Exposé to Master's Thesis:

Exploiting Language Models for Entity Alignment in Knowledge Graphs

by

Radina Sofronova

Degree Course: Informatics M.Sc.

Matriculation Number: 1960185

Institute for Program Structures and Data Organization (IPD),

Department of Informatics

Institute of Applied Informatics and Formal Description Methods (AIFB), Department of Economics and Management

Advisor:	Prof. Dr. Ralf H. Reussner
Second Advisor:	Prof. Dr. Harald Sack
Supervisors:	M. Sc. Russa Biswas, Dr. Mehwish Alam

1 Motivation

Knowledge graphs (KGs) represent real-world information about different domains in the form of triples consisting of a subject, a relation and an object. The subject and object are entities which refer to unique objects in the real world and are linked by a relation, e.g. *<dbr:Leonardo_da_Vinci, dbo:education, dbr:Andrea_del_Verrocchio>*. KGs support a variety of machine learning and Natural Language Processing (NLP) based applications. Those often require diverse knowledge which is not present in a single KG.

Most KGs are unlikely to reach full coverage of a domain since they are usually created independently for particular purposes. Furthermore, they are generated differently, e.g. extracted automatically from structured data or human-curated. A way to increase KG completeness is to integrate multiple KGs, regardless whether they are large-scale such as DBpedia [1], Wikidata [2] and YAGO [3] or domain-specific. This is a challenging task since the data in different KGs is modeled to specific needs and individual naming conventions are applied. The same entity may exist in different forms in different KGs. For example, if we consider the entity *dbr:Leonardo da Vinci* in DBpedia it corresponds to *wikidata:Q762* in Wikidata.

The task of **entity alignment** aims to find entities in two KGs that represent the same real-world entity. The property *owl:sameAs* links between identical items across different KGs need to be identified. *owl:sameAs* statements are often used in defining mappings between ontologies.

Currently, increasing attention has been paid to the utilization of **KG embedding-based (KGE) models** for tackling the entity alignment task. KGE models solely depend on the facts in the KG and catch the structure information. From this perspective, KGE models are discarding a valuable source of information. In order to extend the KGE models, external information, e.g. textual information, can be utilized. [4] consider how this textual information can help overcome the limitation. Moreover, **language models (LMs)** have become crucial for achieving state-of-the-art performance in natural language processing tasks. Therefore, we exploit the integration of text-based and structure-based KG representations.

In this thesis, we propose a novel transfer learning based approach that exploits LMs to obtain the latent representations of the entities and relations in a KG, followed by a word-word alignment model which aligns the same entities across different KGs.

2 Goal

Since most KGs are developed independently and many of them are supplementary in contents, one of core challenges of KGs is to align equivalent entity pairs between different KGs. An entity alignment model aims to embed two KGs into a unified vector space by pushing the seed alignments of entities together. The objective of this work is to explore if an unsupervised word embedding alignment model can be modified for the task of entity alignment in KGs.

The main research questions of this thesis are as follows:

- **RQ1**: Can we use a word-word alignment model for entity alignment in KGs?
- **RQ2**: Do static LM-based KGEs improve entity alignment?
- **RQ3**: Do contextual LM-based KGEs improve entity alignment?

3 Related work

In recent years, due to the development of KG representation learning (RL) methods, entity alignment methods based on RL gained much attention. They first represent the entities and relations of each KG as low-dimensional vectors and then discover the mappings between different embedding spaces by calculating the similarity between entity vectors. The existing methods can be divided into two categories, the **semantic matching-based models** and the **graph neural network (GNN)-based models**.

The semantic matching-based models exploit the semantics of each entity to learn its low-dimensional vector representation. Examples of entity alignment models based on semantic matching are MTransE [5], PTransE [6], IPTransE [7]. Those methods use a KGE model, the translational model TransE [8], to learn the vector space of a single KG and then an alignment model learns a mapping for both entities and relations across different embedding spaces to get a unified vector space. BootEA [9] is another TransE-based model that tackles the problem of insufficient existing pre-aligned entity pairs as training data by iteratively selecting possible entity pairs for training. JAPE [10] uses TransE to represent entities and also learns attribute representations, since it is based on the assumption that entities with similar attributes have a greater probability of being equivalent.

In contrast, the GNN-based entity alignment models use a GNN to learn low-dimensional vector representations of different entities based on the graph structure of the KG. GCN-Align [11] is the first usage of GNN for entity alignment which utilizes a graph convolutional network (GCN) to embed entities of each KG into a unified vector space without

the prior knowledge of relations. GMNN [12] introduces sub-graph alignment. HMAN [13] applies GCNs to combine information of entities, relations, and attributes to learn entity embeddings.

These approaches require seed entity pairs to project entity embeddings from different KGs into a unified space. However, such labeled data are hard to obtain in practical settings. This thesis proposes an entity alignment model for heterogeneous KGs based on an unsupervised approach, which does not require pre-aligned entities or triples as seeds for training.

4 Methodology

At present, entity alignment based on representation learning (RL) has achieved promising results. However, the existing models rely on pre-aligned entity pairs to a large extent. In this work, we investigate whether it is possible to align two KGs without parallel data.

4.1 Preliminaries

KG embedding models.

TransE [8] is a translational model which embeds entities and relations in the same embedding space and is created with the learning assumption $h + r \approx t$, given a triple (h, r, t). Therefore, the structural information of entities is preserved and entities that share similar neighbors have close representations in the embedding space.

DistMult [14] is a semantic matching model which uses a bilinear transformation to calculate the reasonableness score of the fact triples. It constraints the relationship matrix to be diagonal, which enhances the generalization ability of the model.

ConvE [15] is a NN-based model which uses a 2D Convolutional Neural Network (CNN). It concatenates the embedding of the head entity and the relation into an input matrix which is fed into the convolution layer. The output is mapped into a vector, which is dot-producted with the embedding of the tail entity to return the reasonableness score of the triple.

R-GCN [16] is a NN-based model which introduces a relationship matrix in Graph Convolutional Networks (GCNs) as the mapping transformation while the entity merges neighbor features.

Language models.

Word2Vec [17] aims to learn the distributed representation for words reducing the high dimensional word representations in large corpus. It comprises of two model architectures,

Continuous Bag of Words (CBOW) and Skip-gram.
GloVe [18] is a word embedding model which exploits the global word-word co-occurrence statistics in the corpus. The main idea is that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning.
KG-BERT [19] is a language modeling method for KG completion using a pre-trained contextual language model BERT [20]. Triples are treated as textual sequences and BERT is fine-tuned on these sequences for predicting the plausibility of a triple or a relation.

Alignment model.
MUSE [21] is an unsupervised multilingual word embedding alignment model. Starting from a simple unsupervised word-by-word translation model, the model is iteratively improved based on a reconstruction loss, and using a discriminator to align latent distributions of both the source and the target languages.

4.2 The Proposed Approach

We first explore MUSE [21] which is an unsupervised machine translation approach that aligns word embedding spaces based on unaligned datasets of each language. Having two sets of embeddings trained independently on monolingual data, the model builds a common latent space between the two domains. It learns a mapping between the two sets such that similar words are close in the shared space and translates by reconstructing in both domains. MUSE learns a rotation matrix which roughly aligns the two distributions of word embeddings.

Similarly, our goal is to align KG embedding spaces. We first train two KGs that need to be aligned with one another with the same KGE model (see 4.1) with same hyperparameters such that the embedding spaces generated for both the KGs exhibit similar characteristics. Then, using the strategy of MUSE, our model builds a common latent space between the two embedding spaces. We learn a rotation matrix W by a two-player game with a generator and a discriminator where the discriminator aims at maximizing its ability to identify the origin of an embedding, and the rotation matrix W aims at preventing the discriminator from making accurate predictions. The result is a common latent space where same as well as similar entities from the two KGs are represented close to each other.

Furthermore, we would like to explore if extending the KGE models with external knowledge creates a positive impact. KGE models have proven that they can capture the structure information of the KG, but the fusion of other significant information except fact triples can enhance the KG representation. Since we are modifying a language model based alignment method for the task of entity alignment, we would utilize textual infor-

mation in the KG embedding. LMs (see 4.1) are used to generate word embeddings which will serve as pre-initialization of the KGE models.

5 Expected results

The **main desired result** is that the same or related entities from different KGs are aligned. This will be achieved by adapting MUSE [21], an unsupervised word embedding alignment model, for the task of entity alignment in KGs. Following previous works, we adopt three frequently utilized **EA datasets** [22]: (1) DBP15K, which includes three crosslingual KG pairs extracted from DBpedia; (2) SRPRS, which comprises two cross-lingual and two mono-lingual KG pairs extracted from DBpedia, Wikidata and YAGO; and (3) DWY100K which comprises two mono-lingual KG pairs extracted from DBpedia, Wikidata and YAGO. As **evaluation metrics**, following convention, Mean Reciprocal Rank (MRR) and Hits@N will be used. MRR is the rank of correctly aligned entities, and Hits@N is the proportion of correctly aligned entities whose rank is not greater than N. The higher the Hits@N and MRR, the better the performance. Our results will be compared with other existing state-of-the-art entity alignment methods.

6 Schedule

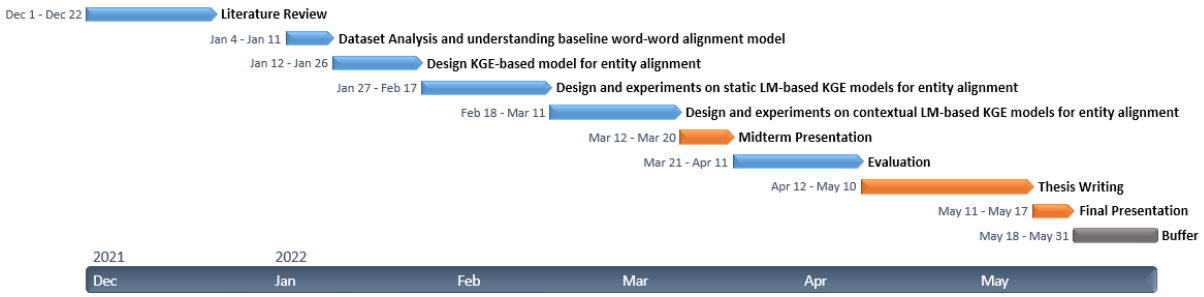

Figure 1: Planned schedule for thesis tasks

References

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al., Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web, vol. 6 (2015). doi:10.3233/sw-140134.

[2] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85. doi:10.1145/2629489.

[3] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: A large ontology from wikipedia and wordnet (2008).

[4] D. Daza, M. Cochez, P. Groth, Inductive entity representations from text via link prediction, CoRR abs/2010.03496 (2020). arXiv:2010.03496.
URL https://arxiv.org/abs/2010.03496

[5] M. Chen, Y. Tian, M. Yang, C. Zaniolo, Multilingual knowledge graph embeddings for cross-lingual knowledge alignment (2017). arXiv:1611.03954.

[6] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, S. Liu, Modeling relation paths for representation learning of knowledge bases, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 705–714. doi:10.18653/v1/D15-1082.
URL https://aclanthology.org/D15-1082

[7] H. Zhu, R. Xie, Z. Liu, M. Sun, Iterative entity alignment via joint knowledge embeddings, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 4258–4264. doi:10.24963/ijcai.2017/595.
URL https://doi.org/10.24963/ijcai.2017/595

[8] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 26, Curran Associates, Inc., 2013, p. 2787–2795.
URL https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

[9] Z. Sun, W. Hu, Q. Zhang, Y. Qu, Bootstrapping entity alignment with knowledge graph embedding, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 4396–4402. doi:10.24963/ijcai.2018/611.
URL https://doi.org/10.24963/ijcai.2018/611

[10] Z. Sun, W. Hu, C. Li, Cross-lingual entity alignment via joint attribute-preserving embedding, in: C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web – ISWC 2017, Springer International Publishing, Cham, 2017, pp. 628–644.

[11] Z. Wang, Q. Lv, X. Lan, Y. Zhang, Cross-lingual knowledge graph alignment via graph convolutional networks, in: Proceedings of the 2018 Conference on Empirical

Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 349–357. doi:10.18653/v1/D18-1032.
URL https://aclanthology.org/D18-1032

[12] K. Xu, L. Wang, M. Yu, Y. Feng, Y. Song, Z. Wang, D. Yu, Cross-lingual knowledge graph alignment via graph matching neural network, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3156–3161. doi:10.18653/v1/P19-1304.
URL https://aclanthology.org/P19-1304

[13] H.-W. Yang, Y. Zou, P. Shi, W. Lu, J. Lin, X. Sun, Aligning cross-lingual entities with multi-aspect information, in: Proc. of EMNLP, 2019, pp. 4431–4441.

[14] B. Yang, W. tau Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases (2015). arXiv:1412.6575.

[15] T. Dettmers, M. Pasquale, S. Pontus, S. Riedel, Convolutional 2d knowledge graph embeddings, in: Proceedings of the 32th AAAI Conference on Artificial Intelligence, 2018, pp. 1811–1818.
URL https://arxiv.org/abs/1707.01476

[16] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks (2017). arXiv:1703.06103.

[17] T. Mikolov, G. Inc, I. Sutskever, G. Inc, K. Chen, G. Inc, G. Corrado, G. Inc, J. Dean, Distributed representations of words and phrases and their compositionality, in: In NIPS, 2013, pp. 3111–3119.

[18] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation (2014).

[19] L. Yao, C. Mao, Y. Luo, Kg-bert: Bert for knowledge graph completion (2019). arXiv:1909.03193.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[21] G. Lample, A. Conneau, L. Denoyer, M. Ranzato, Unsupervised machine translation using monolingual corpora only (2018). arXiv:1711.00043.

[22] X. Zhao, W. Zeng, J. Tang, W. Wang, F. Suchanek, An experimental study of state-of-the-art entity alignment approaches, IEEE Transactions on Knowledge and Data Engineering (2020) 1–1doi:10.1109/TKDE.2020.3018741.