INITIALIZING ENTITY REPRESENTATIONS IN RELATIONAL MODELS

Teng Long, Ryan Lowe, Jackie Cheung & Doina Precup
School of Computer Science
McGill University
teng.long@mail.mcgill.ca
{ryan.lowe,jcheung,dprecup}@cs.mcgill.ca

ABSTRACT

Recent work in learning vector-space embeddings for multi-relational data has focused on combining relational information derived from knowledge bases with distributional information derived from large text corpora. We propose a simple trick that leverages the descriptions of entities or phrases available in lexical resources, in conjunction with distributional semantics, in order to derive a better initialization for training relational models. Applying this trick to the TransE model results in faster convergence of the entity representations, and achieves small improvements on Freebase for raw mean rank. More surprisingly, it results in significant new state-of-the-art performances on the WordNet dataset, decreasing the mean rank from the previous best 212 to 51. We find that there is a trade-off between improving the mean rank and the hits@10 with this approach. This illustrates that much remains to be understood regarding performance improvements in relational models.

1 BACKGROUND

A key challenge of intelligent machines is the need to communicate with humans and understand relationships between objects described in unstructured text. The goal of our work is to find ways of integrating structured knowledge bases and word embeddings, which remains an open problem (Faruqui et al., 2015; Xu et al., 2014; Fried & Duh, 2015; Yang et al., 2015; Labutov & Lipson, 2013). More concretely, we address the problem of knowledge base completion, in which the goal is to generalize relationships between entities in a structured dataset. Perhaps the most well-known existing approach is *Translating Embeddings* (TransE) (Bordes et al., 2013), which takes a pre-existing semantic hierarchy as input and embeds its relational information into a vector space, where linear relationships between entities are learned. For example, given a relation such as $won(Germany, FIFA\ Worldcup)$, the TransE model learns vector representations for *won*, *Germany*, and *FIFA Worldcup* such that $Germany + won \approx FIFA\ Worldcup$.

Existing work that uses distributional vectors for knowledge base completion assumes that reliable distributional vectors are always available for all of the entities in the hierarchy being modeled. Unfortunately, this assumption does not hold in practice; when moving to a new domain with a new knowledge base, there will likely be many entities or phrases for which there is no or very little distributional information in the training corpus. For example, 50-80% of entities from the benchmark WordNet and Freebase datasets are missing from the embedding dictionaries derived using word2Vec and GloVe models. Thus, a method to derive entity representations that works well for both common and rare entities is needed.

Fortunately, knowledge bases typically contain a short description or definition for each of the entities or phrases. For example, WordNet contains synset glosses, and Freebase contains descriptions for entities. We propose a simple and efficient trick that converts short entity descriptions into vector space representations, with the help of existing word embedding models. These vectors are then used as the input for further training with TransE, in order to incorporate structural information.

2 ARCHITECTURE OF THE APPROACH

2.1 THE TRANSE MODEL

The Translating Embedding (TransE) model (Bordes et al., 2013) has become one of the most popular multi-relational models due to its relative simplicity, scalability to large datasets, and (until recently) state-of-the-art results. It uses a simple additive interaction between vector representations of entities and relations. More precisely, assume a given relationship triplet (h, l, t) is valid; then, the embedding of the object t should be very close to the embedding of the subject h plus some vector in $\mathbb{R}^k$ that depends on the relation l.

For each positive triplet $(h, l, t) \in S$, a negative triplet $(h^{'}, l, t^{'}) \in S^{'}$ is constructed by randomly sampling an entity from E to replace either the subject h or the object t of the relationship. The training objective of TransE is to minimize the dissimilarity measure $d(h + l, t)$ of a positive triplet while ensuring that $d(h^{'} + l, t^{'})$ for the corrupted triplet remains large. This is accomplished by minimizing the hinge loss over the training set:

$$L = \sum_{(h,l,t)\in S} \sum_{(h',l,t')\in S'} [\gamma + d(h + l, t) - d(h^{'} + l, t^{'})]_+$$

where γ is the hinge loss margin and $[x]_+$ represents the positive portion of x. There is an additional constraint that the L_2-norm of entity embeddings (but not relation embeddings) must be 1, which prevents the training process to trivially minimize L by artificially increasing the norms of entity embeddings.

2.2 INITIALIZING REPRESENTATIONS WITH ENTITY DESCRIPTIONS

We propose to leverage some external lexical resource to improve the quality of the entity vector representations. In general, this could consist of product descriptions in a product database, or information from a web resource. For example, in a medical dataset with many technical words, the Wikipedia pages, dictionary definitions, or medical descriptions via a site such as `medilexicon.com` could be leveraged as lexical resources. Similarly, when building language models for social media, resources such as `urbandicionary.com` could be used for information about slang words. For the WordNet and Freebase datasets, we use *entity descriptions* which are readily available.

Although there are many possible ways to incorporate this information, we propose a simple initialization trick which we show to have empirical benefits. In particular, we first decompose the description of a given entity into a sequence of word vectors, and combine them into a single embedding by averaging. We then reduce the dimensionality using principle component analysis (PCA), which we found experimentally was important to avoid overfitting. We obtain these word vectors using distributed representations computed using the skip-gram model (Mikolov et al., 2013), and the GloVe model (Pennington et al., 2014). Approximating compositionality by averaging vector representations is simple, yet has some theoretical justification Tian et al. (2015) and can work well in practice Wieting et al. (2015). This approach is similar to that found in (Chen et al., 2014) and elsewhere, but we show that it has a surprising effectiveness when applied to relational models.

3 EXPERIMENTS

3.1 TRAINING AND TESTING SETUP

We perform experiments on the WordNet (WN) (Miller, 1995) and Freebase (FB15k) (Bollacker et al., 2008) datasets used by the original TransE model. TransE hyperparameters include the learning rate λ for stochastic gradient descent, the margin γ for the hinge loss, the dimension of the embeddings k, and the dissimilarity metric d. We use the optimal hyperparameters from (Bordes et al., 2013): for WN, $\lambda = 0.01$, $\gamma = 2$, $k = 20$, and $d = L_1$-norm; for FB15k, $\lambda = 0.01$, $\gamma = 0.5$, $k = 50$, and $d = L_2$-norm. The values of k were further tested to ensure that $k = 20$ and $k = 50$ were optimal. The distributional vectors used in the entity descriptions are of dimension 1000 for the word2vec vectors with Freebase vocabulary, and dimension 300 in all other cases. Dimensionality

		WN				FB15k			
		Mean rank		Hits@10		Mean rank		Hits@10	
		Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
Prev. models	SE (Bordes et al., 2011)	1,011	985	68.5%	80.5%	273	162	28.8%	39.8%
	TransD (unif) (Ji et al., 2015)	242	229	79.2%	**92.5%**	211	**67**	49.4%	74.2%
	TransD (bern) (Ji et al., 2015)	224	212	**79.6%**	92.2%	194	91	**53.4%**	**77.3%**
	TransE random init.	266	254	76.1%	89.2%	195	92	41.2%	55.2%
	TransE W2V init.	—	—	—	—	195	91	41.3%	55.4%
Our models	TransE W2V entity defs. (NS)	210	192	78.5%	92.1%	195	91	41.6%	55.7%
	TransE GloVe entity defs. (NS)	**63**	**51**	64.6%	73.2%	194	90	41.7%	55.8%
	TransE W2V entity defs.	191	179	77.8%	91.6%	195	91	41.6%	55.6%
	TransE GloVe entity defs.	71	59	75.3%	88.0%	**193**	90	41.8%	55.8%

Table 1: Comparison between random initialization and using the entity descriptions. 'NS' tag indicates stopword removal from the entity descriptions'TransE W2V init' model uses word2vec pre-trained with the Freebase vocabulary.

reduction with PCA was then applied to reduce this to $k = 30$ for WN, and $k = 55$ for FB15k. The model is trained with minibatch SGD and early stopping on the validation set.

We use the same train/test/validation split and evaluation procedure as (Bordes et al., 2013): for each test triplet *(h, l, t)*, we remove entity *h* and *t* in turn, and rank each entity in the dictionary by similarity according to the model. We evaluate using *i)* the *mean* of the predicted ranks, and *ii) hits@10*, which represents the percentage of correct entities found in the top 10 list.

3.2 Results and Analysis

Table 1 summarizes the experimental results, compared to baseline and state-of-the-art relational models. We see that the mean rank is greatly improved for the TransE model with strategic initialization over random initialization. More surprisingly, all of our models achieve state-of-the-art performance for both raw and filtered data, compared to the recently developed TransD model. These results are highly significant with $p < 10^{-3}$ according to the Mann-Whitney U test. Thus, even though our method is simple and straightforward to apply, it can still beat all attempts at more complicated structural modifications to the TransE model on this dataset.

Also interesting is the relationship between the mean rank and hits@10. By changing our model, we are able to increase one at the expense of the other. For example, using word2vec without stopwords gives similar hits@10 to TransD with better mean rank, while using GloVe further improves the mean rank at a cost to hits@10. We conjecture that our model helps avoid 'disasters' where some correct entities are ranked very low, which improves mean rank significantly.

For Freebase, our models slightly outperform the TransE model with random initialization, with p-values of 0.173 and 0.410 for initialization with descriptions (including stopwords) using GloVe and word2vec, respectively. We also see improvements over the case of direct initialization with word2vec. Further, we set a new state-of-the-art for mean rank on the raw data, though the improvement is marginal. The difference in performance between datasets can be partly explained by their different nature: WordNet relations are general and are meant to provide links between concepts, while the Freebase relations are specific and denote relationships between named entities.

4 Conclusion and Future Work

Our initialization trick is simple and leads to significant improvements on WordNet mean rank. More complex methods initialization methods could easily be devised, e.g. by using inverse document frequency (idf) weighted averaging, or by applying the work of Le & Mikolov (2014) on paragraph vectors. Alternatively, distributional semantics could be used as a regularizer, similar to Labutov & Lipson (2013), with learned embeddings being penalized for how far they stray from the pre-trained GloVe embeddings. However, *even with intuitive and straightforward methodology*, leveraging lexical resources can have a significant impact on the results of models for multi-relational data. These insights are perhaps most transferable to domains with many out-of-vocabulary (OOV) words.

REFERENCES

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, 2008.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Proceedings of AAAI*, 2011.

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

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *EMNLP*, pp. 1025–1035. Citeseer, 2014.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, 2015.

Daniel Fried and Kevin Duh. Incorporating both distributional and relational semantics in word representations. In *In Proceedings of ICLR*, 2015.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of ACL*, 2015.

Igor Labutov and Hod Lipson. Re-embedding words. In *Proceedings of ACL*, 2013.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of ICML*, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, 2013.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 2014.

Ran Tian, Naoaki Okazaki, and Kentaro Inui. The mechanism of additive composition. *arXiv preprint arXiv:1511.08407*, 2015.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of CIKM*, 2014.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of ICLR*, 2015.