Comparing Similarity Measures for Distributional Thesauri

Muntsa Padró[1], Marco Idiart[2], Aline Villavicencio[1], Carlos Ramisch[3]

[1] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[2] Institute of Physics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[3] Aix Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille, France
muntsa.padro@inf.ufrgs.br, marco.idiart@gmail.com,
alinev@gmail.com, carlos.ramisch@lif.univ-mrs.fr

Abstract

Distributional thesauri have been applied for a variety of tasks involving semantic relatedness. In this paper, we investigate the impact of three parameters: similarity measures, frequency thresholds and association scores. We focus on the robustness and stability of the resulting thesauri, measuring inter-thesaurus agreement when testing different parameter values. The results obtained show that low-frequency thresholds affect thesaurus quality more than similarity measures, with more agreement found for increasing thresholds. These results indicate the sensitivity of distributional thesauri to frequency. Nonetheless, the observed differences do not transpose over extrinsic evaluation using TOEFL-like questions. While this may be specific to the task, we argue that a careful examination of the stability of distributional resources prior to application is needed.

Keywords: distributional thesauri, frequency filters, similarity measures

1. Introduction

Distributional thesauri have been used as the basis for representing semantic relatedness between words. Manually constructed thesauri such as WordNet (Fellbaum, 1998) are not available for all domains and languages, or lack the necessary coverage for many applications. Therefore, one of the main advantages of distributional thesauri over standard resources like WordNet is that they provide inexpensive and fast alternatives for automatically creating large scale resources.

Distributional thesauri are based on the distributional hypothesis (Harris, 1954; Firth, 1957), according to which words are characterized by the contexts in which they appear. To construct these thesauri, the contexts in which a target word occurs are extracted from corpora. The frequencies of those co-occurring *target-context* pairs in the corpus play an important role in determining similarity between words, and may affect how neighbor sets vary (Weeds et al., 2004). A variety of distributional measures has been proposed, first to calculate the degree of association between a target word and its contexts, and second between two target words based on their contexts (Grefenstette, 1994; Lin, 1998; Weeds et al., 2004; Ferret, 2012).

Several parameters of the thesaurus construction methodology can influence the quality of the resulting resource, like:

1. the definition of co-occurring context (e.g. document, sliding window, syntactic frame),
2. the association scores between the target word and its contexts (e.g. co-occurrence frequency, pointwise mutual information),
3. the context filtering or dimensionality reduction method used (e.g. thresholds)
4. the similarity measure used to compare context sets (e.g. cosine, Lin)

This paper focuses on the last three parameters. Our goal is to quantify how sensitive the resulting thesauri are to different configurations of *association scores*, *context filters* and *similarity measures*. As for the first item, our target words are *English verbs* and their contexts are represented by the set of nouns that are syntactically related to the verb in a corpus sentence.

Given the Zipfian distribution of word counts in corpora, a large number of pairs will be very infrequent, potentially introducing noise in the resulting thesaurus. A simple and popular solution for this problem is to set a frequency threshold for removing the long tail of low-frequency words (Lin, 1998; McCarthy et al., 2003; Ferret, 2007). This assumes that the remaining pairs have reliable counts, and can be used as basis for deciding how similar two target words are. However, this only works if a thesaurus is robust to threshold settings. That is, small threshold fluctuations should not provoke drastic changes in the resulting neighbor sets. In this paper we assess the impact of different thresholds in thesaurus robustness, in terms of inter-thesaurus agreement.

The quality of thesauri has been evaluated extrinsically through performance in a variety of semantic tasks, like lexical substitution (McCarthy and Navigli, 2009) and TOEFL-like questions (Freitag et al., 2005; Ferret, 2012). In addition to calculating inter-thesaurus agreement, we analyze how the task performance actually reflects the robustness of a thesaurus used in the task. As an upper bound, we compare our results with WordNet-based thesauri built automatically, using semantic distance measures that exploit the WordNet graph (Fellbaum, 1998).

This paper is structured as follows: it starts with a review of related work (§2.), followed by a description of experimental settings used (§3.). We discuss the results obtained (§4.) and finish with conclusions and future work (§5.).

2. Related Work

To construct distributional thesauri, the contexts in which a target word appears in a sentence or document can be defined in terms of a window of co-occurring (content) words surrounding the target (Freitag et al., 2005; Ferret, 2012) or in terms of the syntactic dependencies in which the target appears (Lin, 1998; McCarthy et al., 2003; Weeds et al., 2004). Baroni and Lenci (2010) propose a unified context model, showing that the appropriate context definition is usually dependent on the task, and task-specific settings are required for obtaining state-of-the-art performance.

Regardless of the context definition, counts for these co-occurring pairs are collected from corpora. The result is a vector for each target word containing its counts with collocated contexts, and the strength of association between word and contexts is calculated with an *association score* or *weight function* (Curran and Moens, 2002) like Pointwise Mutual Information (PMI), χ^2 and t-score. Having a weighted context vector for each target word, a *similarity measure* or *measure function* (Curran and Moens, 2002) calculates similarity, distance or divergence between target words. Similarity measures like cosine, Jensen-Shannon, Dice or Jaccard can be used to rank the set of potential neighbors (all other target words). Target word pairs that have the highest similarity values (lowest distance) for the similarity measure are assumed to be semantically related. Thus the resulting *distributional thesaurus* is a list that associates, to each target word, a list of semantically related neighbors, ranked by decreasing similarity. When the list of target words and corresponding neighbors are identical (as in our case), the thesaurus can be seen as a symmetric matrix, where both rows and columns are the target words and cell values indicate the similarity between words.

Evaluation of the quality of automatically generated large-scale thesauri is a well know problem in the area, and both intrinsic and extrinsic evaluation setups have been adopted. For the former, Lin (1998) analyses the the agreement between thesauri produced by different similarity measures, looking at the average and standard deviation of the similarity of word pairs in different thesauri. Weeds et al. (2004) also advocate for a careful analysis of the properties of the sets of neighbors proposed by different measures. They adapt Lin's (1998) evaluation methodology to calculate the extent to which the set of neighbors of a target word in two thesauri overlap and whether they are in the same order, looking at the neighbor sets for 2000 nouns. Curran and Moens (2002) compare several weight functions (e.g. PMI, t-score, χ^2) and measure functions (e.g. Dice, cosine, Jaccard, Lin). They consider the first 200 neighbors for a set of 70 nouns randomly selected from WordNet in order to cover a wide range of word frequencies, number of senses, specificity and concreteness. They measure precision of the top neighbors, inverse ranking of synonyms and direct matches based on a gold standard consisting of the combination of 3 manually-built thesauri (Macquarie, Roget's and Moby).

Extrinsic evaluation of distributional thesauri have been carried out, for instance, using the WordNet-Based Synonymy Test (WBST), an extended TOEFL-like test automatically generated from WordNet (Freitag et al., 2005). It contains 7,398 verb questions out of a total of 23,570 test questions, with average polysemy of 10.4. The best performance for the verb questions was 63.8%, using a 1 billion word corpus. Ferret (2007) used the same data set for evaluating the identification of synonyms for nouns, where the best results were for high-frequency words and decreased for frequencies lower than 100. Ferret (2007) suggests that the ability of these approaches to capture semantic relatedness seems to be closely correlated with the frequency of these words in the corpus.

Our evaluation follows this line of research by performing an in-depth evaluation of distributional thesauri. However, we are also interested in robustness in terms of different association scores, similarity measures and frequency thresholds. Robustness is estimated by measuring agreement between thesauri and by a task-based extrinsic evaluation.

3. Methodology

The main goal of this work is to compare distributional thesauri generated with different methods and parameters. Therefore, we create two kinds of thesauri: corpus-based distributional thesauri and WordNet based thesauri. The former are the target of our evaluation. The later are used to compare the obtained results with an upper bound, obtained from a manually constructed resource. Once the thesauri are created, we compare them using metrics for rank comparison.

3.1. WordNet-Based Thesauri

To use as upper bound reference for this work, we built WordNet thesauri (Fellbaum, 1998) containing only verbs that are also in the distributional thesauri, using two similarity measures [1]:

- s_{wn-lch}: Leacock Chodorow Similarity (1998). Computes how similar two word senses are based on the shortest path that connects the senses in the taxonomy and the maximum depth of the taxonomy in which the senses occur.
- s_{wn-wup}: Wu-Palmer Similarity (1994). Estimates similarity based on the depth of the two senses in the taxonomy and that of their least common subsumer.

As WordNet metrics are computed over sense (synset) pairs rather than over word pairs, for a polysemous word we adopt the maximum of the similarities calculated for each of its possible sense pairs.

3.2. Distributional Thesauri

The thesauri were constructed from the syntactic dependencies involving verbs in the RASP-parsed (Briscoe et al., 2006) British National Corpus (BNC) (Burnard, 2000), using a threshold of 50 occurrences to discard low-frequency verbs, and removing all relations involving pronouns. We use as a starting point the method proposed by Lin (1998), which calculates the similarity between two words on the basis of the dependencies that they share.

[1] Implemented in NLTK toolkit (Bird et al., 2009). As Path Distance Similarity (Rada et al., 1989) produced a similar thesaurus to s_{wn-lch}, we only discuss the latter.

A *dependency triple* (v, r, n) is the combination of verb v and noun n, with the syntactic relation r in a sentence. The number of occurrences of a dependency triple in a corpus is represented by $||v, r, n||$. For instance the sentence *the pilot drove the car* generates two triples (v="drive", r=dobj, n="car") and (v="drive", r= ncsubj, n="pilot"). As described in Section 4., we filter out triples whose number of occurrences is below a given frequency threshold. We can estimate the probability of a noun n appearing for a given verb-relation pair as

$$p(n|v, r) \simeq \frac{||v, r, n||}{||v, r, *||} \quad (1)$$

where $*$ indicates a sum over all possible values of that variable, or

$$||v, r, *|| = \sum_{n_i} ||v, r, n_i||$$

Following the distributional hypothesis, we could posit that the similarity between two different verbs is a measure of the closeness of their noun distributions. However, a possible problem of this method is that these distributions tend to be dominated by very frequent words that in general are polysemic and may combine with many verbs. Lin (1998) proposes that what has to be compared is not the relative frequency of the words but the information content of the triple measured by Pointwise Mutual Information (PMI), which is defined by

$$\begin{aligned} I(v, r, n) &= \log \frac{p(v, n|r)}{p(v|r)p(n|r)} \\ &\simeq \log \frac{||v, r, n|| \cdot ||*, r, *||}{||v, r, *|| \cdot ||*, r, n||} \end{aligned} \quad (2)$$

PMI indicates how the the frequency of v and n observed together departs from random chance, for a given relation r. As a consequence, it eliminates spurious high correlation due to very frequent words. Therefore Lin's version of the distributional hypothesis states that two words (verbs in our case) are similar if they have similar information content for all pairs (r, n). However, PMI is asymmetric with respect to association: perfect correlation has an upper bound of $-\log p$ where p is the probability of the most frequent word, but anti-correlation is negative and unbounded, reaching $-\infty$ for perfect anti-correlation. To avoid attributing excessive importance for anti-correlated triples, Lin assumes that only positive PMIs should be compared, and proposes the following similarity measure

$$s_{lin}(v_1, v_2) = \frac{\sum_{(r,n)\in\Omega} I(v_1, r, n) + I(v_2, r, n)}{\sum_{(r,n)} I_+(v_1, r, n) + I_+(v_2, r, n)} \quad (3)$$

where $I_+(v, r, n) = I(v, r, n)$ if $I(v, r, n) > 0$ and zero otherwise, and Ω is the set of all pairs (r, n) such both $I(v_1, r, n)$ and $I(v_2, r, n)$ are positive. Since anti-correlations can be as informative as correlations, a reasonable extension of this idea is, instead of using PMI, to use its normalized version, nPMI (Bouma, 2009).

$$I_n(v, r, n) \simeq \frac{\log \frac{||v,r,n|| \cdot ||*,r,*||}{||v,r,*|| \cdot ||*,r,n||}}{-\log \frac{||v,r,n||}{||*,r,*||}} \quad (4)$$

The score I_n (or nPMI) corresponds to PMI normalized by its maximum value when the words always co-occur, that is, when $p(v|r) = p(n|r) = p(v, n|r)$, PMI $= -\log p(v, n|r)$. Its value is bounded between $[-1, +1]$, resulting in -1 for perfect anti-correlation and $+1$ for complete co-occurrence. For our experiments, we generate three distributional thesauri variants. The first one is Lin's traditional setting, based on positive PMI scores (equation 3). The second one, noted $s_{lin-norm}$, allows us to assess the impact of normalization in Lin's measure, since PMI is replaced by nPMI in the similarity measure. The third one uses a novel cosine similarity measure based on nPMI

$$s_{cosine}(v_1, v_2) = \frac{\sum_{(r,n)} I_n(v_1, r, n) \cdot I_n(v_2, r, n)}{\sqrt{\sum_{(r,n)} I_n(v_1, r, n)^2 \sum_{(r',n')} I_n(v_2, r', n')^2}} \quad (5)$$

that takes into full account correlations and anti-correlations in the triples. For building the verb thesauri, we keep only neighbor verbs whose similarity measure with the target is greater than zero.

3.3. Calculating Inter-Thesaurus Agreement

Given two thesauri built with different settings, we would like to quantify how much they agree on the ranking proposed for each target verb. Each verb in a thesaurus has a list of neighbor verbs ranked by decreasing similarity. We determine thesauri agreement in terms of overlapping elements and their ranks for each verb, where agreement at the top of the ranks is particularly important, given the decreasing similarities and potential noise for positions further away from the top of the ranks. We examine agreement in different sub-ranks of length *k* (i.e. the first *k* elements), using 3 measures: $jaccard$, intersection metric (im) and Kendall τ_b.[2] All inter-thesaurus agreement measures were calculated on a target verb basis, and then averaged for all target verbs in the thesaurus.

3.3.1. Jaccard Index

For determining the degree of overlap between two ranks we use jaccard index (Jaccard, 1901) ($jaccard$), which provides a ratio between their intersection and their union. As we remove neighbors with similarity less than or equal to 0, $jaccard$ will essentially tell us the proportion of neighbors that had a positive similarity with the target in both thesauri. It is a set-based measure that does not take into account the rank of neighbors. That is, we may find perfect overlap even if the neighbor sets contain the same elements ranked in reverse order. $jaccard$ ranges from 0 (empty intersection) to 1 (perfect overlap).

3.3.2. Intersection Metric

Since $jaccard$ does not reflect to what extend the order in the ranks is preserved, we also use Intersection Metric (Fagin et al., 2003) (im). For each i in $1, ..., k$, the overlap proportion of the two ranks up to i is computed and then all

[2] Given that two ranks may contain different verbs (they may not be a permutation of each other), rank comparison must take into account incomplete lists.

overlap proportions for increasing i values are averaged, to result in a single agreement measure.[3]This metric also captures whether rankings agree at the top positions, and if that is the case they have im values close to 1. This is particularly important when evaluating distributional thesauri in which only a few top-i candidates will be used in a semantic task.[4] im ranges from 0 to 1, with 1 when all elements are the same and in the same order.

3.3.3. Kendall τ_b

We also use Kendall τ (Kendall and Gibbons, 1948) to determine the agreement between two neighbor rankings (Voorhees, 1998; Voorhees, 2001; Yilmaz and Aslam, 2006). We adopt the variant proposed by Fagin et al. (2003) for rankings with different elements, and Kendall τ_b to account for ties when several verbs have the same similarity. τ_b ranges from 1 for identical rankings to -1 for inverse rankings, with 0 if they are not correlated.

4. Experiments and Results

A total of 18 distributional thesauri were evaluated, created from 3 similarity measures (s_{lin}, $s_{norm-lin}$, s_{cosine}) and 6 low-frequency thresholds (th) for triples, varying from 1 (all triples) to 50. There are 12 reference WordNet thesauri from 2 measures (s_{wn-lch}, s_{wn-wup}) and the same 6 threshold values.

Table 1 shows the decrease in the size of thesauri with the increase in threshold, both in terms of verbs and of their average number of neighbors[5]. There is a large variation in the number of neighbors per verb, as indicated by the average and the standard deviation of the number of neighbors. This is due, again, to the Zipfian distribution of data, and choosing a fixed threshold value will determine the recall-precision balance of the thesaurus.

Threshold th	# verbs	# neighbors			
		s_{lin}		s_{wn-lch}	
		mean	stdev	mean	stdev
1	9,251	3,640	2,265	769	891
3	5,412	838	870	775	715
5	3,898	525	570	726	592
10	2,559	283	327	625	445
20	1,669	147	179	524	322
50	814	59	72	322	171

Table 1: Thesauri size with varying threshold (th) for s_{lin} and WordNet s_{wn-lch}

For calculating inter-thesaurus agreement, we use $jaccard$, im and Kendall τ_b (§ 4.1.). These measures are computed for each pair of thesauri comparing the ranked list of neighbors for each target verb contained in both thesauri. To verify whether there is more agreement in the beginning of the ranks we also report figures for different rank lengths (k), and for the overlap of these sub-ranks. Furthermore, we also measure the performance of the thesauri in the WBST test set (§ 4.2.) to evaluate them.

4.1. Inter-Thesaurus Agreement

As similarity measures produce different thesauri, we examine (a) how different these thesauri are from each other, (b) whether these differences are related to low-frequency words, and (c) if they are larger than what would be expected if we compared WordNet thesauri. The thesauri built with s_{lin} and $s_{norm-lin}$ show high agreement values for all measures (higher than those obtained comparing the two WordNet thesauri). This indicates that the use of normalized PMI has little impact on a thesaurus constructed with Lin's similarity measure. Thus, our analysis focuses on the comparison of different similarity measures.

A comparison of inter-thesaurus agreement for the s_{lin} and s_{cosine} at increasing thresholds is shown in Figure 1. The upperbound agreement obtained between the two WordNet-based thesauri is shown as a solid line. It contains the verbs in the distributional thesauri for $th = 1$. For s_{lin} and s_{cosine} overlap of neighbors ($jaccard$) increases with the threshold. A threshold of 50 occurrences results in a $jaccard$ agreement between s_{lin} and s_{cosine} that is even higher than that for WordNet thesauri. This is significantly higher than for the lower thresholds, even for the 10 first neighbors.

The agreement between the distributional rankings also increases with threshold (im and τ_b) towards those for WordNet, and with longer ranks (k). When only the overlapping neighbors are evaluated (the two lower left plots in Figure 1) the agreement between thesauri is less sensitive to threshold and rank length.

These results indicate that, if what we want is robust results, independently of the thesaurus, then longer ks need to be adopted (around 100), as well as higher thresholds. However, as discussed in the next section, higher thresholds also mean less coverage for the resulting resource.

As the agreement between distributional thesauri can be almost as similar as that between WordNet thesauri, we would like to know how much a distributional thesaurus agrees with WordNet. Figure 2 addresses this question. The low agreement between s_{lin} with s_{wn-lch} (representative of all comparisons) may be explained by differences in coverage with the subset of WordNet used and its impact in $jaccard$. The last 2 plots for the overlap between the thesauri show that the distributional thesaurus agrees much less with WordNet than with another distributional thesauri.

4.2. Extrinsic Evaluation

Given the variation in thesaurus content, the next question is whether these differences have an impact on the application of a thesaurus for a given task. For the extrinsic evaluation, we use the WBST set for verbs (Freitag et al., 2005) which contains 7,398 questions. The task consists of choosing the most suitable synonym for a word among a set of four options, i.e. ranking the answers using their similarity scores with the question word, and selecting the top answer as synonym. To assess the difficulty of the task,

[3]The overlap proportion at i is a modified version of $jaccard$, computed as the number of overlapping elements up to the i-th rank, divided by i.

[4]In McCarthy et al. (2003), for example, only the first 50 neighbors are considered.

[5]Only the values for s_{lin} and s_{wn-lch} are reported since those for the other distributional and WordNet thesauri, respectively, are similar.

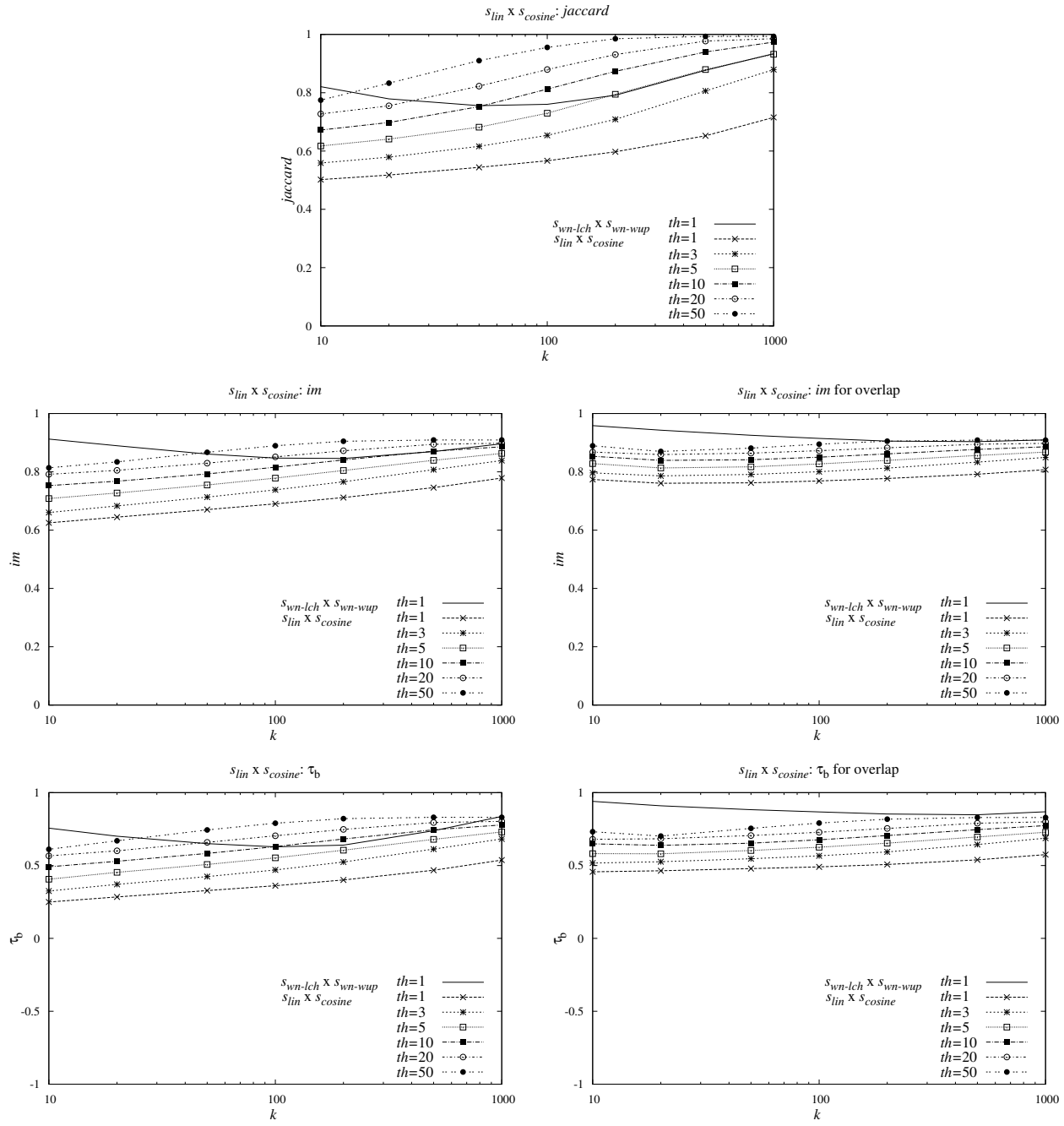

Figure 1: $jaccard$, im, and τ_b values for s_{lin} vs s_{cosine} thesauri, along with their overlap, and WordNet thesauri as upperbound reference

100 of these questions from each PoS were answered by 6 native speakers with average performance of 88.4% and one non-native speaker with 80.3%, as discussed by Freitag et al. (2005). The baseline performance was 25% (without any semantic information), or 35.2% always choosing the most frequent option (34.5% for verbs).

Table 2 shows the results obtained with s_{lin}[6], s_{cosine} and s_{wn-lch} similarities to rank the candidates in two conditions. In the first, the *strict condition* only considers questions for which the target and four candidates were in the thesauri, while the second, the *flexible condition* uses all questions for which at least one candidate is in the thesauri, and assumes that absent candidates are at the end of the synonym rank for the target word. The high accuracy of WordNet thesauri was expected, as WordNet was used to generate WBST. The results for the distributional thesauri are compatible with those obtained by Freitag et al. (2005) (63.8%), and improve for higher thresholds, in the flexible condition. However, the number of verb types decreases almost 10 fold as the threshold goes from 1 to 50, and the verbs that are left are the most frequent ones. Therefore, the increase in performance may just indicate that the test becomes less challenging with the shift in the frequency profile of the verbs.

[6]As expected the results obtained with $s_{norm-lin}$ and s_{lin} were very similar and only the latter is discussed.

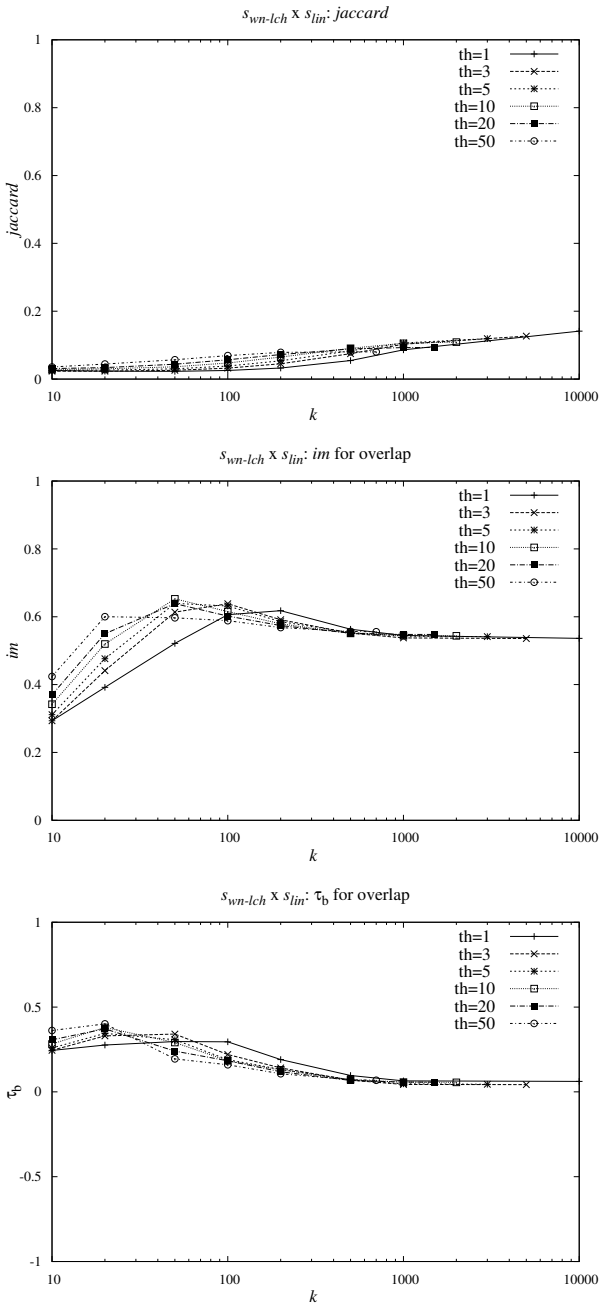

Figure 2: $jaccard$, im, and τ_b values when comparing s_{lin} and s_{wn-lch} similarities

There is no statistically significant difference between the performances of the thesauri built using different similarity measures. This indicates that, even though the thesauri are different, this is not reflected in this particular task. Indeed, if we look at the top-10 neighbors for three example words of different frequency ranges, the variation in their ranks is clear (Table 3). This unstable behavior may be explained because the similarity values have a flat distribution, and adjacent neighbors have very close similarity values (Figure 3). Moreover, this instability may be stronger for particular frequency ranges. For example, in Table 3, the low-frequency word *rush* has more changes in neighbors and in ranks than the other words, with less agreement across thresholds than across similarity measures. A more detailed examination of frequency profiles is planned as future work.

Thres.	Strict condition	Flexible condition
WordNet s_{wn-lch}		
1	0.990 ± 0.006	0.978 ± 0.002
s_{lin}		
1	0.65 ± 0.01	0.648 ± 0.009
3	0.62 ± 0.03	0.65 ± 0.01
5	0.61 ± 0.06	0.66 ± 0.01
10	0.6 ± 0.2	0.69 ± 0.01
20	–	0.74 ± 0.01
50	–	0.76 ± 0.02
s_{cosine}		
1	0.67 ± 0.01	0.660 ± 0.009
3	0.64 ± 0.04	0.66 ± 0.01
5	0.59 ± 0.07	0.67 ± 0.01
10	0.6 ± 0.2	0.69 ± 0.01
20	–	0.74 ± 0.01
50	–	0.76 ± 0.02

Table 2: Results in WBST task, strict and flexible conditions, s_{lin} vs s_{cosine}.

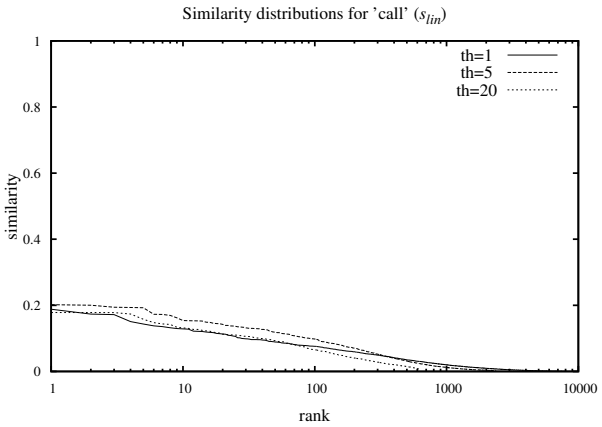

Figure 3: Similarity profile for the neighbors for *call*, a high frequency verb in s_{lin}.

5. Conclusions and Future Work

Distributional thesauri can be invaluable tools to access the meaning of words. This is particularly interesting since they can in principle be generated from a corpus with minimal human intervention. However there is no consensus on how to generate a thesaurus: as different proposals in the literature address a variety of tasks, it transpires that distinct approaches should be used in each of these circumstances, which is far from an ideal situation. Another important concern is how robust a thesaurus is with respect to changes in the corpus and in the settings used to build it.

In this paper, we studied in depth a set of thesauri: Lin's (1998) original proposal and a cosine measure with normalized pointwise mutual information. We looked at the impact of low-frequency thresholds, and examined the agreement between the resulting thesauri. We also compared the agreement between WordNet-based thesauri as an upper bound reference.

s_{lin}					
begin		drop		rush	
th=3	*th=5*	*th=3*	*th=5*	*th=3*	*th=5*
start	start	pick up	pick up	hurry	hurry
continue	continue	lift	throw	speed	trickle
end	end	place	push	trickle	splash
complete	complete	fall	place	flow	filter
come	seem	throw	slip	come in	gush
seem	conduct	put	put	lap	pollute
go	launch	lower	grab	gush	drain
stop	come	rise	fall	swirl	race
follow	follow	put down	lower	surge	bubble
take	go	carry	rise	comfort	sprinkle
s_{cosine}					
begin		drop		rush	
th=3	*th=5*	*th=3*	*th=5*	*th=3*	*th=5*
start	start	pick up	pick up	gush	gush
continue	continue	lower	detonate	trickle	splash
end	end	lift	soar	flow	trickle
resume	resume	put down	lower	surge	hurry
commence	conduct	toss	throw	lap	pulse
conduct	complete	throw	slip	flow down	congeal
complete	launch	place	bounce	congeal	spurt
undertake	commence	fall	fluctuate	wend	cake
launch	initiate	slip	grab	pump	meander
initiate	undertake	rise	hold up	flow out	bubble

Table 3: Top 10 neighbors for high, medium and low frequency target verbs.

The results show that increasing threshold values also increase agreement between thesauri, at the expense of coverage. Although the agreement of distributional thesauri can almost reach the same levels as those between WordNet thesauri, they differ somewhat for the smaller rank lengths: while the latter have the expected higher agreement for the top neighbors, the former have more agreement for larger rank lengths (around 100 to 200 neighbors). This indicates that tasks that use a small number of top-k neighbors in a distributional thesaurus are very sensitive to the choice of the thesaurus.

We also use normalized PMI (Bouma, 2009) as a weight function instead of co-occurrence counts for calculating the cosine similarity, taking into account negative associations. We compared Lin's thesauri with PMI and nPMI to evaluate the effect of normalization as a variable and concluded that the resulting thesauri were very similar.

The largest differences were found between distributional and WordNet thesauri, which suggests that measures involving explicit ranks of neighbors are very different when performed in a distributional or in a WordNet thesaurus.

Finally, the equivalent performances of distributional thesauri in a TOEFL-like test do not seem to reflect the observed differences. This is partly an effect of the type of test, as it consists of searching for specific words in the ranks without limiting the number of neighbors. For future work, we plan to extend this analysis taking into account specific frequency profiles of the target and candidate words.

Acknowledgements

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

6. References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Comp. Ling.*, 36(4):673–721.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Leeuwarden, The Netherlands.

E. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia.

L. Burnard. 2000. Users reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '03, pages 28–36, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic*

Lexical Database (Language, Speech, and Communication). The MIT Press, illustrated edition edition, May.

Olivier Ferret. 2007. Finding document topics for improving topic segmentation. In *Proc. of the 45th ACL (ACL 2007)*, pages 480–487, Prague, Czech Republic, Jul. ACL.

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

John R. Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

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

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Maurice Kendall and Jean D. Gibbons. 1948. *Rank Correlation Methods*. Charles Grifn, 1st edition.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellfaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774. Association for Computational Linguistics.

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, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. In *IEEE Transactions on Systems, Man and Cybernetics*, pages 17–30.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 315–323, New York, NY, USA. ACM.

Ellen Voorhees. 2001. Evaluation by highly relevant documents. In *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82. ACM Press.

Julie Weeds, David J. Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *COLING*, Barcelona, Spain, Jul. ACL.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emine Yilmaz and Javed A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 102–111, New York, NY, USA. ACM.