Computing semantic relatedness of words and texts in Wikipedia-derived semantic space

Evgeniy Gabrilovich and **Shaul Markovitch**
Department of Computer Science
Technion—Israel Institute of Technology, 32000 Haifa, Israel
{gabr,shaulm}@cs.technion.ac.il

Abstract

Adequate representation of natural language semantics requires access to vast amounts of common sense and domain-specific world knowledge. Prior work in the field was either based on purely statistical techniques that did not make use of background knowledge or on huge manual efforts, such as the CYC projects. Here we propose a novel method, called Explicit Semantic Analysis (ESA), for fine-grained semantic interpretation of unrestricted natural language texts. Our method represents meaning in a high-dimensional space of concepts derived from Wikipedia, the largest encyclopedia in existence. We use machine learning techniques that allow us to explicitly represent the meaning of any text in terms of Wikipedia-based concepts. We evaluate the effectiveness of our method on automatically computing the degree of semantic relatedness between fragments of natural language text. Compared with the previous state of the art, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgments: from $r = 0.56$ to 0.75 for individual words and from $r = 0.60$ to 0.72 for texts. Consequently, we anticipate ESA to give rise to the next generation of natural language processing tools. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.

1 Introduction

As computers become increasingly prevalent in our lives, it is essential that humans can converse with them in natural language. Achieving this aim is a major challenge, which requires endowing machines with the ability to understand the meaning of language utterances. In turn, adequate representation of language semantics requires access to vast amounts of common sense and domain-specific world knowledge[Buchanan and Feigenbaum, 1982; Lenat and Guha, 1990].

Prior work in the field was based on purely statistical techniques that did not make use of background knowledge[Baeza-Yates and Ribeiro-Neto, 1999; Sebastiani, 2002; Deerwester *et al.*, 1990], or on long-term endeavors such as CYC[Lenat *et al.*, 1990; Lenat, 1995], which begin with painstaking encoding of numerous nuggets of common sense.

Here we propose a novel method, called Explicit Semantic Analysis (ESA), for fine-grained semantic interpretation of unrestricted natural language texts. Our method represents meaning in a high-dimensional space of concepts derived from Wikipedia[Wikipedia, 2006], the largest encyclopedia in existence. We use machine learning techniques that allow us to explicitly represent the meaning of any text *in terms of* Wikipedia-based concepts. We evaluate the effectiveness of our method on automatically computing the degree of semantic relatedness between fragments of natural language text. The ability to quantify semantic relatedness of texts underlies many fundamental tasks in computational linguistics, including word sense disambiguation, information retrieval, word and text clustering, and error correction[Budanitsky and Hirst, 2006].

2 Background: Semantic Relatedness

How related are "cat" and "mouse"? And what about "preparing a manuscript" and "writing an article"? Reasoning about semantic relatedness of natural language utterances is routinely performed by humans but remains an unsurmountable obstacle for computers. Humans do not judge text relatedness merely at the level of text words. Words trigger reasoning at a much deeper level that manipulates *concepts*—the basic units of meaning that serve humans to organize and share their knowledge. Thus, humans interpret the specific wording of a document in the much larger context of their background knowledge and experience. Lacking such elaborate resources, computers need alternative ways to represent texts and reason about them.

Prior work on computing semantic relatedness pursued three main directions: comparing text fragments as bags of words in vector space, using lexical resources, and using Latent Semantic Analysis (LSA)[Deerwester *et al.*, 1990]. The former technique is the simplest, but performs sub-optimally when the texts to be compared share few words, for instance, when the texts use synonyms to convey similar messages. This technique is also trivially inappropriate for comparing individual words. The latter two techniques attempt to circumvent this problem.

Lexical databases such as WordNet [Fellbaum, 1998] or Roget's Thesaurus[Roget, 1852] encode relations between words such as synonymy, hypernymy, and meronymy. Quite a few metrics have been defined that compute relatedness using various properties of the underlying graph structure of these resources[Budanitsky and Hirst, 2006; Jarmasz, 2003]. The obvious drawback of this approach is that creation of lexical resources requires lexicographic expertise as well as a lot of time and effort, and consequently such resources cover only a small fragment of the language lexicon. Specifically, such resources contain few proper names, neologisms, slang, and domain-specific technical terms. Furthermore, these resources have strong lexical orientation in that they predominantly contain information about individual words, but little world knowledge in general.

On the other hand, LSA[Deerwester *et al.*, 1990] is a purely statistical technique, which leverages word cooccurrence information from a large unlabeled corpus of text. LSA does not rely on any human-organized knowledge; rather, it "learns" its representation by applying Singular Value Decomposition (SVD) to the words-by-documents cooccurrence matrix. LSA is essentially a dimensionality reduction technique that identifies a number of most prominent dimensions in the data, which are assumed to correspond to "latent concepts". Meanings of words and documents are then compared in the space defined by these concepts. Latent semantic models are notoriously difficult to interpret, since the computed concepts cannot be readily mapped into natural concepts manipulated by humans. In the next section we will present our novel Explicit Semantic Analysis method that circumvents this problem.

3 Explicit Semantic Analysis

Our approach is inspired by the desire to augment text representation with massive amounts of world knowledge. We represent texts as a weighted mixture of a predetermined set of *natural* concepts, which are defined by humans themselves and can be easily explained. To achieve this aim, we use concepts defined by Wikipedia[Wikipedia, 2006] articles, e.g., COMPUTER SCIENCE or UNITED KINGDOM. The choice of encyclopedia articles as concepts is quite natural, as each article is focused on a single issue, which it discusses in detail. An important advantage of our approach is thus the use of vast amounts of highly organized human knowledge encoded in Wikipedia. Compared to LSA, our methodology explicitly uses the knowledge collected and organized by humans. Compared to lexical resources such as WordNet, our methodology leverages knowledge bases that are orders of magnitude larger and more comprehensive. Furthermore, Wikipedia undergoes constant development so its breadth and depth steadily increase over time.

We opted to use Wikipedia because it is currently the largest knowledge repository on the Web. Wikipedia is available in dozens of languages, while its English version is the largest of all with 400+ million words in over one million articles, contributed by over 300,000 volunteer editors[Wikipedia, 2006]. Even though Wikipedia editors are not required to be established researchers or practitioners, the open editing approach yields remarkable quality. A recent study[Giles, 2005] found Wikipedia accuracy to rival that of Encyclopaedia Britannica. Another benefit of this openness is scalability—Britannica is about an order of magnitude smaller, with 44 million words in 65,000 articles[Encyclopaedia Britannica, 2006]. Importantly, Wikipedia articles are heavily cross-linked, in a way reminiscent of linking on the Web. These links encode many interesting relations between the concepts, and constitute an important source of information in addition to the article texts.

We use machine learning techniques to build a *semantic interpreter*, which maps fragments of natural language text into a weighted sequence of Wikipedia concepts ordered by their relevance to the input. This way, input texts are represented as weighted vectors of concepts, called *interpretation vectors*. The meaning of a text fragment is thus interpreted in terms of its affinity with a host of Wikipedia concepts. Computing semantic relatedness of texts then amounts to comparing their vectors in the space defined by the concepts, for example, using the cosine metric[Zobel and Moffat, 1998]. Our semantic analysis is *explicit* in the sense that we manipulate manifest concepts grounded in human cognition, rather than "latent concepts" used by LSA.

Observe that input texts are given *in the same form* as Wikipedia articles, that is, as plain text. Therefore, we can use conventional text classification algorithms[Sebastiani, 2002] to rank the concepts represented by these articles according to their relevance to the given text fragment. It is this key observation that allows us to use encyclopedia directly, without the need for deep language understanding or pre-cataloged common-sense knowledge.

Specifically, each Wikipedia concept is represented as an attribute vector. Entries of these vectors were assigned weights using TF.IDF scheme with pivoted unique length normalization[Salton and McGill, 1983; Singhal *et al.*, 1996]. These weights quantify the strength of association between an individual word and a concept.

To speed up semantic interpretation, we then built an *inverted index*, which maps each word into a list of concepts in which it appears. We also used the inverted index to discard insignificant associations between words and concepts by removing those concepts whose weights for a given word are too low.

We implemented the semantic interpreter as a centroid-based classifier[Han and Karypis, 2000], which, given a text fragment, ranks all the Wikipedia concepts by their relevance to the fragment. Given a text fragment, we first represent it as a vector using TF.IDF scheme[Salton and McGill, 1983]. The semantic interpreter iterates over the text words, retrieves corresponding entries from the inverted index, and merges them into a weighted vector of concepts that represents the given text. Let $T = \{w_i\}$ be input text, and let $\langle v_i \rangle$ be its TF.IDF vector, where v_i is the weight of word w_i. Let $\langle k_j \rangle$ be an inverted index entry for word w_i, where k_j quantifies the strength of association of word w_i with Wikipedia concept $c_j, \{c_j \in c_1, \ldots, c_N\}$ (where N is the total number of Wikipedia concepts). Then, the semantic interpretation vector V for text T is a vector of length N, in which the weight of each concept c_j is defined as $\sum_{w_i \in T} v_i \cdot k_j$. Entries of this

#	Input: *"equipment"*	Input: *"investor"*
1	Tool	Investment
2	Digital Equipment Corporation	Angel investor
3	Military technology and equipment	Stock trader
4	Camping	Mutual fund
5	Engineering vehicle	Margin (finance)
6	Weapon	Modern portfolio theory
7	Original equipment manufacturer	Equity investment
8	French Army	Exchange-traded fund
9	Electronic test equipment	Hedge fund
10	Distance Measuring Equipment	Ponzi scheme

Table 1: First ten concepts in the interpretation vectors for sample words.

vector reflect the affinity of the corresponding concepts to text T. Computing semantic relatedness between such vectors can be done using standard metrics, e.g., the cosine metric[Zobel and Moffat, 1998].

Our method is similar to the one used by Gabrilovich and Markovitch[2006] for generating features for text categorization. Here, however, we can not use information-gain based methods for filtering the Wikipedia concepts. Furthermore, while in their work, Gabrilovich and Markovitch generated several concepts for each text input, here we map each text to a multi-dimensional vector in the space of all Wikipedia concepts.

To illustrate our approach, we show the first ten entries of the interpretation vectors for several text fragments. Table 1 contains the first entries in the vectors of individual words ("equipment" and "investor", respectively), while Table 2 show the prefixes of vectors for longer passages. It is particularly interesting to juxtapose the interpretation vectors for fragments that contain ambiguous words. Table 3 shows the (prefixes of) vectors for phrases that contain ambiguous words "bank" and "jaguar". As can be readily seen, our semantic interpretation methodology is capable of performing word sense disambiguation, by considering ambiguous words in the context of their neighbors.

4 Explicit Semantic Analysis vs. WikiRelate!

Just before submitting this paper we encountered a very recent paper [Strube and Ponzetto, 2006] that also uses Wikipedia for computing semantic relatedness. The method used by Strube and Ponzetto, however, is radically different than ours.

Given a pair of words w_1 and w_2, WikiRelate! will search for Wikipedia articles, p_1 and p_2, with w_1 and w_2 in their titles respectively. The semantic relatedness is then based on various distance measures between p_1 and p_2. These measures either rely on the texts of the pages, or path distances within the category hierarchy of Wikipedia. Our approach, on the other hand, will represent each word as a weighted vector of Wikipedia concepts. Semantic relatedness will then be computed by comparing the two concept vectors.

Thus, the differences between the two approaches are:

1. WikiRelate! can only process words that are in titles of Wikipedia articles. ESA only requires that the word appear within the text of Wikipedia articles.
2. WikiRelate! is limited to single words while ESA can process text of any length.
3. WikiRelate! represents the semantics of a word by either the text of the article associated with it, or by the node in the category hierarchy. ESA has a much more structured semantic representation consisting of a vector of Wikipedia concepts.

In the next section we will see that, indeed, the richer representation of ESA yields much better results.

5 Experimental Evaluation

We implemented our ESA approach using a Wikipedia snapshot as of March 26, 2006. After parsing the Wikipedia XML dump, we obtained 2.9 Gb of text in 1,187,839 articles. Upon removing small and overly specific concepts, 241,393 articles were left. We processed the text of these articles by removing stop words and rare words, and stemmed the remaining words using Porter's algorithm[Porter, 1980]; this yielded 389,202 distinct terms, which served for representing Wikipedia concepts as attribute vectors.

To better evaluate Wikipedia-based semantic interpretation, we also implemented a semantic interpreter based on another large-scale knowledge repository—the Open Directory Project (ODP)[ODP, 2006]. The ODP is the largest Web directory to date, where concepts correspond to categories of the directory, e.g., TOP/COMPUTERS/ARTIFICIAL INTELLIGENCE/MACHINE LEARNING. In this case, interpretation of a text fragment amounts to computing a weighted vector of ODP concepts, ordered by their affinity to the input text.

We built the ODP-based semantic interpreter using an ODP snapshot as of April 2004. After pruning the *Top/World* branch that contains non-English material, we obtained a hierarchy of over 400,000 concepts and 2,800,000 URLs. Textual descriptions of the concepts and URLs amounted to 436 Mb of text. In order to increase the amount of training information, we further populated the ODP hierarchy by crawling all of its URLs, and taking the first 10 pages encountered at each site. After eliminating HTML markup and truncating overly long files, we ended up with 70 Gb of additional textual data. After removing stop words and rare words, we obtained 20,700,000 distinct terms that were used to represent ODP nodes as attribute vectors. Up to 1000 most informative attributes were selected for each ODP node using the document frequency criterion[Sebastiani, 2002]. A centroid classifier was then trained, whereas the training set for each concept was combined by concatenating the crawled content of all the URLs classified under this concept.

5.1 Datasets and evaluation procedure

Humans have an innate ability to judge semantic relatedness of texts. Human judgements on a reference set of text pairs can thus be considered correct by definition, a kind of "gold standard" against which computer algorithms are evaluated. Several studies measured inter-judge correlations and found them to be consistently high[Budanitsky and Hirst, 2006; Jarmasz, 2003; Finkelstein *et al.*, 2002a], $r = 0.88 - 0.95$. These findings are to be expected—after all, it is this consensus that allows people to understand each other.

#	Input: *"U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials."*	Input: *"The development of T-cell leukaemia following the otherwise successful treatment of three patients with X-linked severe combined immune deficiency (X-SCID) in gene-therapy trials using haematopoietic stem cells has led to a re-evaluation of this approach. Using a mouse model for gene therapy of X-SCID, we find that the corrective therapeutic gene IL2RG itself can act as a contributor to the genesis of T-cell lymphomas, with one-third of animals being affected. Gene-therapy trials for X-SCID, which have been based on the assumption that IL2RG is minimally oncogenic, may therefore pose some risk to patients."*
1	Iraq disarmament crisis	Leukemia
2	Yellowcake forgery	Severe combined immunodeficiency
3	Senate Report of Pre-war Intelligence on Iraq	Cancer
4	Iraq and weapons of mass destruction	Non-Hodgkin lymphoma
5	Iraq Survey Group	AIDS
6	September Dossier	ICD-10 Chapter II: Neoplasms; Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism
7	Iraq War	Bone marrow transplant
8	Scott Ritter	Immunosuppressive drug
9	Iraq War- Rationale	Acute lymphoblastic leukemia
10	Operation Desert Fox	Multiple sclerosis

Table 2: First ten concepts of the interpretation vectors for sample text fragments.

#	Ambiguous word: "Bank"		Ambiguous word: "Jaguar"	
	"Bank of America"	*"Bank of Amazon"*	*"Jaguar car models"*	*"Jaguar (Panthera onca)"*
1	Bank	Amazon River	Jaguar (car)	Jaguar
2	Bank of America	Amazon Basin	Jaguar S-Type	Felidae
3	Bank of America Plaza (Atlanta)	Amazon Rainforest	Jaguar X-type	Black panther
4	Bank of America Plaza (Dallas)	Amazon.com	Jaguar E-Type	Leopard
5	MBNA	Rainforest	Jaguar XJ	Puma
6	VISA (credit card)	Atlantic Ocean	Daimler	Tiger
7	Bank of America Tower, New York City	Brazil	British Leyland Motor Corporation	Panthera hybrid
8	NASDAQ	Loreto Region	Luxury vehicles	Cave lion
9	MasterCard	River	V8 engine	American lion
10	Bank of America Corporate Center	Economy of Brazil	Jaguar Racing	Kinkajou

Table 3: First ten concepts of the interpretation vectors for texts with ambiguous words.

In this work, we use two such datasets, which are to the best of our knowledge the largest publicly available collections of their kind. To assess word relatedness, we use the WordSimilarity-353 collection[Finkelstein *et al.*, 2002b; 2002a], which contains 353 word pairs. Each pair has 13–16 human judgements, which were averaged for each pair to produce a single relatedness score. Spearman rank-order correlation coefficient[Press *et al.*, 1997] was used to compare computed relatedness scores with human judgements.

For document similarity, we used a collection of 50 documents from the Australian Broadcasting Corporation's news mail service[Lee *et al.*, 2005]. These documents were paired in all possible ways, and each of the 1,225 pairs has 8–12 human judgements. When human judgements have been averaged for each pair, the collection of 1,225 relatedness scores have only 67 distinct values. Spearman correlation is not appropriate in this case, and therefore we used Pearson's linear correlation coefficient.

5.2 results

Table 4 shows the results of applying our methodology to estimating relatedness of individual words. As we can see, both ESA techniques yield substantial improvements over prior studies. ESA also achieves much better results than the other Wikipedia-based method recently introduced.

Table 5 shows the results for computing relatedness of entire documents. On both test collections, Wikipedia-based semantic interpretation is superior to that of the ODP-based one. Two factors contribute to this phenomenon. First, axes of a multi-dimensional interpretation space should ideally be as independent as possible. The hierarchical organization of the ODP reflects the generalization relation between concepts and obviously violates this independence requirement. Second, to increase the amount of training data for building the ODP-based semantic interpreter, we crawled all the URLs listed in the ODP. This allowed us to increase the amount of textual data by several orders of magnitude, but also brought about a non-negligible amount of noise, which is common in

Algorithm	Correlation with human judgements
WordNet-based techniques[Jarmasz, 2003]	0.33–0.35
Roget's Thesaurus-based technique[Jarmasz, 2003]	0.55
LSA[Finkelstein *et al.*, 2002a]	0.56
WikiRelate![Strube and Ponzetto, 2006]	0.19 – 0.48
ESA-Wikipedia	0.75
ESA-ODP	0.65

Table 4: Correlation of word relatedness scores with human judgements on the WordSimilarity-353 collection.

Algorithm	Correlation with human judgements
Bag of words[Lee *et al.*, 2005]	0.1–0.5
LSA[Lee *et al.*, 2005]	0.60
ESA-Wikipedia	0.72
ESA-ODP	0.69

Table 5: Correlation of text relatedness scores with human judgements on Lee et al.'s document collection.

Web pages. On the other hand, Wikipedia articles are virtually noise-free, and mostly qualify as Standard Written English.

In this paper we deal with "semantic relatedness" rather than "semantic similarity" or "semantic distance", which are also often used in the literature. In their extensive survey of relatedness measures, Budanitsky & Hirst[Budanitsky and Hirst, 2006] argued that the notion of relatedness is more general than that of similarity, as the former subsumes many different kind of specific relations, including meronymy, antonymy, functional association, and others. They further maintained that computational linguistics applications often require measures of relatedness rather than the more narrowly defined measures of similarity. For example, word sense disambiguation can use any *related* words from the context, and not merely *similar* words. Budanitsky & Hirst[Budanitsky and Hirst, 2006] also argued that the notion of semantic distance might be confusing due to the different ways it has been used in the literature.

Prior work in the field mostly focused on semantic *similarity* of words, using R&G[Rubenstein and Goodenough, 1965] list of 65 word pairs and M&C[Miller and Charles, 1991] list of 30 word pairs. When only the similarity relation is considered, using lexical resources was often successful enough, reaching the correlation of 0.70–0.85 with human judgements[Budanitsky and Hirst, 2006; Jarmasz, 2003]. In this case, lexical techniques even have a slight edge over ESA-Wikipedia, whose correlation with human scores is 0.723 on M&C and 0.816 on R&G. [1] However, when modelling semantic relatedness, lexical techniques yield substantially inferior results (see Table 4). WordNet-based technique, which only consider the generalization ("is-a") relation between words, achieve correlation of only 0.33–0.35 with human judgements[Budanitsky and Hirst, 2006; Jarmasz, 2003]. Jarmasz & Szpakowicz's ELKB system[Jarmasz, 2003] based on Roget's Thesaurus[Roget, 1852] achieves a higher correlation of 0.55 due to its use of a richer set if relations.

[1] WikiRelate! [Strube and Ponzetto, 2006] achieved very low scores on these domains of 0.31–0.54.

6 Conclusions

The contributions of this paper are twofold. First, we presented a novel technique, called Explicit Semantic Analysis, for representing semantics of natural language texts using natural concepts. This was made possible by using very large scale knowledge repositories, such as Wikipedia and the ODP, which contain hundreds of thousands of human-defined concepts as well as provide a cornucopia of information about each concept. ESA is much more computationally efficient than LSA as there is no need to compute the SVD transformation. Moreover, using natural concepts makes our model easy to interpret, as can be seen in the examples we provided. Second, we evaluated ESA on a prototypical natural language processing task, namely, computing semantic relatedness of texts.

Empirical evaluation confirms that using natural concepts leads to substantial improvements in estimating word and text relatedness. Compared with the previous state of the art, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgements: from $r = 0.56$ to 0.75 for individual words and from $r = 0.60$ to 0.72 for texts. Consequently, we anticipate ESA to give rise to the next generation of natural language processing tools. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.

It is essential to note that Wikipedia is available in numerous languages, while different language versions are cross-linked at the level of concepts. We believe this information can be leveraged to use Wikipedia-based semantic interpretation for improving machine translation.

7 Acknowledgments

We thank Michael D. Lee and Brandon Pincombe for making available their document similarity data. This work was partially supported by funding from the EC-sponsored MUSCLE

Network of Excellence.

References

[Baeza-Yates and Ribeiro-Neto, 1999] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, NY, 1999.

[Buchanan and Feigenbaum, 1982] B. G. Buchanan and E. A. Feigenbaum. Forward. In R. Davis and D. B. Lenat, editors, *Knowledge-Based Systems in Artificial Intelligence*. McGraw-Hill, 1982.

[Budanitsky and Hirst, 2006] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Encyclopaedia Britannica, 2006] Encyclopaedia Britannica, 2006. `http://store.britannica.com` (visited on May 12, 2006).

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Finkelstein *et al.*, 2002a] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January 2002.

[Finkelstein *et al.*, 2002b] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. WordSimilarity-353 test collection, 2002.

[Gabrilovich and Markovitch, 2006] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, Boston, MA, 2006.

[Giles, 2005] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[Han and Karypis, 2000] Eui-Hong (Sam) Han and George Karypis. Centroid-based document classification: Analysis and experimental results. In *PKDD'00*, September 2000.

[Jarmasz, 2003] Mario Jarmasz. Roget's thesaurus as a lexical resource for natural language processing. Master's thesis, University of Ottawa, 2003.

[Lee *et al.*, 2005] Michael D. Lee, Brandon Pincombe, and Matthew Welsh. An empirical evaluation of models of text document similarity. In *CogSci2005*, pages 1254–1259, Austin, TX, 2005.

[Lenat and Guha, 1990] D. Lenat and R. Guha. *Building Large Knowledge Based Systems*. Addison Wesley, Reading, MA, 1990.

[Lenat *et al.*, 1990] Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. CYC: Towards programs with common sense. *Communications of the ACM*, 33(8), August 1990.

[Lenat, 1995] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), November 1995.

[Miller and Charles, 1991] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[ODP, 2006] Open Directory Project, 2006.

[Porter, 1980] M.F. Porter. An algorithm for suffix stripping. *Program - automated library and information systems*, 14(3):130–137, 1980.

[Press *et al.*, 1997] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1997.

[Roget, 1852] Peter Roget. *Roget's Thesaurus of English Words and Phrases*. Longman Group Ltd., Harlow, Essex, England, 1852.

[Rubenstein and Goodenough, 1965] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[Salton and McGill, 1983] G. Salton and M.J. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Sebastiani, 2002] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Singhal *et al.*, 1996] Amit Singhal, Gerard Salton, Mandar Mitra, and Chris Buckley. Document length normalization. *Information Processing and Management*, 32(5):619–633, 1996.

[Strube and Ponzetto, 2006] Michael Strube and Simon Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, Boston, MA, 2006.

[Wikipedia, 2006] Wikipedia, the free encyclopedia, 2006.

[Zobel and Moffat, 1998] Justin Zobel and Alistair Moffat. Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34, 1998.