Irreflexive and Hierarchical Relations as Translations

Antoine Bordes  
Nicolas Usunier  
Alberto García-Durán  
Heudiasyc UMR CNRS 7253, Université de Technologie de Compiègne, Compiègne, France  
Jason Weston  
Oksana Yakhenko  
Google, 111 8th avenue, New York, NY, USA

Abstract

We consider the problem of embedding entities and relations of knowledge bases in low-dimensional vector spaces. Unlike most existing approaches, which are primarily efficient for modeling equivalence relations, our approach is designed to explicitly model irreflexive relations, such as hierarchies, by interpreting them as translations operating on the low-dimensional embeddings of the entities. Preliminary experiments show that, despite its simplicity and a smaller number of parameters than previous approaches, our approach achieves state-of-the-art performance according to standard evaluation protocols on data from WordNet and Freebase.

1. Introduction

Multi-relational data, which refers to directed graphs whose nodes correspond to entities and edges represent relations that link these entities, plays a pivotal role in many areas such as recommender systems, the Semantic Web, or computational biology. Relations are modeled as triplets of the form (head, label, tail), where label indicates the type of link between the entities head and tail. Relations are thus of several types and can exhibit various properties (symmetry, transitivity, irreflexivity, etc.). Such graphs are popular tools for encoding data via knowledge bases (KBs), semantic networks or any kind of database following the Resource Description Framework format. Hence, they are widely used in the Semantic Web (e.g. Freebase\(^1\) or Google Knowledge Graph\(^2\)) but also for knowledge management in bioinformatics (e.g. GeneOntology\(^3\)) or natural language processing (e.g. WordNet\(^4\)).

Despite their appealing ability for representing complex data, multi-relational databases remain complicated to manipulate because of the heterogeneity of the relations (frequencies, connectivity), their inherent noise (collaborative or semi-automatic creation) and their very large dimension (up to millions of entities and thousands of relation types).

In this paper, we introduce a distributed model, which learns to embed such data in a vector space, where entities are modeled as low-dimensional embeddings. Many existing approaches (e.g. from Sutskever et al. (2009); Nickel et al. (2011)) interpret relations as linear transformations of these embeddings: when \((h, \ell, t)\) holds, then the embeddings of head \(h\) and tail \(t\) should be close (in the embedding space) after transformation by a linear operator that depends on the label \(\ell\). With such an interpretation, the model implies that the relation is reflexive since the embedding of \(h\) will always be its nearest neighbor, and because of the triangle inequality, the model will, to some extent, imply some form of transitivity of the relation.

While this interpretation is fine for equivalence relations (such as WordNet’s \_similar\_to\), it is inadequate for irreflexive relations that represent hierarchies, such as WordNet’s \_hypernym\) or Freebase’s type hierarchy. Indeed, taking the simplest example of entities organized in a tree with two relations, “sibling” and “parent”, the embeddings of siblings should be close to each other (since it essentially is an equivalence relation), but enforcing the constraint that parent nodes should be close to their child nodes will lead the embedding of the whole tree to collapse to a small

\[^{1}\text{freebase.com}\]  \[^{2}\text{google.com/insidesearch/features/search/knowledge.html}\]  \[^{3}\text{geneontology.org}\]  \[^{4}\text{wordnet.princeton.edu}\]
region of the space where the siblings and parent of a
given node are impossible to distinguish.

Since hierarchical and irreflexive relations are ex-
tremely common in KBs, we propose a simple model to
efficiently represent them, by interpreting relations as
translations in the embedding space: if \((h, \ell, t)\) holds,
then the embedding of \(t\) should be close to the embed-
ding of \(h\) plus some vector that depends on \(\ell\). This
approach is motivated by the natural representation
of trees (i.e. embeddings of the nodes in dimension 2):
while siblings are close to each other and nodes at a
given height are organized on the \(x\)-axis, the parent-
child relation corresponds to a translation on the \(y\)-
axis. Another, secondary, motivation comes from the
recent work of Mikolov et al. (2013), in which the au-
thors learn word embeddings from free text, and some
one-to-one relations between entities of different types,
such as the relation “capital of” between countries and
cities, are (coincidentally rather than willingly) repre-
sented by the model as translations in the embedding
space. Our approach may then be used in the context
of learning word embeddings in the future to reinforce
this kind a structure of the embedding space.

Apart from the main line of algorithms to learn em-
beddings of KBs, a number of recent approaches deal
with the asymmetry of the relations at the expense
of an explosion of model parameters. We present an
empirical evaluation on data dumps of WordNet and
Freebase, in which our model achieves strong results
compared to such algorithms, with much fewer param-
eters and even lower dimensional embeddings.

In the remainder of the paper, we describe some of the
related work in Section 2. We then describe our model
in Section 3, and discuss its connections with related
methods. We report preliminary experimental results
on WordNet and Freebase in Section 4. We finally
sketch some future work directions in Section 5.

2. Related work

Most previous methods designed to model relations
in multi-relational data rely on latent representations
or embeddings. The simplest form of latent at-
ttribute that can be associated to an entity is a la-
tent class. Several clustering approaches have been
proposed. Kemp et al. (2006) considered a non-
parametric Bayesian extension of the stochastic block-
model allowing to automatically infer the number of
latent clusters; Kok & Domingos (2007) introduced
clustering in Markov-Logic networks; Sutskever et al.
(2009) used a non-parametric Bayesian clustering of
entities embedding in a collective matrix factorization
formulation. All these models cluster not only entities
but relation labels as well.

These methods can provide interpretations and anal-
ysis of the data but are slow and do not scale to
large databases, due to the high cost of inference. In
terms of scalability, models based on tensor factoriza-
tion (like those from (Harshman & Lundy, 1994) or
(Nickel et al., 2011)) have shown to be efficient. How-
ever, they have been outperformed by energy-based
models (Bordes et al., 2011; Jenatton et al., 2012; Bor-
des et al., 2013; Chen et al., 2013). These methods
represent entities as low-dimensional embeddings and
relations as linear or bilinear operators on them and
are trained via an online process, which allows them
to scale well to large numbers of entities and relation
types. In Section 4, we compare our new approach to
SE (Bordes et al., 2011) and SME (Bordes et al., 2013).

3. Translation-based model

We now describe our model and discuss its relationship
to existing approaches.

3.1. Our model

Given a training set \(S\) of labeled arcs \((h, \ell, t)\), our goal
is to learn vector embeddings for all values of \(h, \ell\) and
t. We assume all nodes and labels appear at least
once in the training set. The embeddings take values
in \(\mathbb{R}^k\) (\(k\) is a model hyperparameter) and are denoted
with the same letter, in boldface characters. The basic
idea behind our model is that the functional relation
induced by the \(\ell\)-labeled arcs corresponds to a trans-
lation of the embeddings, i.e. we want that \(h + \ell \approx t\)
when \((h, \ell, t)\) holds, while \(h + \ell\) should be far away
from \(t\) otherwise.

To learn such embeddings, we minimize the following
margin-based ranking criterion over the training set:

\[
\sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'} \left[ \gamma + d(h + \ell, t) - d(h' + \ell, t') \right]_+ \tag{1}
\]

where \([x]_+\) denotes the positive part of \(x\), \(\gamma > 0\) is
a margin hyperparameter, \(d(x, y)\) is some dissimilari-
ty function on \(\mathbb{R}^k\), e.g. the euclidian distance or the
squared euclidian distance, and

\[
S'_{(h,\ell,t)} = \{(t',\ell,t)|h' \in N\} \cup \{(h,\ell,t')|t' \in N\}. \tag{2}
\]

The set of “negative” examples we sample according to
Equation 2 is basically the training (“positive”) triple
with either the head or tail replaced by a random entity
(but not both at the same time). The loss function (1)
favors low values of dissimilarity between head+label

Irreflexive and Hierarchical Relations as Translations
and tail for positive triplets, and large values for negative triplets, and is thus a natural implementation of the intended criterion.

The minimization is carried out by stochastic gradient descent, over the possible $h, \ell$ and $t$, with the additional constraints that the $L_2$-norm of the embeddings of the entities is 1 (no regularization or norm constraints are given to the label embeddings $\ell$).

3.2. Relationship to previous approaches

Section 2 described a large body of work on embedding KBs. We detail here the relationships between our model and those of Bordes et al. (2011) (Structured Embeddings or SE) and Chen et al. (2013).

**SE** (Bordes et al., 2011) embeds nodes into $\mathbb{R}^k$, and labels into two matrices $L_1 \in \mathbb{R}^{k \times k}$ and $L_2 \in \mathbb{R}^{k \times k}$ such that $d(L_1 h, L_2 t)$ is small for positive triplets $(h, \ell, t)$ (and large otherwise). The basic idea is that when two nodes belong to the same edge, their embeddings should be close to each other in some subspace that depends on the label. This basic idea would imply $L_1 = L_2$, and using two different projection matrices for the head and for the tail is intended to account for the possible asymmetry of relation $\ell$. When the dissimilarity function takes the form of $d(x, y) = g(x - y)$ for some $g : \mathbb{R}^k \rightarrow \mathbb{R}$ (e.g., $g$ is a norm), then the model of SE with an embedding of size $k + 1$ is strictly more expressive than our model with an embedding of size $k$, since linear operators in dimension $k + 1$ can reproduce affine transformations in a subspace of dimension $k$ (by constraining the $k + 1$st dimension of all node embeddings to be equal to 1). SE, with $L_2$ as the identity matrix and $L_1$ taken so as to reproduce a translation is then equivalent to our model. Despite the lower expressiveness of our model, we still reach better performance than this model in our experiments because (1) our model is a more direct way to represent the true properties of the relations, and (2) regularization, and more generally any form of capacity control, is difficult in embedding models: greater expressiveness may then be more synonymous to overfitting than to better performance.

Another related model is the Neural Tensor Model of Chen et al. (2013). A special case of that model (which would actually boil down to a “Neural Matrix Model”) corresponds to learn scores $s(h, \ell, t)$ (higher scores for positive triplets) of the form:

$$s(h, \ell, t) = h^T L t + \ell_1^T h + \ell_2^T t$$  \hspace{1cm} (3)

where $L \in \mathbb{R}^{k \times k}$, $L_1 \in \mathbb{R}^k$ and $L_2 \in \mathbb{R}^k$, all of them depending on $\ell$.

If we consider our model with the squared distance as dissimilarity function, we have:

$$d(h+\ell, t) = \|h\|^2 + \|\ell\|^2 + \|t\|^2 - 2(h^T t + \ell_1^T (t-h)).$$

Considering our norm constraints ($\|h\|^2 = \|t\|^2 = 1$) and the ranking criterion (1), in which $\|\ell\|^2$ does not play any role in comparing positive and negatives triplets, our model thus corresponds to the scoring triplets according to $h^T t + \ell_1^T (t-h)$, and thus corresponds to Chen et al. (2013)’s model (Equation (3)) where $L$ is the identity matrix, and $\ell = \ell_1 = -\ell_2$.

We could not run experiments with that model, but once again our model has much fewer parameters: this should ease the training and prevent overfitting, and hence compensate for a lower expressiveness.

4. Experiments

Our approach is evaluated against the methods SE and SME (Semantic Matching Energy) from (Bordes et al., 2011; 2013) on two data sets and using the same ranking setting for evaluation.

We measure the mean and median predicted ranks and the top-10, computed with the following procedure. For each test triplet, the head is removed and replaced by each of the entities of the dictionary in turn. Energies (i.e. dissimilarities) of those corrupted triplets are computed by the model and sorted by ascending order and the rank of the correct entity is stored. This whole procedure is also repeated when removing the tail instead or the head. We report the mean and median of those predicted ranks and the top-10, which is the proportion of correct entities in the top 10 ranks.

4.1. Data

We used data from two KBs; their statistics are given in Table 1.

**Table 1.** Statistics of the data sets used in this paper.

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>WORDNET</th>
<th>FREEBASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTITIES</td>
<td>40,943</td>
<td>14,951</td>
</tr>
<tr>
<td>REL. TYPES</td>
<td>18</td>
<td>1,345</td>
</tr>
<tr>
<td>TRAIN. EX.</td>
<td>141,442</td>
<td>483,142</td>
</tr>
<tr>
<td>VALID. EX.</td>
<td>5,000</td>
<td>50,000</td>
</tr>
<tr>
<td>TEST. EX.</td>
<td>5,000</td>
<td>59,071</td>
</tr>
</tbody>
</table>

**WordNet** This knowledge base is designed to produce an intuitively usable dictionary and thesaurus, and support automatic text analysis. Its entities (termed synsets) correspond to word senses,
Table 2. Some example predictions on the Freebase test set using our approach. Bold indicates the test triple’s true tail and italics other true tails present in the training set. Actual Freebase identifiers have been replaced by readable strings.

<table>
<thead>
<tr>
<th>Input (Head and Label)</th>
<th>Predicted Tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthony LaPaglia performed in</td>
<td>Lantana, Summer of Sam, Happy Feet, The House of Mirth, Unfaithful, Legend of the Guardians, Naked Lunch, X-Men, The Namesake</td>
</tr>
<tr>
<td>Camden County adjoins</td>
<td>Burlington County, Atlantic County, Gloucester County, Union County, Essex County, New Jersey, Passaic County, Ocean County, Bucks County</td>
</tr>
<tr>
<td>The 40-Year-Old Virgin nominated for</td>
<td>MTV Movie Award for Best Comedic Performance, BFCA Critics’ Choice Award for Best Comedy, MTV Movie Award for Best On-Screen Duo, MTV Movie Award for Best Breakthrough Performance</td>
</tr>
<tr>
<td>David Foster has the genre</td>
<td>Pop music, Pop rock, Adult contemporary music, Dance music, Contemporary R&amp;B, Soft rock, Rhythm and blues, Easy Listening</td>
</tr>
<tr>
<td>Costa Rica football team has position</td>
<td>Forward, Defender, Midfielder, Goalkeepers, Pitchers, Infielder, Outfielder, Center, Defensorman</td>
</tr>
<tr>
<td>Lil Wayne born in</td>
<td>New Orleans, Atlanta, Austin, St. Louis, Toronto, New York City, Wellington, Dallas, Puerto Rico</td>
</tr>
<tr>
<td>WALL-E has the genre</td>
<td>Adventure film, Science Fiction, Fantasy, Stop motion, Satire, Drama</td>
</tr>
<tr>
<td>Richard Crenna has cause of death</td>
<td>Pancreatic cancer, Cardiovascular disease, Meningitis, Cancer, Prostate cancers, Stroke, Liver tumour, Brain tumor, Multiple myeloma</td>
</tr>
</tbody>
</table>

Freebase

Freebase is a huge and growing database of general facts; there are currently around 1.2 billion triplets. To make a small data set to experiment on we selected the subset of entities that are also present in the Wikilinks database and that also have at least 100 mentions in Freebase (for both entities and relations). We also removed negative relations like '/people/person/nationality' which just reverses the head and tail compared to the relation '/people/person/nationality'. This resulted in 592,213 triplets with 14,951 entities and 1,345 relations which were randomized and split as shown in Table 1.

4.2. Implementation

We implemented our model using the SME library, which already proposes code for SE and SME. The dissimilarity measure \( d \) was set to the \( L_1 \) distance, mostly because it led to a faster training.

Table 3. Link prediction results on WordNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
<th>Mean</th>
<th>Med.</th>
<th>Top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td>317</td>
<td>26</td>
<td></td>
<td>35.1%</td>
</tr>
<tr>
<td>SE</td>
<td>1,011</td>
<td>3</td>
<td></td>
<td>68.5%</td>
</tr>
<tr>
<td>SME LINEAR</td>
<td>559</td>
<td>5</td>
<td></td>
<td>65.1%</td>
</tr>
<tr>
<td>SME BILINEAR</td>
<td>526</td>
<td>8</td>
<td></td>
<td>54.7%</td>
</tr>
<tr>
<td>Our Approach</td>
<td>263</td>
<td>4</td>
<td></td>
<td>75.4%</td>
</tr>
</tbody>
</table>

For this preliminary set of experiments, we did not perform an extensive search for hyperparameters. For experiments of our method on WordNet, we fixed the learning rate for the stochastic gradient descent to 0.01, the dimension \( k \) of the embeddings to 20 and chosen the margin \( \gamma \) among \{1, 2, 10\} with the validation set (optimal value was 2). We report results for SE and SME extracted from (Bordes et al., 2013) where those models have been trained using a much more thorough hyperparameter search. For experiments on Freebase, we ran all experiments using the SME library with fixed values for the learning rate \( (= 0.01) \), \( k \) \( (= 50) \) and \( \gamma \) \( (= 1) \). For both datasets, the training time was limited to at most 1,000 epochs over the training set. The best model was selected using the mean predicted rank on the validation set.

4.3. Results

Tables 3 and 4 displays the results on both data sets for our method, compared to SE, to two versions of

---

Footnotes:

5 WordNet is composed of senses, its entities are termed by the concatenation of a word, its part-of-speech tag and a digit indicating which sense it refers to i.e. \(_{score\_NN\_1}\) encodes the first meaning of the noun “score”.

6 googleresearch.blogspot.com/2013/03/learning-from-big-data-40-million.html

7 https://github.com/glorotxa/SME

---
SME and to Unstructured, a simple model which only uses the dot-product between \( h \) and \( t \) as dissimilarity measure for a triplet \((h, \ell, t)\), with no influence of \( \ell \). Table 2 gives examples of nearest link prediction results of our approach on the Freebase test set.

Our method greatly outperforms all counterparts on all metrics, with particularly good results for the top-10 metric. We believe that such remarkable performance is due to an appropriate design of the model according to the data, but also to its relative simplicity. Hence, even if the problem is non-convex, it can be optimized efficiently with stochastic gradient. We showed in Section 3.2 that SE is more expressive than our proposal. However, its complexity makes it quite hard to train as shown in the results of tables 3 and 4.

Table 2 illustrates the capabilities of our model. Given a head and a label, the top predicted tails (and the true one) are depicted. The examples come from the Freebase test set. Even if the good answer is not always top-ranked, the predictions reflect common-sense.

### 5. Conclusion and future work

We proposed a new approach to learn embeddings of KBs, focusing on the minimal parametrization of the model to accurately represent hierarchical and irreflexive relations. This short paper is essentially intended to be a proof-of-concept that translations are adequate to model such relations in a multi-relational setting. It can be improved and better validated in several ways. For the experimental evaluation, this paper is the first one to present link prediction on this dump of Freebase. More benchmarking is needed, such as the comparison with models of Chen et al. (2013) and Jenatton et al. (2012). We also intend to consider learning translations of word embedding, either from free text as in (Mikolov et al., 2013) or from \((subject, \text{verb}, object)\) triplets as in (Bordes et al., 2011).

Finally, regarding modeling relations, equivalence relations in our approach are represented by a 0 translation vector, and thus enforces all members of an equivalence class to be close to each other in the embedding space (whatever the relation). Some additional degrees of freedom may be given by adding a projection matrix to each relation, so that equivalence relations only enforce entities to be close to each other in some subspace of the embedding space. However, this would increase the number of parameters, and we believe that regularization and optimization techniques should be further studied to achieve optimal performance.

### Acknowledgments

We thank Thomas Strohmann and Kevin Murphy for useful discussions. This work was supported by the French ANR (EVEREST-12-JS02-005-01).

### References


