

Early and Late Combinations of Criteria for Reranking Distributional Thesauri

Olivier Ferret

CEA, LIST, Vision and Content Engineering Laboratory,
Gif-sur-Yvette, F-91191 France.
olivier.ferret@cea.fr

Abstract

In this article, we first propose to exploit a new criterion for improving distributional thesauri. Following a bootstrapping perspective, we select relations between the terms of similar nominal compounds for building in an unsupervised way the training set of a classifier performing the reranking of a thesaurus. Then, we evaluate several ways to combine thesauri reranked according to different criteria and show that exploiting the complementary information brought by these criteria leads to significant improvements.

1 Introduction

The work presented in this article aims at improving thesauri built following the distributional approach as implemented by (Grefenstette, 1994; Lin, 1998; Curran and Moens, 2002). A part of the work for improving such thesauri focuses on the filtering of the components of the distributional contexts of words (Padró et al., 2014; Polajnar and Clark, 2014) or their reweighting, either by turning the weights of these components into ranks (Broda et al., 2009) or by adapting them through a bootstrapping method from the thesaurus to improve (Zhitomirsky-Geffet and Dagan, 2009; Yamamoto and Asakura, 2010). The other part implies more radical changes, including dimensionality reduction methods such as Latent Semantic Analysis (Padó and Lapata, 2007), multi-prototype (Reisinger and Mooney, 2010) or exemplar-based models (Erk and Pado, 2010), neural approaches (Huang et al., 2012; Mikolov et al., 2013) or the adoption of a Bayesian viewpoint (Kazama et al., 2010; Dinu and Lapata, 2010).

Our work follows (Ferret, 2012), which proposed a different way from (Zhitomirsky-Geffet and Dagan, 2009) to exploit bootstrapping by se-

lecting in an unsupervised way a set of semantically similar words from an initial thesaurus and training from them a classifier to rerank the semantic neighbors of the initial thesaurus entries. More precisely, we propose a new criterion for this selection, based on the similarity relations of the components of similar compounds, and we show two modes – early and late – of combination of thesauri reranked from different criteria, including ours, leading to significant further improvements.

2 Reranking a distributional thesaurus

Distributional thesauri are characterized by heterogeneous performance in their entries, even for high frequency entries. This is a favorable situation for implementing a bootstrapping approach in which the results for “good” entries are exploited for improving the results of the other ones. However, such idea faces two problems: first, detecting “good” entries; second, learning a model from them for improving the performance of the other entries.

The first issue consists in selecting without supervision a set of positive and negative examples of similar words that represents a good compromise between its error rate and its size. Straightforward solutions such as using the similarity value between an entry and its neighbors or relying on the frequency of entries are not satisfactory in terms of error rate. Hence, we propose in Section 3 a new method, based on the semantic compositionality hypothesis of compounds, for achieving this selection in a more indirect way and show the interest to combine it with the criterion of (Ferret, 2012) for building a large training set with a reasonable error rate.

We address the second issue by following (Hagiwara et al., 2009), which defined a Support Vector Machine (SVM) model for deciding whether two words are similar or not. In our context, a positive example is a pair of nouns that are

semantically similar while a negative example is a pair of non similar nouns. The features of each pair of nouns are built by summing the weights of the elements shared by their distributional representations, which are vectors of weighted cooccurents. Cooccurents not shared by the two nouns are given a null weight.

This SVM model is used for improving a thesaurus by reranking its semantic neighbors as follows: for each entry E of the thesaurus, the representation as an example of the word pair $(E, neighbor)$ is built for each of the neighbors of E and submitted to the SVM model in classification mode. Finally, all the neighbors of E are reranked according to the value of the decision function computed for each neighbor by the SVM model.

3 Unsupervised example selection

The evaluation of distributional thesauri shows that a true semantic neighbor is more likely to be found when the thesaurus entry is a high frequency noun and the neighbor has a low rank. However, relying only on these two criteria doesn't lead to a good enough set of positive examples. For instance, taking as positive examples from the initial thesaurus of Section 4 the first neighbor of its 2,148 most frequent entries, the number of positive examples of (Hagiwara et al., 2009), only leads to 44.3% of correct examples. Moreover, this percentage exceeds 50% only when the number of examples is less than 654, which represents a very small training set for this kind of task.

Hence, we propose a more selective approach for choosing positive examples among high frequency nouns to get a more balanced solution between the number of examples and their error rate. This approach exploits a form of semantic compositionality hypothesis of compounds. While much work has been done recently for defining the distributional representation of compounds by composing the distributional representations of their components (Mitchell and Lapata, 2010; Paperno et al., 2014), we adopt a kind of reverse viewpoint by exploiting the possibility to link the meaning of a compound to the meaning of its components. More precisely, we assume that the mono-terms of two semantically related compounds with the same syntactic role in their compound are likely to be semantically linked themselves.

In this work, we only consider compounds having one of these three term structures (with their

percentage of the vocabulary of compounds):

$$(a) \langle \text{noun} \rangle_{mod} \langle \text{noun} \rangle_{head} \quad (30)$$

$$(b) \langle \text{adjective} \rangle_{mod} \langle \text{noun} \rangle_{head} \quad (58)$$

$$(c) \langle \text{noun} \rangle_{head} \langle \text{preposition} \rangle \langle \text{noun} \rangle_{mod} \quad (12)$$

Each compound C_i is represented as a pair (H_i, M_i) , where H_i stands for the head of the compound whereas M_i represents its modifier (*mod*). According to the assumption underlying our selection procedure, if a compound (H_2, M_2) is a semantic neighbor of a compound (H_1, M_1) (*i.e.* at most its c^{th} neighbor in a distributional thesaurus of compounds), we can expect H_1 and H_2 on one hand and M_1 and M_2 on the other hand to be semantically similar. Since distributional thesauri of compounds are far from being perfect, we added constraints on the matching of the components of two compounds. More precisely, the positive examples of semantically similar nouns (noun pairs after \rightarrow) are selected by the three following rules, where $H_1 = H_2$ means that H_1 is the same word as H_2 and $H_1 \equiv H_2$ means that H_2 is at most the m^{th} neighbor of H_1 in the initial thesaurus of mono-terms (but is different from H_1):

$$(1) H_1 \equiv H_2 \ \& \ M_1 = M_2 \rightarrow (H_1, H_2)$$

$$(2) M_1 \equiv M_2 \ \& \ H_1 = H_2 \rightarrow (M_1, M_2)$$

$$(3) M_1 \equiv M_2 \ \& \ H_1 \equiv H_2 \rightarrow (H_1, H_2), (M_1, M_2)$$

The selection of negative examples is also an important issue but benefits from the fact that the number of semantic neighbors of an entry that are actually semantically linked to this entry in a distributional thesaurus quickly decreases as their rank increase. In the experiments of Section 4, we built negative examples from positive examples by turning each positive example (A,B) into two negative examples: (A, *rank 10 A neighbor*) and (B, *rank 10 B neighbor*). Choosing neighbors with a higher rank would have guaranteed fewer false negative examples but taking neighbors with a rather small rank for building negative examples is more useful in terms of discrimination.

4 Experiments and evaluation

4.1 Building of distributional thesauri

The first step of the work we present is the building of two distributional thesauri: the thesaurus of mono-terms to improve (A2ST) and a thesaurus of compounds (A2ST-comp). Similarly to (Ferret, 2012), they were both built from the AQUAINT-2 corpus, a 380 million-word corpus of news articles in English. The building procedure, defined

by (Ferret, 2010), was also identical to (Ferret, 2012), with distributional contexts compared with the Cosine measure and made of window-based lemmatized cooccurents (1 word before and after) weighted by Positive Pointwise Mutual Information (PPMI). For the thesaurus of compounds, a preprocessing step was added to identify nominal compounds in texts. This identification was done in two steps: first, a set of compounds were extracted from the AQUAINT-2 corpus by relying on a restricted set of morpho-syntactic patterns applied by the Multiword Expression Toolkit (`mwetoolkit`) (Ramisch et al., 2010); then, the most frequent compounds in this set (frequency > 100) were selected as reference and their occurrences in the AQUAINT-2 corpus were identified by applying the longest-match strategy to the output of the *TreeTagger* part-of-speech tagger (Schmid, 1994)¹. Finally, distributional contexts made of mono-terms and compounds were built as stated above and neighbors were found for 29,174 compounds.

4.2 Example selection

We applied the three rules of Section 3 with all the entries of our thesaurus of compounds and the upper half in frequency of our mono-term entries. For mono-terms, we only took the first neighbor ($m = 1$) of each entry because of the rather low performance of the initial thesaurus while for compounds, a larger value ($c = 3$) was chosen for enlarging the number of selected examples since neighbors were globally more reliable (see results of Table 2). As the selection method makes the definition of a development set quite difficult, the values of these two parameters were chosen in a conservative way.

Table 1 gives for each rule and two combinations of them the number of selected positive examples (#pos. ex.) and the percentage of positive (%good pos.) and negative examples (%bad neg.) found in our Gold Standard resource for thesaurus evaluation. This resource results from the union of the synonyms of WordNet 3.0 and the associated words of the Moby thesaurus. Table 1 also gives the same data for examples selected by the method of (Ferret, 2012) (*symmetry* row, *sym.* for short), based on the fact that as similarity relations are

¹Longest-match strategy: if C1 is a reference compound that is part of a reference compound C2, the identification of an occurrence of C2 blocks out the identification of the associated occurrence of C1.

method	%good pos.	%bad neg.	#pos. ex.
<i>symmetry</i>	59.7	12.4	796
(1)	56.9	16.1	921
(2)	44.7	14.7	308
(3)	46.2	16.9	40
rules (1,2)	53.0	16.1	1,115
rules (1,2,3)	52.4	15.9	1,131
sym. + (1,2)	54.3	15.0	1,710
sym. + (1,2,3)	53.9	14.5	1,725

Table 1: Selection of examples.

symmetric, a pair of words (A,B) are more likely to be similar if the first neighbor of A is B and the first neighbor of B is A. The data for the union of the examples produced by the two methods also appear in Table 1.

Concerning the method we propose, Table 1 shows that rule (3), which is a priori the least reliable of the three rules as it only requires similarity and not equality for both heads and modifiers, actually produces a very small set of examples that tends to degrade global results. As a consequence, only the combination of rules (1) and (2) is used thereafter (row in bold). Table 1 also suggests that the heads of two semantically linked compounds are more likely to be actually linked themselves if they have the same modifier than the modifiers of two semantically linked compounds having the same head. This confirms our expectation that the head of a compound is more related to the meaning of the compound than its modifier. More globally, Table 1 shows that the *symmetry* method has higher results than the second one but their association produces an interesting compromise between the number of examples, 1,710, and its error rate, 45.7. The fact that the two methods only share 201 noun pairs also illustrates their complementarity.

4.3 Reranking evaluation

For our SVM models, we adopted the RBF kernel, as (Hagiwara et al., 2009), and a grid search strategy for optimizing both the γ and C parameters by applying a 5-fold cross validation procedure to our training set and adopting the precision measure as the evaluation function to optimize. The models were built with LIBSVM (Chang and Lin, 2001) and then applied to the neighbors of our initial thesaurus.

Table 2 gives the results of the reranking for both the method we propose, *compound* (*comp.* for short), with examples selected by rules (1) and

(2), and the one of (Ferret, 2012), *symmetry*. In either case, they correspond to an intrinsic evaluation achieved by comparing the semantic neighbors of each thesaurus entry with the synonyms and related words of our Gold Standard resource for that entry. 12,243 entries with frequency > 10 were present in this resource and evaluated in such a way. As the neighbors are ranked according to their similarity value with their entry, we adopted the classical evaluation measures of Information Retrieval by replacing documents with synonyms and queries with entries: R-precision (R-prec.), Mean Average Precision (MAP) and precision at different cut-offs (1, 5 and 10).

More precisely, the *initial* row of Table 2 gives the values of these measures for our initial thesaurus of mono-terms while its *A2ST-comp* row corresponds to the measures for our thesaurus of compounds. It should be note that in the case of the A2ST-comp thesaurus, the number of evaluated entries is very small, restricted to 813 entries, with also a very small number of reference synonyms by entry. Hence, the results of the evaluation of A2ST-comp have to be considered with caution even if their high level for the very first semantic neighbors tends to confirm the positive impact of the low level of ambiguity of compounds compared to mono-terms.

The two following rows gives the results of the thesauri built from the best models of (Baroni et al., 2014), *B14-count* for the count model, whose main parameters are close or identical to ours, and *B14-predict* for the predict model, built from (Mikolov et al., 2013). These results first illustrate the known importance of corpus size, as the (Baroni et al., 2014)’s corpus is more than 7 times larger than ours, and the fact that for building thesauri, the count model is superior to the predict model. This last observation is confirmed by the results of the skip-gram model of (Mikolov et al., 2013) with its best parameters² for our corpus (5th row), which clearly exhibits worst results than *initial*. For this *Mikolov* thesaurus and the following reranked ones, each value corresponds to the difference between the measure for the considered thesaurus and the measure for the initial thesaurus. All these differences were found statistically significant according to a paired Wilcoxon test with p-value < 0.05 .

²word2vec -cbow 0 -size 600 -window 10 -negative 0 -hs 0 -sample 1e-5

Thesaurus	R-prec.	MAP	P@1	P@5	P@10
initial (A2ST)	7.7	5.6	22.5	14.1	10.8
A2ST-comp	32.7	39.5	34.9	12.3	7.1
B14-count	12.5	9.8	31.9	19.6	15.2
B14-pred	10.9	8.5	30.3	18.4	13.8
Mikolov	-2.2	-1.4	-6.2	-4.6	-3.8
symmetry	+0.3	+0.1	+2.1	+0.8	+0.6
compound	+0.1	+0.0	+2.0	+0.9	+0.6
sym.+comp.	+0.3	+0.2	+2.8	+1.2	+0.9
RRF	+0.7	+0.6	+3.7	+1.9	+1.4
borda	+0.7	+0.5	+3.6	+1.7	+1.3
condorcet	+0.5	+0.4	+3.4	+1.6	+1.2
CombSum	+0.9	+0.8	+4.7	+2.2	+1.5
CS-w-Mik	+1.2	+1.4	+4.2	+2.0	+1.5

Table 2: Evaluation of our initial thesaurus and its reranked versions (values = percentages).

The analysis of the next two rows of Table 2 first shows that each criterion used for reranking our initial thesaurus leads to a global increase of results. The extent of this increase is quite similar for the two criteria: *symmetry* slightly outperforms *compound* but the difference is not significant. This increase is higher for $P@\{1,5,10\}$ than for R-precision and MAP, which can be explained by the high number of synonyms and related words, 38.7 on average, that an entry of our initial thesaurus has in our reference. Hence, even a significant increase of $P@\{1,5,10\}$ may have a modest impact on R-precision and MAP as the overall recall, equal to 9.8%, is low.

4.4 Thesaurus fusion

Having several thesauri reranked according to different criteria offers the opportunity to apply ensemble methods. Such idea was already experimented in (Curran, 2002) for thesauri built with different parameters (window or syntactic based cooccurents, etc). We tested more particularly two general strategies for data fusion (Atrey et al., 2010): early and late fusions. The first one consists in our case in fusing the training sets built from our two criteria. As for each criterion, a classifier is then built from the fused training set and applied for reranking the initial thesaurus (see the *sym.+comp.* row of Table 2).

Table 3 illustrates qualitatively the impact of this first strategy for the entry *esteem*. Its **WordNet** row gives all the synonyms for this entry in WordNet while its Moby row gives the first related words for this entry in Moby. In our *initial*

WordNet	respect, admiration, regard
<u>Moby</u>	admiration, appreciation, acceptance, dignity, regard, respect, account, adherence, consideration, estimate, estimation, fame, greatness, homage + 79 words more
initial	cordiality, gratitude, admiration , comradeship, back-scratching, perplexity, respect , ruination, appreciation , neighbourliness ...
sym.+comp.	respect , admiration , trust, recognition, gratitude, confidence, affection, understanding, solidarity, dignity , appreciation , regard , sympathy, acceptance ...

Table 3: Reranking for the entry *esteem* with the early fusion strategy.

thesaurus, the first two neighbors of *esteem* that are present in our reference resources are *admiration* (rank 3) and *respect* (rank 7). The reranking produces a thesaurus in which these two words appear as the first two neighbors of the entry while its third synonym in WordNet raises from rank 22 to rank 12. Moreover, the number of neighbors among the first 14 ones that are present in Moby increases from 3 to 6.

The late fusion strategy relies on the methods used in Information Retrieval for merging ranked lists of retrieved documents. More precisely, we experimented the Borda, Condorcet (Nuray and Can, 2006) and Reciprocal Rank (RRF) (Cormack et al., 2009) fusions based on ranks and the CombSum fusion based on similarity values, normalized in our case with the Zero-one method (Wu et al., 2006). The corresponding thesauri were built by fusing, entry by entry, the lists of neighbors coming from the *initial*, *symmetry* and *compound* thesauri.

Table 2 first shows that all the thesauri produced by our ensemble methods outperform our first three thesauri, which confirms that *initial*, *symmetry* and *compound* can bring complementary information, exploited by the fusion. It also shows that our late fusion methods are more effective than our early fusion method. However, no specific element advocates at this stage for a generalization of this observation. The evaluation reported by Table 2 also suggests that for fusing distributional thesauri, the similarity of a neighbor with its entry is a more relevant criterion than its rank. Among the rank based methods, we observe

that *RRF* is clearly superior to *condorcet* but only weakly superior to *borda*. Finally, the last row of Table 2 – *CS-w-Mik* – illustrates one step further the interest of ensemble methods for distributional thesauri: whereas the “Mikolov thesaurus” gets the worst results among all the thesauri of Table 2, adding it to the *initial*, *symmetry* and *compound* thesauri in the *CombSum* method leads to improve both R-precision and MAP, with a only small decrease of P@1 and P@5. From a more global perspective, it is interesting to note that our best method, *CombSum*, clearly outperforms the reranking method of (Ferret, 2013) with the same initial starting point.

5 Conclusion and perspectives

In this article, we have presented a method based on bootstrapping for improving distributional thesauri. More precisely, we have proposed a new criterion, based on the relations of mono-terms in similar compounds, for the unsupervised selection of training examples used for reranking the semantic neighbors of a thesaurus. We have evaluated two different strategies for combining this criterion with an already existing one and showed that a late fusion approach based on the merging of lists of neighbors is particularly effective compared to an early fusion approach based on the merging of training sets.

We plan to extend this work by studying how the combination of the unsupervised selection of examples and their use for training supervised classifiers can be exploited for improving distributional thesauri through feature selection. We will also investigated the interest of taking into account word senses in this framework, as in (Huang et al., 2012) or (Reisinger and Mooney, 2010).

References

- Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379.
- Marco Baroni, Georgiana Dinu, and Germán 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)*, pages 238–247, Baltimore, Maryland.
- Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-Based Transformation in Measur-

- ing Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, pages 187–190.
- Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 758–759.
- James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA.
- James Curran. 2002. Ensemble methods for automatic thesaurus extraction. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 222–229.
- Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1162–1172, MIT, Massachusetts, USA.
- Katrin Erk and Sebastian Pado. 2010. Exemplar-based models for word meaning in context. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), short paper*, pages 92–97, Uppsala, Sweden, July.
- Olivier Ferret. 2010. Testing semantic similarity measures for extracting synonyms from a corpus. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
- Olivier Ferret. 2012. Combining bootstrapping and feature selection for improving a distributional thesaurus. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, pages 336–341, Montpellier, France.
- Olivier Ferret. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 561–571, Sofia, Bulgaria.
- Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.
- Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2009. Supervised synonym acquisition using distributional features and syntactic patterns. *Information and Media Technologies*, 4(2):59–83.
- Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 873–882.
- Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 247–256, Uppsala, Sweden.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, pages 768–774, Montreal, Canada.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 746–751, Atlanta, Georgia.
- Jeffrey Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.
- Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, 42(3):595–614.
- Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.
- Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 419–424, Doha, Qatar.
- Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 90–99, Baltimore, Maryland.
- Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 230–238, Gothenburg, Sweden.
- Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May.

- Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010)*, pages 109–117, Los Angeles, California, June.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.
- Shengli Wu, Fabio Crestani, and Yaxin Bi. 2006. Evaluating score normalization methods in data fusion. In *Third Asia Conference on Information Retrieval Technology (AIRS'06)*, pages 642–648. Springer-Verlag.
- Kazuhide Yamamoto and Takeshi Asakura. 2010. Even unassociated features can improve lexical distributional similarity. In *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIC 2010)*, pages 32–39, Beijing, China.
- Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35(3):435–461.