# Unsupervised Entity Alignment Model via Optimal Transport

Xin Zhao Takuo Hamaguchi* Masashi Shimbo† Yuji Matsumoto
Nara Institute of Science and Technology
{zhao.xin.zv1, takuo-h, shimbo, matsu}@is.naist.jp

## Abstract

The task of entity alignment aims at finding corresponding entities that have the same real-world semantics across different knowledge graphs. Recently, embedding-based methods have been proposed for this task. Such models assume that a large number of aligned entities are known. However, in real-world dataset, such supervised data are difficult to obtain. To tackle this problem, we explore unsupervised methods by modeling entity alignment as an optimal transport problem and propose a model using Gromov-Wasserstein distance. In our experiment, we demonstrate that our model can achieve good performance.

## 1 Introduction

Cross-lingual knowledge graphs play a key role in cross-lingual NLP application. However, it is difficult to use cross-lingual knowledge graphs (KGs) without a known corresponding entity-relationship. Finding aligned entities across different knowledge graphs is a key technology to utilize cross-lingual KGs.

The supervised knowledge graph alignment task aims to use information in knowledge graphs themselves and the alignment information of some entities, to obtain unknown aligned entities. In recent years, embedding-based entity alignment methods have been proposed [Sun, 2017, Zhu2017, Chen2016, Sun2018, Trivedi2018]. They use the alignment information between knowledge graphs to jointly learn entity embedding and the mapping function between them so that the learned mapping function can be applied to discover unknown aligned entities.

These existing entity alignment models assume a large number of aligned entities are available. However, it is difficult to obtain aligned entity pairs in the actual scene. Thus, it is meaningful to discover corresponding entities from the graph structure. In this paper, we seek for the unsupervised setting of entity alignment task.

We regard the entity alignment task as an optimal transport (OT) problem and propose a method based on Gromov-Wasserstein distance [Mémoli2011]. OT is a general mathematical toolbox used to evaluate correspondence-based distances and establish mappings between probability distributions. It is also generalized applied in domain adoption applications. [Alvarez-Melis, 2018] show that OT model demonstrates its superiority on unsupervised cross-lingual word alignment.

While OT works well on word alignment, it cannot be applied directly to multi-channel data like KGs. Thus, we propose a multi-channel optimal transport framework for entity alignment. We propose two models, one assuming the information of aligned relations, and the other assuming no such information. Due to the limited computational resources, we performed experiments on a subset of the entities in KGs. The experiments proved the effectiveness with and without known aligned relations.

## 2 Proposed method

### 2.1 Problem Formulation

In this section, we firstly introduce the notation and problem definition, then describe the proposed multi-channel optimal transport model for entity alignment in the multi-relational graph.

**Knowledge Graph Alignment** In the unsupervised setting of knowledge graph alignment, we consider a source knowledge graph $\mathcal{G}_s$ and a target knowledge graph $\mathcal{G}_t$ with their own entities $\mathcal{E}_s, \mathcal{E}_t$ and relations $\mathcal{R}_s, \mathcal{R}_t$, respectively, where $\mathcal{E}_s = \{e_i^s\}$, $\mathcal{E}_t = \{e_j^t\}$ and $\mathcal{R}_s = \{r_i^s\}, \mathcal{R}_t = \{r_j^t\}$, each entity and relation $e, r$ have their corresponding low dimension vector $\mathbf{e} \in \mathbb{R}^{m\times 1}, \mathbf{r} \in \mathbb{R}^{n\times 1}$. To simplify notation, we assume that the embedding of all entities and relations share the same dimension $d$. Thus, source and target embedding space can be represented by

*Currently with The Institute of Statistical Mathematics.
†Currently with STAIR Lab, Chiba Institute of Technology.

$\{E^s \in \mathbb{R}^{|\mathcal{E}_s|\times d}, R^s \in \mathbb{R}^{|\mathcal{R}_s|\times d}\}, \{E^t \in \mathbb{R}^{|\mathcal{E}_t|\times d}, R^t \in \mathbb{R}^{|\mathcal{R}_t|\times d}\}$ respectively.

The problem we need to tackle is finding a mapping function from entities in source KG to target KG:

$$P : \mathcal{E}_s \to \mathcal{E}_t \tag{1}$$

## 2.2 Multi-channel Optimal Transport model for entity alignment

In this section, we introduce the method to find the alignment between the entities of the two graphs based on Gromov-Wasserstein distance. The whole algorithm is shown in Alg. 1. The intention and notation of each line are introduced in the following paragraphs.

To utilize Gromov-Wasserstein distance, we first need to define the distance function between the entities of $\mathcal{G}_s$ and $\mathcal{G}_t$ individually. However, due to the existence of multi-relations in KGs, the distance between two entities cannot be calculated directly. To solve the problem, we regard each relation in KGs as a channel and apply an embedding-based distance. We firstly embedding graphs into low-dimensional spaces and acquire entity embedding $\mathbf{E}_s, \mathbf{E}_t$ and relation embedding $\mathbf{R}_s, \mathbf{R}_t$ in line 1. Then, to capture similarity of entities according to all the relations, we measure all distance of possible triplets as 3-way tensors $d_X, d_Y$ in line 2, where $d_X^{ijk}$ determine the distance of triplet $(e_i, r_k, e_j)$.

Then, as shown in line 3, we apply distance tensors to our multi-channel optimal transport model and obtain $\pi$, which $\pi \in \Gamma$ and $\Gamma$ contains all possible admissible couplings between two knowledge graphs.

$$\Gamma = \{\pi \in \mathbb{R}_+^{|\mathcal{E}_s|\times|\mathcal{E}_t|} \text{ s.t.} \overset{\text{s.t.}}{\sum_i} \pi_{i,j} = p_i, \overset{\text{s.t.}}{\sum_j} \pi_{i,j} = q\} \tag{2}$$

where $\mathbb{R}_+$ stands for the set of non-negative real numbers, $\pi$ is the doubly stochastic matrix measuring the transition probability of entities in two KGs and $p, q$ are the vectors of probability weights associated with each entity. Thus $p, q$ must satisfy the equation:

$$\sum_i p_i = 1, \quad \sum_j p_j = 1 \tag{3}$$

Since calculating distance between entities in $\mathcal{G}_s$ and $\mathcal{G}_t$ directly are impossible, we need to calculate the cross-graph distance function by measuring all possible entity pairwise distance $D\big((i,k),(j,l)\big)$. $D\big((i,k),(j,l)\big)$ are distance between entities $i, k$ $j, l$ in two graphs, it also can be understood as the cost of matching entity pair from $\mathcal{E}_s$ to $\mathcal{E}_t$.

Finally, we utilize $\pi$ acquired by minimizing (4) defined as below, we can obtain best-matched entity pairs in line 4. The detail of each step is explained in the following sections.

$$\pi = \operatorname*{argmin}_{\pi\in\Gamma} \mathcal{L}(d_X, d_Y, \pi) \tag{4}$$

where

$$\mathcal{L}(d_X, d_Y, \pi) = \sum_{ij} \pi_{ij} \sum_{kl} \pi_{kl} \cdot D(d_X^{ik}, d_Y^{jl}) \tag{5}$$

**Algorithm 1** Unsupervised entity alignment with Gromov-wasserstein

**Input**: Two knowledge graphs
$\mathcal{G}_s = \{\mathcal{E}_s, \mathcal{R}_s\}, \mathcal{G}_t = \{\mathcal{E}_t, \mathcal{R}_t\}$
**Output**: aligned entity pairs $\{(e_i^s, e_j^t)\}$

1: $\mathbf{E}_s, \mathbf{R}_s \leftarrow \mathrm{KGE}(\mathcal{G}_s)$, $\quad \mathbf{E}_t, \mathbf{R}_t \leftarrow \mathrm{KGE}(\mathcal{G}_t)$
2: $d_X \leftarrow \mathrm{Dist}(\mathbf{E}_s, \mathbf{R}_s)$, $\quad d_Y \leftarrow \mathrm{Dist}(\mathbf{E}_t, \mathbf{R}_t)$
3: $\pi \leftarrow \operatorname*{argmin}_{\pi\in\Gamma(x,y)} \mathcal{L}(d_X, d_Y, \pi)$
4: Obtain $\{(e_i^1, e_j^2)\}$ from $\pi$

### 2.2.1 KGs Embedding Models

We performed our experiments on TransE [Bordes2013], one of the most popular KGE models. To improve the quality of the calculated distance, inspired by [Sun, 2019] and [Sun2018], we adopt self-adversarial negative sampling and loss function sensitive to negative samples.

**Self-adversarial negative sampling** is a negative sampling strategy for providing a robust negative sample during the training process. As shown in [Sun, 2019], all candidate entities are sampled according to the following distribution, $\alpha$ is the hyper-parameter.

$$p(h_i', r, t_i | (h, r, t)) = \frac{\exp \alpha \cdot f(h', r, t')}{\sum_i \exp \alpha \cdot f(h_i', r, t_i')} \tag{6}$$

**Sigmoid loss function** [Sun2018] noted that margin-based loss can not make the scores of positives triples lower than some of negative ones. Thus, we adopt sigmoid function to make the scores of negative samples and positive examples as large as possible.

$$\begin{aligned} L = &- \log \sigma(\lambda - f(h, r, t)) \\ &- \sum_{i=1}^{n} p(h_i', r, t_i') \log \sigma(f(h', r, t') - \lambda) \end{aligned} \tag{7}$$

where $\lambda$ is the margin value, the negative samples are generated by replacing the head or tail entity according to (6).

#### 2.2.2 Distance tensor calculation

After obtaining the embedding of the two knowledge graphs, we use the distance function defined by KGE model $f(h,r,t)$ to measure the distances of all possible triplets.

$$d_X \in \mathbb{R}^{|\mathcal{E}_s|\times|\mathcal{E}_s|\times|\mathcal{R}_s|} \tag{8}$$

where for any arbitrary $(e_i, r_m, e_k)$, $d_X^{ikm} = f(e_i, r_m, e_k)$. The same for $d_Y$.

Further, we adopt the ensemble method to eliminate the impact of the initial value on the knowledge graph embedding and improve robustness. Firstly, we learn several sets of KG embedding with different initial values. Then, after calculating the distance tensors separately, we use the element-wise average of them as our final distance tensor.

#### 2.2.3 Multi-Channel Optimal Transport

In many cross-knowledge graphs, correspondence between some relations is available. In order to make fuller use of the information we have, we utilize the known aligned relations to help to learn. Also, there are still many knowledge graphs that do not have a shared relation. We discuss the situation with and without aligned relations in the following paragraphs.

**With aligned relations** We denote the set of known aligned relation pairs as $\mathrm{A} = \{(a_s, a_t)\}$, where $a_s$ and $a_t$ stand for same relation. In this setting, we denote distance function across graph as below:

$$D\big((i,k),(j,l)\big) = \sum_{(m,n)\in \mathrm{R}} ||d_X^{ikm} - d_Y^{jln}||^2 \tag{9}$$

**Without aligned relations** In the optimization without any aligned relations, we simply perform a mean pooling operation on the relation-axis (3rd-axis) of the distance tensor during optimization. However, the distance of the all possible triplets cannot be perfectly modeled according to the score function $f$. For some triplets which entities and relations have low correlation, their distance lacks accuracy when used as a feature in Gromov-Wasserstein optimization. To reduce the impact of such triples, as shown in (10), we rescale distance tensor $d_X, d_Y$ before the mean pooling operation.

$$D\big((i,k),(j,l)\big) = ||P\big(R(d_X^{ikn})\big)) - P\big(R(d_Y^{jl})\big)||^2 \tag{10}$$

where

$$R(d) = \exp(-\alpha \cdot d), \quad P(d) = \frac{-\log(\sum_m d^{ikm})}{\alpha} \tag{11}$$

Finally, for the optimization, we adopt optimization method for Gromov-Wasserstein distance proposed by [Peyré2016], which greatly reduce the complexity of the algorithm.

# 3 Experiments

In this section, we seek to: (1) Introduce experiment setup (2) Perform an experiment on datasets and evaluate their performance on benchmark entity alignment task.

## 3.1 Datasets

In our experiment, we use two types-six datasets to verify the performance of our models. Three of them are dummy datasets and others are real dataset. We obtain the three dummy datasets by segmenting FB15K into two sub-graphs with different overlap rates. The definition of overlap rate are shown below:

$$\#\text{overlap\_rate} = \frac{|T_1 \cap T_2|}{|T_1|} = \frac{|T_1 \cap T_2|}{|T_2|} \tag{12}$$

where $T_1, T_2$ are triplets set of source graph and target graph.

We also use DBP15K - a small version of multilingual KB DBpedia which includes (En $\leftrightarrow$ Ja), (En $\leftrightarrow$ Fr), (En $\leftrightarrow$ Zh) three subsets. The statistics are summarized in Table 3.1 and Table 3.2.

Due to the limited computation resource, we train KG embedding on all entities and relation yet only use embedding of top frequency(1K entities and 500 relations) for alignment. After obtaining transformation matrix $\pi$, we find the best matching entity for $e_s^i$ in source graph by $e_t^j = \underset{e_t^j \in \mathcal{E}_t}{\text{argmin}}\ \pi_{ij}$. Otherwise, for each run, we report two common metrics, hits at 1/10 (hit@1 and hit@10).

Table 1: Statistics of the Dummy Datasets.

| Dataset | FB15k-80 | | FB15k-50 | | FB15k-20 | |
|---|---|---|---|---|---|---|
| | src | tgt | src | tgt | src | tgt |
| #Entities | 14951 | 14927 | 14906 | 14911 | 14875 | 14867 |
| #Relations | 1289 | 1262 | 1260 | 1235 | 1242 | 1210 |
| #Overlap_rate | 0.8 | | 0.5 | | 0.2 | |
| #Shared_Triplets | 394808 | | 197404 | | 65801 | |
| #Shared_Entities | 14927 | | 14866 | | 14791 | |
| #Shared_Relations | 1262 | | 1199 | | 1146 | |

Table 2: Statistics of the Cross-lingual KGs.

| Dataset | DBP15K(Ja $\leftrightarrow$ En) | | DBP15K(Fr $\leftrightarrow$ En) | | DBP15K(Zh $\leftrightarrow$ En) | |
|---|---|---|---|---|---|---|
| | src | tgt | src | tgt | src | tgt |
| #Entities | 19814 | 19780 | 19661 | 19993 | 19388 | 19572 |
| #Relations | 1701 | 1323 | 903 | 1208 | 1701 | 1323 |
| #Shared_Entities | 15000 | | 15000 | | 15000 | |
| #Shared_Relations | 582 | | 75 | | 891 | |

## 3.2 Results

The evaluation results are presented in Table 3,4. The algorithm column stands for model settings. For example, 'TransE/w-rel/single' means that the model is using KG embedding trained by TransE, the second part indicate whether to use aligned relations obtained by string matching, which 'w-rel' stands for using and 'n-rel' means non-using. Finally, the third part denotes whether to use ensemble strategy. In our experiment, we use 6 groups of embedding to produce ensemble distance tensor.

We report experiment on six models for cross-lingual datasets as in 4, and only perform experiment on models without aligned relations for dummy datasets, which are shown in 3. Due to our limitation on the number of alignable entities, there is no prior research that can participate in comparison, we only report our own experimental results. From these evaluation results, we have the following findings:

Table 3: Results on Dummy Dataset.

| Algorithm | FB15k-80 | | | FB15k-50 | | | FB15k-20 | | |
|---|---|---|---|---|---|---|---|---|---|
| | H@1 | H@5 | H@10 | H@1 | H@5 | H@10 | H@1 | H@5 | H@10 |
| TransE/n-rel/single | 82.94 | 94.62 | 97.93 | 64.86 | 88.07 | 93.29 | 46.76 | 72.56 | 82.11 |
| TransE/n-rel/ensemble | **90.07** | **96.38** | **98.97** | **74.12** | **89.35** | **94.14** | **55.76** | **76.95** | **84.63** |

Table 4: Results on Cross-Lingual Dataset.

| Algorithm | DBP15k(Ja $\leftrightarrow$ En) | | | DBP15k(Fr $\leftrightarrow$ En) | | | DBP15k(Zh $\leftrightarrow$ En) | | |
|---|---|---|---|---|---|---|---|---|---|
| | H@1 | H@5 | H@10 | H@1 | H@5 | H@10 | H@1 | H@5 | H@10 |
| TransE/n-rel/single | 8.70 | 21.04 | 34.78 | 4.97 | 19.34 | 25.60 | 9.06 | 27.19 | 39.27 |
| TransE/n-rel/ensemble | 8.52 | 23.13 | 35.65 | 1.47 | 7.37 | 10.68 | 10.42 | 29.61 | 41.99 |
| TransE/w-rel/single | 16.17 | 35.83 | 48.52 | 8.84 | 24.68 | **35.17** | 18.28 | 42.45 | 54.68 |
| TransE/w-rel/ensemble | **18.78** | **38.78** | **51.65** | **8.84** | **24.86** | 32.78 | **19.64** | **43.05** | **55.29** |

The experiment result shows that our approach can achieve quite a high accuracy on dummy datasets. Even with the setting that only 20% triplets are shared. For the cross-lingual dataset, even we can not perform a result of supervised methods, as a trial of the completely unsupervised method, we achieve an acceptable result. Overall, our approach also provides a domain adaption framework to multi-channel data and reveal that the aligned channels are critical to alignment.

# 4 Conclusion

This research explores the unsupervised method on entity alignment task. We proposed a multi-channel optimal transport framework for unsupervised domain adaption on knowledge graph, which is able to leverage heuristic information to find corresponding information across two domain without any supervised entities. In practice, we perform an experiment on six datasets, including three dummy datasets and three cross-lingual Knowledge graphs and show its effectiveness. Since the model is limited by computational performance, in the future, we plan to stretch our model to more general settings and solve the defect in distinguish similar semantics.

# References


[Zhu2017] Zhu, Hao, et al. "Iterative Entity Alignment via Joint Knowledge Embeddings." IJCAI. 2017.

[Chen2016] Chen, Muhao, et al. "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment." arXiv preprint arXiv:1611.03954 (2016).

[Sun2018] Sun, Zequn, et al. "Bootstrapping Entity Alignment with Knowledge Graph Embedding." IJCAI. 2018.

[Trivedi2018] Trivedi, Rakshit, et al. "Linknbed: Multi-graph representation learning with entity linkage." arXiv preprint arXiv:1807.08447 (2018).

[Alvarez-Melis, 2018] Alvarez-Melis, David, and Tommi S. Jaakkola. "Gromov-wasserstein alignment of word embedding spaces." arXiv preprint arXiv:1809.00013 (2018).

[Peyré2016] Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-Wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. 2016.

[Mémoli2011] Mémoli, Facundo. "Gromov–Wasserstein distances and the metric approach to object matching." Foundations of computational mathematics 11.4 (2011): 417-487.

[Bordes2013] Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in neural information processing systems. 2013.

[Sun, 2019] Sun, Zhiqing, et al. "Rotate: Knowledge graph embedding by relational rotation in complex space." arXiv preprint arXiv:1902.10197 (2019).

[Sun, 2017] Sun, Zequn, Wei Hu, and Chengkai Li. "Cross-lingual entity alignment via joint attribute-preserving embedding." International Semantic Web Conference. Springer, Cham, 2017.