Leveraging Lexical Resources for Learning Entity Embeddings in Multi-Relational Data

Teng Long, Ryan Lowe, Jackie Chi Kit 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 approach 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 initialization to the TransE model results in significant new state-of-the-art performances on the WordNet dataset, decreasing the mean rank from the previous best of 212 to 51. It also results in faster convergence of the entity representations. 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 Introduction

A surprising result of work on vector-space word embeddings is that word representations that are learned from a large training corpus display semantic regularities in the form of linear vector translations. For example, Mikolov et al. (2013b) show that using their induced word vector representations, $king - man + woman \approx queen$. Such a structure is appealing because it provides an interpretation to the distributional vector space through lexical-semantic analogical inferences.

Concurrent to that work, Bordes et al. (2013) proposed *translating embeddings* (TransE), which takes a pre-existing semantic hierarchy as input and embeds its structure into a vector space. In their model, the linear relationship between two entities that are in some semantic relation to each other is an explicit part of the model's objective function. 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* ≈ *FIFA Worldcup*.

Dataset	Total	W2V found%	GloVe found%
WN	40943	9.7%	51.3%
FB15k	14951	4.0%	20.3%

Table 1: The percentage of WN and FB15k entities that can be found in the pre-trained word2vec (W2V) and GloVe vectors. This does not include the W2V embeddings trained with the FB15k vocabulary[2], which covers 93% of the FB15k entities.

A natural next step is to attempt to integrate the two approaches in order to develop a representation that is informed by both unstructured text and a structured knowledge base (Faruqui et al., 2015; Xu et al., 2014; Fried and Duh, 2015; Yang et al., 2015). However, existing work makes a crucial assumption—that reliable distributional vectors are 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, for example, there will likely be many entities or phrases for which there is no distributional information in

[2] This means that word2vec was trained in the usual way on a large textual corpus, but the vocabulary was truncated to include as many entities from Freebase as possible. Indeed, this is the reason for the small overlap between W2V, GloVe, and the relational databases: after training the word embeddings, the vocabulary must be truncated to a reasonable size, which leaves out many entities from these datasets.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 112–117,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

the training corpus. This important problem is illustrated in Table 1, where most of the entities from WordNet and Freebase are seen to be missing from the distributional vectors derived using Word2Vec and GloVe trained on the Google News corpus. Even when the entities are found, they may not have occurred enough times in the training corpus for their vector representation to be reliable. What is needed is a method to derive entity representations that works well for both common and rare entities.

Fortunately, knowledge bases typically contain a short description or definition for each of the entities or phrases they contain. 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 (see Table 2).

In this paper, we propose a simple and efficient procedure to convert these short descriptions into a vector space representation, with the help of existing word embedding models. These vectors are then used as the input to further training with the TransE model, in order to incorporate structural information. Our method provides a better initialization for the TransE model, not just for the entities that do not appear in the data, but in fact for *all* entities. This is demonstrated by achieving state-of-the-art mean rank on an entity ranking task on two very different data sets: WordNet synsets with lexical semantic relations (Miller, 1995), and Freebase named entities with general semantic relations (Bollacker et al., 2008).

2 Related Work

Dictionary definitions were the core component of early methods in word sense disambiguation (WSD), such as the Lesk algorithm (1986). Chen et al. (2014) build on the use of synset glosses for WSD by leveraging lexical resources. Our work goes further to tie these glosses together with relational semantics, a connection that has not been drawn in the literature before. The integration of lexical resources into distributional semantics has been studied in other lexical semantic tasks, such as synonym expansion (Sinha and Mihalcea, 2009), relation extraction (Kambhatla, 2004), and calculating the semantic distance between concepts (Mohammad, 2008; Marton et al., 2009). We aim to combine lexical resources and other semantic knowledge, but we do so in the context of neural network-based word embeddings, rather than in specific lexical semantic tasks.

WordNet Descriptions	
photography#3	*the occupation of taking and printing photographs or making movies*
transmutation#2	*a qualitative change*

Freebase Descriptions	
Stephen Harper	*Stephen Joseph Harper is a Canadian politician who is the 22nd and current Prime Minister of Canada and the Leader of the Conservative Party...*
El Paso	*El Paso is the county seat of El Paso County, Texas, United States, and lies in far West Texas...*

Table 2: Sample entity descriptions from WordNet and Freebase. As Freebase descriptions are lengthy paragraphs, only the first sentence is shown.

Bordes et al. (2011) propose the Structured Embeddings (SE) model, which embeds entities into vectors and relations into matrices. The relation connection between two entities is modeled by the projection of their embeddings into a different vector space. Rothe and Schütze (2015) use Wordnet as a lexical resource to learn embeddings for synsets and lexemes. Perhaps most related to our work are previous relational models that initialize their embeddings via distributional semantics calculated from a larger corpus. Socher et al. (2013) propose the Neural Tensor Network (NTN), and Yang et al. (2015) the Bilinear model using this technique. Other approaches modify the objective function or change the structure of the model in order to integrate distributional and relational information (Xu et al., 2014; Fried and Duh, 2015; Toutanova and Chen, 2015). Faruqui et al. (2015) retrofit word vectors *after* they are trained according to distributional criteria. We propose a method that does not necessitate post-processing of the embeddings, and can be applied orthogonally to the previously mentioned improvements.

3 Architecture of the Approach

3.1 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 assumes 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[3].

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.

3.2 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 the WordNet and Freebase datasets, we use *entity descriptions* which are readily available.

Although there are many ways to incorporate this, we propose a simple method whereby the entity descriptions are used to *initialize* the entity representations of the model, 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 to reduce overfitting. We obtain these word vectors using distributed representations computed using word2vec (Mikolov et al., 2013a) and GloVe (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).

Additional decisions need to be made concerning which parts of the entity description to include. In particular, if an entity description or word definition is longer than several sentences, using the entire description could cause a 'dilution' of the desired embedding, as not all sentences will be equally pertinent. We solve this by only considering the first sentence of any entity description, which is often the most relevant one. This is necessary for Freebase, where the description length can be several paragraphs.

4 Experiments

4.1 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. For the TransE model with random initialization, 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. For the TransE model with strategic initialization, we used different embedding dimensions. 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 reduction with PCA was then applied to reduce this to $k = 30$ for WN, and $k = 55$ for FB15k, which were empirically found to be optimal. PCA was necessary in this case as pre-trained vectors from word2vec and GloVe are not available for all dimension values.

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

[3] Note that we use $h, l, t \in \mathbb{R}^k$ to denote both the entities and relations, in addition to the vector representations of the entities and relations

		WN					FB15k				
		k	Mean rank		Hits@10		*k*	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.	20	266	254	76.1%	89.2%	50	195	92	41.2%	55.2%
	TransE Freebase W2V init.	—	—	—	—	—	50	195	91	41.3%	55.4%
Our models	TransE W2V entity defs. (NS)	30	210	192	78.5%	92.1%	55	195	91	41.6%	55.7%
	TransE GloVe entity defs. (NS)	30	**63**	**51**	64.6%	73.2%	55	194	90	41.7%	55.8%
	TransE W2V entity defs.	30	191	179	77.8%	91.6%	55	195	91	41.6%	55.6%
	TransE GloVe entity defs.	30	71	59	75.3%	88.0%	55	**193**	90	41.8%	55.8%

Table 3: Comparison between random initialization and using the entity descriptions. 'NS' tag indicates stopword removal from the entity descriptions'TransE Freebase W2V init' model uses word2vec pretrained with the Freebase vocabulary, and thus was not tested on WN.

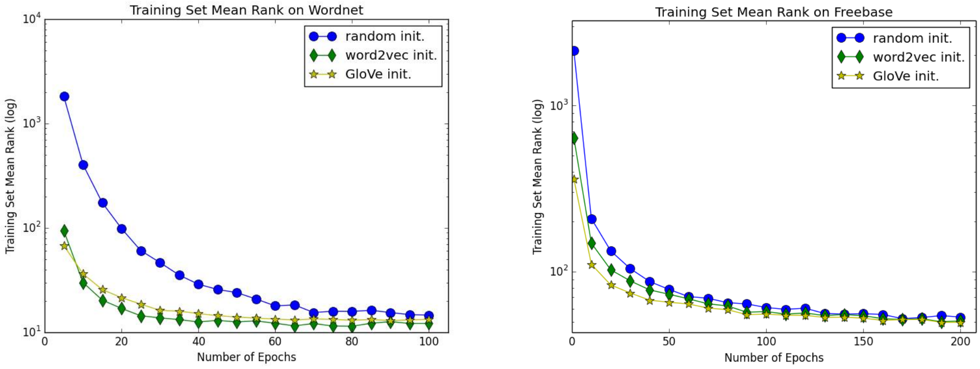

Figure 1: Learning curves for the mean ranks on the training set for WordNet (left) and Freebase (right).

by similarity according to the model. We evaluate using the original and most common metrics for relational models: *i)* the *mean* of the predicted ranks, and *ii) hits@10*, which represents the percentage of correct entities found in the top 10 list; however, other metrics are possible, such as mean reciprocal rank (MRR). We evaluate in both the filtered setting, where other correct responses are removed from the lists ranked by the model, and the raw setting, where no changes are made.

We compare against the TransE model with random initialization, and the SE model (Bordes et al., 2011). We also compare against the state-of-the-art TransD model (Ji et al., 2015). This model uses two vectors to represent each entity and relation; one to represent the meaning of the entity, and one to construct a mapping matrix dynamically. This allows for the representation of more diverse entities.

4.2 Results and Analysis

Table 3 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. Further, the fact that our optimal embedding dimensions are larger (30 and 55 vs. 20 and 50) suggests that our initialization approach helps avoid overfitting.

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.

WordNet Relations
_hyponym
_derivationally_related_form
_member_holonym

Freebase Relations
/award/award_nominee/award_nominations./award/ award_nomination/nominated_for
/broadcast/radio_station_owner/ radio_stations
/medicine/disease/notable_people_with _this_condition

Table 4: Sample relations from WordNet and Freebase. The relations from Freebase are clearly much more specific as they relate named entities.

Finally, we see in Figure 1 that the TransE model converges more quickly during training when initialized with our approach, compared to random initialization. This is particularly true on WordNet.

Mean rank and hits@10 discrepancy It is interesting to note 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. The exact nature of this tradeoff isn't clear, and is an interesting avenue for future work.

However, there are potential reasons for the results discrepancy betweeen mean rank and hits@10. We conjecture that our model helps avoid 'disasters' where some correct entities are ranked very low. For TransE with random initialization, these disasters cause a large decrease in mean rank, which is significantly improved by our model. On the other hand, reducing the number of correct entities that are poorly ranked may not significantly affect the hits@10, since this only considers entities near the top of the ranking.

Note also that using hits@10 to evaluate relational models is not ideal; a model can rank reasonable alternative entities highly, but be penalized because the target entity is not in the top 10. For example, given "rabbit IS-A", both "animal" and "mammal" fit as target entities. This is alleviated by filtering, but is not completely eliminated due to the sparsity of relations in the dataset (which is the reason we require the link prediction task). Thus, we believe the mean rank is a more accurate measure of the performance of a model, particularly on raw data.

Dataset differences It is also interesting to note the discrepancy between the results on the WordNet and Freebase datasets. Although using the entity descriptions leads to a significantly lower mean rank for the WordNet dataset, it only results in a faster convergence rate for Freebase. However, the relations presented in these two datasets are significantly different: WordNet relations are quite general and are meant to provide links between concepts, while the Freebase relations are very specific, and denote relationships between named entities. This is shown in Table 4. It seems that incorporating the definition of these named entities does not improve the ability of the algorithm to answer very specific relation questions. This would be the case if the optimization landscape for the TransE model had fewer local minima for Freebase than for WordNet, thus rendering it less sensitive to the initial condition. It is also possible that the TransE model is simply not powerful enough to achieve a filtered mean rank lower than 90, no matter the initialization strategy.

5 Conclusion and Future Work

We have shown that leveraging external lexical resources, along with distributional semantics, can lead to both a significantly improved optimum and a faster rate of convergence when applied with the TransE model for relational data. We established new state-of-the-art results on WordNet, and obtain small improvements to the state-of-the-art on raw relational data for Freebase. Our method is quite simple and could be applied in a straightforward manner to other models that take entity vector representations as input. Further research is needed to investigate whether performance on other NLP tasks can be improved by leveraging available lexical resources in a similar manner.

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 et al. (2014) on paragraph vectors. Alternatively, distributional semantics could be used as a regularizer, similar to (Labutov and 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.

References

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 SIGMOD*.

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

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

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

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

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

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

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of ACL on Interactive Poster and Demonstration Sessions*.

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

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

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC*.

Yuval Marton, Saif Mohammad, and Philip Resnik. 2009. Estimating semantic distance using soft semantic constraints in knowledge-source-corpus hybrid models. In *Proceedings of EMNLP*.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*.

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

Saif Mohammad. 2008. *Measuring semantic distance using distributional profiles of concepts*. Ph.D. thesis, University of Toronto.

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

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *Proceedings of ACL*.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of RANLP*.

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

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

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*.

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

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

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