Optimal Transport Methods for Aligning Knowledge Graph Triples with Natural Language in Unsupervised Settings

Alexander Kalinowski[1]

Drexel University, Philadelphia, PA 19104, USA ajk437@drexel.edu
https://github.com/akalino

Abstract. Frameworks for aligning embeddings of text and embeddings of knowledge graphs (KG) have been used for generating mappings for test-to-text alignment and KG-to-KG alignment, but little has been done for alignment between these two domains. In this dissertation proposal, I aim to create a framework for KG-to-text alignment that utilizes little to no training data to learn these correspondences. Additionally, motivated by the semantic geometries of these embedding spaces, I propose a new line of research into generating explicit embeddings of triples from a knowledge graph.

Keywords: Knowledge representation · Automated metadata generation · Embeddings · Optimal transport

1 Importance

Knowledge graphs (KGs) and ontologies form the computational backbone of the modern Semantic Web, curated by taxonomists and ontologists in conjunction with domain subject matter experts. Collaboration between these parties is a bottleneck in large-scale organizations due to coordination of people and sourcing of relevant information and terminologies for ontologists to massage into an enterprise-wide standard. This bottleneck is felt both in developing ontological terminologies and populating the knowledge graph with facts and assertions about the domain being described.

Industrial terminologies are buried deep in policy documents or technical white papers, making the job of the ontologist one of synthesizing these documents into a conscise, machine-readable set of interlinked terminologies (T-Box). For large-scale knowledge graphs, such as those leveraged in popular search engines, the information scales past terminologies and additionally covers facts or assertions (A-Box) about the objects described in the graph. Validating the accuracy of assertions in the knowledge graph is critical for auditing the trustworthiness of those claims, which is especially relevant in highly regulated industries such as banking or pharmatecuticals. As facts (in the form of $\langle s, p, o \rangle$ triples) are added to a knowledge graph, either through automated methods such as link prediction techniques or human generated annotations, there is an additional opportunity to enrich the knowledge graph with metadata about these

triples, such as a source of evidence from a text document. Developing methodologies for linking the T-Box and A-Box to textual evidence will provide a set of tools to allow ontologists and knowledge graph developers to expedite their work while ensuring the highest degree of accuracy and auditability, and thus is the focus of this work.

For example, an ontologist working in the financial services domain may wish to develop a standardized definition of a fixed-float interest rate swap. They may begin by inheriting the structure of previously defined terminologies, namely those related to interest rates and swaps, deriving the triples

$$\langle fixed\_float\_interest\_rate\_swap, has\_type, swap\_contract\rangle,$$
$$\langle fixed\_float\_interest\_rate\_swap, has\_leg, fixed\_interest\_rate\rangle \text{ and}$$
$$\langle fixed\_float\_interest\_rate\_swap, has\_leg, floating\_interest\_rate\rangle.$$

However, without a background in financial terminology, they may miss the fact that a fixed-float interest rate swap is more commonly refered to as a vanilla interest rate swap. This fact can be inferenced from a variety of textual sources, such as a sentence like `'A vanilla interest rate swap allows two counterparties to hedge against interest rate volatility by trading a floating rate for a fixed rate.'`, given the set of seed triples the ontologist has developed, helping to expand the coverage of the knowledge graph.

Of critical importance in the development of a text to KG methodology is its ability to rapidly adapt to new source ontologies and data domains. Additionally, such a system should not be bottlenecked by reliance on abundant training data, a point of failure for many ML projects. This motivates the use of unsupervised techniques for knowledge representation, and I propose an exploration of cross-domain optimal transport between embeddings of natural language and embeddings of a knowledge graph. Given this motivation, I present the following formalism.

2 Problem Statement

Suppose $S = \{s_1, s_2, \ldots, s_n\}$ is a set of sentences and $T = \{t_1, t_2, \ldots, t_n\}$ is a set of knowledge graph triples, each triple of the form $t_i = \langle s, p, o\rangle$. Next, define two functions f and g that create low-dimensional (latent) representations of each sentence and triple, respectively, such that $f(s_i) = e_{s_i} \in \mathbb{R}^n$ (the *source* space) and $g(t_j) = e_{t_j} \in \mathbb{R}^m$ (the *target* space) where $n \ll |S|$, $m \ll |T|$. In order to link triples to their most relevant sentences, both for validating terminology and assertions, we seek a mapping function $\Psi : \mathbb{R}^n \to \mathbb{R}^m$ to *transport* one set to another with minimal loss, i.e. $\Psi(T) \approx S$, such that for new triples $t_i \notin T$ we can find a supporting sentence representation $\Psi(t_i) \approx s_j \notin S$ in a large-scale text corpus C that reflects the same semantic information. Learning such a Ψ should not be taxed by reliance on an abundance of paired labeled data points, i.e. a set of labels $L = \{(t_k, s_k), \ldots, (t_l, s_l)\}$. Instead, I assume no such set exists at training time, framing the learning of Ψ as an unsupervised task.

The desire to limit the reliance on paired labeled data points helps inform the potential choices of Ψ. Specifically, methods from optimal transport (OT) theory lend themselves well to this problem by exploiting the structure of each embedding space using pairwise distance metrics rather than relying on data with paired labels. Instead, couplings of objects from each respective space are inferred through probabilistic transport maps, shifting the focus from gathering labeled sentence-triple pairs for supervised learning to refining the representations of these objects in their respective latent spaces. Optimal transport also provides a probabilistic framework for mapping assignments; the Kantorivitch relaxation of Monge's original statement admits a solution where the mass at any point in the source space can be dispatched to several locations in the target space [18], fitting the problem setting as a single KG triple may admit infinitely innumerable sentence representations. The probabilistic and rigorous mathematical approach of OT add to the understandability of results in opposition to black-box models such as generative adversarial networks (GANs).

One drawback of OT techniques is the requirement of a cost matrix defined between the spaces $\mathcal{X}$ and $\mathcal{Y}$. Defining such a cost matrix requires labeled data points beween the two spaces, although a weaker assumption can be used to avoid this by defining two inter-domain distance matrices $D \in \mathbb{R}^{n \times n}$ and $D' \in \mathbb{R}^{m \times m}$. Such matrices can then be aligned through the following formulation of the Gromov-Wasserstein problem

$$GW((a, D), (b, D'))^2 = \min_{P \in U(a,b)} \mathcal{E}_{D,D'}(P)$$

$$\mathcal{E}_{D,D'}(P) = \Sigma_{i,j,i',j'} |D_{i,i'} - D'_{j,j'}|^2 P_{i,j} P_{i',j'}$$

.

Using this formalism, the problem of interest can then be framed as follows:

Problem Statement: *What are the optimal choices of embedding functions f and g to establish distance matrices D and D' such that*

1. *for pairs of triples $(t_i, t_{i'}) \in T$ that contain some notion of semantic similarity, $D_{i,i'} = d(f(t_i), f(t_{i'}))$ is minimized*
2. *while simultaneously minimizing $D'_{j,j'} = d(g(s_j), g(s_{j'}))$ for semantically similar sentences $(s_j, s_{j'}) \in S$*

for some distance metric d, thus allowing optimal couplings between t_i and s_i to be established via the Gromov-Wasserstein distance?

3 Research Questions, Hypotheses and Research Plan

Prior research in this area shows that optimal transport of embedding spaces has been successful within a given data domain (i.e. unsupervised alignment of word embeddings [2], unsupervised alignment of knowledge graph entities [17]). My research questions seek to extend these methodologies to the cross-domain task in order to align embeddings of knowledge graph triples with embeddings of semantically related sentences.

RQ-I: *Is there an accurate, unsupervised technique for aligning a set of knowledge graph triples to a set of semantically similar sentences?*

H-I: *Optimal transport methods provide a mathematical framework for unsupervised alignment based on intra-domain pairwise similarities. Successful application of OT is dependent on how similarities of like-objects are represented in each respective space.*

To accomplish this, properties of each embedding space that make them useful for such alignment must first be established. In my prior work [9], these properties are explored for sentence embeddings while keeping the knowledge graph embeddings fixed as the concatenation of head, tail and relation vectors generated by TransE. It remains to be seen how changing the structure of knowledge graph embeddings, via changing the algorithm selection (such as using more expressive models like ComplEx [20] or ConvE [7]), incorporating additional information such as literals [11] or changing the way entity and relation vectors are combined to represent a triple, help or hurt the ability to generate high-quality alignments, motivating the next research question.

RQ-II: *How can current knowledge graph embedding methods be extended past representing entities and relations as separate objects and instead focus on embedding triples as the target objects?*

As the majority of current methods focus almost exclusively on the link prediction task, these methods may not be well-suited for establishing embeddings of triples, leading to the following hypothesis.

H-II: *Triple embeddings built from aggregations of entity and relation embeddings do not sufficiently encode the underlying semantics of such triples.*

Building upon the work of [8], treating triples as walks on the knowledge graph and weighting the strength of each relationship may help to create a semantic embedding space that will assist in alignment. The following section details how I will approach measuring the amount of semantic information captured by these methods.

4 Approach and Evaluation

Motivated by work on word embedding regularities [14], I wish to probe both sentence and KG embedding spaces generated by a variety of embedding algorithms and measure the degree to which they exhibit an underlying structure that can be leveraged for aligning these resources. Lurking beneath the above research questions is the fuzzily-defined notion of "semantic similarity," but metrics exist to make this quantification concrete. These metrics are used to define how well semantic similarity is encoded in the latent representations of both triples and sentences, and they are important to capture with the goal of defining optimal pair-wise distance matrices D and D' in mind. To formalize the notion of structure, I introduce a definition of clusterability, following the work of [1]. For some dataset $X \in \mathbb{R}^n$, a description of the clusterability of X is a function $c : X \to v$ where $v \in \mathbb{R}$ is a real value. Here, v is a measure of how strong a clustering presence is in the underlying set X.

To test the clusterability hypothesis, I use the spatial histogram (SpatHist) approach to measure the clusterability of each space [1]. The SpatHist approach compares the data binned in all d-dimensions to samples randomly generated in the same d-dimensional space. As many of these bins may end up empty in high-dimensional embedding spaces, I perform principal component analysis (PCA) to project down to the two most informative dimensions, split the data into n equal-width bins, and compute the empirical joint probability mass function (EPMF). The same is then done for 500 sets of uniformly generated points with the same feature dimensionality, and the differences are compared using the Kullback-Leibler (KL) divergence – higher KL divergence indicates more clusterability. I report the mean and standard deviation of each of these experiments as my final estimates of clusterability. Additionally, I apply the Hopkins test of uniformity [12]. As the Hopkins test statistic tends to zero, the underlying data exhibits less uniformity, indicating that clustering may be a good way of exploring the data in an unsupervised way. As the test statistic increases, the data tends to be more uniformly distributed, exhibiting less of the structure I seek to exploit.

For evaluating the quality of learned alignments, I follow in the tradition of knowledge graph embedding literature [21] and evaluate these results for the Hits@5 and Hits@10 metrics. Analyzing the results of the top 5 and top 10 closest matches allows for a nearest neighborhood analysis of each aligned embbedding instance, helping to pinpoint areas for future improvement, such as the mitigation of the influence of hubs [13] .

5 Preliminary Results

To establish a baseline for this task, prior work [9] tested the ability for a low-capacity linear model to learn a mapping between sentence and knowledge graph representations. The purpose of this work is in evaluating sentence representations, measuring the extent to which they are able to create structure in the low-dimensional embedding spaces by evaluating how well they cluster together around their semantic content, in this case the expression of a particular relationship. Findings on the clustering capacity of a selection of sentence embedding methods are reproduced below.

The results demonstrate the dramatic differences in the efficacy of sentence embedding methodologies. In particular, the geometrically motivated GEM algorithm vastly outperforms all others in terms of semantic clusterability, especially those using more complex deep neural models. In addition, the GEM algorithm outperforms all others in terms of Hits@5 and Hits@10 when performing a simple linear map for alignment (Linear@5,10), and all alignments show improvement when replacing the linear alignment with optimal transport techniques (OT@5,10). Utilizing these results gives a clue as to how to build knowledge graph triple embeddings: by focusing on the novelty of each predicate and the entities involved, they can be "pushed" into respective areas of the low-dimensional embedding space, leading to increased cluster cohesion and

higher within cluster semantic similarity. Additional insights and recommended improvements are presented in [9].

Model	Dim.	Linear@5	Linear@10	OT@5	OT@10	SpatHist μ	Hopkins
Random	300	0.0762	0.0943	0.008	0.021	0.0018	0.2365
GloVe-mean	300	0.1175	0.1509	0.202	0.236	2.1680	0.1262
GloVe-DCT	300	0.0249	0.0345	0.105	0.150	1.2412	0.1306
GEM	300	**0.2417**	**0.3111**	**0.360**	**0.397**	**4.9716**	**0.0727**
SkipThought	4800	0.0538	0.0665	0.073	0.101	2.9001	0.1162
QuickThought	2048	0.0319	0.0418	0.067	0.204	1.2082	0.1716
LASER	1024	0.0956	0.1290	0.115	0.159	2.0528	0.1686
InferSentV1	4096	0.0560	0.0824	0.104	0.395	2.0010	0.2068
InferSentV2	4096	0.0598	0.0859	0.118	0.252	2.3067	0.2051
SentBERT	768	0.1038	0.1307	0.132	0.220	1.2141	0.1647

6 Related Work

I detail related work in the following two areas: word alignment methods and graph alignment methods. A complete survey of word, sentence and knowledge graph embeddings as well as methods for aligning each domain can be found in [10].

6.1 Natural Language Representation Alignment

Regression models for word-to-word alignment were first proposed by [15] as a means of capturing geometric patterns between embeddings across embedding spaces. Inconsistencies in this approach were noted by [22] who in turn modified the regression process to add unit length normalization and constrain map to be orthogonal. Applications of pre-processing and orthogonal constraints spurred further research into ways to manipulate the source and target embedding spaces to further express their geometric structures in [3] and [4]. Building 'pseudo-dictionaries' as a means of reducing the amount of necessary training data is suggested in [19]. The work of [16] further explores iterative learning, alternating between supervised alignment and unsupervised distribution matching, as well as introducing novel metrics to assess the orthogonality assumptions used in supervised approaches. A key approach for unsupervised learning is described in [6] where the authors propose leveraging an adversarial learning paradigm. While the adversarial method directly leverages word frequencies, an alternative unsupervised method in [5] captures these patterns by analyzing the similarity distributions of the word vectors themselves. Using the Gromov–Wasserstein distance, [2] transform the alignment problem to one of finding an optimal transport from source X to target Z.

6.2 Graph Representation Alignment

The majority of research in the area of knowledge graph embeddings focus on one specific task, namely knowledge graph completion, which seeks to make predictions of the following form: given $\langle s, p, ? \rangle$, make the best prediction for an object o such that the triple is a valid one in the context of the greater graph. The state-of-the-art methods in this space leverage knowledge graph embeddings, low-dimensional representations of the entities and relations between them as vectors. The majority of research in this area focuses on representing the nodes, or entities, of the graph, with considerable less emphasis on how the relations are represented, and representations of the entire $\langle s, p, o \rangle$ triple are rarely considered. Recent research into representing the entire triple is presented in [8], yet the results are limited to explorations in clustering and recommendation systems. This work can be extended to further use cases and eventually tied in with alignment methods to link knowledge graphs to text documents.

7 Reflection and Future Work

Based on the initial results presented herein, there is clearly room for improvement in methodologies for representation learning of triples in a knowledge graph. Additionally, the alignment of cross-domain representations – those spanning both text and knowledge graph, is not currently well explored. Exploration in this area can provide a brigde between the two respective data realms and provide tooling for unsupervised automation of ontology and knowledge graph development. Future work will involve measuring the clusterability of multiple existing knowledge graph embedding algorithms, evaluating their efficacy in alignments with sentence embeddings, and proposing new, semantically grounded approaches to embedding triples as singular objects.

References

1. Adolfsson, A., Ackerman, M., Brownstein, N.C.: To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition **88**, 13–26 (Apr 2019). https://doi.org/10.1016/j.patcog.2018.10.026, http://dx.doi.org/10.1016/j.patcog.2018.10.026
2. Alvarez-Melis, D., Jaakkola, T.S.: Gromov-wasserstein alignment of word embedding spaces. CoRR **abs/1809.00013** (2018), http://arxiv.org/abs/1809.00013
3. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. pp. 2289–2294 (01 2016). https://doi.org/10.18653/v1/D16-1250
4. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 5012–5019 (February 2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16935
5. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings (2018)

6. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. CoRR **abs/1710.04087** (2017), http://arxiv.org/abs/1710.04087
7. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings (2017)
8. Fionda, V., Pirrò, G.: Triple2vec: Learning triple embeddings from knowledge graphs. CoRR **abs/1905.11691** (2019), http://arxiv.org/abs/1905.11691
9. Kalinowski, A., An, Y.: A comparative study on structural and semantic properties of sentence embeddings (2020)
10. Kalinowski, A., An, Y.: A survey of embedding space alignment methods for language and knowledge graphs (2020)
11. Kristiadi, A., Khan, M.A., Lukovnikov, D., Lehmann, J., Fischer, A.: Incorporating literals into knowledge graph embeddings (2019)
12. Lawson, R.G., Jurs, P.C.: New index for clustering tendency and its application to chemical problems. Journal of Chemical Information and Computer Sciences **30**(1), 36–41 (1990). https://doi.org/10.1021/ci00065a010, https://pubs.acs.org/doi/abs/10.1021/ci00065a010
13. Lazaridou, A., Dinu, G., Baroni, M.: Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics). pp. 270–280. Association for Computational Linguistics, Beijing, China (2015). https://doi.org/10.3115/v1/P15-1027, https://www.aclweb.org/anthology/P15-1027
14. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013)
15. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013), http://arxiv.org/abs/1309.4168
16. Patra, B., Moniz, J.R.A., Garg, S., Gormley, M.R., Neubig, G.: Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 184–193. Association for Computational Linguistics, Florence, Italy (2019), https://www.aclweb.org/anthology/P19-1018
17. Pei, S., Yu, L., Zhang, X.: Improving cross-lingual entity alignment via optimal transport. pp. 3231–3237 (08 2019). https://doi.org/10.24963/ijcai.2019/448
18. Peyré, G., Cuturi, M.: Computational optimal transport (2020)
19. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax (2017)
20. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning (ICML). vol. 48, pp. 2071–2080 (2016)
21. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **PP**, 1–1 (09 2017). https://doi.org/10.1109/TKDE.2017.2754499
22. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 1006–1011. Association for Computational Linguistics, Denver, Colorado (2015), https://www.aclweb.org/anthology/N15-1104