End-to-End Scientific Knowledge Graph Completion via Word Embedding based Entity Type Classification

Qin Dai[1], **Naoya Inoue**[1,2], **Paul Reisert**[2], **Kentaro Inui**[1,2]
[1]Tohoku University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
{daiqin, naoya-i, preisert, inui}@ecei.tohoku.ac.jp

Abstract

Knowledge Graph (KG) such as FreeBase (Bollacker et al., 2008) has been playing an essential role for numerous Natural Language Processing (NLP) tasks. However, the sparsity of existing KG has stimulated the research in Knowledge Graph Completion (KGC), which is proposed to predict the missing knowledge triplet (*h*, *r*, *t*), via projecting KG into a continuous vector space. Most existing methods are only evaluated on general knowledge triplet e.g., (*Tokyo*, *capitalOf*, *Japan*) rather than scientific knowledge triplet e.g., (*hypotension*, *may be treated by*, *dopamine*). Additionally, exiting method merely collects entity type information from existing knowledge base rather than from semantic feature of entity. In this paper, we investigate the effectiveness of some representative KGC models on scientific domain and proposed a new KGC model that predicts missing knowledge triplets by incorporating a word embedding based entity type classification model into a state-of-the-art KGC model. Our experiments on scientific dataset prove the effectiveness of the proposed model on scientific KGC.

1 Introduction

Knowledge Graphs such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015) are extremely crucial for many NLP tasks such as Question Answering (QA), Information Retrieval (IR), Relation Extraction (RE), etc. (Schuhmacher and Ponzetto, 2014). They provide large collections of relations between entities, typically stored as (h, r, t) triplets, where h = *head entity*, r = relation and t = *tail entity*, e.g., (*Tokyo*, *capitalOf*, *Japan*). However, the sparsity of KGs impedes their usefulness in real world applications.

Knowledge Graph Completion (KGC) aims to automatically infer missing facts by examining the latent regularities in existing KG (Trouillon et al., 2016). One of the most promising method for capturing the latent regularities is to embed KG into a low-dimensional continuous vector space and estimate the plausibility of potential knowledge triplets via the vector based algebraic calculation. Recent years, various KG embedding models show excellent performance on KGC, which include TransE (Bordes et al., 2013), TransD (Ji et al., 2015), ComplEx (Trouillon et al., 2016) and SimplE (Kazemi and Poole, 2018), etc. However, most KGC models are merely evaluated by general knowledge, such as (*Turtle Diary*, */film/film/country*, *United States*), rather than scientific knowledge, such as (*pain*, *may be prevented by*, *naproxen*). For investigating their performance on scientific domain and facilitating scientific KGC, in the paper, we select some representative KGC models and assess their performance on scientific KG.

For enhancing KGC, a variety of external information is potentially useful, among which Entity Type (ET) information, such as ET `author` for entity *William Shakespeare*, has been proved to be an effective information for KGC (Xie et al., 2016). However, existing approach merely collects ET information from existing ET knowledge base, rather than applying semantic feature of entity. Distributional representation, typically large corpus trained word embedding, has shown to contain the semantic information about word categories or types (Yaghoobzadeh and Schütze, 2016). For leveraging the effective ET information and simultaneously getting rid of the dependency on an incomplete ET knowledge base, in this work, we proposed an end-to-end KGC model. The proposed model is capable of iden-

tifying ET via the word embedding of target entity and incorporating the predicted ET into a state-of-to-art KGC model to evaluate the plausibility of potential knowledge triplets. Our experiments on scientific KGC show that our end-to-end KGC model has significant improvements compared to all baselines.

2 Base Model for Scientific KGC

We select SimplE (Kazemi and Poole, 2018) as our base scientific KGC model, since it is simple and strong, achieving state-of-the-art performance in general domain. Specifically, suppose we have a KG containing a set of knowledge triplets $\mathcal{O} = \{(e_1, r, e_2)\}$, where each knowledge triplet consists of two entities $e_1, e_2 \in \mathcal{E}$ and their relation $r \in \mathcal{R}$. Here $\mathcal{E}$ and $\mathcal{R}$ stand for the set of entities and relations respectively. SimplE then encodes each entity $e \in \mathcal{E}$ into two vectors h_e, $t_e \in R^d$ and each relation $r \in \mathcal{R}$ into two vectors v_r, $v_{r^{-1}} \in R^d$ respectively, where d is the dimensionality of the embedding space. h_e captures the entity e's behaviour as the *head entity* of a knowledge triplet and t_e captures e's behaviour as the *tail entity* of a knowledge triplet. v_r represents r in triplet (e_1, r, e_2), while $v_{r^{-1}}$ represents its inverse relation r^{-1} in the triplet (e_2, r^{-1}, e_1). The KG scoring function of SimplE for a triplet (e_1, r, e_2) is defined as $\phi(e_1, r, e_2) = \frac{1}{2}(\langle h_{e_1}, v_r, t_{e_2}\rangle + \langle h_{e_2}, v_{r^{-1}}, t_{e_1}\rangle)$. $\langle v, w, x\rangle$ is defined as $\langle u, v, w\rangle = \sum_{n=1}^{d}[u]_n[v]_n[w]_n$, where $[\cdot]_n$ is the n-th entry of a vector. Besides the standard SimplE model, there is a simplified SimplE called SimplE-ignr, where the scoring function is simplified as $\phi(e_1, r, e_2) = \langle h_{e_1}, v_r, t_{e_2}\rangle$.

3 Proposed Model for Scientific KGC

The proposed end-to-end KGC model is illustrated in Figure 1. It includes ET classification part (below) and KGC part (above). In ET classification part, a multi-layer perceptron (MLP) with one hidden layer is applied to identify ET based on word embedding of target entity. In KGC part, *head entity* and *tail entity* along with their predicted ETs and their relation are projected into corresponding KG embeddings, which are then fed to a KG scoring function.

3.1 ET Classification Part

In this work, we use a MLP network to classify ET for *head entity* and *tail entity*. The architecture of our MLP network is as bellow:

$$h = sigmoid(Wx^e) \tag{1}$$

$$y = sigmoid(Vh) \tag{2}$$

where W is a word embedding projection matrix, which is initialized by the pre-trained word embedding that is trained on scientific corpus via Gensim word2vec tool, x^e is a one-hot entity representation, h is the hidden vector, V is ET projection matrix and y is the output vector containing the prediction probabilities of all target ETs. W and V are weight matrices to optimize.

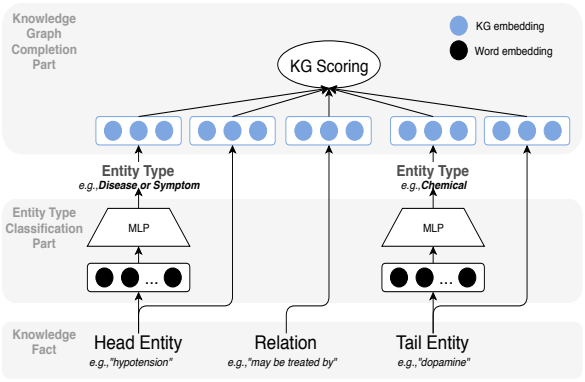

Figure 1: Overview of the proposed end-to-end KGC model

3.2 KGC Part

We extend the base model introduced in Section 2 by incorporating the ET information from the ET classification part. Specifically, given knowledge triplet and predicted ET pair ET_1 (for e_1) and ET_2 (for e_2), the proposed model project them into their corresponding KG embeddings namely h_{e_1}, t_{e_1}, v_r, $v_{r^{-1}}$, h_{e_2}, t_{e_2}, h_{ET_1}, t_{ET_1}, h_{ET_2} and t_{ET_2} respectively, where h_{ET_1} (or t_{ET_1}) represents the KG embedding of ET for e_1 when e_1 acts as the *head entity* (or *tail entity*) in a knowledge triplet. The KG scoring function is defined via Equation 3.

$$\begin{aligned}\phi(e_1, r, e_2) = \frac{1}{4}(&\langle h_{e_1}, v_r, t_{e_2}\rangle \\ &+\langle h_{e_2}, v_{r^{-1}}, t_{e_1}\rangle \\ &+\langle h_{ET_1}, v_r, t_{ET_2}\rangle \\ &+\langle h_{ET_2}, v_{r^{-1}}, t_{ET_1}\rangle)\end{aligned} \tag{3}$$

4 Evaluation

4.1 Data

The scientific KG used in this work is extracted from the Unified Medical Language Sys-

#Entity	#Relation	#Train	#Valid	#Test
27,702	358	89,236	5,089	10,239

Table 1: Statistics of scientific KG in this work.

tem (Humphreys et al., 1998) (UMLS), which includes medical concepts, relations, definitions, etc. As the RE task proposed by (Wang et al., 2014), we only collect the knowledge triplet with RO categories (RO stands for "has Relationship Other than synonymous, narrower, or broader"), which covers the interesting relations like *may treat*, *my prevent*, etc. From the UMLS 2018 release, we extract about 35 thousand entity-ET pairs (e.g., *heart rates*-`Clinical Attribute`) for training ET classification model and about 100 thousand knowledge triplets, which are then randomly divided into training, validation and testing sets for KGC. The statistics of the extracted KG is shown in Table 1. Since we address the KGC in medical domain, we use the MEDLINE corpus to train the domain specific word embedding.

4.2 Setup

To learn the proposed KGC model, we use AdaGrad with mini-batches. We use the default setting of the base KGC model. Specifically, we use the initial learning rate of 0.1, mini-batch size of 1415 and the embedding dimension of 200 for both entity and relation. We sample 1 negative entity for each ground truth entity and generate a labelled batch LB by labelling positive triplets as $+1$ and negatives as -1. We optimize the $L2$ regularized negative log-likelihood of the batch, namely $min_\theta \sum_{((h,r,t),l)\in LB} softplus(-l \cdot \phi(h,r,t)) + \gamma\|\theta\|_2^2$, where θ represents the parameters and $\gamma = 2.0$.

4.3 Result and Discussion

We choose the link prediction task to evaluate the performance of KGC models. In this task, the KGC models predict the missing entity, given an entity and a relation. Specifically, the task predicts *tail entity* t given both *head entity* h and relation r, e.g., $(h, r, *)$, or predict *head entity* h given $(*, r, t)$.

We report the mean reciprocal rank (MRR) for the evaluated models. Table 2 presents the performance of the selected and proposed KGC models. It can be viewed that the proposed KGC model does a good job compared with the existing baselines on the scientific dataset. This proves the effectiveness of the proposed end-to-end architecture for scientific KGC. Specifically, incorporating word embedding based ET classification model is an effective approach to improve the performance of KGC. The table also shows that the KGC models that give state-of-the-art results on general domain, such as ComplEx and SimplE, equally achieve good performance on scientific KGC. This indicates the domain independency of these KGC Models, thus, we could apply them into scientific KGC.

5 Related Work

(Nickel et al., 2011; Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015; Trouillon et al., 2016; Liu et al., 2017; Kazemi and Poole, 2018) propose KG embedding models to calculate plausibility of potential knowledge triplets using structural information from existing knowledge triplets. Aside from the existing knowledge triples, external information is applied to improve KGC. The external information includes surrounding text (Riedel et al., 2013; Wang et al., 2014; Zhao et al., 2015), entity type and relation domain (Guo et al., 2015; Chang et al., 2014; Xie et al., 2016), logical rules (Wang et al., 2015; Rocktäschel et al., 2015) and cross-lingual triples(Klein et al., 2017). However, these methods have not utilized the word embedding based ET classification model for KGC.

6 Conclusion

In this paper, we propose an end-to-end KGC model to address the scientific KGC. The proposed model could classify ET based on word embedding and leverage the ET information to enhance the KGC performance. Our experiments not only indicate the domain independency of the existing KGC models, but also show our model achieves substantial improvements against the state-of-the-art baselines.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1513, Japan and KAKENHI Grant Number 16H06614.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a col-

	MRR		Hit@		
Model	**Raw**	**Filter**	**1**	**3**	**10**
TransE	0.143	0.175	0.117	0.191	0.283
TransD	0.103	0.106	0.068	0.111	0.175
ComplEx	0.117	0.186	0.115	0.220	0.324
SimplE-ignr	0.121	0.186	0.118	0.213	0.324
SimplE (Base model)	0.145	0.229	0.162	0.259	0.356
Proposed model	**0.200**	**0.295**	**0.222**	**0.329**	**0.430**

Table 2: Results on scientific KGC.

laboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Kai-Wei Chang, Scott Wen-tau Yih, Bishan Yang, and Chris Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction.

Shu Guo, Quan Wang, Bin Wang, Lihong Wang, and Li Guo. 2015. Semantically smooth knowledge graph embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 84–94.

Betsy L Humphreys, Donald AB Lindberg, Harold M Schoolman, and G Octo Barnett. 1998. The unified medical language system: an informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 687–696.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs.

Patrick Klein, Simone Paolo Ponzetto, and Goran Glavaš. 2017. Improving neural knowledge base completion with cross-lingual projections. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. *arXiv preprint arXiv:1705.02426*.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1119–1129.

Michael Schuhmacher and Simone Paolo Ponzetto. 2014. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 543–552. ACM.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, pages 1112–1119.

Quan Wang, Bin Wang, Li Guo, et al. 2015. Knowledge base completion using embeddings and rules. In *IJCAI*, pages 1859–1866.

Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation learning of knowledge graphs with hierarchical types. In *IJCAI*, pages 2965–2971.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Corpus-level fine-grained entity typing using contextual information. *arXiv preprint arXiv:1606.07901*.

Yu Zhao, Zhiyuan Liu, and Maosong Sun. 2015. Representation learning for measuring entity relatedness with rich information. In *IJCAI*, pages 1412–1418.