An effective, low-cost measure of semantic relatedness obtained from Wikipedia links

David Milne **Ian H. Witten**

Department of Computer Science, University of Waikato
Private Bag 3105, Hamilton, New Zealand
{dnk2, ihw}@cs.waikato.ac.nz

Abstract

This paper describes a new technique for obtaining measures of semantic relatedness. Like other recent approaches, it uses Wikipedia to provide structured world knowledge about the terms of interest. Our approach is unique in that it does so using the hyperlink structure of Wikipedia rather than its category hierarchy or textual content. Evaluation with manually defined measures of semantic relatedness reveals this to be an effective compromise between the ease of computation of the former approach and the accuracy of the latter.

Introduction

How are *cars* related to *global warming*? What about *social networks* and *privacy*? Making judgments about the semantic relatedness of different terms is a routine yet deceptively complex task. To perform it, people draw on an immense amount of background knowledge about the concepts these terms represent. Any attempt to compute semantic relatedness automatically must also consult external sources of knowledge. Some techniques use statistical analysis of large corpora to provide this. Others use hand-crafted lexical structures such as taxonomies and thesauri. In either case it is the background knowledge that is the limiting factor; the former is unstructured and imprecise, and the latter is limited in scope and scalability.

These limitations are the motivation behind several new techniques which infer semantic relatedness from the structure and content of Wikipedia. With over two million articles and thousands of contributors, this massive online repository of knowledge is easily the largest and fastest growing encyclopedia in existence. With its extensive network of cross-references, portals and categories it also contains a wealth of explicitly defined semantics. This rare combination of scale and structure makes Wikipedia an attractive resource for this work (and for other NLP applications).

This paper describes a new technique—the Wikipedia Link-based Measure—which calculates semantic relatedness between terms using the links found within their corresponding Wikipedia articles. Unlike other techniques based on Wikipedia, WLM is able to provide accurate measures efficiently, using only the links between articles rather than their textual content. Before describing the details, we first outline the other systems to which it can be compared. This is followed by a description of the algorithm, and its evaluation using several manually defined gold standards. The paper concludes with a discussion of the strengths and weaknesses of the new approach.

Related Work

The purpose of semantic relatedness measures is to allow computers to reason about written text. They have many applications in natural language processing and artificial intelligence (Budanitsky, 1999), and have consequently received a lot of attention from the research community. Table 1 shows the performance of various semantic relatedness measures according to their correlation with a manually defined gold standard; namely Finkelstein *et al*'s (2002) WordSimilarity-353 collection.

The central point of difference between the various techniques is their source of background knowledge. For the first two entries in the table, this is obtained from manually created thesauri. WordNet and Roget have both been used for this purpose (McHale, 1998). Thesaurus-based techniques are limited in the vocabulary for which they can provide relatedness measures, since the structures they rely on must be built by hand.

measure	accuracy
Thesuarus based	
Wordnet	0.33-0.35
Roget	0.55
Corpus based	
Latent Semantic Analysis (LSA)	0.56
Wikipedia based	
WikiRelate	0.19-0.48
Explicit Semantic Analysis (ESA)	0.75

Table 3: Performance of existing semantic relatedness measures (from Gabrilovich and Markovitch, 2007)

Corpus-based approaches obtain background knowledge by performing statistical analysis of large untagged document collections. The most successful and well known of these techniques is Latent Semantic Analysis (Landauer *et al.*, 1998), which relies on the tendency for related words to appear in similar contexts. LSA offers the same vocabulary as the corpus upon which it is built. Unfortunately it can only provide accurate judgments when the corpus is very large, and consequently the pre-processing effort required is significant.

Strube and Ponzetto (2006) were the first to compute measures of semantic relatedness using Wikipedia. Their approach—WikiRelate—took familiar techniques that had previously been applied to WordNet and modified them to suit Wikipedia. Their most accurate approach is based on Leacock & Chodorow's (1998) path-length measure, which takes into account the depth within WordNet at which the concepts are found. WikiRelate's implementation does much the same for Wikipedia's hierarchical category structure. While the results are similar in terms of accuracy to thesaurus based techniques, the collaborative nature of Wikipedia offers a much larger—and constantly evolving—vocabulary.

Gabrilovich and Markovitch (2007) achieve extremely accurate results with ESA, a technique that is somewhat reminiscent of the vector space model widely used in information retrieval. Instead of comparing vectors of term weights to evaluate the similarity between queries and documents, they compare weighted vectors of the Wikipedia articles related to each term. The name of the approach—Explicit Semantic Analysis—stems from the way these vectors are comprised of manually defined concepts, as opposed to the mathematically derived contexts used by Latent Semantic Analysis. The result is a measure which approaches the accuracy of humans. Additionally, it provides relatedness measures for any length of text: unlike WikiRelate, there is no restriction that the input be matched to article titles.

Obtaining Semantic Relatedness from Wikipedia Links

We have developed a new approach for extracting semantic relatedness measures from Wikipedia, which we call the Wikipedia Link-based Measure (WLM). The central difference between this and other Wikipedia based approaches is the use of Wikipedia's hyperlink structure to define relatedness. This theoretically offers a measure that is both cheaper and more accurate than ESA: cheaper, because Wikipedia's extensive textual content can largely be ignored, and more accurate, because it is more closely tied to the manually defined semantics of the resource.

Wikipedia's extensive network of cross-references, portals, categories and info-boxes provide a huge amount of explicitly defined semantics. Despite the name, Explicit Semantic Analysis takes advantage of only one property: the way in which Wikipedia's text is segmented into individual topics. It's central component—the weight between a term and an article—is automatically derived rather than explicitly specified. In contrast, the central component of our approach is the link: a manually-defined connection between two manually disambiguated concepts. Wikipedia provides millions of these connections, as

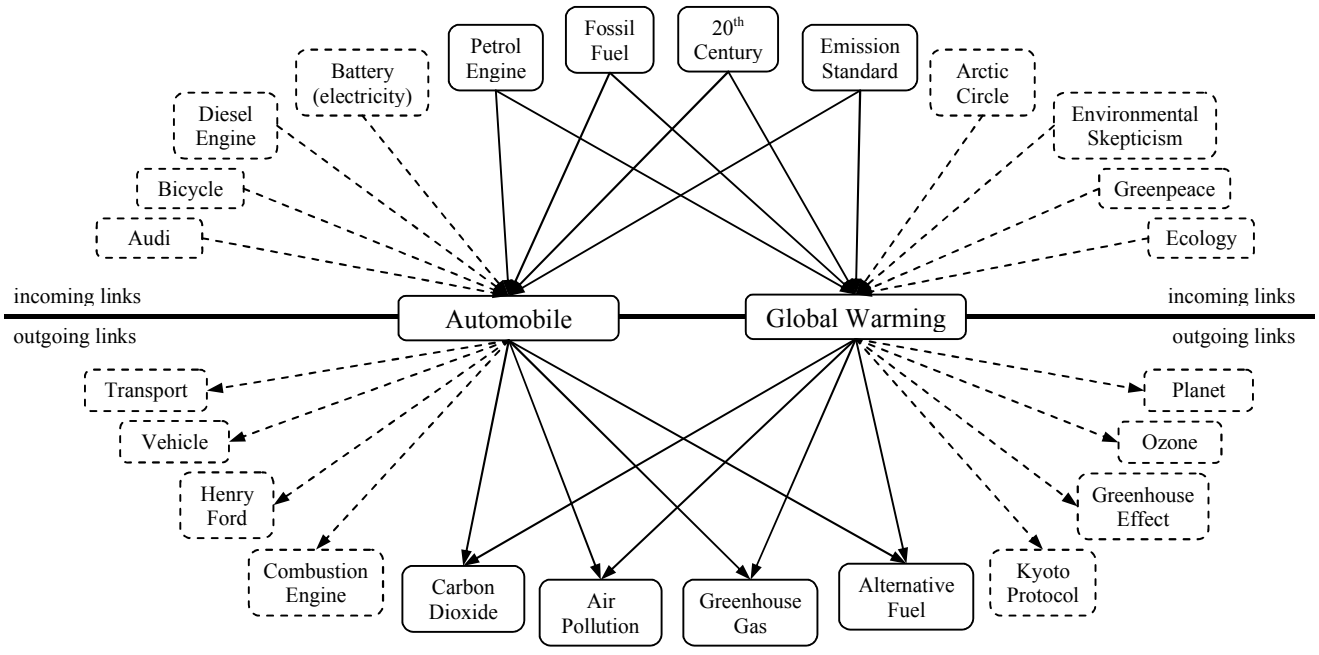

Figure 1: Obtaining a semantic relatedness measure between Automobile *and* Global Warming *from Wikipedia links.*

Figure 1 illustrates by attempting to answer the question posed at the start of the paper. It displays only a small sample—a mere 0.34%—of the links available for determining how *automobiles* are related to *global warming*. While the category links used by WikiRelate are also manually defined, they are far less numerous. On average, articles have 34 links out to other articles and receive another 34 links from them, but belong to only 3 categories.

The remainder of this section elaborates on our approach, and the various options we experimented with. These are revisited in the evaluation section, which examines the options individually and identifies the most effective ones.

Identifying candidate articles

The first step in measuring the relatedness between two terms is to identify the concepts they relate to: in Wikipedia's case, the articles which discuss them. This presents two problems: ambiguity and polysemy.

Ambiguity is the tendency for terms to relate to multiple concepts: for example *plane* might refer to a fixed-wing aircraft, a theoretical surface of infinite area and zero depth, or a tool for flattening wooden surfaces. The correct sense depends on the context of the term to which we are comparing it to; consider the relatedness of *plane* to *wing*, and *plane* to *surface*.

Polesemy is the tendency for concepts to be known by multiple names: a *plane* may also be referred to as *fixed wing aircraft, airplane* or *aeroplane*. It must be possible navigate to the appropriate article (and thus obtain the same relatedness measure) with any of these synonyms.

We use anchors—the terms or phrases in Wikipedia articles texts to which links are attached—to identify candidate articles for terms. Wikipedia's documentation dictates that any term or phrase that relates to a significant topic should be linked to the article that discusses it. Consequently it provides a vast number of anchor texts which capture both ambiguity and polysemy: *plane* links to different articles depending on the context in which it is found, and *plane, airplane* and *aeroplane* are all used to link to the same article.

Measuring relatedness between articles

The next step in obtaining a similarity measure between two terms is to judge the similarity between their representative articles. We have experimented with two measures. One is based the links extending out of each article, the other on the links made to them. These correspond to the bottom and top halves of Figure 1.

The first measure is defined by the angle between the vectors of the links found within the two articles of interest. These are almost identical to the TF×IDF vectors used extensively within information retrieval. The only difference is that we use link counts weighted by the probability of each link occurring, instead of term counts weighted by the probability of the term occurring. This probability is defined by the total number of links to the target article over the total number of articles. Thus if *s* and *t* are the source and target articles, then the weight *w* of the link $s \to t$ is:

$$w(s \to t) = \log\left(\frac{|W|}{|T|}\right) \quad \text{if } s \in T,\ 0 \text{ otherwise}$$

where *T* is the set of all articles that link to *t*, and *W* is the set of all articles in Wikipedia. In other words, the weight of a link is the inverse probability of any link being made to the target, or 0 if the link does not exist. Thus links are considered less significant for judging the similarity between articles if many other articles also link to the same target. The fact that two articles both link to *science* is much less significant than if they both link to a specific topic such as *atmospheric thermodynamics*.

These link weights are used to generate vectors to describe each of the two articles of interest. The set of links considered for the vectors is the union of all links made from either of the two source articles. The remainder of the approach is exactly the same as in the vector space model: the similarity of the articles is given by the angle (cosine similarity) between the vectors. This ranges from 0^o if the articles contain identical lists of links to 90^o if there is no overlap between them.

The second measure we use is modeled after the Normalized Google Distance (Cilibrasi and Vitanyi, 2007), which is based on term occurrences on web-pages. The name stems from the use of the Google search engine to obtain pages which mention the terms of interest. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite. Our measure is based on Wikipedia's links rather than Google's search results. Formally, the measure is:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where *a* and *b* are the two articles of interest, *A* and *B* are the sets of all articles that link to *a* and *b* respectively, and—as before—*W* is the entire Wikipedia.

Measuring relatedness between terms

Once the candidate articles have been identified and the relatedness between them calculated, we work backwards to identify the relatedness between the original pair of terms. As with the previous step, there are several options.

First, one can use the two terms involved to disambiguate each other. For example, when identifying the relatedness of *jaguar* and *car* it makes sense to use *car* to determine that we are talking about the automobile manufacturer *Jaguar Cars Ltd*, rather than the species of cat. This amounts to selecting the two candidate senses which most closely relate to each other.

Second, one can make a snap decision by using the most common sense of each term. For example, when making a judgment between *Israel* and *Jerusalem,* one would consider only the nation and its capital city. The obscure

but strong connection between two townships in Ohio with the same names would be completely ignored. The commonness of a sense is defined by the number of times the term is used to link to it: e.g. 95% of *Israel* anchors link to the nation, 2% to the football team, 1% to the ancient kingdom, and a mere 0.1% to the Ohio township.

Finally, we can also consider the case where two words are closely related because they belong together in the same phrase: e.g. *family planning* is a well-known phrase, and consequently *family* and *planning* are given a high semantic relatedness even though their respective concepts are relatively disjoint. For these cases we simply concatenate the terms and add the logarithm of the frequency with which this is used as an anchor to the relatedness score of the original terms.

Evaluation

Our natural ability as humans to disambiguate topics and judge their relatedness can be considered the gold standard against which any automatically generated semantic relatedness measure should be evaluated. We evaluated our approach on three datasets of term pairs and manually defined relatedness measures: Miller and Charles' (1991) list of 30 term pairs, Rubenstein and Goodenough's (1965) 65 pairs, and the WordSimilarity-353 dataset described under Related Work.

The version of Wikipedia used to obtain our measures was released on November 20, 2007. At this point it contained approximately 13GB of uncompressed XML markup. This relates to just under two million articles, which constitute the various concepts for which semantic relatedness judgments were available. We also mined over five million distinct anchors, which defines the vocabulary of terms by which these concepts can be accessed.

Evaluation of algorithm components

This section evaluates the options presented above, in order to identify the best ones and define the final algorithm. This was done over a subset of 50 randomly selected term pairs from the WordSimilarity-353 collection, to avoid over-fitting the algorithm to the data. To gain further insights into the decisions the algorithm was making, the terms within the subset were manually disambiguated to the appropriate article. This provides a gold standard of *article* pairs—as opposed to *term* pairs—and manually defined measures of relatedness between them.

Recall that the candidate selection process depends on anchor text. All of the 95 distinct terms in the 50 pair subset were used as anchors in Wikipedia, and in all cases the correct sense of the term was one of the anchor's destinations. All but one of the terms were ambiguous, with an average of 42 senses per term and a maximum of 371 senses (for *king*). This highlights one of the weaknesses of using anchors in this way: they are often linked to instances of the concept (in this case, specific kings) rather than senses of the term. Consequently, for efficiency and accuracy's sake we only consider articles which receive at least 1% of the anchor's links. This theoretically leaves up to 100 candidates to be examined, but in practice the distribution of links for each anchor follows the power law, meaning that the vast majority are made to a handful of candidates. In the sample, the largest number of candidates examined for a term was 26.

In our description above we proposed two measures of relatedness between articles: a *TF×IDF inspired* comparison of the links found within the articles, and a *Google Distance inspired* comparison of the links made to each article. To evaluate these measures independently from the disambiguation task, we manually identified the correct articles for each term pair in the test set, and computed the correlation between manually defined gold standard measures and those provided by each approach. Table 2 shows the results, where accuracy is defined as the correlation coefficient between the automatic and manually defined measures. It clearly identifies *Google Distance* as the more accurate measure. It also shows that a modest gain can be made by taking the average of the measures; this is the approach used in the remainder of the paper.

Table 3 shows the correlation with manual judgments when the relatedness between terms is automatically inferred from the relatedness between candidate concepts. Good results are obtained when disambiguation is not performed at all, but instead all the different candidate senses contribute to the final relatedness value. That is, the *weighted average relatedness* (weighted by commonness of senses) across all candidate senses is just as accurate as the relatedness between manually identified ideal senses (in Table 2, row 3).

Merely selecting the most well known sense for each term (*most common pair*) performs surprisingly well. Selecting the *most closely related pair* of senses performs better, but is marred by the number of obscure senses available. Evenly weighting the candidate senses by their commonness and relatedness and choosing the pair with the highest weight—*highest (commonness+relatedness)*—gives exactly the same results as with manual

measure	accuracy
TF×IDF inspired	0.66
Google Distance inspired	0.72
combined (average)	0.74

Table 2: Accuracy of semantic relatedness measures between manually selected articles

measure	accuracy
weighted average relatedness	0.74
most common pair	0.68
most closely related pair	0.69
highest (commonness + relatedness)	0.74
sequential decision	0.75
final relatedness measure	0.78

Table 3: Accuracy of relatedness measures (and disambiguation strategies) between original terms

disambiguation (Table 2, row 3). Interestingly we can improve upon the gold standard by making a simple *sequential decision*, which first groups the most related pairs of candidates (within 40% of the most related pair) and then chooses the most common pair. This makes subtly different—but equally valid—choices from those in the gold standard. Given the term *environment* and the context *ecology*, for example, the system selects *ecosystem* as the representative article rather than *natural environment*, and consequently obtains a more accurate relatedness score. The approach is further improved by adding the normalized frequency of the concatenated anchor, to give our *final relatedness measure*, which has a correlation of 0.78 with manual judgments over the sample dataset.

Comparison to alternative approaches

Now we evaluate our approach as a whole by comparing it to the other methods described in Related Work. Table 4 compares the algorithm with its two main competitors—WikiRelate and ESA—by their correlation with gold standard manually defined judgments. Only the best measures obtained by the different approaches are shown. Across all three datasets, we see a consistent trend: WLM is better than WikiRelate but worse than ESA. The final row in the table combines the results across the three datasets, with correlations weighted by the size of each dataset. This shows WLM outperforming WikiRelate by 0.19, and in turn being outperformed by ESA by 0.08.

The third row in Table 4 shows the performance of the algorithms over the WordSimilarity-353 collection. The accuracy of 0.69 for our system can be directly compared to the results in Table 1, which were obtained from the same dataset. Our algorithm outperforms all others except ESA by at least 0.13.

It is interesting to note the drop in WLM's performance on 50 term sample used in the previous section (0.78) and the full WordSimilarity-353 collection used here (0.69). Much of the drop may be due to over-fitting the algorithm to the sample set. Analysis of the results, however, reveals another reason: WLM differs most from the gold standard when the terms being compared cannot be resolved to suitable Wikipedia articles. For example, there is no article for the concept *defeat*; the anchor points only to specific military encounters. These cases are common in the full 353 set but were, by chance, excluded from the sample.

Figure 2 plots the performance of WLM as successively more pairs are discarded from the 353 collection set so that only the most well defined terms are considered. We use anchor frequency as a simple indicator of how well a term is defined; if a term is not used to make a sufficient number of links, it is considered problematic. From this, WLM's performance can clearly be seen to approach the benchmark of 0.75 set by ESA when the terms involved are well defined as individual articles in Wikipedia.

dataset	WikiRelate	ESA	WLM
Miller and Charles	0.45	0.73	0.70
Rubenstein and Goodenough	0.52	0.82	0.64
WordSimilarity-353	0.49	0.75	0.69
Weighted average	*0.49*	*0.76*	*0.68*

Table 4: Accuracy of semantic relatedness measures for three standard datasets.

Discussion

The previous section clearly identifies ESA as the best measure. It is less brittle, because it only requires that articles mention the terms involved. WLM, however, achieves competitive levels of accuracy when the terms involved correspond to distinct Wikipedia articles. Given Wikipedia's sheer size and rate of growth, one can expect this to hold true whenever the terms represent topics which one could reasonably write an article about. This is the case for most applications in the literature, which deal primarily with topics: named entities (Bunescu and Pasca, 2006; Cucerzan, 2007); key phrases (Mihalcea and Csomai, 2007); categories (Gabrilovich and Markovitch, 2006); or entries in existing ontologies (Medelyan and Legg, 2008) and thesauri (Ruiz-Casado *et al.*, 2005). In these applications, we can expect WLM to be competitive with the state of the art.

The advantage of our approach is that it requires far less data and resources. To obtain measures from ESA, one must preprocess a vast amount of textual data; 13Gb as of November 2007. Each term must be matched to the articles in which it is found, and each of the resulting lists of articles must be weighted, sorted, and pruned. One assumes (given the sorting requirement) that this is a log-linear at best. By comparison, WLM requires only the link structure of Wikipedia (450 Mb) and the statistics of how often anchors are used to link to different concepts (140Mb). No preprocessing is required other than to extract this information from Wikipedia's XML dumps. This is a straight-forward task that can be achieved in linear time (assuming constant hash-table operations).

Another advantage is the accessibility of our approach. At the time of writing, ESA is not publicly available. The only known re-implementation is based on the smaller

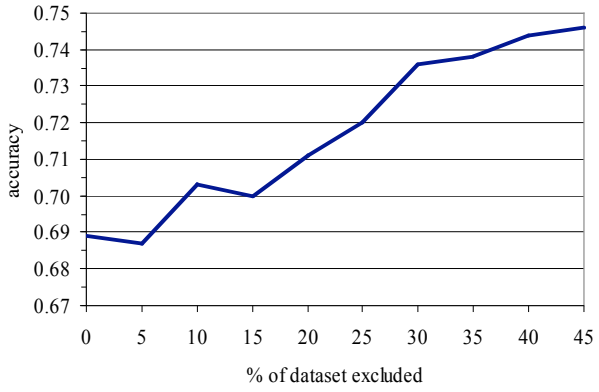

Figure 2: Accuracy of WLM with poorly defined terms excluded.

German version of Wikipedia and a very restricted vocabulary of 10,000 terms (Jacobs, 2007). WLM in contrast is readily available as part of the open source WikipediaMiner toolkit.[1] This implementation provides measures for the full anchor vocabulary of whatever version of Wikipedia it is applied to: currently more than 5 million distinct phrases for the English language version. Other language versions have not yet been tested, but in principle the approach is language independent.

Conclusion

In this paper we propose and evaluate WLM, a novel approach to computing semantic relatedness with the aid of Wikipedia. Our approach is most similar to WikiRelate and ESA, which also exploit the online encyclopedia for this purpose. WLM forms a compromise between these very different methods by utilizing Wikipedia's network of inter-article links, rather than the comparatively small category hierarchy used by the former system, or the full textual content used by the latter.

Our measure consistently outperforms WikiRelate and all previous approaches across all datasets. ESA remains the best measure in terms of robustness; however, we are able to match its accuracy when the terms involved correspond to topics that are well-defined in Wikipedia. Given the number of potential applications for which this requirement holds, we consider WLM to be a valuable contribution. For many tasks we expect it to be competitive with ESA, while using far less data and resources.

Future work will involve applying WLM to various tasks in order to investigate its utility more fully. Strube and Ponzetto (2006) have rightly pointed out the danger in using a few subjective and relatively small datasets for evaluation. Like them, we hope to apply our work to a host of NLP tasks that will require hundreds of thousands of relatedness judgments to be made, and thus provide a more reliable evaluation. Please, download the code, apply it to your own NLP problems, and help us in this endeavor!

References

Budanitsky, A. (1999) *Lexical Semantic Relatedness and its Application in Natural Language Processing*. Ph.D. thesis, University of Toronto, Ontario.

Bunescu, R. and Pasca, M. (2006) Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy.

Cilibrasi, R.L. and Vitanyi, P.M.B. (2007) The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering 19*(3), 370-383.

Cucerzan, S. (2007) Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2007)*, Prague, Czech Republic.

Finkelstein, L., Gabrilovich, Y.M., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002) Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*(1).

Gabrilovich, E. and Markovitch, S. (2006) 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.

Gabrilovich, E. and Markovich, S. (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India.

Jacobs, H. (2007) *Explicit Semantic Analysis (ESA) using Wikipedia*. Retrieved March 14, 2008 from http://www.srcco.de/v/wikipedia-esa

Landauer, T.K. and Foltz, P.W. and Laham, D. (1998) An introduction to latent semantic analysis. *Discourse Processes 25*(2-3), pp 259-284.

Leacock, C. & M. Chodorow (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet. An Electronic Lexical Database*, Chp. 11, pp. 265.283. Cambridge, Mass.: MIT Press.

McHale, M. (1998) A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity, In *Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montréal, Canada. pp. 115-120.

Medelyan, O. and Legg, C. (2008) What CYC can learn from Wikipedia. Submitted to *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008)*.

Mihalcea, R. and Csomai, A. (2007) Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on Information and Knowledge Management (CIKM)*, Lisbon, Portugal. pp 233-242.

Miller, G. A. & W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*(1):1-28.

Rubenstein, H. & J. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10):627.633.

Ruiz-Casado, M., Alfonseca, E., Castells, P. (2005) Automatic assignment of Wikipedia Encyclopedic Entries to WordNet synsets. In *Proceedings of Advances in Web Intelligence*, Lodz, Poland.

Strube, M. and Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, pp.1419-1424.

[1] An open source implementation of WLM is available at www.sourceforge.net/WikipediaMiner