A Principled Approach for Multilingual Knowledge Graph Completion with Zero Seed Alignment

Group Members: Ashwath Radhachandran, Rakesh Bal, Rustem Aygun and Rahul Kapur

Mentors: Zijie Huang (zijiehuang@cs.ucla.edu) and Roshni Iyer (roshnigiyer@ucla.edu)

Problem & Goal:

The knowledge graph completion (KGC) task is an important research area since most knowledge graphs (KGs) are often incomplete since human annotation is an expensive task. The problem of knowledge graph incompleteness is exacerbated when working with data with a multilingual representation. This is especially seen in languages where human annotations are rare or challenging to acquire. Efforts to narrow language gaps have been started on multilingual KG embedding methods but the lack of seed alignment (mappings between similar concepts in two KGs) prevents effective knowledge transfer. Sometimes the acquisition of alignment labels is a costly and noisy process, but simply using unsupervised methods (correlation or similarity between entities) can fall short of effective language alignment in KGs. We are aiming to address the setting where there is zero seed alignment present (as opposed to there being some pre existing alignment pairs for knowledge transfer), since this is more difficult and more practical for real world situations. And to reiterate, this problem is important because most current KGs are incomplete, and being able to fill in missing facts in these KGs can improve current technology (such as Google's SEOs).

Expected Outcome:

Past research has demonstrated the potential of KGC with limited alignment annotation but this has typically not been in the multilingual context. We address the task of Multilingual KGC (MKGC), which utilizes KGs across multiple languages to achieve a more accurate KG completion task on each individual KG. Unlike previous settings of MKGC which assume some graph alignment pairs are present to perform knowledge transfer, we focus on addressing the setting where there is zero seed alignment present, which is a more challenging task and is practical in many real-world scenarios. Hence, our expected outcome from the project is to implement and test performance of approaches for MKGC with zero seed alignment and hence provide for a more general solution to the MKGC problem.

Dataset:

- DBP-5L - Contains five language-specific KGs from DBpedia i.e. English (EN), French (FR), Spanish (ES), Japanese (JA) and Greek (EL)

Proposed Solutions:

1. Learn alignment pairs in an unsupervised way so that existing techniques can be applied. We can also add some self-supervision signals to guarantee the generation quality of such alignment pairs.

2. We would like to build a model that embeds all KGs in a common space and extract some common-sense knowledge from them so that even without seed alignment, those shared knowledge can facilitate the knowledge reasoning in each single-KG. To embed all KGs in the shared space and in the meantime distinguish them, we can use adversarial training to teach the model how to tell whether two entities are from the same KG or not. We would also utilize both

the textual information for entities and relations, as well as the graph info (triples (h,r,t)) to do the reasoning.

Evaluation Plan:

- Evaluation Protocol - Given each query (e_h, r, ?e_t), compute the plausibility scores f(e_h, r, $\breve{e}_t$) for triples formed by each possible tail entity (ĕt) in the test candidate set and rank them.

- Evaluation Metrics - a) Mean reciprocal ranks (MRR), b) Accuracy (Hits@1), c) The proportion of correct answers ranked within the top 10 (Hits@10)

- Baseline Baselines: a) Monolingual Baselines - TransE, RotatE, DisMult, KG-BERT, b) Multilingual Baselines - KEnS, CG-MuA, AlignKGC and SS-AGA

Timeline:

1. *End of Week 5:* Complete a thorough literature survery in this domain i.e. compile and read a list of relevant papers to understand the existing KG models and some recent multilingual knowledge graph papers. Also, read and understand the literature in unsupervised entity alignment/ multi-modal /cross-lingual tasks without additional alignment.
2. *End of Week 6:* Reproduce SoTA results on the baselines mentioned on the DBP-5L dataset.
3. *End of Week 7:* Design techniques to implement the approach 1 mentioned in the proposed approaches section. Also conduct the experiments on the DBP-5L dataset for this approach.
4. *End of Week 8:* Design techniques to implement the approach 2 mentioned in the proposed approaches section. Also conduct the experiments on the DBP-5L dataset for this approach.
5. *End of Week 9:* Analyze the results and come up with potential solutions to the problems arising in the proposed approaches. Also, implement the improvements on the dataset.
6. *End of Week 10:* Wrap up results, complete report and prepare presentation.

Workload Distribution:

1. Literature Review - Rustem, Rakesh, Rahul and Ashwath
2. Reproducing SoTA on monolingual baselines - Rustem and Rakesh
3. Reproducing SoTA on multilingual baselines - Rahul and Ashwath
4. Design and conduct experiments on approach 1: Rustem and Rakesh
5. Design and conduct experiments on approach 2: Rahul and Ashwath
6. Writing Report and Presentation - Rustem, Rakesh, Rahul and Ashwath

References:

1. Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun and Carlo Zaniolo. Multilingual Knowledge Graph Completion via Ensemble Knowledge Transfer. EMNLP 2021.
2. Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), pages 5429– 5435.
3. Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc. VLDB Endow., page 2326–2340.
4. Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, Wei Wang. Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment (ACL 2022)