Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis

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

Abstract

Computing semantic relatedness of natural language texts requires access to vast amounts of common-sense and domain-specific world knowledge. We propose Explicit Semantic Analysis (ESA), a novel method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia. We use machine learning techniques to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts. Assessing the relatedness of texts in this space amounts to comparing the corresponding vectors using conventional metrics (e.g., cosine). 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. Importantly, due to the use of natural concepts, the ESA model is easy to explain to human users.

1 Introduction

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.

It has long been recognized that in order to process natural language, computers require access to vast amounts of common-sense and domain-specific world knowledge [Buchanan and Feigenbaum, 1982; Lenat and Guha, 1990]. However, prior work on semantic relatedness was based on purely statistical techniques that did not make use of background knowledge [Baeza-Yates and Ribeiro-Neto, 1999; Deerwester *et al.*, 1990], or on lexical resources that incorporate very limited knowledge about the world [Budanitsky and Hirst, 2006; Jarmasz, 2003].

We propose a novel method, called Explicit Semantic Analysis (ESA), for fine-grained semantic representation of unrestricted natural language texts. Our method represents meaning in a high-dimensional space of natural concepts derived from Wikipedia (http://en.wikipedia.org), the largest encyclopedia in existence. We employ text classification 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 contributions of this paper are threefold. First, we present Explicit Semantic Analysis, a new approach to representing semantics of natural language texts using natural concepts. Second, we propose a uniform way for computing relatedness of both individual words and arbitrarily long text fragments. Finally, the results of using ESA for computing semantic relatedness of texts are superior to the existing state of the art. Moreover, using Wikipedia-based concepts makes our model easy to interpret, as we illustrate with a number of examples in what follows.

2 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 articles, e.g., COMPUTER SCIENCE, INDIA, or LANGUAGE. An important advantage of our approach is thus the use of vast amounts of highly organized human knowledge encoded in Wikipedia. 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 (compared to 44 million words in 65,000 articles in Encyclopaedia Britannica[1]). Interestingly, the open editing approach yields remarkable quality—a recent study [Giles, 2005] found Wikipedia accuracy to rival that of Britannica.

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

We use machine learning techniques to build a *semantic interpreter* that 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 Latent Semantic Analysis.

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. The choice of encyclopedia articles as concepts is quite natural, as each article is focused on a single issue, which it discusses in detail.

Each Wikipedia concept is represented as an attribute vector of words that occur in the corresponding article. Entries of these vectors are assigned weights using TFIDF scheme [Salton and McGill, 1983]. These weights quantify the strength of association between words and concepts.

To speed up semantic interpretation, we build an *inverted index*, which maps each word into a list of concepts in which it appears. We also use 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 TFIDF scheme. 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 TFIDF 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 vector reflect the relevance of the corresponding concepts to text T. To compute semantic relatedness of a pair of text fragments we compare their vectors using the cosine metric.

Figure 1 illustrates the process of Wikipedia-based semantic interpretation. Further implementation details are available in [Gabrilovich, In preparation].

In our earlier work [Gabrilovich and Markovitch, 2006], we used a similar method for generating features for text categorization. Since text categorization is a supervised learning task, words occurring in the training documents serve as valuable features; consequently, in that work we used Wikipedia concepts to augment the bag of words. On the other hand, computing semantic relatedness of a pair of texts is essentially a "one-off" task, therefore, we replace the bag of words representation with the one based on concepts.

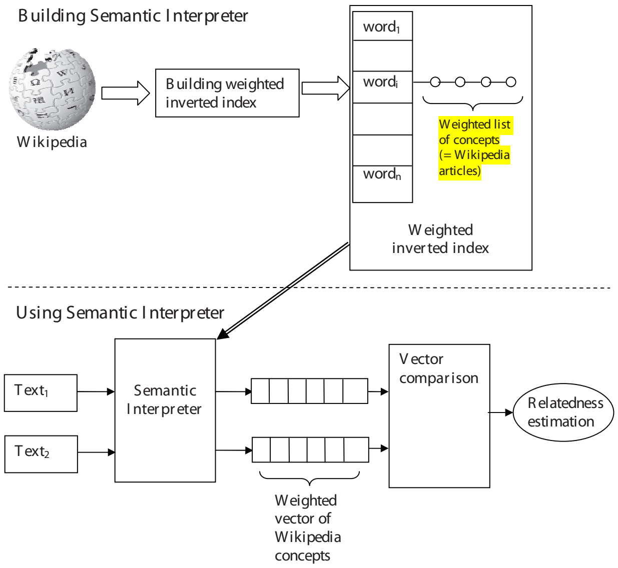
Building Semantic Interpreter
Wikipedia
Building weighted inverted index
word$_1$
word$_i$
word$_n$
Weighted list of concepts (= Wikipedia articles)
Weighted inverted index
Using Semantic Interpreter
Text$_1$
Text$_2$
Semantic Interpreter
Vector comparison
Relatedness estimation
Weighted vector of Wikipedia concepts

Figure 1: Semantic interpreter

#	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 sample interpretation vectors.

To illustrate our approach, we show the ten highest-scoring Wikipedia concepts in the interpretation vectors for sample text fragments. When concepts in each vector are sorted in the decreasing order of their score, the top ten concepts are the most relevant ones for the input text. Table 1 shows the most relevant Wikipedia concepts for individual words ("equipment" and "investor", respectively), while Table 2 uses longer passages as examples. It is particularly interesting to juxtapose the interpretation vectors for fragments that contain ambiguous words. Table 3 shows the first entries in the 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.

3 Empirical 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

#	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.

removing small and overly specific concepts (those having fewer than 100 words and fewer than 5 incoming or outgoing links), 241,393 articles were left. We processed the text of these articles by removing stop words and rare words, and stemming the remaining words; 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, `http://www.dmoz.org`). The ODP is the largest Web directory to date, where concepts correspond to categories of the directory, e.g., Top/Computers/Artificial Intelligence. 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 cataloged under this concept. Further implementation details are available in [Gabrilovich and Markovitch, 2005].

Using world knowledge requires additional computation. This extra computation includes the (one-time) preprocessing step where the semantic interpreter is built, as well as the actual mapping of input texts into interpretation vectors, performed online. On a standard workstation, the throughput of the semantic interpreter is several hundred words per second.

3.1 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.*, 2002], $r = 0.88 - 0.95$. These findings are to be expected—after all, it is this consensus that allows people to understand each other.

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[2] [Finkelstein *et al.*, 2002], which contains 353 word pairs.[3] Each pair has 13–16 human judgements, which were averaged for each pair to produce a single relatedness score. Spearman rank-order correlation coefficient 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.

3.2 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 [Strube and Ponzetto, 2006]. 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 orthogonal as possible. However, the hierarchical organization of the ODP defines the generalization relation between concepts and obviously violates this orthogonality requirement. Second, to increase the amount of training data for building the ODP-based semantic interpreter, we crawled all the URLs cataloged 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 Web pages. On the other hand, Wikipedia articles are virtually noise-free, and mostly qualify as Standard Written English.

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

Table 4: Computing word relatedness

Algorithm	Correlation with humans
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: Computing text relatedness

4 Related Work

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]. Prior work in the field pursued three main directions: comparing text fragments as bags of words in vector space [Baeza-Yates and Ribeiro-Neto, 1999], 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 limitation.

Lexical databases such as WordNet [Fellbaum, 1998] or Roget's Thesaurus [Roget, 1852] encode relations between words such as synonymy, hypernymy. 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; Banerjee and Pedersen, 2003; Resnik, 1999; Lin, 1998; Jiang and Conrath, 1997; Grefenstette, 1992]. 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 and mainly contain information about individual words but little world knowledge in general.

WordNet-based techniques are similar to ESA in that both approaches manipulate a collection of concepts. There are, however, several important differences. First, WordNet-based methods are inherently limited to individual words, and their

[2] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353`

[3] Despite its name, this test collection is designed for testing word relatedness and not merely similarity, as instructions for human judges specifically directed the participants to assess the *degree of relatedness* of the words. For example, in the case of antonyms, judges were instructed to consider them as "similar" rather than "dissimilar".

adaptation for comparing longer texts requires an extra level of sophistication [Mihalcea *et al.*, 2006]. In contrast, our method treats both words and texts in essentially the same way. Second, considering words in context allows our approach to perform word sense disambiguation (see Table 3). Using WordNet cannot achieve disambiguation, since information about synsets is limited to a few words (gloss); in both ODP and Wikipedia concept are associated with huge amounts of text. Finally, even for individual words, ESA provides much more sophisticated mapping of words to concepts, through the analysis of the large bodies of texts associated with concepts. This allows us to represent the meaning of words (or texts) as a weighted combination of concepts, while mapping a word in WordNet amounts to simple lookup, without any weights. Furthermore, in WordNet, the senses of each word are mutually exclusive. In our approach, concepts reflect different aspects of the input (see Tables 1–3), thus yielding weighted multi-faceted representation of the text.

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. The Explicit Semantic Analysis method we proposed circumvents this problem, as it represents meanings of text fragments using natural concepts defined by humans.

Our approach to estimating semantic relatedness of words is somewhat reminiscent of distributional similarity [Lee, 1999; Dagan *et al.*, 1999]. Indeed, we compare the meanings of words by comparing the occurrence patterns across a large collection of natural language documents. However, the compilation of these documents is not arbitrary, rather, the documents are aligned with encyclopedia articles, while each of them is focused on a single topic.

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 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 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, whose correlation with human scores is 0.723 on M&C and 0.816 on R&G.[4] However, when the entire language wealth is considered in an attempt to capture more general semantic relatedness, lexical techniques yield substantially inferior results (see Table 1). WordNet-based techniques, 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 & Szpakowicz's ELKB system [Jarmasz, 2003] based on Roget's Thesaurus achieves a higher correlation of 0.55 due to its use of a richer set if relations.

Sahami and Heilman [2006] proposed to use the Web as a source of additional knowledge for measuring similarity of short text snippets. A major limitation of this technique is that it is only applicable to short texts, because sending a long text as a query to a search engine is likely to return few or even no results at all. On the other hand, our approach is applicable to text fragments of arbitrary length.

Strube and Ponzetto [2006] also used Wikipedia for computing semantic relatedness. However, their method, called WikiRelate!, is radically different from ours. Given a pair of words w_1 and w_2, WikiRelate! searches for Wikipedia articles, p_1 and p_2, that respectively contain w_1 and w_2 in their titles. Semantic relatedness is then computed using 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. On the other hand, our approach represents each word as a weighted vector of Wikipedia concepts, and semantic relatedness is then computed by comparing the two concept vectors.

Thus, the differences between ESA and WikiRelate! are:

1. WikiRelate! can only process words that actually occur in titles of Wikipedia articles. ESA only requires that the word appears within the text of Wikipedia articles.
2. WikiRelate! is limited to single words while ESA can compare texts 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 sophisticated semantic representation based on a weighted vector of Wikipedia concepts.

Indeed, as we have shown in the previous section, the richer representation of ESA yields much better results.

5 Conclusions

We proposed a novel approach to computing semantic relatedness of natural language texts with the aid of very large

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

scale knowledge repositories. We use Wikipedia and the ODP, the largest knowledge repositories of their kind, which contain hundreds of thousands of human-defined concepts and provide a cornucopia of information about each concept. Our approach is called Explicit Semantic Analysis, since it uses concepts explicitly defined and described by humans.

Compared to LSA, which only uses statistical cooccurrence information, 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.

Empirical evaluation confirms that using ESA leads to substantial improvements in computing word and text relatedness. Compared with the previous state of the art, using ESA results in notable 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. Furthermore, due to the use of natural concepts, the ESA model is easy to explain to human users.

6 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.

[Banerjee and Pedersen, 2003] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, pages 805–810, 2003.

[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.

[Dagan *et al.*, 1999] Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, 1999.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

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

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

[Gabrilovich and Markovitch, 2005] Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI'05*, pages 1048–1053, 2005.

[Gabrilovich and Markovitch, 2006] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI'06*, pages 1301–1306, July 2006.

[Gabrilovich, In preparation] Evgeniy Gabrilovich. *Feature Generation for Textual Information Retrieval Using World Knowledge*. PhD thesis, Department of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel, In preparation.

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

[Grefenstette, 1992] Gregory Grefenstette. SEXTANT: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *ACL'92*, pages 324–326, 1992.

[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.

[Jiang and Conrath, 1997] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING'97*, 1997.

[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, 2005.

[Lee, 1999] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the ACL*, 1999.

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

[Lin, 1998] Dekang Lin. An information-theoretic definition of word similarity. In *ICML'98*, 1998.

[Mihalcea *et al.*, 2006] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI'06*, July 2006.

[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.

[Resnik, 1999] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *JAIR*, 11:95–130, 1999.

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

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

[Sahami and Heilman, 2006] Mehran Sahami and Timothy Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW'06*. ACM Press, May 2006.

[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 Comp. Surv.*, 34(1):1–47, 2002.

[Strube and Ponzetto, 2006] Michael Strube and Simon Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI'06*, Boston, MA, 2006.

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