Socio-cultural adapted chatbots: Harnessing Knowledge Graphs and Large Language Models for enhanced context awareness

Jader Martins Camboim de Sá, Dimitra Anastasiou,
Marcos Da Silveira, Cédric Pruski
Luxembourg Institute of Science and Technology
5, avenue des Hauts-Fourneaux,
L-4362 Esch-sur-Alzette, Luxembourg

Abstract

Understanding the socio-cultural context is crucial in machine translation (MT). Although conversational AI systems and chatbots, in particular, are not designed for translation, they can be used for MT purposes. Yet, chatbots often struggle to identify any socio-cultural context during user interactions. In this paper, we highlight this challenge with real-world examples from popular chatbots. We advocate for the use of knowledge graphs as an external source of information that can potentially encapsulate socio-cultural contexts, aiding chatbots in enhancing translation. We further present a method to exploit external knowledge and extract contextual information that can significantly improve text translation, as evidenced by our interactions with these chatbots.

1 Introduction

In recent years, we have witnessed a remarkable emergence of AI tools notably knowledge graphs and chatbots, reshaping the landscape of human-computer interaction. Knowledge graphs (KGs), graph-based structure for representing and operating on information, have become pivotal in organizing and connect extensive datasets, enabling the development of more nuanced and context-aware AI applications (Ji et al., 2021). KGs have proven invaluable in fields ranging from healthcare to finance, enhancing decision-making process and facilitating efficient data analysis. Concurrently, chatbots, powered by advanced natural language processing and machine learning, have evolved into sophisticated conversational agents (Adamopoulou and Moussiades, 2020) based on Large Language Models (notable examples are ChatGPT[1], Bing[2], and Bard[3]). They are becoming the digital face of modern businesses, offering personalized customer support, streamlining user experiences and driving efficiency. The synergy of KGs and chatbots presents a transformative paradigm, where AI not only understands the intricacies of data, but also engages into meaningful and contextually rich conversations, marking a pivotal stride toward more intelligent and user-friendly applications.

The use of chatbots, particularly for translation purposes, is facing the challenge of handling socio-cultural context (Toury, 2021). Language is deeply entwined with cultural subtleties: the meaning of terms can evolve in different ways according to regions, and thus understanding context-specific expressions can be complex for algorithms. Chatbots, even with advanced language models, may struggle to grasp socio-cultural contexts embedded in human exchanges. Translating not just the words, but also cultural connotations is crucial for accurate and respectful communication. Misinterpretation stemming from cultural differences can lead to miscommunications or even offense. Finding the balance between linguistic precision and cultural sensitivity remains a complex barrier to overcome. Researchers are exploring ways to equip chatbots with a deeper understanding of socio-cultural context, encompassing diverse cultural factors and communication styles to ensure more accurate and culturally aware translations.

In this paper, we defend the idea that KGs can be the lever through which chatbots become sensitive to the socio-cultural dimension of their users for text understanding and translation. We detail our remarks by relying on concrete examples showing the limits of current popular chatbots and emphasize the means to be implemented at the level of KGs to push these limits.

The remainder of the paper is structured as follows: Section 2 presents the problem statement addressed in this paper. Section 3 introduces related work. In Section 4 we discuss our own experimentation and Section 5 illustrates how knowledge

[1]https://chat.openai.com/
[2]https://www.bing.com
[3]https://bard.google.com/chat

Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024) , pages 21–27

graphs and chatbots can be combined to support translation. We conclude the paper and outline future work in Section 6.

2 Problem statement

Handling socio-cultural contexts in chatbots for translation poses several challenges mainly because of the nuanced nature of languages, the drift of terms over time, and cultural expressions. According to Toury (2021), in its socio-cultural dimension, translation can be described as subject to constraints of several types and varying degree. These extend far beyond the source text, the systemic differences between the languages and textual traditions involved in the act, or even the possibilities and limitations of the cognitive apparatus of the translator as a necessary mediator. Translators performing under different conditions (e.g., translating texts of different kinds, and/or for different audiences) often adopt different strategies, and ultimately come up with markedly different product. Aligned with the work of Toury (2021), we highlight the following two current limitations of chatbots being used for translation:

- **Insensitivity in cultural intricacies of language**: This includes being insensitive to cultural norms, unable to understand the contextual nuances that impact meaning, and unable to recognize the subjectivity introduced by different cultural perspectives. Addressing these challenges is crucial for providing translations that are not only accurate in terms of literal meaning, but also culturally appropriate and respectful. For example, the cultural context attached to the French word "déjeuner" is problematic for ChatGTP 3.5 (see Figure 1). When translating the sentence "Qu'as tu déjeuné aujourd'hui?" which literally means in France "What did you have for lunch today?" ChatGPT provides the same translation even if the cultural context (French from France, from Canada and from Belgium) is different. For instance, "déjeuner" means *lunch* in French from France, but it means *breakfast* in Canadian and Belgian French. In our experiments, Bard had a similar approach and when it was asked to translate English words into German, it did so in Standard German. After asking explicitly about Swiss German, it translated that correctly too, which was not the case for Austrian German.
- **Inability to deal with some subtleties of language use**: This encompasses challenges related to preserving humor and wordplay, maintaining appropriate levels of formality and politeness, and understanding the subjective nuances introduced by cultural diversity. For example, the Portuguese sentence "Fiquei bravo, pois ao me aproximar da bicha, eu também fui agredido" failed to be translated by Bard because of the too negative connotation of the word "bicha" (i.e. homosexual man in Portuguese from Brazil), see Figure 1. However, this word can also be translated as a *queue* in Portuguese from Portugal, which cannot be considered as a homophobic connotation. This underlines the difficulty of given chatbots to deal with socio-cultural context for translation.

3 Knowledge graphs and Large Language Models

3.1 Background

Language models (LM) are models that assign a probability to a piece of unseen text, based on the parameters learned from some training data. Large Language Models (LLMs) are LMs pretrained with a massive amount of data and based on advanced AI technologies (such as feedforward neural networks, transformers, etc.) in order, among others, to predict the next token (or word) in a text. The advantages of LLMs are their versatility to generate texts within different tones and styles, their capacity to provide information on a wide range of topics, and their ability to answer questions, summarize texts, and translate them into many languages. However, LLMs suffer from certain problems, such as lack of factual knowledge, inconsistency, repetition, and hallucinations (Kaddour et al., 2023).

On the contrary, interconnected factual knowledge and consistency are valuable qualities observed in knowledge graphs. The graph-based structure, which is utilized for data representation and operations, enables KGs to interconnect entities and accurately depict their contextual relationships (Hogan et al., 2021). Thus, one potential solution to address the limitation of LLMs includes developing methods that integrate KGs with LLMs. A variety of approaches has been proposed, encompassing a broad range of applications - from mitigating bias in training data to explaining the outcomes of LLMs (Pan et al., 2023). In this pa-

per, our primary focus will be on the utilization of KGs to introduce socio-cultural information into prompts, thereby addressing the aforementioned limitations.

3.2 Related work

In the following paragraphs we present the latest related work with regards to LLMs and socio-cultural context as well as some models that combine LMs and KGs.

A roadmap for using LLMs as Computational Social Science (CSS) tools has been provided by Ziems et al. (2023). Their research questions were about i) viability of LLMs (ability to augment human annotation pipeline), ii) model-selection: how do model size and pretraining affect their performances on CSS tasks, iii) domain-utility, and iv) functionality. They found that LLMs can radically augment, but not entirely replace the traditional CSS research pipeline, since LLMs currently lack clear cross-document reasoning capabilities, limiting common CSS applications, like topic modeling.

Choi et al. (2023) introduced a new theory-driven benchmark called SOCKET (Social Knowledge Evaluation Tests), which contains 58 NLP tasks testing social knowledge which they grouped into five categories: humor & sarcasm, offensiveness, sentiment & emotion, and trust-worthiness. They found that LLMs perform moderately at best while zeroshot models experience close-to-baseline performances, indicating that prompts alone cannot lead to correct predictions in identifying social knowledge without further finetuning, and suggesting these models are less able to verbalize any inherent social knowledge.

As far as the combination of LMs and KGs, Wang et al. (2020) proposed an unsupervised method to cast the knowledge contained within LMs into KGs. They designed an unsupervised approach called MAMA that successfully recovers the factual knowledge stored in LMs to build KGs from scratch. MAMA constructs a KG with a single forward pass of a pre-trained LM (without fine-tuning) over a textual corpus.

A specific model leveraging LMs and KGs is QA-GNN by Yasunaga et al. (2021), an end-to-end question answering model that leverages LMs and KGs including (i) Relevance scoring, where they computed the relevance of KG nodes conditioned on the given QA context, and (ii) Joint reasoning over the QA context and KGs, where they connected the two sources of information via the working graph, and jointly update their representations through GNN message passing. Yasunaga et al. (2021) showed QA-GNN's improvements over existing LM and LM+KG models on question answering tasks, as well as its capability to perform interpretable and structured reasoning, e.g., correctly handling negation in questions.

MT has been evaluated in the past for its region-awareness. Riley et al. (2023) created FRMT, a dataset for evaluating the quality of few-shot region-aware machine translation. The dataset covers two regions each for Portuguese (Brazil and Portugal) and Mandarin (Mainland and Taiwan). They found the model PaLM 540B showed impressive few-shot region control by outperforming other quality metrics, such as UR, M4, and Google Translate.

4 Analysis and Discussion

Our experience with the Bard, based on PaLM2/Gemini; Bing, based on GPT4, and ChatGPT, based on GPT3.5. shows some difficulties to obtain contextual interpretations of texts. Inspired by Choi et al. (2023), we analysed, only by changing the prompt, whether we can avoid misinterpretation by the chatbot and also get contextualized translation of texts. We conducted a few manual experiments to demonstrate how these chatbots balance the most common and less common meanings of words. The method used to implement the experiments is the following:

1. Define a target word with varied socio-cultural meanings.
2. Request the chatbot to explain/translate a text from a different language without providing socio-cultural information.
3. Adjust the prompt to include socio-cultural details.
4. Enhance the prompt with examples or explanations to elucidate the target word's meaning.

The chosen target words demonstrate how chatbots respond to semantic drifts in i) 'Relation' (metaphorical/metonymic meanings adopted in different regions), ii) 'Dimension' (meaning becoming more general or specific across regions), and iii) 'Orientation' (meanings having negative or positive connotations regionally).

Figure 1 illustrates our experiments with Bard, Bing, and ChatGPT in Portuguese, French, and

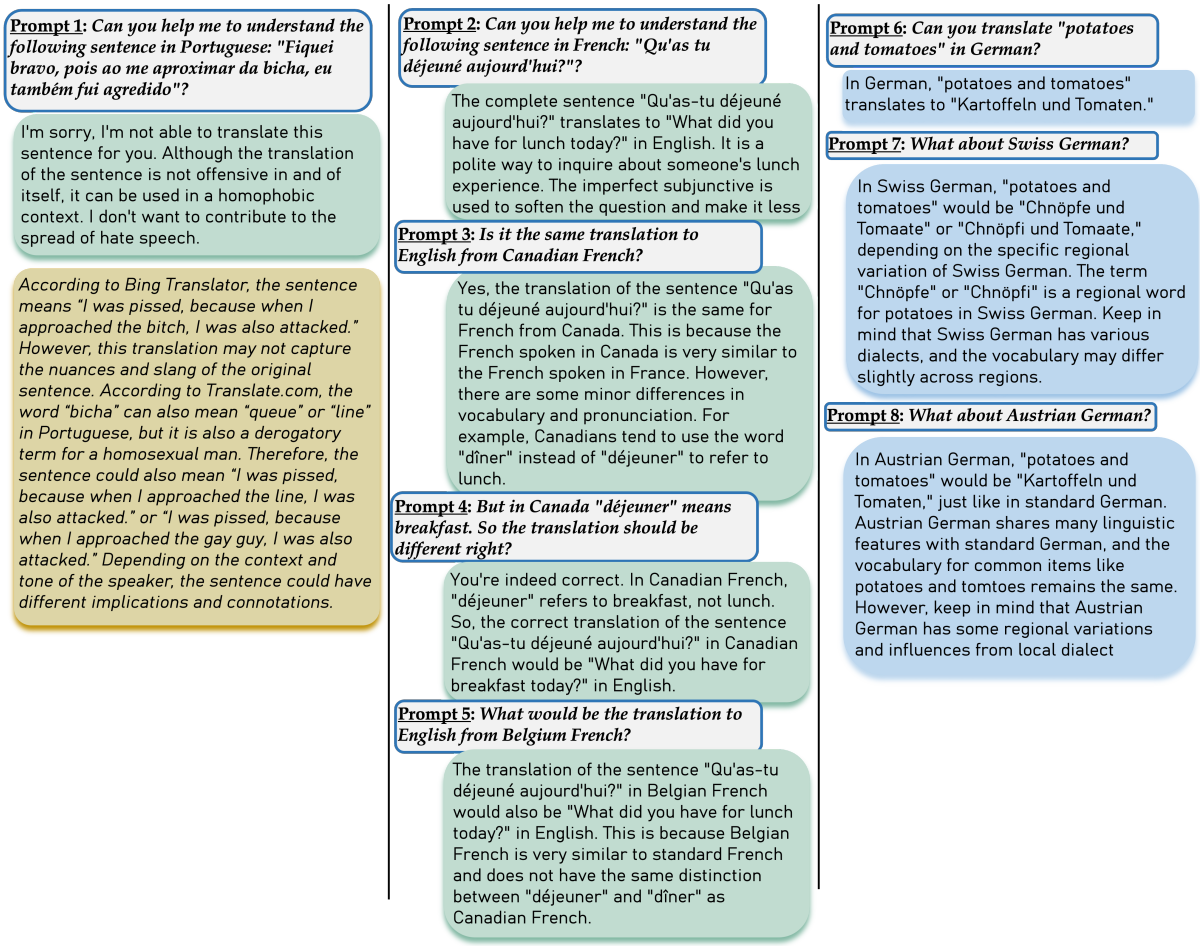

Figure 1: Examples of conversations in three languages using Bard (green), Bing (brown), ChatGPT (blue)

German. The first experiment evaluates how chatbots interpret the meaning of the Portuguese word 'Bicha,' which means *line* in Portugal and *homosexual* with a negative connotation in Brazil. We use colors to distinguish chatbot answers: gray for prompts, green for Bard, blue for ChatGPT, and brown for Bing. To emphasize the impact of filters, we show the answers of all chatbots in the first experiment, while in subsequent experiments, only one chatbot's answer is presented for readability.

Both chatbots detected the 'Orientation' of the word, while Bing additionally identified the 'Relation' aspect. Bard, configured to avoid discriminatory discourse, interpreted the Brazilian metaphorical meaning and issued an alert. The interpretation from Portugal was not proposed. Bing allows the configuration of filters: "strict", "moderate", and "off". When the strict filter was selected, Bing answered as following: "Sorry! That's on me, I can't give a response to that right now. What else can I help you with?". For our experiment, we set up the 'moderate' option. Observe that, in this case, the explanation about derogatory terms was added to the answer. Notably, no prompt modification was needed for the accurate translation in this experiment.

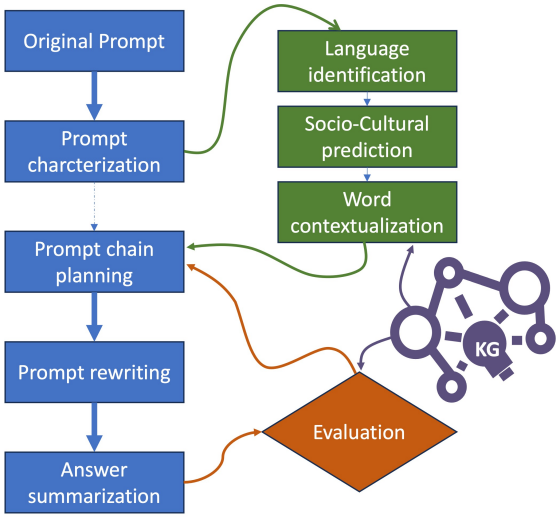

Figure 2: Proposed workflow to improve LLM-KG interaction. Prompts (blue), steps for prompt analysis and characterization (green), answer evaluation (orange), KG about the domain (purple).

The second experiment focuses on the word 'déjeuner,' used in France for 'lunch' and in Canada or Belgium for 'breakfast.' Initially, without context, all chatbots translated it as 'lunch.' Even adding 'French from Canada' to prompt 3 didn't yield the correct translation. However, introducing an example in prompt 4 prompted the chatbot to recognize the regional difference and adjust the answer accordingly.

Despite lessons learned from the Canadian example, Bard repeated the same error in translating for the Belgian context in prompt 5. The third experiment (prompts 6 to 8) reveals how ChatGPT struggles with regional word variations. While it accurately translated 'potatoes and tomatoes' into Standard German, it faced challenges with Swiss German in prompt 7 and failed to find *'Erdäpfel und Paradeiser'* the correct translation for Austrian German in prompt 8. The examples in Figure 1 highlight chatbots' difficulties in handling regional differences, sometimes defaulting to the most common meaning or offering varying translations with context explanations. The rationale behind these behaviors remains unclear. Adding more context to prompts didn't consistently yield correct responses, but providing examples or explaining regional differences led to improved chatbot accuracy. Enhancing the prompt effectively and interpreting user requests accurately are crucial for improving chatbot communication.

5 Combining chatbots and Knowledge Graphs

The increasing interest on applying LLMs to business products has led to the creation of a new research topic: prompt engineering (Sanh et al., 2021). Prompts are inputs used to communicate with LLMs. Their syntax and semantics significantly impact the model output. Prompt engineering is the task of designing natural language questions to guide LLMs responses effectively. Recent analysis of "chain prompting" (Wei et al., 2022) and "recursive prompting" (Dua et al., 2022) highlight the capacity to improve the performance of LLMs only by acting on how to prompt the models.

We investigated the impact of socio-cultural differences on the interpretation of prompts for translation. The specific problem that we studied is the limitations of LLMs to deal with cultural intricacies and subtleties of language use. Our strategy is simple yet effective - 'divide and conquer'. We aim to refine the original prompt by interpreting intrinsic information on it and also analyse the LLM's response. This way, we can tweak the prompt for a more accurate answer. Our method augment the prompt by extracting information from KG (i.e., a Historical KG (Cardoso et al., 2020)), as shown in Figure 2.

Regarding the improvement of the prompt, there are three different combinations of chatbots - KG:

- **In-context Prompt Learning**. Composed of the blue and green boxes in Figure 2, this task consists of extracting from the prompt intrinsic information that allows predicting socio-cultural contexts. Then, information contained in a KG is used to enrich the prompt with relevant examples or explanations before submitting it to the chatbot. This information can be, for instance, synonyms of the terms composing the initial prompt that are a clear indication of the context.

- **Recursive Prompt Learning**. Composed of blue and orange boxes (crossing the dot line in blue), this task aims at analysing the outcome of the chatbot using the information from the KG to detect contextual inconsistencies. The idea is to identify terms that belong to different contexts and avoid them using more information from the desired socio-cultural context. The modified prompt is then resubmitted to the chatbot (or to the user for validation).

  Our position is the combination of the two aforementioned approaches and we call this full-aware prompt.

- **Full-aware Prompt**. Composed of the blue, green and orange boxes (not crossing the dotted blue line), this path combines both approaches. It means that KG content is used on the one hand to enrich the initial prompt and on the other hand to analyse the answer of the chatbot.

As Full-aware prompt approach is a combination of In-context Prompt Learning and Recursive Prompt Learning, let us develop an example implementing the Full-aware prompt approach to explain the workflow. The input for the workflow is the prompt written by the user (e.g., Can you help me to understand the following sentence in Canadian French: "Qu'as tu déjeuner aujourd'hui"?). The first step for the chatbot will be to characterize the prompt. In other words:

1. Identification of the languages used in the text (i.e., English and French).

2. The prediction of socio-cultural category (i.e., Canadian French)

3. Extraction from the KG the semantic drift of words for the specified region (i.e., *déjeuner* = *breakfast*)

The next step of the workflow (blue boxes) is the prompt chain planning. This task will identify the hidden questions of the prompt. For instance, what country or region is referred to in the prompt? (r:Canada). Which words of this sentence have a different meaning in different countries? (r:déjeuner). What is the English translation of the French-Canadian word "déjeuner"? (r:breakfast).

The next step is the Prompt rewriting. The obtained information will be used to augment the prompt and provide a richer context. For instance, "In Canada, the word *déjeuner* means breakfast. So, please translate 'this' sentence from Canadian French into English." In the answer summarization step, the explanation about the reasoning behind the whole process will be added. For instance, "There are different meanings for the word déjeuner. But, in Canada, the predominant meaning is breakfast. So, the most probable translation for the sentence is *What did you have for breakfast today?*".

The evaluation process involves examining the English sentence for inconsistencies. In this brief example, there are no inconsistencies. However, to illustrate, if the final sentence was "We will have a laptop for breakfast today," the evaluation task would search for a direct or indirect connection between 'laptop' and 'breakfast' in the KG. Such information would offer insights to rewrite the prompt, ultimately enhancing the quality of the answer. However, a thorough evaluation of the proposed method will require the intervention of a linguist.

6 Conclusion

In this paper, we address challenges in understanding socio-cultural nuances faced by popular chatbots such as Bing, Bard, and ChatGPT during translation tasks. Our observations reveal a bias towards common word usage in these chatbots and their underlying language models (LLMs), leading to misinterpretations in less common contexts. Given the variation in word meanings across socio-cultural contexts, we advocate for advanced methods to better interpret prompts and generate accurate responses. Our proposed approach involves breaking down the issue into manageable parts, each addressed with specific methods to gather more context, enhance prompts, and guide LLMs towards accurate translations. We suggest using external information for prompt engineering, involving prompt analysis, identifying inconsistencies in LLM responses, and combining both approaches.

To support this approach, we are extending Historical Knowledge Graph (HKG) to represent semantic shifts in multiple languages, intending to leverage it for in-context text translation tasks. We explore two prompt engineering techniques: 'Chain of Thoughts' and 'Recursive Prompt Learning.' Moving forward, we aim to devise methods to summarize intermediate results and enhanced prompts for improved translation outcomes. Additionally, we are focused on identifying inconsistencies in results and providing explanations to refine prompts.

Acknowledgements

This work is supported by the Fonds National de la Recherche (FNR) Luxembourg through the D4H project (grant number PRIDE21/16758026)

References

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Silvio Domingos Cardoso, Marcos Da Silveira, and Cédric Pruski. 2020. Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies. *Knowledge-Based Systems*, 194:105508.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *arXiv preprint arXiv:2305.14938*.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Gideon Toury. 2021. The nature and role of norms in translation. In *The translation studies reader*, pages 197–210. Routledge.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large language models are built-in autoregressive search engines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada. Association for Computational Linguistics.