Effective Use of BERT in Graph Embeddings for Sparse Knowledge Graph Completion

Xinglan Liu
Know-Center GmbH
Graz, Austria
lliu@know-center.at

Hussain Hussain
Graz University of Technology
Graz, Austria
hhussain@know-center.at

Houssam Razouk
Graz University of Technology
Graz, Austria
houssam.razouk@student.tugraz.at

Roman Kern
Graz University of Technology
Graz, Austria
rkern@know-center.at

ABSTRACT

Graph embedding methods have emerged as effective solutions for knowledge graph completion. However, such methods are typically tested on benchmark datasets such as Freebase, but show limited performance when applied on sparse knowledge graphs with orders of magnitude lower density. To compensate for the lack of structure in a sparse graph, low dimensional representations of textual information such as word2vec or BERT embeddings have been used. This paper proposes a BERT-based method (BERT-ConvE), to exploit transfer learning of BERT in combination with a convolutional network model ConvE. Comparing to existing text-aware approaches, we effectively make use of the context dependency of BERT embeddings through optimizing the features extraction strategies. Experiments on ConceptNet show that the proposed method outperforms strong baselines by 50% on knowledge graph completion tasks. The proposed method is suitable for sparse graphs as also demonstrated by empirical studies on ATOMIC and sparsified-FB15k-237 datasets. Its effectiveness and simplicity make it appealing for industrial applications.

KEYWORDS

Knowledge graph embedding, sparse knowledge graph, language model, context aware embedding, BERT

ACM Reference Format:
Xinglan Liu, Hussain Hussain, Houssam Razouk, and Roman Kern. 2022. Effective Use of BERT in Graph Embeddings for Sparse Knowledge Graph Completion. In *Proceedings of ACM SAC Conference (SAC'22).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3477314.3507031

1 INTRODUCTION

Knowledge graphs (KG) are the foundation for many NLP applications such as information retrieval and question answering. To improve the completeness and correctness of knowledge graphs, various knowledge graph embedding techniques (see e.g. survey [7]) are devised, which show good performance for knowledge graph completion tasks on benchmark knowledge graphs like Freebase.

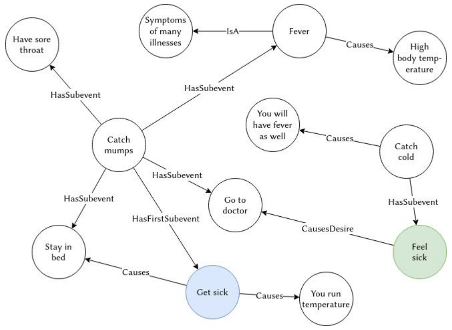

Figure 1: Example graph from ConceptNet100k. The semantically similar nodes "get sick" and "feel sick" are not connected, and would appear unrelated based on a graph embedding model. Our approach fine-tunes BERT on the node text attributes within the context of its graph neighborhood, before combined with ConvE for knowledge graph completion. As a result, both semantic similarity and graph neighborhood similarity are encoded.

Most knowledge graphs are often noisier and sparser than these benchmark knowledge graphs. For example, ConceptNet is a knowledge graph built via crowd sourcing and other sources [19]. Table 1 provides statistics on ConceptNet100k [9], in comparison to FB15k-237 [20]. We see that ConceptNet100k is two orders of magnitude sparser than FB15k-237.

Figure 1 depicts a small subset of Conceptnet100k and demonstrates the sparse nature of this knowledge graph. Here we see that nodes like "feel sick" and "get sick" are not directly connected. At the same time, one would expect a high similarity between these nodes. Based on this intuition, incorporating text attributes of nodes into the knowledge graph embedding models could be especially beneficial for sparse knowledge graphs.

To effectively capture the rich information stored in the text attributes, language models such as BERT are preferred choices, as they have been shown to consistently improve the performance of a

SAC'22, April 25 –April 29, 2022, Brno, Czech Republic

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507031

	# triples	# nodes	# rel types	density
ConceptNet100k	99,997	78,088	34	1.6e-5
ATOMIC	609,233	255,786	9	9.3e-6
FB15k-237	272,115	14,505	237	1.3e-3
FB15k-237-10k	10,328	14,505	237	4.9e-5

Table 1: Stats on knowledge graphs ConceptNet100k, ATOMIC, and FB15k-237, which are commonly used in literature. To study the sparsity, we add FB15k-237-10k, a subset from FB15k-237, where we randomly sampled 10k triples from FB15k-237 with the constraint that the number of nodes and relation types remain the same.

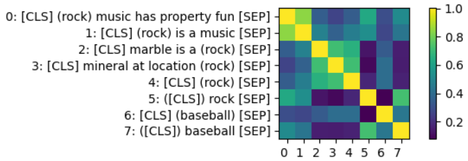

Figure 2: Pairwise cosine similarities of BERT embeddings for words *[CLS]*, *rock*, and *baseball* (shown in parenthesis) in different contexts (shown as axis label). Higher similarity (green) is observed for *rock* in similar context. When embeddings of the *[CLS]* token is used, *rock* is more similar to *baseball*, than to the other mentions of *rock*.

large variety of NLP tasks [3]. For instance, using BERT embeddings to represent nodes allows semantically similar nodes to be closer to each other in the embedding space, even if there is no link directly connecting them within the knowledge graph.

Indeed, in a recent work on knowledge graph completion [11], BERT embeddings are used for representing node text attributes. As a result, the authors have achieved significant gain in model performance. However, to extract BERT embeddings of each node, only the *[CLS]* token of the node text attribute is used without future context. We are interested in studying if that can be improved upon since, similar to ELMo [14], meaningful BERT embeddings of a word only exist within a given context.

Figure 2 illustrates the context dependency of BERT embeddings. Here a pre-trained BERT base model[1] is used to extract embeddings for the word *rock* in different contexts (shown as axis label), and we visualize their pairwise cosine similarities as a heat map. For comparison, we also include similarities to the word *baseball*. For example, the embedding in the context *rock is a music* shows higher similarity to that in similar context (*rock music has property fun*) than that in different context (*marble is a rock*). However, the *[CLS]* token within the context *rock* is more similar to that of the *[CLS]* token for *baseball* than any other mentions of *rock*.

In other words, the common practice where BERT embeddings are extracted either without contexts or only from the *[CLS]* token is sub-optimal, as is also pointed out in ELMo [14] and Sentence-BERT [16], respectively.

In this work[2] , in order to arrive at a knowledge graph embedding model for sparse knowledge graph completion that combines both textual and structural information, we extract BERT embeddings within the contexts of the local neighborhood, before feeding them into a graph embedding model ConvE.

Our main contributions are:

- Leverage transfer learning and the context dependency of BERT embeddings to learn context-aware node embeddings suitable for KG completion.
- Empirically demonstrate the effectiveness of the proposed approach of combining textual and structural information on sparse knowledge graphs.
- Establish new state of the art performance for real-life sparse KG completion task on Conceptnet100k.

2 RELATED WORK

Graph embedding techniques have been used to effectively solve graph-related tasks [4]. For knowledge graphs, where relations have different types/features, specific embedding approaches have been introduced such as TransE [1], RGCN [18], and ConvE [2]. However, these embedding methods are not particularly text-aware. Attempts to build text-aware graph embeddings are usually made using pre-computed text representation like bag-of-keywords in [8, 24], word2vec embeddings [12] in [6, 23], BERT embeddings in [11, 21, 25] and XLNet embeddings in [10]. Complementary to the pre-computed representations, research following a neural-symbolic have been investigated [13]. Making use of the node text attributes, which appear in knowledge graphs, some existing works also use rule based methods to densify the graph, where the rules include cut off on cosine similarity between BERT embeddings of nodes [11] and word overlap [10].

3 METHOD

3.1 Problem Formulation

Preliminaries. Given a knowledge graph $G = (E, R, T)$ with a set of entities E, a set of relations R, and a set of relation types T. Each entity $e \in E$ has a textual representation $\bar{e}$. Each relation $(h, r, t) \in R \subseteq E \times T \times E$, is formed of a triple of a head entity, relation type, and a tail entity. Each triple in a knowledge graph then represents a fact.

Graph pre-processing. Similar to Dettmers et al. [2], for each relation $(h, r, t) \in R$, an inverse relation (t, r^{-1}, h) is added to R.

Knowledge graph completion task. Given a knowledge graph, and an incomplete triple $(h, r) \in E \times T$, the task is to predict an entity $t \in E$ which completes the triple to (h, r, t).

Evaluation. As score functions, we use the mean rank (MR), mean reciprocal rank (MRR), and the hits@k with $k \in \{1, 10\}$. We compute these scores in a filtered setting, i.e. candidates that form valid known triples are excluded from ranking together with their inverse triples [2].

3.2 BERT Fine-tune and Feature Extraction

The workflow for BERT-ConvE is summarized in Figure 3. BERT model is fine-tuned on text sequences that represent the node within a triple, before extracting embeddings. The rationale is twofold: (1) Fine-tuning BERT in-domain has shown to improve performance of

[1] https://huggingface.co/bert-base-uncased

[2] We release all code and data for future studies at https://github.com/tugraz-isds/kd

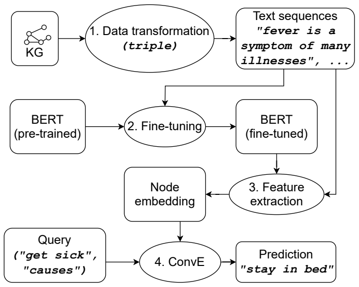

Figure 3: BERT-ConvE work flow.

specific tasks [5]; (2) As BERT embeddings are context dependent, we exploit this property in order to better represent a node within its local textual and structural contexts.

Data transformation. Firstly, a graph is transformed into text sequences, by representing each triple as a single text sequence, where relation types are split into the words that make them up (e.g. *HasSubevent* becomes *"has subevent"*). And an example generated text sequence can be *"catch mumps has subevent stay in bed"*.

Fine-tuning. BERT model is then fine-tuned on the transformed text sequences, using the transformers library from Hugging Face [22], with the weights of the complete model being updated.

Feature Extraction. After finetuning, BERT embeddings for each node is extracted within the context of the transformed text sequences. For example, given the sequence "*catch mumps has subevent stay in bed*" as the text representation for the node "*catch mumps*", we first obtain the last hidden state of each token in the sequence by making a single pass through the BERT encoder. Then we take the average embedding of the three tokens that make up the text attribute of the node: *"catch"*, *"mum"*, and *"##ps"*.

In addition, for nodes of degree 2 or more, the extracted embeddings from different transformed sequences are averaged. For example, given the node "*catch mumps*" with two text representations: "*catch mumps has subevent stay in bed*" and "*catch mumps has subevent have sore throat*", two different embedding vectors will be extracted from these two sequences. These two embedding vectors are averaged to obtain the final embedding for this node.

3.3 ConvE

ConvE [2] is used for modelling graph structure for tasks like KG completion. To incorporate the textual information, BERT embeddings of nodes are used to replace the randomly initialized embedding layer used in the original ConvE model. During the training of ConvE, BERT embeddings are frozen.

In summary, the proposed BERT-ConvE method combines BERT and ConvE to achieve a text and structure aware KG embedding model, with special attentions on construction the contexts for feature extraction. Without the context-aware aspect, the proposed method is conceptually similar to Malaviya et al. [11].

4 EXPERIMENTS AND DISCUSSIONS

Data. For the experiments in this paper, we use the dataset "ConceptNet100k" as constructed for knowledge base completion task by Li et al. [9][3], ATOMIC [17][4], and FB15k-237 [20]. For the node text attributes for FB15k-237, we use the same as in Yao et al. [25]. The standard validation and test splits are used for all datasets. Statistics of the training split for all benchmark datasets are shown in Table 1.

	ConceptNet100k				
	MRR	MR	Hits@1	Hits@3	Hits@10
ConvE	20.00	6554.65	13.41	21.83	32.77
Malaviya et al. [11] best	51.11	-	39.42	59.58	73.59
BERT-ConvE best	**76.91**	**57.45**	**69.14**	**82.03**	**91.41**

	ATOMIC				
	MRR	MR	Hits@1	Hits@3	Hits@10
ConvE	10.38	60620.91	8.44	10.63	13.83
Malaviya et al. [11] best	13.88	-	11.50	14.44	18.38
BERT-ConvE best	13.20	5903.38	10.23	13.74	**18.75**

Table 2: Performance on ConceptNet100k and ATOMIC test data for knowledge base completion. ConvE represents a model just based on structure information. The second approach [11] is comparable with ours, as it also employs BERT to capture textual information. Our BERT-ConvE method, provides the best results for ConceptNet100k and shows comparable performance for ATOMIC.

Results. We first establish a baseline via ConvE, which uses exclusively structural information and is unaware of the node textual attributes. Next, we compare BERT-ConvE to a similar approach [11], where also BERT embeddings (without contexts) are used.Table 2 lists the performance of the models on the task of knowledge graph completion on the test data.

In comparison with the purely structure-based baseline ConvE, the addition of textual information (BERT-ConvE and Malaviya et al. [11]) improves the performance significantly on ConceptNet100k and ATOMIC. Furthermore, BERT-ConvE improves by 50% in MRR over the current state of the art on ConceptNet100k, highlighting its effectiveness.

Ablation. As a direct evidence for the importance of context in BERT feature extraction, we extracted BERT embeddings using only the node text attributes (similar to [11, 21]). The resulting MRR is 46.07. Namely, context optimization accounts for 65% improvement in MRR.

Sparsity dependency. In order to find out how dense a graph needs to be for it to be less beneficial to incorporate textual information, we generate several graphs with different density by down-sampling FB15k-237, similar to Pujara et al. [15]. This procedure not only sparsifies the graph, but also decreases the number of training examples.

Figure 4 compares the performance of both ConvE and BERT-ConvE on FB15k-237 for different sparsity values. We observe that, while for the original FB15k-237 dataset, BERT-ConvE is 2 points behind that of ConvE, BERT-ConvE surpasses ConvE as the sparsity increases. The experiment confirms the previous observations on ConceptNet100k and ATOMIC, that the addition of BERT to ConvE yields a better performance on sparse knowledge graphs.

[3] https://ttic.uchicago.edu/~kgimpel/commonsense.html

[4] https://homes.cs.washington.edu/~msap/atomic/

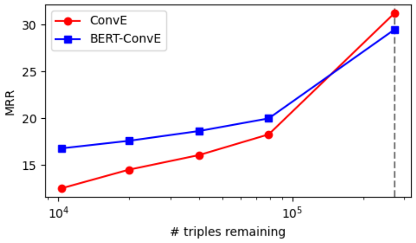

Figure 4: MRR of ConvE and BERT-ConvE in terms of the number of remaining triples in FB15k-237 subgraph. Lower number of triples indicates higher sparsity.

5 CONCLUSION

In this work, we introduced BERT-ConvE, a model that effectively exploits transfer learning and context-dependency of BERT in combination with a convolutional network model, ConvE, for the knowledge graph completion task.

The model achieved substantial improvement over the state of the art on sparse knowledge graph completion tasks on ConceptNet100k. The main contributing factor to the performance gain is the inclusion of the textual attributes of the neighborhood as contexts for fine-tuning and BERT embedding extraction. Experiments on sparsified FB15k-237 dataset show that BERT-ConvE is suitable for sparse knowledge graphs, where structural information is limited and textual information is informative for reasoning over the graph. As a result, our work recommends to use both, the node text attributes and the graph context (i.e., neighborhoods), to effectively exploit BERT for sparse KG completion.

Integrating the language model together with the graph embedding model in an end-to-end fashion is a possible future direction.

ACKNOWLEDGEMENTS

The research was conducted under the framework of the ECSEL AI4DI "Artificial Intelligence for Digitising Industry" project. The project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The Know-Center is funded within the Austrian COMET Program—Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. We acknowledge useful comments and assistance from our colleagues at Know-Center and at Infineon.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*. 1–9.

[2] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[3] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1, Mlm (2019), 4171–4186. arXiv:1810.04805

[4] Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (2018), 78–94.

[5] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. (2020). http://arxiv.org/abs/2004.10964

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. Nips (2017), 1–11.

[7] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2020. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *arXiv e-prints* (2020), arXiv–2002.

[8] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[9] Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense Knowledge Base Completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1445–1455. https://doi.org/10.18653/v1/P16-1137

[10] Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2019. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. (2019). https://doi.org/10.1609/aaai.v34i05.6364 arXiv:1909.05311

[11] Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2019. Commonsense Knowledge Base Completion with Structural and Semantic Context. (2019). arXiv:1910.02915 http://arxiv.org/abs/1910.02915

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:cs.CL/1301.3781

[13] Farhad Moghimifar, Lizhen Qu, Terry Yue Zhuo, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2021. Neural-Symbolic Commonsense Reasoner with Relation Predictors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 797–802. https://doi.org/10.18653/v1/2021.acl-short.100

[14] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[15] Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1751–1756. https://doi.org/10.18653/v1/D17-1184

[16] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[17] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. arXiv:cs.CL/1811.00146

[18] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[19] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *CoRR* arXiv:1612.03975 (2016). arXiv:1612.03975 http://arxiv.org/abs/1612.03975

[20] Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*. Association for Computational Linguistics, Beijing, China, 57–66. https://doi.org/10.18653/v1/W15-4007

[21] Bin Wang, Guangtao Wang, Jing Huang, Jiaxuan You, Jure Leskovec, and C. C. Jay Kuo. 2021. Inductive Learning on Commonsense Knowledge Graph Completion. arXiv:cs.AI/2009.09263

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[23] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2018. Representation Learning of Knowledge Graphs with Entity Attributes and Multimedia Descriptions. *2018 IEEE 4th International Conference on Multimedia Big Data, BigMM 2018* (2018), 2659–2665. https://doi.org/10.1109/BigMM.2018.8499179

[24] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. 2016. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*. PMLR, 40–48.

[25] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. arXiv:cs.CL/1909.03193