Relieving Polysemy Problem for Synonymy Detection

Gaël Dias and Rumen Moraliyski

University of Beira Interior, Covilhã 6201-001, Portugal
ddg@di.ubi.pt, rumen@hultig.di.ubi.pt

Abstract. In order to automatically identify noun synonyms, we propose a new idea which opposes classical polysemous representations of words to monosemous representations based on the "*one sense per discourse*" hypothesis. For that purpose, we apply the attributional similarity paradigm on two levels: corpus and document. We evaluate our methodology on well-known standard multiple choice synonymy question tests and evidence that it steadily outperforms the baseline.

1 Introduction

Identifying and extracting synonyms is a crucial issue for NLP systems. Indeed, synonyms are widely used, even in narrow domains, to refer to the same concept thus avoiding repetition.

Discovering close synonyms is a difficult task. Some methods detect broad range of semantic relatedness including but not limited to synonymy such as in [1]. Other methods make use of syntactic contexts which tighten the set of detected relations but they tend to correctly solve less polysemous cases as shown in [2]. Due to the fact that only about 25% of the words are polysemous, this kind of methods solve the problem for the most part of the vocabulary. Nevertheless, most frequent words are usually highly polysemous and represent a large proportion of the texts [3]. As a consequence, there is a critical need for methods capable of separating the different meanings of a polysemous word in distinct representations.

In this paper, we propose a methodology to measure syntactic oriented attributional similarity based on the "*one sense per discourse*" hypothesis [4]. Instead of relying exclusively on corpus distributions, we build noun representations and compare them within document limits. This paper presents extended evaluation and analysis of our previous work [5].

In order to test this assumption, we implement the vector space model based on the cosine similarity measure over Term Frequency weighted by Inverse Document Frequency, Pointwise Mutual Information [6] and Conditional Probability [7]. We also implement two probabilistic similarity measures: the Ehlert model [8] and the Lin model [9]. Finally, we evaluate our methodology on a *noun subset* of well-known standard multiple choice synonymy question tests and evidence that it steadily outperforms the baseline.

2 Related Work

Most related research works use multiple choice synonymy questions for evaluation. Each question consists of five words: the target word and four response words, one of

which is the correct answer and the other ones are called decoys. In this context, the problem aims at defining a function that highlights the best synonym candidate.
One of the most famous works is proposed in [1] where document distribution is used to measure word similarity. They show that the accuracy of Latent Semantic Analysis (LSA) is statistically indistinguishable from that of a population of non-native English speakers on the same questions.
More recent works have focused on the window based vector space model (VSM). A context vector built on co-occurrence basis within the entire corpus is associated to each word from the multiple choice test. For instance, in [6], a variety of similarity metrics and weighting schemes of contexts are studied and their DR-PMI achieves a statistical tie compared to the PMI-IR proposed by [10].
The PMI-IR is one of the first works to propose a hybrid approach to deal with synonym detection. Indeed, it uses a combination of evidences such as the Pointwise Mutual Information (PMI) and Information Retrieval (IR) features like the "NEAR" and "NOT" operators of a search engine to measure similarity between pairs of words. At this point, it is important to notice that this work does not follow the attributional similarity paradigm but rather proposes a heuristic to measure semantic distances. Later, [11] refined the PMI-IR algorithm and proposed a module combination to include new features such as LSA and thesaurus evidences. However, the introduction of thesaurus features biases radically the test of synonym detection.
In parallel, some works have used linguistic resources to measure similarity. Results for a number of relatively sophisticated thesaurus-based methods which look at path lengths between words in the heading classifications of Roget's Thesaurus are given in [12]. However, this methodology does not follow the attributional similarity paradigm.
In the syntactic attributional similarity paradigm, word context vectors associated to all target words of the test are indexed by the words they co-occur with within a given corpus for a given syntactic relation. For example, *(good, adjective)* and *(have, direct-obj)* are attributes of the noun "*idea*" as illustrated in [13]. Unfortunately, to our knowledge, unlike window based approaches, syntactic based methodologies have not been evaluated against synonymy test. Rather, they have been used to build linguistic resources.
All the systems reviewed so far share the difficulty to deal with the polysemous words since they merge all the senses of a word in a single statistical representation. The importance of sense information for the purpose of semantic relation discovery is illustrated by the attempt of [14] at word sense induction (WSI). For a word w he looks for pairs of strongly associated to w words such that their context vectors sum up to the one of w and at the same time are as different as possible. This pair of words and their context vectors describe two distinct senses of w. Thus on the TOEFL task vectors pertaining to specific senses are compared and the method achieves 92.5% accuracy.
In order to summarize the most significant works proposed so far, in Table 1 we present the different results over the TOEFL question set.

3 Attributional Similarity

Theoretically, an attributional similarity measure can be defined as follows. Suppose that $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip})$ is a row vector of observations on p variables

Work	Best result
Landauer and Dumais 1997	64.40%
Sahlgren 2001	72.00%
Turney 2001	73.75%
Jarmasz and Szpakowicz 2003	78.75%
Terra and Clarke 2003	81.25%
Elhert 2003	82.00%
Freitag et al. 2005	84.20%
Rapp 2003	92.5%
Turney et al. 2003	97.50%

Table 1. Accuracy on TOEFL question set.

(or attributes) associated with a label i, the similarity between two units i and j is defined as $S_{ij} = f(X_i, X_j)$ where f is some function of the observed values. In our context, we must evaluate the similarity between two nouns which are represented by their respective word context vectors.

For our purpose, the attributional representation of a noun consists of tuples $\langle v, r\rangle$ where r is an object or subject relation and v is a given verb appearing within this relation with the target noun. For example, if the noun "*brass*" appears with the verb "*cast*" within a object relation, we will have the following triple $\langle$*brass, cast, object*$\rangle$ and the tuple $\langle$*cast, object*$\rangle$ will be an attribute of the word context[1] vector associated to the noun "*brass*".

As similarity measures are based on real-value attributes, our task is two-fold. First, we must define a function which will evaluate the importance of a given attribute $\langle v,\ r\rangle$ for a given noun. Our second goal is to find the appropriate function f that will accurately evaluate the similarity between two verb context vectors.

3.1 Weighting Attributes

In order to construct more precise representations of word meanings, numerous weighting schemas have been proposed. In this section, we will point at the most common ones although many others could be used.

Word Frequency and IDF The simplest form of the vector space model treats a noun n as a vector which attribute values are the number of occurrences of each tuple $\langle v, r\rangle$ associated to n i.e. $tf(n, \langle v, r\rangle)$. However, the usual form of the vector space model introduces the inverse document frequency defined in the context of syntactic attribute similarity paradigm in Equation 1 where n is the target noun, $\langle v, r\rangle$ a given attribute, N is the set of all the nouns and $|.|$ the cardinal function.

$$tf.idf(n,\langle v,r\rangle) = tf(n,\langle v,r\rangle) \times \log_2 \frac{|N|}{|\{n_i \in N | \exists (n_i, v, r)\}|} \quad (1)$$

[1] From now on, we will talk about verb context vectors instead of word context vectors.

Pointwise Mutual Information The value of each attribute $\langle v, r \rangle$ can also be seen as a measure of association with the noun being characterized. For that purpose, [6,10] have proposed to use the Pointwise Mutual Information (PMI) as defined in Equation 2 where n is the target noun and $\langle v, r \rangle$ a given attribute.

$$PMI(\langle n|r \rangle, \langle v|r \rangle) = \log_2 \frac{P(n,v|r)}{P(n|r)P(v|r)} \tag{2}$$

Conditional Probability Another way to look at the relation between a noun n and a tuple $\langle v, r \rangle$ is to estimate their conditional probability of co-occurrence as in Equation 3. In our case, we are interested in knowing how strongly a given attribute $\langle v, r \rangle$ may evoke the noun n.

$$P(n|v,r) = \frac{P(n,v,r)}{P(v,r)} \tag{3}$$

3.2 Similarity Measures

There exist many similarity measures in the context of the attributional similarity paradigm [7]. They can be divided into two main groups: (1) metrics in a multi-dimensional space also called vector space model, (2) measures which calculate the correlations between probability distributions.

Vector Space Model To quantify similarity between two words, the Cosine similarity measure is usually applied and estimates to what extent two vectors point along the same direction. It is defined in Equation 4.

$$cos(n_1, n_2) = \frac{\sum_{k=1}^{p} n_{1k} n_{2k}}{\sqrt{\sum_{k=1}^{p} {n_{1k}}^2} \sqrt{\sum_{k=1}^{p} {n_{2k}}^2}} \tag{4}$$

Probabilistic Models Probabilistic measures can be applied to evaluate the similarity between words when they are represented by a probabilistic distribution. In this paper, we present two different measures i.e. the Ehlert and the Lin models.
Ehlert model: Equation5 proposed in [8] evaluates the probability to interchange two word context vectors (i.e. what is the probability that the first word is changed for the second one).

$$Ehl(n_1|n_2) = \sum_{\langle v,r \rangle \in A} \frac{P(n_1|v,r)P(n_2|v,r)P(v,r)}{P(n_2)} \tag{5}$$

with $A = \{\langle v,r \rangle | \exists (n_1,v,r) \wedge \langle v,r \rangle | \exists (n_2,v,r)\}$.

Lin model: [9] defines similarity as the ratio between the amount of information needed to state the commonality of two words and the total information available about them and is defined in Equation 6.

$$Lin(n_1, n_2) = \frac{2 \times \sum_{\langle v,r \rangle \in A} \log_2 P(v,r)}{\sum_{\langle v,r \rangle \in B} \log_2 P(v,r) + \sum_{\langle v,r \rangle \in C} \log_2 P(v,r)} \tag{6}$$

with
A={$\langle v,r\rangle|\exists(n_1,v,r)\wedge\langle v,r\rangle|\exists(n_2,v,r)$},
B={$\langle v,r\rangle|\exists(n_1,v,r)$},
C={$\langle v,r\rangle|\exists(n_2,v,r)$}.

3.3 Global and Local Attributional Similarity

The approaches reviewed earlier which build context attributional representations of words do so from a corpus as one huge text and do not respect the document limits. We call *Global similarities* ($Gsim$), the similarity estimations obtained in this manner.
However, this approach poses many problems for polysemous nouns as contexts which are pertinent to different meanings are gathered into a single global representation when they should be differentiated. In this context, [2] attempt to introduce a measure of difficulty of tests based on polysemy. They automatically build a number of test cases by taking two words from a synset of WordNet [15] and three randomly other words for decoys. As a result, they find strong positive correlation between polysemy and error level. However, they do not take into account the decoys. They divide their tests with respect to the sum of polysemy of the target word and the correct answer and observe that the more polysemous the test is, the more difficult it is to be solved. The conclusion is that polysemy level is characteristic of the difficulty of the test.
According to [4] "*... if a polysemous word such as 'sentence' appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense*". From this assumption follows that if a word representation is built out of single discourse evidences it probably describes just one sense of the word. Hence if we obey document borders we can avoid mixing all word senses together.
On the other hand Turney [10] demonstrates that synonyms tend to co-occur in texts more often than by chance. On a similar supposition is grounded [1] which seek for synonyms among words that co-occur in the same set of documents.
Thus, our proposal to apply the "*one sense per discourse*" paradigm takes advantage of the fact that people do not tend to repeat words; yet they repeat ideas. As a consequence, we compare attributional representations of nouns only within document's limits. Apparently, statistics gathered from a unique short text may not be reliable. In order to obtain more stable results, we average attributional similarity values over the set of documents in which both nouns occur and introduce the $Lsim(.,.)$ function in Equation 7, where $sim(.,.)$ is any function from Section 3.2. We call this value *Local similarity* ($Lsim$).

$$Lsim(n_1,n_2)=\frac{\sum_{d\in D} sim(n_1,n_2)}{|D|} \tag{7}$$

As a result, nouns that co-occur in a document but with different meanings will rarely share contexts and will end with low similarity. On the other hand, nouns that co-occur as synonyms will share contexts with greater probability hence will receive higher similarity estimations.
Rapp [14] observed that when multiple first order associates taken as a vector are used as a sense descriptor only the most frequent sense of the word tends to be reflected. Therefor he uses as sense descriptors the context vectors of set of strong first order associate words and ensures they specify as narrow as possible senses. On the other hand

Local similarity relies on association vector that is monosemous and directly related to the meaning described. It can be seen as extension to the method of Rapp as he found that the multidimensional descriptor is more stable with respect to sampling errors.
In this paper, we also propose that Global and Local approaches may have properties that complement each other. In order to take advantage of both heuristics, we propose the *Product similarity* ($Psim$) measure, a multiplicative combination of both Local and Global similarities as defined in Equation 8.

$$Psim(n_1,n_2)=Gsim(n_1,n_2)^{\alpha} \times Lsim(n_1,n_2)^{(1-\alpha)} \tag{8}$$

In fact, Equation 8 is a generalization of all similarity measures. When $\alpha = 0$ only the Local similarity is taken into account, while for $\alpha = 1$ only the Global similarity is applied. We will see in section 5 that the combination of both similarity measures provides improved results in particular situations.

4 Corpus

Any work based on the attributional similarity paradigm depends on the corpus used to determine the attributes and to calculate their values. [6] use a terabyte of web data that contains 53 billion words and 77 million documents, [16] - a 10 million words balanced corpus with a vocabulary of 94 thousand words and [2,8] - the 256 million words North American News Corpus (NANC). For our experiments, we used the Reuters Corpus Volume I (RCV1) [17]. However, our proposal needs co-occurrences of both synonym candidates to appear a few times each within a single document and we observed that substantial proportion of word pairs have zero occurrence in RCV1. As we did not want to reduce our test set, we decided to build a corpus suitable to the problem at hand.
To build the Web Corpus for Synonym Detection (WCSD), we used the Google API and queried the search engine with set of different pairs of words. For each test case, we built 4 queries i.e. the target word and one of the candidate-synonyms. Subsequently, we collected all of the seed results and followed a set of selected links to gather more textual information about the queried pairs. The overall collection of web pages was then shallow parsed using the MontyLingua software [18]. Thus, the WCSD consists of 500M words in 110K documents in which each sentence is a predicate structure. The benefit of thus gathered corpus is to maximize the ratio of the observed instances to the volume of the text processed.

5 Results and Discussion

To illustrate the results of our methodology, we use 145 noun test cases i.e. all 23 noun questions taken from the ESL (English as a Second Language) multiple choice test, all 19 noun cases from the TOEFL (Test of English as Foreign Language) [1]. We also add the subset of all 103 noun questions out of the 301 manually collected test cases provided by Peter Turney [11] and referred as RD (Reader's Digest). The success over synonymy tests does not guarantee success in real-world applications and the tests also show problematic issues as shown in [2]. However, the scores have an intuitive appeal,

they are easily interpretable, and the expected performance of a random guesser (25%) and of a typical non-native speaker are both known (64.5%), thus making TOEFL-like tests a good basis for evaluation.
The Table 2 shows the differences in terms of accuracy obtained by comparing the RCV1 with the WCSD. As expected, the WCSD allows significant improvement in terms of accuracy. These results clearly show that the corpus used to compute the measures influences drastically the performance of any experiment and comparisons of different methodologies should always be made based on the same statistical evidences. As a consequence, the results given in Table 1 are only indicative as better or worse results may be obtained on different experimental frameworks.

	Global		**Local**		**Product**	
	RCV1	WCSD	RCV1	WCSD	RCV1	WCSD
Cos TfIdf	38%	68%	42%	74%	42%	72%
Cos PMI	40%	63%	42%	66%	42%	63%
Cos Prob	36%	54%	40%	68%	39%	68%
Ehlert	41%	61%	44%	68%	44%	68%
Lin	38%	54%	41%	62%	42%	61%

Table 2. Comparison between RCV1 and WCSD.

In Table 2, we present the overall results of our experiments. All the models proposed in this paper were tested on set of 145 noun questions. The figures in the table show the accuracy level by test set and measure. The Cos TfIdf as a Local similarity evidences the overall best result with an accuracy of 74% and 108 correct answers. The table shows that the Local similarity approach improves over the Global similarity for all measures. In parallel, the worst results were obtained by the Lin model and the Cos Prob for the Global approach reaching 54%. However, those are the measures that benefit most from the introduction of the Local approach respectively improving by 14% (20 additional correct guesses) and 8% (12). This situation is in accord with the finding of [19] that the performance of Lin's distributional similarity score decreases more significantly than other measures for low frequency nouns [20]. Thus Global similarity fails to realize its advantage for the less polysemous yet less frequent cases and leaves more room for improvement by Local similarity.
In order to have a better understanding of the results, we applied the measures to all tests individually, i.e. RD, TOEFL and ESL. The results are shown in Table 3. The best results are obtained by (1) the Cos TfIdf as Local similarity with 69% for the RD multiple choice test, (2) the Cos Prob as Product similarity[2] with 84% for the TOEFL test and (3) the Cos TfIdf as Local similarity with 96% for the ESL. These results clearly show that the type of the test influences the overall performance and also point out the fact that different measures can be tuned for different tests. As a consequence, it is important to understand the behavior of each measure as well as the characteristics of each

[2] The parameter $\alpha = 0.46$ from Equation 8 is tuned through ten-fold cross validation.

	Cos						Cos				
All 145	TfIdf	PMI	Prob	Ehlert	Lin	TOEFL 19	TfIdf	PMI	Prob	Ehlert	Lin
Global	68%	63%	54%	61%	54%	Global	74%	68%	47%	53%	74%
Local	**74%**	66%	68%	68%	62%	Local	79%	74%	68%	74%	74%
Product	72%	63%	68%	68%	61%	Product	74%	68%	**84%**	68%	68%
L - G	6%	3%	**14%**	7%	**8%**	L - G	5%	6%	21%	21%	0%
RD 103	TfIdf	PMI	Prob	Ehlert	Lin	ESL 23	TfIdf	PMI	Prob	Ehlert	Lin
Global	64%	62%	52%	60%	50%	Global	78%	61%	70%	74%	52%
Local	**69%**	61%	65%	65%	56%	Local	**96%**	83%	83%	74%	78%
Product	69%	61%	60%	65%	56%	Product	83%	70%	83%	78%	74%
L - G	5%	-1%	13%	5%	6%	L - G	18%	22%	13%	0%	26%

Table 3. Accuracy by test.

test set.

Freitag and colleagues [2] were the first to introduce a measure to evaluate the difficulty of a test based on its polysemy. In this paper, we go further in this analysis by taking into account the level of polysemy of the correct answer compared to the decoys. The first column of Table 4 shows the level of polysemy[3] when the correct answer is the most polysemous noun among all the alternatives. Similarly, the second column shows the level of polysemy when the correct answer is the second most polysemous noun from all the decoys, and so on. The results show that on average all test sets have simi-

	Most	Second	Third	Least	Avg
All	6.9(38)	4.3(28)	2.8(31)	1.6(48)	3.8(145)
RD	6.5(29)	4.0(19)	2.6(20)	1.5(35)	3.6(103)
TOEFL	6.2(6)	4.0(6)	2.8(4)	1.6(3)	4.0(19)
ESL	9.6(3)	5.9(3)	3.8(7)	1.9(10)	4.0(23)

Table 4. Correct answers by polysemy rank.

lar polysemy level. However, the distribution of the correct answers over the polysemy categories (i.e. Most, Second, Third and Least) is different especially for the TOEFL. This situation is crucial to understand why the Cos Prob, applied as a Product, gives better results for the TOEFL compared to other measures and test sets. In fact, the biggest proportion of correct answers is highly polysemous in TOEFL compared to RD and ESL. As a consequence, this similarity measure seems to benefit the extraction of most polysemous answers. This is indeed an important observation as most complicated cases to solve are the polysemous ones. However, we can not draw definitive conclusions from this first evidence. Indeed, we also need to understand better the behavior

[3] The level of polysemy is calculated as in [2] and is the sum of the polysemy of the target noun and the correct answer given by WordNet.

of each measure with respect to polysemy and frequency. In each column of Table 5, we can find the accuracy level, the number of correct answers in brackets and the type of similarity approach which gives the best result (g for global, l for local and p for product). Similarly to what we observed from the results and from the analysis of the

By Frequency	Most	Second	Third	Least
Cos TfIdf	80% l	72% l	78% l	66% l
Cos PMI	66% l	65% l	74% l	**69% g**
Cos Prob	**81% p**	**77% p**	61% l	59% l
Ehlert	73% l	66% l	61% l	**72% g**
Lin	59% l	51% l	78% l	59% l

By Polysemy	Most	Second	Third	Least
Cos TfIdf	76% l	**68% p**	77% l	77% l
Cos PMI	66% l	50% l	74% l	**81% g**
Cos Prob	74% l	**71% p**	71% l	63% l
Ehlert	66% l	61% l	74% l	**79% g**
Lin	55% l	57% l	71% l	65% l

Table 5. Accuracy by test characteristics.

test sets, Table 5 shows that the Product similarity only appears as the best result when the correct answer is frequent word or polysemous one. In this case, the Cos Prob and the Cos TfIdf are the only measures to show this behavior. In parallel, the Global similarity provides best results when the correct answer is the least frequent or the least polysemous alternative. In this case, the Ehlert model and the Cos PMI are the only two elected measures.

Finally, in order to understand better the differences between the Product, the Local and the Global similarity, we illustrate their behaviors by polysemy rank for the Cos Prob in Figure 1. The reason why Local similarity performs better than the Global similarity

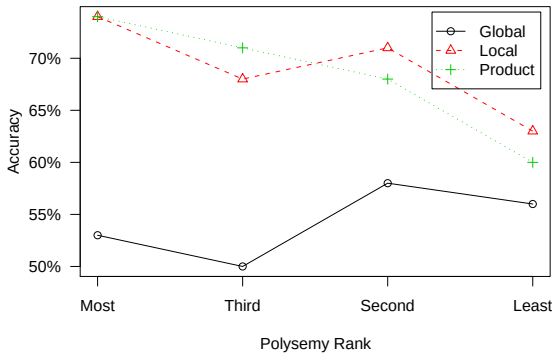

Fig. 1. Performance by polysemy rank.

for more polysemous cases is that it always compares monosemous representations and is not influenced by polysemous nouns. With the decreasing of the polysemy, Global similarity becomes enough informed to solve the test correctly while Local similarity

looses because less polysemous nouns are less frequent and thus become less probable to be encounter both nouns in the same documents. This is illustrated by the convergence of the Local and Global lines in Figure 1. In terms of Product similarity, its value depends on the rate of the polysemy of the test case. As we have seen before, the Product similarity is more tuned to solve more polysemous cases. Subsequently, in Figure 1, we see that the Product similarity overtakes the similarity values of both the Global and the Local similarities for most polysemous cases. However, when polysemy decreases Local similarity performs better than all other measures. At the same time, the more meanings in the test set, the more the ability of the Local similarity to implicitly disambiguate is valued and the lower the value of α is. On the other hand more weight to Global similarity should be given when the test set contains rare words and there is no enough statistics for the Local similarity.

All the observations made so far show that does not exist any single methodology to accomplish synonymy detection alone. Depending on the type of the multiple choice question set, different measures and weighting schemas may be applied to improve overall performance. However, we already saw that some measures seem highly correlated such as the Ehlert model and the Cos PMI for the Global similarity. In this case, the combination of these measures will not benefit the overall performance. For that purpose, we propose to measure the correlation between pairs of similarity measures with the Pearson Product-Moment Correlation test [21] (See Table 6). Additionally, we

	Pearson	**Overlap**	**Optimal**
Cos Prob & Ehlert	**0.132**	**0.572**	**86% (124)**
Cos Prob & Cos PMI	0.153	0.565	84% (122)
Cos Prob & Cos TfIdf	0.334	0.677	83% (121)
Cos Prob & Lin	0.245	0.581	81% (117)
Cos PMI & Cos TfIdf	0.507	0.739	79% (115)
Ehlert & Cos TfIdf	0.555	0.773	79% (115)
Ehlert & Lin	0.448	0.675	77% (111)
Ehlert & Cos PMI	0.551	0.745	76% (110)
Lin & Cos TfIdf	0.610	0.763	76% (110)
Lin & Cos PMI	0.606	0.750	72% (104)

Table 6. Inter-measure correlation.

compute the overlap of correct answers in the second column and finally calculate the possible optimal performance that could be obtained by combining two measures both in percentage and number of possible correct answers.

The results clearly show that all similarity measures would benefit the most from their association with the Cos Prob. In particular, the optimal case could achieve 86% accuracy by combining the Cos Prob and the Ehlert model. Indeed, both measures share the second smallest proportion of correct test cases i.e. 57.2% and the smallest correlation i.e. 0.132. Moreover, by looking at Table 5, both measures individually show three of the best four results for the frequency distribution i.e. Cos Prob = 81% for the most

frequent, Cos Prob = 77% for the second most frequent and Ehlert = 72% for the least frequent.
To conclude, this exhaustive evaluation allows us to say that (1) the corpus size matters, (2) the "*one sense per discourse*" paradigm improves steadily over the baseline i.e. the Global similarity, (3) different measures provide uncorrelated results and (4) the combination of similarity measures would lead to improved results.

6 Conclusion

In this paper, we presented a new heuristic based on the attributional similarity paradigm in attempt to alleviate word polysemy problem in synonymy discovery without performing explicit word sense disambiguation. Our method proved to gain greatest advantage over Global similarity namely in most polysemous cases.
In particular, we obtain 96% accuracy on ESL, 84% on TOEFL, 69% on RD and 74% over the all joined test cases.
Further experiments have also been conducted to evidence result differences between web corpora over standard collections of texts motivated by recent discussions in the NLP area and show that optimal performance is obtained with web text collections tailored for our specific task.
The main contribution is certainly the exhaustive evaluation and categorization of the different similarity measures that were tested. Indeed, as previously shown by [2], the tests show problematic issues that can influence the results of different similarity measures. For that purpose, we have first shown the differences in terms of polysemy between the RD, the TOEFL and the ESL. Then, based on these results, we have conducted further experiments that showed that some similarity metrics are more tailored to solve polysemous cases than others (e.g. the Cos Prob). Finally, by looking at the Pearson Product-Moment Correlation coefficient between pairs of measures, we clearly evidence that multiple choice question tests for synonymy detection should be solved by the optimization of a learning function based on the combination of similarity measures.

Acknowledgment

We would like to thank P. Turney, T. Landauer and D. Freitag for providing us with their collections of tests.

This work is supported by the VIPACCESS project funded by the Portuguese Agency for Research (Fundasoção para a Ciência e a Tecnologia) with the reference PTDC/PLP/72142/2006.

References

1. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review **104**(2) (1997) 211–240

2. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, Michigan (2005) 25–32
3. Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: HLT '94: Proceedings of the workshop on Human Language Technology, Morristown, NJ, USA, Association for Computational Linguistics (1994) 240–243
4. Gale, W., Church, K.W., Yarowsky, D.: One sense per discourse. In: HLT '91: Proceedings of the workshop on Speech and Natural Language, Morristown, NJ, USA (1992) 233–237
5. Moraliyski, R., Dias, G.: One sense per discourse for synonymy extraction. (2006)
6. Terra, E., Clarke, C.: Frequency estimates for statistical word similarity measures. In: Proceedings of HTL/NAACL 2003, Edmonton, Canada (2003) 165–172
7. Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity. In: Proceedings of COLING 2004, Geneva, Switzerland (2004)
8. Ehlert, B.: Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics. Master's thesis, University of California, San Diego (2003)
9. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA (1998) 296–304
10. Turney, P.D.: Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. Lecture Notes in Computer Science **2167** (2001) 491–502
11. Turney, P.D., Littman, M.L., Bigham, J., Shnayder, V.: Combining independent modules in lexical multiple-choice problems. In: Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003. (2003) 101–110
12. Jarmasz, M., Szpakowicz, S.: Roget's thesaurus and semantic similarity. In: Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria (2004) 212–219
13. Curran, J.R., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, USA (2002) 59–66
14. Rapp, R.: Word sense discovery based on sense descriptor dissimilarity. In: Proceedings of the Ninth Machine Translation Summit. (2003) 315–322
15. Fellbaum, C., ed.: WordNet: an electronic lexical database. The MIT Press (1998)
16. Sahlgren, M., Karlgren, J.: Vector-based semantic analysis using random indexing for cross-lingual query expansion. In: CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, London, UK (2002) 169–176
17. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5** (2004) 361–397
18. Liu, H.: Montylingua: An end-to-end natural language processor with common sense. (2004) Available at: `web.media.mit.edu/~hugo/montylingua`.
19. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical distributional similarity. Computational Linguistic **31**(4) (2005) 439–475
20. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Unsupervised acquisition of predominant word senses. Comput. Linguist. **33**(4) (2007) 553–590
21. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. In: Biometrika. Volume 10. (1915) 507–521