Combination of Global and Local Attributional Similarities 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.

1 Introduction

The task of recognizing synonyms can be defined as in [Turney 2001]: *"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 [Landauer and Dumais 1997], [Turney 2001], [Sahlgren and Karlgren 2001], [Terra and Clarke 2003], [Ehlert 2003], [Turney et al. 2003], [Jarmasz and Szpakowic 2003], [Freitag et al. 2005] have been proposed and evaluated using the multiple-choice synonym questions taken from the Test of English as Foreign Language (TOEFL).

Most of the work proposed so far explore the attributional similarity paradigm [Medin et al. 1990] which computes similarities between attributionally described words. So, when two words have a high degree of attributional similarity, they can be called synonyms [Turney 2006][1].

Numerous approaches to construct attributional representation of a word have been developed: window oriented [Sahlgren and Karlgren 2001], [Terra and Clarke 2003], [Ehlert 2003], [Freitag et al. 2005], lexicon oriented [Banerjee and Pedersen 2003], syntactic oriented [Grefenstette 1993], [Curran and Moens 2002], [Weeds et al. 2004], [Yang and Powers 2006], document oriented [Landauer and Dumais 1997].

[1] We prefer to call them synonym candidates.

In order to classify these approaches, [Kilgarriff and Yallop 2000] use the terms loose and tight. On one hand, methodologies which use document as a context[2] only seem to identify loose associative kinds of semantic relationships. For example, the words "doctor" and "disease" are likely to be linked in an associative way. On the other hand, methods using syntactic information tend to identify tight semantic relationships between words. For example, such methods are likely to recognize a semantic similarity between the words "doctor" and "dentist", but not between "doctor" and "hospital".

In this paper, we propose different method to measure syntactic oriented attributional similarity[3]. 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. For example, one may consider the words "association" and "organization" as synonyms. However, only one of their meanings clearly evidences their synonymy relation. Indeed, if one takes the seventh sense of "organization" in WordNet *(i.e. synset 04712979)*, one could say *"his compulsive organization was not an endearing quality"*[4] . It is obvious that the word "association" cannot be used as a synonym of the word "organization" in this sentence.

[Rapp 2004] attempts to utilize separate corpus evidences from 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. Here we propose somewhat different approach to measure semantic similarity between words based on the *"one sense per discourse"* paradigm. Instead of relying exclusively on global distributions, we compare word representations within single document.

If we go on with our previous example, the words "organization" and "association" are unlikely to appear in the same text as polysemous thus solving part of the problem plaguing the global approach - mixing many senses in a single representation. Instead, by building the word representation within a single document, the attributional similarity is calculated between two senses.

We argue that our proposal coupled with the global approach lead to improved results. In order to test that, we implemented the vector space model over term frequency, term frequency weighted by inverse document frequency, Pointwise Mutual Information [Terra and Clarke 2003] and conditional probability [Weeds et al. 2004]. We also implemented two probabilistic similarity measures: the Ehlert model [Ehlert 2003] and Lin model [Lin 1998]. All models were tested 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) kindly provided by P. Turney. The best results were obtained by

[2] A sliding window over raw text corpora.

[3] Indeed, this work is included in a global project aiming at automatically building ontologies based on hierarchical soft clustering where each cluster may contain only noun synonyms or noun near synonyms [Dias et al. 2006].

[4] This sentence is taken from Wordnet 2.1.

the vector space model over the conditional probability which scored 91.30% accuracy (i.e. 21 out of 23 nouns questions).

2 Related Work

Previous research into corpus-analytic approaches to synonymy has used the TOEFL and ESL which consist of set of multiple-choice questions. Each question involves five words: the problem word and four response words, one of which is a synonym of the target and the other ones are called decoys. In this context, a distance function must be defined to order the correct answer word in front of the decoys.

One of the most famous work is proposed by [Landauer and Dumais 1997] 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. [Terra and Clarke 2003] studied a variety of similarity metrics and weightings of contexts and achieved a statistical tie with their DR-PMI compared to the PMI-IR proposed by [Turney 2001].

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. [Turney et al. 2003] 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. [Jarmasz and Szpakowic 2003] 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 [Curran and Moens 2002], who use syntactic context relations.

In the 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, *(adjective, good)* and *(direct-obj, have)* are attributes of the noun "idea" as illustrated in [Curran and Moens 2002].

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 [Yang and Powers 2006]. As a summary, Table 1 presents the results achieved by most of the mentioned

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.

methodologies[5].

By analyzing the related work, there are clearly two different approaches for synonym detection evaluated against a TOEFL-like test set: the attributional similarity paradigm [Landauer and Dumais 1997], [Sahlgren and Karlgren 2001], [Terra and Clarke 2003], [Ehlert 2003], [Freitag et al. 2005] and the definition of ad hoc similarity measures [Turney 2001], [Turney et al. 2003], [Jarmasz and Szpakowic 2003]. Although, one could think that best results should be obtained by theoretically founded metrics, results show the contrary. The best results are obtained by [Turney et al. 2003] who use a combination of document features (LSA), linguistic knowledge (thesaurus), information retrieval features (specific operators and connectors) and a co-occurrence measure (Pointwise Mutual Information).

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 [Gale et al. 1992]. 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 effect of this 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.

We combine the local similarity with the global one to achieve improved performance over it. In fact, the global similarity should work as an indicator

[5] All 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 [Sahlgren 2006].

that two words are similar and the average local similarity confirms that both words are not just only similar, but instead good synonym candidates.

In order to evaluate our proposal, for the global and local approach, we will implement the vector space model over term frequency, term frequency weighted by inverse document frequency, Pointwise Mutual Information [Terra and Clarke 2003] and conditional probability [Weeds et al. 2004]. We will also implement two probabilistic similarity measures: the Ehlert model [Ehlert 2003] and Lin model [Lin 1998]. In particular, we will use the syntactic oriented attributional similarity paradigm to be able to find tight relations between words.

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. [Terra and Clarke 2003] use a terabyte of web data that contains 53 billion words and 77 million documents, [Sahlgren and Karlgren 2001] a 10 million words balanced corpus with a vocabulary of 94 thousands words and [Ehlert 2003], [Freitag et al. 2005] a 256 million words North American News Corpus (NANC). As mentioned by [Ehlert 2003], [Terra and Clarke 2003], 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. Corpus of scale of the one used by [Terra and Clarke 2003] probably would do, but the processing of such a corpus is a heavy task. As a result we decided to build a corpus satisfying our specific necessities which is available at http://hultig.di.ubi.pt/∼rumen/Corpus/Index.html.

4.2 Construction

To build our corpus, we used the Google API and queried the search engine with 92 different pairs of words. For each ESL test case, we built each query based on the target word and one of the proposed variants . Subsequently, we collected all of the seed results, lemmatized them using the MontyLingua software [Liu, 2004] and followed a set of selected links to gather more textual information about the queried pairs. In order to choose which links to follow, we defined a Text Quality function $TQ(.)$ for each text which value would decide upon the selection of links. If the $TQ(.)$ of text t (i.e. $TQ(t)$) is low, then it is useless to follow the links in t. Otherwise, we should follow the links until enough textual data has been gathered for statistical evidence. This restriction function is defined in Equation 1 where t is a web page and p is a pair of words,

T is the set of texts retrieved so far, P is the set of all word pairs and $c1$ is a tuning constant[6].

$$TQ(t) = \frac{\sum_{p_i \in t} tf(p_i, t) \times idf(p_i)}{\max_{p_i \in t}(tf(p_i, t) \times idf(p_i)) \times card(\{p_i | tf(p_i, t) > 2\})} \qquad (1)$$

where

$$idf(p) = \log_2 \frac{\left(\frac{\sum_{p_i \in P} card(\{t_j \in T | tf(p_i, t_j) > 2\})}{card(P) + c_1} \right)}{card(\{t_i \in T | tf(p, t_i) > 2\})}$$

The basic idea of Equation 1 is to give preference to texts where only the rarest pairs occur. Indeed, if there is one rare pair with high $tf(.,.).idf(.)$ and many others for which we already have many texts (i.e. with low idf), then the $TQ(.)$ value will be low. As a result, this will lead to choose only a few links from this page 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 extracted web pages will provide useful text material as well as useful links for further crawling, we propose a simple heuristic defined in Equation 2 which integrates the $TQ(.)$ function. We call it the Page Quality function and denote it $PQ(.)$ where $c2$ is a tuning constant[7].

$$PQ(t) = \frac{TQ(t) \times TextLengthInCharacters}{c_2 \times LinksCount} \qquad (2)$$

In order to build our final corpus, we selected those documents that contained at least one of the test pairs. Thus, the corpus consists of 38.794.161 words and 122.665 types. The overall corpus was finally shallow parsed using the MontyLingua software [Liu, 2004] to obtain a predicate structure for each sentence.

5 Syntactic 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 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 syntactic attributional similarity approach is implemented as follows: each variable of the word context vector is a tuple $< r, v >$ where r is an object or subject relation, and v is a given verb appearing within

[6] In our experiments, we used $c1 = 500$, causing the crawler to be always greedy for 500 more documents.

[7] In our experiments, we used $c2 = 300$. This requires that the web page contains at least 300 characters per link.

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 *(brass, subject, press)* and the tuple <*subject, press*> will be an attribute of the word context[8] vector associated to the noun *"brass"*.

As similarity measures are based on real-value attributes, our task is twofold. First, we must define a function which will evaluate the importance of a given attribute $< r, v >$ 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. Here, we will present the term frequency, the term frequency weighted by inverse document frequency, the pointwise mutual information and the conditional probability.

5.1.1 Word Frequency and Inverse Document Frequency

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 $< r, v >$ associated to n i.e. $tf(< r, v >, n)$. 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 3 where n is the target noun, $< r, v >$ a given attribute and N the set of all the nouns.

$$tf.idf(< r, v >, n) = tf(< r, v >, n) \times \log_2 \frac{card(N)}{card(\{n_i \in N | \exists(n_i, r, v)\})} \tag{3}$$

However, the vector space model can be defined with other weighting schemas: association measures or probabilities.

5.1.2 Pointwise Mutual Information

The value of each attribute $< r, v >$ can also be seen as a measure of association. For that purpose, [Turney 2001], [Terra and Clarke 2003] have proposed to use the Pointwise Mutual Information (PMI). The PMI is defined in Equation 4 where n is the target noun and $< r, v >$ a given attribute.

$$PMI(< n|r >, < v|r >) = \log_2 \frac{P(n, v|r)}{P(n|r)P(v|r)} \tag{4}$$

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

5.1.3 Conditional Probability

Another way to look at the relation between a noun n and a tuple $< r, v >$ is to estimate their conditional probability of co-occurrence. In our case, we are interested in knowing how strong a given attribute $< r, v >$ may select the noun n. This can easily be interpreted in terms of conditional probability as expressed in Equation 5.

$$P(n| < r, v >) = \frac{P(n, r, v)}{P(< r, v >)} \tag{5}$$

5.2 Similarity Measures

There exist many similarity measures in the context of the attributional similarity paradigm [Weeds et al. 2004]. They can be divided into two main groups: (1) measures which calculate the angles between vectors in a high dimensional space also called Hyperspace Analogue to Language [Ehlert 2003], (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 6.

$$cos(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} X_{jk}}{\sqrt{\sum_{k=1}^{p} {X_{ik}}^2} \sqrt{\sum_{k=1}^{p} {X_{jk}}^2}} \tag{6}$$

5.2.2 Probabilistic Measures

Probabilistic measures can be applied to evaluate the similarity between nouns when there are represented by a probabilistic distribution. In this paper, we will experiment two different measures.
Ehlert model: [Ehlert 2003] proposes a 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). This measure is presented in Equation 7.

$$P(n_1, n_2) = \sum_{<r,v>} \frac{P(n_1| < r, v >)P(n_2| < r, v >)P(< r, v >)}{P(n_2)} \tag{7}$$

Lin model: [Lin 1998] defines similarity in terms of information theory. This model is universal as it is applicable as long as the domain has a probabilistic distribution and it is theoretically justified. This measure is defined in Equation 8.

$$Lin(n_1, n_2) = \frac{2 \times \sum_{<r,v> \in A} \log_2 P(< r, v >)}{\sum_{<r,v> \in B} \log_2 P(< r, v >) + \sum_{<r,v> \in C} \log_2 P(< r, v >)} \tag{8}$$

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

5.3 Global and Local Similarity

According to [Terra and Clarke 2003] large enough corpora are necessary for human level performance on TOEFL synonymy test. But the common attributional similarity approach of gathering statistics from large corpora discards the information within single texts which has shown promising results as in [Turney 2001]. So, instead of relying exclusively on global distributional similarities between pairs of words, we believe that candidate synonyms must be compared not only based on global distributions, but one document at a time as well.

The global attributional similarity approach builds the verb context vectors based on the overall corpus by treating it as a single huge text. So, statistics are calculated on this basis where all different word meanings are gathered into a single representation. As a consequence, the corpus document structure is not taken into account.

The local attributional similarity approach aims at introducing the document dimension to evaluate the similarity between nouns. As a consequence, different noun meanings are not merged into the same vector thus implementing *"one sense per discourse"* paradigm. For that purpose, we propose a simple way to evaluate similarity in large corpora as defined in Equation 9 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.

$$Lsim(n_1, n_2) = \frac{\sum_{d \in D} sim(n_1, n_2)}{card(D)} \tag{9}$$

This slight modification implies some necessary adjustments for the calculation of the above mentioned similarity measures. In particular, when dealing with the local similarity measure $Lsim(.,.)$, all probabilities as well as term frequencies and inverse document frequencies are calculated within each document and not over the all corpus.

Finally, we propose another measure which gathers both global and local similarity. Indeed, the global similarity should work as an indicator that two words are similar and the local similarity confirms that two words are not just only similar, but instead good synonym candidates. For that purpose, we just multiply both global and local similarities as shown in Equation 10 where $Gsim(.,.)$ is any similarity measure computed over the entire corpus, discarding the corpus document structure.

$$Psim(n_1, n_2) = Gsim(n_1, n_2) \times Lsim(n_1, n_2) \tag{10}$$

6 Results and Discussion

The success over the ESL test does not guarantee success in real-word applications and the test also shows problematic issues [Freitag et al. 2005]. 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.

			Global	**Local**	**Product**
HAL	tf	1	39.13%	73.91%	73.91%
		4		73.91%	69.57%
	tf.idf1	1	52.17%	73.91%	65.22%
		4		69.57%	69.57%
	tf.idf2	1		73.91%	73.91%
		4		78.26%	78.26%
	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.

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.

For the local attributional similarity, two adaptations must be introduced. On one hand, we propose two measures of the $tf.idf$. The first one ($tf.idf1$) is the usual measure where the *idf* is calculated over the entire corpus and the tf.idf2 adapts the usual *idf* by calculating it for each text. So, while the *idf* is unique for a given attribute in the first case, it changes from text to text in the second case (see Equation 3).

Moreover, 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.

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[9].

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 92 nouns of the 23 noun questions of the 50 multiple-choice synonym questions taken from the ESL. The impact of the other nouns in the corpus will be on (1) the marginal probabilities of the probabilistic models and the PMI, and (2) on the *idf* for the HAL models. So in Table 3, we present the results obtained by just looking at the 92 nouns.

			Global	**Local**	**Product**
HAL	tf	1	39.13%	73.91%	73.91%
		4		73.91%	69.57%
	tf.idf1	1	73.91%	69.57%	73.91%
		4		65.22%	65.22%
	tf.idf2	1		60.87%	69.57%
		4		60.87%	73.91%
	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%	56.22%	69.57%
		4		78.26%	69.57%

Table 3: Performance for ESL subset of the noun vocabulary.

Interestingly, the overall best results were obtained in this case by the cosine of conditional probability with 91.30% accuracy (21 correct answers over 23) when conjugated with the $Psim(.,.)$ similarity measure. 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 approaches. PMI shows a tendency to perform worse than random guesser. This observation is not surprise since the synonyms tend to co-occur more often than by chance and so they receive lower weights by this scheme than when two

[9] An important question arises here: is the ESL test oriented to rare pairs of words?

non relevant words co-occur in a document. In this manner the synonymous words result with lower similarity than non-synonymous ones.

The important conclusion to draw from these results is (1) that the cosine of conditional probability provides a powerful measure to detect near synonyms within single texts, (2) that global similarity approach is necessary to improve the results when the noun space is not enough representative and (3) that local attributional similarity proves to lead to improved results compared to the classical global attributional similarity approach.

7 Conclusions

According to [Terra and Clarke 2003] 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 texts. On the other hand, [Turney 2001] shows that synonyms co-occur in texts more often than by chance. In this paper, we proposed an approach 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 work presented here, despite the very short test - 23 cases, encourages us to perform larger scale evaluation and experiments in Word Sense Induction and disambiguation.

References

[Banerjee and Pedersen 2003] Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the 18th International Joint Conference on Artificial Intelligence. 805-810.

[Curran and Moens 2002] James R. Curran and Marc Moens. 2002. Improvements in Automatic Thesaurus Extraction. In Proceedings of the Workshop of the Special Interest Group on the Lexicon (SIGLEX) in collaboration with the 40th Conference of Computational Linguistics. 59-66.

[Dias et al. 2006] Dias, G., Santos, C., and Cleuziou, G. (2006). Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining. In Proceedings of the Workshop on Information Extraction Beyond the Document associated to the Joint Conference of the International Committee of Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006). Sydney, Australia, July 22. pp. 36-47. ISBN: 1-932432-74-4.

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

[Freitag et al. 2005] 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 9th Conference on Computational Natural Language Learning, Ann Arbor, Michigan, USA. 25-31.

[Gale et al. 1992] William A. Gale, Kenneth W. Church and David Yarowsky. One sense per Discourse. 1992. In Proceedings of the Workshop on Speech and Natural Language of the Human Language Technology, New York, USA. 233-237.

[Grefenstette 1993] Gregory Grefenstette. 1993. Automatic thesaurus generation from raw text using knowledge-poor techniques. In Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research.

[Jarmasz and Szpakowic 2003] Mario Jarmasz and stan Szpakowic. 2003. Roget's Thesaurus and Semantic Similarity. In Proceedings of Recent Advances of Natural Language Processing, Borovets, Bulgaria. 212-219.

[Kilgarriff and Yallop 2000] Adam Kilgarriff and Collin Yallop. 2000. What's in a thesaurus? In Proceedings of the Second Conference on Language Resource an Evaluation. 1371-1379.

[Landauer and Dumais 1997] Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104(2):211-240.

[Lin 1998] Dekang Lin. 1998. An Information-theoretic Definition of Similarity. In Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, USA. 296-304.

[Liu, 2004] Hugo Liu. 2004. MontyLingua: An end-to-end natural language processor with common sense. Available at: http://web.media.mit.edu/ hugo/montylingua

[Medin et al. 1990] Douglas L. Medin, Robert L. Goldstone and Dedre Gentner. 1990. Similarity involving attributes and relations: judgments of similarity and differences are not inverses. Psychological Science, 1(1):64-69.

[Rapp 2004] Reinhard Rapp. 2004. Utilizing the One-Sense-per-Discourse Constraint for Fully Unsupervised Word Sense Induction and Disambiguation. In Proceedings of the Forth Language Resources and Evaluation Conference,LREC 2004.

[Sahlgren and Karlgren 2001] Magnus Sahlgren and J. Karlgren. 2001. Vector-Based Semantic Analysis Using Random Indexing for Cross-lingual Query Expansion. In Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum. 169-176.

[Sahlgren 2006] Magnus Sahlgren. 2006. Towards pertinent evaluation methodologies for word-space models. In Proceedings of the 5th International Conference on Language Resources and Evaluation.

[Sugawara et al. 1985] K.M. Sugawara, K. Nishimura, K. Toshioka, M. Okachi and T. Kaneko. 1985. Isolated word recognition using hidden markov models. In Proceedings of the ICASSP-1985. 1-4.

[Terra and Clarke 2003] Egidio Terra and Charlie L.A. Clarke. 2003. Frequency Estimates for Statistical Word Similarity Measures. In Proceedings of Human Language Technology North American Association for Computational Linguistics. 165-172.

[Turney 2001] Peter. D. Turney. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the12fth European Conference on Machine Learning. 491-502.

[Turney et al. 2003] Peter. D. Turney, Michael L. Littman, Jeffrey Bigham and Victor Shnayder. 2003. Combining Independent Modules in Lexical Multiple-Choice Problems. In Recent Advances in Natural Language Processing III:Selected Papers from RANLP 2003. John Benjamins. 101-110.

[Turney 2006] Peter. D. Turney. 2006. Similarity of Semantic Relations. Computational Linguistics, 32(3):379-416.

[Weeds et al. 2004] Julie Weeds, David Weir and Diana. McCarthy. 2004. Characterising measures of lexical distributional similarity. In Proceedings of COLING 2004.

[Yang and Powers 2006] Dongqiang Yang, David M.W. Powers. 2006. Distributional Similarity in the Varied Order of Syntactic Spaces. In Proceedings of the First International Conference on Innovative Computing, Information and Control. 406-409.