Combining Discourse Markers and Cross-lingual Embeddings for Synonym–Antonym Classification

Michael Roth[1], **Shyam Upadhyay**[2]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Department of Computer and Information Science, University of Pennsylvania
rothml@ims.uni-stuttgart.de, shyamupa@seas.upenn.edu

Abstract

It is well-known that distributional semantic approaches have difficulty in distinguishing between synonyms and antonyms (Grefenstette, 1992; Padó and Lapata, 2003). Recent work has shown that supervision available in English for this task (e.g., lexical resources) can be transferred to other languages via cross-lingual word embeddings. However, this kind of transfer misses monolingual distributional information available in a target language, such as contrast relations that are indicative of antonymy (e.g., *hot...while...cold*). In this work, we improve the transfer by exploiting monolingual information, expressed in the form of co-occurrences with discourse markers that convey contrast. Our approach makes use of less than a dozen markers, which can easily be obtained for many languages. Compared to a baseline using only cross-lingual embeddings, we show absolute improvements of 4–10% F_1-score in Vietnamese and Hindi.

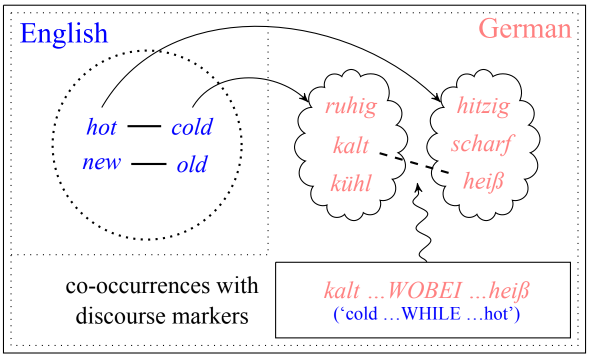

Figure 1: Supervision for distinguishing antonyms from synonyms can be derived using discourse markers. Here, antonyms available in English (denoted by solid edge between *hot* and *cold*) are translated to German via cross-lingual word embeddings. Using co-occurrences with discourse markers indicative of antonymy (shown in box), we can identify pairs of words in the n-best translations (clouds) that are also antonymous (e.g., dashed edge between *kalt* and *heiß*).

1 Introduction

Recent work has shown that monolingual word embeddings in different languages can be aligned in an unsupervised manner (Artetxe et al., 2018; Kementchedjhieva et al., 2018). The resulting *cross-lingual embeddings* can be used to share supervision for lexical classification tasks across languages, when annotated data is not available in one language. For instance, a model for distinguishing lexical relations such as hypernymy and meronymy can be transferred to other languages (Glavaš and Vulić, 2018). However, this kind of transfer, using only cross-lingual embeddings, misses useful monolingual information available in the target language.

In this paper, we consider one lexical classification task, namely the distinction between synonyms and antonyms, which is important for downstream applications such as contradiction detection (Harabagiu et al., 2006; Marneffe et al., 2008; Voorhees, 2008) and machine translation (Marton et al., 2011). To facilitate better transfer, we propose to use monolingual information in the form of word co-occurrences in contrast relations, in addition to cross-lingual embeddings (see Figure 1). In particular, we utilize the fact that discourse markers conveying contrast are more likely to be surrounded by antonyms than synonyms (e.g., *hot...while...cold*), as shown by Roth and Schulte im Walde (2014).

Our analysis reveals that (1) fine-grained semantic information that is required to distinguish synonyms from antonyms is insufficiently preserved cross-lingually in word embeddings, but (2) such information can be recovered (at least partially) by relying on linguistic intuitions about contrast relations in discourse.

Proceedings of NAACL-HLT 2019, pages 3899–3905
Minneapolis, Minnesota, June 2 - June 7, 2019.

2 Related Work

The identification of paradigmatic relations such as synonymy and antonymy has been a task of interest for more than a decade. Early work focused on the identification of instances of a single relation: for example, Charles and Miller (1989) investigated co-occurrences as an indicator for antonymy, Hearst (1992) introduced a pattern-based approach to identify hypernymy, and Edmonds and Hirst (2002) applied distributional statistics to identify synonymy.

Beyond the *identification* of instances of a particular relation, more recent approaches attempt to *distinguish* relations such as synonymy and antonymy, based on lexico-syntactic patterns, distributional information, or combinations of both (Lin et al., 2003; Shwartz and Dagan, 2016). As supervision for weighting different features, most recent work makes use of lexical resources and/or lists of affix patterns that indicate contrast morphologically (Yih et al., 2012; Mohammad et al., 2013; Ono et al., 2015).

Supervision for lexical classification tasks is not available in all languages. To overcome this difficulty, some approaches (Mrkšić et al., 2017; Glavaš and Vulić, 2018) combine resources for English and cross-lingual word embeddings to distinguish lexical relations in other languages. That is, they train and test one model *across* different languages. In contrast, we propose to first use unsupervised translation techniques to *transfer supervision* into a target language, given resources available in English. More specifically, our approach transfers supervision using a combination of unsupervised cross-lingual embeddings (Artetxe et al., 2018) and word co-occurrences with discourse markers that indicate contrast. In doing so, our approach generalizes the discourse-marker based model for relation classification by Roth and Schulte im Walde (2014) to a cross-lingual setting.

3 Our Approach

In this work, we address the task of distinguishing synonyms and antonyms cross-lingually: given a word pair (a, b) in a target language, a model trained in a different source language determines whether this constitutes a synonym or an antonym pair. Our approach consists of two main ingredients: a *translation module* that creates training data in a target language (see Figure 2), and a *classification module* that uses this data in a supervised fashion. Our focus is on the first part, which we describe in more detail below. For classification, we re-use the LexNet model by Shwartz and Dagan (2016), which is a feed-forward network that uses distributional word representations as input.

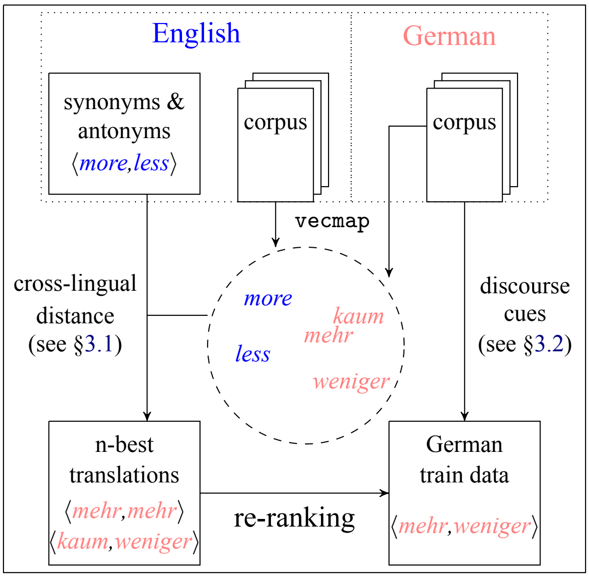

Figure 2: In our approach, we create training data in multiple languages by translating English data via cross-lingual word embeddings and discourse cues.

3.1 Cross-lingual Embeddings for Translation

Our transfer approach addresses a general setting that only assumes availability of monolingual data. Accordingly, we induce a translation dictionary of synonyms and antonyms in an unsupervised way, instead of relying on manually crafted dictionaries. We achieve this by finding n-best translations of English words in a cross-lingual word embedding space. We create this space using `vecmap`, a method that aligns and maps embedding spaces in two languages (Artetxe et al., 2018). Formally, we define the n^{th} best translation $w_t^{(n)}$ of a source word w_s to be the n^{th} nearest neighbor in the joint semantic space (based on cosine similarity).

3.2 Re-ranking n-best Translations

While 1$^{\text{st}}$ best translations are often accurate, erroneous translations occur for example when two related words in a source language are close to the same words in the target language (e.g., *less* and *more* share the same nearest neighbor in Figure 2: *mehr* 'more'). In such cases, we aim to use monolingual information about lexical relations in the form of discourse cues. Specifically, given two antonymous words a_s and b_s (henceforth $a_s \perp b_s$)

Language	Corpus size (in words)	Data size (in pairs)	Ratio (#syn:#ant)
English (train)	2.1 B	10 932	50:50
German (dev)	0.7 B	811	52:48
Hindi	0.5 B	1 000	50:50
Vietnamese	0.2 B	353	45:55

Table 1: Data statistics of the classification datasets as well as the corpora used to create word representations. Note that the data used here consist of only synonyms and antonyms (approximately in equal proportions).

in the source language, we want to choose translations $a_t^{(i)}$ and $b_t^{(j)}$ that are most likely antonymous as well. We achieve this by re-ranking translations $a_t^{(i)}$ and $b_t^{(j)}$ such that the probability $\Pr(a_t^{(i)} \perp b_t^{(j)})$ of them appearing in contrast relations is maximized. We only assume a set of discourse markers M that indicate contrast as knowledge about the target language. We approximate $\Pr(a_t^{(i)} \perp b_t^{(j)})$ by computing the pointwise mutual information (PMI) between two terms, conditioned on co-occurrence *around* a discourse marker $m \in M$:

$$\Pr(a_t^{(i)} \perp b_t^{(j)}) \sim \min_{m \in M} \log \frac{N(m)\, n(a_t^{(i)}, b_t^{(j)}|m)}{n(a_t^{(i)}|m)\, n(b_t^{(j)}|m)}$$

where $N(m)$ is the frequency of discourse marker m in the corpus, $n(w|m)$ is the number of times term w occurs with m, and $n(w_1, w_2|m)$ is the number of co-occurrences $w_1 ... m ... w_2$. For simplicity, we count all joint occurrences within a sentence, regardless of sentence length and distance between words. In practice, the set M is constructed by manually translating eight discourse markers that frequently indicate contrast relations according to the Penn Discourse Treebank 2.0 (Prasad et al., 2007): *although*, *by comparison*, *by contrast*, *however*, *nevertheless*, *nonetheless*, *though*, and *thus*.

4 Experiments

Our experiments test the performance of an off-the-shelf system for lexical relation classification under different cross-lingual settings. In particular, we evaluate the performance of unsupervised cross-lingual word embeddings and assess the benefits of translation and re-ranking, taking into account word co-occurrences in contrast relations.

Experimental setup. For our experiments, we use four languages: English, German, Hindi and Vietnamese. The antonymy and synonymy dataset by Nguyen et al. (2017) is used for training and estimating an upper bound in English. For cross-lingual development and hyper-parameter selection, we use the German dataset by Glavaš and Vulić (2018). Specifically, we select the number of hidden layers {0, <u>1</u>}, learning rate {<u>0.001</u>,…,0.1}, dropout rate {0.0, <u>0.5</u>}, and whether to include information on paths between words {yes, <u>no</u>}.[1] As examples of under-resourced languages, we evaluate on the Vietnamese dataset ViCon by Nguyen et al. (2018) and on a new Hindi dataset, which we crawled from `hindi2dictionary.com`.[2] Note that annotated data in these languages is held out exclusively for testing.

The task addressed here is to distinguish synonymous from antonymous words. Accordingly, we consider only word pairs for training and testing that are marked as antonyms or synonyms, even if the original dataset also contains unrelated words or other relations. Statistics of all considered word pairs, the ratio between synonyms and antonyms, as well as sizes of the text corpora used in our experiments are given in Table 1.

Models. We test our proposed approach against two baselines: **NoTrans** and **BestTrans**. **NoTrans** simply uses the training data in English and allows us to test in how far semantic properties related to the distinction between synonymy and antonymy are preserved cross-lingually. **BestTrans** uses 1st best translations of the English training data in a target language, as described in Section 3.1.

Our own approach, henceforth called **LingTrans**, is based on n-best translations and exploits word co-occurrences in contrast relations for re-ranking, as described in Section 3.2.

All three models take cross-lingual word embeddings as input. We created these as follows. First, the unsupervised morphological analyzer of Xu et al. (2018) is used to lemmatize the monolingual corpora for the respective languages—the German Wikipedia, the Vietnamese portion of the LORELEI pack (Strassel and Tracey, 2016), and the HindEnCorp corpus (Bojar et al., 2014). This step ensures that we can compare the (stems of)

[1] Since we do not assume access to syntactic paths in the target languages, we only experimented with lexical surface paths and found them to consistently degrade performance.

[2] To verify the correctness of the crawled data, we asked a native speaker to validate all instances. Since synonym/antonym labels were derived automatically, validation was only performed to remove noisy instances.

LexNet-crosslingual	de (dev)	hi	vi
NoTrans	59.9	54.6	42.2
BestTrans	62.2	59.2	46.9
LingTrans	**64.2***	**63.5***	**56.8****
LexNet-monolingual (en)		70.1	
+ path embeddings		**74.3**	

Table 2: Macro-averaged F_1-scores for crosslingual synonymy/antonymy distinction in German (de), Hindi (hi), and Vietnamese (vi). Significant differences from NoTrans are marked by asterisk(s) (* $p<0.1$, ** $p<0.01$). Monolingual results in English (en) are only shown for comparison.

word tokens in text to those of the word forms that actually appear in the synonymy/antonymy datasets. Note that, while we use an unsupervised morphological analyzer, a stemmer can also be used, if available for that language. Next, we created monolingual embeddings for each language using fastText (Bojanowski et al., 2017). Finally, we applied the unsupervised variant of `vecmap` (Artetxe et al., 2018) to compute alignments and cross-lingual mappings.

The word and discourse marker co-occurrence counts required for our approach are computed on the same monolingual corpora used for training monolingual embeddings.

Results. Table 2 shows macro-averaged F_1-scores for crosslingual synonymy/antonymy distinction (top part) as well as monolingual results in English for comparison (bottom part). The monolingual results show that word embeddings provide appropriate information for classification in most instances, achieving an F_1-score of 70.1. The comparatively low cross-lingual results by **NoTrans** indicate that aligning and mapping embedding spaces does not preserve all semantic properties relevant to the distinction between synonymy and antonymy. The use of first-best translations in **BestTrans** alleviates this issue partially and improves F_1 by up to 4.6 points. Our intuition regarding discourse cues in **LingTrans** leads to further improvements of up to 9.9 additional points in F_1 and considerably closes the gap between monolingual and cross-lingual results (56.8−64.2 vs. 70.1 F_1). Based on an approximate randomization test over the respective test items (Yeh, 2000), we find the improvements of our proposed approach **LingTrans** over **NoTrans** to be significant in Vietnamese ($p<0.01$), German and Hindi ($p<0.1$). In contrast, there is no significant difference in performance between **BestTrans** and **NoTrans**, confirming that both translation *and* reranking are required to achieve consistent gains.

5 Analysis

We examine how far our proposed approach affects performance on the task of synonymy–antonymy distinction and discuss remaining shortcomings.

A first observation concerns the general variance of results across languages. In addition to differences in terms of available corpus sizes, we observe different challenges in each language and dataset. Notably, each dataset is created with its own linguistically motivated definitions of antonymy and synonymy, some of which are more relaxed than others. For example, Nguyen et al. (2018) consider "words which are strongly associated but highly dissimilar to each other" as antonyms and "words that are highly similar in meaning" as synonyms.

The performance of our baseline **NoTrans** suggests that the distinction between synonyms and antonyms in the crosslingual space is harder for Vietnamese than for the other languages. At least partially, this is due to errors and noise from preprocessing (i.e., lemmatization). Also, Vietnamese belongs to a different language family than English. Consequently, information relevant to the synonymy–antonymy distinction may be distributed differently in the target language space than in the source language. In Hindi, some errors specific to the data involve synonyms that denote the same mythological deity (e.g., हनुमान / बजरंगबली 'Hanuman/Bajrangbali') and thus require factual knowledge for classification.

Even for our development language German, results suggest that there is room for improvement. We examine the improvements achieved and remaining errors in German in more detail below.

Classification improvements. The main improvement of our **LingTrans** approach over **BestTrans** is reflected in an increased number of correctly classified antonyms, mostly including antonymous adjectives (e.g., *ständig* 'steadily' vs. *sporadisch* 'sporadically') but also mutually exclusive nouns (e.g., *Privatgelände* 'private grounds' vs. *Öffentlichkeit* 'public'). Compared to the **NoTrans** baseline, we find **BestTrans** to substantially increase the number of correctly classified synonyms (by almost 20%). In particular, this affects words derived from the same stem (*unzählig/zahllos* 'countless'), synonymous verbs (*pfuschen/schummeln* 'cheat') as well as (near-)synonymous adjectives (*verfügbar/verwendbar* 'available/usable').

Translation improvements. The observed quantitative improvements are in line with qualitative improvements that we see in the automatically generated training data. For example, *more* and *less* have the same 1st-best translation in German according to the cross-lingual space (*mehr* 'more'); as illustrated in Figure 2, reranking is required to find the correct translation for the latter word (i.e., *weniger* 'less'). To quantify translation improvements of our proposed reranking method, we count the number of words overlapping between the automatically translated data and data available in the target languages. Compared to the **BestTrans** baseline, we find our **LingTrans** approach to improve overlap for German, Hindi, and Vietnamese by 2%, 5% and 5%, respectively. Despite the increase in word overlap, it is worth noting that the overlap between training and testing in terms of actual data instances (i.e., pairs of words) remains constant between 0% and 1%. A partial explanation for the improved results could thus be that our reranking approach lets us find translations that are *generally* more likely to have an antonym (or synonym)—regardless of which other word they are presented with. A similar observation was made by Levy et al. (2015) regarding hypernymy classification, in which they found some words to be generally more likely to be "category words". Whether similar biases exist in synonymy–antonymy classification will be subject of future work.

Remaining errors. We categorized a randomly selected subset of 30 mis-classifications. We found 30% of the errors to be related to sparsity. In these cases, at least one word was infrequent or the word pair never co-occurred in the corpus (e.g., *unversöhnbar/unvereinbar* 'unforgiving/irreconcilable'). 27% of the errors can be attributed to problems with the test data because it contains "synonyms" that are only indirectly related (e.g., *parieren/ausweichen* 'parry/dodge').

The 42% remaining cases involve a number of challenging antonyms and synonyms, but also a few remarkably easy cases. The latter include morphologically marked antonyms (e.g., *Richtigkeit* 'correctness' vs. *Unrichtigkeit* 'incorrectness') and instances of synonyms that involve two identical words (e.g., *Verwandtschaft/Verwandtschaft* 'kinship'), which presumably is an artifact of the automatic translation step used in creating the German dataset (see Glavaš and Vulić (2018)). Both types of errors could be identified by simple heuristics but are incorrectly classified by the model. On the other hand, hard cases involve words of different linguistic register, which are expected to be distributed differently and therefore hard to capture by distributional methods (e.g., *mittellos/verarmt* 'penniless/impoverished').

In summary, we find substantial improvement through our transfer method, compared to the baselines. However, further improvements could be achieved, for example, taking into account register (e.g., formal vs. colloquial) and morphological marking (e.g., negation affixes).

6 Conclusion

We proposed to combine unsupervised cross-lingual embeddings and discourse cues to generate supervision for distinguishing synonyms and antonyms in under-resourced languages. Compared to a baseline that uses only cross-lingual embeddings, we showed that the use of a small set of discourse markers indicating contrast can yield absolute improvements of up to 10% F_1-score. The simplicity of our approach allows to easily incorporate other features (e.g., morphological marking) and it can be extended to further languages or lexical relations. For example, discourse markers that indicate specification and instantiation relations (e.g. *specifically*, *for instance*) could be used to detect hypernymy (cf. Roth and Schulte im Walde (2014)). Beyond classification, another direction for future work is to extend our approach to distinguish synonyms and antonyms from unrelated word pairs. An interesting direction to pursue would be to use multiple related languages, to aid lexical relation classification in an under-resourced language, instead of transferring supervision from a single language (English).

Acknowledgements

This research was supported by the German Research Foundation (DFG EXC 285), and by Contract HR0011-15-2-0025 and Agreement HR0011-15-2-0023 with the US DARPA.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Walter G Charles and George A Miller. 1989. Contexts of Antonymous Adjectives. *Applied psycholinguistics*, 10(3):357–375.

Philip Edmonds and Graeme Hirst. 2002. Near-Synonymy and Lexical Choice. *Computational linguistics*, 28(2):105–144.

Goran Glavaš and Ivan Vulić. 2018. Discriminating between Lexico-Semantic Relations with the Specialization Tensor Model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 181–187. Association for Computational Linguistics.

Gregory Grefenstette. 1992. Finding Semantic Similarity in raw text: The Deese Antonyms. In *Fall Symposium Series, Working Notes, Probabilistic Approaches to Natural Language*, pages 61–65.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In *AAAI*, volume 6, pages 755–762.

M. A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976. Association for Computational Linguistics.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying Synonyms among Distributionally Similar Words. In *IJCAI*, volume 3, pages 1492–1493.

Marie-Catherine Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding Contradictions in Text. In *Proceedings of ACL-08: HLT*, pages 1039–1047.

Yuval Marton, Ahmed El Kholy, and Nizar Habash. 2011. Filtering Antonymous, Trend-contrasting, and Polarity-dissimilar Distributional Paraphrases for Improving Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 237–249. Association for Computational Linguistics.

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

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 Association for Computational Linguistics*, 5:309–324.

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: Volume 1, Long Papers*, pages 76–85. Association for Computational Linguistics.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2018. Introducing Two Vietnamese Datasets for Evaluating Semantic Models of (Dis-)Similarity and Relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 199–205. Association for Computational Linguistics.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989.

Sebastian Padó and Mirella Lapata. 2003. Constructing Semantic Space Models from Parsed Corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 annotation manual. Technical report.

Michael Roth and Sabine Schulte im Walde. 2014. Combining Word Patterns and Discourse Markers for Paradigmatic Relation Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 524–530, Baltimore, Maryland. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2016. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 24–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In *LREC*.

Ellen M Voorhees. 2008. Contradictions and Justifications: Extensions to the Textual Entailment Task. *Proceedings of ACL-08: HLT*, pages 63–71.

Hongzhi Xu, Mitchell Marcus, Charles Yang, and Lyle Ungar. 2018. Unsupervised Morphology Learning with Statistical Paradigms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 44–54, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics, Volume 2*, pages 947–953.

Wen-tau Yih, Geoffrey Zweig, and John C Platt. 2012. Polarity Inducing Latent Semantic Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1212–1222. Association for Computational Linguistics.