# Jointly Embedding Relations and Mentions for Knowledge Population

Miao Fan[†,‡,*], Kai Cao[‡], Yifan He[‡] and Ralph Grishman[‡]
[†] CSLT, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University, Beijing, 100084, China.
[‡]Proteus Group, New York University, NY, 10003, U.S.A.
[*]fanmiao.cslt.thu@gmail.com

## ABSTRACT

This paper contributes a jointly embedding model for predicting relations between a pair of entities in the scenario of knowledge population. It differs from most stand-alone approaches which separately perform on either knowledge bases or free texts. The proposed model simultaneously learns low-dimensional vector representations for both triplets in knowledge repositories and the mentions of relations in free texts, so that we can leverage the evidences from both of the two resources to make more accurate predictions. We use NELL to evaluate the performance of our approach, compared with most of cutting-edge methods. Results of extensive experiments show that our model achieves significant improvement on relation extraction.

## Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Relation systems; I.2.7 [**Natural Language Processing**]: Text analysis

## General Terms

Theory

## Keywords

Knowledge population, embedding representations, relation extraction.

## 1. INTRODUCTION

Relation extraction [1, 10, 15], which aims at discovering the relationships between a pair of entities, is a significant research direction on discovering more beliefs for knowledge bases. Most stand-alone approaches, however, either use local graph patterns in knowledge repositories, or extract features from text mentions, to individually help predict relations between two entities. The heterogeneity brings about a gap between structured repositories and unstructured free texts, which breaks the dream of sharing the evidences from both knowledge and natural language.

For studies in decades, scientists either compete the performance of their methods on the public text datasets such as ACE[1] [11] and MUC[2] [16], or look for effective approaches [9, 12] on improving the accuracy of link prediction within knowledge bases such as NELL[3] [5] and Freebase[4] [2]. Thanks to a grateful research of distantly supervised relation extraction [6, 14] which facilitates the manual annotation via automatically aligning with the relation mentions in free texts, NELL can not only extract triplets, i.e. $\langle head\_entity, relation, tail\_entity \rangle$, but also collect the texts between two entities as the evidence of relation mention. We take an example from NELL which originally records a belief: $\langle concept : city : caroline, concept : citylocatedinstate, concept : stateorprovince : maryland, County\ and\ State\ of \rangle$, where "$County\ and\ State\ of$" is the mention between the head entity $concept : city : caroline$, and the tail entity $concept : stateorprovince : maryland$, to indicate the relation $concept : citylocatedinstate$.

Fortunately, the embedding techniques [7, 13] enlighten us to break through the limitation of heterogeneous resources, and to establish a connection between a relation and its corresponding mention via learning a specific vector representation for each of the elements, including the entities and relations in triplets, and the words in mentions. More specifically, we propose a joint relation mention embedding (JRME) model in this paper, which simultaneously learns low-dimensional vector representations for entities and relations in knowledge repositories, and in the meanwhile, each word in the relation mentions is also trained a dedicated embedding. This model helps us take advantage of the benefits from the two resources to make more accurate predictions. We use two different datasets extracted from NELL to evaluate the performance of JRME, compared with most of cutting-edge methods. It turns out that our model achieves significant improvement on relation extraction.

## 2. RELATED WORK

We group some recent works on relation extraction into two categories, i.e. text-based approaches and knowledge-based methods. Generally speaking, both of the parties seek better evidences to make more accurate prediction. The text-based community focuses on linguistic features such as the words combined with POS tags that indicate the relations, but the other side conducts relation inference depending on the local connecting patterns between en-
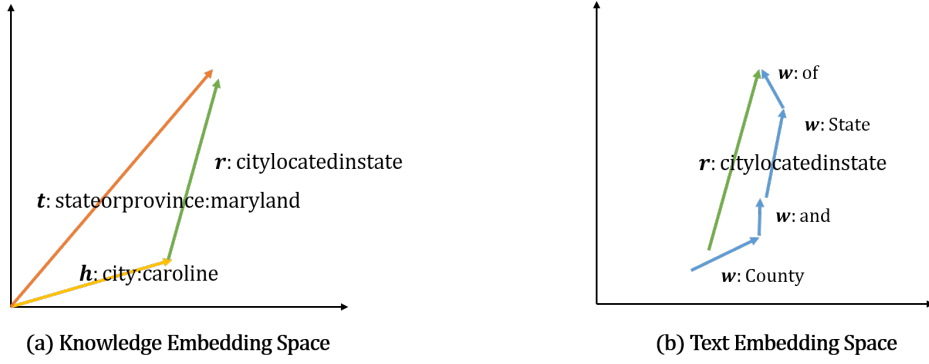
---

[1]http://www.itl.nist.gov/iad/mig/tests/ace/

[2]http://www.itl.nist.gov/iaui/894.02/related projects/muc/

[3]http://rtw.ml.cmu.edu/rtw/

[4]http://www.freebase.com/

**Figure 1: Given a belief, h** : $city : caroline$, **r** : $citylocatedinstate$, **t** : $stateorprovince : maryland$ **and m** : $County\ and\ State\ of$ **in NELL, (a) shows the distributed representations of a triplet in the knowledge space, and (b) illustrates word embeddings in the text space.**

tity pairs learnt from the knowledge graph which is established by beliefs.

## 2.1 Text-based Approaches

It is believed that the texts between two recognized entities in a sentence indicate their relationships to some extent. To implement a relation extraction system guided by supervised learning, a key step is to annotate the training data. Therefore, two branches emerge as follows,

- *Relation extraction with manual annotated corpora*: Traditional approaches compete the performance on the public text datasets which are annotated by experts, such as ACE and MUC. They choose different features extracted from the texts, like kernel features [16] or semantic parser features [11], and there is a comprehensive survey [15] which shows more details about this branch.

- *Relation extraction with distant supervision*: Due to the limited scale and tedious labor caused by manual annotation, scientists explore an alternative way to automatically generate large-scale annotated corpora, named by distant supervision [14]. Even though this cutting-edge technique solves the issue of lacking annotated corpora, we still suffer from the problem of noisy and sparse features [6].

## 2.2 Knowledge-based Methods

Knowledge bases contains millions of entries which are usually represented as triplets, i.e. $\langle head\_entity, relation, tail\_entity \rangle$, which intuitively inspire us to regard the whole repository as a graph, where entities are nodes and relations are edges. Therefore, one research community looks forward to predicting unknown relations which may exist between two entities via learning the linking patterns, and another promising research group tries to learn structured embeddings of knowledge bases.

- *Relation prediction with graph patterns*: Some canonical studies [9, 12] adopt a data-driven random walk model, which follows the paths from the head entity to the tail entity on the local graph structure to generate non-linear feature combinations to represent relations, and then uses logistic regression to select the significant features that contribute to classifying other entity pairs which also have the given relation.

- *Relation prediction with embedding representations*: Bordes et al. [3, 4] propose an alternative way that embed-

ding the whole knowledge graph via learning a specific low-dimensional vector for each entity and relation, so that we just need simple vector calculation instead to predict relations.

Our model (JRME) benefits more from the latest and state-of-the art embedding approaches, TransE [3] and IIKE [8]. Therefore, we re-implement them as the rival methods, and conduct extensive comparisons in the subsequent experiments.

## 3. MODEL

The heterogeneity between free texts and knowledge bases brings about a challenge that we can hardly take advantage of the features uniformly, since they locate in different spaces and have varies dimensions. Thankfully, the embedding techniques [7, 13] leave an idea that almost all the elements, including words, entities, relations, can be learnt and assigned distributed representations, and the mission remain for us is to jointly learn embeddings for entities, relations, and the words in the same feature space.

We arrange the subsequent content as follows: Section 3.1 and 3.2 describe how to model the knowledge and texts individually, and we finally talk about the proposed jointly embedding model in Section 3.3.

## 3.1 Knowledge Relation Embedding

Inspired by TransE [3], we regard the relation $r$ between a pair of entities, i.e. $h$ and $t$, as a transition, due to the hierarchical structure of knowledge graphs. Therefore, we use $D_r(h, r, t)$ as follows to denote the plausibility of a triplet $(h, r, t)$ illustrated by Figure 1(a):

$$D_r(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||_1, \tag{1}$$

where the closer $\mathbf{h} + \mathbf{r}$ is to $\mathbf{t}$, the more likely the triplet $(h, r, t)$ exists. The bold fonts indicate the vector representations, e.g. the embedding of the head entity $h$ is $\mathbf{h} \in \mathbb{R}^d$ where $d$ is short for dimension.

Assume that $R$ is the set of relations. Given a correct triplet $(h, r, t)$, we aim at pushing all the possible corrupt triplets with wrong relations $\{r'|r' \in R\ \&\ r' \neq r\}$ away. Therefore, we adopt a margin-based ranking loss function with a block $\alpha$ to separate all the negative triplets in the corrupted base $K'$ from all the positives in the

correct knowledge base $K$:

$$\arg\min_{r,r'} \mathcal{L}_r = \sum_{(h,r,t)\in K} \sum_{(h,r',t)\in K'} [\alpha + D_r(h,r,t) - D_r(h,r',t)]_+,$$
(2)

in which $[\,]_+$ is a hinge loss function, i.e. $[x]_+ = max(0,x)$.

## 3.2 Text Mention Embedding

Similar to the Knowledge Relation Embedding (KBE), we can also find an approach to measure the distance between the mention $m$ and its corresponding relation $r$ in Text Mention Embedding (TME). To denote the embedding of mention $\mathbf{m}$, we sum all the embeddings of words included by $m$ as shown by Equation (3). Thanks to representing all the words and relations in vectors with the same dimension which is demonstrated by Figure 1(b), we adopt inner product function shown by Equation (4) to calculate their similarity.

$$\mathbf{m} = \sum_{w \in m} \mathbf{w},$$
(3)

$$D_m(r,m) = -\mathbf{r}^T \mathbf{m}.$$
(4)

Before using the margin-based ranking loss function to learn, we need to construct negative set $T'$ for each pair of relation mention $(r,m)$ which appears in the correct training set $T$. To generate the negative pairs $(r',m)$, we keep the mention $m$ but iteratively change other relations in the set of relations $R$. The subsequent Formula (5) helps to discriminate the two opponent sets with a margin $\beta$,

$$\arg\min_{r,m,r'} \mathcal{L}_m = \sum_{(r,m)\in T} \sum_{(r',m)\in T'} [\beta + D_m(r,m) - D_m(r',m)]_+.$$
(5)

## 3.3 Joint Relation Mention Embedding

Due to the uniform modeling standard of KBE and TME, we can jointly embedding the relations and corresponding mentions (JRME) with Equation (6),

$$\arg\min_{r,m,r'} \mathcal{L} = \sum_{(h,r,t,m)\in KT} \sum_{(h,r',t,m)\in KT'} [\gamma + D_r(h,r,t)$$
$$- D_r(h,r',t) + D_m(r,m) - D_m(r',m)]_+,$$
(6)

in which each belief $(h,r,t,m)$ belonging to $KT$ contains two entities, the relation and its corresponding mention.

If we achieve the learnt embeddings for all the entities, relations and words in mentions, we can simply use Equation (7) to measure the rationality of a relation $r$ appearing between a pair of entities $h,t$ with the evidence of $m$:

$$Score(h,r,t,m) = D_r(h,r,t) + D_m(r,m)$$
(7)

## 4. EXPERIMENTS

We set up three objects to the experiments for evaluating the effectiveness of JRME, which are:

- testing the effectiveness of JRME in terms of different evaluation protocols/metrics;

- comparing the performances of JRME with other cutting-edge approaches;

- judging the robustness of the proposed model by using a larger but noisy dataset.

Therefore, Section 4.1 and 4.2 display the different datasets and the varies protocols we use to measure the performances compared with several state-of-the-art approaches, i.e TransE [3] and IIKE [8]. Section 4.3 will show the results of the extensive experiments.

## 4.1 Datasets

We prepare two datasets with different statistical characteristics. As illustrated by Table 1, both of them are generated by NELL [5], a Never-Ending Language Learner which aims to automatically extract beliefs from the Web. NELL-50K is a medium size dataset, and each of the belief, which contains the head entity $h$, the tail entity $t$, the relation $r$ between them, and the mention $m$ indicate the relation, is validated by experts. However, NELL-5M is a much larger one with five million uncertain training examples automatically learnt from the Web by NELL.

| DATASET | NELL-50K | NELL-5M |
|---|---|---|
| #(ENTITIES) | 29,904 | 177,635 |
| #(RELATIONS) | 233 | 236 |
| #(TRAINING EX.) | 57,356 | 5,000,000 |
| #(VALIDATING EX.) | 10,710 | 47,335 |
| #(TESTING EX.) | 10,711 | 47,335 |

**Table 1: Statistics of the datasets used for relation prediction task.**

## 4.2 Protocols

The scenario for experiments is that given a pair of entities, a short text/mention to indicate the correct relations and a set of candidate relations, we compare the performances among our models and other state-of-the-art approaches, with the metrics as follows,

- *Average Rank*: Each candidate relation will gain a score calculated by Equation (7). We sort them in ascent order and compare with the corresponding ground-truth belief. For each belief in the testing set, we get the rank of the correct relation. The average rank is an aggregative indicator to some extent, to judge the overall performance on relation extraction of an approach.

- *Hit@10*: Besides the average rank, scientists from the industrials concern more about the accuracy of extraction when selecting Top10 relations. This metric shows the proportion of beliefs that we predict the correct relation ranked in Top10.

- *Hit@1*: It is a more strict metric that can be referred by automatic system, since it demonstrates the accuracy when just picking the first predicted relation in the sorted list.

## 4.3 Hyperparameters

Before displaying the evaluation results, we need to elaborate the hyperparameters that have been tried, and show the best combination of hyperparameters we choose. Another advantage of embedding-based model is that it is unnecessary to tune many hyperparameters. For our model, we just need to set four, which are the uniform dimension $d$ of entities, relations and the words in mentions, the margin $\alpha$ of KBE, the margin $\beta$ of TME and the margin $\gamma$ of JRME. To decide the ideal set of hyperparameters, we use the validation set to pick the best combination from $d \in \{10, 20, 50, 100, 200\}$, $\alpha \in \{0.1, 1.0, 2.0, 5.0, 10.0\}$, $\beta \in \{0.1, 1.0, 2.0, 5.0, 10.0\}$ and $\gamma \in \{0.1, 1.0, 2.0, 5.0, 10.0\}$. Finally, we choose $d = 100, \alpha = 1.0, \beta = 1.0$ and $\gamma = 2.0$ to train the embeddings, as this combination of hyperparameters helps perform best on the validation set.

## 4.4 Performances

Table 2 and 3 illustrate the results of experiments on NELL-50K and NELL-5M, respectively. Both of them show that JRME performs best among all the approaches we implemented. We can also figure out that text mentions contribute a lot on predicting the correct relations. Moreover, Table 3 also demonstrates that not only IIKE is robust to the noise in NELL-5M dataset, which consists with its characteristics emphasized by Fan et al. [8], but also TME and JRME share this special "gene".

| APPROACH | AVG. R. | HIT@10 | HIT@1 |
|----------|---------|--------|-------|
| TransE [3] | 54.6 | 6.5% | 1.5% |
| KRE | 25.8 | 16.8% | 5.8% |
| IIKE [8] | 6.3 | 34.8% | 24.3% |
| TME | 3.3 | 45.9% | 32.4% |
| JRME | **2.1** | **47.5%** | **36.8%** |

**Table 2: Performance of TransE, KRE, IIKE, TME and JRME on the metrics of Average Rank, Hit@10 and Hit@1 in NELL-50K dataset.**

| APPROACH | AVG. R. | HIT@10 | HIT@1 |
|----------|---------|--------|-------|
| TransE [3] | 50.8 | 0.52% | 0.18% |
| KRE | 45.8 | 5.7% | 0.6% |
| IIKE [8] | 4.5 | 41.3% | 26.6% |
| TME | 1.8 | 48.4% | 32.4% |
| JRME | **1.5** | **48.6%** | **35.8%** |

**Table 3: Performance of TransE, KRE, IIKE, TME and JRME on the metrics of Average Rank, Hit@10 and Hit@1 in NELL-5M dataset.**

## 5. CONCLUSION

We engage in bridging over the gap between unstructured free texts and structured knowledge bases to predict more accurate relations via proposing a jointly embedding model between any given entity pair for knowledge population. The results of extensive experiments with varies evaluation protocols on both medium and large NELL datasets effectively demonstrate that our model (JRME) outperform other state-of-the-art approaches. Because of the uniform low-dimensional vector representations for entities, relations and even the words, evidences for prediction are compressed into embeddings to facilitate the information exchange and computing, which finally leads a huge leap forward in relation extraction.

There still remains, however, several open questions on this promising research direction in the future, such as exploring better ways to embed the whole beliefs or mentions without losing too much regularities of knowledge and linguistics.

## 6. ACKNOWLEDGMENTS

## 7. REFERENCES

[1] N. Bach and S. Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2007.

[2] K. Bollacker, R. Cook, and P. Tufts. Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pages 1962–1963, 2007.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795, 2013.

[4] A. Bordes, J. Weston, R. Collobert, Y. Bengio, et al. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[6] M. Fan, D. Zhao, Q. Zhou, Z. Liu, T. F. Zheng, and E. Y. Chang. Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[7] M. Fan, Q. Zhou, E. Chang, and T. F. Zheng. Transition-based knowledge graph embedding with relational mapping properties. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*, pages 328–337, Phuket,Thailand, December 2014. Department of Linguistics, Chulalongkorn University.

[8] M. Fan, Q. Zhou, and T. F. Zheng. Learning embedding representations for knowledge inference on imperfect and incomplete repositories. *arXiv preprint arXiv:1503.08155*, 2015.

[9] M. Gardner, P. P. Talukdar, B. Kisiel, and T. M. Mitchell. Improving learning and inference in a large knowledge-base using latent syntactic cues. In *EMNLP*, pages 833–838. ACL, 2013.

[10] R. Grishman. Information extraction: Techniques and challenges. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE '97, pages 10–27, London, UK, UK, 1997. Springer-Verlag.

[11] Z. GuoDong, S. Jian, Z. Jie, and Z. Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[12] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[15] S. Sarawagi. Information extraction. *Foundations and trends in databases*, 1(3):261–377, 2008.

[16] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.