Semantic Entity Enrichment by Leveraging Multilingual Descriptions for Link Prediction

Genet Asefa Gesese[1,2], Mehwish Alam[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@kit.edu

Abstract. Most Knowledge Graphs (KGs) contain textual descriptions of entities in various natural languages. These descriptions of entities provide valuable information that may not be explicitly represented in the structured part of the KG. Based on this fact, some link prediction methods which make use of the information presented in the textual descriptions of entities have been proposed to learn representations of (monolingual) KGs. However, these methods use entity descriptions in only one language and ignore the fact that descriptions given in different languages may provide complementary information and thereby also additional semantics. In this position paper, the problem of effectively leveraging multilingual entity descriptions for the purpose of link prediction in KGs will be discussed along with potential solutions to the problem.

1 Introduction

Knowledge Graphs (KGs) such as Freebase [1], DBpedia [10], and Wikidata [13] have been created in order to share linked data which describes entities and the relationships between them. The availability of these various cross-domain KGs has sparked interest in undertaking research directions such as KG completion using tasks like link prediction. Hence, different Knowledge Graph Embedding (KGE) approaches, which map KGs to a low dimensional vector space, based on a link prediction task have been published. Link prediction is widely used because in the Open World Assumption the knowledge explicitly represented in a KG is never complete, there are always missing facts which can be predicted using link prediction.

DistMult [17] and ConvE [4] are among those KGE models which are trained on a link prediction task but without making use of the textual descriptions of entities. On the other hand, there are some models such as DKRL [15] which leverage the textual descriptions of entities for the link prediction task on (monolingual) datasets like FB15K [2] and FB15K-237 [12] and have demonstrated that using the textual descriptions enhance the embeddings of entities [5]. However,

despite the fact that the entities in these KGs have multilingual entity descriptions, all the existing models which use descriptions of entities focus on using descriptions written in only one natural language.

In most of the popular KGs, a single entity can have descriptions in two or more languages where the contents of the descriptions are different. This fact is demonstrated in Figure 1 using, as an example, a triple from FB15K with some of the descriptions of its head and tail entities extracted from Freebase. In this example, it can be seen that, for both the head ('m.02rcdc2') and tail ('m.019f4v') entities, the description provided in one language contains information that is not available in the description given in the other language. Hence, a KGE model which uses entity descriptions only in one language discards the extra information provided in the descriptions in other languages.

Fig. 1: A triple in FB15K with multilingual descriptions of the head and tail entities from Freebase.

2 Related Work

Few attempts have been made to combine the structured part of KGs with entity descriptions to learn KGE models. Among these models, DKRL [15], MKBE [11], and Jointly [16] use neural network encoders (either CNN or LSTM) to represent entity descriptions. The other models are SSP [14] and LiteralE [8] which rely on document embedding approaches to get representations for entity descriptions. In DKRL, CNN is used to encode entity descriptions using word embeddings as

an input. MKBE, same as in DKRL, uses CNN to encode textual descriptions of entities. The descriptions that are used in both DKRL and MKBE are provided in only one language (English).

Jointly [16] is a KGE method which combines structural and textual encoding as in DKRL but using (attentive) LSTM encoder instead of CNN. In this approach, the embedding of a word is initialized by taking the average of the embeddings of the entities whose description include this word. Initialising in this way does not work well for multilingual descriptions because it is not capable of capturing words which are from different languages but semantically similar and are not linked to the same set of entities.

SSP is another KGE approach which jointly learns from structured information and entity descriptions. This method adopts the Non-negative Matrix Factorization (NMF) topic model to generate a representation for an entity based on its description, i.e., treating each entity description as a document and taking the topic distribution of the document as the representation of the corresponding entity. However, the approach would not perform well with multilingual entity descriptions since the adopted topic model does not deal with multilinguality. In LiteralE, entity descriptions are represented using a document embedding technique proposed in [9]. This document embedding technique works by first mapping the whole document (i.e., entity description) and also every word present in the document into corresponding unique vectors and then taking the average or concatenation of the paragraph vector and word vectors so as to predict the next word in a context.

3 Proposed Methodology

In order to address the issues with the existing KGE models in using multilingual descriptions, this study provides the following insights into the potential solutions.

3.1 Applying Language Translators

The straight forward way to incorporate multilingual entity descriptions in the existing neural network encoder based KGE models (i.e., DKRL, MKBE, and Jointly) is first to convert all the descriptions into one language (English) with a language translator and then to pass as inputs to the encoder pre-trained embeddings of the words present in the descriptions. The pre-trained word embeddings can be obtained from any monolingual word embedding model. The main challenge with this method is the errors that occur during machine translation (converting multilingual descriptions into one language) will be propagated to the encoder. One way to address this issue is to use the multilingual word embeddings instead of applying machine translation.

3.2 Using Multilingual Word Embeddings

KDCoE [3] is a KGE approach which leverages a weakly aligned multilingual KG for semi-supervised cross-lingual learning using descriptions of entities. With this approach, the authors have demonstrated that a very good performance can be achieved for an entity alignment task by using an Attentive Gated Recurrent Unit encoder (AGRU) to encode multilingual descriptions with multilingual word embeddings as inputs. The multilingual embedding model used is a cross-lingual Bilbowa word embedding [6] trained on the cross-lingual parallel corpora Europarl v7 [7] and monolingual corpora of Wikipedia dump. The results from KDCoE show that multilingual text encoders can benefit from multilingual word embeddings. It would also be interesting to adopt the same approach, which is used to encode multilingual entity descriptions for cross-lingual entity alignment task in KDCoE, for a link prediction task on different monolingual datasets such as FB15K and FB15K-237. This approach allows to capture as much information as possible from entity descriptions present in multiple languages.

Furthermore, the existing models such as DKRL and Jointly can be improved by leveraging multilingual entity descriptions by passing as inputs to the encoders the embeddings of the words in the descriptions obtained by a multilingual word embedding model like MUSE [3]. For instance, Figure 2 shows how the CNN encoder part of DKRL can be modified to take pre-trained word embeddings from multilingual descriptions as inputs.

4 Conclusion

In this position paper, the problem of leveraging multilingual entity descriptions for link prediction task on KGs is discussed. As mentioned in Section 1 and Section 2, the available link prediction models on monolingual datasets such as FB15K-237 use only monolingual entity descriptions and ignore the fact that the descriptions in other languages may contain additional semantics. Thus, in this study, some insights into potential solutions to this problem are provided. These solutions enable the existing link prediction models to leverage multilingual entity descriptions. The solutions proposed in this study for link prediction task can also be adopted for other KG completion tasks such as triple classification and entity classification. In order to come up with an even better solution to the problem, conducting detailed analysis on the nature and quality of the multilingual entity descriptions available in different KGs such as DBpedia, Wikidata, and Freebase would be beneficial.

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

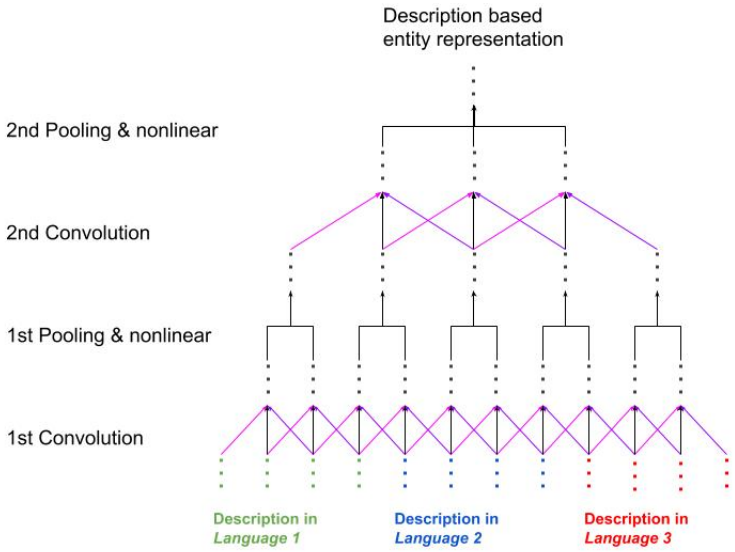

Fig. 2: Passing pre-trained multilingual word embeddings to a CNN encoder which is adopted from DKRL [15], in order to encode multilingual entity descriptions.

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250 (2008)
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating Embeddings for Modeling Multi-Relational Data. In: NIPS (2013)
3. 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)
4. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. 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)
6. Gouws, S., Bengio, Y., Corrado, G.: Bilbowa: Fast bilingual distributed representations without word alignments (2015)
7. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86. Citeseer (2005)
8. 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)
9. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
10. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
11. 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
12. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference. In: Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality. pp. 57–66. Association for Computational Linguistics (Jul 2015), https://www.aclweb.org/anthology/W15-4007
13. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
14. 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)
15. Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M.: Representation learning of knowledge graphs with entity descriptions. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
16. 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
17. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)