Contextual Pattern Embeddings for One-shot Relation Extraction

Abiola Obamuyide
Department of Computer Science
The University of Sheffield
Sheffield, UK
avobamuyide1@sheffield.ac.uk

Andreas Vlachos
Department of Computer Science
The University of Sheffield
Sheffield, UK
a.vlachos@sheffield.ac.uk

Abstract

We consider the task of learning extractors for knowledge base relations from little training data. This learning setup, also referred to as one-shot learning, is challenging for models that assume the availability of substantial amounts of training data from which to learn patterns that can generalize to unseen instances at test time. Nevertheless, it is also of practical importance, as many real-world knowledge bases are incomplete, and need to be extended to new relations for which there are typically limited learning examples. Previous work proposed the use of logic rules combined with matrix factorization in order to improve predictive accuracy when only a few training examples are available. In this work, we instead propose and show how explicit modeling of contextual patterns, within a factorization machine-based model, can be effectively utilized for this task. We test our approach on a standard relation extraction dataset, and find that with limited training data, our approach obtains relative improvements of more than 3% points in area under the weighted Mean Average Precision (wMAP) curve, compared to state-of-the-art approaches that utilize matrix factorization combined with additional supervision signals in the form of propositional logic rules.

1 Introduction

Extracting the relations between entities of interest plays a useful role in many natural language understanding systems, including those used for various tasks such as question answering and automatic knowledge base population, resulting in several methods and techniques being used for this task (Zelenko, Aone, and Richardella, 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2006; Mintz et al., 2009; Surdeanu et al., 2012; Riedel et al., 2013).

Riedel et al. (2013) for instance, proposed an approach based on matrix factorization. This approach casts the problem of extracting relations between entities as one of link prediction over a universal schema consisting of the union of surface patterns, knowledge base (KB) relations and entities. In this framework, facts from the knowledge base stipulating that a certain relation holds among two entities provide supervision signal for learning relation extractors.

However, the need often arises in practice to learn extractors for new relations for which there is limited training data, as is the case when a knowledge base needs to be extended to new relations for which there are only a few known facts. This learning setting, where a model is allowed access to limited learning instances per class, has also been referred to as one-shot learning (Miller, Matsakis, and Viola, 2000; Fei-Fei, Fergus, and Perona, 2006).

In order to learn extractors for knowledge base relations for which there is limited existing training data, we propose and evaluate the use of *contextual patterns*, which we define as the surface patterns which co-occur with knowledge base entities within a corpus of text. We learn embeddings for these

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

patterns and knowledge base relations jointly within the framework of a Factorization Machine (FM) model.

Our contributions are twofold: (i) We demonstrate that by jointly modeling the correlations between knowledge base relations and contextual patterns in a single framework, we are able to exploit information which is readily available in the text and achieve superior accuracy, without making use of additional human supervision in the form of propositional rules, as is the case in current approaches (ii) We present evaluations showing that the proposed approach leads to improved area under the weighted Mean Average Precision (wMAP) curve, when compared to state-of-the-art approaches.

2 Background

2.1 One-shot Learning

The notion of one-shot learning, which has also been explored in computer vision (Miller, Matsakis, and Viola, 2000; Fei-Fei, Fergus, and Perona, 2006), is used to describe the learning setting where the model is required to generalize from one or few example instances per class. This is a realistic scenario when there are classes for which training data is limited, for instance when having to learn a classifier for a new type of object in computer vision or when a knowledge base needs to be extended to new relations for which limited learning examples are available. In contrast with the zero-shot learning (Larochelle, Erhan, and Bengio, 2008) setting, where the model is not allowed any labeled examples, one-shot allows for one or few example labels per class. We consider the one-shot learning setting to be realistic, as limited supervision can often be easily obtained for new classes, for instance by asking the user to provide some examples for the new relation.

In a similar vein, bootstrapped learning approaches, for instance Carlson et al. (2010), learn extractors for different relations starting from a few initial seeds as training data. Learning then proceeds in several rounds, by progressively re-training extractors with the union of the previous training instances and the current model predictions on unlabeled data. However, such approaches are often troubled by the noise in the model predictions, a phenomenon referred to as semantic drift (Curran, Murphy, and Scholz, 2007).

2.2 Factorization Machines (FM)

Rendle (2010) proposed factorization machines in the context of recommender systems as a way to learn effective scoring functions with sparse inputs, in order to assess how likely is that a user-item combination occurs in reality. More concretely, a FM of order 2 models the scoring of a possibly sparse, real-valued input feature vector $\boldsymbol{f} \in \Re^d$ according to the following equation:

$$s(\boldsymbol{f}) = \sum_{m=1}^{d} b_m f_m + \sum_{m=1}^{d} \sum_{n=m+1}^{d} \langle \phi_m, \phi_n \rangle f_m f_n \tag{1}$$

The first summand is a linear model, where each feature f_m is weighted by a corresponding feature weight $b_m \in \Re$. The second summand captures the interaction between all possible feature pairs under a low-rank assumption. Each feature f_m has a corresponding embedding $\phi_m \in \Re^k$ with $k << d$, and the interaction between two features is captured via their dot product $\langle \phi_m, \phi_n \rangle$ multiplied by the product of their values in the instance $f_m f_n$.

While FM models have been explored for relation extraction by Petroni, Del Corro, and Gemulla (2015) and Weibl, Bouchard, and Riedel (2016), their effectiveness was not investigated within the context of limited supervision data.

2.3 Relation Extraction with Universal Schemas

Universal Schema (Riedel et al., 2013) is an approach to relation extraction that jointly embeds surface patterns, knowledge base relations and entities in a common embedding space through matrix factorization. It sidesteps the problem of aligning relations to sentences from the training corpus, which can lead to semantic drift in distantly supervised relation extraction approaches. It achieves this by performing joint inference across surface patterns, knowledge base relations and entities.

Rocktäschel, Singh, and Riedel (2015) and Demeester, Rocktäschel, and Riedel (2016) inject prior knowledge in the form of logical rules to improve relation extraction learning for new relations with zero or few training labels. While their experiments were also carried out within the framework of universal schema-based relation extraction, they considered the use of propositional logic rules, which for instance, can be mined from external knowledge bases (which are often incomplete themselves), obtained from a domain expert or from ontologies such as WordNet (Miller, 1995) (both of which may not be readily available, especially for a new domain). We instead investigate the use of information which is available from the text itself, and does not require consulting any additional external data sources to obtain.

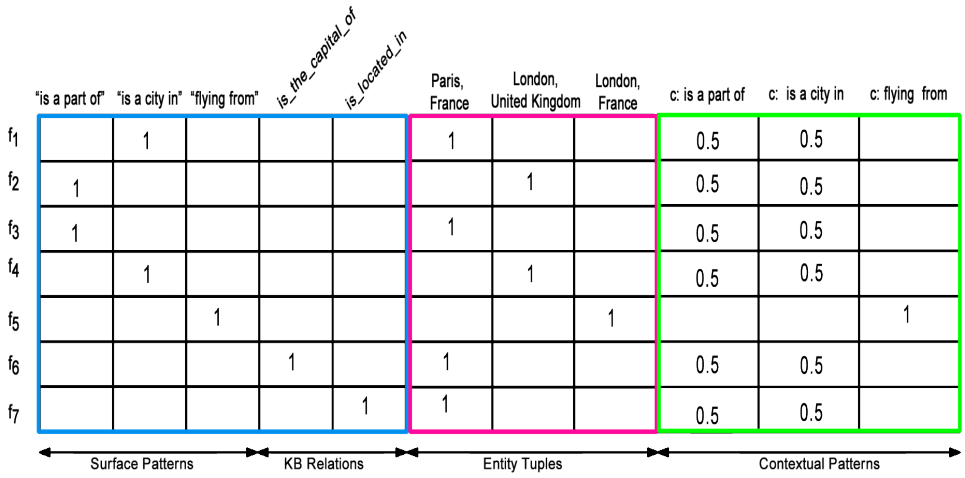


Figure 1: Input observations as a matrix with contextual pattern information. Each row $(f_1, f_2, f_3, ...)$ represents a (candidate) fact. The surface patterns have been simplified from their lexicalized dependency representations for readability.

3 Proposed Approach

Let $\mathcal{T}$ and $\mathcal{R}$ be the set of entity pairs and relations respectively, where $\mathcal{R}$ is the union of KB relations and surface patterns $\mathcal{S}$. We represent a fact as a triple (r, t, c^t) consisting of a relation $r \in \mathcal{R}$, an entity pair $t \in \mathcal{T}$ and a vector of counts of contextual patterns $c^t \in \mathcal{S}$. The contextual pattern c^t vector represents the counts of surface patterns that have been observed together with tuple t in a text corpus, normalized to sum to one. We generate $\mathbf{f}$, a fact's feature vector, by concatenating vectors encoding each of r,t and c^t.

The contextual patterns can be thought of as indicators of the textual contexts in which t is likely to be found. By explicitly modeling the contextual patterns we are able to capture the correlations between them and the KB relations. The contextual patterns not only provide evidence of the surface patterns that are descriptive of the entity pair in the text corpus, but crucially, they allow the model to learn which combinations of surface patterns are indicative of certain knowledge base relations. This enables the model to draw on statistical evidence from surface patterns across a text corpus in order to derive more reliable estimates for the interaction factors of relations. This also gives us the benefit of making the most of surface relations, which are easily obtained but noisy, to learn with very few annotation labels for relations. We can thus exploit any abundant text resource (the web, for instance) to learn relation extractors with very few supervision labels from the KB for a new relation.

For instance, the first row in Figure 1 represents that the tuple *Paris,France* was observed with the surface relation *"is a city in"* and that the same tuple was observed with two contextual patterns, *"is a city in"* and *"is a part of"*, hence each of them have a value of 0.5. Similarly, the sixth row represents that the same tuple *Paris,France* has the KB relation *is_the_capital_of*. This allows the model to learn the interaction between the surface patterns *"is a city in"* and *"is a part of"* and the KB relation *is_the_capital_of* more reliably. Furthermore, consider that we want to predict which is a more likely entity tuple between *London, United Kingdom* and *London, France* for the knowledge base

relation *"is_located_in"*. Observe that the tuples *London, United Kingdom* and *Paris, France* have more contextual pattern overlap than the tuple *London, France*. The proposed approach would be aware of such correlations to give a higher score for the fact *(London,is_located_in,United Kingdom)* than *(London,is_located_in,France)*.

We next describe how we model the score of fact candidates with a factorization machine model. We encode the relation r and tuple t as one-hot feature vectors of dimensionality $|\mathcal{R}|$ and $|\mathcal{T}|$ respectively. The feature vector $\mathbf{f}$ is made up of the one-hot encoded KB relations/surface patterns, entity tuples and contextual patterns. However, most of the surface patterns in each contextual pattern have 0 value, hence $\mathbf{f}$ is very sparse. We exploit this in order to accelerate the computation of Equation 1 for a candidate fact by ignoring the features with value of 0 and considering only the active ones A and their corresponding vector representations, which yields the following scoring function for a fact:

$$s(\mathbf{f}) = \sum_{a \in A} b_a f_a + \sum_{a \in A, a' \in A \setminus a} \langle \phi_a, \phi_{a'} \rangle f_a f_{a'} \tag{2}$$

3.1 Objective Formulation

Given a text corpus, we aim to extract relations between entities of interest, with limited training data from the knowledge base and learn a model that can differentiate between true and false facts, i.e. assign high scores to the former and lower scores to the latter using equation 2. However, only examples of observed true relations between entities (positive facts) are available at training time. In order for the model to effectively discriminate between positive and negative facts, it needs to have also seen examples of negative facts. One way to achieve this is to treat observed relations as true facts and all unobserved relations between entities as false facts. However since the facts we seek to extract are unobserved, this carries the risk that we treat plausible relations between entities as negative, which can consequently lead to inferior model performance. Following previous work (Riedel et al., 2013; Petroni, Del Corro, and Gemulla, 2015), we make use of an alternative approach, which is to instead treat unobserved facts as unknowns, and left for the model to infer. This is achieved using a ranking-based objective, which optimizes to rank observed facts higher that unobserved ones. Concretely, we make use of the Bayesian Personalized Ranking (BPR) (Rendle et al., 2009) objective, which optimizes to maximize difference between the score of observed and unobserved facts. Given a set of observed F^+ and unobserved F^- facts, we estimate model parameters Θ that satisfy the following objective:

$$\arg\min_{\Theta} \; -\sum_{\substack{\mathbf{f}^+ \in F^+ \\ \mathbf{f}^- \in F^-}} \log\left(1 + e^{\delta(\mathbf{f}^+, \mathbf{f}^-)}\right) + \lambda \|\Theta\|^2 \tag{3}$$

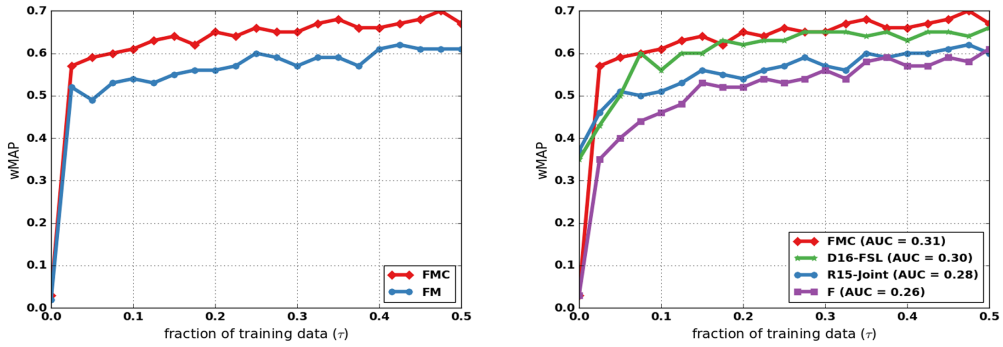
where $\delta(\mathbf{f}^+, \mathbf{f}^-) = \mathbf{s}(\mathbf{f}^+) - \mathbf{s}(\mathbf{f}^-)$ and λ is a regularization hyper parameter. The objective (3) essentially maximizes the difference $\delta(\mathbf{f}^+, \mathbf{f}^-)$ between the scores of observed and unobserved facts.

Note that the set F^- is unobserved and is generated automatically from F^+ by random sampling. Specifically, in each iteration and for every positive fact $\mathbf{f}^+$ in the current batch, we fix the relation r and randomly select an entity pair $t' \in E$, such that the triple $(r, t', c^{t'})$ has not been observed.

4 Training and Evaluation

For all experiments, we make use of a latent dimension size of 100, L_2 regularization penalty of 0.01, and ran our model for 1000 epochs. Our system is implemented in Tensorflow (Abadi et al., 2015), and uses Adam (Kingma and Ba, 2014) for optimization, with a learning rate of 1×10^{-4} and batch size of 1024. We sample one unobserved fact at random per positive fact during training.

We make use of the same evaluation setup as Riedel et al. (2013), by retrieving for each relation the top 1000 entity tuples from each system, the top 100 of which is pooled and manually annotated. These provided a set of results that is used to compute precision measures for each system. We computed Mean Average Precision (MAP) and weighted Mean Average Precision (wMAP) for each run. While MAP computes the expectation of average precision scores across all the relations for each



(a) (b)

Figure 2: (a) : One-shot comparison between two model variants : without contextual pattern embeddings (*FM*) and with contextual pattern embeddings (*FMC*). (b) : One-shot comparison of model *FMC* with previous work. Results obtained from Demeester, Rocktäschel, and Riedel (2016).

system, weighted MAP takes into account the number of true facts for each relation in computing this expectation.

5 Experiments and Results

For our experiments, we make use of the dataset of Riedel et al. (2013), which consists of data from the *New York Times (NYT)* corpus (Sandhaus, 2008). The corpus has been preprocessed with a named entity recognizer and the entities have been linked, where possible, with their corresponding Freebase (Bollacker et al., 2008) entities. The shortest dependency path between each pair of entities in a sentence has also been extracted as the surface pattern.

In the one-shot experiments, we perform evaluations with a fraction $\tau \in [0, 0.5]$ of the training labels for each relation. We make use of the same dimensionality for the embeddings and the same pre-processing (named entity recognition and linking, syntactic parsing) as the approaches we compare with in order to ensure a fair comparison.

Figure 2a presents the results of one-shot experiments for the model that utilizes contextual pattern information (*FMC*), and a variant of it that does not (*FM*). The figure shows that the difference in performance between models *FMC* and *FM* is wider when less supervision data is available. These results demonstrate that the contextual pattern information in model *FMC* enhanced its performance when less supervision labels are available.

Figure 2b presents results of model (*FMC*) compared to state-of-the-art models from Rocktäschel, Singh, and Riedel (2015) (*R15-Joint*) and Demeester, Rocktäschel, and Riedel (2016) (*D16-FSL*). Our approach does not make use of any rules as extra supervision data, and this affected its performance in the zero-shot setting. Nevertheless, it was still able to obtain better coverage, as measured by the wMAP AUC, despite not using any extra supervision. This is because it was able to utilize the contextual pattern representations to better model the relationship between entities, thus requiring less supervision.

In order to assess how well model *FMC* generalizes to the fully supervised setting, we also perform a diagnostic experiment making use of the full training set. Results of this experiment as shown in Table 1 for several models from the literature (M09: Mintz et al. (2009), Y11: Yao, Riedel, and Mccallum (2011), S12: Surdeanu et al. (2012), R13-*: Riedel et al. 2013). We observe that model *FMC* compares favorably with other the models in this setting as well, demonstrating the usefulness of explicit modeling of contextual patterns in both the limited and full supervision settings.

Relation	#	M09	Y11	S12	R13-N	R13-F	R13-NF	R13-NFE	FMC
person/company	104	0.67	0.63	0.69	0.72	0.75	0.75	0.78	**0.80**
location/containedby	75	0.48	0.51	0.53	0.42	*0.68*	0.66	*0.68*	*0.68*
person/nationality	30	0.13	**0.38**	0.12	0.13	0.18	0.18	0.20	0.20
author/works_written	29	0.50	0.51	0.52	0.45	0.61	0.63	**0.69**	0.67
parent/child	19	0.14	0.25	0.62	0.46	0.76	0.78	0.76	**0.79**
person/place_of_death	19	0.79	0.79	0.86	**0.89**	0.83	0.85	0.86	0.83
person/place_of_birth	18	0.78	0.75	0.82	0.50	0.83	0.81	**0.89**	0.81
neighborhood/neighborhood_of	12	0.00	0.00	0.08	0.43	0.65	0.66	**0.72**	0.62
person/parents	7	0.24	0.27	*0.58*	0.56	0.53	*0.58*	0.39	0.56
company/founders	4	0.25	0.25	0.53	0.24	0.77	**0.80**	0.68	0.67
film/directed_by	4	0.06	0.15	0.25	0.09	0.26	0.26	**0.30**	0.07
sports_team/league	4	0.00	0.43	0.18	0.21	0.59	**0.70**	0.63	0.48
team/arena_stadium	3	0.00	0.06	0.06	0.03	0.08	*0.09*	0.08	*0.09*
team_owner/teams_owned	2	0.00	0.50	0.70	0.55	0.38	0.61	**0.75**	0.63
roadcast/area_served	2	*1.00*	0.50	*1.00*	0.58	0.58	0.83	*1.00*	0.58
structure/architect	2	0.00	0.00	*1.00*	0.27	*1.00*	*1.00*	*1.00*	*1.00*
composer/compositions	2	0.00	0.00	0.00	0.50	0.67	*0.83*	0.12	*0.83*
person/religion	1	0.00	*1.00*	*1.00*	0.50	*1.00*	*1.00*	*1.00*	*1.00*
film/produced_by	1	*1.00*	*1.00*	*1.00*	*1.00*	0.50	0.50	0.33	*1.00*
MAP		0.32	0.42	0.55	0.45	0.61	0.66	0.63	0.65
Weighted MAP		0.48	0.51	0.56	0.52	0.66	0.66	0.68	0.68

Table 1: Results using the full training dataset. The # column is the number of true facts in the test pool. Winners are in bold, tied winners in italics.

6 Conclusion

We considered the task of learning to extract relations with few annotated labels. We proposed a model that utilized contextual patterns, which is readily available within the text itself. We showed that our approach improved in extraction accuracy compared to previous approaches. While we have represented each surface pattern within a contextual pattern with a single low-rank representation, a future direction for our work is investigating the use of compositional representations for the surface patterns, which have been shown to lead to better modeling of knowledge base relations (Toutanova et al., 2015; Verga et al., 2016). To encourage further work in this area, we make our code and data publicly available[1].

Acknowledgments

We are grateful to the reviewers for their comments and suggestions for future work. We are also grateful to Sebastian Riedel and members of the University College London's Machine Reading Group for helpful discussions. This work is supported by the European Union's H2020 research and innovation project SUMMA under grant agreement No 688139, and by NVIDIA through a hardware grant.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data* 1247–1250.

[1] Code and data will be available after review at https://github.com/sheffieldnlp/contextual_pattern_embeddings

Bunescu, R. C., and Mooney, R. J. 2006. Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems* 18:171.

Carlson, A.; Betteridge, J.; Wang, R. C.; Hruschka Jr, E. R.; and Mitchell, T. M. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, 101–110. ACM.

Culotta, A., and Sorensen, J. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 423. Association for Computational Linguistics.

Curran, J. R.; Murphy, T.; and Scholz, B. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*.

Demeester, T.; Rocktäschel, T.; and Riedel, S. 2016. Lifted Rule Injection for Relation Embeddings. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)* 1389–1399.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(4):594–611.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, 3.

Miller, E.; Matsakis, N.; and Viola, P. 2000. Learning from one example through shared densities on transforms. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition.* 1.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* 1003–1011.

Petroni, F.; Del Corro, L.; and Gemulla, R. 2015. CORE: Context-Aware Open Relation Extraction with Factorization Machines. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 1763–1773.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* 452–461.

Rendle, S. 2010. Factorization machines. *Proceedings - IEEE International Conference on Data Mining, ICDM* 995–1000.

Riedel, S.; Yao, L.; Marlin, B. M.; and McCallum, A. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*.

Rocktäschel, T.; Singh, S.; and Riedel, S. 2015. Injecting Logical Background Knowledge into Embeddings for Relation Extraction. *North American Association for Computational Linguistics* 1119–1129.

Sandhaus, E. 2008. The New York Times Annotated Corpus LDC2008T19. DVD. Linguistic Data Consortium, Philadelphia.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance Multi-label Learning for Relation Extraction. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP '12* 455–465.

Toutanova, K.; Chen, D.; Pantel, P.; Poon, H.; Choudhury, P.; and Gamon, M. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, volume 15, 1499–1509.

Verga, P.; Belanger, D.; Strubell, E.; Roth, B.; and McCallum, A. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of NAACL-HLT*, 886–896.

Weibl, J.; Bouchard, G.; and Riedel, S. 2016. A factorization machine framework for testing bigram embeddings in knowledgebase completion. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC)*.

Yao, L.; Riedel, S.; and Mccallum, A. 2011. Structured Relation Discovery using Generative Models. *Technology* 1456–1466.

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