Injecting Lexical Contrast into Word Vectors by Guiding Vector Space Specialisation

Ivan Vulić and **Anna Korhonen**
Language Technology Lab, University of Cambridge, UK
{iv250,alk23}@cam.ac.uk

Abstract

Word vector space specialisation models offer a portable, light-weight approach to fine-tuning arbitrary distributional vector spaces to discern between synonymy and antonymy. Their effectiveness is drawn from external linguistic constraints that specify the exact lexical relation between words. In this work, we show that a careful selection of the external constraints can steer and improve the specialisation. By simply selecting appropriate constraints, we report state-of-the-art results on a suite of tasks with well-defined benchmarks where modeling lexical contrast is crucial: 1) true semantic similarity, with highest reported scores on SimLex-999 and SimVerb-3500 to date; 2) detecting antonyms; and 3) distinguishing antonyms from synonyms.

1 Introduction

Representation models grounded in the distributional hypothesis (Harris, 1954) generally fail to distinguish highly contrasting words (*antonyms*) from highly similar ones (*synonyms*), due to similar word co-occurrence signatures in text corpora (Turney and Pantel, 2010; Mohammad et al., 2013).[1] In addition to antonymy and synonymy being fundamental lexical relations that are central to the organisation of the mental lexicon (Miller and Fellbaum, 1991; Murphy, 2010), this undesirable property of distributional word vector spaces has grave implications on their application in NLP reasoning and understanding tasks. As shown in prior work (Pham et al., 2015; Mrkšić et al., 2016; Kim et al., 2016; Nguyen et al., 2017b; Mrkšić et al., 2017, *i.a.*), explicitly modeling the *lexical contrast* benefits text entailment, dialogue state tracking, spoken language understanding, language generation, etc.[2]

A popular solution to address the limitation concerning lexical contrast is to move beyond standalone unsupervised learning. Post-processing procedures have been designed that leverage external lexical knowledge available in human- and automatically-constructed lexical resources (e.g., PPDB, WordNet): these methods *fine-tune* input word vectors to satisfy *linguistic constraints* from the external resources (Faruqui et al., 2015; Jauhar et al., 2015; Rothe and Schütze, 2015; Wieting et al., 2015; Mrkšić et al., 2016; Mrkšić et al., 2017; Vulić et al., 2017b, *i.a.*). This process has been termed *retrofitting* or *vector space specialisation*.

As one advantage, the post-processing methods are applicable to arbitrary input vector spaces. They are also "light-weight", that is, they do not require large corpora for (re-)training, as opposed to *joint specialisation models* (Yu and Dredze, 2014; Kiela et al., 2015; Pham et al., 2015; Nguyen et al., 2016) which integrate lexical knowledge directly into distributional training objectives.[3]

The main driving force of the retrofitting models are the external constraints, which specify which words should be close to each other in the specialised vector space (i.e., the so-called ATTRACT constraints), and which words should be far apart in the space (REPEL). By manipulating the constraints, one can steer the specialisation goal: e.g., Vulić et al. (2017a) use verb relations from VerbNet (Kipper, 2005) to accentuate VerbNet-style syntactic-semantic relations in the vector space.

[1] As pointed out by Cruse (1986), antonyms have a paradoxical nature: on the one hand, they constitute the two opposites of a meaning continuum, and therefore could be seen as semantically remote; on the other hand, they are paradigmatically similar, having almost identical distributions.

[2] Using a simple example, users asking for *a cheap pub in northern Seattle* do not want a virtual personal assistant to recommend *an expensive restaurant in southern Portland*.

[3] An additional advantage of post-processors is their better overall performance across a range of tasks when compared to the "heavy-weight" joint models (Mrkšić et al., 2016).

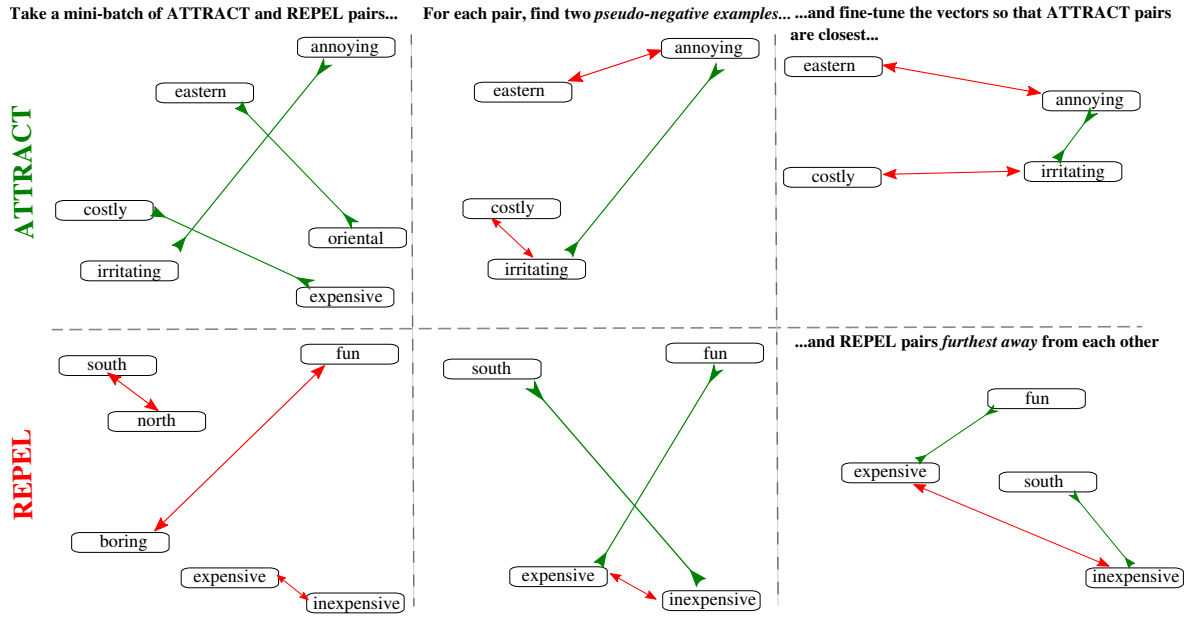

Figure 1: An illustration of specialisation for lexical contrast with toy examples. The specialisation model operates with two sets of external linguistic constraints: 1) ATTRACT word pairs, which have to be as close as possible in the fine-tuned vector space (e.g., *irritating* and *annoying*); and 2) REPEL word pairs, which have to be as far away from each other as possible (e.g., *expensive* and *inexpensive*).

Contributions. In this work, we investigate how different constraints affect specialisation. We show that a careful selection of external constraints can guide specialisation models to emphasise lexical contrast in the fine-tuned vector space: e.g., we indicate that direct (i.e., 1-step) WordNet hypernymy-hyponymy pairs are useful for boosting lexical contrast. Our specialised word vector spaces yield state-of-the-art results on a range of tasks where modeling lexical contrast is crucial: **1)** true semantic similarity; **2)** antonymy detection; and **3)** distinguishing antonyms from synonyms. Our SimLex-999 (Hill et al., 2015) and SimVerb-3500 (Gerz et al., 2016) scores are the highest reported results on these datasets to date: the result on SimLex-999 is the first result on the dataset surpassing the ceiling of mean inter-annotator agreement.

2 Methodology

Specialisation Model. Post-processing models are generally guided by two broad sets of constraints: **1)** ATTRACT constraints (AC) specify which words should be close to each other in the fine-tuned vector space; **2)** REPEL (RC) constraints describe which words should be pulled away from each other. The nomenclature is adopted from Mrkšić et al. (2017). Earlier post-processors (Faruqui et al., 2015; Jauhar et al., 2015; Wieting et al., 2015) operate only with ATTRACT constraints, and are therefore not suited to model both aspects of lexical contrast. In this work, we employ the state-of-the-art specialisation model of Mrkšić et al. (2017) which integrates both sets of constraints into its fine-tuning process. Here, we provide only a high-level description of the model, also illustrated by Figure 1, while we refer the interested reader to the original paper for a full (technical) description.

`syn` (AC)	`hypl` (AC)	`antexp` (RC)
(outburst, outbreak)	(discordance, dissonance)	(smooth, shake)
(safe, secure)	(postmen, deliverymen)	(clear, obscurity)
(cordial, warmhearted)	(employee, worker)	(relief, pressure)
(answer, response)	(swap, exchange)	(half, full)

Table 1: Examples of linguistic constraints.

In short, the model trains over batches of ATTRACT and REPEL pairs and contains three terms in its objective function. First, the ATTRACT term pushes two words from each ATTRACT constraint closer to each other (in terms of the cosine similarity) than to any other word present in the current batch by a margin δ_{att}. Second, the REPEL term pulls away two words from each REPEL constraint so that they are further away from each other than from any other word present in the current batch (again, by a margin δ_{rpl}): see Figure 1 again. Third, a regularisation term is used to preserve the useful semantic content originally present in the distribu-

tional space, as long as this information does not contradict the injected external knowledge.

Linguistic Constraints. The constraints are in fact word pairs (x_i, x_j), $x_i, x_j \in V$, where V is the vocabulary represented in the input distributional space. First, the conflation of synonymy and antonymy relations in the input space can be obviously mitigated by assigning synonymy pairs (`syn`) to the ATTRACT set, and antonymy pairs (`ant`) to the REPEL set. Further, similar to Ono et al. (2015), it is possible to extend the (typically less exhaustive) list of antonyms by combining the available knowledge from `syn` and `ant` word pairs. If (x_i, x_j) are a pair of synonyms, and (x_i, x_k) are a pair of antonyms, one can add another pair (x_j, x_k) to the expanded list of antonyms: this yields a larger set (`antexp`) to serve as REPEL constraints.

Finally, as the analysis of Hill et al. (2015) shows, the taxonomic hypernymy-hyponymy IS-A relation is often mistaken by true synonymy by humans. Therefore, we also experiment with direct (i.e. 1-step) IS-A pairs (`hyp1`) from Wordnet as another set included in the ATTRACT pairs for lexical contrast specialisation. To the best of our knowledge, the `hyp1` pairs were not used before for lexical contrast modeling. A selection of constraints from different sets is shown in Table 1. In what follows, we test how these different configurations of constraints influence the specialisation process.

3 Experimental Setup

Training Setup and Constraints. We train the state-of-the-art specialisation model of Mrkšić et al. (2017) using suggested settings:[4] Adagrad (Duchi et al., 2011) is used for stochastic optimisation, batch size is 50, and we train for 15 epochs. To emphasise lexical contrast in the specialised space we set the respective ATTRACT and REPEL margins δ_{att} and δ_{rpl} to the same value: 1.0. We use large 300-dim skip gram vectors with bag-of-words contexts and negative sampling (SGNS-GN) (Mikolov et al., 2013), pre-trained on the 100B Google News corpus. As all other components of the model are kept fixed, the difference in performance can be attributed to the difference in the constraints used.

We experiment with external constraints employed in prior work (Zhang et al., 2014; Ono et al., 2015): these were extracted from WordNet (Fellbaum, 1998) and the Roget thesaurus (Kipfer, 2009), and comprise 1,023,082 synonymy (`syn`) pairs and 380,873 `ant` pairs. The expanded `antexp` set of antonyms contains a total of 10,334,811 word pairs. Finally, the `hyp1` set extracted from WordNet contains 326,187 word pairs.

We evaluate all specialised spaces in three standard tasks with well-defined benchmarks where modeling lexical contrast is beneficial: **1)** semantic similarity, **2)** antonymy detection, and **3)** distinguishing antonyms from synonyms. For each task, we compare against a representative selection of baselines, currently holding peak scores on the respective benchmarks. Due to a large space of models in our comparison, we refer the interested reader to the original papers for their full descriptions.

Task 1: Word Similarity. We evaluate all models on the SimLex-999 dataset (Hill et al., 2015), and SimVerb-3500 (Gerz et al., 2016), a recent verb pair similarity dataset with 3,500 verb pairs.[5] The evaluation metric is Spearman's ρ rank correlation.

Task 2: Antonymy Detection. For this task, we rely on the widely used Graduate Record Examination (GRE) dataset (Mohammad et al., 2008, 2013). The task, given an input cue word, is to select the best antonym from five options. Given a word vector space, we take the word with the largest cosine distance to the cue as the best antonym. The GRE dataset contains 950 questions in total. We report balanced F_1 scores on the entire dataset.

Task 3: Synonymy vs. Antonymy. In this binary classification task, the system must decide whether the relation between two words is synonymy or antonymy. We use the recent dataset of Nguyen et al. (2017b), comprising 1,020 noun (N) test pairs, 908 verb (V) pairs, and 1,986 adjective (A) pairs, with the equal number of synonymy and antonymy pairs in each test subset. A classification threshold decides on the relation: all word pairs with their cosine similarity above the threshold are considered synonyms, all the others are antonyms.[6]

[4] https://github.com/nmrksic/attract-repel

[5] Unlike WordSim-353 (Finkelstein et al., 2002) or MEN (Bruni et al., 2014), SimLex and SimVerb provide explicit guidelines to discern between true semantic similarity and (more broad) conceptual relatedness, so that related but non-similar words (e.g. *tiger* and *jungle*) have a low rating.

[6] Similar to the work on hypernymy detection (Santus et al., 2014; Nguyen et al., 2017a; Vulić and Mrkšić, 2018), we tune the threshold on a validation set of 206 N pairs, 182 V pairs, and 398 A pairs, also used by Nguyen et al. (2017b).

MODEL	SimLex	SimVerb
SGNS-GN (Mikolov et al., 2013)	0.414	0.348
Symmetric Patterns (Schwartz et al., 2015)	0.563	0.328
Non-distributional (Faruqui and Dyer, 2015)	0.578	0.596
Joint Specialisation (Nguyen et al., 2016)	0.590	0.516
Paragram-SL999 (Wieting et al., 2015)	0.690	0.540
Counter-fitting (Mrkšić et al., 2016)	0.740	0.628
AR: BabelNet (Mrkšić et al., 2017)	0.751	0.674
RC: ant	0.596	0.589
RC: antexp	0.606	0.551
AC: syn	0.748	0.728
AC: hyp1	0.546	0.387
AC: syn, RC: ant	0.778	0.767
AC: syn, RC: antexp	0.736	0.708
AC: syn+hyp1, RC: ant	**0.791**	**0.770**
AC: syn+hyp1, RC: antexp	0.751	0.710
Mean inter-annotator agreement	0.779	0.864

Table 2: **Task 1.** Results on two word similarity benchmarks (Spearman's ρ). Best-scoring baseline models from the literature are reported. The dashed line separates purely distributional models from the ones leveraging external lexical knowledge.

MODEL	GRE: F_1
Constraints Lookup (ANT)	0.62
SGNS-GN (Mikolov et al., 2013)	0.48
Polarity LSA (Yih et al., 2012)	0.81
Bayesian Tensor Factor. (Zhang et al., 2014)	0.82
Joint Specialisation Model (Ono et al., 2015)	0.89
RC: ant	0.79
RC: antexp	0.80
AC: syn	0.33
AC: hyp1	0.44
AC: syn, RC: ant	0.90
AC: syn, RC: antexp	0.83
AC: syn+hyp1, RC: ant	**0.92**
AC: syn+hyp1, RC: antexp	0.85

Table 3: **Task 2.** Results (F_1 scores) on the full GRE multiple-choice antonymy detection dataset.

4 Results and Discussion

Task 1: Word Similarity. A summary of the results is provided in Table 2. The most striking findings are new state-of-the-art correlation scores on both benchmarks: both are obtained by combining syn and hyp1 into ATTRACT constraints, and using the unexpanded list of antonyms as REPEL constraints. This suggests that: **1)** both ATTRACT and REPEL constraints are required to provide the synergistic effect during specialisation; **2)** a larger (and noisier) set of antonymy pairs is not necessarily more effective; **3)** the hyp1 pairs are useful for modeling lexical contrast. When included as ATTRACT constraints, these pairs lead to small but consistent gains across all three tasks (see also Tables 3-4).

MODEL	A	V	N
Symmetric Patterns (Schwartz et al., 2015)	0.718	0.584	0.482
(Roth and Schulte im Walde, 2014)	0.717	0.788	0.832
AntSynNET (Nguyen et al., 2017b)	0.784	0.777	0.855
RC: ant	0.956	0.938	0.854
RC: antexp	0.899	0.915	0.809
AC: syn	0.876	0.845	0.773
AC: hyp1	0.678	0.678	0.681
AC: syn, RC: ant	0.959	0.969	0.872
AC: syn, RC: antexp	0.951	0.955	0.871
AC: syn+hyp1, RC: ant	**0.969**	**0.975**	**0.879**
AC: syn+hyp1, RC: antexp	0.953	0.947	0.872

Table 4: **Task 3.** Results (F_1) on the synonymy-vs-antonymy evaluation set (Nguyen et al., 2017b).

The reported high score on SimLex of **0.791** is the first correlation score moving beyond mean human performance on the dataset (0.779), thus questioning the further usability of the benchmark in semantic modeling evaluation. The gain on SimVerb is even more substantial: from the previous high score of 0.674 (Mrkšić et al., 2017) to **0.770**.[7] The difference is again attributed to the use of higher-quality constraints: Mrkšić et al. (2017) relied on a noisier and smaller set from BabelNet, verifying the importance of guiding specialisation by the correct choice of constraints. In short, the specialisation model simply encodes the provided external knowledge into the input vector space, and as such it is critically tied to the constraints.

Task 2: Antonymy Detection. A summary of the results is provided in Table 3. The results suggest that antonymous REPEL constraints are more beneficial for this task, which is easily explained by the nature of the task, but the synergistic effect is again observed: both types of constraints are essential to boost the scores. The best performing configuration of constraints outperforms two strong baselines (Zhang et al., 2014; Ono et al., 2015) which also rely on the same external lexical knowledge (minus hyp1 pairs). Importantly, the results also suggest that the specialisation model indeed learns useful relationships in the specialised space beyond a simple baseline model that lookups into constraints: large gains over this baseline are reported with a variety of configurations. Distributional SGNS-GN vectors coalesce antonymy and synonymy: as a consequence, they are not a competitive baseline in any of the three evaluation tasks.

[7] We have also verified that the specialisation process is robust to the chosen distributional vector space. The best configuration of constraints from Table 2 with two other starting spaces, GLOVE (Pennington et al., 2014) and FASTTEXT (Bojanowski et al., 2017), yields respective correlation scores of 0.787 and 0.774 on SimLex and 0.764 and 0.744 on SimVerb.

The model which uses a large set of ANTEXP again cannot match performance of the model which relies on the original ANT. We see this as an interesting finding which suggests that the massive expansion of lexical constraints decreases the strength of originally provided word relationships, which were hand-crafted by linguistic experts.

Task 3: Synonymy vs. Antonymy. A summary of the results with strongest baselines from prior work is provided in Table 4: specialisation again outperforms the competitors.[8] The score differences between best-performing configurations are not as pronounced as in the other two tasks: we attribute this to the reduced task complexity. However, the results again indicate that: **1)** both types of constraints are important for distinguishing between the coalesced relations of synonymy and antonymy, with the synergistic effect again observed; **2)** the noisy and large ANTEXP set of antonyms falls short of the smaller, more accurate ANT set; and **3)** the same configuration as in the two other tasks (AC: SYN+HYP1, RC: ANT) again leads to peak performance.

5 Conclusion

We have demonstrated that post-processing specialisation models serve as a powerful tool for injecting lexical contrast knowledge into distributional word vector spaces. We have verified the hypothesis that a careful selection of external constraints is crucial for guiding the specialisation by improving state-of-the-art scores on three standard tasks used for evaluation of lexical contrast modeling: detecting antonyms, distinguishing antonyms from synonyms, and word similarity.

The post-processing specialisation models such as ATTRACT-REPEL fine-tune only vectors of words present in the external constraints. In the follow-up work, we have proposed a method which can propagate the useful external signal also to the full vocabulary (Vulić et al., 2018), leading to additional gains with specialised vectors in downstream language understanding applications. In future work, we will further investigate the full-vocabulary specialisation approaches.

[8] However, note that the specialization model cannot be directly and fairly compared to the baselines in this task, which do not use any supervision signal. The reported performance of the specialisation model can be seen as an upper bound to such distributional approaches.

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL (no 648909). The authors would like to thank the anonymous reviewers for their helpful suggestions.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

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

Alan D. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of ACL*, pages 464–469.

Christiane Fellbaum. 1998. *WordNet*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL-HLT*, pages 683–693.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *Proceedings of SLT*.

Barbara Ann Kipfer. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

George A. Miller and Christiane Fellbaum. 1991. Semantic networks of English. *Cognition*, 41(1):197 – 229.

Saif Mohammad, Bonnie J. Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of EMNLP*, pages 982–991.

Saif Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

M. Lynne Murphy. 2010. *Lexical Meaning*. Cambridge University Press.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017a. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of EMNLP*, pages 233–243.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of ACL*, pages 454–459.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017b. Distinguishing antonyms and synonyms in a pattern-based neural network. In *Proceedings of EACL*, pages 76–85.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Nghia The Pham, Angeliki Lazaridou, and Marco Baroni. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of ACL*, pages 21–26.

Michael Roth and Sabine Schulte im Walde. 2014. Combining word patterns and discourse markers for paradigmatic relation classification. In *Proceedings of ACL*, pages 524–530.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pages 1793–1803.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT*.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of NAACL-HLT*.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017a. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*, pages 2546–2558.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017b. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL*, pages 56–68.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Wen-tau Yih, Geoffrey Zweig, and John Platt. 2012. Polarity inducing Latent Semantic Analysis. In *Proceedings of EMNLP*, pages 1212–1222.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. 2014. Word semantic representations using bayesian probabilistic tensor factorization. In *Proceedings of EMNLP*, pages 1522–1531.