Enhancing Machine Translation Experiences with Multilingual Knowledge Graphs

Simone Conia[1], Daniel Lee[2], Min Li[3], Umar Farooq Minhas[3], Yunyao Li[4]

[1]Sapienza University of Rome, Italy
[2]University of Calgary, Canada
[3]Apple
[4]Adobe
simone.conia@uniroma1.it, daniel.lee1@ucalgary.ca, min_li6@apple.com, ufminhas@apple.com, yunyaol@adobe.com

Abstract

Translating entity names, especially when a *literal translation* is not correct, poses a significant challenge. Although Machine Translation (MT) systems have achieved impressive results, they still struggle to translate cultural nuances and language-specific context. In this work, we show that the integration of multilingual knowledge graphs into MT systems can address this problem and bring two significant benefits: i) improving the translation of utterances that contain entities by leveraging their human-curated aliases from a multilingual knowledge graph, and, ii) increasing the interpretability of the translation process by providing the user with information from the knowledge graph.

Introduction and Related Work

Over the years, researchers in MT have steadily faced numerous challenges and have consequently presented new systems that are not only increasingly robust and fluent but also support a growing number of languages (Tang et al. 2021; Fan et al. 2021; Costa-jussà et al. 2022). However, there are several challenges that current approaches have yet to overcome; one of them is translating text that contain entity names (Wan et al. 2022). Entity names can be challenging, as their translation in the target language may not be a *literal translation* from the source language but a *transcreation* (Díaz-Millón and Olvera-Lobo 2023), which requires an adaptation to language- and cultural-specific aspects to maintain the original intent, style, tone, and context. It can be the case for movies ($\textit{Moana}_{\text{EN}} \rightarrow \textit{Vaiana}_{\text{FR}}$), books ($\textit{The Catcher in the Rye}_{\text{EN}} \rightarrow \textit{Il Giovane Holden}_{\text{IT}}$), people ($\textit{Copernicus}_{\text{EN}} \rightarrow \textit{Kopernikus}_{\text{DE}}$), and also animals, sports, and TV series, among other entity types, whose names may be significantly different across languages because of cultural, geographical, and socioeconomic factors (Bai 2018).

A common approach to mitigate the issue of entity name translation is augmenting the training dataset to cover as many entities as possible, allowing an MT system to memorize entity name translations at training time (Hu et al. 2022). However, data augmentation for entity name translation comes with noteworthy downsides: i) the size of the augmented dataset may become infeasibly large, as there may be millions of entities and hundreds of language pairs to take into account; and, ii) new entities (e.g., movies) appear every day and existing entities may need to be updated (e.g., change of name), requiring re-building the augmented training set and re-training the MT system periodically. Therefore, we argue that re-training or fine-tuning on augmented datasets for entity name translation is a sub-optimal strategy to update the "memory" of an MT system. On the contrary, updating a knowledge graph is usually easy and quick, as it can be done by a non-expert human in a matter of minutes.

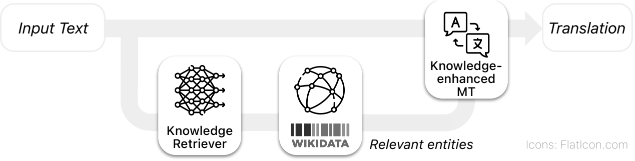


Figure 1: Wiki-MT is composed of two main components: i) a *knowledge retriever*, which collects relevant information from Wikidata, and, ii) a *knowledge-enhanced machine translator*, which leverages the retrieved knowledge to produce better translations for inputs that contain entity names.

System Overview

To address the above-mentioned issues, we introduce Wiki-MT (Wikidata-enhanced MT), a demo that showcases how Wikidata (Vrandečić 2012) – one of the most popular multilingual knowledge graphs – can be integrated into an MT system. This integration brings two significant benefits to state-of-the-art MT systems: i) instead of memorizing name translations across languages during the training process, Wiki-MT can query Wikidata to retrieve human-curated entity names and their translations to provide high-quality translations of a text that contain entity names; ii) the information retrieved from Wikidata can be displayed to users, allowing them to better understand the translation.

Wiki-MT features two main components: (1) a *knowledge retriever*, which retrieves the most relevant information about an input query q from Wikidata; and (2) a *knowledge-enhanced machine translator*, which is an MT system adapted to use the retrieved knowledge to better translate the entity names appearing in q.

Knowledge Retriever. Given an input query q that we want to translate, the main goal of the knowledge retriever is to collect from Wikidata the top-n entities $E = \{e_1, \ldots, e_n\}$ that are most relevant to q. Our knowledge retriever performs this operation by computing the vector representation $\mathbf{v}_q$ of q and retrieving the top-3 most-relevant entities from Wikidata, i.e., those entities whose vector representation $\mathbf{v}_e$ produces the highest cosine similarity with $\mathbf{v}_q$ among all the entities in Wikidata. The knowledge retriever computes the vector representation $\mathbf{v}_e$ of an entity e by taking into account its Wikidata name and description. For example, for the entity *Q183883*, the knowledge retriever obtains the entity representation $\mathbf{v}_{Q183883}$ by encoding the text "*The Catcher in the Rye: novel by J.D. Salinger*".

Knowledge-Enhanced Machine Translator. Given a source language l_s, a target language l_t, and an input query q that we want to translate from l_s to l_t, our machine translation component also takes in input the set $E = \{e_1, e_2, e_3\}$ of the top-3 entities that are most related to q, according to the knowledge retriever. For each entity e_i in E, we obtain from Wikidata its name in l_s (e.g., *The Catcher in the Rye* for *Q183883* in English), and its translation in l_t (e.g., *Il Giovane Holden*[1] in Italian). This information is appended to the original query q to obtain a knowledge-augmented query q':

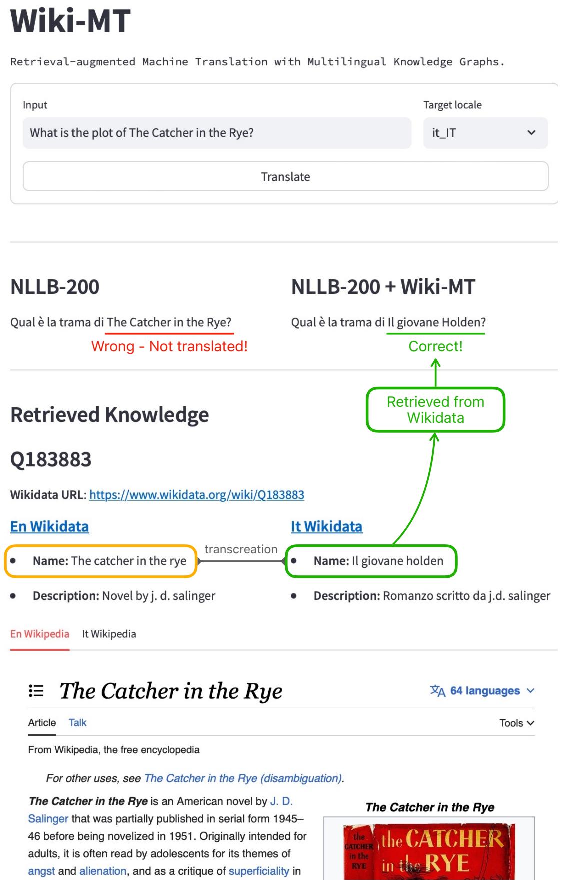
$$\begin{aligned} q' = q \,&\oplus \text{[META] } \mathrm{name}(e_1, l_s) \text{ [AS] } \mathrm{name}(e_1, l_t) \\ &\oplus \text{[META] } \mathrm{name}(e_2, l_s) \text{ [AS] } \mathrm{name}(e_2, l_t) \\ &\oplus \text{[META] } \mathrm{name}(e_3, l_s) \text{ [AS] } \mathrm{name}(e_3, l_t) \end{aligned}$$

where $\oplus$ is the append operator, and [META] and [AS] are special tokens used to separate entities and name translations, respectively. Since we add new special tokens, the machine translator needs to be fine-tuned on a dataset.

Implementation details. Wiki-MT is implemented using PyTorch, PyTorch Lightning, and HuggingFace Transformers. More specifically, the knowledge retriever is built on top of multilingual Contriever (Izacard et al. 2021), a multilingual retriever trained using contrastive learning. The finetune the knowledge retriever on a Wikidata dump from May 2023 in 11 languages: Arabic, English, French, German, Italian, Japanese, Korean, Spanish, Thai, Traditional Chinese, and Turkish. As for the underlying MT system in Wiki-MT, we adopt NLLB-200 (Costa-jussà et al. 2022), a many-to-many multilingual MT model. Finally, we fine-tune Wiki-MT on Mintaka (Sen, Aji, and Saffari 2022), which provides translations of knowledge-seeking queries. While Wiki-MT supports 11 languages, it can be expanded to all those supported by NLLB-200 and Wikidata.

Wiki-MT's UI. The UI is divided into three main parts:

- **Top – user input**, where users can type an English query q to translate, and select the target language l_t.
- **Middle – system comparison**, where users can compare the translations produced by Wiki-MT with and without the information from the knowledge retriever. This side-by-side comparison allows users to directly assess the effect of using name translations in the MT component.



Wiki-MT

Retrieval-augmented Machine Translation with Multilingual Knowledge Graphs.

Input: What is the plot of The Catcher in the Rye?

Target locale: it_IT

NLLB-200

Qual è la trama di The Catcher in the Rye?

NLLB-200 + Wiki-MT

Qual è la trama di Il giovane Holden?

Retrieved Knowledge

Q183883

Wikidata URL: https://www.wikidata.org/wiki/Q183883

En Wikidata

- **Name:** The catcher in the rye
- **Description:** Novel by j. d. salinger

It Wikidata

- **Name:** Il giovane holden
- **Description:** Romanzo scritto da j.d. salinger

The Catcher in the Rye

From Wikipedia, the free encyclopedia

For other uses, see The Catcher in the Rye (disambiguation).

The Catcher in the Rye is an American novel by J. D. Salinger that was partially published in serial form 1945–46 before being novelized in 1951. Originally intended for adults, it is often read by adolescents for its themes of angst and alienation, and as a critique of superficiality in

The Catcher in the Rye

Figure 2: User interface of Wiki-MT: users can translate text, visualize the effect of using the knowledge retrieved from Wikidata, and learn more about the Wikidata entities collected by the knowledge retriever to gain deeper insights into the translation process.

- **Bottom – retrieved knowledge**, where the user can read about the entities fetched by the knowledge retriever. This allows the user to gain insights into the translation process, e.g., whether Wiki-MT was able to correctly recognize the entities E that are relevant for a given q and what information Wikidata provides about E.

Conclusion and Future Work

Wiki-MT showcases the integration of multilingual knowledge graphs into MT systems to enhance user experiences from two perspectives: i) generating higher-quality translations of texts that contain entity names; and, ii) increasing the interpretability of the translation process, as users can directly see which entities have been recognized by Wiki-MT and learn more about them. Wiki-MT hints at promising directions to improve existing MT models, where future work may explore i) a deeper integration of retrievers into MT systems to include richer semantics in the translation process, and ii) improving the quality of the textual information in multilingual knowledge graphs (Conia et al. 2023).

[1] The literal translation in Italian of this book means *The Young Holden*, i.e., it is completely different from *The Catcher in the Rye*.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback, Saloni Potdar for her engaging conversations, Ihab Ilyas for his initial comments, and many other people at Apple for their invaluable help. The majority of this work has been carried out while Simone Conia and Daniel Lee were interns at Apple. Simone Conia gratefully acknowledges the PNRR MUR project PE0000013-FAIR, which fully funds his fellowship since September 2023.

References

Bai, Z. 2018. On Translation Strategies of English Movie Titles. *Journal of Language Teaching & Research*, 9(1).

Conia, S.; Li, M.; Lee, D.; Minhas, U. F.; Ilyas, I.; and Li, Y. 2023. Increasing Coverage and Precision of Textual Information in Multilingual Knowledge Graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. Singapore: Association for Computational Linguistics.

Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Díaz-Millón, M.; and Olvera-Lobo, M. D. 2023. Towards a definition of transcreation: a systematic literature review. *Perspectives*, 31(2): 347–364.

Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; Goyal, N.; Birch, T.; Liptchinsky, V.; Edunov, S.; Grave, E.; Auli, M.; and Joulin, A. 2021. Beyond English-Centric Multilingual Machine Translation. 22(1).

Hu, J.; Hayashi, H.; Cho, K.; and Neubig, G. 2022. DEEP: DEnoising Entity Pre-training for Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1753–1766. Dublin, Ireland: Association for Computational Linguistics.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning.

Sen, P.; Aji, A. F.; and Saffari, A. 2022. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1604–1619. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Tang, Y.; Tran, C.; Li, X.; Chen, P.-J.; Goyal, N.; Chaudhary, V.; Gu, J.; and Fan, A. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3450–3466. Online: Association for Computational Linguistics.

Vrandečić, D. 2012. Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, 1063–1064. New York, NY, USA: Association for Computing Machinery. ISBN 9781450312301.

Wan, Y.; Yang, B.; Wong, D. F.; Chao, L. S.; Yao, L.; Zhang, H.; and Chen, B. 2022. Challenges of Neural Machine Translation for Short Texts. *Computational Linguistics*, 48(2): 321–342.