MG-BERT: Multi-Graph Augmented BERT for Masked Language Modeling

Parishad BehnamGhader and **Hossein Zakerinia** and **Mahdieh Soleymani Baghshah**
Sharif University of Technology
Tehran, Iran
{pbehnamghader, hzakerynia}@ce.sharif.edu, soleymani@sharif.edu

Abstract

Pre-trained models like Bidirectional Encoder Representations from Transformers (BERT), have recently made a big leap forward in Natural Language Processing (NLP) tasks. However, there are still some shortcomings in the Masked Language Modeling (MLM) task performed by these models. In this paper, we first introduce a multi-graph including different types of relations between words. Then, we propose Multi-Graph augmented BERT (MG-BERT) model that is based on BERT. MG-BERT embeds tokens while taking advantage of a static multi-graph containing global word co-occurrences in the text corpus beside global real-world facts about words in knowledge graphs. The proposed model also employs a dynamic sentence graph to capture local context effectively. Experimental results demonstrate that our model can considerably enhance the performance in the MLM task.

1 Introduction

In recent years, pre-trained models have led to promising results in various Natural Language Processing (NLP) tasks. Recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) has received much attention as a pre-trained model that can be easily fine-tuned for a wide range of NLP tasks. BERT is pre-trained using two unsupervised tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In (Ettinger, 2019), some psycholinguistic diagnostics are introduced for assessing the linguistic capacities of pre-trained language models. These diagnostic tests consist of commonsense and pragmatic inferences, role-based event prediction, and negation. Ettinger (2019) observes some shortcomings in BERT's results and demonstrates that although BERT sometimes predicts the first candidate for the masked token almost correctly, some of its top candidates contradict each other. Besides, in the tests targeting commonsense and pragmatic inference, it is illustrated that BERT can not precisely fill the gaps based on just the input context (Ettinger, 2019).

In this paper, we incorporate co-occurrences and global information about words through graphs describing relations of words along with local contexts considered by BERT. The intention is to find more reliable and meaningful embeddings that result in better performance in MLM task. Utilizing external information about the corpus and the world in the form of graphs helps the model fill the gaps in the MLM task more easily and with more certainty. We take advantage of the rich information source accessible in knowledge graphs and also condensed information of words co-occurrence in graphs using Relational Graph Convolutional Network (R-GCN) to enrich the embedding of tokens. We also utilize the words in the current context as a dynamic complete graph using an attention mechanism. These graphs can considerably influence the performance of BERT in the MLM task as shown in the experiments.

2 Related Work

Knowledge graphs (KGs) are valuable sources of facts about real-world entities. Many studies have been recently introduced to utilize knowledge graphs for various purposes, such as recommender systems (Wang et al., 2019a,b; He et al., 2020) or link prediction (Feng et al., 2016; Nguyen et al., 2018; Sun et al., 2019; Zhang et al., 2020). Recently, using BERT along with knowledge graphs has also been attended for knowledge graph completion and analysis. Yao et al. (2019) employ KG-BERT in triple classification, link prediction, and relation prediction tasks. Furthermore, knowledge graphs are used in NLP tasks such as text classification (K M et al., 2018; Ostendorff et al., 2019; Zhang et al., 2019a), named entity recognition (Dekhili et al., 2019), and language

Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), pages 125–131

modeling (Ahn et al., 2016; Logan et al., 2019). ERNIE (Zhang et al., 2019b) is an enhanced language representation model incorporating knowledge graphs. In addition to BERT's pre-training objectives, it uses an additional objective that intends to select appropriate entities from the knowledge graph to complete randomly masked entity alignments. Moreover, named entity mentions in the text are recognized and aligned to their corresponding entities in KGs.

Other types of graphs have also been utilized in NLP tasks in some studies. For instance, Text GCN (Yao et al., 2018) applies Graph Convolutional Network (GCN) to the task of text classification. This paper's employed graph is a text graph created based on token co-occurrences and document-token relations in a corpus. Moreover, VGCN-BERT (Lu and Nie, 2019) enriches the word embeddings of an input sentence using the text graph inspired by Text GCN (Yao et al., 2018) and examines the obtained model in FIRE hate language detection tasks (Mandl et al., 2019).

In this paper, we aim to improve BERT's performance (in the MLM task) by incorporating a static multi-graph that includes both the knowledge graph and global co-occurrence graphs derived from the corpus as well as a dynamic graph including input sentence tokens. Static text graphs have been recently employed in VGCN-BERT (Lu and Nie, 2019) via a modified version of GCN that extends the input by a fixed number of embeddings. However, the modification of embeddings in this work is only based on input tokens. Neither other vocabularies in the static text graphs nor real-world facts (available in KGs) affect the final embeddings of tokens. On the other hand, while ENRIE (Zhang et al., 2019b) and KEPLER (Wang et al., 2019c) utilize KGs to reach an improved model, they do not employ other graphs derived from the corpus. Also, ERNIE does not learn graph-based embedding during representation learning and only adopts embeddings trained by TransE (Bordes et al., 2013). However, in our model, since we incorporate a multi-graph by extending BERT architecture and providing a graph layer of an R-GCN module and attention mechanism, a multi-graph augmented representation learning model is obtained.

3 Preliminaries

GCN (Kipf and Welling, 2017) is one of the most popular models for graph node embedding. R-GCN (Schlichtkrull et al., 2018) extends GCN to provide node embedding of multi-relational graphs:

$$h_i^{(l+1)} = \sigma(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}),$$

where $h_i^{(l)}$ is the l-th layer's hidden state of node v_i, $W_r^{(l)}$ is the weight matrix for relation r in layer l, and W_0 is the weight matrix for self-loops. $\mathcal{N}_i^r$ is the set of v_i's neighbours under relation r and $c_{i,r}$ is a normalization constant.

4 Methodology

This section presents the overall architecture of our model, called Multi-Graph augmented BERT (MG-BERT). MG-BERT takes advantage of BERT's power in capturing context of an input text as well as a graph module including an R-GCN layer over a static multi-graph and a graph attention layer over a dynamic sentence graph. This static multi-graph includes global information about words available as facts in KGs in addition to dependencies between tokens of the input text and other words in the vocabulary which are discovered by computing co-occurrence statistics in the corpus. Two graphs are used to condense co-occurrences of words in the corpus inspired by Text GCN (Yao et al., 2018) that are also employed by VGCN-BERT (Lu and Nie, 2019). One of these graphs includes local co-occurrences of terms that is computed based on point-wise mutual information (PMI) of terms i and j which is calculated by:

$$p(i) = \frac{\#W(i)}{\#W}, \quad p(i,j) = \frac{\#W(i,j)}{\#W},$$
$$\text{PMI}(i,j) = log\frac{p(i,j)}{p(i)p(j)}. \qquad (1)$$

In the above equations, $\#W(i)$ and $\#W(i,j)$ denote the number of fixed size windows containing term i and both of the terms i and j, respectively. $\#W$ is the whole number of windows in the corpus. The other graph includes the document level co-occurrence of tokens in the corpus computed based on term frequency-inverse document frequency (TF-IDF). The knowledge graph is also incorporated in this multi-graph. Formally, the weighted edges between token i and token j for three types of relations $\mathcal{R} = \{\text{KG}, \text{PMI}, \text{TF-IDF}\}$

in the multi-graph are:

$$\begin{cases} A_{ij}^{TF-IDF} = \lambda_T \sum\limits_{d \in docs} \mathrm{T}_{id}\mathrm{T}_{jd} & \\ A_{ij}^{KG} = \lambda_K \sum\limits_{e \in \mathrm{KG}} \mathrm{KG}_{ie}\mathrm{KG}_{ej} & \text{if } i,j \in \mathrm{KG} \\ A_{ij}^{PMI} = \lambda_P \, \mathrm{PMI}(i,j) & \text{if } \mathrm{PMI}(i,j) > 0 \\ A_{ij}^{*} = 1 & \text{if } i = j \end{cases} \tag{2}$$

where T_{id} denotes the TF-IDF of token i in document d, $\mathrm{PMI}(i,j)$ shows PMI calculated by Eq. 1, and $\mathrm{KG}_{e_1 e_2}$ is nonzero when a relation between these two entities exists in the knowledge graph. Note that we add a self-connection relation to our knowledge graph for maintaining one-hop links, while also considering two hops as in Eq. 2 to employ indirect relations through paths of length two in the knowledge graph. λ_K, λ_P, and λ_T are also hyperparameters that can control the impact of three types of relations on tokens' embeddings. To utilize the multi-graph introduced above, we add a single-layer R-GCN described in Section 3 to the BERT model.

Furthermore, we use a graph attention mechanism to capture local information via a dynamic and complete graph in which nodes represent all tokens of the input sentence. The complete dynamic graph is used in order to obtain context-dependent new embeddings while the R-GCN layer itself provides the same new embeddings for a specific token even if the token appears in different contexts. This happens because the single R-GCN layer always performs on the same static multi-graph.

As shown in Fig. 1, the whole graph module is placed immediately after the BERT token embeddings layer since the hidden states of the whole vocabulary are available in this layer. We pass the entire multi-graph to the R-GCN module so that the global dependencies would affect embeddings of tokens properly using Eq. 3. We also use an attention mechanism as in Eq. 4 to consider the local context. The new embedding of token i in sentence s is computed as:

$$h_i^{'} = (1 - \lambda_{dyn}) \sum_{r \in R} \hat{A}_i^r h_i W_r \tag{3}$$

$$+ \lambda_{dyn} \left(\Big\Vert_{k=1}^{K} \sum_{j \in s} \alpha_{ij}^k h_j W_k^{Val} \right) W^O, \tag{4}$$

$$\alpha_{ij}^k = \frac{\exp((h_i W_k^{Query}).(h_j W_k^{Key}))}{\sum_{t \in s} \exp((h_i W_k^{Query}).(h_t W_k^{Key}))}, \tag{5}$$

where $\hat{A}^r$ refers to the normalized adjacency matrix of relation r, Ws are trainable weight matrices (i.e. W_rs denote parameters of the R-GCN layer and W_k^{Query}, W_k^{Key}, and W_k^{Val} denote the attention parameters), and h_i is the i^{th} token's embedding from the BERT token embeddings layer.

Next, we aggregate the obtained tokens' embeddings by the graph module with position embeddings and segment embeddings (similar to BERT). Afterward, we feed these representations to BERT encoders to find final embeddings. The proposed model architecture is shown in Figure 1.

In the training phase, a token from each sentence is randomly masked, and the model is trained to predict the masked token based on both the context and the incorporated static multi-graph.

5 Experiments

In this section, we explain the details of training MG-BERT and conduct experiments to evaluate and compare our model with the related methods recently proposed.

Datasets. During training, we use the WN18 knowledge graph, derived from WordNet, as an unlabeled graph (Bordes et al., 2014). We also experiment MG-BERT and other recent models on CoLA, SST-2, and Brown datasets (Warstadt et al., 2019; Socher et al., 2013; Francis and Kucera, 1979). The detailed description of these datatsets is given in Appendix A.

Parameter Setting. In order to capture word co-occurence statistics of the corpus, we use the BERT's tokenizer on sentences and set the sliding window size to 20 when calculating the PMI value. The whole BERT module in MG-BERT is first initialized with the pre-trained *bert-base-uncase* version of BERT in PyTorch and the model is trained on the MLM task with cross entropy loss (Wolf et al., 2019). Regarding Eq. 2, different hyper-parameter settings have been used for each dataset. λ_K, λ_P, and λ_T are set to 0.01, 0.001, and 0.001, respectively in both CoLA and Brown datasets and 0.001, 1.0, and 0.001 in SST-2 dataset. The hyperparameter λ_{dyn} is also set to 0.8. The graph attention mechanism is performed with 12 heads. The R-GCN and graph attention layers' output dimension are also set to 768 that equals to the dimension of the BERT token embeddings layer to substitute easily BERT's token embeddings with the embeddings

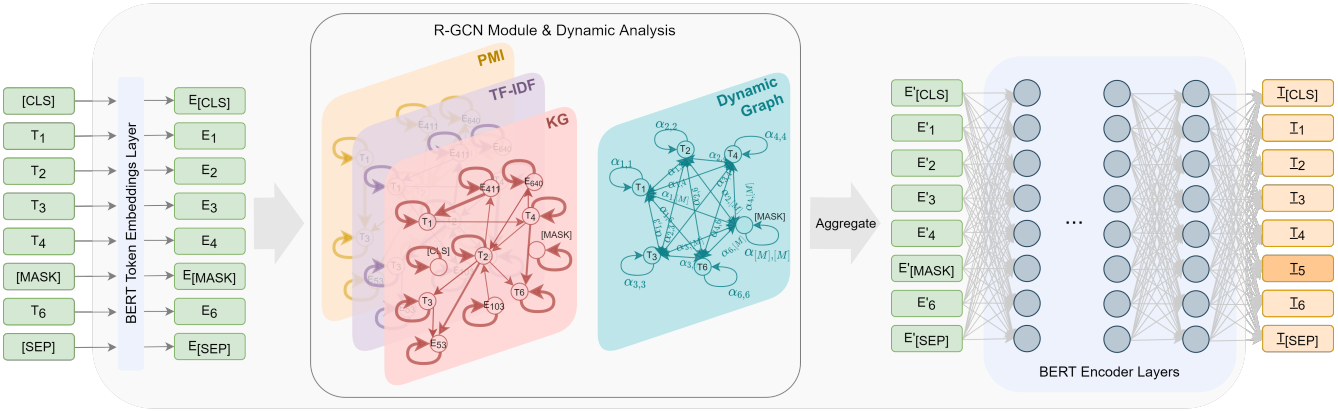

Figure 1: The architecture of MG-BERT. The "Aggregate" phase includes an aggregation of new tokens' embeddings with the position embeddings and the segment embeddings of the BERT model.

derived from the graph module. We also employ the normalization trick introduced in GCN (Kipf and Welling, 2017) to normalize each adjacency matrix in the multi-graph.

Compared methods. To assess our model, we compare it with BERT as the baseline. Moreover, ERNIE and VGCN-BERT are being compared as the recent methods utilizing knowledge graph and text graph, respectively (Zhang et al., 2019b; Lu and Nie, 2019). We also compare MG-BERT with MG-BERT(base) which doesn't use the dynamic graph incorporating the context according to Eq. 4. All these models are fine-tuned on the text datasets for a fair evaluation.

Results. We evaluate our model using Hits@1 and Hits@5 metrics. Hits@k shows the proportion of correct tokens appearing in the top k results for each sample. In Table 1, we report the results of evaluations performed on the test sets of CoLA, SST-2, and Brown datasets. These results demonstrate that the proposed method outperforms other models and taking advantage of the graph module with dataset-specific hyper-parameters improves the performance.

The reason to our superiority over VGCN-BERT (Lu and Nie, 2019) is that it doesn't take advantage of real-world facts (available in KGs). Moreover, as opposed to MG-BERT, it modifies initial embeddings of tokens only based on input tokens of each sentence and other vocabularies in the text graphs don't influence the final embeddings of tokens. On the other hand, ERNIE (Zhang et al., 2019b) doesn't take full advantage of graphs since it doesn't use graphs derived from the corpus. Besides, it does not learn graph-based embeddings during representation learning. It is worth mentioning that the entity embedding model used in ERNIE has been trained on a huge subset of Wikidata[1], which is almost 120 times bigger than WN18 knowledge graph employed in our method.

The superiority of MG-BERT over MG-BERT(base) demonstrates the importance of the dynamic sentence graph and the results of MG-BERT(base) itself shows that utilizing the static multi-graph has been useful.

Graphs	Hits@1	Hits@5
K	70.51 ± 1.28	86.27 ± 0.31
P	69.63 ± 1.78	85.75 ± 0.91
T	69.78 ± 1.18	84.78 ± 1.10
KP	70.37 ± 1.41	85.66 ± 0.92
KT	70.50 ± 0.99	85.54 ± 0.83
PT	70.60 ± 1.32	85.22 ± 0.80
KPT	70.94 ± 1.20	85.12 ± 1.20

Table 2: Experimental results of variations of MG-BERT(base) using different graphs on CoLA dataset. The symbols K, P, and T stand for employing KG, PMI, and TF-IDF relations, respectively.

In addition, evaluation results of different variations of MG-BERT(base) on CoLA dataset, considering different graphs, are represented in Table 2, demonstrating the effect of each graph on the performance. The experimental results indicate the role of exploiting various graphs in language representation learning.

We also compare MG-BERT and MG-BERT(base) with other models using perplexity

[1]https://www.wikidata.org/

Model	CoLA		SST-2		Brown	
	Hits@1	Hits@5	Hits@1	Hits@5	Hits@1	Hits@5
BERT (Devlin et al., 2019)	68.50 ±1.49	84.53 ±1.18	80.48 ±0.85	88.42 ±0.70	58.31 ±1.17	76.38 ±0.49
ERNIE (Zhang et al., 2019b)	69.57 ±0.89	84.58 ±0.72	81.17 ±0.77	88.26 ±0.57	57.42 ±0.73	75.34 ±0.65
VGCN-BERT (Lu and Nie, 2019)	69.03 ±0.78	84.81 ±0.62	80.85 ±0.48	88.37 ±0.58	57.97 ±0.97	76.17 ±0.67
MG-BERT(base)	70.95 ±1.20	85.12 ±1.20	81.28 ±0.51	88.56 ±0.79	**58.66** ±0.11	**76.64** ±0.61
MG-BERT	**71.72** ±0.97	**86.67** ±0.51	**83.07** ±0.47	**89.13** ±0.39	58.38 ±0.60	76.59 ±0.67

Table 1: Hits@k esults on CoLA, SST-2, and Brown datasets. The best score is highlighted in bold and the second best score is highlighted with underline.

Model	CoLA	SST-2	Brown
BERT (Devlin et al., 2019)	1.33 ±0.01	1.43 ±0.01	1.66 ±0.02
ERNIE (Zhang et al., 2019b)	**1.23** ±0.01	**1.20** ±0.01	1.71 ±0.02
VGCN-BERT (Lu and Nie, 2019)	1.32 ±0.01	1.41 ±0.01	1.75 ±0.02
MG-BERT(base)	1.26 ±0.02	1.45 ±0.01	1.82 ±0.01
MG-BERT	**1.23** ±0.01	1.25 ±0.01	**1.63** ±0.01

Table 3: Perplexity results on CoLA, SST-2, and Brown datasets. The best score is highlighted in bold.

metric in Table 3. In this paper, the perplexity is only calculated on the masked tokens as:

$$PPL = \exp\left(\sum_{i=1}^{n} -\log \hat{y}_i^{[MASK]}\right),$$

where $\hat{y}_i^{[MASK]}$ is the predicted probability of the masked token in the i-th sample. A model with higher perplexity allocates lower probability to the correct masked tokens, which is not desired. The results shown in Table 3 generally demonstrate the fact that both MG-BERT and ERNIE solve the MLM task with more certainty compared to BERT and VGCN-BERT.

We also illustrate some examples of MLM task performed by MG-BERT(base) and BERT in Appendix B. These examples demonstrate that real-world information of knowledge graph and global information of co-occurrence graphs remarkably compensate BERT's shortage.

6 Conclusion

In this paper, we proposed a language representation learning model that enhances BERT by augmenting it with a graph module (i.e. an R-GCN layer over a static multi-graph, including global dependencies between words, and a graph attention layer over a dynamic sentence graph). The static multi-graph utilized in this work consists of a knowledge graph as a source of information about real-world facts and two other graphs built based on word co-occurrences in local windows and documents in the corpus. Therefore, the proposed model utilizes the local context, the corpus-level co-occurence statistics, and the global word dependencies (through incorporating a knowledge graph) to find the input tokens' embeddings. The results generally show the superiority of the proposed model in the Masked Language Modeling task compared to both the BERT model and the recent models employing knowledge or text graphs.

References

Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *CoRR*, abs/1608.00318.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, pages 233–259.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.

2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Gaith Dekhili, Tan Ngoc Le, and Fatiha Sadat. 2019. Augmenting named entity recognition with commonsense knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, page 142, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Jun Feng, Yang Huang, Minlie andd Yang, and Xiaoyan Zhu. 2016. GAKE: Graph aware knowledge embedding. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 641–651, Osaka, Japan. The COLING 2016 Organizing Committee.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Xiangnan He, K. H. Deng, Xiang Wang, Yaliang Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. *ArXiv*, abs/2002.02126.

Annervaz K M, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. 2018. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 313–322, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Zhibin Lu and Jian-Yun Nie. 2019. Raligraph at hasoc 2019: Vgcn-bert: Augmenting bert with graph embedding for offensive language detection. In *FIRE (Working Notes)*, volume 2517 of *CEUR Workshop Proceedings*, pages 221–228. CEUR-WS.org.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2018. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333, New Orleans, Louisiana. Association for Computational Linguistics.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julián Moreno Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification. *ArXiv*, abs/1909.08402.

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 *The Semantic Web - 15th International Conference, ESWC 2018, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 593–607. Springer/Verlag.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *CoRR*, abs/1902.10197.

Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019a. Multi-task feature learning for knowledge graph enhanced recommendation. In *The World Wide Web Conference*, WWW '19, page 2000–2010, New York, NY, USA. Association for Computing Machinery.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019b. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 950–958, New York, NY, USA. Association for Computing Machinery.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019c. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *CoRR*, abs/1911.06136.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *CoRR*, abs/1809.05679.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *ArXiv*, abs/1909.03193.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019a. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3065–3072. AAAI Press.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.