

# Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction

**Jason Weston**  
Google  
111 8th avenue  
New York, NY, USA  
jweston@google.com

**Antoine Bordes**  
Heudiasyc  
UT de Compiègne  
& CNRS  
Compiègne, France  
bordes@uttc.fr

**Oksana Yakhnenko**  
Google  
111 8th avenue  
New York, NY, USA  
oksana@google.com

**Nicolas Usunier**  
Heudiasyc  
UT de Compiègne  
& CNRS  
Compiègne, France  
nusunier@uttc.fr

## Abstract

This paper proposes a novel approach for relation extraction from free text which is trained to jointly use information from the text and from existing knowledge. Our model is based on scoring functions that operate by learning low-dimensional embeddings of words, entities and relationships from a knowledge base. We empirically show on New York Times articles aligned with Freebase relations that our approach is able to efficiently use the extra information provided by a large subset of Freebase data (4M entities, 23k relationships) to improve over methods that rely on text features alone.

## 1 Introduction

Information extraction (IE) aims at generating structured data from free text in order to populate Knowledge Bases (KBs). Hence, one is given an incomplete KB composed of a set of triples of the form  $(h, r, t)$ ;  $h$  is the left-hand side entity (or *head*),  $t$  the right-hand side entity (or *tail*) and  $r$  the relationship linking them. An example from the Freebase KB<sup>1</sup> is  $(/m/2d3rf, \langle \text{director-of} \rangle, /m/3/324)$ , where  $/m/2d3rf$  refers to the director “Alfred Hitchcock” and  $/m/3/324$  to the movie “The Birds”.

This paper focuses on the problem of learning to perform relation extraction (RE) under weak supervision from a KB. RE is sub-task of IE that considers that entities have already been detected by a different process, such as a named-entity recognizer. RE then aims at assigning to a relation mention  $m$

(i.e. a sequence of text which states that some relation is true) the corresponding relationship from the KB, given a pair of extracted entities  $(h, t)$  as context. For example, given the triple  $(/m/2d3rf, \langle \text{wrote and directed} \rangle, /m/3/324)$ , a system should predict  $\langle \text{director-of} \rangle$ . The task is said to be weakly supervised because for each pair of entities  $(h, t)$  detected in the text, all relation mentions  $m$  associated with them are labeled with all the relationships connecting  $h$  and  $t$  in the KB, whether they are actually expressed by  $m$  or not.

Our key contribution is a novel model that employs not only weakly labeled text mention data, as most approaches do, but also leverages triples from the known KB. The model thus learns the plausibility of new  $(h, r, t)$  triples by generalizing from the KB, even though this triple is not present. A ranking-based embedding framework is used to train our model. Thereby, relation mentions, entities and relationships are all embedded into a common low-dimensional vector space, where scores are computed. We show that our method can successfully take into account information from a large-scale KB (Freebase: 4M entities, 23k relationships) to improve over systems that are only using text features.

This paper is organized as follows: Section 2 presents related work, Section 3 introduces our model and its main influences, and experimental results are displayed in Section 4.

## 2 Previous Work

Learning under weak supervision is common in natural language processing, especially for tasks where the annotation costs are significant such as in se-

<sup>1</sup>www.freebase.com

mantic parsing (Kate and Mooney, 2007; Liang et al., 2009; Bordes et al., 2010; Matuszek et al., 2012). This is also naturally used in IE, since it allows to train large-scale systems without requiring to label numerous texts. The idea was introduced by (Craven et al., 1999), which matched the Yeast Protein Database with PubMed abstracts. It was also used to train open extractors based on Wikipedia infoboxes and corresponding sentences (Wu and Weld, 2007; Wu and Weld, 2010). Large-scale open IE projects (e.g. (Banko et al., 2007)) also rely on weak supervision, since they learn models from a seed KB in order to extend it.

Weak supervision is also a popular option for RE: Mintz et al. (2009) used Freebase to train weakly supervised relational extractors on Wikipedia, an approach generalized by the multi-instance learning frameworks (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). All these works only use textual information to perform extraction.

Lao et al. (2012) proposed the first work aiming to perform RE employing both KB data and text, using a rule-based random walk method. Recently, Riedel et al. (2013) proposed another joint approach based on collaborative filtering for learning entity embeddings. This approach connects text with Freebase by learning shared embeddings of entities through weak supervision, in contrast to our method where no joint learning is performed. We do not compare to these two approaches since they use two different evaluation protocols that greatly differ from those used in all aforementioned previous works. Nevertheless, our method is easier to integrate into existing systems than those, since KB data is used via the addition of a scoring term, which is trained separately beforehand (with no shared embeddings). Besides, we demonstrate in our experimental section that our system can handle a large number of relationships, significantly larger than that presented in (Lao et al., 2012; Riedel et al., 2013).

### 3 Embedding-based Framework

Our work concerns energy-based methods, which learn low-dimensional vector representations (*embeddings*) of atomic symbols (words, entities, relationships, etc.). In this framework, we learn two models: one for predicting relationships given re-

lation mentions and another one to encode the interactions among entities and relationships from the KB. The joint action of both models in prediction allows us to use the connection between the KB and text to perform relation extraction. One could also share parameters between models (via shared embeddings), but this is not implemented in this work. This approach is inspired by previous work designed to connect words and Wordnet (Bordes et al., 2012).

Both submodels end up learning vector embeddings of symbols, either for entities or relationships in the KB, or for each word/feature of the vocabulary (denoted  $\mathcal{V}$ ). The set of entities and relationships in the KB are denoted by  $\mathcal{E}$  and  $\mathcal{R}$ , and  $n_v$ ,  $n_e$  and  $n_r$  denote the size of  $\mathcal{V}$ ,  $\mathcal{E}$  and  $\mathcal{R}$  respectively. Given a triple  $(h, r, t)$  the embeddings of the entities and the relationship (vectors in  $\mathbb{R}^k$ ) are denoted with the same letter, in boldface characters (i.e.  $\mathbf{h}$ ,  $\mathbf{r}$ ,  $\mathbf{t}$ ).

#### 3.1 Connecting Text and Relationships

The first part of the framework concerns the learning of a function  $S_{m2r}(m, r)$ , based on embeddings, that is designed to score the similarity of a relation mention  $m$  and a relationship  $r$ .

Our scoring approach is inspired by previous work for connecting word labels and images (Weston et al., 2010), which we adapted, replacing images by mentions and word labels by relationships. Intuitively, it consists of first projecting words and features into the embedding space and then computing a similarity measure (the dot product in this paper) between this projection and a relationship embedding. The scoring function is then:

$$S_{m2r}(m, r) = \mathbf{f}(m)^\top \mathbf{r}$$

with  $\mathbf{f}$  a function mapping words and features into  $\mathbb{R}^k$ ,  $\mathbf{f}(m) = \mathbf{W}^\top \Phi(m)$ .  $\mathbf{W}$  is the matrix of  $\mathbb{R}^{n_v \times k}$  containing all word embeddings  $\mathbf{w}$ ,  $\Phi(m)$  is the (sparse) binary representation of  $m$  ( $\in \mathbb{R}^{n_v}$ ) indicating absence or presence of words/features, and  $\mathbf{r} \in \mathbb{R}^k$  is the embedding of the relationship  $r$ .

This approach can be easily applied at test time to score (mention, relationship) pairs. Since this type of learning problem is weakly supervised, Bordes et al. (2010) showed that a convenient way to train it is by using a ranking loss. Hence, given a data set  $\mathcal{D} = \{(m_i, r_i), i = 1, \dots, |\mathcal{D}|\}$  consisting of (mention, relationship) training pairs, one could learn the

embeddings using constraints of the form:

$$\forall i, \forall r' \neq r_i, \mathbf{f}(m_i)^\top \mathbf{r}_i > 1 + \mathbf{f}(m_i)^\top \mathbf{r}' , \quad (1)$$

where 1 is the margin. That is, we want the relation that (weakly) labels a given mention to be scored higher than other relation by a margin of 1. Then, given any mention  $m$  one can predict the corresponding relationship  $\hat{r}(m)$  with:

$$\hat{r}(m) = \arg \max_{r' \in \mathcal{R}} S_{m2r}(m, r') = \arg \max_{r' \in \mathcal{R}} (\mathbf{f}(m)^\top \mathbf{r}').$$

Learning  $S_{m2r}(\cdot)$  under constraints (1) is well suited when one is interested in building a permutation prediction system. However, performance metrics of relation extraction are sometimes measured using precision recall curves aggregated for all mentions concerning the same pair of entities, as in (Riedel et al., 2010). In that case the scores across predictions for different mentions need to be calibrated so that the most confident ones have the higher scores. This can be better encoded with constraints of the following form:

$$\forall i, j, \forall r' \neq r_i, r_j, \mathbf{f}(m_i)^\top \mathbf{r}_i > 1 + \mathbf{f}(m_j)^\top \mathbf{r}' .$$

In this setup, scores of pairs observed in the training set should be larger than that of any other prediction across all mentions. In practice, we use “soft” ranking constraints (optimizing the hinge loss), i.e. we minimize:

$$\forall i, j, \forall r' \neq r_i, r_j, \max(0, 1 - \mathbf{f}(m_i)^\top \mathbf{r}_i + \mathbf{f}(m_j)^\top \mathbf{r}').$$

Finally, we also enforce a (hard) constraint on the norms of the columns of  $\mathbf{W}$  and  $\mathbf{r}$ , i.e.  $\forall i, \|\mathbf{W}_i\|_2 \leq 1$  and  $\forall j, \|\mathbf{r}_j\|_2 \leq 1$ . Training is carried out by Stochastic Gradient Descent (SGD), updating  $\mathbf{W}$  and  $\mathbf{r}$  at each step, following (Weston et al., 2010; Bordes et al., 2013). That is, at the start of training the parameters to be learnt (the  $n_v \times k$  word/feature embeddings in  $W$  and the  $n_r \times k$  relation embeddings  $r$ ) are initialized to random weights. We initialize each  $k$ -dimensional embedding vector randomly with mean 0, standard deviation  $\frac{1}{k}$ . Then, we iterate the following steps to train them:

1. Select at random a positive training pair  $(m_i, r_i)$ .

2. Select at random a secondary training pair  $(m_j, r_j)$ , used to calibrate the scores.
3. Select at random a negative relation  $r'$  such that  $r' \neq r_i$  and  $r' \neq r_j$ .
4. Make a stochastic gradient step to minimize  $\max(0, 1 - \mathbf{f}(m_i)^\top \mathbf{r}_i + \mathbf{f}(m_j)^\top \mathbf{r}')$ .
5. Enforce the constraint that each embedding vector is normalized, i.e. if  $\|\mathbf{W}_i\|_2 > 1$  then  $\mathbf{W}_i \leftarrow \mathbf{W}_i / \|\mathbf{W}_i\|_2$ .

### 3.2 Encoding Structured Data of KBs

Using only weakly labeled text mentions for training ignores much of the prior knowledge we can leverage from a large KB such as Freebase. In order to connect this relational data with our model, we propose to encode its information into entity and relationship embeddings. This allows us to build a model which can score the plausibility of new entity relationship triples which are missing from Freebase. Several models have been recently developed for that purpose (e.g. in (Nickel et al., 2011; Bordes et al., 2011; Bordes et al., 2012)): we chose in this work to follow the approach of (Bordes et al., 2013), which is simple, flexible and has shown very promising results on Freebase data.

Given a training set  $\mathcal{S} = \{(h_i, r_i, t_i), i = 1, \dots, |\mathcal{S}|\}$  of relations extracted from the KB, this model learns vector embeddings of the entities and of the relationships using the idea that the functional relation induced by the  $r$ -labeled arcs of the KB should correspond to a translation of the embeddings. That is, given a  $k$ -dimensional embedding of the left-hand side (head) entity, adding the  $k$ -dimensional embedding of a given relation should yield the point (or close to the point) of the  $k$ -dimensional embedding of the right-hand side (tail) entity. Hence, this method enforces that  $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$  when  $(h, r, t)$  holds, while  $\mathbf{h} + \mathbf{r}$  should be far away from  $\mathbf{t}$  otherwise. The model thus gives the following score for the plausibility of a relation:

$$S_{kb}(h, r, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2.$$

A ranking loss is also used for training  $S_{kb}$ . The ranking objective is designed to assign higher scores

to existing relations versus any other possibility:

$$\begin{aligned} \forall i, \forall h' \neq h_i, \quad S_{kb}(h_i, r_i, t_i) &\geq 1 + S_{kb}(h', r_i, t_i), \\ \forall i, \forall r' \neq r_i, \quad S_{kb}(h_i, r_i, t_i) &\geq 1 + S_{kb}(h_i, r', t_i), \\ \forall i, \forall t' \neq t_i, \quad S_{kb}(h_i, r_i, t_i) &\geq 1 + S_{kb}(h_i, r_i, t'). \end{aligned}$$

That is, for each known triple  $(h, r, t)$ , if we replaced the (i) head, (ii) relation or (iii) tail with some other possibility, the modified triple should have a lower score (i.e. be less plausible) than the original triple. The three sets of constraints defined above encode the three types of modification. As in Section 3.1 we use soft constraints via the hinge loss, enforce constraints on the norm of embeddings, i.e.  $\forall h, r, t, \|h\|_2 \leq 1, \|r\|_2 \leq 1, \|t\|_2 \leq 1$ , and training is performed using SGD, as in (Bordes et al., 2013).

At test time, one may again need to calibrate the scores  $S_{kb}$  across entity pairs. We propose a simple approach: we convert the scores by ranking all relationships  $\mathcal{R}$  by  $S_{kb}$  and instead output:

$$\tilde{S}_{kb}(h, r, t) = \Phi\left(\sum_{r' \neq r} \delta(S_{kb}(h, r, t) > S_{kb}(h, r', t))\right),$$

i.e. a function of the rank of  $r$ . We chose the simplified model  $\Phi(x) = 1$  if  $x < \tau$  and 0 otherwise;  $\delta(\cdot)$  is the Kronecker function.

### 3.3 Implementation for Relation Extraction

Our framework can be used for relation extraction in the following way. First, for each pair of entities  $(h, t)$  that appear in the test set, all the corresponding mentions  $\mathcal{M}_{h,t}$  in the test set are collected and a prediction is performed with:

$$\hat{r}_{h,t} = \operatorname{argmax}_{r \in \mathcal{R}} \sum_{m \in \mathcal{M}_{h,t}} S_{m2r}(m, r).$$

The predicted relationship can either be a valid relationship or NA – a marker that means that there is no relation between  $h$  and  $t$  (NA is added to  $\mathcal{R}$  during training and is treated like other relationships). If  $\hat{r}_{h,t}$  is a relationship, a composite score is defined:

$$S_{m2r+kb}(h, \hat{r}_{h,t}, t) = \sum_{m \in \mathcal{M}_{h,t}} S_{m2r}(m, \hat{r}_{h,t}) + \tilde{S}_{kb}(h, \hat{r}_{h,t}, t)$$

That is, only the top scoring non-NA predictions are re-scored. Hence, our final composite model favors predictions that agree with both the mentions and the KB. If  $\hat{r}_{h,t}$  is NA, the score is unchanged.

## 4 Experiments

We use the training and test data, evaluation framework and baselines from (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012).

**NYT+FB** This dataset, developed by (Riedel et al., 2010), aligns Freebase relations with the New York Times corpus. Entities were found using the Stanford named entity tagger (Finkel et al., 2005), and were matched to their name in Freebase. For each mention, sentence level features are extracted which include part of speech, named entity and dependency tree path properties. Unlike some of the previous methods, we do not use features that aggregate properties across multiple mentions. We kept the 100,000 most frequent features. There are 52 possible relationships and 121,034 training mentions of which most are labeled as no relation (labeled “NA”) – there are 4700 Freebase relations mentioned in the training set, and 1950 in the test set.

**Freebase** Freebase is a large-scale KB that has around 80M entities, 23k relationships and 1.2B relations. We used a subset restricted to the top 4M entities for scalability reasons – where top is defined as the ones with the largest number of relations to other entities. We used all the 23k possible relationships in Freebase. To make a realistic setting, we did not choose the entity set using the NYT+FB data set, so it may not overlap completely. For that reason, we needed to keep the set rather large. Keeping the top 4M entities gives an overlap of 80% with the entities in the NYT+FB test set. Most importantly, we then removed all the entity pairs present in the NYT+FB test set from Freebase, i.e. all relations they are involved in independent of the relationship. This ensures that we cannot just memorize the true relations for an entity pair – we have to learn to generalize from other entities and relations.

As the NYT+FB dataset was built on an earlier version of Freebase we also had to translate the deprecated relationships into their new variants (e.g. “/p/business/company/place\_founded” → “/organization/organization/place\_founded”) to make the two datasets link (then, the 52 relationships in NYT+FB are now a subset of the 23k from Freebase). We then trained the  $S_{kb}$  model on the remaining triples.

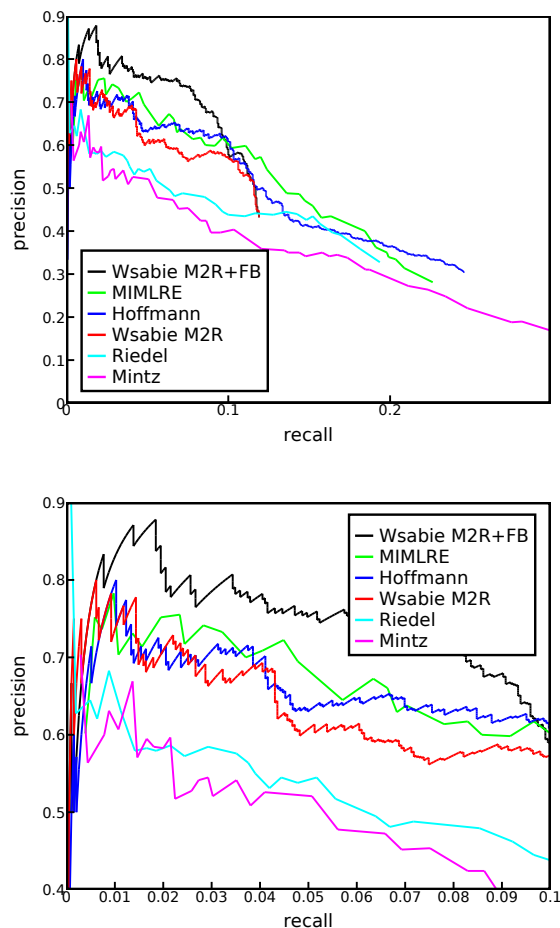


Figure 1: Top: Aggregate extraction precision/recall curves for a variety of methods. Bottom: the same plot zoomed to the recall [0-0.1] region.  $Wsabie_{M2R}$  is our method trained only on mentions,  $Wsabie_{M2R+FB}$  uses Freebase annotations as well.

**Modeling** Following (Bordes et al., 2013) we set the embedding dimension  $k$  to 50. The learning rate for SGD was selected using a validation set: we obtained 0.001 for  $S_{m2r}$ , and 0.1 for  $S_{kb}$ . For the calibration of  $\hat{S}_{kb}$ ,  $\tau = 10$  (note, here we are ranking all 23k Freebase relationships). Training  $S_{m2r}$  took 5 minutes, whilst training  $S_{kb}$  took 2 days due to the large scale of the data set.

**Results** Figure 1 displays the aggregate precision / recall curves of our approach  $Wsabie_{M2R+FB}$  which uses the combination of  $S_{m2r} + S_{kb}$ , as well as  $Wsabie_{M2R}$ , which only uses  $S_{m2r}$ , and existing state-of-the-art approaches: HOFFMANN (Hoffmann

et al., 2011)<sup>2</sup>, MIMLRE (Surdeanu et al., 2012), RIEDEL (Riedel et al., 2010) and MINTZ (Mintz et al., 2009).

$Wsabie_{M2R}$  is comparable to, but slightly worse than, the MIMLRE and HOFFMANN methods, possibly due to its simplified assumptions (e.g. predicting a single relationship per entity pair). However, the addition of extra knowledge from other Freebase entities in  $Wsabie_{M2R+FB}$  provides superior performance to all other methods, by a wide margin, at least between 0 and 0.1 recall (see bottom plot).

Performance of  $Wsabie_{M2R}$  and  $Wsabie_{M2R+FB}$  for recall  $> 0.1$  degrades rapidly, faster than that of other methods. This is also caused by the simplifications of  $Wsabie_{M2R}$  that prevent it from reaching high precision when the recall is greater than 0.1. We recall that Freebase data is not used to detect relationships i.e. to discriminate between NA and the rest, but only to select the best relationship in case of detection. That is  $Wsabie_{M2R+FB}$  only improves precision, not recall, so both versions of  $Wsabie$  are similar w.r.t. recall. This could be improved by borrowing ideas from HOFFMANN (Hoffmann et al., 2011) or MIMLRE (Surdeanu et al., 2012) for dealing with the multi-label case. Our approach, which uses Freebase to increase precision, is general and could improve any other method.

## 5 Conclusion

In this paper we described a framework for leveraging large scale knowledge bases to improve relation extraction by training not only on (mention, relationship) pairs but using all other KB triples as well. We empirically showed that it allows to significantly improve precision on extracted relations. Our modeling approach is general and should apply to other settings, e.g. for the task of entity linking.

## Acknowledgments

This work was carried out in the framework of the Labex MS2T (ANR-11-IDEX-0004-02), and funded by the French National Agency for Research (EVEREST-12-JS02-005-01).

<sup>2</sup>There is an error in the plot from (Hoffmann et al., 2011), which we have corrected. The authors acknowledged this issue.

## References

- Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.
- Antoine Bordes, Nicolas Usunier, and Jason Weston. 2010. Label ranking under ambiguous supervision for learning semantic correspondences. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 103–110.
- Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *Proc. of the 25th Conf. on Artif. Intel. (AAAI)*.
- Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *Proc. of the 15th Intern. Conf. on Artif. Intel. and Stat.*, volume 22, pages 127–135. JMLR.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS 26)*.
- Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 541–550.
- Rohit J Kate and Raymond J Mooney. 2007. Learning language semantics from ambiguous supervision. In *AAAI*, volume 7, pages 895–900.
- Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W Cohen. 2012. Reading the web with learned syntactic-semantic inference rules. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1017–1026. Association for Computational Linguistics.
- Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. 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 1-Volume 1*, pages 91–99. Association for Computational Linguistics.
- Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the International Conference on Machine Learning*.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. 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.
- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 809–816.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.
- Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL-HLT*, pages 74–84.
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.
- Jason Weston, Samy Bengio, and Nicolas Usunier. 2010. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35.
- Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.
- Fei Wu and Daniel S Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.