Natural Language Engineering **16** (4): 439–467. © Cambridge University Press 2010
doi:10.1017/S135132491000015X

Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis

GAËL DIAS[1], RUMEN MORALIYSKI[1], JOÃO CORDEIRO[1], ANTOINE DOUCET[2] and HELENA AHONEN-MYKA[3]

[1]*Centre for HLT and Bioinformatics, Department of Computer Science, University of Beira Interior, 6201-001 - Covilhã, Portugal*
emails: ddg@di.ubi.pt, rumen@penhas.di.ubi.pt, jpaulo@di.ubi.pt
[2]*Campus Côte de Nacre, Boulevard du Maréchal Juin, University of Caen, BP 5186 - 14032 - Caen CEDEX, France*
email: doucet@info.unicaen.fr
[3]*Department of Computer Science, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), FI-00014, Helsinki, Finland*
email: helena.ahonen-myka@cs.helsinki.fi

(*Received 15 July 2009; revised 28 February 2010*)

Abstract

Thesauri, which list the most salient semantic relations between words, have mostly been compiled manually. Therefore, the inclusion of an entry depends on the subjective decision of the lexicographer. As a consequence, those resources are usually incomplete. In this paper, we propose an unsupervised methodology to automatically discover pairs of semantically related words by highlighting their local environment and evaluating their semantic similarity in local and global semantic spaces. This proposal differs from all other research presented so far as it tries to take the best of two different methodologies, i.e. semantic space models and information extraction models. In particular, it can be applied to extract close semantic relations, it limits the search space to few, highly probable options and it is unsupervised.

1 Introduction

Thesauri, which list the most salient semantic relations between words, have mostly been compiled manually. Therefore, the inclusion of an entry depends on the subjective decision of the lexicographer. Unfortunately, those resources are incomplete. Indeed, thesauri unlikely include syntagmatic semantic relations.[1] Levin (1993) is certainly the most comprehensive effort, to date, to categorize the verb part of the vocabulary with respect to the kind of constructions a word can participate in. Consider the following simple sentence: *The words of a phrase relate in many ways*

[1] In this sense, syntagmatic relations are the various ways in which words within the same sentence may be related to each other.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

to each other. Probably only the pair <*word, phrase*> would be listed in a manual resource with its semantic relation, but interpretation clues for a polysemous word like *way* would be more difficult to code in a thesaurus. In text understanding, humans are capable, up to a variable extent, of uncovering those relations. Natural language processing systems, however, need either a complete inventory of the semantic relations or a module able to infer them from the text in order to perform human-like interpretation.

Numerous attempts in automatic thesaurus construction are known (Grefenstette 1994; Lin 1998a; Curran and Moens 2002). The entries they extract comprise long lists of terms related to the head in unspecified ways. An attempt to partially annotate such thesauri with semantic information following the distributional lexical semantics paradigm is the work described in Lin *et al.* (2003). Apparently, applying various classifiers and filters consecutively improves the precision but at cost of recall. Other works, make use of exhaustive search over the vocabulary to induce semantic relations (Heylen *et al.* 2008).

The exhaustive search is the obvious way to verify all the possible connections between words of the vocabulary. However, comparison based on word usage can only highlight those terms that are *highly* similar in meaning. This method of representation is usually unable to distinguish between *middle strength* and *weak* semantic relations (Rubenstein and Goodenough 1965). Thus, the relative success of the Vector Space Model paradigm on synonymy tests (Landauer and Dumais 1997; Ehlert 2003; Rapp 2003; Terra and Clarke 2003; Turney *et al.* 2003; Jarmasz and Szpakowicz 2004; Freitag *et al.* 2005; Sahlgren 2006)[2] is due to the tests' structure, i.e. a pair of unequivocally synonymous words and a small set of mostly unrelated decoys. As a matter of fact, the results on synonymy tests depend very much on the number of the candidates among which choice has to be made. We conducted a simple experiment with a set of 1,000 random test cases, created in the manner described in Freitag *et al.* (2005), with up to ten decoy words and one synonym. We then solved the test cases using a contextual similarity measure. In particular, we used the Cosine similarity measure with features weighted by the Pointwise Mutual Information (see Section 4). The increase of the number of decoys caused a rapid drop of the probability to rank first the synonym as shown in Figure 1.

Thus, the exhaustive search is only capable of finding the most salient semantic relations, the ones that are established in the language and are frequent enough to be well represented, the ones that are usually included in the manually built thesauri as well. At the same time, neologisms, recently adopted foreign words and names which consist that part of the current vocabulary that needs constant update, elude characterization since they are not always well established and represented by written media.

To overcome difficulties encountered by exhaustive global search in semantic spaces, some works propose to exploit local patterns to extract hyponymy (Hearst 1992; Caraballo 1999; Snow, Jurafsky and Ng 2005, 2006), synonymy (Bollegala, Matsuo and Ishizuka 2007), or meronymy relations (Berland and Charniak 1999). Most of these studies, to the exception of Hearst (1992), Caraballo (1999) and

[2] to name but a few.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

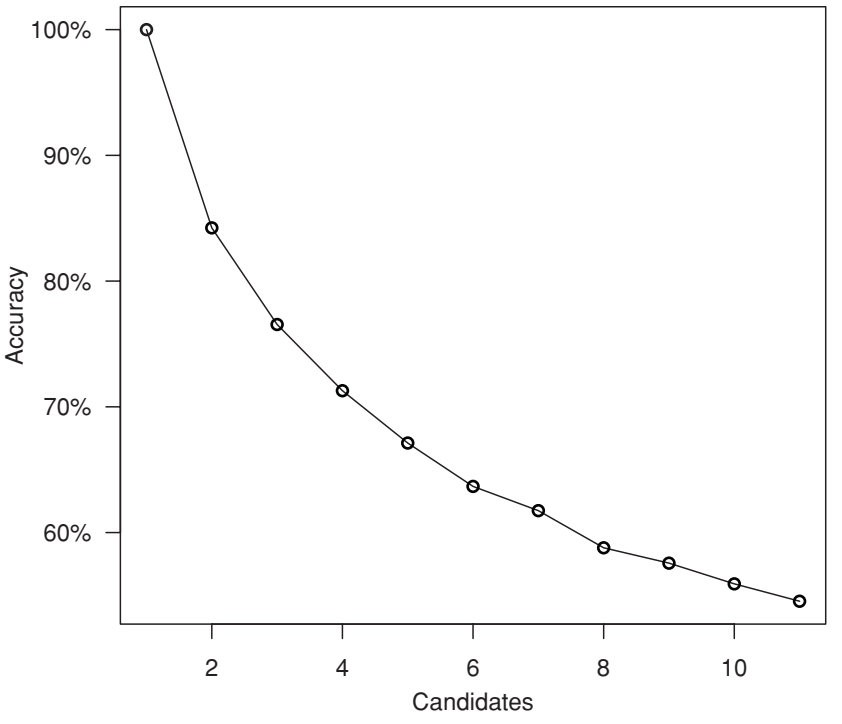


Fig. 1. Accuracy by candidates count.

Berland and Charnaik (1999), who study manual patterns, propose to automatically acquire relevant local patterns based on supervised learning to leverage manual work. Although, these approaches present successful results, they do not avoid critical issues. Unlike the distributional approach, information extraction solutions suffer from supervision as training data are necessary thus narrowing their application to one language and one exclusive semantic relation.

The problem is still worse when the relation between a pair of words is established not through their most common meaning but by some uncommon usage or they have related meanings only within a certain domain. For example, the words ⟨*node*, *vertex*⟩ are near-synonyms only in the context of graph theory, while the pair ⟨*spring*, *leap*⟩ is synonymous only through the fourth most frequent meaning of *spring* in WordNet (Fellbaum 1998). Discovering those relations is a very difficult task for purely statistical methods (Bordag 2003) as the rare events are obscured by the most frequent ones.

In order to discover pairs of semantically related words that may be used in figurative or rare sense, we need to have them highlighted by their environment as in the information extraction strategy and evaluate their semantic similarity by looking at their local and global distributional representations. Actually, we hypothesize that if semantically similar words have similar contexts then words that have similar contexts are semantically similar as well. The method we propose in the following sections aims at creating Test of English as a Foreign Language (TOEFL) like tests of one target word plus an as short as possible list of words that are

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

predominantly in paradigmatic relations with the target.[3] Eventually, a candidate word may be interchangeable with the target word in context. To this end, we propose to align paraphrases, automatically crawled from the web, and discover words that are possibly substitutable for one another in context. Then, we introduce a contextual similarity measure and an exhaustive characterization of the ability of various geometric and probabilistic models to highlight close semantic relations.

This proposal differs from all other research presented so far as it tries to take the best of two different methodologies, i.e. semantic space models and information extraction models. In particular, it is language independent, it can be applied to extract different semantic relations, it extracts relations between unfrequent word senses, it limits the search space and it is completely unsupervised.

This paper is divided into three main parts. In the first part of our argumentation, we review classical work on distributional semantics. In the second part, we explain how the TOEFL-like tests are created by crawling, identifying, clustering and aligning paraphrases. Then we propose a new similarity measure based on local and global contexts. Finally, we discuss our results.

2 Related work

Many earlier works (Hirschman, Grishman and Sager 1975; Hindle 1990; Grefenstette 1994) make attempt to discover related words in an automated manner. Usually, they employ some syntactic parser in order to reduce irrelevant statistical evidences and to minimize computational complexity. However, a common problem encountered is the size of the corpus available and for this reason they used to focus on restricted domains. As we extract our candidates from current news stories we are bound to use the Web as a corpus instead of some static collection (see Section 5).

Later, Lin (1998a) uses a broad-coverage parser to extract dependency triples from a 64-million-word text corpus in order to calculate word-to-word similarities and builds lists of similar words. According to the proposed evaluation, the created resource is more similar to WordNet than Roget's thesaurus (Roget 1852). We use very similar syntactic information; however, our aim is to sift out very specific semantic relation.

An account in Curran and Moens (2002) compares a number of weighting schemas and finds that formulae with logarithm (e.g. Pointwise Mutual Information) build representations that are more sensitive to rare events, which are otherwise obscured by the frequent ones. Complementarily, Weeds, Weir and McCarthy (2004) make a similar categorization with respect to the similarity measures.

A more recent work (Heylen *et al.* 2008) shows that the behavior of contextual similarity measures depends on frequency but as well on semantic specificity and semantic classes of words. Although strong conclusions cannot be drawn, since

[3] Paradigmatic are those relations that factorize a set in classes of interchangeable units. For example, the parts of speech classes are paradigmatic classes and the substitution of a word in a sentence for another one from the same syntactic class does not change the syntactic correctness of the sentence. Put into another way, we want to automatically discover paradigmatic classes of semantically related words – synonyms, hypernyms, hyponyms.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

comparison with the corresponding WordNet quantities is missing, it is still apparent that contextual similarity measures have a tendency to detect semantic relations beyond mere synonymy.

The thesaurus entries, extracted by the methodologies described so far, comprise long lists of terms related to the head in unspecified ways. Here emerge two diverging families of work. One is established by Landauer and Dumais (1997) with the introduction of the synonymy part of the TOEFL as an evaluation problem for synonymy discovery techniques. Although this practice has been repeatedly criticized, it affords straightforward results comparison and certain baselines such as random guesser, 25 per cent, and non-native speaker, 64.5 per cent, performance. Along this direction go as well the following works: Sahlgren and Karlgren (2002), Ehlert (2003), Jarmasz and Szpakowicz (2004), Terra and Clarke (2003), Rapp (2004) and Freitag *et al.* (2005).

The other direction takes a more general view and aims at automatically annotating existing lexicons with semantic information or at building, from scratch, resources listing only pairs of words in some specific semantic relation. Hearst (1992) describes the first work to find specific semantic relations from free text based on patterns instead of on a distributional approach. A set of words in specific relations are manually selected from a lexicon. Then, the words are sought to be found in syntactic proximity in a corpus. From those instances they manually select patterns that convey the desired semantic relations and subsequently more pairs of words that fit are gathered. This methodology works best for the *Is-a* relation but Berland and Charniak (1999), by coupling patterns with probability and confidence estimates, achieve reasonable results for meronymy discovery as well.

One of the hardest challenges to the semantic relation extraction is that the patterns are ambiguous and only few of the correct relations might be found in text expressed overtly by them (Cederberg and Widdows 2003). Thus, the system described in Caraballo (1999) first builds an unlabeled hierarchy of nouns using agglomerative bottom-up clustering of vectors of noun coordination information. To each node at upper levels are assigned hypernyms with the assistance of the lexico-syntactic patterns from Hearst (1992), according to a vote of the subsumed nodes. Similarly, Lin *et al.* (2003) look for pairs of contextually similar words that fit in specific patterns in order to highlight possible synonyms.

The main drawbacks of the methodologies proposed so far is the semi-supervision as manual work is necessary. In order to overcome this limitation Snow *et al.* (2005) use machine-learning techniques to automatically replace hand-built knowledge. Given a training set of texts containing known hypernym pairs, their algorithm automatically extracts useful dependency paths and applies them to new corpora to identify novel pairs. However, the most interesting work is certainly proposed by Bollegala *et al.* (2007) who extract patterns in two steps. First, they find lexical relationships between synonymous pairs based on snippets counts and apply wildcards to generalize the acquired knowledge. Then, they apply a Support Vector Machine (SVM) classifier to determine whether a new pair shows a relation of synonymy or not, based on a feature vector of lexical relationships.

In order to overcome the difficulties outlined so far, we take the most of both strategies, i.e. the one looking at common patterns and the other one using

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

distributional semantics analysis. We achieve this in several steps. First, we extract similar sentences from parallel news stories. Since news tend to be very focused, the pairs of lexically similar sentences convey almost the same information, the differences usually being in the form of words substitutions for synonyms or hyponym/hypernym, word order changes, or an accentuation on different details. Thus, in the next step, we align the corresponding parts of the sentences and the parts that differ are where we look for different semantic usage of words. From those parts we create TOEFL-like test cases that we then solve by some contextual similarity measure. In the following sections, we give motivation, technical details and evaluation of this process.

3 Creating test cases

The Distributional Hypothesis, formulated in Harris (1968), suggests that counting the contexts that two words share improves the chance for correct guessing whether they express the same meaning or not. The plausibility of this assumption is supported by the psycho-linguistic research (Kaplan 1950; Rubenstein and Goodenough 1965; Charles 2000) and by numerous empirical studies as well (see Section 2).

However, the words prove to be rather promiscuous with respect to the semantic frames in which they can fit. This specific behavior has primary origin in polysemy, the capacity of the words to have more than one meaning, and in the creative use of language. Locally, a single context may not be enough to select a word's sense. Rather, the semantic relations between the words within a sentence and discourse select their meanings (Kaplan 1950). Following the same idea, Charles (2000) collected a number of sentences, removed a word from each of them and asked two groups of human subjects to recover the missing words when presented with a list of sentences and with a list of words taken from the same sentences. He observed that sentences impose stronger lexical preference than disconnected words, and thus were more reliable evidence for measuring semantic similarity of pairs of words.

Therefore, in this work, we aim at finding pairs of sentences in which one word is substituted by another one and then to make confident decisions whether both words share meanings or not. Detecting of paraphrases provides an elegant solution to the first part of the problem. Paraphrases are sentences sharing an essential idea while written in different ways. As such, from paraphrases, we hope to learn TOEFL-like tests, i.e. clusters of words where there is a target word and a *short* list of semantically related candidates, predominantly in paradigmatic relations with the target.

In this section, we propose an unsupervised, language-independent methodology to automatically extract, cluster and align paraphrases which will help creating the test cases in an automatic way.

3.1 Paraphrase extraction

Paraphrase is a restatement of a text or passage, using other words. This is often accomplished by replacing words with their synonyms, hyponyms or hypernyms

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

1. *Kazakhs are outraged by the wildly anticipated mock documentary featur Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*
2. *The news follows controversy surrounding the comedy film Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan which cut so close to the funny bone.*
3. *Meanwhile Borat is leaping to the big screen in the mockumentary Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan.*

Fig. 2. A sample set of three paraphrases.

and changing word order. For example, the sentences in Figure 2, taken from web news stories excerpts, are paraphrases of a news about the release of a comic movie and show that '*feature*' can be substituted by *'news'*, *'controversy'*, *'comedy'* or *'film'* and as such may share common meanings. As a consequence, the extraction of paraphrases can lead to the identification of semantically related words in a microworld compared to a macroworld used by exhaustive search strategies which would seek for candidates in the entire vocabulary.

A few unsupervised metrics have been applied to automatic paraphrase identification and extraction (Barzilay and Lee 2003; Dolan, Quirk and Brockett 2004). However, these unsupervised methodologies show a major drawback by extracting quasi-exact or even exact match pairs of sentences as they rely on classical string similarity measures such as the Edit Distance in the case of Dolan *et al.* (2004) and Word N-gram Overlap for Barzilay and Lee (2003). For these functions, the more similar two strings are the more likely they will be classified as paraphrases. At the extreme, the 'best' pair will be precisely two exactly equal strings. This is clearly naïve and we may state that the more similar two strings are the poorer will be the paraphrase quality they generate. It is desirable to identify paraphrases which have certain level of dissimilarity, because this is precisely what will open room for semantic relation discovery.

The Edit Distance is rather problematic for paraphrase identification as true paraphrase sentence pairs having a considerable amount of word reordering due to distinct syntactic structures are likely to be considered as non-paraphrases. For example, the sentences

(1) *Due to high energy prices, our GDP may continue to fall, said Prime Minister, early morning.*
(2) *Early morning, Prime Minister said that our GDP may continue to fall, due to growing energy values.*

are in fact paraphrases; however, the Edit Distance by returning a high value would indicate a great dissimilarity.

To overcome the difficulties faced by the existent functions, new paraphrase identification functions have been investigated by Cordeiro, Dias & Brazdil (2007b) such as Gaussian functions (see (1)), Parabolic functions (see (2)), Trigonometric

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

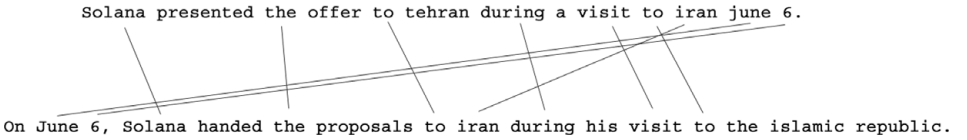


Fig. 3. Exclusive links between a sentence pair.

functions (see (3)), Triangular functions (see (4)) and Entropy functions (see (5)).

$$f_{Gauss}(x) = ae^{-\frac{(x-b)^2}{2c^2}} \quad (1)$$

$$f_{Parabolic}(x) = 4x - 4x^2 \quad (2)$$

$$f_{Trigonometric}(x) = sin(\pi x) \quad (3)$$

$$f_{Triangular}(x) = 1 - 2 \times |x - 0.5| \quad (4)$$

$$f_{Entropy}(x) = -xlog_2(x) - (1-x)log_2(1-x) \quad (5)$$

The x value represents some connection feature value, counted between the candidate paraphrase sentences, like, for example, word n-gram overlaps. In Cordeiro *et al.* (2007b), this value is calculated based on lexical exclusive links counts. For a given sentence pair, an exclusive link is a connection between two equal words from each sentence. When such a link holds then each word becomes bound and cannot be linked to any other word. This is illustrated in figure 3, where, for example, the determinant *'the'* in the first sentence has only one link to the first occurrence of *'the'* in the second sentence and the second occurrence of *'the'* remains unconnected.

In (1)–(5) x is defined as $x = \sqrt{\frac{\lambda}{m} * \frac{\lambda}{n}}$, where the number of exclusive links binding two sentences is represented by λ, m is the number of words in the longer sentence and n the number from the shorter one. Unlike the classical functions, these ones share the common characteristic of having a hill-shaped graph curve, with zero or near zero values near the domain boundaries and a maximum value reached in between. The important property of this type of hill functions is not the exact form of how they are calculated but the general shape of their graphs. These graphs convey a common meaning, since the maximum value is reached strictly inside the $[0, 1]$ interval, in some cases near the 0.5 value, which means, on one hand, that a certain degree of dissimilarity between the paraphrase sentences is 'desirable' and, on the other hand, that either the excessive dissimilarity or similarity tend to be penalized as we have the same property of zero approximation, on their boundaries, i.e: $\lim_{x\to 0} f_{hill}(x) = 0$ and $\lim_{x\to 1} f_{hill}(x) = 0$ (see Figure 4).

The main difference with the classical paraphrase detection functions (e.g. Edit Distance and Word N-gram Overlap) is that these latter have $\lim_{x\to 1} f(x) = 1$. An example of a paraphrase that would have high value with classical functions and low value for f_{hill} functions, is shown in Figure 5. From the Distributional Hypothesis standpoint, this example contains obviously very low utility.

The results of Cordeiro *et al.* (2007b) suggest that the hill-shaped functions ((1)–(5)) perform better than the classical ones and better paraphrases were extracted. It is clear, however, that with the current techniques paraphrase extraction is a

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

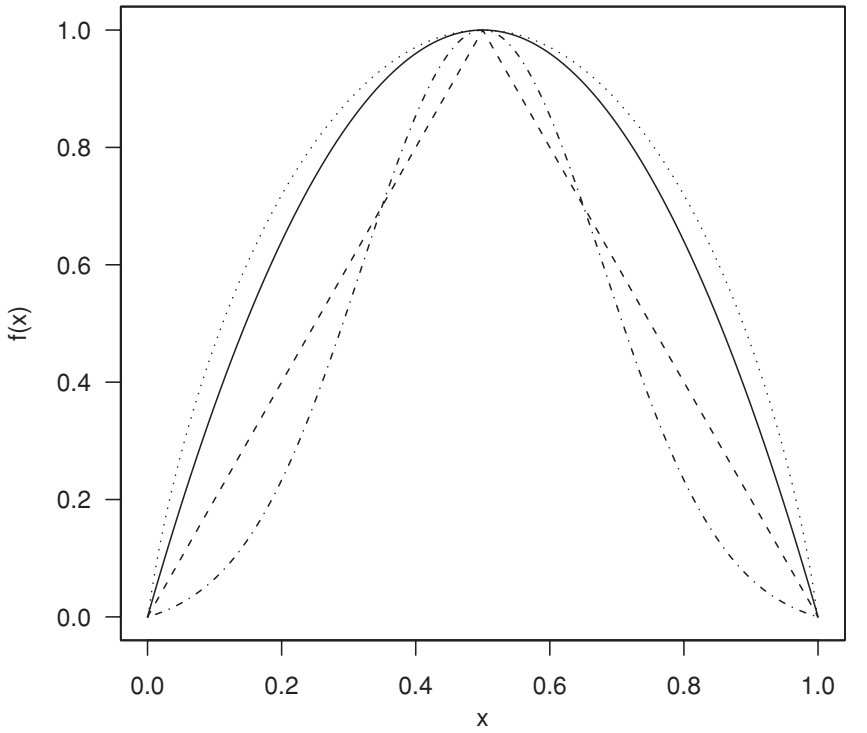


Fig. 4. *Hill shape* functions for paraphrase identification.

1. *The stock has gained 9.6 percent this year.*
2. *The stock has gained 9.6% this year.*

Fig. 5. A too similar paraphrase example.

difficult problem. In order to accomplish this task in a more dependable way (Jing and McKeown 2000; Dolan *et al.* 2004) and others make use of parallel or aligned monolingual corpus. Following those works, Cordeiro *et al.* (2007b) evaluate a number of metrics on extraction of paraphrases from clusters of news stories. Indeed, clustering of complete stories is more robust than extracting just pairs of lexically similar sentences since it relies on more statistical evidence. Thus, extracting paraphrases from stories that are already known to deal with the same subject improves the probability that lexically similar sentences have the same focus.

Within the scope of this work, Cordeiro *et al.* (2007b) proposed another function, with similar characteristics as the functions in (1)–(5), but performing even better than any other one in most of the standard corpora (Dolan *et al.* 2004; Cordeiro *et al.* 2007b). This function is called the *Sumo-Metric*, and for a given sentence pair, where each sentence has m and n words, respectively, and with λ exclusive links between the sentences, the *Sumo-Metric* is defined as in (6)

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

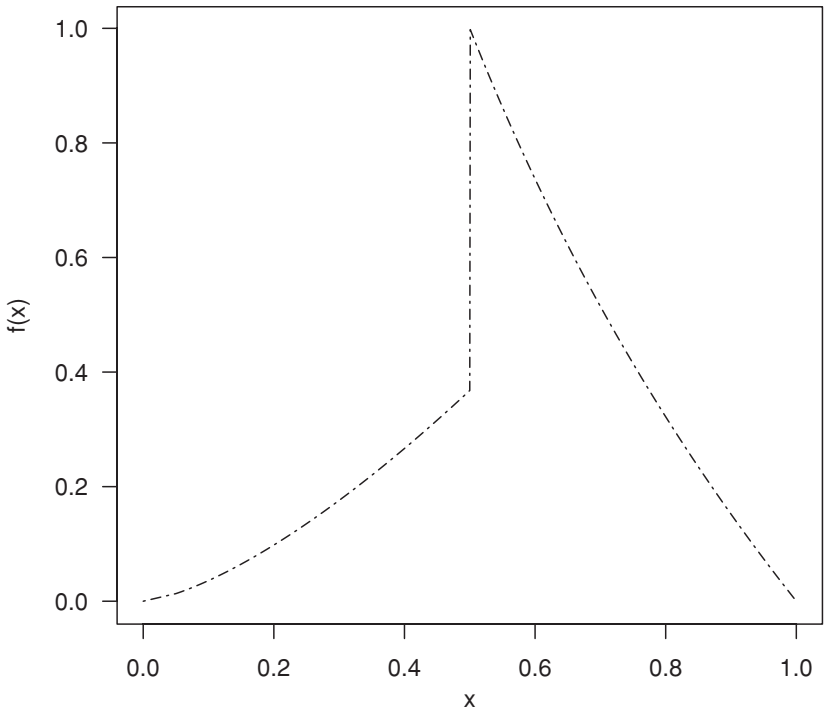


Fig. 6. *Sumo-Metric* function for paraphrase identification.

and (7).

$$S(S_a, S_b) = \begin{cases} S(m,n,\lambda) & if \;\; S(m,n,\lambda) < 1.0 \\ 0 & if \;\; \lambda = 0 \\ e^{-k*S(m,n,\lambda)} & otherwise \end{cases} \tag{6}$$

where

$$S(m,n,\lambda) = \alpha \log_2 \left(\frac{m}{\lambda}\right) + \beta \log_2 \left(\frac{n}{\lambda}\right), \alpha, \beta \in [0,1], \alpha + \beta = 1 \tag{7}$$

In particular, (Cordeiro *et al.* 2007b) show that the *Sumo-Metric* outperforms all state-of-the-art functions, over the tested corpora and allows to identify similar sentences with high probability to be paraphrases by defining a noncontinuous function of paraphrase similarity as shown in Figure 6 compared to the hill-shape-curve functions.

3.2 Paraphrase clustering

Literature shows that there are two main reasons to apply clustering for paraphrase extraction. On the one hand, as Barzilay and Lee (2003) evidence, clusters of paraphrases can lead to better learning of text-to-text rewriting rules compared to just pairs of paraphrases. On the other hand, clustering algorithms may lead to better performance than stand-alone similarity measures as they may take

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

advantage of the different structures of sentences in the cluster to detect new similar sentences.

So, instead of extracting only sentence pairs from corpora, one may consider the extraction of paraphrase clusters. There are many well-known clustering algorithms, which may be applied to a corpus of sentences $S = \{s_1, \ldots, s_n\}$. Clustering implies the definition of a similarity or (distance) matrix $A_{n \times n}$, where each element a_{ij} is the similarity (distance) between sentences s_i and s_j. In our context, the similarity measure is the *Sumo-Metric* between two sentences extracted from automatically crawled news stories.

The conclusion from Cordeiro, Dias and Brazdil (2007a) is that clustering tends to achieve worse results than simple paraphrase pair extraction, in terms of precision. In their work, a cluster of sentences is evaluated as a correct one, if and only if each pair of sentences in the cluster is a correct paraphrase pair. The baseline they adopted was to simply extract all sentence pairs through a similarity function $\sigma(.,.)$ (e.g. Edit Distance, Word N-gram Overlap, *Sumo-Metric*) under the condition that the similarity between sentences exceeded a given threshold value ϵ, which meant that two sentences, s_i and s_j, were considered as a paraphrase pair, if and only if $\sigma(s_i, s_j) > \epsilon$. In this case, a paraphrase pair may be viewed as a 'trivial' paraphrase cluster with only two sentences.

However, among the clustering algorithms experimented by Cordeiro *et al.* (2007a) for sentence clustering, the *Quality Threshold* (QT) algorithm (Heyer, Kruglyak and Yooseph 1999) achieved better precision, 64 per cent, than other clustering algorithms, 57 per cent, that do not need in advance the expected number of clusters. Therefore, in this work we decided to apply the QT clustering algorithm to the similarity matrix based on the *Sumo-Metric* to obtain clusters of paraphrases. The QT algorithm was originally conceived to tackle the problem of gene clustering and despite the fact that it requires more computational power than other clustering algorithms, like hierarchical clustering, it enables more control directed toward the specific problem. For the particular task of paraphrase clustering, Cordeiro *et al.* (2007a) manually selected number of true paraphrases and random sentence pairs and empirically established the optimal threshold at 0.2. Since *Sumo-Metric* takes values in the [0, 1] interval, this means that in each generated cluster, and for each sentence pair in that cluster, we will have $S(s_i, s_j) > 0.8$.

QT creates the largest possible nonoverlapping clusters. It was designed to avoid a sometimes inadequate behavior of other clustering algorithms, K-means for example, that do not take into account whether a unit of a set to be clustered possibly belongs to any cluster, but force each unit into the nearest one. This property of QT coupled with the *Sumo-Metric* avoids grouping together very dissimilar sentences, but it avoids also grouping together sentences that are too similar and leaves out from the cluster the sentences that do not contribute new information. Thus QT with *Sumo-Metric* naturally deal with possible redundant processing.

Finally, once paraphrase clusters have been extracted, we need to align the sentences in the clusters in order to extract lists of interchangeable words to be able to create TOEFL-like test cases in an automatic way. For that purpose, we

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

implement the methodology of Doucet and Ahonen-Myka (2006) who propose to extract Maximal Frequent Sequences.

3.3 Aligning paraphrases

In this section, our goal is to align the paraphrases inside each cluster, detecting their common parts so as to evidence what differentiates them. Our approach considers sentences as word sequences and therefore reduces the resulting problem to that of multiple sequence alignment (Notredame 2007). In the field of bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA or proteins to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations, for instance. The alignment of sequences is performed to evidence their common and distinctive parts, possibly taking gaps into account.

Similarly, in the field of natural language processing, sequence alignment allows to observe variations in language use, and is particularly useful for similar text fragments, such as paraphrases (Barzilay and Lee 2003). But while there are several efficient techniques for multiple sequence alignment in the field of bioinformatics, they actually aim at slightly different problems. Indeed, biosequences to be aligned are typically few, very long, and with limited vocabulary (e.g. there are only twenty amino acids, and only four nitrogenous bases present in the nucleic acids DNA and RNA, designated by the letters A, C, G and T). In comparison, paraphrases are more numerous, shorter, with a larger vocabulary and very few words are repeated within the same sentence. As a consequence, the techniques optimized for biosequences alignment turn out to be inappropriate for paraphrases (Barzilay and Lee 2003).

In this paper, we present a two-phase approach to efficiently align a set of paraphrases. In the first phase, we extract the Maximal Frequent Sequence set (MFS) of the paraphrases, that we later use as a pivot for the one-pass alignment of the paraphrases.

3.3.1 Maximal frequent sequences

Maximal Frequent Sequences were defined by Ahonen-Myka (1999). A frequent sequence is defined as a noncontiguous sequence of words that must occur in the same order more often than a given sentence-frequency threshold. MFSs are constructed by expanding a frequent sequence to the point where the frequency drops below the threshold. This expansion is done through a greedy algorithm extensively described in the aforementioned reference (Ahonen-Myka 1999). It is worth to note that this technique does not require any preprocessing. For instance, neither stemming nor stopword removal are necessary. This way, we can assign a set of MFSs to each set of paraphrases. In the following of this section, we formally define the notions of MFS.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Definition 1
A sequence $p = a_1 \cdots a_k$ is a *subsequence* of a sequence q if all the items $a_i, 1 \leq i \leq k$, occur in q and they occur in the same order as in p. If p is a subsequence of q, we also say that p *occurs* in q and that q is a *supersequence* of p.

For instance, the sequence *'Glorious Nation of Kazakhstan'* can be found in all of the three sentences in Figure 2 and as such is a subsequence of each sentence.

Definition 2
A sequence p is *frequent* in a set of fragments S if p is a subsequence of at least σ fragments of S, where σ is a given frequency threshold.

Definition 3
A sequence p is a *maximal frequent (sub)sequence* in a set of fragments S if there does not exist any sequence p' in S such that p is a subsequence of p' and p' is frequent in S.

As a consequence, in the example presented in Figure 2, the sequence *'Glorious Nation of Kazakhstan'* is not maximal, since it is a subsequence of the frequent sequence *'Borat: Cultural learnings of America for Make Benefit Glorious Nation of Kazakhstan'*. This latter sequence is maximal. With this simple example, we already get a glimpse of the compact descriptive power of MFSs. We do not restrict ourselves to the extraction of word pairs. Indeed, the twelve-word sequence *'Borat: Cultural learnings of America for Make Benefit Glorious Nation of Kazakhstan'* would need to be replaced by $\binom{12}{2} = 66$ word pairs. With MFSs, no restriction is put on the maximal length of the phrases. Thus, we can obtain a very compact representation of the regularities of texts. So, by extracting the MFSs of a cluster of paraphrases, we obtain a compact sequential description of the corresponding paraphrases, i.e. a 'skeleton' of the cluster that may be used for alignment.

3.3.2 Multiple sequence alignment

Given the corresponding set of MFSs, we can extract the commons and specifics of a set of sentences, very efficiently, in one pass. For instance, let us assume we are to align the three paraphrases in Figure 2.

This set contains one MFS of frequency 3: '*Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan*'. Once this MFS has been extracted, we can rely on the sequential property of MFSs to know that, passing in parallel through each of the paraphrases, we are bound to encounter the word '*Borat:*', and that any word encountered before is not common to all sentences. Once '*Borat:*' is encountered, we know that we are bound to encounter the second word of the MFS, '*Cultural*'. Since the MFS allows gaps, any word encountered between two successive words of the MFS is not common to all sentences. So, the process continues until the last word of the MFS is reached. This way, in only one pass over the word sequences, we can obtain the resulting alignment presented in Figure 7.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

[{1:*Kazakhs are outraged by the wildly anticipated mock documentary feature*} {2:*The news follows controversy surrounding the comedy film*} {3:*Meanwhile Borat is leaping to the big screen in the mockumentary*}] Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan [{2:*which cut so close to the funny bone*}]

Fig. 7. The alignment corresponding to the sentences of Figure 2. Word sequences without brackets are common to both sentences. The sequences in curly brackets are specific to the sentences with the corresponding numbers.

1. *kazakh* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*
2. *feature* | *news* | *controversy* | *film* | *borat* | *screen* | *mockumentary*

Fig. 8. Two TOEFL-like test cases.

3.4 Forming the test cases

The final step is to form TOEFL-like test cases from the aligned segments in the clusters. The notion of test implies one word in a specific position, or target word, for which we are searching matches among a list of candidates. In this section, we show how to create tests that consist of words with high probability of being in a paradigmatic relation.

So far, we have clusters of sentences saying nearly the same thing, but slightly differing in expression and with the corresponding parts aligned. We now need to search for candidates among the words which appear out of the MFS, the different parts of the paraphrases, i.e. words in between the brackets in Figure 7.

In order to extract lists of interchangeable words, we first lemmatize and assign part-of-speech tags to the aligned paraphrases with MontyLingua (Liu 2004). This step is necessary since we are interested in nominal semantic relations and only open class words with the same part-of-speech are eligible candidates.

Those parts of the paraphrases that lie between two successive parts of a MFS have different orthographic appearance; nevertheless, we assume that they have similar meanings since they are both parts of paraphrase sentences and share left and right MFS contexts. Therefore, here is the place where we search for word substitutions. Precisely, the construction of the candidate tests goes like the following algorithm:

```
For each aligned sub-segment
    For each open class word
        Create a list of candidates from
        the rest of the segments that share
        left and right MFS contexts.
    End
End
```

For example, from the words from the first aligned paraphrase in Figure 7 we extract two test cases for the target words '*kazakh*' and '*feature*' as shown in Figure 8. Six more test cases would be extracted from this paraphrase cluster for the nouns '*news*', '*controversy*', '*film*', '*Borat*', '*screen*' and '*mockumentary*'.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

In this first part, we propose to create TOEFL-like tests with a short list of candidates that are predominantly in paradigmatic relations with the target word. Eventually, a candidate word can be interchangeable with the target word in context. In particular, this methodology is language independent and completely unsupervised which may allow the study of different languages. In the remainder of this paper, we propose an exhaustive study of local and global similarity measures over the Vector Space Model and Probabilistic Models to identify the semantic relations between the words inside the TOEFL-like tests.

4 Measuring similarity between words

Now as we have the lists of candidate words, as shown in Figure 8, we need a method to select the best candidate. As we outlined in Section 1, we followed the thoroughly studied hypothesis that words that share more contexts are more probable to be semantically similar. Now, we calculated the contextual similarities through the geometric metaphor of the Vector Space Model to solve the TOEFL-like tests.

In this context, we must evaluate the similarity between two nouns, which are represented by their respective word context vectors $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip})$ of observations on p variables (or attributes). The similarity between the two units i and j is defined as $S_{ij} = f(X_i, X_j)$, where f is some function of the observed values.

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 '*controversy*' appears with the verb '*surrounding*' within a subject relation, we will have the following triple[4] $\langle$*controversy, surround, subject*$\rangle$ and the tuple $\langle$*surround, subject*$\rangle$ will be an attribute of the word context vector associated to the noun '*controversy*'.

4.1 Weighting attributes

The simplest form of the vector space model treats a noun n as a vector which attribute values are the number of occurrences of n in the context of each of the tuples $\langle v, r \rangle$. However, not all the contexts are equally informative therefore numerous weighting schemas have been proposed. In this section, we will list the most common ones although many others could be used.

4.1.1 Word frequency and IDF

Inverse document frequency was introduced in order to weight index terms for Information Retrieval (IR) (Spärck-Jones 1972). In the context of the syntactic attribute similarity paradigm, we define it as in (8) where n is the target noun, $\langle v, r \rangle$

[4] due to lemmatization

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

a given attribute, N is the set of all the nouns and $|.|$ is the cardinal function.

$$tf.idf(n, \langle v, r \rangle) = tf(n, \langle v, r \rangle) \times \log_2 \frac{|N|}{|\{n_i \in N | \exists (n_i, v, r)\}|} \tag{8}$$

4.1.2 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, Turney (2001) and Terra and Clarke (2003) have proposed to use the Pointwise Mutual Information (PMI) as defined in (9) where n is the target noun and $\langle v, r \rangle$ a given attribute.

$$PMI(\langle n|r \rangle, \langle v|r \rangle) = \log_2 \frac{P(n, v|r)}{P(n|r)P(v|r)} \tag{9}$$

4.1.3 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 cooccurrence as in (10). In our case, we are interested in knowing how strongly a given attribute $\langle v, r \rangle$ may evoke the noun n.

$$P(n|v, r) = \frac{P(n, v, r)}{P(v, r)} \tag{10}$$

4.2 Similarity measures

Numerous similarity measures have been evaluated in Terra and Clarke (2003) and Weeds *et al.* (2004). 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.

4.2.1 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 (11).

$$cos(n_1, n_2) = \frac{\sum_{k=1}^{p} n_{1k} n_{2k}}{\sqrt{\sum_{k=1}^{p} {n_{1k}}^2} \sqrt{\sum_{k=1}^{p} {n_{2k}}^2}} \tag{11}$$

4.2.2 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.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Ehlert model: Equation (12) proposed in Ehlert (2003) evaluates the probability that the first word was changed for the second one.

$$Ehl(n_1|n_2) = \sum_{\langle v,r\rangle \in A} \frac{P(n_1|v,r)P(n_2|v,r)P(v,r)}{P(n_2)} \tag{12}$$

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

Lin model: Lin (1998b) 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 (13).

$$Lin(n_1,n_2) = \frac{2 \times \sum_{\langle v,r\rangle \in A} \log_2 P(v,r)}{\sum_{\langle v,r\rangle \in B} \log_2 P(v,r) + \sum_{\langle v,r\rangle \in C} \log_2 P(v,r)} \tag{13}$$

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

4.3 Global and local attributional similarity

The approaches reviewed so far 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, Freitag *et al.* (2005) found high correlation between polysemy and error level and conclude that polysemy level is characteristic of the difficulty of a test. Moreover, they suggest that the Global similarity measures bear affinity to less polysemous pairs.

According to Gale, Church and Yarowsky (1992) '*... 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. Turney (2001) also demonstrates that synonyms tend to cooccur in texts more often than by chance. A similar supposition is grounded in Landauer and Dumais (1997), who seek for synonyms among words that cooccur in the same set of documents.

As a consequence, we apply the '*one sense per discourse*' paradigm and compare nouns only within a single document based on the attributional similarity paradigm.

Apparently statistics gathered from a unique short text may not be reliable. As a consequence, in order to obtain more stable results, we average attributional similarity values over the set of documents in which two nouns occur and introduce the *Lsim*(., .) function in (14), where *sim*(., .) is any function from Section 4.2. Such

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

an approach uses similarity measures in a local (document) context and is called *Local similarity* (*Lsim*).

$$Lsim(n_1, n_2) = \frac{\sum_{d \in D} sim(n_1, n_2)}{|D|} \tag{14}$$

We claim that in most cases Local similarities compare statistical representations of word meanings as opposed to words and thus it is similar to the measures of concept-distance proposed in Agirre and Lacalle (2003) and Mohammad and Hirst (submitted). However, it is radically different from those approaches as we do not make use of any preexisting knowledge source in order to build the distributional representations of the concepts, rather we relay uniquely on an automatically acquired corpus.

In this paper, we 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 (15).

$$Psim(n_1, n_2) = Gsim(n_1, n_2)^{\gamma} \times Lsim(n_1, n_2)^{(1-\gamma)}, \gamma \in [0, 1] \tag{15}$$

In fact, (15) is a generalization of all similarity measures. Indeed, when $\gamma = 0$, only the Local similarity is taken into account while when $\gamma = 1$ only the Global similarity is applied.

5 Corpus

Any work based on the attributional similarity paradigm depends on the corpus used to determine the attributes and to calculate their values. Terra and Clarke (2003) use a terabyte of web data that contains 53 billion words and 77 million documents; Sahlgren and Karlgren (2002) – a 10 million words balanced corpus with a vocabulary of 94,000 words; and Freitag *et al.* (2005) and Ehlert (2003) – the 256 million words North American News Corpus (NANC). As mentioned in Ehlert (2003) and Terra and Clarke (2003), the bigger the corpus is, the better the results are. For our experiments, we used the Reuters Corpus Volume I (RCV1) (Lewis *et al.* 2004). However, our proposal to calculate local similarity requires cooccurrences of both candidates to appear a few times each within a single document, and we observed that a substantial proportion of word pairs have zero occurrence in RCV1. RCV1 consists of more than 800,000 stories produced by Reuters journalists between August 20, 1996 and August 19, 1997, while our paraphrases are extracted from news stories gathered during three days in November 2006. Apparently in 10 years the main players and subjects changed radically in dynamic genre as is the news.

As we did not want to reduce our test set, we decided to build a corpus suitable to the problem at hand. For this purpose we used the Google API and queried the search engine with set of different pairs of words. For each test case we built all the pairs that consist of the target word and one of the candidates. Subsequently, we collected all of the seed results and followed a set of selected links to gather more

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

textual information about the queried pairs. The overall collection of web pages was then shallow parsed using the MontyLingua software in order to extract the predicate triples as described in Section 4. Thus, the corpus consists of 500 million words in 110,000 documents in which each sentence is a predicate structure. The benefit of such a corpus is to maximize the ratio of the observed instances to the volume of the text processed.

6 Results and discussion

In this section, we propose to evaluate our methodology over a set of web news stories extracted automatically on a daily basis. This environment proves to be very fruitful for paraphrase extraction, since many sentences convey the same message but in a different form.

For this experiment, we gathered three days of news from the Google News website.[5] From these texts, 178,000 sentences were extracted as paraphrase candidates, which formed 27,000 clusters of sentences. Finally, 183,000 alignments were produced which, then, yielded a set of 22,000 TOEFL-like test cases with an average of 4.6 candidates.

6.1 Paraphrase extraction

The paraphrase extraction and clustering methodology that we adopted in this work was proposed and formally evaluated in Cordeiro *et al.* (2007a). Here we examine its properties from the viewpoint of the semantic relations detection problem.

Bad preprocessing, which means that HTML or XML tags were taken as tokens or partial sentences were extracted, accounted for the most part of wrong paraphrase classification. Indeed the *Sumo-Metric* is very optimistic with respect to the short sequences. For example, the pair

(1) *He is a superstar Texas senior cornerback Aaron Ross said.*
(2) *What he is doing now is being a great leader Texas coach Mack Brown said.*

was classified as a paraphrase based on the pronoun '*he*', the verb '*to be*', the name '*Texas*' and a very common citation frame in the news writing style which is '*said*'.[6]

Few wrong paraphrases could have been avoided with the assistance of named entity recognition or multiword unit extraction that would give a single count to a unique reference or concept as, for example, in the following example where '*kazakh authorities*' and '*legal action*' should be treated as single concepts.

(1) *Many months later the funny bruised fruits of his labor Borat: cultural learnings of America for make benefit glorious nation of Kazakhstan are poised to hit the collective american conscience with a juicy splat.*

[5] `http://news.google.com/`
[6] Other examples are *confirmed* or *said in a statement*.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

(2) *Borat: cultural learnings of America for make benefit glorious nation of Kazakhstan opens in the United States on Friday but the run in with kazakh authorities who even threatened legal action generated huge pre release publicity.*

This results in the following erroneous test *splat* | *action* | *authority* | *kazakh* | *publicity* | *release*.

Even a relatively big number of overlaps cannot guarantee that the pair of sentences have the same communication intent as shown below:

(1) *Luke broke onto the screen under Washington's direction in Antwone Fisher, then went on to Friday Night Lights and Glory Road.*
(2) *Luke, best known for his work in Antwone Fisher and Friday Night Lights, is the versatile and commanding young leading man Hollywood needs.*

Although this mode of paraphrasing might seem very productive, a much more common source of tests without any perceivable semantic relation are perfectly aligned paraphrases which make accent on different details such as in the following paraphrase:

(1) *Federline released his debut CD on October 31.*
(2) *Federline released his debut CD in which he raps about his rise from obscurity.*

Although the essential information is the same, deeper interpretation would be necessary as to make clear that the adverbial and the subordinate clauses are not subject to semantic alignment. This kind of paraphrasing was the major source of tests void of semantic relations.

However, the average sentence length in our news corpus is twenty-four words and most of them are correctly classified. This step alone is responsible for about 35 per cent of the wrong test cases or 23 per cent of all the tests.

6.2 Aligning paraphrases

The alignment phase is based on finding an as long as possible sequence that is common for both sentences. However, the alignment failed in number of cases when the paraphrasing effect is achieved through word order change. For example, although the sentences

[{1:*The median price of an existing single family home dropped* 2.5 *per cent from September 2005, the biggest year on year drop since record keeping began in 1969*}] the national association of realtors said [{1:*in Washington*} {2:*existing home sales declined for the sixth consecutive month in September while the median price fell 2.5 per cent year over year, the biggest decline on record*}]

are perfect paraphrases, the only possible alignment results in the test *washington* | *month* | *price* | *record* | *september* | *year* while from these sentences one could infer similarity between the nouns '*drop*' and '*decline*'. This is a common case when two long sentences are aligned around a single sequence that refers to the common

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

agent. Even when the alignment is anchored in many points there is still the option of conjunction rearrangements or even syntactic structure alterations such as the following alignment.

[{1:*He found*} {2:*This revealed*}] that [{2:*their*}] sperm [{1:*count, viability, motility*} {2:*declined steadily in number, quality*}] and [{1:*shape declined*} {2:*ability to swim*}] as mobile phone usage increased.

Paraphrase classification and alignment can occur based on secondary details as well. For example, the sentences

[{1:*The gloomy prediction follows*} {2:*Marine species are disappearing at an accelerating rate posing a serious threat to human health and wellbeing*}] a four year [{1:*multinational*}] study of the state of the world's [{1:*seas and*}] oceans [{2:*has concluded*}].

are paraphrases and correctly aligned; however, the essential information is not explicitly present in the first one. In order to avoid this kind of alignment and the consequent bad test cases, discourse analysis is necessary as well as reconstruction of the intended message by means of anaphora resolution.

6.3 TOEFL-like tests

In order to keep the evaluation manageable, we retained at random 1,000 clusters of sentences and from them extracted 1,058 noun test cases. Few clusters yielded more than one test. Then we manually classified them into five classes with respect to whether the test contained a pair of words in one of the following relations: *Synonymy*, *Siblings (Co-Hyponymy)*, *Is-a*, *Instance-of*. Otherwise, we labeled it as *None*. Only afterward we used the test set for evaluation as to avoid any influence of the automatic classification on the manual labeling.

In order to classify a test, we first disambiguated all the words in the contexts of the source paraphrase cluster and the original news story, when necessary. If there was a candidate that referred to the same concept as the target we subsequently classified the test with respect to the perceived relation. For example, the *Instance-of* relation *aisawa* | *legislator* was extracted from the following paraphrase.

(1) *Aisawa declined to elaborate.*
(2) *The japanese legislator declined to elaborate.*

In Table 1, we present the distribution of the tests per category. It is interesting to observe, that the *Synonymy* together with *Co-Hyponymy* are more populous than the other two categories. It is no surprise, though, that same level words are preferred substitutes for the sake of paraphrasing.

The reason to undertake the manual annotation is that news exhibit very creative use of language and rare synonymy relations like *leader – godfather* that are not present in WordNet or foreign words, for example, '*madrassa*' when narrating about religious school in Pakistan, appear regularly in the texts. Out of the 117 pairs that we found to be in synonymy roles only twenty-nine were present in WordNet as

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Table 1. *Classification of the test cases*

Synonyms	Siblings	Is-A	Instance Of	None	Overall
117	108	61	86	686	1058

Table 2. *Manually annotated tests. The respective relations hold between the first and the second words of each test*

Synonyms:	*body* | *panel*
	michael | *mike*
	administration | *government*
	condition | *disease* | *treatment*
	seat | *place* | *american* | *congress* | *election*
Siblings:	*idea* | *plan*
	amazon | *ebay*
	journalist | *videographer*
	blaze | *wildfire* | *santa*
	reality | *point* | *campaign*
Is-A:	*conspiracy* | *obstruction*
	capability | *repair*
	status | *fame* | *fortune*
	game | *play* | *room* | *sideline*
	allegation | *statement* | *admission* | *family* | *friday*
Instance Of:	*july* | *month*
	community | *un*
	patriot | *team* | *right*
	fedex | *company* | *order*
	schwarzenegger | *star* | *film* | *terminator*

such. An excerpt of the annotated tests is given in Table 2. They all contain a pair of words that could be regarded in a given semantic relation in a given context.

The manual annotation and disambiguation process was instructive as for the strength of the semantic relations. Although the *Siblings* in the *Is-a* hierarchy of WordNet are connected by longer paths, they tend to be perceived as more similar to each other than are the words in the *Is-a* relation. This subjective judgement seems to be confirmed by the persistently higher contextual similarity between the former compared to the latter category (see Section 6.4).

The figures in Table 3 show the distribution of the test cases over nine categories with respect to the number of the candidate words, thus the first column represents those cases, which come from paraphrase pairs in which only one noun is substituted by another one and so on and so forth.

A substantial part of the tests have four, five or six candidates. However, this does not indicate the most common mode of paraphrasing because these same sets contain the lowest ratio of paradigmatic semantic relations. It is natural to expect, that the more candidates a given test has, the higher the probability that any of them will be in any of the specified semantic relations with the target word. The tests with more candidates come from paraphrases with greater absolute number

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Table 3. *Proportion of good tests by test size*

	1	2	3	4	5	6	7	8	9
Good	41%	37%	36%	30%	31%	34%	41%	41%	35%
All	120	87	119	204	155	159	100	74	40

of differences. For such a pair to be taken as a paraphrase, it also needs a greater number of common subsequences, thus implying that more information is shared between the sentences. This is why, after the level of greatest linguistic variety in the middle of the specter, test extraction improves. In particular, we were able to only extract ten and six tests with ten and eleven candidates, respectively.

A certain portion of the tests, 65 per cent, do not belong to any of these semantic categories. Some of them are due to wrong alignments as in *understanding* | *Lipunga* | *onion* | *tomato* | *village*. About 25 per cent of the tests are void of any perceivable semantic relation.

Further, bad part-of-speech tagging caused a set of good candidates to be lost and replaced by words that were actually used in another part-of-speech role. For example, from the following aligned paraphrase that reports on infant disease

[{1:*right now the*}{2:*currently the brain*}] defects cannot be detected until after death

the test *right* | *brain* is extracted where '*right*' is indeed an adverb.

Finally, the rest of the tests in this group could be classified in some more loose semantic category such as in the following case: *study* | *caution* | *Washington* | *finding*. It is also important to notice that we did not encounter words in antonymy relation. Table 4 shows sets of extracted pairs with their corresponding categories.

6.4 Similarity measures

In order to quantify the feasibility of the methodology, we retained only the 372 test cases labeled with a specific semantic relation and performed a comparative study. For all the similarity measures and the respective weighting schemes (i.e. (1) the Cosine similarity measure associated to the Tf.Idf, the Pointwise Mutual Information or the Conditional Probability, and (2) the Lin and Ehlert models for the Conditional Probability), we solved each test using the Global (*Gsim*), Local (*Lsim*) and Product similarities (*Psim*). In particular, the γ parameter from Equation 15 was trained using well-known synonymy tests: the TOEFL (Landauer and Dumais 1997), the ESL (Turney *et al.* 2003), the Reader Digest (Turney *et al.* 2003) and the Freitag set (Freitag *et al.* 2005). The results are summarized in Tables 5, 6 and 7.

The first observation we can make from Table 5 is that the combination Cosine with PMI is nearly sufficient to extract the closest semantic relations. However, none of the Global measures achieves results different from random guessing for the category *Instance Of*. This is no surprise since, in order to be solved, most of the cases in this category reduce to a problem of finding the most salient property associated

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Table 4. *Candidate thesaurus relations*

Synonyms		Siblings	
administration	government	Amazon	eBay
agency	association	battle	race
CAT	CT	candidate	challenger
commander	gen	culture	habit
condition	disease	department	government
godfather	leader	draft	resolution
imagery	model	flaw	issue
madrassa	school	idea	plan
marine	navy	journalist	videographer
Michael	Mike	mother	wife
planning	policy	poll	survey
Rudolph	Rudy	report	review
shia	shiite	today	Monday
Is-A		**Instance Of**	
ability	breathing	agency	IAEA
agency	custom	Aisawa	legislator
agreement	deal	Bush	president
blaze	fire	community	UN
brain	tissue	company	FedEx
cancer	disease	country	Germany
case	lawsuit	Dec	month
cleric	sheik	group	Nirvana
conspiracy	obstruction	host	Winfrey
envoy	negotiator	Ortega	Sandinista
foe	opponent	Russia	state
force	marine	Schwarzenegger	star

Table 5. *Accuracy of Global on 372 tests*

	Lin (%)	Ehlert (%)	Cond Pr (%)	PMI (%)	TfIdf (%)
Synonyms	42	58	42	**75**	50
Siblings	**65**	29	53	**65**	47
Is-A	42	29	46	**58**	54
Instance Of	14	26	23	**30**	26
Overall	33	34	36	**49**	39

to a proper name (see Table 4). For example, the pair *president* - *Luiz* refers to '*Luiz Inácio Lula da Silva*'. However, '*Luiz*' is a common name and as such is very polysemous word. Here is where the Local comes to play. Since it always compares monosemous representations, it is bound to associate '*president*' with '*Luiz*' in those documents where the President of Brazil is the subject. As a result, the performance of the Local similarities show statistically significant improvements over the Global for the *Instance Of* test cases.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Table 6. *Accuracy of Local on 372 tests*

	Lin (%)	Ehlert (%)	Cond Pr (%)	PMI (%)	TfIdf (%)
Synonyms	54	58	**71**	50	58
Siblings	**59**	47	47	53	41
Is-A	38	42	42	42	**46**
Instance Of	40	42	35	40	**49**
Overall	43	46	46	44	**48**

Table 7. *Accuracy of Product on 372 tests*

	Lin (%)	Ehlert (%)	Cond Pr (%)	PMI (%)	TfIdf (%)
Synonyms	50	**63**	46	58	58
Siblings	**71**	41	41	59	64
Is-A	46	42	50	**58**	46
Instance Of	33	37	35	34	**42**
Overall	43	44	40	46	**48**

Table 8. *Best methodology by Category*

	Lin	Ehlert	Cond Pr	PMI	TfIdf
Synonyms	–	–	–	Global	–
Siblings	Product	–	–	–	–
Is-A	–	–	–	Global or Product	–
Instance Of	–	–	–	–	Local
Overall	–	–	–	Global	–

Moreover, the results evidence that a single measure cannot solve the entire problem. The *Synonym* relation is best treated by Global values, the *Instance Of* relation is best treated by Local values, while the Lin model deals best with the *Siblings* for the Product values. We summarize all these results by kind of semantic relation and measure in Table 8.

7 Conclusions and future work

In this paper, we presented an innovative approach for word semantic relation extraction. This proposal differs from all other research presented so far as it tries to take the best of two different methodologies, i.e. semantic space models and information extraction models. In particular, it is language independent, it can be applied to extract different semantic relations, it extracts relations between unfrequent word senses, it limits the search space and it is completely unsupervised.

To achieve this result, we first extract paraphrases from parallel news stories and cluster them into meaningful clusters of similar sentences with high lexical overlap. Then, we align the corresponding parts of the sentences to look for substitutions of

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

words. From these alignments, we finally create small-sized TOEFL-like test cases that we solve with contextual similarity measures.

In particular, as many as 35 per cent of the constructed TOEFL-like test cases contain close semantic relations. The methodology is also not hindered by low-frequency words and discovered 88 synonymous word pairs not listed in WordNet. Compared to other methods that create long lists of words related in unspecified way, our methodology extracts very short lists of candidates in paradigmatic relation with the head. Those lists can be easily scrutinized by a human expert in computer aided thesaurus construction.

We applied a number of contextual similarity measures over the set of 372 tests. The fact that the preceding step yielded tests with few candidates allowed recall of 75 per cent on detecting *Synonyms* and 58 per cent on *Is-a* by the Global strategy over the Cosine-PMI combination, 71 per cent on *Siblings* by the Product strategy over the Lin model and 49 per cent on *Instance Of* by the Local strategy over the combination Cosine-TfIdf. The results suggest that informed use of combinations of measures may lead to improved results.

However, the methodology still produces erroneous tests mostly resulting from bad text preprocessing and unreliable part-of-speech tagging. Some incorrect tests are also a consequence of incorrect paraphrasing. Finally, wrong alignments may also give rise to wrong test cases. With respect to preprocessing and part-of-speech tagging, efforts can be made to take advantage of the best tools in the field and should not be a major obstacle. However, one major improvement can be obtained by the normalization of the corpus, i.e. by detecting multiword units or named entities in a unsupervised way with the SENTA software (Dias, Guilloré and Lopes 1999).

Common alignment techniques do not deal with sentence reordering which mainly induce wrong alignments. As future work, we aim at testing a new alignment technique proposed by Cordeiro, Dias and Cleuziou (2007c), who use a combination of local and global biology-based alignment algorithms which deals with sentence reordering in an elegant way.

Finally, the extraction of paraphrases still remains an open problem. New techniques have been proposed, but they mostly rely on the use of huge linguistic resources or tools. We believe that the method proposed by Cordeiro *et al.* (2007a) can be improved by weighting the exclusive links so that more emphasis is made to exclusive links between meaningful words.

Acknowledgements

This reaserch was supported by the MEDON project funded by the Portuguese Agency for Research (Fundação para a Ciência e a Tecnologia) with the reference PTDC/EIA/80772/2006.

References

Agirre, E., and Lacalle, O. L. 2003. Clustering WordNet word senses. In *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, Borovets, Bulgaria, pp. 11–18.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Ahonen-Myka, H. 1999. Finding all frequent maximal sequences in text. In *Proceedings of ICML-99 Workshop on Machine Learning in Text Data Analysis*, Bled, Slovenia, pp. 11–17.

Barzilay, R., and Lee, L. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *HLT-NAACL 2003: Main Proceedings*, Edmonton, Canada, pp. 16–23.

Berland, M., and Charniak, E. 1999. Finding parts in very large corpora. In *Proceedings of ACL*, pp. 57–64. College Park, MD: Association for Computational Linguistics.

Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, Banff, Alberta, Canada, pp. 757–766.

Bordag, S. 2003. Sentence co-occurrences as small-world graphs: a solution to automatic lexical disambiguation. In *CICLing'03: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2588 of Lecture Notes in Computer Science, Mexico City, Mexico, pp. 329–332.

Caraballo, S. A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 120–126. College Park, MD: Association for Computational Linguistics.

Cederberg, S., and Widdows, D. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In W. Daelemans, M. Osborne (eds.), *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 111–118. Stroudsburg, PA: Association for Computational Linguistics.

Charles, W. G. 2000. Contextual correlates of meaning. *Applied Psycholinguistics* **21**(04): 505–524.

Cordeiro, J., Dias, G., and Brazdil, P. 2007a. Learning paraphrases from WNS corpora. In *Twentieth International FLAIRS Conference*, Key West, Florida, pp. 193–198.

Cordeiro, J., Dias, G., and Brazdil, P. 2007b. New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software* **2**(4): 12–23.

Cordeiro, J., Dias, G., and Cleuziou, G. 2007c. Biology based alignments of paraphrases for sentence compression. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing (ACL-PASCAL / ACL2007)*, Prague, Czech Republic, pp. 177–184.

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

Dias, G., Guilloré, S., and Lopes, J. G. P. 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of 6me Confrence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, Cargèse, France, pp. 333–339.

Dolan, W. B., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 1–7.

Doucet, A., and Ahonen-Myka, H. 2006. Probability and expected document frequency of discontinued word sequences, an efficient method for their exact computation. *Traitement Automatique des Langues (TAL)* **46**(2): 13–37.

Ehlert, B. 2003. *Making Accurate Lexical Semantic Similarity Judgments Using Word-Context Co-Occurrence Statistics*. Master's thesis, University of California, San Diego, CA.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

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

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Gale, W., Church, K. W., and Yarowsky, D. 1992. One sense per discourse. In *HLT '91: Proceedings of the Workshop on Speech and Natural Language*, Harriman, NY, pp. 233–237.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Norwell, MA: Kluwer.

Harris, Z. S. 1968. *Mathematical Structures of Language*. Wiley, New York, NY.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, pp. 539–545.

Heyer, L. J., Kruglyak, S., and Yooseph, S. 1999. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research* **9**(11): 1106–1115.

Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. 2008. Modelling word similarity: An evaluation of automatic synonymy extraction algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp. 3243–3249.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, Pittsburgh, PA, pp. 268–275.

Hirschman, L., Grishman, R., and Sager, N. 1975. Grammatically-based automatic word class formation. *Information Processing and Management* **11**(1–2): 39–57.

Jarmasz, M., and Szpakowicz, S. 2004. Roget's thesaurus and semantic similarity. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 212–219, Borovets, Bulgaria.

Jing, H., and McKeown, K. R. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics conference*, Seattle, WA, pp. 178–185.

Kaplan, A. 1950. An experimental study of ambiguity and context. *Mechanical Translation* **2**(2): 39–46. (Published as: Abraham, K. 1955. An experimental study of ambiguity and context. *Mechanical Translation* **2**(2): 39–46.)

Landauer, T. K., and Dumais, S. T. 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.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. 2004. RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research* **5**: 361–397.

Lin, D. 1998a. Automatic retrieval and clustering of similar words. In *COLING-ACL*, Montreal, QC, Canada, pp. 768–774.

Lin, D. 1998b. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, pp. 296–304.

Lin, D., Zhao, S., Qin, L., and Zhou, M. 2003. Identifying synonyms among distributionally similar words. In G. Gottlob, T. Walsh, G. Gottlob and T. Walsh (eds.), *Proceedings of IJCAI-03: Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, pp. 1492–1493.

Liu, H. 2004. MontyLingua: an end-to-end natural language processor with common sense. Available at: `web.media.mit.edu/~hugo/montylingua`

Mohammad, S. 2008. *Measuring Semantic Distance using Distributional Profiles of Concepts*. PhD thesis, University of Toronto, Toronto, Canada.

Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology* **3**(8):1405–1408.

Rapp, R. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, pp. 315–322.

Rapp, R. 2004. 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*, Lisbon, Portugal, pp. 1–7.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press

Roget, P. M. (ed.) 1852. *Roget's Thesaurus of English Words and Phrases*. Harlow, Essex, UK: Longman Group Ltd.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* **8**(10): 627–633.

Sahlgren, M. 2006. *The Word-Space Model*. PhD thesis, Stockholm University, Stockholm, Sweden.

Sahlgren, M., and Karlgren, J. 2002. 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*, pp. 169–176. London, UK.

Snow, R., Jurafsky, D., and Ng, A. Y. 2005. Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss and L. Bottou(eds.), *Advances in Neural Information Processing Systems 17*, pp. 1297–1304. MIT Press.

Snow, R., Jurafsky, D., and Ng, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia, pp. 801–808.

Spärck-Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**(1): 11–21.

Terra, E., and Clarke, C. 2003. Frequency estimates for statistical word similarity measures. In *NAACL '03: Proceedings of the 2003 Conference of HTL/NAACL*, Edmonton, Canada, pp. 165–172.

Turney, P. D. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, Freiburg, Germany, pp. 491–502.

Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. 2003. Combining independent modules in lexical multiple-choice problems. In *Recent Advances in Natural Language Processing III*, Borovets, Bulgaria, pp. 101–110.

Weeds, J., Weir, D., and McCarthy, D. 2004. Characterising measures of lexical distributional similarity. In *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 1015–1021.

https://doi.org/10.1017/S135132491000015X Published online by Cambridge University Press