# Generalizing Representations of Lexical Semantic Relations

**Anupama Chingacham**
SFB 1102, Saarland University
Saarbrucken ,66123, Germany
anu.vgopal2009@gmail.com

**Denis Paperno**
CNRS, LORIA, UMR 7503
Vandoeuvre-lès-Nancy, F-54500, France
denis.paperno@loria.fr

## Abstract

**English.** We propose a new method for unsupervised learning of embeddings for lexical relations in word pairs. The model is trained on predicting the contexts in which a word pair appears together in corpora, then generalized to account for new and unseen word pairs. This allows us to overcome the data sparsity issues inherent in existing relation embedding learning setups without the need to go back to the corpora to collect additional data for new pairs.

**Italiano.** *Proponiamo un nuovo metodo per l'apprendimento non supervisionato delle rappresentazioni delle relazioni lessicali fra coppie di parole (word pair embeddings). Il modello viene allenato a prevedere i contesti in cui compare uns coppia di parole, e successivamente viene generalizzato a coppie di parole nuove o non attestate. Questo ci consente di superare i problemi dovuti alla scarsità di dati tipica dei sistemi di apprendimento di rappresentazioni, senza la necessità di tornare ai corpora per raccogliere dati per nuove coppie di parole.*



## 1 Introduction

In this paper we address the problem of unsupervised learning of lexical relations between any two words. We take the approach of unsupervised representation learning from distribution in corpora, as familiar from word embedding methods, and enhance it with an additional technique to overcome data sparsity.

Word embedding models give a promise of learning word meaning from easily available text data in an unsupervised fashion and indeed the resulting vectors contain a lot of information about the semantic properties of words and objects they refer to, cf. for instance Herbelot and Vecchi (2015). Based on the distributional hypothesis coined by Z. S. Harris (1954), word embedding models, which construct word meaning representations as numeric vectors based on the co-occurrence statistics on the word's context, have been gaining ground due to their quality and simplicity. Produced by efficient and robust implementations such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), modern word vector models are able to predict whether two words are related in meaning, reaching human performance on benchmarks like WordSim353 (Agirre et al., 2009) and MEN (Bruni et al., 2014).

On the other hand, lexical knowledge includes not only properties of individual words but also relations between words. To some extent, lexical semantic relations can be recovered from the word representations via the vector offset method as evidenced by various applications including analogy solving, but already on this task it has multiple drawbacks (Linzen, 2016) and has a better unsupervised alternative (Levy and Goldberg, 2014).

Just like a word representation is inferred from the contexts in which the word occurs, information about the relation in a given word pair can be extracted from the statistics of contexts in which the two words of the pair appear together. In our model, we use this principle to learn high-quality pair embeddings from frequent noun pairs, and on their basis, build a way to construct a relation representation for an arbitrary pair.

Note that we approach the problem from the viewpoint of lerning general-purpose semantic knowledge. Our goal is to provide a vector representation for an arbitrary pair of words $w_1, w_2$. This is a more general task than *relation extraction*, which aims at identifying the semantic rela-

tion between the two words in a particular context. Modeling such general relational knowledge is crucial for natural language understanding in realistic settings. It may be especially useful for recovering the notoriously difficult bridging relations in discourse since they involve understanding implicit links between words in the text.

Representations of word relations have applications in many NLP tasks. For example, they could be extremely useful for resolving bridging, especially of the lexical type (Rösiger et al., 2018). But in order to be useful in practice, word relation models must generalize to rare or unseen cases.

## 2 Related Work

Our project is related to the task of relation extraction that has been in focus of various complex models (Mintz et al., 2009; Zelenko et al., 2003) including recurrent (Takase et al., 2016) and convolutional neural network architectures (Xu et al., 2015; Nguyen and Grishman, 2015; Zeng et al., 2014), although the simple averaging or summation of the context word vectors seems to produce good results for the task (Fan et al., 2015; Hashimoto et al., 2015). The latter work by Hashimoto et al. bears the greatest resemblance to the approach to learning semantic relation representations that we utilize here. Hashimoto et al. train noun embeddings on the task of predicting words occurring in between the two nouns in text corpora and use these embeddings along with averaging-based context embeddings as input to relation classification.

There are numerous studies dedicated to characterizing relations in word pairs abstracted away from the specific context in which the word pair appears. Much of this literature focuses on one specific lexical semantic relation at a time. Among these, lexical entailment (hypernymy) has probably been the most popular since Hearst (1992) with various representation learning approaches specifically targeting lexical entailment (Fu et al., 2014; Anh et al., 2016; Roller and Erk, 2016; Bowman, 2016; Kruszewski et al., 2015) and the antonymy relation has also received considerable attention (Ono et al., 2015; Pham et al., 2015; Shwartz et al., 2016; Santus et al., 2014). Another line of work in representing the compositionality of meaning of words using syntactic structures(like Adjective-Noun pairs) is another approach towards semantic relation representations. (Baroni and Zamparelli, 2010; Guevara, 2010).

The kind of relation representations we aim at learning are meant to encode general relational knowledge and are produced in an unsupervised way, even though it can be useful for identification of specific relations like hypernymy and for relation extraction from text occurrences (Jameel et al., 2018). The latter paper documents a model that produces word pair embeddings by concatenating Glove-based word vectors with relation embeddings trained to predict the contexts in which the two words of the pair co-occur. The main issue with Jameel et al.'s models is scalability: as the authors admit, it is prohibitively expensive to collect all the data needed to train all the relation embeddings. Instead, their implementation requires, for each individual word pair, going back to the training corpus via an inverse index and collecting the data needed to estimate the embedding of the pair. This strategy might not be efficient for practical applications.

## 3 Proposed Model

We propose a simple solution to the scalability problem inherent in word relation embedding learning from joint cooccurrence data, which also allows the model to generalize to word pairs that never occur together in the corpus, or occur too rarely to accumulate significant relational cues information. The model is trained in two steps.

First, we apply the skip-gram with negative sampling algorithm to learn relation vectors for pairs of nouns $n_1, n_2$ with high individual and joint occurrence frequencies. In our experiments, all word pairs with pair frequency more than 100 and its individual word frequency more than 500 are considered as *frequent pair*s. To estimate the **SkipRel** vector of the pair, we adapted the learning objective of skip-gram with negative sampling, maximizing

$$log\sigma(v_c'^T.u_{n_1:n_2}) + \Sigma_{i=1}^{k} \mathbb{E}_{c_i^* \sim P_n(c)}[log\sigma(-v_{c_i^*}'^T.u_{n_1:n_2})] \quad (1)$$

where $u_{n_1:n_2}$ is the SkipRel embedding of a word pair, $v_c'$ is the embedding of a context word occurring between $n_1$ and $n_2$, and $k$ is the number of negative samples.

High-quality SkipRel embeddings can only obtained for noun pairs that co-occur frequently. To allow the model to generalize to noun pairs that do not co-occur in our corpus, we estimated an inter-

polation $\tilde{u}_{n_1:n_2}$ of the word pair embedding

$$\tilde{u}_{n_1:n_2} = relU(Av_{n_1} + Bv_{n_2}) \quad (2)$$

where $v_{n_1}, v_{n_2}$ are pretrained word embeddings for the two nouns and the matrices $A, B$ encode systematic correspondences between the embeddings of a word and the relations it participates in. Matrices $A, B$ were estimated using stochastic gradient descent with the objective of minimizing the square error for the SkipRel vectors of frequent noun pairs $n_1, n_2$

$$\frac{1}{\mid P \mid}\Sigma_{n_1:n_2 \in P}(\tilde{u}_{n_1:n_2} - u_{n_1:n_2}) \quad (3)$$

We call $\tilde{u}_{n_1:n_2}$ the generalized SkipRel embedding (g-SkipRel) for the noun pair $n_1, n_2$. **RelWord**, the proposed relation embedding, is the concatenation of the **g-SkipRel** vector $\tilde{u}_{n_1:n_2}$ and the **Diff** vector $v_{n_1} - v_{n_2}$.

## 4 Experimental setup

We trained relation vectors on the ukWAC corpus (Baroni et al., 2009) containing 2 bln tokens of web-crawled English text. SkipRel is trained on noun pair instances separated by at most 10 context tokens with embedding size of 400 and mini-batch size of 32. Frequency filtering is performed to control the size of pair vocabulary ($|P|$). Frequent pairs are pre-selected using pair and word frequency thresholds. For pretrained word embeddings we used the best model from Baroni et al. (2014).

The experimental setup is built and maintained on GPU clusters provided by GRID5000 (Cappello et al., 2005). The code for model implementation and evaluation is publicly available at `https://github.com/Chingcham/SemRelationExtraction`

## 5 Evaluation

If our relation representations are rich enough in the information they encode, they will prove useful for any relation classification task regardless of the nature of the classes involved. We evaluate the model with a supervised softmax classifier on 2 labeled multiclass datasets, BLESS (Baroni and Lenci, 2011) and EVALuation1.0 (Santus et al., 2015), as well as the binary classification EACL antonym-synonym dataset (Nguyen et al., 2017). BLESS set consists of 26k triples of concept and relata spanned across 8 classes of semantic relation and EVALuation1.0 has 7.5k datasets spanned across 9 unique relation types. From EACL_2017 dataset, we used a list of 4062 noun pairs.

| Model | BLESS | EVAL | EACL |
|---|---|---|---|
| Diff | 81.15 | 57.83 | 71.25 |
| g-SkipRel | 59.07 | 48.06 | 70.31 |
| RelWord | 80.94 | **59.05** | **73.88** |
| Random | 12.5 | 11.11 | 50 |
| Majority | 24.71 | 25.67 | 50.4 |

Table 1: Semantic relation classification accuracy

Since we aim at recognizing whether the information relevant for relation identification is present in the representations in an easily accessible form, we choose to employ a simple, one-layer SoftMax classifier. The classifier was trained for 100 epochs, and the learning rate for the model is defined through crossvalidation. L2 regularization is employed to avoid over-fitting and the l2 factor is decided through empirical analysis. The classifier is trained with mini-batches of size 16 for BLESS & EVALuation1.0 and 8 for EACL_2017. SGD is utilized for optimizing model weights.

We prove the efficiency of RelWord vectors, we contrast them with the simpler representations of (g-)SkipRel and to Diff, the difference of the two word vectors in a pair, which is a commonly used simple method. We also include two simple baselines: random choice between the classes and the constant classifier that always predicts the majority class.

## 6 Results

All models outperform the baselines by a wide margin (Table 1). RelWord model compares favorably with the other options, outperforming them on EVAL and EACL datasets and being on par with the vector difference model for BLESS. This result signifies a success of our generalization strategy, because in each dataset only a minority of examples had pair representations directly trained from corpora; most WordRel vectors were interpolated from word emeddings.

Now let us restrict our attention to word pairs that frequently co-occur (Table 2). Note that the composition of classes, and by consequence the majority baseline, is different from Table 1, so the accuracy figures in the two tables are not di-

| Model | BLESS | EVAL | EACL |
|---|---|---|---|
| Diff | 77.13 | 44.61 | 66.07 |
| SkipRel | 73.37 | 48.40 | **83.03** |
| RelWord | **83.27** | **54.47** | 79.46 |
| Random | 12.5 | 11.11 | 50 |
| Majority | 33.22 | 26.37 | 63.63 |

Table 2: Semantic relation classification accuracy for frequent pairs

| pair | gold | Diff | RelWord |
|---|---|---|---|
| bottle, can | antonym | hasproperty | hasa |
| race, time | hasproperty | hasa | antonym |
| balloon, hollow | hasproperty | antonym | hasa |
| clear, settle | isa | antonym | synonym |
| develop, grow | isa | antonym | synonym |
| exercise, move | entails | antonym | isa |
| fact, true | hasproperty | antonym | synonym |
| human, male | isa | synonym | hasproperty |
| respect, see | isa | antonym | synonym |
| slice, hit | isa | antonym | synonym |

Table 3: Ten random examples in which RelWord and Diff make different errors. In the first one, the two models make predictions of comparable quality. In the second one, Diff makes a more intuitive error. In the remaining examples, RelWord's prediction is comparatively more adequate.

rectly comparable. For these frequent pairs we can rely on SkipRel relation vectors that have been estimated directly from corpora and have a higher quality; we also use SkipRel vectors instead of g-SkipRel as a component of RelWord. We note that for these pairs the performance of the Diff method dropped uniformly. This presumably happened in part because the classifier could no longer rely on the information on relative frequencies of the two words which is implicitly present in Diff representations; for example, it is possible that antonyms have more similar frequencies than synonyms in the EACL dataset. For BLESS and EVAL, the drop in the performance of Diff could have happened in part because the classes that include more frequent pairs such as isA, antonyms and synonyms are inherently harder to distinguish than classes that tend to contain rare pairs. In contrast, the comparative effectiveness of RelWord is more pronounced after frequency filtering. The usefulness of relation embeddings is especially impressive for the EACL dataset. In this case, vanilla SkipRel emerges as the best model, confirming that word embeddings *per se* are not particularly useful for detecting the synonymy-antonymy distinction for this subset of EACL, getting an accuracy just above the majority baseline, while pair embeddings go a long way.

Finally, quantitative evaluation in terms of classification accuracy or other measures does not fully characterize the relative performance of the models; among other things, certain types of misclassification might be worse than others. For example, a human annotator would rarely confuse synonyms with antonyms, while mistaking `has_a` for `has_property` could be a common point of disagreement between annotators. To do a qualitative analysis of errors made by different models, we selected the elements of EVAL test partition where Diff and RelWord make distinct predictions that are both different from the gold standard label. We manually annotated for each of the 53 examples of this kind, which model is more a acceptable according to a human's judgment. In a majority of cases (28) the WordRel model makes a prediction that is more human-like than that of Diff. For example, WordRel predicts that *shade* is part of *shadow* rather than its synonym (gold label); indeed, any part of a shadow can be called *shade*. The Diff model in this case and in many other examples bets on the antonym class, which does not make any sense semantically; the reason why `antonym` is a common false label is probably that it is simply the second biggest class in the dataset. The examples where Diff makes a more meaningful error than RelWord are less numerous (6 out of 53). There are also 15 examples where both system's predictions are equally bad (for example, for *Nice,France* Diff predict `isa` label and WordRel predicts `synonym`) and 4 examples where the two predictions are equally reasonable. For more examples, see Table 3. We note that sometimes our model's prediction seems more correct than the gold standard, for example in assigning `hasproperty` rather than `isa` label to the pair *human, male*.

## 7 Conclusion

The proposed model is simple in design and training, learning word relation vectors based on co-occurrence with unigram contexts and extending to rare or unseen words via a non-linear mapping. Despite its simplicity, the model is capable of capturing lexical relation patterns in vector representations. Most importantly, RelWord extends straightforwardly to novel word pairs in a

manner that does not require recomputing cooccurrence counts from the corpus as in related approaches (Jameel et al., 2018). This allows for an easy integration of the pretrained model into various downstream applications.

In our evaluation, we observed that learning word pair relation embeddings improves on the semantic information already present in word embeddings. With respect to certain semantic relations like synonyms, the performance of relation embedding is comparable to that of word embeddings but with an additional cost of training a representation for a significant number of pair of words. For other relation types like antonyms or hypernyms, in which words differ semantically but share similar contexts, learned word pair relation embeddings have an edge over those derived from word embeddings via simple subtraction. While in practice one has to make a choice based on the task requirements, it is generally beneficial to combine both types of relation embeddings for best results in a model like RelWord.

Our current model employs pretrained word embeddings and learns the word pair embeddings and a word-to-relation embedding mapping separately. In the future, we plan to train a version of the model end-to-end, with word embeddings and the mapping trained simultaneously. As literature suggests (Hashimoto et al., 2015; Takase et al., 2016), such joint training might not only benefit the model but also improve the performance of the resulting word embeddings on other tasks.

## Acknowledgments

This research is supported by CNRS PEPS grant ReSeRVe. We thank Roberto Zamparelli, Germán Kruszewski, Luca Ducceschi and anonymous reviewers who gave feedback on previous versions of this work.

## References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Tuan Luu Anh, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germn Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, volume 1, pages 238–247, 06.

Samuel Ryan Bowman. 2016. *Modeling natural language semantics in learned representations*. Ph.D. thesis, Ph. D. thesis, Stanford University.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Franck Cappello, Eddy Caron, Michel J. Dayd, Frdric Desprez, Yvon Jgou, Pascale Vicat-Blanc Primet, Emmanuel Jeannot, Stphane Lanteri, Julien Leduc, Nouredine Melab, Guillaume Mornet, Raymond Namyst, Benjamin Qutier, and Olivier Richard. 2005. Grid'5000: a large scale and highly reconfigurable grid experimental testbed. In *GRID*, pages 99–106. IEEE Computer Society.

Miao Fan, Kai Cao, Yifan He, and Ralph Grishman. 2015. Jointly embedding relations and mentions for knowledge population. *arXiv preprint arXiv:1504.01683*.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1199–1209.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '10, pages 33–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2015. Task-oriented learning of word embeddings for semantic relation classification. *arXiv preprint arXiv:1503.00095*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. Technical Report S2K-92-09.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. 2018. Unsupervised learning of distributional relation vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23–33. Association for Computational Linguistics.

German Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving boolean structures from distributional vectors. *Transactions of the Association for Computational Linguistics*, 3:375–388.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In Phil Blunsom, Shay B. Cohen, Paramveer S. Dhillon, and Percy Liang, editors, *VS@HLT-NAACL*, pages 39–48. The Association for Computational Linguistics.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 76–85, Valencia, Spain.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *HLT-NAACL*, pages 984–989.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nghia The Pham, Angeliki Lazaridou, Marco Baroni, et al. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 21–26.

Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. *CoRR*, abs/1605.05433.

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528.

Enrico Santus, Qin Lu, Alessandro Lenci, and Churen Huang. 2014. Unsupervised antonym-synonym discrimination in vector space.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2016. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. *arXiv preprint arXiv:1612.04460*.

Sho Takase, Naoaki Okazaki, and Kentaro Inui. 2016. Modeling semantic compositionality of relational patterns. *Engineering Applications of Artificial Intelligence*, 50:256–264.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. *CoRR*, abs/1506.07650.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.