Leveraging Literals for Knowledge Graph Embeddings

Genet Asefa Gesese

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
Karlsruhe Institute of Technology, Institute AIFB, Germany
Genet-Asefa.Gesese@fiz-karlsruhe.de

Abstract. Nowadays, Knowledge Graphs (KGs) have become invaluable for various applications such as named entity recognition, entity linking, question answering. However, there is a huge computational and storage cost associated with these KG-based applications. Therefore, there arises the necessity of transforming the high dimensional KGs into low dimensional vector spaces, i.e., learning representations for the KGs. Since a KG represents facts in the form of interrelations between entities and also using attributes of entities, the semantics present in both forms should be preserved while transforming the KG into a vector space. Hence, the main focus of this thesis is to deal with the multimodality and multilinguality of literals when utilizing them for the representation learning of KGs. The other task is to extract benchmark datasets with a high level of difficulty for tasks such as link prediction and triple classification. These datasets could be used for evaluating both kind of KG Embeddings, those using literals and those which do not include literals.

Keywords: Knowledge Graph Embedding · Knowledge Graph Completion · Link Prediction · Literals · Benchmark Datasets

1 Introduction

Knowledge Graphs (KGs) consist of facts about any discipline in the real world in the form of entities, attributes of entities, and interrelations between entities. Various KGs have been published so far such as Wikidata [17], DBpedia [9], and YAGO [13], which have become crucial for different applications in the area of natural language processing, machine learning, information retrieval, and etc. As discussed in [7], due to the rigorous symbolic frameworks used by KGs it is difficult to use their data for other systems [2] and also the complexity of different important graph mining algorithms on KGs are proven to be NP-complete. Hence, to deal with these issues, it is beneficial to learn latent representation of the KGs while preserving the semantics present in these graphs.

When learning latent representations of KGs, it is necessary to capture the semantics contained in all elements of the KGs, i.e., from both relational and attributive triples. Relational triples are triples with relations between entities (object properties) whereas attributive triples are those with attributes (datatype

properties). Figure 1 presents, as an example, a graph which is part of the content about the entity *Covid-19* from Wikidata and Wikipedia[1]. In this graph, *Covid-19* and *Wuhan* are entities connected with the relation *location of discovery* creating a relational triple <*Covid-19 location-of-discovery Wuhan*>. The attributes *official name*, *short name*, and *also known as* could be considered as attributes taking short text literals as values whereas *description from Wikipedia* takes long text literals. The other attributes except *image*, take *datetime* values or *measurements* with/without units. These attribute values are either numeric or they could easily be converted to numeric.

The different type of literals associated with the entity *Covid-19* as shown in the example KG, hold important information which could not be found with just relational triples. Hence, a KG Embedding (KGE) model which makes use of all these literal values (i.e., Multimodal KGEs) would be able to learn better representations that are rich in semantics for the entities *Covid-19* and *Wuhan* as compared to models that do not use literals (Unimodal KGEs). Therefore, this thesis focuses on i) conducting a survey about current KGE models by performing experimentally-supported comparative analysis as discussed in Section 5.1, ii) building benchmark datasets which would be appropriate for evaluating both Unimodal KGEs and Multimodal KGEs (refer to Section 5.2 for details), and then proposing a new KGE model which addresses the shortcomings with existing models in terms of utilizing literals.

2 Importance

Most KGs contain a significant amount of information represented in the form of literals. It is common to find different types of literals such as measurements and date values in various KGs. For instance, in Wikidata, DBpedia, and YAGO there are date values associated with different events (birth date, date of death, ...) and measurements of length, size, density, and so on. When learning KGEs, it is important to handle literals effectively as they contain additional or complementary information to what is already present in the relational triples.

There are even more specific areas that would highly benefit from properly incorporating literals into the representation learning. One of them is KGs for IoT (Internet of Things) where there exist an enormous amount of literals such as measurements collected from sensors, like *datetime*, *latitude*, *longitude*, and *temperature* values. The Unimodal KGEs, those which do not make use of literals, would not perform well in such cases. Therefore, it is necessary to design a Multimodal KGE model to capture the semantics present in literals.

3 Related Work

Some attempts have been made to incorporate literals into the representation learning of KGs. Detailed analysis of these approaches is presented in a survey [7]

[1] `https://www.wikipedia.org/`

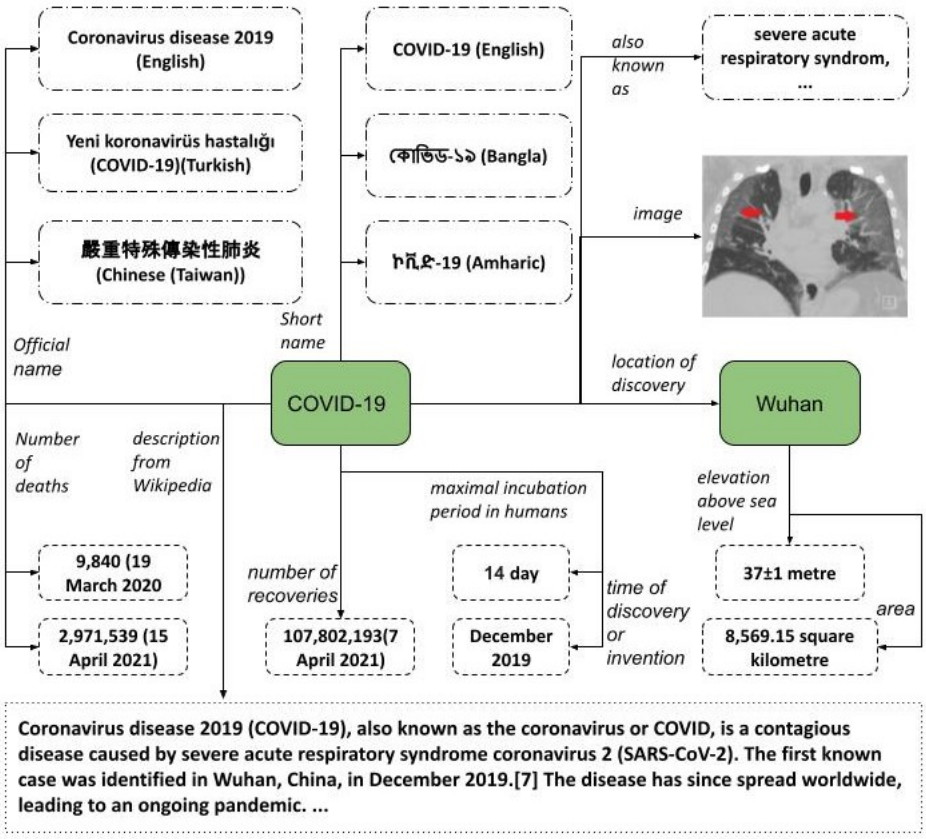

Fig. 1: An example graph containing part of literals extracted from Wikidata and Wikipedia about the virus Covid-19

conducted as part of this thesis. These KGE approaches could be grouped into different categories based on the kind of literals they use as follows:

Text literals: The approaches that make use of text literals are DKRL [20], Jointly [22], SSP [19], KDCoE [3], and KGloVe with literals [4]. DKRL is an extension of TransE [1], combining relational triples with textual descriptions to learn KG representations by encoding the descriptions using CNN. Similarly, Jointly extends TransE by capturing semantics from entity descriptions but it uses Attentive LSTM instead of CNN. SSP also combines relational triples and textual descriptions of entities for the embedding task by applying first-order constraints to capture the correlations of the triples and the descriptions. On the other hand, KDCoE learns KG representations using an entity alignment task. It applies a multilingual KG embedding model and a multilingual entity description embedding model over a weakly aligned multilingual KG for semi-supervised cross-lingual learning. KGloVe with literals is an attempt to incorporate entity

descriptions into the KGloVe KG embedding approach. One common drawback with these KGEs is the fact that they focus on long texts as literals and do not give attention to short text literals such as names and labels.

Numeric literals: MT-KGNN [14], KBLRN [5], LiteralE [8], and TransEA [18] are the ones using numeric literals. MT-KGNN applies a binary (pointwise) classifier for relational triple prediction and regression task for non-discrete attribute value prediction to learn embeddings for KGs. KBLRN uses relational, latent, and numerical feature types and trains them jointly end to end via a probabilistic PoE (Product of Experts) method. LiteralE works by incorporating literals into other existing unimodal KGE models. In this approach, two kinds of vectors for each entity are created, *entity vector* by the given unimodal KGE and *literal vector* from the entity's corresponding attribute values. These two vectors would be mapped to a new literal enriched entity vector using a learnable transformation function. On the other hand, TransEA extends TransE with an attributive embedding model which is based on a linear regression task. One of the drawbacks of these models is the fact that they fail to interpret datatypes of attributes. Besides, most of these models do not handle multi-valued attributes properly.

Others: IKRL [21] and MTKGRL [10] use images of entities in addition to the relational triples. On the other hand, MKBE [11] leverages both text and numeric literals along with images. As in the models with numeric literals, MKBE is not capable of capturing the semantics present in the data types/units of attribute values. For more details about all of the KGE models discussed in this section, refer to the survey [7].

4 Research Questions

In the previous sections, the advantages of using literals for KGEs and the current embedding models leveraging literals are discussed. Here, based on the shortcomings of the existing approaches in making use of literals, the following research questions are formulated.

- **RQ1:** *Which ones of the current KGEs using literals perform better for the task of link prediction?*
  It is beneficial to perform experiment-based comparison in order to better analyze and understand the capability of the existing Multimodal KGE models for the task of link prediction. Hence, experiments have been conducted on KGE models which use numeric and/or text literals and the results obtained are reported in [7]. More details on the models and evaluations are provided in Section 5.1.
- **RQ2:** *How to extract benchmark datasets from popular KGs such as Wikidata, focusing primarily on literals?*
  High quality benchmark datasets containing literals are required in order to properly evaluate Multimodal KGE models. Hence, some details about the collection of benchmark datasets LiterallyWikidata [6] are given in 5.2.

- **RQ3:** *How to effectively use literals together with relations between entities to learn KG representations?*
  Here, the main goal is to properly deal with the multimodality and multilinguality of literals when combining relational triples and attributive triples into the representation learning. Providing such a Multimodal KGE model which also addresses the weaknesses of the current models is yet to be done.
- **RQ4:** *How to evaluate KG embeddings on downstream tasks?*
  Evaluating KG embeddings on the tasks different from what they are trained on, would give insights into the re-usability of the embeddings for other tasks.

Therefore, the main contributions of this thesis would be:

- Providing an extensive survey of existing Multimodal KGE approaches, which includes experiments on the task of link prediction.
- A set of benchmark datasets *LiterallyWikidata* extracted from Wikidata and Wikipedia.
- A novel Multimodal KGE approach which leverages both relational triples and attributive triples for the link prediction and triple classification tasks.

5 Preliminary Results

In this section, the works that have been done towards solving the research questions defined in section 4 are presented.

5.1 Comparative analysis of existing approaches on link prediction

We have conducted an extensive survey on KG embedding models which use literals [7]. This survey presents a detailed analysis of the models in terms of the scoring function, the tasks used for training or evaluation, and model complexity. Furthermore, experimental results on the task of link prediction with models taking numeric and/or text literals are also presented. The models using numeric literals are the different varaities of LiteralE (ComplEx-LiteralE$_g$, ComplEx-LiteralE$_g$, and ComplEx-LiteralE$_g$), KBLN, MTKGNN, TransEA whereas the model DKRL$_{Bern}$ is the one with text literals. In addition to these models, DistMult-LiteralE$_g$-text (i.e., another variety of LiteralE) is also included in the experiments as a model making use of both numeric and text literals. Table 1 presents the results reported in the survey with these models on the dataset FB15K-237 [15], refer to the survey for the details about the dataset. As the result indicates, DistMult-LiteralE$_g$-text performs better than DistMult-LiteralE$_g$ which means combining numeric gives better results than just numeric.

5.2 Benchmark datasets for Knowledge Graph Completion (KGC) with literals

The ways existing KGC datasets such as FB15K-237 [15] and CoDEx [12] are created do not give attention to literals. In order to address this problem, we have

Table 1: Link prediction results on FB15K-237 dataset using filtered setting

		Mrr	Hits@1	Hits@10
Numeric	DistMult-LiteralE$_g$	0.314	0.228	0.481
	ComplEx-LiteralE$_g$	0.27	0.191	0.427
	ConvE-LiteralE$_g$	0.224	0.149	0.378
	KBLN	0.296	0.211	0.463
	MTKGNN	0.287	0.207	0.448
	TransEA	0.158	0.172	0.456
Text	DKRL$_{Bern}$	**0.327**	0.214	**0.546**
Numeric&Text	DistMult-LiteralE$_g$-text	0.319	**0.232**	0.489

created a collection of KGC benchmark datasets named LiterallyWikidata[6], with a primary focus on numeric and text literals. LiterallyWikidata contains three datasets varying in size and structure, namely, LitWD1K, LitWD19K, and LitWD48K. These datasets contain relational triples and attributive (numerical) triples together with entity/relation/attribute labels, aliases, and descriptions from Wikidata. Furthermore, LiterallyWikidata contains textual descriptions for the entities from their corresponding summary sections of English, German, Chinese, and Russian Wikipedia pages. Benchmarking experiments are conducted on the task of link prediction with the models DistMult [23], ComplEx [16], and DistMultLiteral [8]. The statistics of the datasets are given in Table 2. The LiterallyWikidata benchmark paper is currently under review at a conference.

Table 2: LiterallyWikidata KGC benchmarking datasets

	LitWD1K	LitWD19K	LitWD48K
#Entities	1,533	18,986	47,998
#Relations	47	182	257
#Attributes	81	151	297
#Structured Triples	29,017	288,933	336,745
#Numerical Attributive Triples	10,988	63,951	324,418
#Train	26,115	260,039	303,117
#Valid	1,451	14,447	16,838
#Test	1,451	14,447	16,838

6 Evaluation

When conducting the experiments on link prediction with existing Multimodal KGEs for addressing the research question RQ1 as discussed in Section 5.1, the datasets FB15K and FB15K-237 are used to evaluate the models. The evaluation metrics used are MR, MRR, and Hits@K. The same procedure is followed

to evaluate the quality of the benchmark datasets LiterallyWikidata which is created to solve RQ2. On the other hand, for evaluating the solution that will be provided for RQ3, more tasks other than link prediction such as triple classification would be used and evaluated with the same metrics.

7 Conclusion and Future Work

As the results discussed in Section 5.1 indicate, the current Multimodal KGEs suffer from various drawbacks such as not handling multi-valued attributes well and failing to capture the semantics in data types/units. Hence, there arises the necessity to design and implement a Multimodal KGE model which addresses these drawbacks and leverages literals for better KG embedding. Besides, the discussion in Section 5.2 shows the need for better benchmark datasets for Multimodal KGEs. Therefore, this thesis work provides a collection of benchmark datasets extracted from Wikidata and Wikipedia, named LiterallyWikidata.

Developing the Multimodal KGE model, to address the research question RQ3, is yet to be done and hence, it is part of the future work. Besides, the model would be evaluated on KGs from other domain such as scholarly articles on downstream tasks. This would be a solution for the research question RQ4.

Acknowledgements I would like to thank my supervisors Prof. Dr. Harald Sack and Dr. Mehwish Alam for their invaluable mentoring and support.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating Embeddings for Modeling Multi-Relational Data. In: NIPS (2013)
2. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. In: AAAI (2011)
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. Cochez, M., Garofalo, M., Lenßen, J., Pellegrino, M.A.: A first experiment on including text literals in kglove. In: Joint Proceedings of ISWC 2018 Workshops SemDeep-4 and NLIWOD-4 (2018)
5. García-Durán, A., Niepert, M.: Kblrn: End-to-end learning of knowledge base representations with latent, relational, and numerical features. In: Globerson, A., Silva, R. (eds.) Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence. pp. 372–381. AUAI Press (2018)
6. Gesese, G.A., Alam, M., Sack, H.: LiterallyWikidata - A Benchmark for Knowledge Graph Completion using Literals (Apr 2021). https://doi.org/10.5281/zenodo.4701190, `https://doi.org/10.5281/zenodo.4701190`
7. 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)

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. 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)
10. Mousselly-Sergieh, H., Botschen, T., Gurevych, I., Roth, S.: A multimodal translation-based approach for knowledge graph representation learning. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 225–234. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-2027, `https://www.aclweb.org/anthology/S18-2027`
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)
12. Safavi, T., Koutra, D.: CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Nov 2020)
13. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th International Conference on the World Wide Web. pp. 697–706 (2007)
14. Tay, Y., Tuan, L.A., Phan, M.C., Hui, S.C.: Multi-task neural network for non-discrete attribute prediction in knowledge graphs. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. p. 1029–1038. Association for Computing Machinery (2017)
15. 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 (2015)
16. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex embeddings for simple link prediction. p. 2071–2080. ICML'16, JMLR.org (2016)
17. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
18. Wu, Y., Wang, Z.: Knowledge graph embedding with numeric attributes of entities. In: Proceedings of The Third Workshop on Representation Learning for NLP. pp. 132–136. Association for Computational Linguistics (2018)
19. 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)
20. Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M.: Representation learning of knowledge graphs with entity descriptions. In: AAAI (2016)
21. Xie, R., Liu, Z., Luan, H., Sun, M.: Image-embodied knowledge representation learning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. p. 3140–3146. IJCAI'17, AAAI Press (2017)
22. Xu, J., Qiu, X., Chen, K., Huang, X.: Knowledge graph representation with jointly structural and textual encoding. pp. 1318–1324 (08 2017)
23. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: International Conference on Learning Representations (ICLR) (2015)