On the Inter-Linguistic Disparity of Knowledge Graphs: Bridging the Gap between English and Non-English Languages

Simone Conia
Sapienza University of Rome
simone.conia@uniroma1.it

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

The exponential growth of Large Language Models (LLMs) has taken the world of Natural Language Processing (NLP) by storm. Despite their capability of generating impressively fluent text, LLMs still exhibit gaps in producing factual, coherent, and relevant outputs in complex scenarios (Augenstein et al., 2023). To mitigate this issue, the research community has introduced techniques for grounding LLMs in knowledge graphs (KGs), where each node usually represents a concept (e.g., *universe*, *weather*, or *president*) or a named entity (e.g., *Albert Einstein*, *Rome*, or *The Legend of Zelda*), and each edge between two nodes represents a fact (e.g., "*Rome* is the capital of *Italy*" or "*The Legend of Zelda* is a *video game series*"). The synergy between KGs and LLMs has become successful for two main reasons: their relational information, i.e., the links between nodes, and their textual information, i.e., the lexicalizations of the concepts and entities.

However, despite the advances in grounding language models to KGs (Schneider et al., 2022), these efforts are significantly limited by the stark discrepancy between English and non-English textual information (Kaffee et al., 2023). This discrepancy manifests on two fronts: a disparity in *coverage*, where non-English languages are limited in the number of entities for which at least one lexicalization is provided, and a disparity in *precision*, as the quality of non-English textual information is usually lower. This gap in data coverage and precision severely limits the applicability of recent approaches to multilingual applications (Peng et al., 2023).

In this paper, we present our contributions aimed at addressing the problems of coverage and precision of textual information in multilingual KGs (Conia et al., 2023), with a particular focus on Wikidata, one of the most widely used KGs in the NLP community. More specifically, we analyze the gap between English and non-English languages in the entity names in Wikidata, describe a novel benchmark for evaluating automatic systems on narrowing this gap, present a methodology to mitigate the above-mentioned issues, and discuss how our efforts can improve a set of downstream applications. With this discussion, we aim to stimulate conversations on novel research directions and foster future research on bridging the gap between English and non-English languages in KGs.

2 Measuring the Inter-Linguistic Gap in Multilingual Knowledge Graphs

While relational information in KGs is usually language-agnostic (e.g., "AI" is a field of "Computer Science" independently of the language we consider), textual information is usually language-dependent (e.g., the lexicalizations of "AI" and "Computer Science" vary across languages). With the growing number of languages supported by KGs, it becomes increasingly challenging for human editors to maintain their content up-to-date in all languages (Kaffee et al., 2019). Therefore, it is important to invest in developing and evaluating systems that can support humans in curating textual information across languages.

Coverage. Ideally, we would like every entity in Wikidata to be "covered" in all languages, i.e., we would like Wikidata to provide at least one entity name for each language supported by the knowledge graph. However, this is not currently the case for Wikidata, as we can observe in Figure 1, which provides an overview on the availability of entity names in nine non-English languages.

Precision. While achieving inter-linguistic parity in terms of coverage of entity names is fundamental, another crucial aspect is the accuracy of such information. It is not uncommon to find inaccurate textual information in Wikidata, including human mistakes (e.g., spelling errors), outdated entries (e.g., name changes), and under-specific information (e.g., generic job roles, such as "musician" instead of "pianist").

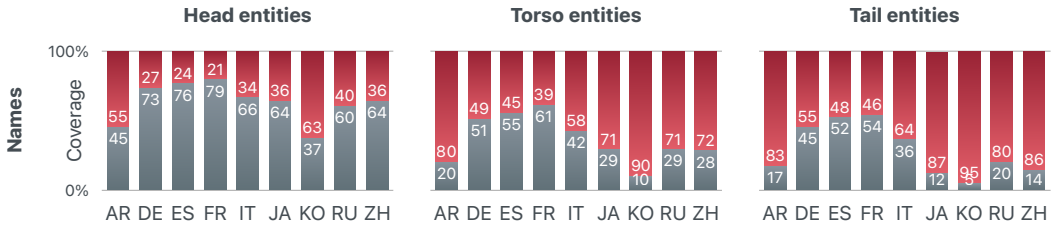

Figure 1: Coverage of non-English entity names compared to English in Wikidata. Best seen in color.

	AR	DE	EN	ES	FR	IT	JA	KO	RU	ZH	All
Entities	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	10,000
Entity names in WikiKGE-10	4,213	3,498	2,837	4,320	3,548	3,156	2,999	3,874	3,901	4,088	36,434
- Entity names in Wikidata	2,521	2,336	2,090	2,732	2,330	1,840	2,235	2,136	2,706	2,569	23,495
- Entity name errors in Wikidata	320	491	219	571	530	236	486	329	507	830	4,663

Table 1: Overview of WikiKGE-10, which features 10 languages – Arabic (AR), German (DE), English (EN), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Russian (RU), simplified Chinese (ZH).

WikiKGE-10. To assess the severity of the above-mentioned issues, we created WikiKGE-10, a novel benchmark to evaluate systems for improving the quality of textual information in multilingual KGs. WikiKGE-10 covers 10 diverse languages – English, Arabic, German, Italian, French, Spanish, Korean, Japanese, Russian, and Chinese – and provides more than 35 thousand manually-graded entity names for 1000 entities in each of the 10 languages, as shown in Table 1. The creation of WikiKGE-10[1] is especially relevant for UniDive in the context of WG1 for corpus annotation of complex multilingual datasets and WG4 for promoting truly inter-linguistic resources, and future work may aim at extending and adapting this methodology for low-resource languages.

3 Bridging the Inter-Linguistic Gap in Multilingual Knowledge Graphs

A key objective of our work is to also evaluate the capability of three broad categories of approaches that can be applied to bridge the inter-linguistic gap in multilingual KGs, namely, machine translation (MT), web search (WS), and LLMs. These three categories are often strong candidates for the development of multilingual technology tools, which is the scope of WG3. In our experiments, we observe that MT with NLLB-200 (Costa-jussà et al., 2022) (41% in coverage; 58% in precision), WS with Google Web Search (28% in coverage; 45% in precision), and LLM prompting with GPT-4 (42% in coverage; 58% in precision) struggle to generate high-quality entity names in the ten languages of WikiKGE-10. To address this issue, we also introduce M-NTA (Multi-source Naturalization, Translation, and Alignment), a novel method for combining the predictions from MT, WS, and LLMs. M-NTA shows promising gains (53.9% in coverage; 80.1% in precision), demonstrating that combining knowledge across languages is essential to improve the quality of language resources.

4 Impact on Downstream Applications

Finally, our endeavors also include an investigation on the impact of improving the textual information provided by KGs on a set of three downstream tasks, namely, multilingual Entity Linking (MEL), multilingual Knowledge Graph Completion (MKGC), and multilingual Knowledge-Graph Question Answering (mKGQA). For each of these tasks, we measure the performance of a system in two settings: i) using the original textual data (entity names and descriptions) from Wikidata; ii) using the textual data from Wikidata enhanced with M-NTA (see Section 3). Our experiments show that increasing the quantity and quality of textual information in a multilingual KG improves the performance of mGENRE (De Cao et al., 2022) for MEL by 1.2% points in F1 score and AlignKGC (Chakrabarti et al., 2022) for MKGC by 1.3% in MRR, while decreasing the number of unanswerable questions by 36.8% in the MKQA benchmark (Longpre et al., 2021) for mKGQA. We provide more details on the experimental results in Conia et al. (2023). We hope our contributions to these three tasks will encourage future studies on the potential impact of KGs in other areas.

[1] https://github.com/apple/ml-kge

Acknowledgements

Simone Conia would like to thank the anonymous reviewers for their feedback, Min Li for her technical and scientific contributions, Daniel Lee for his work on the evaluation, Ihab Ilyas for his thought-provoking discussions and brainstorming sessions, and Yunyao Li for her invaluable mentorship. His thanks also go to Saloni Potdar for her continuous support, Behrang Mohit for his feedback, and many other people at Apple for their invaluable help on this project.

Simone Conia gratefully acknowledges the PNRR MUR project PE0000013-FAIR / PE01-FAIR-SPOKE-5-DIAG CUP B53C22003980006, which currently funds his postdoctoral research fellowship (RTD-A) in full.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain, and Mausam. 2022. Joint completion and alignment of multilingual knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11922–11938, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simone Conia, Min Li, Daniel Lee, Umar Minhas, Ihab Ilyas, and Yunyao Li. 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*, pages 1612–1634, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Lucie-Aimée Kaffee, Russa Biswas, C. Maria Keet, Edlira Kalemi Vakaj, and Gerard de Melo. 2023. Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, 1(1):10:1–10:19.

Lucie-Aimée Kaffee, Kemele M. Endris, and Elena Simperl. 2019. When humans and machines collaborate: Cross-lingual label editing in wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration*, OpenSym '19, New York, NY, USA. Association for Computing Machinery.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, pages 1–32.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.