Enhancing Machine Translation Experiences with Multilingual Knowledge Graphs

Simone Conia¹, Daniel Lee², Min Li³, Umar Farooq Minhas³, Yunyao Li⁴
¹Sapienza University of Rome, Italy
²University of Calgary, Canada
³Apple
⁴Adobe

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 contains entity names; ii) the information retrieved from Wikidata can be displayed to users, translations of a text that contain entity names; ii) the information retrieved from Wikidata can be displayed to users, translations of a text that contain entity names; ii) the information retrieved from Wikidata can be displayed to users, translations of a text that contain entity names.

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 $v_q$ of $q$ and retrieving the top-3 most-relevant entities from Wikidata, i.e., those entities whose vector representation $v_e$ produces the highest cosine similarity with $v_q$ among all the entities in Wikidata. The knowledge retriever computes the vector representation $v_e$ 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 $v_{Q183883}$ 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 Holden1 in Italian). This information is appended to the original query $q$ to obtain a knowledge-augmented query $q'$:

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

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 fine-tune 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.

1The literal translation in Italian of this book means *The Young Holden*, i.e., it is completely different from *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.](image)

- **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).
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