains consist of about 2,000, 1,000,000 and 3,000,000 triplets respectively, which are automatically mined from domain parallel corpora (Zhang et al., 2022).

4.1.3 Metrics

We adopt the commonly used BLEU score (Papineni et al., 2002) as our primary metric, which is calculated by the toolkit `SacreBLEU` (Post, 2018). And we use the COMET (Rei et al., 2020) metric as reference. Specifically, we use the reference-based metric COMET-22 (`wmt22-COMET-da`) (Rei et al., 2022). Additionally, we report the translation accuracy of entities (ACC) in the translations.

4.2 Experimental Results

Our NER model achieves very good recognition results on all the three test datasets, which is helpful for domain entity translation correctness. The NER results of the source sentences in the

[2]`https://translate.google.com`

Item	Content
Source	现有的免疫功能定义如免疫过度、免疫抑制、免疫麻痹、免疫耐受均停留在描述性的概念，缺乏定量的诊断标准，结合临床表现及生物学指标将是未来最有效的量化方式。
Reference	The existing definitions of immune function, such as excessive immunity, immunosuppression, immune paralysis and immune tolerance, remain in descriptive concepts and lacks quantitative diagnostic criteria. Clinical manifestations combined with biological indicators will be the most effective quantitative method in the future.
Entities from MKG	<免疫过度, excessive immunity, zh-en> <免疫抑制, immunosuppression, zh-en> <免疫麻痹, immune paralysis, zh-en> <免疫耐受, immune tolerance, zh-en>
ChatGPT	The current definition of immune function, such as immune overreactivity, immune suppression, immune paralysis, and immune tolerance, remains descriptive and lacks quantitative diagnostic criteria. Combining clinical presentations and biological markers will be the most effective way of quantification in the future.
ChatGPT+ MKG	The existing definitions of immune functions, such as excessive immunity, immunosuppression, immune paralysis, and immune tolerance, are still descriptive concepts lacking quantitative diagnostic criteria. Combining clinical manifestations and biological indicators will be the most effective way to quantify these concepts in the future.
Google Translate	Existing definitions of immune function, such as immune hyperactivity, immunosuppression, immune paralysis, and immune tolerance, remain descriptive concepts and lack quantitative diagnostic criteria. Combining clinical manifestations and biological indicators will be the most effective quantitative method in the future.

Table 3: A case that contains four domain entities in the biomedical science domain.

Item	Content
Source	综述颈椎结核手术治疗中不同内植物材料的特点及使用方法，以及新型内植物的研究进展。
Reference	To review the characteristics and applications of different implantation in cervical tuberculosis surgery and the research progress of the new implantation.
Entities from MKG	<内植物, implantation, zh-en> <颈椎结核, cervical tuberculosis, zh-en>
ChatGPT	This article reviews the characteristics and usage of different internal implant materials in surgical treatment for cervical spine tuberculosis, as well as the research progress of new internal implants.
ChatGPT+ MKG	A review of the characteristics and usage methods of different implant materials in surgical treatment of cervical tuberculosis, as well as the research progress of new implants.
Google Translate	This paper reviews the characteristics and application methods of different implant materials in the surgical treatment of cervical tuberculosis, as well as the research progress of new implants.

Table 4: A case that contains three domain entities in the biomedical science domain.

three test datasets are shown in Table 1, where the precision, recall and $F1$ scores are included.

The three metrics (BLEU, ACC and COMET) of ChatGPT with or without MKG and Google Translate are illustrated in Table 2. It could be seen that MKG-based translation prompts not only improve the BLEU scores for ChatGPT, but also significantly improve the ACC values (improving 43.2%, 30.2% and 37.9% absolute points for the biomedical science, VLC and wireless domains respectively). Although the BLEU scores of ChatGPT with MKG are lower than those of Google Translate in the biomedical science and VLC domains, the ACC values of ChatGPT with MKG are significantly higher in all the three domains (40.5%, 24.6% and 35.7% absolute points higher in the biomedical science, VLC and wireless domains respectively). Due to the strong particularity of the wireless domain, ChatGPT with MKG has made the most

Item	Content
Source	那在可见光波段，它的衰减是非常非常大的，单模光纤很难传，那怎么能够传可见光?
Reference	In the visible light band, its attenuation is very, very large, and single-mode optical fiber is difficult to transmit, then how to transmit visible light?
Entities from MKG	<可见光波段, visible light band, zh-en> <单模光纤, single-mode optical fiber, zh-en> <可见光, visible light, zh-en>
ChatGPT	In the visible light wavelength, the attenuation is very, very large, and it is difficult for single-mode fibers to transmit. So, how can visible light be transmitted?
ChatGPT+ MKG	In the visible light band, the attenuation is very significant, and it is difficult for single-mode optical fiber to transmit. How can visible light be transmitted?
Google Translate	In the visible light band, its attenuation is very, very large, and single-mode fiber is difficult to transmit, so how can it transmit visible light?

Table 5: A case that contains three domain entities in the VLC domain.

Type	Content
Source	电源线与空开连接器接触不良，如电源线剥线过短导致压绝缘皮等。
Reference	The connection between the power cable and circuit breaker connector is poor. For example, the insulation sheath is pressed because the stripped length is too short for the power cable.
Entities from MKG	<空开连接器, circuit breaker connector, zh-en> <绝缘皮, insulation sheath, zh-en> <电源线, power cable, zh-en>
ChatGPT	Poor contact between power cord and circuit breaker connector, such as insufficiently stripped wires of the power cord causing pressure on the insulation.
ChatGPT+ MKG	The power cable has poor contact with the circuit breaker connector, which may be caused by the stripping of the power cable being too short and pressing against the insulation sheath.
Google Translate	The power cord is in poor contact with the circuit breaker connector, for example, the stripping of the power cord is too short and the insulation is pressed.

Table 6: A case from the wireless domain, which contains four domain entities.

significant improvement in translation quality (+11.25 BLEU), scoring even 6.2 BLEU points higher than Google Translate. In addition, ChatGPT with MKG gets the best COMET scores in all the three domains (82.8, 82.6 and 82.6 in biomedical science, VLC and wireless domains respectively).

4.3 Case Study

In this section, we provide four cases from three specific domains, which are illustrated in Tables 3 to 6. For each case, the source sentence, its reference, the entities from MKG, and the translations from ChatGPT, ChatGPT+MKG and Google Translate are provided.

In Table 3, four domain entities are highlighted in blue in the source sentence and its reference. ChatGPT (without MKG) can only correctly translate two entities (免疫麻痹 and 免疫耐受), and the translations of the other two entities (免疫过度 and 免疫抑制) are wrong (in red). With the MKG-based translation prompt, ChatGPT can correctly translate all the four entities (in blue). Google Translate has one entity mistranslation in this sentence.

In Table 4, there are three domain entities in the source sentence and ChatGPT (without MKG) cannot correctly translate these entities. Although the correct translations of these entities from MKG are provided in the prompt, ChatGPT still cannot translate the entity "内植物" correctly. This suggests that better prompts need to be designed for ChatGPT.

From Table 5 and Table 6, ChatGPT with MKG can correctly translate all the domain entities, while ChatGPT without MKG and Google Translate cannot. This shows the effectiveness

Dataset	# Sentence	# Entity
WMT19 en-zh	1,997	5,492
WMT19 en-de	1,997	5,045

Table 7: Statistics of the en-zh and en-de datasets from WMT19, where "# Sentence" and "# Entity" denote the number of source sentences and the number of entities in these sentences respectively.

En-Zh translation prompt with MKG:
Based on the following translations for these entities:
Lifeguard: 救生员
Shark: 鲨鱼
Please translate the following sentence to Chinese:
Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, but it was determined not to be a dangerous species of shark.

En-Zh translation prompt without MKG:
Please translate the following sentence to Chinese:
Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, but it was determined not to be a dangerous species of shark.

En-De translation prompt with MKG:
Based on the following translations for these entities:
Lifeguard: Rettungsschwimmer
Shark: Hai
Please translate the following sentence into German:
Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, but it was determined not to be a dangerous species of shark.

En-De translation prompt without MKG:
Please translate the following sentence into German:
Lifeguard Capt. Larry Giles said at a media briefing that a shark had been spotted in the area a few weeks earlier, but it was determined not to be a dangerous species of shark.

Figure 4: Examples of translation prompts with or without MKG for en-zh and en-de datasets from WMT19.

System	WMT19 (en-zh)			WMT19 (en-de)		
	BLEU	ACC	COMET	BLEU	ACC	COMET
ChatGPT	33.02	0.737	84.8	38.70	0.785	**86.2**
ChatGPT+MKG	34.24	**0.905**	84.5	39.72	**0.869**	85.6
Google Translate	**37.68**	0.810	**85.4**	**42.85**	0.805	84.5

Table 8: BLEU scores, ACC values and COMET scores of ChatGPT with or without MKG on the en-zh and en-de datasets from WMT19.

of MKG-based translation prompts.

It is should be pointed out that the entities pairs provided for ChatGPT prompts must be correct, otherwise it is likely to deteriorate the translation results.

4.4 Effects on the News Domain

Since our domain-specific data (parallel sentences and MKG) for experiments is not yet publicly available, we further evaluate the MKG-based translation prompts for ChatGPT on public domain data. To the best of our knowledge, only in the annotated WMT19 news translation

task data by Gekhman et al. Gekhman et al. (2020), both the parallel sentences and MKG are publicly available. In this section, we choose the English-Chinese (en-zh) and English-German (en-de) datasets from WMT19 to evaluate the translation performance of ChatGPT with or without MKG. It should be noted that the translation prompts used for ChatGPT are the English versions as illusatrated in Fig. 4. The statistics of the two datasets are shown in Table 7.

The BLEU scores, ACC values and COMET scores of ChatGPT with or without MKG are illustrated in Table 8. It could be clearly seen that both the BLEU scores and the ACC values are improved by the MKG-based translation prompts for ChatGPT. However, because the news domain is not so particular as the above three specific domains (biomedical science, VLC and wireless domains), the BLEU scores and ACC values are improved modestly (+1.22 BLEU and +16.8 absolute points of ACC for en-zh, and +1.02 BLEU and +8.4 absolute points of ACC for en-de). Nevertheless, the MKG-based translation prompts proposed in this paper are very effective in improving the translation quality of ChatGPT and the translation accuracy of domain entities.

5 Conclusion

In this paper, a novel MKG-based translation prompt is designed for ChatGPT to improve the domain-specific translation quality, especially the translation accuracy of domain entities. Domain entities are first recognized from the source sentences and are used to extract corresponding bilingual entity pairs from MKG. Then MKG-based translation prompts are generated for ChatGPT with the bilingual entity pairs. Experimental results in the three specific domains demonstrate that the ChatGPT empowered by MKG-based translation prompts greatly improves the translation accuracy of the domain entities, even outperforming Google Translate. However, the BLEU scores of ChatGPT with the MKG-based translation prompts in two of the three domains are still lower than those of Google Translate, although they are greatly higher than those of the ChatGPT without MKG. This suggests that better prompt engineering needs to be developed for ChatGPT to further unlock its potential in translation. In addition, the performance of the MKG-based translation prompts for ChatGPT in the general domain (news domain) is also investigated.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gekhman, Z., Aharoni, R., Beryozkin, G., Freitag, M., and Macherey, W. (2020). KoBE: Knowledge-based machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jiao, W., Wang, W., tse Huang, J., Wang, X., and Tu, Z. (2023). Is chatgpt a good translator? a preliminary study. In *ArXiv*.

Kocmi, T. and Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *ArXiv*, abs/2302.14520.

Lu, Q., Qiu, B., Ding, L., Xie, L., and Tao, D. (2023). Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

Lu, Y., Zhang, J., and Zong, C. (2019). Exploiting knowledge graph in neural machine translation. In Chen, J. and Zhang, J., editors, *Machine Translation*, pages 27–38, Singapore. Springer Singapore.

Moussallem, D., Ngonga Ngomo, A.-C., Buitelaar, P., and Arcan, M. (2019). Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 139–146, New York, NY, USA. Association for Computing Machinery.

Neves, M., Jimeno Yepes, A., Siu, A., Roller, R., Thomas, P., Vicente Navarro, M., Yeganova, L., Wiemann, D., Di Nunzio, G. M., Vezzani, F., Gerardin, C., Bawden, R., Estrada, D. J., Lima-Lopez, S., Farre-Maduel, E., Krallinger, M., Grozea, C., and Neveol, A. (2022). Findings of the wmt 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*, pages 694–723, Abu Dhabi. Association for Computational Linguistics.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. (2023). Towards making the most of chatgpt for machine translation. *arxiv preprint*.

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

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? *ArXiv*, abs/2302.06476.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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 *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shi, C., Liu, S., Ren, S., Feng, S., Li, M., Zhou, M., Sun, X., and Wang, H. (2016). Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2245–2254, Berlin, Germany. Association for Computational Linguistics.

Wang, J., Liang, Y., Meng, F., Li, Z., Qu, J., and Zhou, J. (2023). Cross-Lingual Summarization via ChatGPT. *arXiv.org*.

Zhang, M., Peng, S., Yang, H., Zhao, Y., Qiao, X., Zhu, J., Tao, S., Qin, Y., and Jiang, Y. (2022). Entityrank: Unsupervised mining of bilingual named entity pairs from parallel corpora for neural machine translation. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3708–3713.

Zhang, M., Zhao, X., Yanqing, Z., Yang, H., Qiao, X., Zhu, J., Ma, W., Chang, S., Liu, Y., Li, Y., Wang, M., Peng, S., Tao, S., and Jiang, Y. (2023). Leveraging chatgpt and multilingual knowledge graph for automatic post-editing. In *International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*. accepted for publication.

Zhao, Y., Xiang, L., Zhu, J., Zhang, J., Zhou, Y., and Zong, C. (2020a). Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhao, Y., Zhang, J., Zhou, Y., and Zong, C. (2020b). Knowledge graphs enhanced neural machine translation. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4039–4045. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zhao, Y., Zhang, M., Chen, X., Deng, Y., Geng, A., Liu, L., Liu, X., Li, W., Jiang, Y., Yang, H., Han, Y., Tao, S., Li, X., Ma, M., Zhang, Z., and Xie, N. (2023). Human evaluation for translation quality of chatgpt: A preliminary study. In *International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*. accepted for publication.

Zhong, Q., Ding, L., Liu, J., Du, B., and Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint*.