A Factorization Machine Framework for Testing Bigram Embeddings in Knowledgebase Completion

Johannes Welbl, **Guillaume Bouchard** and **Sebastian Riedel**
University College London
London, UK
{j.welbl, g.bouchard, s.riedel}@cs.ucl.ac.uk

Abstract

Embedding-based Knowledge Base Completion models have so far mostly combined distributed representations of *individual* entities or relations to compute truth scores of missing links. Facts can however also be represented using pairwise embeddings, i.e. embeddings for *pairs* of entities and relations. In this paper we explore such bigram embeddings with a flexible Factorization Machine model and several ablations from it. We investigate the relevance of various bigram types on the `fb15k237` dataset and find relative improvements compared to a compositional model.

1 Introduction

Present day Knowledge Bases (KBs) such as YAGO (Suchanek et al., 2007), Freebase (Bollacker et al., 2008) or the Google Knowledge Vault (Dong et al., 2014) provide immense collections of structured knowledge. Relationships in these KBs often exhibit regularities and models that capture these can be used to predict missing KB entries. A common approach to KB completion is via tensor factorization, where a collection of fact triplets is represented as a sparse mode-3 tensor which is decomposed into several low-rank sub-components. Textual relations, i.e. relations between entity pairs extracted from text, can aid the imputation of missing KB facts by modelling them together with the KB relations (Riedel et al., 2013).

The general merit of factorization methods for KB completion has been demonstrated by a variety of models, such as RESCAL (Nickel et al., 2011), TransE (Bordes et al., 2013) and DistMult (Yang et al., 2014). These models learn distributed representations for entities and relations (be it as vector or as matrix) and infer the truth value of a fact by combining embeddings for these constituents in an appropriate composition function.

Most of these factorization models however operate on the level of embeddings for single entities and relations. The implicit assumption here is that facts are *compositional*, i.e. that the subject, relation and object of a fact are its atomic constituents. Semantic aspects relevant for imputing its truth can directly be recovered from its constituents when composing their respective embeddings in a score.

For further notation let E and R be sets of entities and relations, respectively. We denote a fact f stating a relation $r \in$ R between subject $s \in$ E and object $o \in$ E as $f = (s, r, o)$. Our goal is to learn embeddings for larger sub-constituents of f than just $s, r,$ and o: we want to learn embeddings also for the entity pair bigram (s, o) as well as the relation-entity bigrams (s, r) and (r, o). As an example, consider Freebase facts with relation `eating/practicer_of_diet/diet` and object `Veganism`. Overall only two objects are observed for this relation and it thus makes sense to learn a joint embedding for bigrams (r, o) together, instead of distinct embeddings for each atom alone and then having to learn their compatibility.

While Riedel et al. (2013) have trained embeddings only for entity pairs, we will in this paper explore the role of general bigram embeddings for KB completion, i.e. also the embeddings for

other possible pairs of entities and relations. This is achieved using a Factorization Machine (FM) framework (Rendle, 2010) that is modular in its feature components, allowing us to selectively add or discard certain bigram embeddings and compare their relative importance. All models are empirically compared and evaluated on the `fb15k237` dataset from Toutanova et al. (2015).

In summary, our main contributions are: i) Adressing the question of generic bigram embeddings in a KB completion model for the first time; ii) The adaption of Factorization Machines for this matter; iii) Experimental findings for comparing different bigram embedding models on `fb15k237`.

2 Related Work

In the Universal Schema model (model *F*), Riedel et al. (2013) factorize KB entries together with relations of entity pairs extracted from text, embedding textual relations in the same vector space as KB relations. Singh et al. (2015) extend this model to include a variety of other interactions between entities and relations, using different relation vectors to interact with subject, object or both. Jenatton et al. (2012) also recognize the need to integrate rich higher-order interaction information into the score. Like Nickel et al. (2011) however, their model specifies relationships as relation-specific bilinear forms of entity embeddings. Other embedding methods for KB completion include DistMult (Yang et al., 2014) with a trilinear score, and TransE (Bordes et al., 2013) which offers an intriguing geometrical intuition. Among the aforementioned methods, embeddings are mostly learned for individual subjects, relations or objects; merely model *F* (Riedel et al., 2013) constitutes the exception.

Some methods rely on more expressive composition functions to deal with non-compositionality or interaction effects, such as the Neural Tensor Networks (Socher et al., 2013) or the recently introduced Holographic Embeddings (Nickel et al., 2015). In comparison to the otherwise used (generalized) dot products, the composition functions of these models enable richer interactions between unit constituent embeddings. However, this comes with the potential disadvantage of presenting less well-behaved optimisation problems and being slower to train. Factorization Machines have already been applied in a similar setting to ours by Petroni et al. (2015) who use them with contextual features for an Open Relation Extraction task, but without bigrams.

3 Model

3.1 Brief Recall of Factorization Machines

A Factorization Machine (FM) is a quadratic regression model with low-rank constraint on the quadratic interaction terms.[1] Given a sparse input feature vector $\phi = (\phi_1, \ldots, \phi_n)^T \in \mathbb{R}^n$, the FM output prediction $X \in \mathbb{R}$ is

$$ X = \langle \mathbf{v}, \phi \rangle + \sum_{i,j=1}^{n} \langle \mathbf{w}_i, \mathbf{w}_j \rangle \cdot \phi_i \phi_j \qquad (1) $$

where $\mathbf{v} \in \mathbb{R}^n$, and $\forall i, j = 1, \ldots n : \mathbf{w}_i, \mathbf{w}_j \in \mathbb{R}^k$ are model parameters with $k \ll n$ and $\langle \cdot, \cdot \rangle$ denotes the dot product. Instead of allowing for an individual quadratic interaction coefficient per pair (i, j), the FM assumes that the matrix of quadratic interaction coefficients has low rank k; thus the interaction coefficient for feature pair (i, j) is represented by an inner product of k-dimensional vectors $\mathbf{w}_i$ and $\mathbf{w}_j$. The low rank constraint (i) provides a strong form of regularisation to this otherwise over-parameterized model, (ii) pools statistical strength for estimating similarly profiled interaction coefficients and (iii) retains a total number of parameters linear in n. In summary, with a FM one can efficiently harness a large set of sparse features and interactions between them while retaining linear memory complexity.

3.2 Feature Representation for Facts

For the KB completion task we will use a FM with unit and bigram indicator features to learn low-rank embeddings for both. To formalize this, we will refer to the elements of the set $\mathtt{U}_f = \{s, r, o\}$ as *units* of fact f, and to the elements of $\mathtt{B}_f = \{(s, r), (r, o), (o, s)\}$ as *bigrams* of fact f. Let $\iota_u \in \mathbb{R}^{|\mathbb{E}|+|\mathbb{R}|}$ be the one-hot indicator vector that encodes a particular unit[2] $u \in (\mathbb{E} \cup \mathbb{R})$. Furthermore we define $\iota_{(s,r)} \in \mathbb{R}^{|\mathbb{E}||\mathbb{R}|}$, $\iota_{(r,o)} \in \mathbb{R}^{|\mathbb{R}||\mathbb{E}|}$ and $\iota_{(o,s)} \in \mathbb{R}^{|\mathbb{E}|^2}$

---

[1]We disregard the more general extension to higher-order interactions that is described in the original FM paper and only consider the quadratic case. Also, we omit the global model bias as we found that it was not helpful for our task empirically.

[2]For subject and object the same entity embedding is used.

to be the one-hot indicator vectors encoding particular bigrams. Our feature vector $\phi(f)$ for fact $f = (s, r, o)$ then consists of simply the concatenation of indicator vectors for all its units and bigrams:

$$\phi(f) = \text{concat}(\iota_s, \iota_r, \iota_o, \iota_{(s,r)}, \iota_{(r,o)}, \iota_{(o,s)}) \quad (2)$$

This sparse set of features provides a rich representation of a fact with indicators for subject, relation and object, as well as any pair thereof.

3.3 Scoring a Fact

Harnessing the expressive benefits of a sigmoid link function for relation modelling (Bouchard et al., 2015), we define the truth score of a fact as $g(f) = \sigma(X_f)$ where σ is the sigmoid function and X_f is given as output of the FM model (1) with unit and bigram features $\phi(f)$ as defined in (2):

$$X_f = \langle \phi(f), \mathbf{v} \rangle + \sum_{i,j=1}^{n} \langle \mathbf{w}_i, \mathbf{w}_j \rangle \cdot \phi_i(f)\phi_j(f) \quad (3)$$

Since our feature vector $\phi(f)$ is sparse with only six active entries, we can re-express (3) in terms of the activated embeddings which we directly index by their respective units and bigrams:

$$X_f = \sum_{c \in (\mathbf{U}_f \cup \mathbf{B}_f)} v_c + \sum_{c_1, c_2 \in (\mathbf{U}_f \cup \mathbf{B}_f)} \langle \mathbf{w}_{c_1}, \mathbf{w}_{c_2} \rangle \quad (4)$$

This score comprises all possible interactions between any of the units and bigrams of f.

3.4 Model Ablations for Investigating Particular Bigram Embeddings

The score (4) can easily be modified and individual summands removed from it. In particular, when discarding all but one summand, model F is recovered, i.e. with $c_1 = (s, o)$; $c_2 = r$. On the other hand, alternatives to model F with other bigrams than entity pairs can be tested by removing all summands but the one of a single bigram $b \in B_f$ vs. the remaining complementary unit $u \in U_f$:

$$X_f^{u,b} = \mathbf{v}_u + \mathbf{v}_b + \langle \mathbf{w}_u, \mathbf{w}_b \rangle \quad (5)$$

This general formulation offers us a method for investigating the relative impact of all combinations of bigram vs. unit embeddings besides model F, namely the models with $u = s$; $b = (r, o)$ and with $u = o$; $b = (s, r)$.

3.5 Training Objective

Given sets of true training facts Ω^+ and sampled negative facts Ω^-, we minimize the following loss:

$$-\sum_{f \in \Omega^+} \log(1 + e^{X_f}) + \frac{1}{\eta} \sum_{f \in \Omega^-} \log(1 + e^{X_f}) \quad (6)$$

where the parametrization of X_f is learned. We use the hyperparameter $\eta \in \mathbb{R}^+$ for denoting the ratio of negative facts that are sampled per positive fact so that the contributions of true and false facts are balanced even if there are more negative facts than positives. The loss differs from a standard negative log-likelihood objective with logistic link, but we found that it performs better in practice. The intuition comes from the fact that instead of penalizing badly classified positive facts, we put more emphasis (i.e. negative loss) on positive facts that are correctly classified. Since we used an L_2 regularization and the loss is asymptotically linear, the resulting objective is continuous and bounded from below, guaranteeing a well defined local minimum.

4 Experiments

The bigram embedding models are tested on `fb15k237` (Toutanova et al., 2015), a dataset comprising both Freebase facts and lexicalized dependency path relationships between entities.

Training Details and Evaluation We optimized the loss using AdaM (Kingma and Ba, 2015) with minibatches of size 1024, using initial learning rate 1.0 and initialize model parameters from $\mathcal{N}(0, 1)$. Furthermore, a hyperparameter $\tau < 1$ like in (Toutanova et al., 2015) is introduced to discount the importance of textual mentions in the loss. When sampling a negative fact we alter the object of a given training fact (s, r, o) at random to $o' \in \mathbb{E}$, and repeat this η times, sampling negative facts every epoch anew. There is a small implied risk of sampling positive facts as negative, but this is rare and the discounted loss weight of negative samples mitigates the issue further. Hyperparameters (L_2-regularisation, η, τ, latent dimension k) are selected in a grid search for minimising Mean Reciprocal Rank (*MRR*) on a fixed random subsample of size 1000 of the validation set. All reported results are for the test set. We use the competitive unit model

	Model	τ	overall *HITS@*			*MRR*		
			1	3	10	overall	no TM	with TM
	DistMult	0.0	18.2	27.0	37.9	24.8	28.0	16.2
	full FM	0.0	20.1	28.7	38.9	26.4	29.3	18.3
(*)	(s, o) vs. r	1.0	2.1	3.8	6.5	3.5	0.0	13.1
(**)	(r, o) vs. s	0.1	24.9	34.8	45.8	32.0	34.7	24.8
(***)	(s, r) vs. o	0.0	9.0	17.3	29.9	15.6	17.3	10.9
	(*) + (**) + (***)	0.1	**25.9**	**36.2**	**47.4**	**33.2**	**35.0**	**28.3**

Table 1: Test set metrics for different models and varying unit and bigram embeddings on `fb15k237`, all performance numbers in % and best result in bold. The optimal value for τ is indicated as well.

DistMult as baseline and employ the same ranking evaluation scheme as in (Toutanova et al., 2015) and (Toutanova and Chen, 2015), computing filtered *MRR* and *HITS* scores whilst ranking true test facts among candidate facts with altered object. Particular bigrams that have not been observed during training have no learned embedding; a 0-embedding is used for these. This nullifies their impact on the score and models the back-off to using nonzero embeddings.

Results   Table 1 gives an overview of the general results for the different models. Clearly, some of the bigram models can obtain an improvement over the unit DistMult model. In a more fine-grained analysis of model performances, characterized by whether entity pairs of test facts had textual mentions available in training (*with TM*) or not (*without TM*), the results exhibit a similar pattern like in (Toutanova et al., 2015): most models perform worse on test facts with TM, only model F, which can learn very little without relations has a reversed behavior. A side observation is that several models achieved highest overall MRR with $\tau = 0$, i.e. when not using TM.

The sum of the three more light-weight bigram models performs better than the full FM, even though the same types of embeddings are used. A possible explanation is that applying the same embedding in several interactions with other embeddings (as in the full FM) instead of only one interaction (like in (*)+(**)+(***)) makes it harder to learn since its multiple functionalities are competing.

Another interesting finding is that some bigram types achieve much better results than others, in particular model (**). A possible explanation becomes apparent with closer inspection of the test set: a given test fact f usually contains at least one bigram $b \in \mathbb{B}_f$ which has never been observed yet. In these cases the bigram embedding is 0 by design and only

the offset values are used. The proportions of test facts for which this happens are 73%, 10% and 24% respectively for the bigrams (s, o), (r, o), and (s, r). Thus models (**) and (***) already have a definite advantage over model (*) that originates purely from the nature of the data. A trivial but somehow important lesson we can learn from this is that if we know about the relative prevalence of different bigrams (or more generally: sub-tuples) in our dataset, we can incorporate and exploit this in the sub-tuples we choose.

Finally, for the initial example with relation `eating/practicer_of_diet/diet` and object `Veganism`, we indeed find that in all instances model (**) with its (r, o) embedding gives the correct fact in the top 2 predictions, while the purely compositional DistMult model ranks it far outside the top 10. More generally, cases in which only a single object co-appeared with a test fact relation during training had $95, 3\%$ *HITS@1* with model (**) while only $52, 6\%$ for DistMult. This supports the intuition that bigram embeddings of (r, o) are in fact better suited for cases in which very few objects are possible for a relation.

5   Conclusion

We have demonstrated that FM provide an approach to KB completion that can incorporate embeddings for bigrams naturally. The FM offers a compact unified framework in which various tensor factorization models can be expressed, including model F.

Extensive experiments have demonstrated that bigram models can improve prediction performances substantially over more straightforward unigram models. A surprising but important result is that bigrams other than entity pairs are particularly appealing.

The bigger question behind our work is about

compositionality vs. non-compositionality in a broader class of knowledge bases involving higher order information such as time, origin or context in the tuples. Deciding which modes should be merged into a high order embedding without having to rely on heavy cross-validation is an open question.

Acknowledgments

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS 26*.

Guillaume Bouchard, Sameer Singh, and Theo Trouillon. 2015. On approximate reasoning capabilities of low-rank vector spaces. In *AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA. ACM.

Rodolphe Jenatton, Nicolas L. Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *NIPS 25*, pages 3167–3175. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *The International Conference on Learning Representations (ICLR)*.

Maximilian Nickel, Volker Tresp, and Hans peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 809–816, New York, NY, USA. ACM.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2015. Holographic Embeddings of Knowledge Graphs. Technical report, arXiv, October.

Fabio Petroni, Luciano Del Corro, and Rainer Gemulla. 2015. Core: Context-aware open relation extraction with factorization machines. In Llus Mrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*, pages 1763–1773. The Association for Computational Linguistics.

Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 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)*, June.

Sameer Singh, Tim Rocktaschel, and Sebastian Riedel. 2015. Towards combined matrix and tensor factorization for universal schema relation extraction. In *NAACL Workshop on Vector Space Modeling for NLP*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NIPS 26*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference, Banff, Canada*, pages 697–706.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Workshop on Continuous Vector Space Models and Their Compositionality (CVSC)*.

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 *Empirical Methods in Natural Language Processing (EMNLP)*. ACL Association for Computational Linguistics, September.

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