Corpus-based and Knowledge-based Measures of Text Semantic Similarity

Rada Mihalcea and **Courtney Corley**
Department of Computer Science
University of North Texas
{rada,corley}@cs.unt.edu

Carlo Strapparava
Istituto per la Ricerca Scientifica e Tecnologica
ITC – irst
strappa@itc.it

Abstract

This paper presents a method for measuring the semantic similarity of texts, using corpus-based and knowledge-based measures of similarity. Previous work on this problem has focused mainly on either large documents (e.g. text classification, information retrieval) or individual words (e.g. synonymy tests). Given that a large fraction of the information available today, on the Web and elsewhere, consists of short text snippets (e.g. abstracts of scientific documents, imagine captions, product descriptions), in this paper we focus on measuring the semantic similarity of short texts. Through experiments performed on a paraphrase data set, we show that the semantic similarity method outperforms methods based on simple lexical matching, resulting in up to 13% error rate reduction with respect to the traditional vector-based similarity metric.

Introduction

Measures of text similarity have been used for a long time in applications in natural language processing and related areas. One of the earliest applications of text similarity is perhaps the vectorial model in information retrieval, where the document most relevant to an input query is determined by ranking documents in a collection in reversed order of their similarity to the given query (Salton & Lesk 1971). Text similarity has also been used for relevance feedback and text classification (Rocchio 1971), word sense disambiguation (Lesk 1986; Schutze 1998), and more recently for extractive summarization (Salton *et al.* 1997), and methods for automatic evaluation of machine translation (Papineni *et al.* 2002) or text summarization (Lin & Hovy 2003). Measures of text similarity were also found useful for the evaluation of text coherence (Lapata & Barzilay 2005).

With few exceptions, the typical approach to finding the similarity between two text segments is to use a simple lexical matching method, and produce a similarity score based on the number of lexical units that occur in both input segments. Improvements to this simple method have considered stemming, stop-word removal, part-of-speech tagging, longest subsequence matching, as well as various weighting and normalization factors (Salton & Buckley 1997). While successful to a certain degree, these lexical similarity methods cannot always identify the *semantic* similarity of texts. For instance, there is an obvious similarity between the text segments *I own a dog* and *I have an animal*, but most of the current text similarity metrics will fail in identifying any kind of connection between these texts.

There is a large number of word-to-word semantic similarity measures, using approaches that are either knowledge-based (Wu & Palmer 1994; Leacock & Chodorow 1998) or corpus-based (Turney 2001). Such measures have been successfully applied to language processing tasks such as malapropism detection (Budanitsky & Hirst 2001), word sense disambiguation (Patwardhan, Banerjee, & Pedersen 2003), and synonym identification (Turney 2001). For text-based semantic similarity, perhaps the most widely used approaches are the approximations obtained through query expansion, as performed in information retrieval (Voorhees 1993), or the latent semantic analysis method (Landauer, Foltz, & Laham 1998) that measures the similarity of texts by exploiting second-order word relations automatically acquired from large text collections.

A related line of work consists of methods for paraphrase recognition, which typically seek to align sentences in comparable corpora (Barzilay & Elhadad 2003; Dolan, Quirk, & Brockett 2004), or paraphrase generation using distributional similarity applied on paths of dependency trees (Lin & Pantel 2001) or using bilingual parallel corpora (Barnard & Callison-Burch 2005). These methods target the identification of paraphrases in large documents, or the generation of paraphrases starting with an input text, without necessarily providing a measure of their similarity. The recently introduced textual entailment task (Dagan, Glickman, & Magnini 2005) is also related to some extent, however textual entailment targets the identification of a *directional inferential* relation between texts, which is different than textual similarity, and hence entailment systems are not overviewed here.

In this paper, we suggest a method for measuring the semantic similarity of texts by exploiting the information that can be drawn from the similarity of the component words. Specifically, we describe two corpus-based and six knowledge-based measures of word semantic similarity, and show how they can be used to derive a text-to-text similarity metric. We show that this measure of text semantic similarity outperforms the simpler vector-based similarity approach, as evaluated on a paraphrase recognition task.

Text Semantic Similarity

Measures of semantic similarity have been traditionally defined between words or concepts, and much less between text segments consisting of two or more words. The emphasis on word-to-word similarity metrics is probably due to the availability of resources that specifically encode relations between words or concepts (e.g. WordNet), and the various testbeds that allow for their evaluation (e.g. TOEFL or SAT analogy/synonymy tests). Moreover, the derivation of a text-to-text measure of similarity starting with a word-based semantic similarity metric may not be straightforward, and consequently most of the work in this area has considered mainly applications of the traditional vectorial model, occasionally extended to n-gram language models.

Given two input text segments, we want to automatically derive a score that indicates their similarity at *semantic* level, thus going beyond the simple lexical matching methods traditionally used for this task. Although we acknowledge the fact that a comprehensive metric of text semantic similarity should also take into account the structure of the text, we take a first rough cut at this problem and attempt to model the semantic similarity of texts as a function of the semantic similarity of the component words. We do this by combining metrics of word-to-word similarity and word specificity into a formula that is a potentially good indicator of the semantic similarity of the two input texts.

The following section provides details on eight different corpus-based and knowledge-based measures of word semantic similarity. In addition to the similarity of words, we also take into account the *specificity* of words, so that we can give a higher weight to a semantic matching identified between two specific words (e.g. *collie* and *sheepdog*), and give less importance to the similarity measured between generic concepts (e.g. *get* and *become*). While the specificity of words is already measured to some extent by their depth in the semantic hierarchy, we are reinforcing this factor with a corpus-based measure of word specificity, based on distributional information learned from large corpora.

The *specificity* of a word is determined using the inverse document frequency (*idf*) introduced by Sparck-Jones (1972), defined as the total number of documents in the corpus divided by the total number of documents including that word. The *idf* measure was selected based on previous work that theoretically proved the effectiveness of this weighting approach (Papineni 2001). In the experiments reported here, document frequency counts are derived starting with the British National Corpus – a 100 million words corpus of modern English including both spoken and written genres.

Given a metric for word-to-word similarity and a measure of word specificity, we define the semantic similarity of two text segments T_1 and T_2 using a metric that combines the semantic similarities of each text segment in turn with respect to the other text segment. First, for each word w in the segment T_1 we try to identify the word in the segment T_2 that has the highest semantic similarity ($maxSim(w, T_2)$), according to one of the word-to-word similarity measures described in the following section. Next, the same process is applied to determine the most similar word in T_1 starting with words in T_2. The word similarities are then weighted with the corresponding word specificity, summed up, and normalized with the length of each text segment. Finally the resulting similarity scores are combined using a simple average. Note that only open-class words and cardinals can participate in this semantic matching process. As done in previous work on text similarity using vector-based models, all function words are discarded.

The similarity between the input text segments T_1 and T_2 is therefore determined using the following scoring function:

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{T_1\}} (maxSim(w,T_2) * idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in \{T_2\}} (maxSim(w,T_1) * idf(w))}{\sum_{w \in \{T_2\}} idf(w)}\right) \quad (1)$$

This similarity score has a value between 0 and 1, with a score of 1 indicating identical text segments, and a score of 0 indicating no semantic overlap between the two segments.

Note that the maximum similarity is sought only within classes of words with the same part-of-speech. The reason behind this decision is that most of the word-to-word knowledge-based measures cannot be applied across parts-of-speech, and consequently, for the purpose of consistency, we imposed the "same word-class" restriction to all the word-to-word similarity measures. This means that, for instance, the most similar word to the noun *flower* within the text "*There are many green plants next to the house*" will be sought among the nouns *plant* and *house*, and will ignore the words with a different part-of-speech (*be*, *green*, *next*). Moreover, for those parts-of-speech for which a word-to-word semantic similarity cannot be measured (e.g. some knowledge-based measures are not defined among adjectives or adverbs), we use instead a lexical match measure, which assigns a $maxSim$ of 1 for identical occurrences of a word in the two text segments.

Semantic Similarity of Words

There is a relatively large number of word-to-word similarity metrics that were previously proposed in the literature, ranging from distance-oriented measures computed on semantic networks, to metrics based on models of distributional similarity learned from large text collections. From these, we chose to focus our attention on two corpus-based metrics and six knowledge-based different metrics, selected mainly for their observed performance in other natural language processing applications.

Corpus-based Measures

Corpus-based measures of word semantic similarity try to identify the degree of similarity between words using information exclusively derived from large corpora. In the experiments reported here, we considered two metrics, namely: (1) pointwise mutual information (Turney 2001), and (2) latent semantic analysis (Landauer, Foltz, & Laham 1998).

Pointwise Mutual Information The pointwise mutual information using data collected by information retrieval

(PMI-IR) was suggested by (Turney 2001) as an unsupervised measure for the evaluation of the semantic similarity of words. It is based on word co-occurrence using counts collected over very large corpora (e.g. the Web). Given two words w_1 and w_2, their PMI-IR is measured as:

$$\textit{PMI-IR}(w_1, w_2) = log_2 \frac{p(w_1 \& w_2)}{p(w_1) * p(w_2)} \quad (2)$$

which indicates the degree of statistical dependence between w_1 and w_2, and can be used as a measure of the semantic similarity of w_1 and w_2. From the four different types of queries suggested by Turney (2001), we are using the $NEAR$ query (co-occurrence within a ten-word window), which is a balance between accuracy (results obtained on synonymy tests) and efficiency (number of queries to be run against a search engine). Specifically, the following query is used to collect counts from the AltaVista search engine.

$$p_{NEAR}(w_1 \& w_2) \simeq \frac{hits(w_1\ NEAR\ w_2)}{WebSize} \quad (3)$$

With $p(w_i)$ approximated as $hits(w_1)/WebSize$, the following PMI-IR measure is obtained:

$$log_2 \frac{hits(w_1\ AND\ w_2) * WebSize}{hits(w_1) * hits(w_2)} \quad (4)$$

In a set of experiments based on TOEFL synonymy tests (Turney 2001), the PMI-IR measure using the $NEAR$ operator accurately identified the correct answer (out of four synonym choices) in 72.5% of the cases, which exceeded by a large margin the score obtained with latent semantic analysis (64.4%), as well as the average non-English college applicant (64.5%). Since Turney (2001) performed evaluations of synonym candidates for one word at a time, the $WebSize$ value was irrelevant in the ranking. In our application instead, it is not only the ranking of the synonym candidates that matters (for the selection of $maxSim$ in Equation 1), but also the true value of PMI-IR, which is needed for the overall calculation of the text-to-text similarity metric. We approximate the value of $WebSize$ to 7x10^{11}, which is the value used by Chklovski (2004) in co-occurrence experiments involving Web counts.

Latent Semantic Analysis Another corpus-based measure of semantic similarity is the latent semantic analysis (LSA) proposed by Landauer (1998). In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus. For the experiments reported here, we run the SVD operation on the British National Corpus.

SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. In our case, SVD decomposes the term-by-document matrix $\mathbf{T}$ into three matrices $\mathbf{T} = \mathbf{U}\mathbf{\Sigma_k}\mathbf{V}^T$ where $\mathbf{\Sigma_k}$ is the diagonal $k \times k$ matrix containing the k singular values of $\mathbf{T}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$, and $\mathbf{U}$ and $\mathbf{V}$ are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is re-composed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\mathbf{\Sigma_{k'}}\mathbf{V}^T$.

LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. The similarity in the resulting vector space is then measured with the standard cosine similarity. Note also that LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, and texts.

The application of the LSA word similarity measure to text semantic similarity is done using Equation 1, which roughly amounts to the *pseudo-document* text representation for LSA computation, as described by Berry (1992). In practice, each text segment is represented in the LSA space by summing up the normalized LSA vectors of all the constituent words, using also a *tf.idf* weighting scheme.

Knowledge-based Measures

There are a number of measures that were developed to quantify the degree to which two words are semantically related using information drawn from semantic networks – see e.g. (Budanitsky & Hirst 2001) for an overview. We present below several measures found to work well on the WordNet hierarchy. All these measures assume as input a pair of concepts, and return a value indicating their semantic relatedness. The six measures below were selected based on their observed performance in other language processing applications, and for their relatively high computational efficiency.

We conduct our evaluation using the following word similarity metrics: Leacock & Chodorow, Lesk, Wu & Palmer, Resnik, Lin, and Jiang & Conrath. Note that all these metrics are defined between concepts, rather than words, but they can be easily turned into a word-to-word similarity metric by selecting for any given pair of words those two meanings that lead to the highest concept-to-concept similarity[1]. We use the WordNet-based implementation of these metrics, as available in the WordNet::Similarity package (Patwardhan, Banerjee, & Pedersen 2003). We provide below a short description for each of these six metrics.

The **Leacock & Chodorow** (Leacock & Chodorow 1998) similarity is determined as:

$$Sim_{lch} = -\log \frac{length}{2 * D} \quad (5)$$

where $length$ is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

The **Lesk** similarity of two concepts is defined as a function of the overlap between the corresponding definitions, as provided by a dictionary. It is based on an algorithm proposed by Lesk (1986) as a solution for word sense disambiguation. The application of the Lesk similarity measure is not limited to semantic networks, and it can be used in conjunction with any dictionary that provides word definitions.

The **Wu and Palmer** (Wu & Palmer 1994) similarity metric measures the depth of two given concepts in the Word-

[1]This is similar to the methodology used by (McCarthy *et al.* 2004) to find similarities between words and senses starting with a concept-to-concept similarity measure.

Net taxonomy, and the depth of the least common subsumer (LCS), and combines these figures into a similarity score:

$$Sim_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)} \quad (6)$$

The measure introduced by **Resnik** (Resnik 1995) returns the information content (IC) of the LCS of two concepts:

$$Sim_{res} = IC(LCS) \quad (7)$$

where IC is defined as:

$$IC(c) = -\log P(c) \quad (8)$$

and $P(c)$ is the probability of encountering an instance of concept c in a large corpus.

The next measure we use in our experiments is the metric introduced by **Lin** (Lin 1998), which builds on Resnik's measure of similarity, and adds a normalization factor consisting of the information content of the two input concepts:

$$Sim_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)} \quad (9)$$

Finally, the last similarity metric considered is **Jiang & Conrath** (Jiang & Conrath 1997):

$$Sim_{jnc} = \frac{1}{IC(concept_1) + IC(concept_2) - 2 * IC(LCS)} \quad (10)$$

Note that all the word similarity measures are normalized so that they fall within a 0–1 range. The normalization is done by dividing the similarity score provided by a given measure with the maximum possible score for that measure.

A Walk-Through Example

The application of the text similarity measure is illustrated with an example. Given two text segments, as shown below, we want to determine a score that reflects their semantic similarity. For illustration purposes, we restrict our attention to one corpus-based measure – the PMI-IR metric implemented using the AltaVista $NEAR$ operator.

Text Segment 1: When the defendant and his lawyer walked into the court, some of the victim supporters turned their backs on him.
Text Segment 2: When the defendant walked into the courthouse with his attorney, the crowd turned their backs on him.

Starting with each of the two text segments, and for each open-class word, we determine the most similar word in the other text segment, according to the PMI-IR similarity measure. As mentioned earlier, a semantic similarity is sought only between words with the same part-of-speech. Table 1 shows the word similarity scores and the word specificity (idf) starting with the first text segment.

Next, using Equation 1, we combine the word similarities and their corresponding specificity, and determine the semantic similarity of the two texts as 0.80. This similarity score correctly identifies the paraphrase relation between the two text segments (using the same threshold of 0.50 as used throughout all the experiments reported in this paper). Instead, a cosine similarity score based on the same *idf* weights will result in a score of 0.46, thereby failing to find the paraphrase relation.

Text 1	Text 2	maxSim	idf
defendant	defendant	1.00	3.93
lawyer	attorney	0.89	2.64
walked	walked	1.00	1.58
court	courthouse	0.60	1.06
victims	courthouse	0.40	2.11
supporters	crowd	0.40	2.15
turned	turned	1.00	0.66
backs	backs	1.00	2.41

Table 1: Word similarity scores and word specificity (idf)

Although there are a few words that occur in both text segments (e.g. *defendant*, or *turn*), there are also words that are not identical, but closely related, e.g. *lawyer* found similar to *attorney*, or *supporters* which is related to *crowd*. Unlike traditional similarity measures based on lexical matching, our metric takes into account the semantic similarity of these words, resulting in a more precise measure of text similarity.

Evaluation and Results

To test the effectiveness of the text semantic similarity measure, we use it to automatically identify if two text segments are paraphrases of each other. We use the Microsoft paraphrase corpus (Dolan, Quirk, & Brockett 2004), consisting of 4,076 training and 1,725 test pairs, and determine the number of correctly identified paraphrase pairs in the corpus using the text semantic similarity measure as the only indicator of paraphrasing. The paraphrase pairs in this corpus were automatically collected from thousands of news sources on the Web over a period of 18 months, and were subsequently labeled by two human annotators who determined if the two sentences in a pair were semantically equivalent or not. The agreement between the human judges who labeled the candidate paraphrase pairs in this data set was measured at approximately 83%, which can be considered as an upperbound for an automatic paraphrase recognition task performed on this data set.

For each candidate paraphrase pair in the test set, we first evaluate the text semantic similarity metric using Equation 1, and then label the candidate pair as a paraphrase if the similarity score exceeds a threshold of 0.5. Note that this is an unsupervised experimental setting, and therefore the training data is not used in the experiments.

Baselines

For comparison, we also compute two baselines: (1) A random baseline created by randomly choosing a true (paraphrase) or false (not paraphrase) value for each text pair; and (2) A vector-based similarity baseline, using a cosine similarity measure as traditionally used in information retrieval, with *tf.idf* weighting.

Results

We evaluate the results in terms of accuracy, representing the number of correctly identified true or false classifications in

Metric	Acc.	Prec.	Rec.	F
Semantic similarity (corpus-based)				
PMI-IR	69.9	70.2	95.2	81.0
LSA	68.4	69.7	95.2	80.5
Semantic similarity (knowledge-based)				
J & C	69.3	72.2	87.1	79.0
L & C	69.5	**72.4**	87.0	79.0
Lesk	69.3	**72.4**	86.6	78.9
Lin	69.3	71.6	88.7	79.2
W & P	69.0	70.2	92.1	80.0
Resnik	69.0	69.0	96.4	80.4
Combined	**70.3**	69.6	**97.7**	**81.3**
Baselines				
Vector-based	65.4	71.6	79.5	75.3
Random	51.3	68.3	50.0	57.8

Table 2: Text similarity for paraphrase identification

the test data set. We also measure precision, recall and F-measure, calculated with respect to the true values in the test data. Table 2 shows the results obtained. Among all the individual measures of similarity, the PMI-IR measure was found to perform the best, although the difference with respect to the other measures is small.

In addition to the individual measures of similarity, we also evaluate a metric that combines several similarity measures into a single figure, using a simple average. We include all similarity measures, for an overall final accuracy of 70.3%, and an F-measure of 81.3%.

The improvement of the semantic similarity metrics over the vector-based cosine similarity was found to be statistically significant in all the experiments, using a paired t-test ($p < 0.001$).

Discussion and Conclusions

As it turns out, incorporating semantic information into measures of text similarity increases the likelihood of recognition significantly over the random baseline and over the vector-based cosine similarity baseline, as measured in a paraphrase recognition task. The best performance is achieved using a method that combines several similarity metrics into one, for an overall accuracy of 70.3%, representing a significant 13.8% error rate reduction with respect to the vector-based cosine similarity baseline. Moreover, if we were to take into account the upperbound of 83% established by the inter-annotator agreement achieved on this data set (Dolan, Quirk, & Brockett 2004), the error rate reduction over the baseline appears even more significant.

In addition to performance, we also tried to gain insights into the applicability of the semantic similarity measures, by finding their coverage on this data set. On average, among approximately 18,000 word similarities identified in this corpus, about 14,500 are due to lexical matches, and 3,500 are due to semantic similarities, which indicates that about 20% of the relations found between text segments are based on semantics, in addition to lexical identity.

Despite the differences among the various word-to-word similarity measures (corpus-based vs. knowledge-based, definitional vs. link-based), the results are surprisingly similar. To determine if the similar overall results are due to a similar behavior on the same subset of the test data (presumably an "easy" subset that can be solved using measures of semantic similarity), or if the different measures cover in fact different subsets of the data, we calculated the Pearson correlation factor among all the similarity measures. As seen in Table 3, there is in fact a high correlation among several of the knowledge-based measures, indicating an overlap in their behavior. Although some of these metrics are divergent in what they measure (e.g. Lin versus Lesk), it seems that the fact they are applied in a *context* lessens the differences observed when applied at word level. Interestingly, the Resnik measure has a low correlation with the other knowledge-based measures, and a somehow higher correlation with the corpus-based metrics, which is probably due to the data-driven information content used in the Resnik measure (although Lin and Jiang & Conrath also use the information content, they have an additional normalization factor that makes them behave differently). Perhaps not surprising, the corpus-based measures are only weakly correlated with the knowledge-based measures and among them, with LSA having the smallest correlation with the other metrics.

An interesting example is represented by the following two text segments, where only the Resnik measure and the two corpus-based measures manage to identify the paraphrase, because of a higher similarity found between *systems* and *PC*, and between *technology* and *processor*.

Text Segment 1: Gateway will release new Profile 4 systems with the new Intel technology on Wednesday.
Text Segment 2: Gateway 's all-in-one PC , the Profile 4 , also now features the new Intel processor.

There are also cases where almost all the semantic similarity measures fail, and instead the simpler cosine similarity has a better performance. This is mostly the case for the negative (not paraphrase) examples in the test data, where the semantic similarities identified between words increase the overall text similarity above the threshold of 0.5. For instance, the following text segments were falsely marked as paraphrases by all but the cosine similarity and the Resnik measure:

Text Segment 1: The man wasn't on the ice, but trapped in the rapids, swaying in an eddy about 250 feet from the shore.
Text Segment 2: The man was trapped about 250 feet from the shore, right at the edge of the falls.

The small variations between the accuracies obtained with the corpus-based and knowledge-based measures also suggest that both data-driven and knowledge-rich methods have their own merits, leading to a similar performance. Corpus-based methods have the advantage that no hand-made resources are needed and, apart form the choice of an appropriate and large corpus, they raise no problems related to the completeness of the resources. On the other hand, knowledge-based methods can encode fine-grained information. This difference can be observed in terms of precision and recall. In fact, while precision is generally higher with knowledge-based measures, corpus-based measures give in general better performance in recall.

Although our method relies on a bag-of-words approach, as it turns out the use of measures of *semantic* similarity improves significantly over the traditional lexical matching

	Vect	PMI-IR	LSA	J&C	L&C	Lesk	Lin	W&P	Resnik
Vect	1.00	0.84	0.44	0.61	0.63	0.60	0.61	0.50	0.65
PMI-IR		1.00	0.58	0.67	0.68	0.65	0.67	0.58	0.64
LSA			1.00	0.42	0.44	0.42	0.43	0.34	0.41
J&C				1.00	0.98	0.97	0.99	0.91	0.45
L&C					1.00	0.98	0.98	0.87	0.46
Lesk						1.00	0.96	0.86	0.43
Lin							1.00	0.88	0.44
W&P								1.00	0.34
Resnik									1.00

Table 3: Pearson correlation among similarity measures

metrics. We are nonetheless aware that a bag-of-words approach ignores many of the important relationships in sentence structure, such as dependencies between words, or roles played by the various arguments in the sentence. Future work will consider the investigation of more sophisticated representations of sentence structure, such as first order predicate logic or semantic parse trees, which should allow for the implementation of more effective measures of text semantic similarity.

References

Barnard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Barzilay, R., and Elhadad, N. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*.

Berry, M. 1992. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications* 6(1).

Budanitsky, A., and Hirst, G. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*.

Chklovski, T., and Pantel, P. 2004. Verbocean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing.*.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Workshop*.

Dolan, W.; Quirk, C.; and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Jiang, J., and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*.

Landauer, T. K.; Foltz, P.; and Laham, D. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25.

Lapata, M., and Barzilay, R. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*. The MIT Press.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*.

Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference* .

Lin, D., and Pantel, P. 2001. Discovery of inference rules for question answering. *Natural Language Engineering* 7(3).

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conf. on Machine Learning*.

McCarthy, D.; Koeling, R.; Weeds, J.; and Carroll, J. 2004. Finding predominant senses in untagged text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Papineni, K. 2001. Why inverse document frequency? In *Proceedings of the North American Chapter of the Association for Compuatational Linguistics*, 25–32.

Patwardhan, S.; Banerjee, S.; and Pedersen, T. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

Resnik, P. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Rocchio, J. 1971. *Relevance feedback in information retrieval*. Prentice Hall, Ing. Englewood Cliffs, New Jersey.

Salton, G., and Buckley, C. 1997. Term weighting approaches in automatic text retrieval. In *Readings in Information Retrieval*. San Francisco, CA: Morgan Kaufmann Publishers.

Salton, G., and Lesk, M. 1971. *Computer evaluation of indexing and text processing*. Prentice Hall, Ing. Englewood Cliffs, New Jersey. 143–180.

Salton, G.; Singhal, A.; Mitra, M.; and Buckley, C. 1997. Automatic text structuring and summarization. *Information Processing and Management* 2(32).

Schutze, H. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–124.

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

Turney, P. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*.

Voorhees, E. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference*.

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.