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 UMR CNRS 7253
U. Tech. de Compiègne
Compiègne, France
bordesan@utc.fr

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

Nicolas Usunier
Heudiasyc UMR CNRS 7253
U. Tech. de Compiègne
Compiègne, France
usuniern@utc.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 two scoring functions that operate by learning low-dimensional embeddings of words and of 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 existing 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[1] is (/m/2d3rf ,<director-of>, /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 triplet (/m/2d3rf ,“wrote and directed", /m/3/324), a system should predict <director-of>. 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 such a 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 system can successfully take into account information from a large-scale KB (Freebase: 4M entities, 23k relationships) to improve over existing systems, which are only using text features.

Previous work Learning under weak supervision is common in Natural language processing, especially for tasks where the annotations costs are important such as semantic 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 la-

[1] www.freebase.com

bel 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 (Banko et al., 2007; Carlson et al., 2010) 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.

Recently, Riedel et al. (2013) proposed an approach to model jointly KB data and text by relying on collaborative filtering. Unlike our model, this method can not directly connect text mentions and KB relationships, but does it indirectly through joint learning of shared embeddings for entities in text and in the KB. We did not compare with this recent approach, since it uses a different evaluation protocol than previous work in RE.

2 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 relation 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 triplet (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. $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$).

2.1 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 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 windows of words 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 a window of words 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 ($\in \mathbb{R}^{n_v}$) 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 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 $\mathcal{D} = \{(m_i, r_i), i = 1, \ldots, |\mathcal{D}|\}$ consisting of (mention, relationship) training pairs, one could learn the embeddings using constraints of the form:

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

where 1 is the margin. Given any mention m one can predict the corresponding relationship $\hat{r}(m)$ with:

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

Learning $S_{m2r}(\cdot)$ under constraints (1) is well suited when one is interested in building a a per-mention 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:

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

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), and 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. See (Weston et al., 2010; Bordes et al., 2013) for details.

2.2 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 use the approach of (Bordes et al., 2013), which is simple, flexible and has shown very promising results on Freebase data.

Given a training set $\mathcal{S} = \{(h_i, r_i, t_i), i = 1, \ldots, |\mathcal{S}|\}$ 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. 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. Hence such a model 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:

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

As in section 2.1 we use soft constraints, 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:

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

i.e. a function of the rank of r. We chose the simplified model $\Phi(x) = 1$ if $x < t$ and 0 otherwise.

2.3 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:

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

Hence, the composite model favors predictions that agree with both the mentions and the KB. If $\hat{r}_{h,t}$ is NA, the score is unchanged.

3 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 (with the largest number of relations in a preprocessed subset) for scalability reasons. We used all the 23k relationships. 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 " $\rightarrow$ "/organization/organization/place_founded") to make the two datasets link (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.

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 $\hat{S}_{kb}$, $t = 10$ (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.

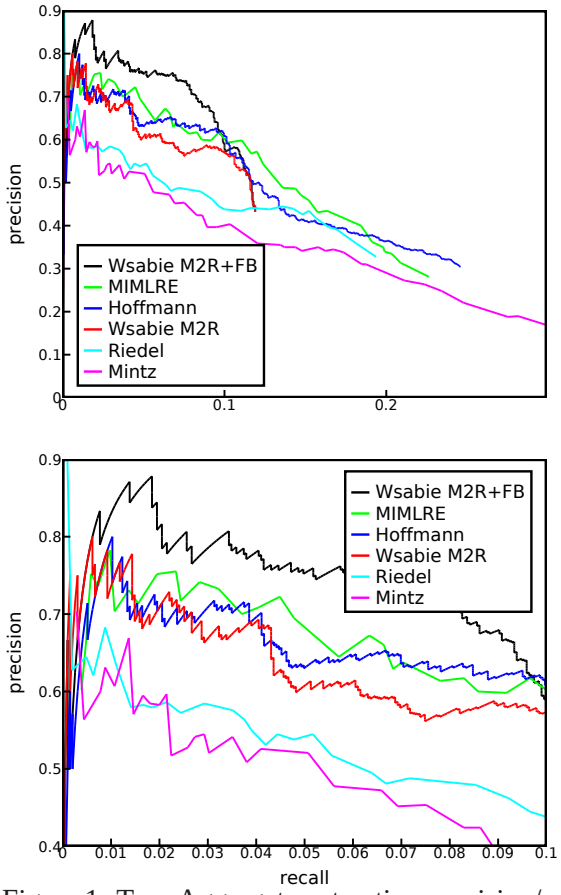

Figure 1: Top: Aggregate extraction precision/recall curves for a variety of methods. Bottom: the same plot zoomed to the recall [0-0.1] region. Wsabie$_{M2R}$ is our method trained only on mentions, Wsabie$_{M2R+FB}$ uses Freebase annotations as well.

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 state-of-the-art: HOFFMANN (Hoffmann et al., 2011)[2], 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

[2]There is an error in the plot from (Hoffmann et al., 2011), which we have corrected. The authors acknowledged the issue.

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).

4 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. Our modeling approach is general and should apply to other settings, e.g. for the task of entity linking.

References

[Banko et al.2007] 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.

[Bordes et al.2010] 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.

[Bordes et al.2011] 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)*.

[Bordes et al.2012] 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.

[Bordes et al.2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Irreflexive and hierarchical relations as translations. *arXiv preprint arXiv:1304.7158*.

[Carlson et al.2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*.

[Craven et al.1999] Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

[Finkel et al.2005] 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.

[Hoffmann et al.2011] 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.

[Kate and Mooney2007] Rohit J Kate and Raymond J Mooney. 2007. Learning language semantics from ambiguous supervision. In *AAAI*, volume 7, pages 895–900.

[Liang et al.2009] 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.

[Matuszek et al.2012] 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*.

[Mintz et al.2009] 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.

[Nickel et al.2011] 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.

[Riedel et al.2010] 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.

[Riedel et al.2013] 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.

[Surdeanu et al.2012] 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.

[Weston et al.2010] 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.

[Wu and Weld2007] 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.

[Wu and Weld2010] 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.