Distributed Prediction of Relations for Entities: The Easy, The Difficult, and The Impossible

Abhijeet Gupta* and **Gemma Boleda**[†] and **Sebastian Padó***
*Stuttgart University, Germany
{abhijeet.gupta,pado}@ims.uni-stuttgart.de
[†]Universitat Pompeu Fabra, Barcelona, Spain
gemma.boleda@upf.edu

Abstract

Word embeddings are supposed to provide easy access to *semantic relations* such as "male of" (*man–woman*). While this claim has been investigated for *concepts*, little is known about the distributional behavior of relations of *(Named) Entities*. We describe two word embedding-based models that predict values for relational attributes of entities, and analyse them. The task is challenging, with major performance differences between relations. Contrary to many NLP tasks, high difficulty for a relation does *not* result from low frequency, but from (a) one-to-many mappings; and (b) lack of context patterns expressing the relation that are easy to pick up by word embeddings.

1 Introduction

A central claim about distributed models of word meaning (e.g., Mikolov et al. (2013)) is that word embedding space provides easy access to *semantic relations*. E.g., Mikolov et al.'s space was shown to encode the "male-female relation" linearly, as a vector $(\overrightarrow{man} - \overrightarrow{woman} = \overrightarrow{king} - \overrightarrow{queen})$.

The accessibility of semantic relations was subsequently examined in more detail. Rei and Briscoe (2014) and Melamud et al. (2014) reported successful modeling of lexical relations such as hypernymy and synonymy. Köper et al. (2015) considered a broader range of relationships,with mixed results. Levy and Goldberg (2014b) developed an improved, nonlinear relation extraction method.

These studies were conducted primarily on *concepts* and their semantic relations, like `hypernym(politician) = person`. Meanwhile, *entities* and the relations they partake in are much less well understood.[1] Entities are instances of concepts, i.e., they refer to specific individual objects in the real world, for example, *Donald Trump* is an instance of the concept *politician*. Consequently, entities are generally associated with a rich set of numeric and relational attributes (for `politician` instances: `size, office`, etc.). In contrast to concepts, the values of these attributes tend to be *discrete* (Herbelot, 2015): while the `size` of `politician` is best described by a probability distribution, the `size` of `Donald Trump` is `1.88m`. Since distributional representations are notoriously bad at handling discrete knowledge (Fodor and Lepore, 1999; Smolensky, 1990), this raises the question of how well such models can capture entity-related knowledge.

In our previous work (Gupta et al., 2015), we analysed distributional prediction of *numeric* attributes of entities, found a large variance in quality among attributes, and identified factors determining prediction difficulty. A corresponding analysis for *relational (categorial) attributes* of entities is still missing, even though entities are highly relevant for NLP. This is evident from the highly active area of *knowledge base completion* (KBC), the task of extending incomplete entity information in knowledge bases such as Yago or Wikidata (e.g., Bordes et al., 2013; Freitas et al., 2014; Neelakantan and Chang, 2015; Guu et al., 2015; Krishnamurthy and Mitchell, 2015).

In this paper, we assess to what extent *relational attributes of entities* are easily accessible from word embedding space. To this end, we define two models that predict, given a target entity (`Star_Wars`) and a relation (`director`), a distributed representation for the relatum (`George_Lucas`). We carry out a detailed per-relation analyses of their performance on seven

[1] The original dataset by Mikolov et al. (2013) did contain a small number of entity-entity relations.

*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 104–109,
Vancouver, Canada, August 3-4, 2017. ©2017 Association for Computational Linguistics

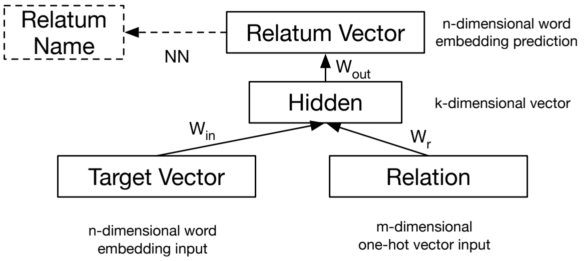
Relatum Name
NN
Relatum Vector
n-dimensional word embedding prediction
W_out
Hidden
k-dimensional vector
W_in
W_r
Target Vector
Relation
n-dimensional word embedding input
m-dimensional one-hot vector input

Figure 1: Nonlinear model (NonLinM) structure

major FreeBase domains and identify what makes a relation difficult by correlating performance with properties of the relations. We find that, contrary to many other NLP tasks, relations are *not* difficult if they are infrequent or sparse, but instead (a) if they relate one target to multiple relata; (b) if they do not give rise to linguistic patterns that can be picked up by bag-of-words distributional models.

2 Two Relatum Prediction Models

Both models predict a vector for a *relatum* r (plural: *relata*) given a *target entity* vector t and a symbolic *relation* ρ.

The Linear Model (LinM) is inspired by Mikolov et al.'s "phrase analogy" evaluation of word embeddings $(\overrightarrow{man} - \overrightarrow{woman} = \overrightarrow{king} - \overrightarrow{queen})$. However, instead of looking at individual words, we extract representations of semantic relations from sets of pairs $T_\rho = \{(t_i, \rho, r_i)\}$ instantiating the relation ρ. For each relation ρ, LinM computes the average (or centroid) difference vector over the set of training pairs:

$$\hat{r}(t, \rho) = t + \sum_{(r,\rho,t)\in T_\rho} (r - t)/N \qquad (1)$$

That is, the predicted $\hat{r}$ for an input (t, ρ) is the sum of the target vector and the relation's prototype. This model should work well if relations are represented additively in the embedding space.

The Nonlinear Model (NonLinM) is a feed-forward network (Figure 1) introducing a nonlinearity, inspired by Levy and Goldberg (2014b) and similar to models used in KBC, e.g., Socher et al. (2013). The relatum vector is predicted as

$$\hat{r}_\theta(t, \rho) = \sigma(\sigma(t \cdot W_{in} + v_\rho \cdot W_r) \cdot W_{out}) \qquad (2)$$

where v_ρ is the relation encoded as an m-dimensional one-hot vector and the three matrices W_{in}, W_r, W_{out} form the model parameters θ. For the nonlinearity σ, we use $tanh$.

In this model, the hidden layer represents a nonlinearly transformed composition of target and relation from which the relatum can be predicted. NonLinM can theoretically make accurate predictions even if relations are not additive in embedding space. Also, its sharing of training data among relations should lead to more reliable learning for infrequent relations. As objective function, we use

$$L(\theta) = \sum_{(t,r)} [\cos(\hat{r}_\theta(t, \rho), r) - \alpha \cdot \cos(\hat{r}_\theta(t, \rho), nc(\hat{r}_\theta(t, \rho)))] \qquad (3)$$

where $nc(v)$ is the *nearest confounder* of v, i.e., the next neighbor of v that is not a relatum for the current target-relation pair. Thus, we minimize the cosine distance between the predicted vector and the gold vector for the relatum while maximizing the cosine distance of the prediction to the closest negative example. We introduce a weight $\alpha \in [0, 1]$ for the negative sampling term as a hyper-parameter optimized on the development set. During training, we apply gradient descent with the adaptive learning rate method AdaDelta (Zeiler, 2012).

3 Experiments

Data. We extract relation data from FreeBase. We follow our earlier work Gupta et al. (2015), but go beyond its limitation to two domains (`country`, `citytown`). We experiment with seven major FreeBase domains: `animal`, `book`, `citytown`, `country`, `employer`, `organization`, `people`. We limit the number of datapoints of very large relation types to 3000 with random sampling for efficiency reasons. We only remove relation types with fewer than 3 datapoints. This results in a quite challenging dataset that demonstrates the generalizability of our models and is roughly comparable, in variety and size, to the FB15K dataset (Bordes et al., 2013).

The distributed representations for all entities come from the 1000-dimensional "Google News" skip-gram model (Mikolov et al., 2013) for FreeBase entities[2] trained on a 100G token news corpus. We only retain relation datapoints where both target and relatum are covered in the Google News vectors. Table 1 shows the numbers of relations and unique objects (target plus relata).

[2] `https://code.google.com/p/word2vec/`

	Size		Performance			Relation-level statistics			
Domain	#R	#Ts+Ra.	BL	LM	NLM	%R>0.3	%R<0.1	ρ(NLM, #In)	ρ(NLM, #RpT)
`animal`	24	3,428	0.11	0.16	**0.29**	38%	42%	.07	-.34
`book`	22	7,014	0.11	0.24	**0.26**	9%	68%	.09	-.15
`citytown`	46	86,551	0.05	0.13	**0.26**	28%	39%	-.12	-.22
`country`	89	191,196	0.04	0.08	**0.18**	20%	52%	-.32	-.23
`employer`	76	14,658	0.05	0.15	**0.23**	30%	45%	.01	-.35
`organization`	53	8,989	0.07	0.17	**0.26**	34%	42%	-.24	-.29
`people`	91	11,397	0.09	0.19	**0.27**	34%	23%	.23	-.25
Micro average			0.06	0.14	**0.22**	25%	45%	-.12	-.25
Macro average			0.08	0.16	**0.23**	28%	44%	-.05	-.26

Table 1: Test set statistics and results. #R: relations; #Ts+Ra: unique targets and relata; BL/LM/NLM: Baseline, linear and nonlinear model (macro-average MRR); %R$\lesseqgtr$x: percent of relations with MRR $\lesseqgtr$x; ρ: Spearman correlation; #In: instances; #RpT: relata per target

We split all domains into training, validation, and test sets (60%–20%–20%). The split applies to each relation type: in test, we face no unseen relation types, but unseen datapoints for each relation.[3]

Hyperparameter settings. The NonLinM model uses an L_2 norm constraint of s=3. We adopt the best AdaDelta parameters from Zeiler (2012), viz. $\rho = 0.95$ and $\epsilon = 10^{-6}$. We optimize the negative sampling weight α (cf. Eq. 3) by line search with a step size of 0.1 on the largest domain, `country`, and find 0.6 to be the optimal value, which we reuse for all domains. Due to the varying dimensionality m of the relation vector per domain, we set the size of the hidden layer to $k = 2n + m/10$ (n is the dimensionality of the word embeddings, cf. Figure 1). We train all models for a maximum of 1000 epochs with early stopping.

Evaluation. Models that predict vectors in a continuous vector space, like ours, cannot expect to predict the output vector precisely. Thus, we apply *nearest neighbor mapping* using the set of all unique targets and relata in each domain (cf. Table 1) to identify the correct relatum name. We then perform an Information Retrieval-style ranking evaluation: We compute the rank of the correct relatum r, given the target t and the relation ρ, in the test set T and aggregate these ranks to compute the *mean reciprocal rank* (MRR):

$$MRR = \frac{1}{||T||} \sum_{(t,\rho,r)\in T} \frac{1}{rank_{t,\rho}(r)} \qquad (4)$$

where $rank$ is the nearest neighbor rank of the relatum vector r given the prediction of the model for the input t, ρ. We report results at the relation level as well as macro- and micro-averaged MRR for the complete dataset.

Frequency Baseline (BL). Our baseline model ignores the target. For each relation, it predicts the frequency-ordered list of all training set relata.

4 Results and Discussion

Overall results. Table 1 shows that the nonlinear model NonLinM consistently gives the best results and statistically outperforms the linear model on all domains according to a Wilcoxon test (α=0.05). Both LinM and NonLinM clearly outclass the baseline. Most MRRs are around 0.25 (micro average 0.22), with one outlier, at 0.18, for `country`, the largest domain. Overall, the numbers may appear disappointing at first glance: these MRRs mean that the correct relatum is typically around the fourth nearest neighbor of the prediction vector. This indicates that open-vocabulary relatum prediction in a space of tens of thousands of words is a challenging task that warrants more detailed analysis. We observe that the nonlinear model achieves reasonable results even for sparse domains (cf. the low baseline), which we take as evidence for its generalization capabilities.

Analysis at relation level. Table 1 shows the number of relations with good MRRs (greater than 0.3) and bad MRRs (smaller than 0.1) for each relation. While the numbers vary across domains, the models tend to do badly on around 40-50% of all relations, and obtain good scores for less than one third of all relations.

Figure 2 shows the distribution for the best domain (`animal`) and the worst one (`country`) . Both plots show a Zipfian distribution with a rel-

[3] The dataset are available at: http://www.ims.uni-stuttgart.de/data/RelationPrediction.html

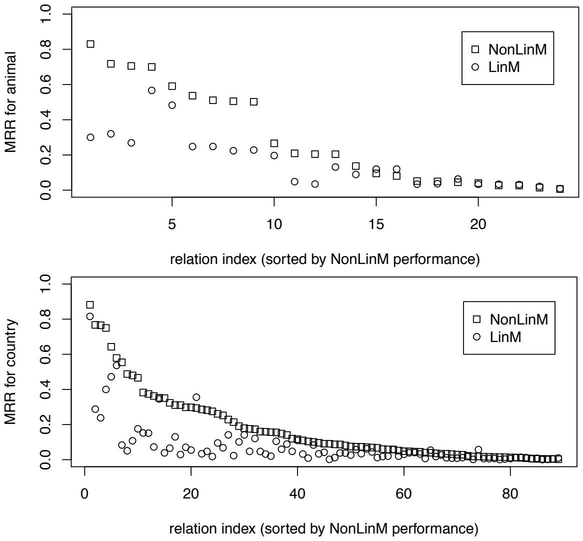

Figure 2: Results by relation for best and worst domains (`animal`, above; `people`, below), sorted by NonLinM performance

	Target	Correct	LinM	NonLinM
`continent`	Japan	Asia	Japan	**Asia**
	Kazakhstan	Asia	Central Asia	**Asia**
	Nicaragua	North America	Latin America	Americas
`capital`	Nepal	Kathmandu	Nepal	Dhaka
	Qatar	Doha	Qatar	Riyadh
	Venezuela	Caracas	**Caracas**	Quito

Table 2: Example predictions for two `country` relations (correct answer in boldface)

atively small set of well-modelled relations and a long tail of poorly modelled ones. NonLinM does better or as well as LinM for almost all relations. The performances of the two models are very tighly correlated for difficult relations; they only differ for the easier ones, where NonLinM's evidently captures the data better.

Qualitatively, the two models differ substantially with regard to prediction patterns at the level of targets. Table 2 shows the first predictions for three targets from two relations: `continent`, where NonLinM outperforms LinM, and `capital`, where it is the other way around. NonLinM's errors consist almost exclusively in predicting semantically similar entities of the correct relatum type, e.g., predicting Quito (the capital of Ecuador) as capital of Venezuela. In contrast, the LinM model has a harder time capturing the correct type, predicting country entities as capitals (e.g., Nepal as the capital of Nepal).

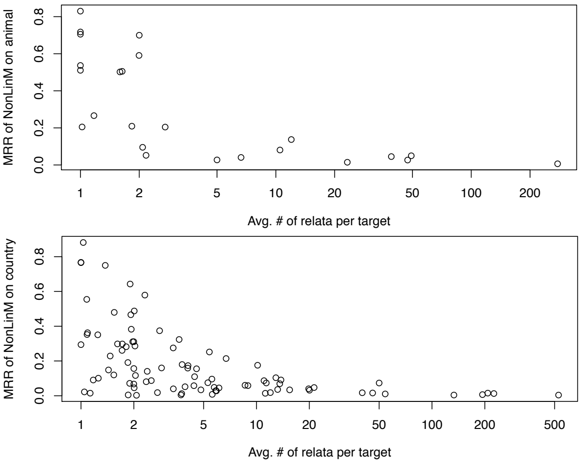

Figure 3: Scatterplot: MRR vs. number of relata per target (above: `animal`, below: `country`)

Analysis of Difficulty. So what makes many FreeBase relations hard to model? To test for sparsity problems, we first computed the correlation between model performance and the "usual suspect" relation frequency (number of instances for each relation). In NLP applications, this typically yields a high positive correlation. The second-to-last column of Table 1 shows that this is not true for our dataset. We find a substantial positive correlation only for `people`, correlations around zero for most domains, and substantial negative ones for `organization` and `country`. For these domains, therefore, frequent relations are actually *harder* to model. Further analysis revealed two main sources of difficulty:

(1) One-to-many relations. Relations with many datapoints tend to be *one-to-many*. We assume this to be a major source of difficulty, since the model is presented with multiple relata for the same target during training and will typically learn to predict a centroid of these relata. As an extreme case, consider a relation like `administrative_divisions` that relates the US to all of its federal states: the resulting prediction will arguably be dissimilar to every individual state. To test this hypothesis, we computed the rank correlation at the relation level between the number of relata per target and NonLinM performance, shown in the last column of Table 1. Indeed, we find a strong negative correlation for every single domain. In addition, Figure 3 plots relation performance (y axis) against the ratio of relata per target (x axis: one-to-one on the left, one-to-many on the right) for `animal` and `country`.

Relation	BL	LinM	NonLinM
tournaments	0.88	0.82	0.88
continent	0.29	0.29	0.77
country	0.25	0.24	0.77
⋮			
disputed_territories	0.00	0.01	0.01
horses_from_here	0.00	0.01	0.01
2nd_level_divisions	0.00	0.00	0.01

Table 3: The three most easy and most difficult relations for the `country` domain

Qualitatively, Table 3 shows examples for the three most easy and difficult relations for `country`. The list suggests that relations tend to be easy when they associate targets with single relata: the relation `country` maps territories and colonies onto their motherlands, and the `tournaments` relation is only populated with a few Commonwealth games (cf. the high baseline). In contrast, relations that map targets on many relata are difficult, such as administrative `divisions` of countries, or a list of `disputed_territories`. Note that this is not an evaluation issue, since MRR can deal with multiple correct answers. Our models do badly because they lack strategies to address these cases.

(2) Lack of contextual support. One-to-many relations are not the only culprit. Strikingly, Figure 2 shows that a low target-relatum ratio is a *necessary* condition for good performance (the upper right corners are empty), but not a *sufficient* one (the lower left corners are not empty). Some relations are not modelled well even though they are (almost) one-to-one. Examples include `currency_formerly_used` or `named_after` for `country` and `place_of_origin` for `animal`. Further analysis indicated that these relations suffer from what Gupta et al. (2015) called *lack of contextual support*: Although they are expressed overtly in the linguistic context of the target and relatum (and often even frequently so), their realizations cannot be tied to individual words or topics. Instead, they are expressed by relatively specific linguistic patterns, often predicate-argument structures (*X used to pay with Y, X is named in the honor of Y*). Such structures are hard to pick up by word embedding models that make the bag-of-words independence assumption among context words.

5 Conclusion

This paper considers the prediction of related entities ("relata") given a pair of a target Named Entity and a relation (`Star Wars, director, ?`) on the basis of distributional information. This task is challenging due to the more discrete behavior of attributes of entities as compared to concepts. We provide an analysis based on two models that use vector representations for both the targets and the relata.

Our results yield new insights into how embedding spaces represent entity relations: they are generally not represented additively, and nonlinearity helps. They also complement insights on the behavior of numeric attributes of entities (Gupta et al., 2015): Relations, like numeric attributes, are difficult to model if they are not specifically expressed in the lingusitic context of target and relatum. A new challenge specific to relations are situations where a single target maps onto many relata. If none of the two problems applies, relations are *easy* to model. If one applies, they are *difficult*. And if both apply, they are essentially *impossible*.

Among the two challenges, the problem of one-to-many relations appears easier to address, since a continuous output vector is, at least in principle, able to be similar to many relata. In the future, we will extend the model to deal better with one-to-many relations. While the lack of contextual support seems more fundamental, it could be addressed by either using syntax-based embeddings (Levy and Goldberg, 2014a) that can better pick up the specific context patterns characteristic for these relations, or by optimizing the input word embeddings for the task. This becomes a similar problem to joint training of representations from knowledge base structure and textual evidence (Perozzi et al., 2014; Toutanova et al., 2015).

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154) and the DFG (SFB 732, Project D10). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko.

2013. Translating embeddings for modeling multi-relational data. In *Proceedings of Neural Information Processing Systems 26*. pages 2787–2795.

Jerry Fodor and Ernie Lepore. 1999. All at Sea in Semantic Space: Churchland on Meaning Similarity. *Journal of Philosophy* 96(8):381–403.

André Freitas, Joao Carlos Pereira da Silva, Edward Curry, and Paul Buitelaar. 2014. A distributional semantics approach for selective reasoning on commonsense graph knowledge bases. In *Natural Language Processing and Information Systems*, Springer, pages 21–32.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 12–21.

Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 318–327.

Aurélie Herbelot. 2015. Mr Darcy and Mr Toad, gentlemen: Distributional names and their kinds. *Proceedings of the 11th International Conference on Computational Semantics* pages 151–161.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces. In *Proceedings of the 11th Conference on Computational Semantics*. London, UK, pages 40–45.

Jayant Krishnamurthy and Tom M Mitchell. 2015. Learning a compositional semantics for freebase with an open predicate vocabulary. *Transactions of the Association for Computational Linguistics* 3:257–270.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan, pages 171–180.

Oren Melamud, Ido Dagan, Jacob Goldberger, Idan Szpektor, and Deniz Yuret. 2014. Probabilistic modeling of joint-context in distributional similarity. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan, pages 181–190.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Lake Tahoe, NV, pages 3111–3119.

Arvind Neelakantan and Ming-Wei Chang. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. Denver, CO, pages 515–525.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York City, NY, pages 701–710.

Marek Rei and Ted Briscoe. 2014. Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan, pages 68–77.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46(1-2):159–216.

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*. Lake Tahoe, CA, pages 926–934.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of EMNLP*. Lisbon, Portugal, pages 1499–1509.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. In *CoRR, abs/1212.5701*.