Nothing like Good Old Frequency: Studying Context Filters for Distributional Thesauri

Muntsa Padró,[♣] Marco Idiart[♡], Carlos Ramisch[♢], Aline Villavicencio[♣]
[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
[♢]Aix Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille (France)
muntsa.padro@inf.ufrgs.br, marco.idiart@gmail.com,
carlos.ramisch@lif.univ-mrs.fr, avillavicencio@inf.ufrgs.br

Abstract

Much attention has been given to the impact of informativeness and similarity measures on distributional thesauri. We investigate the effects of context filters on thesaurus quality and propose the use of cooccurrence frequency as a simple and inexpensive criterion. For evaluation, we measure thesaurus agreement with WordNet and performance in answering TOEFL-like questions. Results illustrate the sensitivity of distributional thesauri to filters.

1 Introduction

Large-scale *distributional thesauri* created automatically from corpora (Grefenstette, 1994; Lin, 1998; Weeds et al., 2004; Ferret, 2012) are an inexpensive and fast alternative for representing semantic relatedness between words, when manually constructed resources like WordNet (Fellbaum, 1998) are unavailable or lack coverage. To construct a distributional thesaurus, the (collocational or syntactic) contexts in which a target word occurs are used as the basis for calculating its similarity with other words. That is, two words are similar if they share a large proportion of contexts.

Much attention has been devoted to refining thesaurus quality, improving informativeness and similarity measures (Lin, 1998; Curran and Moens, 2002; Ferret, 2010), identifying and demoting bad neighbors (Ferret, 2013), or using more relevant contexts (Broda et al., 2009; Biemann and Riedl, 2013). For the latter in particular, as words vary in their collocational tendencies, it is difficult to determine how informative a given context is. To remove uninformative and noisy contexts, filters have often been applied like pointwise mutual information (PMI), lexicographer's mutual information (LMI) (Biemann and Riedl, 2013), t-score (Piasecki et al., 2007) and z-score (Broda et al., 2009). However, the selection of a measure and of a threshold value for these filters is generally empirically determined. We argue that these filtering parameters have a great influence on the quality of the generated thesauri.

The goal of this paper is to quantify the impact of context filters on distributional thesauri. We experiment with different filter methods and measures to assess context significance. We propose the use of simple cooccurrence frequency as a filter and show that it leads to better results than more expensive measures such as LMI or PMI. Thus we propose a cheap and effective way of filtering contexts while maintaining quality.

This paper is organized as follows: in §2 we discuss evaluation of distributional thesauri. The methodology adopted in the work and the results are discussed in §3 and §4. We finish with some conclusions and discussion of future work.

2 Related Work

In a nutshell, the standard approach to build a distributional thesaurus consists of: (i) the *extraction of contexts* for the target words from corpora, (ii) the application of an *informativeness measure* to represent these contexts and (iii) the application of a *similarity measure* to compare sets of contexts. The contexts in which a target word appears can be extracted in terms of a window of cooccurring (content) words surrounding the target (Freitag et al., 2005; Ferret, 2012; Erk and Pado, 2010) or in terms of the syntactic dependencies in which the target appears (Lin, 1998; McCarthy et al., 2003; Weeds et al., 2004). The informativeness of each context is calculated using measures like PMI, and t-test while the similarity between contexts is calculated using measures like Lin's (1998), cosine, Jensen-Shannon divergence, Dice or Jaccard.

Evaluation of the quality of distributional thesauri is a well know problem in the area (Lin,

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 419–424,
October 25-29, 2014, Doha, Qatar.

1998; Curran and Moens, 2002). For instance, for intrinsic evaluation, the agreement between thesauri has been examined, looking at the average similarity of a word in the thesauri (Lin, 1998), and at the overlap and rank agreement between the thesauri for target words like nouns (Weeds et al., 2004). Although much attention has been given to the evaluation of various informativeness and similarity measures, a careful assessment of the effects of filtering on the resulting thesauri is also needed. For instance, Biemann and Riedl (2013) found that filtering a subset of contexts based on LMI increased the similarity of a thesaurus with WordNet. In this work, we compare the impact of using different types of filters in terms of thesaurus agreement with WordNet, focusing on a distributional thesaurus of English verbs. We also propose a frequency-based saliency measure to rank and filter contexts and compare it with PMI and LMI.

Extrinsic evaluation of distributional thesauri has been carried out for tasks such as English lexical substitution (McCarthy and Navigli, 2009), phrasal verb compositionality detection (McCarthy et al., 2003) and the WordNet-based synonymy test (WBST) (Freitag et al., 2005). For comparative purposes in this work we adopt the latter.

3 Methodology

We focus on thesauri of *English verbs* constructed from the BNC (Burnard, 2007)[1]. Contexts are extracted from syntactic dependencies generated by RASP (Briscoe et al., 2006), using nouns (heads of NPs) which have subject and direct object relations with the target verb. Thus, each target verb is represented by a set of *triples* containing (i) the verb itself, (ii) a context noun and (iii) a syntactic relation (object, subject). The thesauri were constructed using Lin's (1998) method. Lin's version of the distributional hypothesis states that two words (verbs v_1 and v_2 in our case) are similar if they share a large proportion of contexts weighted by their information content, assessed with PMI (Bansal et al., 2012; Turney, 2013).

In the literature, little attention is paid to context filters. To investigate their impact, we compare two kinds of filters, and before calculating similarity using Lin's measure, we apply them to remove potentially noisy triples:

- **Threshold (th):** we remove triples that occur less than a threshold th. Threshold values vary from 1 to 50 counts per triple.
- **Relevance (p):** we keep only the top p most relevant contexts for each verb, were relevance is defined according to the following measures: (a) frequency, (b) PMI, and (c) LMI (Biemann and Riedl, 2013). Values of p vary between 10 and 1000.

In this work, we want to answer two questions: (a) Do more selective filters improve *intrinsic evaluation* of thesaurus? and (b) Do they also help in *extrinsic evaluation*?

For *intrinsic evaluation*, we determine agreement between a distributional thesaurus and WordNet as the path similarities for the first k distributional neighbors of a verb. A single score is obtained by averaging the similarities of all verbs with their k first neighbors. The higher this score is, the closer the neighbors are to the target in WordNet, and the better the thesaurus. Several values of k were tested and the results showed exactly the same curve shapes for all values, with WordNet similarity decreasing linearly with k. For the remainder of the paper we adopt $k = 10$, as it is widely used in the literature.

For *extrinsic evaluation*, we use the WBST set for verbs (Freitag et al., 2005) with 7,398 questions and an average polysemy of 10.4. The task consists of choosing the most suitable synonym for a word among a set of four options. The thesaurus is used to rank the candidate answers by similarity scores, and select the first one as the correct synonym. As discussed by Freitag et al. (2005), the upper bound reached by English native speakers is 88.4% accuracy, and simple lower bounds are 25% (random choice) and 34.5% (always choosing the most frequent option).

4 Results

Figure 1 shows average WordNet similarities for thesauri built filtering by frequency threshold th and by p most frequent contexts. Table 1 summarizes the parametrization leading to the best WordNet similarity for each kind of filter. In all cases we show the results obtained for different frequency ranges[2] as well as the results when averaging over all verbs.

[1] Even though larger corpora are available, we use a traditional carefully constructed corpus with representative samples of written English to control the quality of the thesaurus.

[2] In order to study the influence of verb frequency on the results, we divide the verbs in three groups: high-frequency

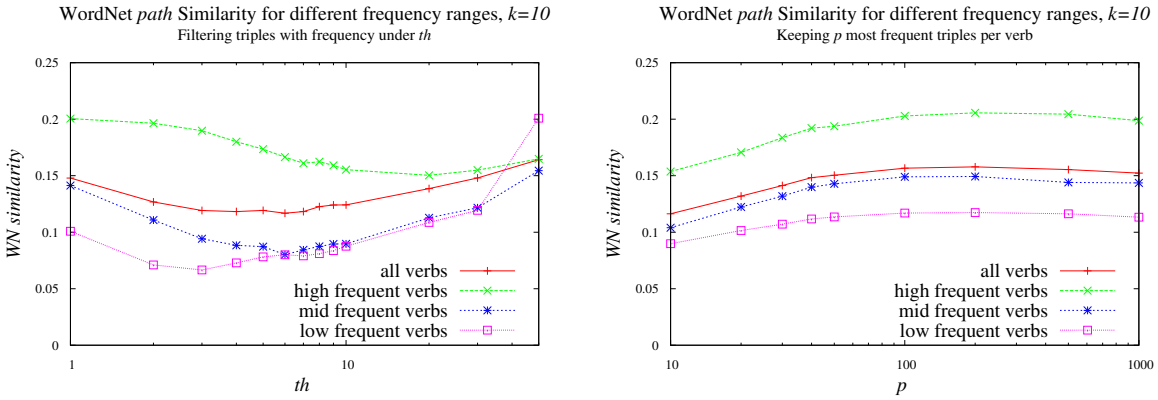

Figure 1: WordNet scores for verb frequency ranges, filtering by frequency threshold th (left) and p most frequent contexts (right).

Filter	All verbs		Frequency range Low		Mid		High	
No filter	-	0.148	-	0.101	-	0.144	-	0.198
Filter low freq. contexts	$th = 50$	**0.164**	$th = 50$	0.202	$th = 50$	0.154	$th = 1$	0.200
Keep p contexts (freq.)	$p = 200$	**0.158**	$p = 500$	0.138	$p = 200$	0.149	$p = 200$	0.206
Keep p contexts (PMI)	$p = 1000$	0.139	$p = 1000$	0.101	$p = 1000$	0.136	$p = 1000$	0.181
Keep p contexts (LMI)	$p = 200$	0.155	$p = 100$	0.112	$p = 200$	0.147	$p = 200$	0.208

Table 1: Best scores obtained for each filter for all verbs and frequency ranges. Scores are given in terms of WordNet path. Confidence interval is arround $\pm$ 0.002 in all cases.

When using a threshold filter (Figure 1 left), high values lead to better performance for mid- and low-frequency verbs. This is because, for high th values, there are few low and mid-frequency verbs left, since a verb that occurs less has less chances to be seen often in the same context. The similarity for verbs with no contexts over the frequency threshold cannot be assessed and as a consequence those verbs are not included in the final thesaurus. As Figure 2 shows, the number of verbs decreases much faster for low and mid frequency verbs when th increases.[3] For example, for $th = 50$, there are only 7 remaining low-frequency verbs in the thesaurus and these tend to be idiosyncratic multiword expressions. One example is *wreak*, and the only triple containing this verb that appeared more than 50 times is *wreak havoc* (71 occurrences). The neighbors of this verb are *cause* and *play*, which yield a good similarity score in WordNet. Therefore, although higher thresholds result in higher similarities for low and mid-frequency verbs, this comes at a cost, as the number of verbs included in the thesaurus decreases considerably.

($||v|| \geq 500$), mid-frequency ($150 \leq ||v|| < 500$) and low-frequency ($||v|| < 150$).

[3]For p most salient contexts, the number of verbs does not vary and is the same shown in Figure 2 for $th = 1$ (no filter).

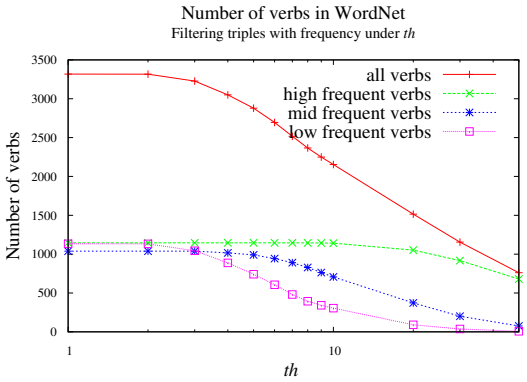

Figure 2: Number of verbs per frequency ranges when filtering by context frequency threshold th

As expected, the best performance is obtained for high-frequency verbs and no filter, since it results in more context information per verb. Increasing th decreases similarity due to the removal of some of these contexts. In average, higher th values lead to better overall similarity among the frequency ranges (from 0.148 with $th = 1$ to 0.164 with $th = 50$). The higher the threshold, the more high-frequency verbs will prevail in the thesauri, for which the WordNet path similarities are higher.

On the other hand, when adopting a relevance

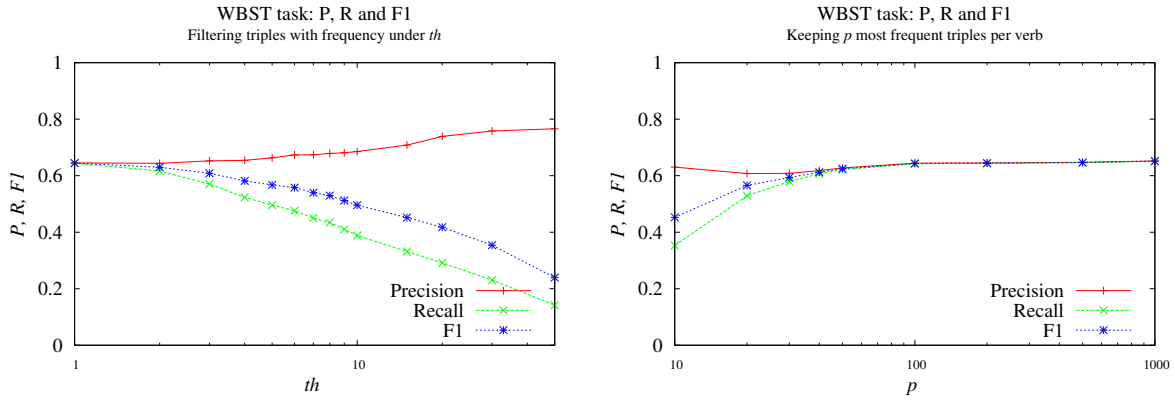

Figure 3: WBST task scores filtering by frequency threshold th (left) and p most frequent contexts (right).

filter of keeping the p most relevant contexts for each verb (Figure 1 right), we obtain similar results, but more stable thesauri. The number of verbs remains constant, since we keep a fixed number of contexts for each verb and verbs are not removed when the threshold is modified. WordNet similarity increases as more contexts are taken into account, for all frequency ranges. There is a maximum around $p = 200$, though larger values do not lead to a drastic drop in quality. This suggests that the noise introduced by low-frequency contexts is compensated by the increase of informativeness for other contexts. An ideal balance is reached by the lowest possible p that maintains high WordNet similarity, since the lower the p the faster the thesaurus construction.

In terms of saliency measure, when keeping only the p most relevant contexts, sorting them with PMI leads to much worse results than LMI or frequency, as PMI gives too much weight to infrequent combinations. This is consistent with results of Biemann and Riedl (2013). Regarding LMI versus frequency, the results using the latter are slightly better (or with no significant difference, depending on the frequency range). The advantage of using frequency instead of LMI is that it makes the process simpler and faster while leading to equal or better performance in all frequency ranges. Therefore for the extrinsic evaluation using WBST task, we use frequency to select the p most relevant contexts and then compute Lin's similarity using only those contexts.

Figure 3 shows the performance of the thesauri in the WBST task in terms of precision, recall and F1.[4] For precision, the best filter is to remove contexts occurring less than th times, but, this also leads to poor recall, since many verbs are left out of the thesauri and their WSBT questions cannot be answered. On the other hand, keeping the most relevant p contexts leads to more stable results and when p is high (right plot), they are similar to those shown in the left plot of Figure 3.

4.1 Discussion

The answer to our questions in Section 3 is *yes*, more selective filters improve intrinsic and extrinsic thesaurus quality. The use of both filtering methods results in thesauri in which the neighbors of target verbs are closer in WordNet and get better scores in TOEFL-like tests. However, the fact that filtering contexts with frequency under th removes verbs in the final thesaurus is a drawback, as highlighted in the extrinsic evaluation on the WBST task.

Furthermore, we demonstrated that competitive results can be obtained keeping only the p most relevant contexts per verb. On the one hand, this method leads to much more stable thesauri, with the same verbs for all values of p. On the other hand, it is important to highlight that the best results to assess the relevance of the contexts are obtained using frequency while more sophisticated filters such as LMI do not improve thesaurus quality. Although an LMI filter is relatively fast compared to dimensionality reduction techniques such as singular value decomposition (Landauer and Dumais, 1997), it is still considerably more expensive than a simple frequency filter.

In short, our experiments indicate that a reason-

[4]Filters based on LMI and PMI were also tested with the same results as intrinsic evaluation: sorting contexts by frequency leads to better results.

able trade-off between noise, coverage and computational efficiency is obtained for $p = 200$ most frequent contexts, as confirmed by intrinsic and extrinsic evaluation. Frequency threshold th is not recommended: it degrades recall because the contexts for many verbs are not frequent enough. This result is useful for extracting distributional thesauri from very large corpora like the UKWaC (Ferraresi et al., 2008) by proposing an alternative that minimizes the required computational resources while efficiently removing a significant amount of noise.

5 Conclusions and Future Work

In this paper we addressed the impact of filters on the quality of distributional thesauri, evaluating a set of standard thesauri and different filtering methods. The results suggest that the use of filters and their parameters greatly affect the thesauri generated. We show that it is better to use a filter that selects the most relevant contexts for a verb than to simply remove rare contexts. Furthermore, the best performance was obtained with the simplest method: frequency was found to be a simple and inexpensive measure of context salience. This is especially important when dealing with large amounts of data, since computing LMI for all contexts would be computationally costly. With our proposal to keep just the p most frequent contexts per verb, a great deal of contexts are cheaply removed and thus the computational power required for assessing similarity is drastically reduced.

As future work, we plan to use these filters to build thesauri from larger corpora. We would like to generalize our findings to other syntactic configurations (e.g. noun-adjective) as well as to other similarity and informativeness measures. For instance, ongoing experiments indicate that the same parameters apply when Lin's similarity is replaced by cosine. Finally, we would like to compare the proposed heuristics with more sophisticated filtering strategies like singular value decomposition (Landauer and Dumais, 1997) and non-negative matrix factorization (Van de Cruys, 2009).

Acknowledgments

We would like to thank the support of projects CAPES/COFECUB 707/11, PNPD 2484/2009, FAPERGS-INRIA 1706-2551/13-7, CNPq 312184/2012-3, 551964/2011-1, 482520/2012-4 and 312077/2012-2.

References

Mohit Bansal, John DeNero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Montréal, Canada, June. Association for Computational Linguistics.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1).

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In James Curran, editor, *Proc. of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sidney, Australia, Jul. ACL.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-based transformation in measuring semantic relatedness. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, Canadian AI '09, pages 187–190, Berlin, Heidelberg. Springer-Verlag.

Lou Burnard. 2007. User Reference Guide for the British National Corpus. Technical report, Oxford University Computing Services, Feb.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proc.of the ACL 2002 Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, Pennsylvania, USA. ACL.

Katrin Erk and Sebastian Pado. 2010. Exemplar-based models for word meaning in context. In *Proc. of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden, Jun. ACL.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press, May. 423 p.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

Olivier Ferret. 2010. Testing semantic similarity measures for extracting synonyms from a corpus. In *Proc. of the Seventh LREC (LREC 2010)*, pages 3338–3343, Valetta, Malta, May. ELRA.

Olivier Ferret. 2012. Combining bootstrapping and feature selection for improving a distributional thesaurus. In *ECAI*, pages 336–341.

Olivier Ferret. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *Proc. of the 51st ACL (Volume 1: Long Papers)*, pages 561–571, Sofia, Bulgaria, Aug. ACL.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In Ido Dagan and Dan Gildea, editors, *Proc. of the Ninth CoNLL (CoNLL-2005)*, pages 25–32, University of Michigan, MI, USA, Jun. ACL.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Springer, Norwell, MA, USA.

Thomas K Landauer and Susan T. Dumais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of the 36th ACL and 17th COLING, Volume 2*, pages 768–774, Montreal, Quebec, Canada, Aug. ACL.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan, Jul. ACL.

Maciej Piasecki, Stanislaw Szpakowicz, and Bartosz Broda. 2007. Automatic selection of heterogeneous syntactic features in semantic similarity of polish nouns. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, pages 99–106, Berlin, Heidelberg. Springer-Verlag.

Peter D. Turney. 2013. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. 1:353–366.

Tim Van de Cruys. 2009. A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 83–90, Athens, Greece, March. Association for Computational Linguistics.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proc. of the 20th COLING (COLING 2004)*, pages 1015–1021, Geneva, Switzerland, Aug. ICCL.