Leveraging Multilingual Descriptions for Link Prediction: Initial Experiments

Genet Asefa Gesese[1,2], Mehwish Alam[1,2], Fabian Hoppe[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
firstname.lastname@fiz-karlsruhe.de

Abstract. In most Knowledge Graphs (KGs), textual descriptions of entities are provided in multiple natural languages. Additional information that is not explicitly represented in the structured part of the KG might be available in these textual descriptions. Link prediction models which make use of entity descriptions usually consider only one language. However, descriptions given in multiple languages may provide complementary information which should be taken into consideration for the tasks such as link prediction. In this poster paper, the benefits of multilingual embeddings for incorporating multilingual entity descriptions into the task of link prediction in KGs are investigated.

1 Introduction

Various Knowledge Graphs (KGs) such as DBpedia and Wikidata have been published to share linked data and have been crucial for many tasks. However, according to the Open World Assumption, KGs are never complete. Due to this fact, different KG completion models which map KGs to a low dimensional vector space based on the task of link prediction have been proposed. However, only some of these models such as DKRL [10], MKBE[7] , Jointly[11] , SSP [9] , and LiteralE [6] leverage the textual descriptions of entities for the link prediction task [4]. Furthermore, most popular KGs contain descriptions in two or more languages for a single entity due to the multilingual community working on these KGs (as in Wikidata) or the multilingual nature of its sources (as in DBpedia). The cultural context and bias associated with each of these descriptions induces a difference with regards to content. However, despite the fact that entity descriptions are available in multiple natural languages, all the existing models including DKRL consider only one language. Figure 1 presents an example scenario showing the differences in contents of multilingual descriptions of a single entity. In this example, the description in German contains information which does not appear in the English or French descriptions. For instance, the fact that the team is the record winner of the U-19 Asian Cup with twelve titles is only mentioned in the German description.

This paper presents part of an ongoing work which is an empirical evaluation of an already published position paper [3]. Specifically, in this poster paper, the performance of the existing model DKRL in leveraging multilingual descriptions using multilingual embeddings has been analysed and the results of the initial experiments are discussed.

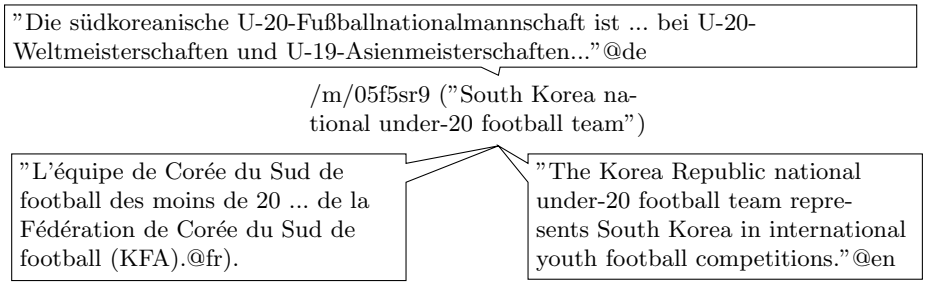

Fig. 1: An entity from Freebase with descriptions from its corresponding English, German, and French Wikipedia pages. For instance, the description in German provides more content that is not in the descriptions of either the English or the French Pages.

2 Multilingual Word Embeddings

As discussed in [3], an entity alignment model named KDCoE [2] has demonstrated the advantage of multilingual word embeddings by using a cross-lingual Bilbowa word embedding [5] to encode multilingual descriptions for the task of cross lingual learning. In this paper, the same approach is adopted to encode multilingual entity descriptions for a link prediction task on a monolingual dataset. In particular, the experiments have been performed with one of the existing models DKRL [10] using pretrained multilingual word embeddings by MUSE[3]. DKRL is an extension of TransE [1], which learns two kinds of vector representations for an entity, i.e., structure-based and description-based representations. DKRL adopts TransE for the structure-based representation and uses CNN to encode entity descriptions for the description-based representations. These two kinds of entity representations are learned simultaneously into the same vector space without forcing them to be unified. In our experiments, in order to effectively utilize multilingual descriptions, the embeddings of the words in the descriptions obtained by MUSE are passed as inputs to the encoder.

MUSE has been chosen because this paper deals with multilingual descriptions and MUSE aligns embeddings (specifically, FastText embeddings) of words

[3] https://github.com/facebookresearch/MUSE

in different languages into the same vector space. Figure 2 shows the CNN encoder part of DKRL with pretrained word embeddings from multilingual descriptions as inputs.

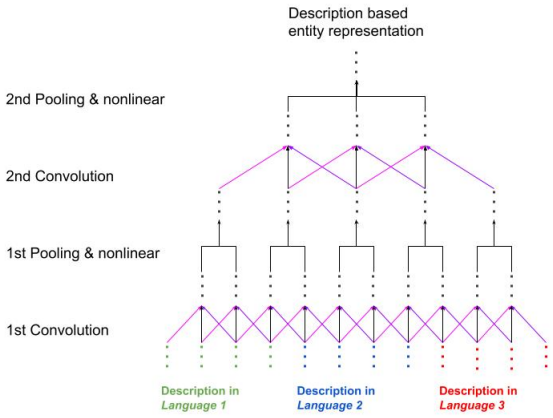

Fig. 2: Passing pretrained multilingual word embeddings to a CNN encoder which is adopted from DKRL [10] and shown in [3], in order to encode multilingual entity descriptions.

3 Experimental Evaluation

In this section, the experiments conducted to incorporate textual descriptions in English, French, and German into DKRL (for the task of link prediction) are presented. Table 1 shows the dataset created out of FB15K-237 [8] for the experiments by removing those triples for which either the head or tail entity does not have descriptions in at least one of the three languages mentioned above or have less than 3 words after preprocessing. Since FB15K-237 is a dataset generated from Freebase and the entity descriptions in Freebase are old, the descriptions in all the three languages have been constructed by taking the information from the summary part of their respective Wikipedia pages. During preprocessing, stop words are removed and all phrases are marked using entity names and also by applying Spacy's[4] named entity recognizer. The created dataset is available at `https://github.com/ISE-FIZKarlsruhe/Link-Prediction-with-Multilingual-Entity-Descriptions`.

For the experiments with DKRL and TransE, the code published at `https://github.com/xrb92/DKRL` by the authors of the DKRL paper and the code

[4] https://spacy.io/

for TransE available at `https://github.com/thunlp/OpenKE` has been used respectively. The DKRL model has been trained on three varieties of the dataset, given the names DKRL$_e$, DKRL$_{eg}$, and DKRL$_{egf}$. For DKRL$_e$, only English descriptions are used whereas for DKRL$_{eg}$ the combination of descriptions in German and English are used. On the other hand, descriptions in all the three languages are used to train DKRL$_{egf}$. The minimum, maximum, and average number of words are 3, 615, and 107.351 for the descriptions in DKRL$_e$, 9, 970, and 140.591 in DKRL$_{eg}$, and 18, 1460, and 192.091 in DKRL$_{egf}$ respectively.

In DKRL$_e$, the words are initialized using FastText pretrained embeddings and for the other two models MUSE pretrained embeddings are used. As shown in Table 2, TransE [1] has also been trained on the new dataset for fair comparison with DKRL. The hyperparameters are chosen from embedding size {50, 100, 150}, margin {1.0, 2.0, 3.0, 4.0, 5.0}, learning rate {0.01, 0.1, 1.0} (following the same procedure as in the paper of TransE). The optimal parameters for TransE on this dataset is embedding size: 100, margin: 4.0, learning rate: 0.1, and epoch: 1000. For all the other models DKRL$_e$, DKRL$_{eg}$, and DKRL$_{egf}$, the same procedure as in the original study DKRL has been adopted. The optimal parameters are entity and relation embedding size: 100, learning rate 0.001, margin: 1.0, window-size: 2, dimension of feature map: 100, and word embedding size: 300, for all the three varieties. DKRL$_e$ is trained for 1000 epochs where as DKRL$_{eg}$ and DKRL$_{egf}$ are trained for 1200 epochs.

As shown in Table 2, incorporating descriptions into the link prediction task brings improvement over TransE. However, when comparing the different varieties of DKRL, it is seen that combining multiple descriptions has only a slight improvement. For instance, hits@10 is the same for DKRL$_{eg}$ and DKRL$_{egf}$. One potential reason for such results could be the out of vocabulary words in the pretrained word embeddings by MUSE. There are 18.4% and 20% out of vocabulary words for DKRL$_{eg}$ and DKRL$_{egf}$ respectively and they are randomly initialized.

Table 1: The statistics of the dataset used for the experiments.

	FB15K-237
#Ent	12729
#Rel	234
#Train	219573
#Valid	13919
#Test	16084

Table 2: Experiment results using transE and DKRL models on the different varieties of the FB15K-237 dataset.

	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	213	0.266	0.175	0.297	0.448
DKRL$_e$	201	0.275	0.189	0.304	0.449
DKRL$_{eg}$	185	0.280	0.191	0.310	0.457
DKRL$_{egf}$	180	0.285	0.196	0.311	0.457

4 Conclusion and Future Work

In this paper, which is based on an already published position paper, preliminary results from an ongoing work to discuss the benefits of leveraging multilingual embeddings for the task of link prediction are presented. As a future work,

more experiments will be conducted by aligning pretrained FastText embeddings which have bigger vocabulary size, using MUSE, to avoid the problem which rises due to out of vocabulary words. Moreover, another way to improve the results will be investigated which is to learn description-based embeddings of an entity separately for each language and then fusing the vectors using different approaches.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating Embeddings for Modeling Multi-Relational Data. In: NIPS (2013)
2. Chen, M., Tian, Y., Chang, K.W., Skiena, S., Zaniolo, C.: Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. arXiv preprint arXiv:1806.06478 (2018)
3. Gesese, G.A., Alam, M., Sack, H.: Semantic entity enrichment by leveraging multilingual descriptions for link prediction. In: DL4KG@ ESWC (2020)
4. Gesese, G.A., Biswas, R., Alam, M., Sack, H.: A survey on knowledge graph embeddings with literals: Which model links better literal-ly? arXiv preprint arXiv:1910.12507 (2019)
5. Gouws, S., Bengio, Y., Corrado, G.: Bilbowa: Fast bilingual distributed representations without word alignments. In: ICML (2015)
6. Kristiadi, A., Khan, M.A., Lukovnikov, D., Lehmann, J., Fischer, A.: Incorporating literals into knowledge graph embeddings. In: International Semantic Web Conference. pp. 347–363. Springer (2019)
7. Pezeshkpour, P., Chen, L., Singh, S.: Embedding multimodal relational data for knowledge base completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3208–3218. Association for Computational Linguistics (Oct-Nov 2018), https://www.aclweb.org/anthology/D18-1359
8. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference. In: 3rd Workshop on Continuous Vector Space Models and their Compositionality. Association for Computational Linguistics (2015)
9. Xiao, H., Huang, M., Meng, L., Zhu, X.: Ssp: semantic space projection for knowledge graph embedding with text descriptions. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
10. Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M.: Representation learning of knowledge graphs with entity descriptions. In: AAAI (2016)
11. Xu, J., Qiu, X., Chen, K., Huang, X.: Knowledge graph representation with jointly structural and textual encoding. pp. 1318–1324 (08 2017). https://doi.org/10.24963/ijcai.2017/183