One Sense Per Discourse for Synonym Detection

Rumen Moraliyski, Gaël Dias
Centre of Human Language Technology and Bioinformatics
University of Beira Interior
rumen@penhas.di.ubi.pt, *ddg@di.ubi.pt*

Abstract

In this paper, we present a new methodology for synonym detection based on the combination of global and local distributional similarities of pairs of words. The methodology is evaluated on the noun space of the 50 multiple-choice synonym questions taken from the ESL and reaches 91.30% accuracy using a conditional probabilistic model associated with the cosine similarity measure.

Keywords

Synonym discovery, similarity measure, discourse

1 Introduction

The task of recognizing synonyms can be defined as in [15]: *"given a problem word and a set of alternative words, choose the member from the set of alternative words that is most similar in meaning to the problem word"*. Based on this definition, many algorithms [3] [4] [6] [7] [13] [14] [15] [16] have been proposed and evaluated using multiple-choice synonym questions taken from the Test of English as Foreign Language (TOEFL).

Most of the works proposed so far explore the attributional similarity paradigm [10].

To construct attributional representation of a word, many approaches have been developed: window oriented [3], [4], [13], [14], lexicon oriented [1], syntactic oriented [2], [17], document oriented [7].

Most of the work proposed so far, independently of their categorization, have in common the fact that the word representation is built on global corpus evidence. As a consequence, all the senses of a polysemous word share a single description. This fact is clearly a drawback for any word meaning analysis. Indeed, this would mean that, to be synonyms, two words should share, as many as possible of their senses, while they usually do share just one.

A first attempt to take into account local corpus evidence is proposed in [11] who separate corpus evidences for distinct word occurrences in a corpus to build a matrix that is afterwards subjected to a SVD and analyzed to discover the major word senses. However, they do not propose any evaluation and validation of their work, neither it is reproducible on a small scale i.e. single texts.

Here, we propose a method to measure syntactic oriented attributional similarity based on the *"one sense per discourse"* paradigm. Instead of relying exclusively on global distributions, we build words representations and compare them within documents limits. In this way, we only compare two specific senses of each word at a time.

We argue that our proposal coupled with the global approach leads to improved results. In order to test this assumption, we implemented the vector space model over term frequency, term frequency weighted by inverse document frequency, Pointwise Mutual Information [14] and conditional probability [17]. We also implemented two probabilistic similarity measures: the Ehlert model [3] and Lin model [8]. The evaluation was conducted on the subset of the 23 noun questions of a 50 multiple-choice synonym questions taken from the ESL (test for students of English as Second Language) provided by P. Turney. The best results were obtained by the vector space model over the conditional probability which scored 91% accuracy (i.e. 21 out of 23 nouns questions).

2 Related Work

Previous research on corpus-analytic approaches to synonymy has used the TOEFL and ESL which consist of set of multiple-choice questions. In this context, a distance function must be defined to order the correct answer word in front of then decoys.

One of the most famous work is proposed by [7] who use document distribution 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 window based vector space model. For that purpose, the word context vectors associated to all the words from the TOEFL are built on co-occurrence basis within the entire corpus. [14] studied a variety of similarity metrics and weighting schemes of contexts and achieved a statistical tie with their DR-PMI compared to the PMI-IR proposed by [15].

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 to measure similarity between pairs of words. This work does not follow the attributional similarity paradigm but rather proposes a heuristic to measure semantic distance. [16] refined the PMI-IR algorithm and proposed a module combination to include new features such as LSA and thesaurus evidences.

In parallel, some works have focused on linguistic features to measure similarity. [6] give results for a number of relatively sophisticated thesaurus-based methods that looked at path length between words in the heading classifications of Roget's Thesaurus. However, this methodology does not follow the attributional similarity paradigm unlike [2], who use syntactic context relations.

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%
Turney et al. 2003	97.50%

Table 1: *Accuracy on TOEFL question set.*

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 [2].

Unfortunately, to our knowledge, unlike window based approaches, syntactic based methodologies have not been tested over TOEFL or ESL. Rather, they have been used to build linguistic resources. As a summary, Table 1 presents the results achieved by most of the mentioned methodologies[1].

3 Proposal

While the attributional similarity paradigm has been used over global corpus evidence, the *ad hoc* metrics have privileged, to some extent, a closer view of the data taking advantage of the *"one sense per discourse"* hypothesis proposed by [5]. To our point of view discarding the corpus structure in terms of documents is a key factor for the "failure" of the attributional similarity measures based on global corpus evidence.

Our proposal consists in implementing *"one sense per discourse"* through comparing two words within a single document at a time and averaging over the documents in which both words were encountered. As a result words that co-occur in a document but with different meanings will rarely share contexts and will end with low similarity. On the other hand words that co-occur as synonyms will share contexts with greater probability hence will receive higher similarity estimation. The value obtained we call local similarity.

Finally, we combine the local similarity with the global one under the syntactic attributional paradigm to achieve improved performance.

4 The Corpus

4.1 Motivation

Any work based on the attributional similarity paradigm depends on the corpus used to calculate the values of the attributes. [14] use a terabyte of web data that contains 53 billion words and 77 million documents, [13] a 10 million words balanced corpus with a vocabulary of 94 thousand words and [3], [4] a 256 million words North American News Corpus (NANC). As mentioned in [3], [14], the size of the corpus does matter and the bigger the corpus is, the better the results are. In our case, we could also have used NANC. However our proposal demands co-occurrence of the two synonym candidates within a single document few times each. It is improbable that general purpose corpus would comprise enough documents containing pairs of our set of words four or more times each. As a result we decided to build a corpus suitable to the problem at hand thus exploring the merits and flaws of the approach as opposed to solving a problem fit to the data available. The corpus is available at http://hultig.di.ubi.pt/.

4.2 Construction

To build our corpus, we used the Google API and queried the search engine with 92 (23 questions × 4 alternatives) different pairs of words. For each ESL test case, we built 4 queries - target word and one of the proposed variants. Subsequently, we collected all of the seed results, lemmatized the text using the MontyLingua software [9] and followed a set of selected links to gather more textual information about the queried pairs. Preference to texts where only the rarest pairs occur was given. Indeed, if in the text there is one rare pair with high $tf(.,.).idf(.)$ and many others for which we already have many examples (i.e. with low $idf(.)$), then we should choose only few links for further crawling as the new textual material would bring more of the same.

One of the problems with web pages is that some of them only consist of link descriptions and do not contain meaningful sentences. In order to be sure that the processed web pages provide useful textual material as well as useful links, we assured that for each link in the page there were at least 300 characters of running text.

For our final corpus we retained those documents that contained at least one of the test pairs. Thus, the corpus consists of 39 million words and 122 thousand word types in nearly 16 thousand documents. The overall corpus was finally shallow parsed using the MontyLingua software [9] to obtain a predicate structure for each sentence.

5 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 (or attributes) associated with a label i, the similarity between two units i and j is defined

[1] The values can not be compared directly as they may not be evaluated (1) on the same corpora or/and (2) the same set of questions. However, these results will give the reader an idea of the expected results for future methodologies. For more information about evaluation see [12].

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 $\langle r, v\rangle$ 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 *"press"* within a subject relation, we will have the following triple $\langle$*brass, press, subject*$\rangle$ and the tuple $\langle$*press, subject*$\rangle$ will be an attribute of the word context[2] 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.

5.1 Weighting Attributes

In order to construct more precise representations of word meanings, numerous weighting schemas have been developed.

5.1.1 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 and N the set of all the nouns.

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

5.1.2 Pointwise Mutual Information

The value of each attribute $\langle r, v\rangle$ can also be seen as a measure of association with the noun being characterized. For that purpose, [15], [14] have proposed to use the Pointwise Mutual Information (PMI) as defined in Equation 2 where n is the target noun and $\langle r, v\rangle$ a given attribute.

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

5.1.3 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. In our case, we are interested in knowing how strongly a given attribute $\langle v, r\rangle$ may evoke the noun n.

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

The conditional probability could also be seen as the $\langle n, v\rangle$ distribution over the possible relations between n and v.

$$P(n, v|r) = \frac{P(n, v, r)}{P(r)} \quad (4)$$

Due to this characteristic, the model would suffer low selectivity - the similarity values calculated based on it would be within very short interval, which would result in unconfident decisions, as we tested and evidenced.

5.2 Similarity Measures

There exist many similarity measures in the context of the attributional similarity paradigm [17]. They can be divided into two main groups: (1) metrics in a high dimensional space also called Hyperspace Analogue to Language (HAL) [3], (2) measures which calculate the correlations between different probability distributions.

5.2.1 Cosine Similarity Measure

To quantify similarity between two words in a vector space model, the cosine metric measures to what extent two verb context vectors point along the same direction. It is defined in Equation 5.

$$cos(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} X_{jk}}{\sqrt{\sum_{k=1}^{p} {X_{ik}}^2} \sqrt{\sum_{k=1}^{p} {X_{jk}}^2}} \quad (5)$$

5.2.2 Probabilistic Measures

Probabilistic measures can be applied to evaluate the similarity between nouns when they are represented by a probabilistic distribution. In this paper, we will employ two different measures.

Ehlert model: Equation 6 presents proposed in [3] measure which evaluates the probability to interchange two word context vectors (i.e. what is the probability that the first noun is changed for the second one).

$$P(n_1|n_2) = \sum_{\langle v, r\rangle \in A} \frac{P(n_1|v, r)P(n_2|v, r)P(v, r)}{P(n_2)} \quad (6)$$

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

Lin model: [8] defines similarity as the ratio between the amount of information needed to state the

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

commonality of the two nouns and the total information available about them.

$$Lin(n_1, n_2) = \frac{2 \times \sum\limits_{\langle v,r\rangle \in A} \log_2 P(v,r)}{\sum\limits_{\langle v,r\rangle \in B} \log_2 P(v,r) + \sum\limits_{\langle v,r\rangle \in C} \log_2 P(v,r)} \quad (7)$$

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

5.3 Global and Local Similarity

The common attributional similarity approach of gathering statistics from large corpora discards the information within single texts which has shown promising results as in [15]. Indeed building the verb context vectors based on the overall corpus by treating it as a single huge text implies the assumption that described words are monosemous.

The local attributional similarity approach, on the other hand, aims at introducing the document dimension to the word meaning acquisition process. As a consequence, different noun meanings are not merged together into single vector. The formal expression of the the local similarity is given in Equation 8 where D is the set of texts in the corpus where both n_1 and n_2 appear and $sim(.,.)$ is any similarity measure described above calculated within the document and not over the entire corpus.

$$Lsim(n_1, n_2) = \frac{\sum\limits_{d \in D} sim(n_1, n_2)}{card(D)} \quad (8)$$

This modification implies that the attribute values are calculated within the document for each member of the sum.

The global similarity works as an indicator that the words n_1 and n_2 are similar and the local similarity confirms that n_1 and n_2 are not just only similar, but instead good synonym candidates. Hence their product reaches maximal value when the words compared are synonyms. In Equation 9 $Gsim(.,.)$ is any similarity measure computed over the entire corpus.

$$Psim(n_1, n_2) = Gsim(n_1, n_2) \times Lsim(n_1, n_2) \quad (9)$$

6 Results and Discussion

The success over the ESL test does not guarantee success in real-world applications and the test also shows problematic issues [4]. However, the scores have an intuitive appeal, they are easily interpretable, and the expected performance of a random guesser (25%) and typical non-native speaker performance are both known (64.5%), thus making TOEFL-like tests a good basis for evaluation.

All the models proposed in this paper were tested on the subset of the 23 noun questions of the 50 multiple-choice synonym questions taken from ESL. Table 2 shows the different results obtained for the HAL models and the Probabilistic models.

			Global	**Local**	**Product**
HAL	tf	1	39.13%	73.91%	73.91%
		4		73.91%	69.57%
	tf.idf	1	52.17%	73.91%	65.22%
		4		69.57%	69.57%
	PMI	1	78.26%	65.22%	78.26%
		4		73.91%	78.26%
	cosPr	1	73.91%	60.87%	73.91%
		4		**82.61%**	**82.61%**
Prob	Ehlert	1	78.26%	65.22%	69.57%
		4		60.87%	73.91%
	Lin	1	60.87%	73.91%	69.57%
		4		78.26%	69.57%

Table 2: *Performance for full noun vocabulary.*

For the local similarity, we make a distinction between the results obtained on the set of documents which contain both words (being compared) at least once or four times (lines marked "1" and "4" in tables 2 and 3).

For the HAL models, the best results are obtained by the cosine of conditional probability reaching 82.61% accuracy (i.e. 19 correct answers out of 23). An interesting characteristic of PMI is the fact that it behaves steadily and does not gain anything by introducing our local similarity measure or the product of similarities. As it is known PMI is biased toward rare events, but here we compare pairs of words in documents where they occur more often than by chance and thus PMI can not manifest its specificity.

The Probabilistic models, likewise the HAL models, give better results for the texts with more occurrences of the examined nouns. The best results are obtained by Lin measure with 78.26% accuracy for *Lsim*. One interesting result is the fact that the Ehlert model gives the best results on the global similarity while it looses greatly when introducing the local similarity. In fact, the Ehlert model is an asymmetric measure, which gives an important part of its weight to the marginal probability of the examined answer word. When dealing globally, the measure shows a tendency to select the word with lowest probability. In fact, like the Pointwise Mutual Information, Ehlert is biased to rare cases. When compared to locally obtained values the figures show that indeed it does not attribute much importance to the contexts. When calculating the local Ehlert measure, the marginal probability of the answer varies from document to document but in fact turns out to be more stable when local similarities are averaged. As a consequence, it loses selectivity.

In this first analysis, we took into account all the nouns of the corpus with their respective verb context vectors. However, the same calculations can be done just by looking at the 94 nouns of the 23 noun questions[3]. The impact of the other nouns in the corpus is only on the marginal probabilities and on the *idf* values. This experiment is reasonable since we want to distinguish between just a limited set of nouns. We need factors that can point out the differences and similarities between them and as a consequence the rest of the noun vocabulary is useless. Table 3 presents the results with the 94 nouns space.

[3] Some of the nouns appear in more than one test case hence 94 instead of $23 \times 5 = 115$

			Global	Local	Product
HAL	tf	1	39.13%	73.91%	73.91%
		4		73.91%	69.57%
	tf.idf	1	73.91%	69.57%	73.91%
		4		65.22%	65.22%
	PMI	1	60.87%	13.04%	30.43%
		4		26.09%	30.43%
	cosPr	1	65.22%	69.57%	86.96%
		4		**82.61%**	**91.30%**
Prob	Ehlert	1	65.22%	60.87%	69.57%
		4		60.87%	69.57%
	Lin	1	56.52%	65.22%	69.57%
		4		78.26%	69.57%

Table 3: *Performance for 94 ESL nouns.*

The overall best results were again obtained by $Psim(.,.)$ of cosine of conditional probability with 91.30% accuracy (21 correct answers over 23). However, almost all other measures loose in accuracy in all cases although they keep the same characteristics as shown in Table 2 when comparing the global, local and product figures. PMI shows a tendency to perform worse than random guesser. This observation is not a surprise since the synonyms tend to cooccur more often than by chance and so they receive lower weights by this scheme than when two unrelated words co-occur in a document. In this manner the synonymous words result with lower similarity than non-synonymous ones. Table 4 illustrates how the global similarity highlights related words yet the local similarity is the measure that selects the correct option.

stem	Global	Local	Product
a) column	0.0066	0.0370	0.0002
b) bark	0.0230	0.0225	0.0005
c) stalk	0.0278	**0.0577**	**0.0016**
d) trunk	**0.0288**	0.0151	0.0004

Table 4: *Global vs. Local cosPr.*

	Global	Local	Product
1	60.87%	65.22%	82.61%
4		78.26%	82.61%

Table 5: *Global PMI for 94 ESL nouns.*

It seems worth to investigate the combination between global association measure and local term representation thus taking advantage of more reliable association values still maintaining the context vector unambiguous. This effect is evidenced for the PMI comparing Tables 3 and 5.

7 Conclusions

According to [14] large enough corpora are necessary for human level performance on TOEFL synonymy test. But the common approach of gathering statistics from large corpora discards the information within single text. On the other hand, [15] shows that synonyms co-occur in texts more often than by chance. In this paper, we proposed a method which combines both approaches by employing global and local evidence of attributional similarity into a single measure. The methodology was evaluated on the noun space of the 50 multiple-choice synonym questions taken from the ESL and reached 91.30% accuracy with the cosine of conditional probability. The results presented here encourages us to perform larger scale evaluation and experiments in word meaning acquisition.

References

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, 2003.

[2] J. Curran and M. Moens. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA, 2002.

[3] B. Ehlert. Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics. Master's thesis, University of California, San Diego, 2003.

[4] D. Freitag, M. Blume, J. Byrnes, E. Chow, S. Kapadia, R. Rohwer, and Z. Wang. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 25–32, Ann Arbor, Michigan, 2005.

[5] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA, 1992.

[6] M. Jarmasz and S. Szpakowicz. Rogets thesaurus and semantic similarity. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 212–219, Borovets, Bulgaria, 2004.

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

[8] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[9] H. Liu. Montylingua: An end-to-end natural language processor with common sense. available at: http://web.media.mit.edu/˜hugo/montylingua, 2004.

[10] D. L. Medin, R. L. Goldstone, and D. Gentner. Similarity involving attributes and relations: judgments of similarity and differences are not inverses. *Psychological Science*, 1(1):64–69, 1990.

[11] R. Rapp. Utilizing the one-sense-per-discourse constraint for fully unsupervised word sense induction and disambiguation. In *Proceedings of Forth Language Resources and Evaluation Conference,LREC*, 2004.

[12] M. Sahlgren. Towards pertinent evaluation methodologies for word-space models. In *Proceedings of LREC 2006: Language Resources and Evaluation*, Genoa, Italy, 2006.

[13] M. Sahlgren and J. Karlgren. 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*, pages 169–176, London, UK, 2002.

[14] E. Terra and C. Clarke. Frequency estimates for statistical word similarity measures. In *Proceedings of HTL/NAACL 2003*, pages 165–172, Edmonton, Canada, 2003.

[15] P. D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167:491–502, 2001.

[16] P. D. Turney, M. L. Littman, J. Bigham, and V. Shnayder. Combining independent modules in lexical multiple-choice problems. In *In Recent Advances in Natural Language Processing III:Selected Papers from RANLP 2003.*, pages 101–110, 2003.

[17] J. Weeds, D. Weir, and D. McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of COLING 2004*.