Entity Embedding Analogy for Implicit Link Discovery

Nada Mimouni, Jean-Claude Moissinac, Anh Vu

Entity Embedding Analogy for Implicit Link Discovery

Nada Mimouni, Jean-Claude Moissinac, and Anh Tuan Vu

Telecom ParisTech, Institut Mines Telecom,
46 Rue Barrault, 75013 Paris, France
nada.mimouni@telecom-paristech.fr,
jean-claude.moissinac@telecom-paristech.fr, anh.vu@telecom-paristech.fr
https://www.telecom-paristech.fr/

Abstract. In this work we are interested in the problem of knowledge graphs (KG) incompleteness that we propose to solve by discovering implicit triples using observed ones in the incomplete graph leveraging analogy structures deducted from KG embedding model. We use a language modelling approach that we adapt to entities and relations. The first results show that analogical inferences in the projected vector space is relevant for link prediction task.

1 Introduction

General purpose knoweldge bases (KB) like Yago, Wikidata and DBPedia are valuable background resources for various AI tasks like recommendation, web search [9] and question answering [10]. However, using these resources pose several problems which are mainly related to their large size and highly incompleteness [4]. Many KB completion approaches have been proposed which aim at predicting whether a relationship not in the KG is likely to be true. Recently, vector-space embedding models for KB completion have been extensively studied for their efficiency and scalability and proven to achieve state-of-the-art link prediction performances [6, 2, 1]. An overview of those models with results for link prediction and triple classification is given in [5]. KG embedding models learn distributed representations [8] for entities and relations which are represented as low-dimensional dense vectors or matrices in continuous vector spaces which are intended to preserve the information in the KG. Particularly, in this work we are interested in adapting the language modelling approach proposed by [3] where relational similarities or linguistic regularities between pairs of words are captured. They are represented as translations in the projected vector space where similar words appear close to each other and allow for arithmetic operations on vectors of relations between pairs of words. For instance, this vector translation example: $v(Germany) - v(Berlin) \approx v(France) - v(Paris)$ shows relational similarity between countries and capital cities. It captures the *capital* relationship that we could represent by a translation vector $v(capital)$ such that: $v(France) + v(capital) - v(Paris) \approx 0$. These examples show the analogical

properties between the embedded words expressed by the analogy "*Berlin* is to *Germany* as *Paris* is to *France*". We propose to apply this property to entities and relations in KGs as represented by diagrams (a) and (b) in figure 1. We use the analogical property for KB completion and show that it is particularly relevant for this task. Our intuition is illustrated by diagrams (b) and (c) in figure 1 where an unobserved triple can be inferred by mirroring its counterpart in the parallelogram. The steps of our approach are described in section 2 and first results are given in section 3.

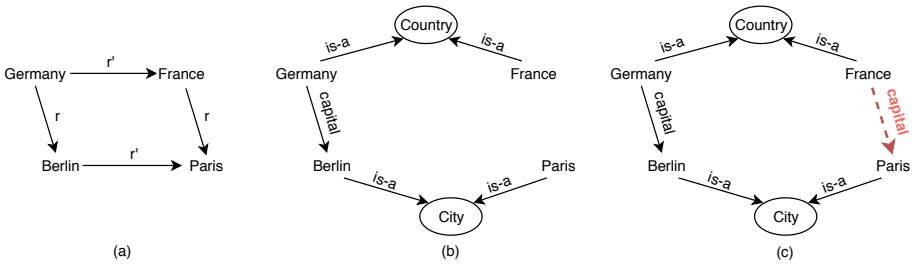

Fig. 1. (a) Analogy relation diagram (parallelogram) between countries and capital cities. In (b) and (c) KGs, r corresponds to the relation *capital* and r' is decomposed into two type relations (*is-a*) to concepts *Country* and *City*.

2 Approach For KB Completion

First we adapt the language modelling approach to KG embedding [1]. We transform the entities and relations in the graph as paths that are considered as sequences of words in natural language. To extract RDF graph sub-structures, we use breath-first algorithm to get all the graph walks and random walks for a limited number N. Let $G = (V, E)$ be an RDF graph where V is the set of vertices and E is the set of directed edges. For each vertex v we generate *all* or N graph walks P_v of depth d rooted in the vertex v by exploring direct outgoing and ongoing edges of v and iteratively direct edges of its neighbours v_i until depth d is reached. The paths after the first iteration follow this pattern $v \rightarrow e_i \rightarrow v_i$ where $e_i \in E$. The final set of sequences for G is the union of the sequences of all the vertices $\bigcup_{v \in V} P_v$.

Second, we train a neural language model which estimates the likelihood of a sequence of entities and relations appearing in the graph and represents them as vectors of latent numerical features. We use the CBOW (continuous bag of words) and Skip-Gram models as described in [3]. CBOW predicts target words w_t from context words within a context window c while Skip-Gram does

[1] A similar approach is described in [7] although we differ in many details that we do not discuss here due to space constraints.

the inverse and tries to predict the context words from the target word. The probability $p(w_t|w_{t-c}...w_{t+c})$ is calculated using the $softmax$ function.

Last, we extract analogical properties from the features space to estimate the existence of new relations between entities. We use the following arithmetic operation on the features vectors (entities of figure 1): $v(Germany) + v(France) - v(Berlin) = v(x)$ that we consider it is solved correctly if $v(x) \approx v(Paris)$. In the left part of the equation, entities with the same type are affected the same sign (+ or -), e.g. *Germany* and *France* have the same type *Country* and are both positive, *Berlin* of different type *City* is negative. The right part of the equation contains the missing corner of the diagram to be predicted. We use cosine similarity measure between the resulting vector $v(x)$ and vectors of all other entities of the same type in the embedding space (discarding original ones in the equation) to rank the results.

3 Experiments and Results

We experiment our approach on a sub-graph of DBpedia representing a target domain, here we worked on museums of Paris. We propose to handle scalability issue by contextualizing the input graphs assuming that more relevant information is centralized within a perimeter of α hops around main entities of this domain (α could be fixed at 2 or 3). We build our KG as the union of individual contextual graphs of all entities representing the input data from the cultural institution *Paris Musées*. The entities are identified on DBpedia.fr and an entity resolution task is performed by exact or approximate label matching [2]. The final graph contains 448309 entities, 2285 relations and 5122879 triples. To generate sequences of entities and relations we use random graph walks with $N = 1000$ for depth $d = 4, 6, 8$ and $N = 2000$ for $d = 4$. We also consider for each entity all walks of depth $d = 2$ (all direct neighbours). We train the Skip-Gram word2vec model on the corpus of sequences with the following parameters: window size = 5, number of iterations = 10, negative samples = 25 (for optimisation) and dimension of entities' vectors = 200. To test the approach we manually built a ground-truth for analogy between entities in the KG (museums, artists, places). Each entry corresponds to a parallelogram as described in figure 1 with one unobserved triple in the KG. We complete the missing triples either by observing the open KB *Joconde* or manually by reading the corresponding entity description in DBPedia or the text article in Wikipedia. Table 1 shows results for top most similar entities of the predicted vectors, filtered by type *Person* and the relation *presentedIn* between an artist and a museum, on a subset of representative entries from the ground-truth with parameters $N = 1000$ and $d = 6$. Most of these results are returned in up to third position of the ranked list resulting in good values of Mean Reciprocal Rank (MRR) = 0.72 and Hits@3 close to 1 which makes our approach performing well for the link discovery and KB completion task.

[2] In the following, we denote the namespace `http://fr.dbpedia.org/resource/entity` shortly as *dbr* : *entity*.

Table 1. Top most similar entities for predicted vectors with analogy property of relation *presentedIn* between an artist and a museum.

Positive	Negative	Prediction	Rank
$dbr:Musee\_Zadkine$ $dbr:Musee\_Bourdelle$	$dbr:Ossip\_Zadkine$	$dbr:Antoine\_Bourdelle$	$\frac{1}{10}$
$dbr:Musee\_Bourdelle$ $Maison\_de\_Victor\_Hugo$	$dbr:Antoine\_Bourdelle$	$dbr:Victor\_Hugo$	$\frac{1}{10}$
$dbr:Musee\_Bourdelle$ $dbr:Musee\_CognacqJay$	$dbr:Antoine\_Bourdelle$	$dbr:Rembrandt$ $dbr:Gustave\_Courbet$	$\frac{2}{10}$ $\frac{3}{10}$
$dbr:Musee\_Zadkine$ $dbr:Musee\_Carnavalet$	$dbr:Ossip\_Zadkine$	$dbr:Gustave\_Courbet$ $dbr:Jacob\_van\_Ruisdael$	$\frac{1}{10}$ $\frac{2}{10}$

4 Conslusion and Future Work

We presented an approach for link discovery in KBs based on the neural language embedding of RDF graphs and leveraging analogical structures extracted from relational similarities which could be used to infer new unobserved triples from observed ones. The test of our approach on a domain-specific ground-truth shows promising results. We are extending experiments to comparison with state-of-the-art approaches for KB completion on the standard baselines.

References

1. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS (2013)
2. Liu, H., Wu, Y., Yang, Y.: Analogical inference for multi-relational embeddings. In: Proceedings of ICML. pp. 2168–2178 (2017)
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
4. Min, B., Grishman, R., Wan, L., Wang, C., Gondek, D.: Distant supervision for relation extraction with an incomplete knowledge base. In: Proceedings of NAACL-HLT. pp. 777–782 (2013)
5. Nguyen, D.Q.: An overview of embedding models of entities and relationships for knowledge base completion. CoRR **abs/1703.08098** (2017)
6. Nickel, M., Rosasco, L., Poggio, T.A.: Holographic embeddings of knowledge graphs. In: AAAI (2016)
7. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: Proceedings of ISWC. p. 498 – 514 (2016)
8. Rumelhart, D.E., McClelland, J.L. (eds.): Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. MIT Press (1986)
9. Szumlanski, S., Gomez, F.: Automatically acquiring a semantic network of related concepts. In: Proceedings of the 19th ACM CIKM. pp. 19–28 (2010)
10. Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., Li, X.: Neural generative question answering. In: Proceedings of IJCAI. pp. 2972–2978. AAAI Press (2016)