Wikipedia is a Practical Alternative to the Web for measuring Co-occurrence based Word Association

Om P. Damani, Pankhil Chedda, Dipak Chaudhari
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{damani,pankhil,dipakc}@cse.iitb.ac.in

Abstract

While the World Wide Web is an attractive resource, few researchers can access or manage a Web-scale corpus. Instead they use search-hit counts as a substitute for direct measurements on a web corpus. In contrast, one can download a small high quality corpus like Wikipedia and carry out exact measurements. By extensive experiments with multiple word-association measures and several public datasets, we show that for exploring document level co-occurrence based word associations, despite being three orders of magnitude smaller in size, the Wikipedia is a reasonable alternative to a web corpus that can only be accessed using search engines.

Further, with Wikipedia, one can carry out measurements at a granularity finer than document scale. Instead of document level co-occurrence, one can consider a word-pair occurrence significant, only if the two words occur within a certain threshold distance of each-other. In general, such fine-grained information cannot be obtained from search engines. Our experiments show that the word level co-occurrence measures perform better than the document level measures. This indicates another practical advantage of the Wikipedia, or any other downloadable corpus, over a Web corpus which can only be accessed using search engines.

1. Introduction

The World Wide Web is an attractive resource for carrying out the NLP research. If one does not need the entire document contents and can just work with the frequency information of certain document types, then using the APIs provided by various search engines, one can use the Web as a corpus and need not collect and manage a corpus. An example area where one can take advantage of these APIs is the measurement of word association based on lexical co-occurrence [1].

The notion of *word association* is important for numerous NLP applications, like, information retrieval, question-answering, word sense disambiguation, optical character recognition, speech recognition, parsing, lexicography, text summarization, natural language generation, and machine translation. In [2] word association is motivated as *the basis for a statistical description of a variety of interesting linguistic phenomena*.

While the traditional co-occurrence based word association measures are formulated in terms of the word frequencies, it is straight-forward to reformulate them as working with the document frequencies. As an example, consider the the popular word association measure PMI [2]. It is defined as:

$$PMI(x,y) = log\frac{p(x,y)}{p(x)p(y)}$$

where $p(x)$ and $p(y)$ are unigram probabilities and $p(x,y)$ is bigram probabilities. These probabilities are obtained by dividing $f(x), f(y), f(x,y)$ by corpus-size in words. $f(x), f(y)$are the number of occurrences of words x and y in the corpus, i.e. the unigram frequencies of x and y, and $f(x,y)$ is the number of occurrences of the word-pair (x,y) in the corpus, i.e. the bigram frequency of (x,y).

To work with web corpus, one can simply replace word-frequencies with document frequencies, provided one knows the number of documents in the corpus, an information that is generally not available when using search engines. Instead, note that we need not work with probabilities. We can directly work with document frequencies since we are only interested in relative rankings of word-pairs and not in their absolute PMI values. Hence, as discussed later in Section 3.1, ignoring the corpus size does not affect any of our ranking based results. Therefore, we redefine PMI as:

$$PMI(x,y) = log\frac{n(x,y)}{n(x)n(y)}$$

where $n(x,y), n(x), n(y)$ are the counts of documents containing both words x and y, only x, and only y respectively. In the same way, we can redefine most

other word-association measures in terms of document frequencies.

Though the large number of documents available on the Web are an attractive resource, Kilgarriff argues in [3] that "Googleology is bad science". One of the reasons cited there is the unreliability of the document counts obtained. After giving that warning, Kilgarriff accepts that "With enormous data, you get better results", and exhorts the readers to "make resources on this scale available".

Given that very few researchers can afford to access or manage a Web-scale corpus, only alternative they are left with is to use search-hit counts as a substitute for doing direct measurements on a Web-scale corpus. However, as argued in [3] and other places, for various performance and cost reasons, search-hit counts provided by search-engines are only crude approximations and poor substitute for actual Web statistics.

Given these limitations of working with Web, in this work, we argue that for applications like determining the word association, the quality of the data is much more important than the quantity. We find that using a Wikipedia dump containing 2.7 million documents gives better word association results than using the Yahoo and Bing search engines which indexed roughly 3.5 billion and 12 billion pages respectively[1] at the time of our experiments. Hence if a researcher cannot afford a Web-scale corpus, then it is better to work with a Wikipedia dump, than to use search-hit counts, at least for measuring word association.

Further, with Wikipedia, one can carry out measurements at a granularity finer than document scale. Instead of document level co-occurrence, one can consider a word-pair occurrence significant, only if the two words occur within a certain threshold distance of each-other. In general, such fine-grained information cannot be obtained from search engines. Our experiments show that the word level co-occurrence measures perform better than the document level measures. This indicates another practical advantage of the Wikipedia, or any other downloadable corpus, over a Web corpus which can only be accessed using search engines.

2. Related Work

The existing word association measures can be divided into three broad categories:

Frequency based measures rely on co-occurrence frequencies of both words in a corpus in addition to the individual unigram frequencies.

Distributional Similarity based measures based on Firth's "You shall know a word by the company it keeps" [4], these measures characterize a word by the distribution of other words around it and compare two words for distributional similarity [5, 6, 7, 8].

Knowledge-based measures rely on knowledge-sources like thesauri, semantic networks, or taxonomies [9, 10, 11, 12, 13, 14].

In this work, our focus is on choice of resources for frequency based co-occurrence measures, and we do not discuss the details of the distributional similarity and knowledge based measures.

Chklovski and Pantel [15] have mined the web for fine-grained semantic relations such as similarity, strength, antonymy, enablement, and temporal happens-before relations between a pair of verbs. Mihalcea et. al. [16] measure the semantic similarity of short texts using several knowledge based and corpus based measures. They use the Microsoft paraphrase corpus [17], which was constructed by automatically collecting potential paraphrases from thousands of news sources on the Web over a period of 18 months. In [18], a new co-occurrence measure called Co-occurrence Significance Ratio is introduced and it is compared with a host of other measures using a Wikipedia corpus.

Although previously mentioned researchers have used the Web [15, 16] or the Wikipedia [18] for computing co-occurrence measures, to our knowledge no-one has performed a comparative study of the Web vs. the Wikipedia.

Information from the Wikipedia, such as its link structure [9], its concepts [11, 12], and its category trees [13] has earlier been used for knowledge-based word association measures. It is not surprising that the Wikipedia has been found useful for the knowledge-based measures. What is somewhat surprising is that the accurate measurements over Wikipedia give better results than the crude search hit counts from the Web for exploring even lexical co-occurrence based word associations, where one would expect that the much bigger corpus would always give better results due to the law of large numbers.

3. Wikipedia vs. Web

The advantage of the Web as a corpus is that it takes very little effort to work with it. However, while it is easy to replicate experiments on traditional corpora, the Web content keeps changing. In addition, the indexing and search strategies of the commercial search engines also change over time. Hence, it is hard to rerun the Web based experiments for reproducibility. Still, given the advantage of size and the ease of effort, it is worth exploring whether co-occurrence measures performs better

[1] Source: http://www.worldwidewebsize.com/

Table 1: Definition of Co-occurrence based word association measures.

Measure	Document Count	Word Count
Dice	$\frac{2n(x,y)}{n(x)+n(y)}$	$\frac{2\hat{f}(x,y)}{f(x)+f(y)}$
Jaccard	$\frac{n(x,y)}{n(x)+n(y)-n(x,y)}$	$\frac{\hat{f}(x,y)}{f(x)+f(y)-\hat{f}(x,y)}$
Ochiai	$\frac{n(x,y)}{\sqrt{n(x)n(y)}}$	$\frac{\hat{f}(x,y)}{\sqrt{f(x)f(y)}}$
PMI	$log\frac{n(x,y)}{n(x)n(y)}$	$log\frac{\hat{p}(x,y)}{p(x)p(y)}$
SCI	$\frac{n(x,y)}{n(x)\sqrt{n(y)}}$	$\frac{\hat{p}(x,y)}{p(x)\sqrt{p(y)}}$

$n(x,y)$	Total number of documents in the corpus having at-least one occurrence of (x,y)
$n(x), n(y)$	the number of documents in the corpus containing at least one occurrence of x and y respectively
$f(x), f(y)$	unigram frequencies of x, y in the corpus
$\hat{f}(x,y)$	span-constrained bigram frequency of x, y in the corpus
N	Total number of tokens in the corpus
$\hat{p}(x,y), p(x), p(y)$	$\hat{f}(x,y)/N, f(x)/N, f(y)/N$

or worse with the Web than with a much smaller corpus like Wikipedia.

3.1. Co-occurrence based Association Measures

To compare the performance of the Web and the Wikipedia, we experiment with six different co-occurrence based word association measures: Dice [19], Jaccard [20], Ochiai [21], Pointwise Mutual Information - PMI [2], and Semi-Conditional Information - SCI [22]. Their definitions are given in Table 1. Except SCI, all other measures are well-established and besides language processing, have been used in several domains like ecology, psychology, and medicine.

The word count based definitions are discussed later in Section 4. In this section, we are concerned with document count based definitions only. It is important to note that the word-count based versions count span-constrained bigram occurrences while the document based versions do not take span into account, since for the Web we do not have access to the span information.

Our results show that the Jaccard and the Dice have almost identical performance, since $[n(x,y) \ll n(x)]$ and $[n(x,y) \ll n(y)]$ for most word pairs. Hence we do not distinguish between these measures when presenting our results.

We have not experimented with other popular measures like the Log Likelihood Ratio - LLR [23], and the T-test, since their definitions require knowing the total number of documents in the corpus. Technically, knowledge of the corpus size in documents is needed even for our chosen measures, but we can ignore the corpus size by working with a scaled versions of these measures. For example, in the definition of PMI given earlier, technically all three terms $n(x,y), n(x)$ and $n(y)$ should be divided by the corpus size, but ignoring the corpus size does not affect any of our ranking based results, while the same cannot be said of the LLR and the T-Test.

As explained later, we evaluate a measure on a given dataset by the Spearman rank correlation coefficient between the word-associations produced by the measure and the gold-standard ratings for the dataset. The Spearman rank correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. Since any monotonic transformation of the word association scores produced by a measure leaves the rankings unchanged, the modified scores obtained by ignoring the corpus size leaves the rankings unchanged.

3.2. Datasets and Resources

The two major types of word associations discussed in literature are *free association* and *semantic relatedness*.

Free association refers to the first response given by a subject on being given a stimulus word [24]. The standard methodology for collecting *free association* data is explained at [24]: "Native speakers are presented with stimulus words and are asked to write down the first word that comes to mind for each stimulus. The degree of free association between a stimulus (S) and response (R) is then quantified by the percentage of test subjects who produced R when presented with S."

We use five different publicly available datasets for measuring free association: Kent [26], Minnesota [27], White-Abrams [28], Goldfarb-Halpern [25], and

Table 2: Characteristics of data sets used. 'Respondents' is the number of individuals who were asked to respond to a given set of stimulus words. 'Word Pairs' is the total number of unique (stimulus,response) pairs generated. 'Filtered Word Pairs' is the size of the subset of the corresponding dataset used in our experiments.

Aspect	Data Set	No. of Respondents	No. of Word-Pairs	No. of Filtered Word-Pairs
Semantic relatedness	wordsim353 [14]	16	353	351
Free-Association	Essli [24]	100	272	272
	Goldfarb-Halpern [25]	316	410	384
	Kent [26]	1,000	14,576	14,086
	Minnesota [27]	1,007	10,447	9,649
	White-Abrams [28]	440	745	652

Essli [24].

The *semantic relatedness* encompasses relations like synonymy, meronymy, antonymy, and functional association [29]. We use the publicly available Wordsim [14] dataset to measure the semantic relatedness. The word association scores in this dataset are the average of the values on a scale from 0 to 10 given by the respondents when they were asked to estimate the relatedness of the words in a given pair.

One could say that free association datasets are asymmetric datasets where one stimulus words occur with multiple response words. In contrast, semantic relatedness datasets are symmetric in that both words in a pair have the same status.

Many of these datasets contain multi-word expressions. We removed word-pairs containing multiword expressions. For data sets with more than 10,000 word-pairs, we filtered out pairs that contain stop words listed in [30]. The details of the dataset after filtering is given in Table 2.

3.3. Corpus

We use the a Wikipedia dump with 2.7 million documents and of size 1.24 Gigawords. We used Lucene[2] APIs to obtain various statistics from the corpus. No function-word removal, lemmatization or any other preprocessing was performed on the corpus by us other than whatever preprocessing is done by default by Lucene.

For the web search, we use Yahoo and Bing search services. For Yahoo, we use BOSS API[3]. For Bing, we issue simple search requests and parse the response pages to obtain the hit count. In both cases, we use boolean conjunctive queries to get the count of documents containing both words in the pair. We could not use Google Search since it allows only 1000 queries a day.

[2]http://lucene.apache.org/

[3]http://developer.yahoo.com/search/boss/

3.4. Evaluation Methodology

For the word-pairs in each dataset, each measure under consideration produces a ranked list of the word association scores. We also have the gold-standard human judgment ranking available for each dataset. We follow the standard methodology of evaluating a word-association measure on a given dataset by the Spearman rank correlation coefficient between the word-associations produced by the measure and the gold-standard ratings for the dataset.

3.5. Results

The purpose of our experiments is not to compare the word association measures but to compare a downloadable Wikipedia corpus with a Web corpus that can be accessed using only search-engine interfaces. That is, we wish to find out how the performance of a given measure on a given dataset changes as we move from the Web to the Wikipedia. The results of our experiments are shown in Table 3. For completeness, we also show the results from [13], the only known document level co-occurrence result from the literature for these datasets.

We can see that for all measures and for five out of the six datasets, the performance always improves as we move from the Web to the Wikipedia. Even for the sixth dataset Goldfarb-Halpern, the Web does not perform better than the Wikipedia, except when the correlations are close to zero, i.e. when things are pretty random.

4. Further Improvement

Another advantage of using the Wikipedia is that unlike the Web, in case of the Wikipedia, a researcher can download the entire Wikipedia and can carry out measurements at a granularity finer than the document scale. In fact, traditionally co-occurrence based measures are defined in terms of the span constrained word-pair oc-

Table 3: Performance of various document based co-occurrence measures while using Wikipedia and the Web. For each measure and each dataset, the *best performing* version has been *highlighted*. Results for the Web-Google [13] are available only for the *wordsim353* dataset. Also, note that we have filtered out the multi-word expressions from each dataset. Hence, for example, we work with only 351 of the 353 pairs in the *w*ordsim353 dataset.

Measure	Corpus	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	words353 (351)	Esslli (272)
PMI	Wikipedia	**0.20**	**0.12**	**0.19**	**0.18**	**0.58**	**0.33**
	Web-Yahoo	0.18	0.04	0.11	0.18	0.35	0.17
	Web-Bing	0.08	0.02	0.07	0.18	0.20	0.10
SCI	Wikipedia	**0.36**	**0.24**	**0.28**	**0.17**	**0.48**	**0.50**
	Web-Yahoo	0.26	0.18	0.14	0.07	0.28	0.27
	Web-Bing	0.21	0.10	0.15	**0.17**	0.23	0.26
Ochiai	Wikipedia	**0.31**	**0.20**	**0.20**	-0.02	**0.41**	**0.31**
	Web-Yahoo	0.24	0.18	0.11	-0.03	0.18	0.14
	Web-Bing	0.22	0.12	0.12	**0.04**	0.29	0.11
Jaccard/ Dice	Wikipedia	**0.31**	**0.20**	**0.18**	-0.01	**0.36**	**0.21**
	Web-Yahoo	0.24	0.14	0.11	-0.01	0.14	0.09
	Web-Bing	0.22	0.12	0.09	**0.04**	0.25	0.05
	Web-Google [13]	-	-	-	-	0.18	-

currences[4]. By span we mean the inter word distance. When querying the Web, we get the counts of documents containing the word pair regardless of the distance between them in the documents. By span constrained occurrence we mean that a word-pair occurrence is counted only if the words occur close enough, that is if their span is less than a given threshold. That is, with every measure, a span-threshold parameter is attached.

4.1. Span Constrained Word Count Performance

We compare the performance of word based and document based version of each measure as given in Table 1. Our methodology of computing ranked correlation for a measure on a dataset remains the same (as described in Section 3.4). Only difference is that the word based version of each measure has span threshold as a parameter.

We follow the standard methodology of evaluating parametrized measures by cross validation. Each dataset is divided into five random partitions, four of which are used for training and one for testings. The span threshold is varied between 5 and 50 words for each measure and the span value that performs best on four training folds is used for the remaining one testing fold. The performance of a measure on a dataset is its average Spearman rank correlation over 5 runs with 5 different test folds.

[4]Note how Church and Hanks define joint probability in their seminal paper [2] that introduced PMI: *Joint probabilities, $P(x, y)$, are estimated by counting the number of times that x is followed by y in a window of w words, $f_w(x, y)$, and normalizing by N.*

4.2. Comparison

The comparison of document based and word based measures are shown in Table 4. From the results, we can see that with Wikipedia, further performance gain is obtained by moving from document counts to span-constrained word counts. We have four measures and six datasets for a total of twenty-four combinations. For eighteen out of the twenty-four combinations, such a performance gain is observed.

As an aside, it is interesting to note that in Tables 3 and 4, regardless of the corpus, performance of the Dice measure is virtually identical to that of the Ochiai measure on the Kent and Minnesota - two largest datasets. This is interesting because the Dice is the harmonic mean while the Ochiai is the geometric mean of the Conditional Probabilities $\frac{n(x,y)}{n(x)}$ and $\frac{n(x,y)}{n(y)}$.

5. Conclusions

By performing extensive experiments with various measures and multiple datasets of varying size, we demonstrate that despite being three orders of magnitude smaller in size, the Wikipedia is a reasonable alternative to the Web for measuring the document level co-occurrence based word association.

Another practical advantage of Wikipedia compared to web is that most researchers can exploit the span

Table 4: Performance comparison of the word based and document based version of each measure on Wikipedia. For each measure and each dataset, the *better performing* version has been *highlighted*. All standard deviations across 5 cross-validation runs for Kent and Minnesota are between 0.01 and 0.02, for White-Abrams were between 0.05 and 0.07, for Goldfarb-Halpern between 0.05 and 0.14, for Wordsim were between 0.02 and 0.11, and for Esslli were between .09 and .17. Note that the word-count based versions count span-constrained bigram occurrences while the document based versions do not take the span information into account.

Measure	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	words351 (351)	Esslli (272)
PMI-doc	0.20	0.12	0.19	**0.18**	0.58	**0.33**
PMI-word	**0.36**	**0.26**	**0.22**	0.11	**0.69**	0.32
SCI-doc	0.36	0.24	**0.28**	**0.17**	**0.48**	**0.50**
SCI-word	**0.38**	**0.27**	0.23	0.06	0.37	0.44
Ochiai-doc	0.31	0.20	0.20	-0.02	0.41	0.31
Ochiai-word	**0.43**	**0.31**	**0.29**	**0.08**	**0.62**	**0.44**
Jaccard/Dice-doc	0.31	0.20	0.18	-0.01	0.36	0.21
Jaccard/Dice-word	**0.43**	**0.32**	**0.21**	**0.09**	**0.59**	**0.35**

and the word count information with Wikipedia but not with the web. Our experiments demonstrate the utility of these information in improving the word association performance. In the current work, we have compared span-constrained word-count based versions with non-span-constrained document-count based versions, since for the Web we do not have access to the span information. In future, we plan to experiment with span-constrained document-count based versions for Wikipedia.

References

[1] P. Pecina and P. Schlesinger, "Combining association measures for collocation extraction," in *Association for Computational Linguistics*, 2006.

[2] K. W. Church and P. Hanks, "Word association norms, mutual information and lexicography," in *Association for Computational Linguistics*, pp. 76–83, 1989.

[3] A. Kilgarriff, "Googleology is bad science," *Computational Linguistics*, vol. 33, no. 1, pp. 147–151, 2007.

[4] J. R. Firth, "A synopsis of linguistics theory," *Studies in Linguistic Analysis*, pp. 1930–1955, 1957.

[5] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *NAAssociation for Computational Linguistics-HLT*, 2009.

[6] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *WWW*, pp. 757–766, 2007.

[7] H.-H. Chen, M.-S. Lin, and Y.-C. Wei, "Novel association measures using web search with double checking," in *Association for Computational Linguistics*, 2006.

[8] T. Wandmacher, E. Ovchinnikova, and T. Alexandrov, "Does latent semantic analysis reflect human associations?," in *European Summer School in Logic, Language and Information (ESSLLI'08)*, 2008.

[9] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Association for Computational Linguistics*, 2008.

[10] T. Hughes and D. Ramage, "Lexical semantic relatedness with random graph walks," in *Conference on Empirical Methods on Natural Language Processing*, 2007.

[11] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *International Joint Conference on Artificial Intelligence*, 2007.

[12] E. Yeh, D. Ramage, C. Manning, E. Agirre, and A. Soroa, "Wikiwalk: Random walks on wikipedia for semantic relatedness," in *Association for Computational Linguistics workshop "TextGraphs-4: Graph-based Methods for Natural Language Processing"*, 2009.

[13] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *Conference on Artificial Intelligence*, pp. 1419–1424, 2006.

[14] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: the concept revisited," *ACM Transactions on Information Systems*, vol. 20, no. 1, pp. 116–131, 2002.

[15] T. Chklovski and P. Pantel, "Verbocean: Mining the web for fine-grained semantic verb relations," in *Conference on Empirical Methods on Natural Language Processing*, pp. 33–40, 2004.

[16] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Conference on Artificial Intelligence*, 2006.

[17] W. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *20th International Conference on Computational Linguistics*, 2004.

[18] D. L. Chaudhari, O. P. Damani, and S. Laxman, "Lexical co-occurrence, statistical significance, and word association," in *Conference on Empirical Methods on Natural Language Processing*, 2011.

[19] L. R. Dice, "Measures of the amount of ecological association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[20] P. Jaccard, "The distribution of the flora of the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.

[21] A. Ochiai, "Zoogeografical studies on the soleoid fishes found in japan and its neighbouring regions-ii," *Bulletin of the Japanese Society of Scientific Fisheries*, vol. 22, 1957.

[22] J. Washtell and K. Markert, "A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations," in *Conference on Empirical Methods on Natural Language Processing*, pp. 628–637, 2009.

[23] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[24] ESSLLI, "*Free association* task at lexical semantics workshop esslli 2008." `http://wordspace.collocations.de/doku.php/workshop:esslli:task`, 2008.

[25] R. Goldfarb and H. Halpern, "Word association responses in normal adult subjects," *Journal of Psycholinguistic Research*, vol. 13, no. 1, pp. 37–55, 1984.

[26] G. Kent and A. Rosanoff, "A study of association in insanity," *American Journal of Insanity*, pp. 317–390, 1910.

[27] W. Russell and J. Jenkins, "The complete minnesota norms for responses to 100 words from the kent-rosanoff word association test," tech. rep., Office of Naval Research and University of Minnesota, 1954.

[28] K. K. White and L. Abrams, "Free associations and dominance ratings of homophones for young and older adults," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 408–420, 2004.

[29] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguists*, vol. 32, no. 1, pp. 13–47, 2006.

[30] StopWordList, "`http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`," *The Information Retrieval Group, University of Glasgow*, 2010. Accessed: November 15, 2010.