Feature-Rich Networks for Knowledge Base Completion

Alexandros Komninos
Department of Computer Science
University of York
York, YO10 5GH
United Kingdom
ak1153@york.ac.uk

Suresh Manandhar
Department of Computer Science
University of York2
York, YO10 5GH
United Kingdom
suresh@cs.york.ac.uk

Abstract

We propose jointly modelling Knowledge Bases and aligned text with Feature-Rich Networks. Our models perform Knowledge Base Completion by learning to represent and compose diverse feature types from partially aligned and noisy resources. We perform experiments on Freebase utilizing additional entity type information and syntactic textual relations. Our evaluation suggests that the proposed models can better incorporate side information than previously proposed combinations of bilinear models with convolutional neural networks, showing large improvements when scoring the plausibility of unobserved facts with associated textual mentions.

1 Introduction

Knowledge Bases (KB) are an important resource for many applications such as question answering (Reddy et al., 2014), relation extraction (Mintz et al., 2009) and named entity recognition (Ling and Weld, 2012). While large collaborative KBs like Freebase (Bollacker et al., 2008) and DBpedia (Auer et al., 2007) contain facts about million of entities, they are mostly incomplete and contain errors. A large amount of research has been dedicated to automatically extend knowledge bases, a task called Entity Linking or Knowledge Base Completion (KBC). Proposed approaches to KBC either reason about the internal structure of the KB, or utilize external data sources that indicate relations between the entities in the KB.

A very successful approach to KBC is latent feature models (Nickel et al., 2011; Bordes et al., 2013; Socher et al., 2013; Nickel et al., 2016). Such models embed the symbols of the KB into a low dimensional space and assign a score to unseen triples as a function of the latent feature representations. Most approaches define a scoring function as a linear or bilinear operator. Latent feature models have shown good performance when considering the internal structure of KBs and are scalable to very large datasets.

Utilizing textual data or other external resources for KBC is a challenging task but has the potential of constantly updating KBs as new information becomes available. A line of work uses the KB as a means to obtain distant supervision to train relation extraction systems that classify textual mentions into one of the KBs relations (Mintz et al., 2009; Hoffmann et al., 2011; Surdeanu et al., 2012). State-of-the-art approaches for KBC with external textual data are obtained by latent feature models that jointly embed the KB symbols and text relations into the same space (Riedel et al., 2013; Toutanova et al., 2015). The benefit of such models over relation extraction systems is that they can combine both the internal structure of the KB and textual information to reason about the plausibility of unobserved facts.

A commonly used approach for augmenting a KBC given an aligned text corpus is by adopting a Universal Schema (Riedel et al., 2013), where extracted textual relations between entities are directly added to the knowledge graph and treated the same as KB relations. This allows application of any latent variable model defined over triples to jointly embed the KB and text relations to the same space. An extension to the Universal Schema approach was proposed by (Toutanova et al., 2015), where representations of text relations are formed compositionally by Convolutional Neural Networks (CNNs) and then composed with entity vectors by a bilinear model to score a fact. However, these models show only moderate improvement when incorporating tex-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 324–329
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2051

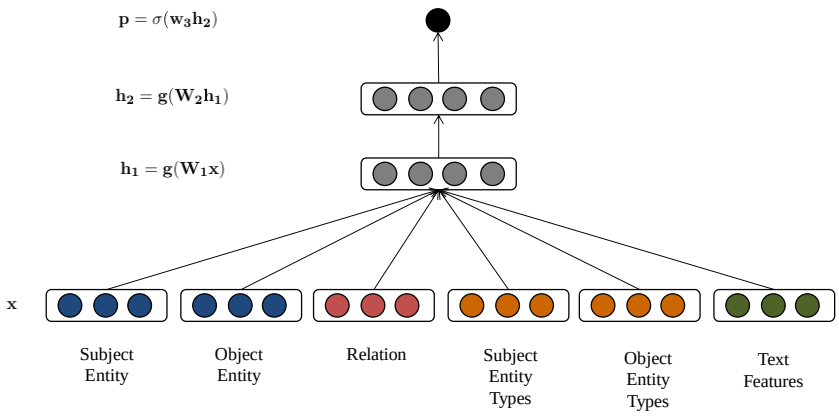


Figure 1: Feature-Rich Network with all the aligned feature types associated with a fact.

tual relations.

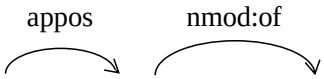
A limitation of the Universal Schema approach for joint embedding of KBs and text is that information about the correspondence between KB and text relations is only implicitly available through their co-occurrence with entities. Text relations can often be noisy and pairs of entities can co-occur in the same sentence without sharing a semantic relation. In addition, there is usually a mismatch in the relations found in the KB and those expressed in text. The model has to learn the alignment between KB and text relations without explicit evidence of co-occurrence between the two, and then propagate that information through the entity embeddings in order to score unseen KB triples.

We propose a different approach to combine KB and textual evidence, where the textual relations are not part of the same graph but are treated as side evidence. In our setting, a fact does not necessarily consist of a (sbj, rel, obj) triple, but as an n-tuple where extra elements are formed by extracting additional information from the KB and aligned side resources such as text. We score the probability of the tuple being true by learning latent representations for each element of the tuple, and then learning a composition and scoring function parameterized by a Multilayer Perceptron (MLP). We choose MLPs as they are a generic method to model interactions between latent features without having to specify the form of a composition operator for tuples of different arity. When scoring the plausibility of unseen facts, all the side evidence associated with that fact becomes explicit through the n-tuple.

We evaluate the ability of the proposed Feature-Rich Networks (FRN) for KBC on the challenging FB15k-237 (Toutanova et al., 2015). We compare the performance of bilinear models to an MLP when facts are represented as simple triples, and the contribution of two additional types of aligned information: entity types and textual relation mentions from a side corpus. We also evaluate the contribution of initializing feature representations from external models. Evaluation suggests that while MLPs and bilinear models perform similarly when treating facts as triples of KB symbols, the proposed approach can better utilize additional textual data than a combination of CNNs with bilinear models, showing large improvements in predicting unseen facts when they have linked relation mentions in text.

2 Model Definition

Knowledge Bases can be represented as a directed graph where nodes are entities $e \in E$ and edges are typed relations $r \in R$. A fact in the KB is encoded as a triple (e_s, r, e_o), where e_s is the subject entity and e_o is the object entity. Starting with an existing KB consisting of a set of observed facts, our goal is to reason about the plausibility of unobserved facts, given some additional external resource. In our proposed model, we expand the representation of a fact to an n-tuple by considering alignments of the additional resource with elements of the triple. Our most expressive model encodes a fact as $\mathcal{X} = (e_s, r, e_o, t_s, t_o, T_{o,s})$, where t_s, t_o are associated representation of types of the two entities, and $T_{o,s}$ is the aligned textual evidence associated with a pair of entities from a side corpus. Representations of entities and entity types are shared between subjects and objects.



subject entity	/m/02vyw
object entity	/m/07g1sm
relation	/film/director/film
subject entity types	/people/person /film/director/ /award/award_winner
object entity types	/film/film /award/award_winning_work
text features	appos^{-1}_Esbj director appos_director of nmod:of^{-1}_director nmod:of_Eobj

Extracted features for a KB fact with a single associated textual relation mention.

2.0.1 Feature Rich Networks

We model the probability of an n-tuple being true with an MLP that learns to compose and score the compatibility of the features associated with it. Features for each individual element of the tuple are assigned low dimensional embeddings which are concatenated to form the input to the MLP. The embeddings are jointly learned with the composition and scoring model through back-propagation. The probability of a fact being true is given by:

$$p(\mathcal{X} = 1) = \sigma(\mathbf{w_3} \cdot g(\mathbf{W_2} \cdot g(\mathbf{W_1} \cdot \mathbf{x})) \quad (1)$$

$$\mathbf{x} = v(e_s); v(r); v(e_o); v(t_s); v(t_o); v(T_{s,o}) \quad (2)$$

where W_1, W_2, w_3 are the weights of the network, $g(\bullet)$ is a non-linear function applied element-wise, $\sigma(\bullet)$ is the sigmoid function and $v(\bullet)$ are latent feature representations of each element of the tuple. We use Rectified Linear Units as non-linearities (Nair and Hinton, 2010).

2.1 Additional Features

We create compositional representations for the entity types and textual relation mentions with simple aggregation functions of their feature embeddings. Although not considered in this work, the overall approach is highly modular allowing for each component to be modelled by a different kind of network.

Freebase Entity Types

Entities in Freebase can have multiple types assigned to them. While entity types are explicitly provided in Freebase, we instead learn type representations by considering observed relations in the training set. Each relation in Freebase is encoded as a domain/type/property of the subject entity. We extract the set of all triples where an entity takes the subject position, and keep the domain/type part as a type feature of that entity. We aggregate embeddings of all the observed discrete features using summation followed by L2-normalization to create the final representation of an entity's type. We use a special UNKNOWN symbol for entities with no observed types in the training set (i.e., entities that do not appear as subject of a triple). We create entity type representations for both subject and object entities and concatenate them to the input vector of the network.

Text Relations

We use a side corpus where pairs of entities are linked to the KB and take the shortest dependency path connecting them as a textual relation mention. Since textual relations are tied to entity pairs, we collect all mentions for a given entity pair and associate them with a fact. This results in a set of phrases that act as textual evidence for relations of an entity pair.

We create a representation of the associated text for each entity pair by using a Neural Bag of Words model augmented with dependency features. A dependency feature is a symbol for a word having a specific dependency relation, such as compound_knowledge, compound^{-1}_base for the knowledge base noun compound. Similar to the Entity Type representations, embeddings of words and dependency features are aggregated by summation followed by L2-normalization, and a special UNKNOWN symbol is assigned to tuples whose pair of entities does

Model	All		With Mentions		Without Mentions	
	MRR	H@10	MRR	H@10	MRR	H@10
KB only						
F	16.9	24.5	26.4	49.1	13.3	15.5
E	33.2	47.6	25.5	37.8	36.0	51.2
DistMult	35.7	52.3	26.0	39.0	39.3	57.2
E + DistMult	37.3	55.2	28.6	42.9	**40.5**	59.8
FRN trp	35.8	55.3	28.7	44.3	38.6	59.7
FRN trp + types	36.0	56.0	28.2	45.0	39.0	60.3
FRN trp + types + init	**37.6**	**57.5**	**30.5**	**48.3**	40.4	**61.1**
KB and text						
Conv-F	19.2	28.4	34.9	63.7	13.3	15.4
Conv-E	33.2	47.6	25.5	37.8	36.0	51.2
Conv-DistMult	36.6	53.5	28.3	43.4	39.7	57.2
Conv-E + Conv-DistMult	40.1	58.1	33.9	49.9	**42.2**	**61.1**
FRN trp + types + text	38.1	58.3	**45.4**	**68.8**	35.2	54.2
FRN trp + types + text + init	**40.3**	**62.0**	44.1	68.3	38.7	59.5

Table 1: Evaluation results on the FB15k-237 dataset. Results for F,E,DistMult and their CNN versions are reported from (Toutanova et al., 2015). With/Without Mentions indicates KB facts with/without aligned textual relations for their entity pair.

not have textual relation mentions. While our text representation component is quite simple, similar models have shown competitive performance on modelling short text (Komninos and Manandhar, 2016).

Initialization with Pre-trained Embeddings
We experiment with pre-trained embeddings to initialize the entity vectors and text feature embeddings of our model. Text feature embeddings are initialized from an available dependency based skip-gram model trained on Wikipedia (Komninos and Manandhar, 2016). Features that are not included in the vocabulary of the pre-trained model are initialized with a random vector from a normal distribution with zero mean and same variance as the set of pre-trained embeddings. For entity vectors, we retrieve the English name of the entity from Freebase and construct a representation by averaging the embeddings of the words appearing in the name. Entities that do not have a name property are initialized randomly.

2.2 Training

The network weights are optimized by minimizing the binary cross-entropy loss over mini-batches using the AdaM optimizer (Kingma and Ba, 2014). To avoid the large computational cost of training with all possible unobserved facts, we make use of negative sampling. The loss function is:

$$L(\Theta) = -\sum_{|\mathcal{X}_p|} \log p(\mathcal{X}_p) - \sum_{|\mathcal{X}_n|} \log(1 - p(\mathcal{X}_n)) \tag{3}$$

where Θ are all the parameters of the network including the feature embeddings, $\mathcal{X}_p$ are the observed facts in the training set and $\mathcal{X}_n$ are randomly drawn unobserved facts. We construct the negative samples by fixing the subject entity and relation, and uniformly sampling an object entity with the restriction that the resulting triple is not included in the training set. We then expand the triple with entity type and text alignments. This negative sampling schedule follows the evaluation procedure, where the network has to rank triples that only differ in the object entity position. Experiments in the validation set indicated that for a fixed number of negative samples, only considering negative samples that differ in the object position performs better than also including negative samples for the subject position.

3 Evaluation

3.1 Dataset and Evaluation Protocol

The FB15k237 dataset consists of about 15k entities and 237 relations derived from the FB15k dataset (Toutanova et al., 2015). This sub-

set of relations does not contain redundant relations that can be easily inferred, resulting in a more challenging task compared to the original FB15k dataset. There are 310,116 triples in the dataset split into 272,115/17,535/20,466 for training/validation/testing. In addition to the KB, the dataset includes dependency paths of approximately 2.7 million relation instances of linked entity mentions extracted from the ClueWeb corpus (Gabrilovich et al., 2013).

Evaluation follows the procedure of (Toutanova et al., 2015). Given a positive fact in the test set, the subject entity and relation are fixed and models have to rank all facts formed by the object entities appearing in the training set. The reported metrics are mean reciprocal rank (MRR) and hits@10. Hits@10 is the fraction of positive facts ranked in the top 10 positions. Positive facts in the training, validation and test set are removed before ranking.

3.2 Implementation Details

Hyperparameters of the model were chosen by maximizing MRR on the validation set. We use two 300-dimensional hidden layers for the MLP, and dimensions of feature embeddings are: 300 for entity and text features, 100 for relations and 20 for entity type features. The number of negative samples was set to 20 as increasing their number only resulted in minor gains, and the batch size was set to 420. Best models were chosen among 20 epochs of training by monitoring validation MRR. Models with embedding initializations converged within the first 10 epochs. Initialization in the text model includes initializing entity and relation embeddings from a model without text mentions.

3.3 Results

We compare our Feature-Rich Networks with the bilinear models F and E (Riedel et al., 2013), model DistMult (Yang et al., 2014) and their CNN augmented versions (Toutanova et al., 2015). Results can be seen in Table 1.

We first observe that when modelling just KB triples, the MLP model outperforms individual bilinear formulations, performing similarly to the best combination of DistMult + E. This shows that an additive combination of bilinear models is a strong baseline even though it does not use additional parameters other than embeddings to compose and score triples. The addition of entity type information has a positive but small contribution to performance. This is not surprising as entity type information is extracted from observed relations, and latent feature models can effectively learn that during training. On the other hand, initializing entity embeddings with averaged word embeddings of their names results in a substantial performance gain of about 1.5 points in both MRR and hits@10. In general, we observe that all models perform worse on facts with textual relation mentions when they have not access to such mentions.

When textual relation mentions are added, we observe that our proposed model increases its performance score about 3 points in MRR and 4.5 in hits@10 compared to the best model that does not include text. Contrary to the conv-bilinear models, the performance gain is much larger for facts with textual mentions, reaching an additional 15/20 in MRR/hits@10 respectively. We attribute this gain to the explicitly represented textual relation alignments with the KB symbols as encoded by the expanded tuple representations, and the non-linear composition of its elements by the MLP. We also notice that embedding initialization performs better overall.

4 Conclusion

In this paper, we propose joint modelling of Knowledge Bases and text with Feature-Rich Networks. Our models can learn to combine information from different sources and better utilize noisy information from text than bilinear models augmented with convolutional neural networks. Besides text, we experiment with entity types and initialization with pre-trained embeddings, getting positive gains in performance. An interesting direction for future work is to combine our models with additional aligned information, such as multiple KBs and to experiment with different components such as CNNs or LSTMs for text encoding.

Acknowledgements

Alexandros Komninos was supported by EPSRC via an Engineering Doctorate in LSCITS. Suresh Manandhar was supported by EPSRC grant EP/I037512/1, A Unified Model of Compositional & Distributional Semantics: Theory and Application.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, Springer, pages 722–735.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, pages 1247–1250.

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*. pages 2787–2795.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by* 5.

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*. Association for Computational Linguistics, pages 541–550.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Alexandros Komninos and Suresh Manandhar. 2016. Dependency based embeddings for sentence classification tasks. In *Proceedings of NAACL:HLT*. Association for Computational Linguistics, pages 1490–1500.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *AAAI*. Citeseer.

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*. Association for Computational Linguistics, pages 1003–1011.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. pages 807–814.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104(1):11–33.

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.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics* 2:377–392.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas .

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*. pages 926–934.

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*. Association for Computational Linguistics, pages 455–465.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*. Citeseer, volume 15, pages 1499–1509.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* .