Using Web-Search Results to Measure Word-Group Similarity

Ann Gledson and John Keane
School of Computer Science,
University of Manchester
Oxford Road, Manchester, UK M13 9PL
{ann.gledson,john.keane}@manchester.ac.uk

Abstract

Semantic relatedness between words is important to many NLP tasks, and numerous measures exist which use a variety of resources. Thus far, such work is confined to measuring similarity between two words (or two texts), and only a handful utilize the web as a corpus. This paper introduces a distributional similarity measure which uses internet search counts and also extends to calculating the similarity within word-groups. The evaluation results are encouraging: for word-pairs, the correlations with human judgments are comparable with state-of-the-art web-search page-count heuristics. When used to measure similarities within sets of 10 words, the results correlate highly (up to 0.8) with those expected. Relatively little comparison has been made between the results of different search-engines. Here, we compare experimental results from Google, Windows Live Search and Yahoo and find noticeable differences.

1 Introduction

The propensity of words to appear together in texts, also known as their distributional similarity is an important part of Natural Language Processing (NLP):

> 'The need to determine semantic relatedness… between two lexically expressed concepts is a problem that pervades much of [NLP].' (Budanitsky and Hirst 2006)

Such requirements are evident in word sense disambiguation (WSD) (Patwardhan et-al 2003), spelling correction (Budanitsky and Hirst 2006) and Lexical Chaining (Morris and Hirst 1991).

As well as measuring the co-occurrence of word-pairs, it is also considered useful to extend these measures to calculate the likelihood of sets of words to appear together. For example, Caracciolo et-al (2004) evaluate two established topic area detection (TAD) algorithms and indicate text homogeneity as a document feature affecting the results. Furthermore, Gliozzo et-al (2004) and McCarthy et-al (2007) highlight the importance of topic features for WSD and Navigli and Velardi (2005) report differing WSD results for 3 types of text: unfocussed, mid-technical (eg finance articles) and overly technical (eg interoperability and computer networks). Measures of word-group similarity can be used to calculate the level of topic cohesion in texts, and the potential for this to be used to benefit NLP areas such as WSD and TAD has been indicated in Gledson and Keane (2008).

We consider web-searching an important part of measuring similarity, as it provides up-to-date information on word co-occurrence frequencies in the largest available collection of English language documents. We propose a simple measure that uses internet search counts by measuring the decline in the number of hits as more words (from the word-set to be measured) are appended to a query string using the 'AND' operator. To offset against the effect of individual word hit-counts, the above gradient is compared to that of the individual word hit counts – arranged in descending order of hits returned.

The paper is structured as follows: in Section 2 we describe related work in the areas of word similarity and the use of search-engine counts in NLP. Section 3 outlines the algorithm to be used for our similarity measure which can utilise any search-engine that is web-service enabled. Sec-

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 281–288
Manchester, August 2008

tion 4 describes and discusses the results of the evaluation techniques used, one for evaluating word-pair similarities, which compares with previous work and human judgements, and three for evaluating word-group similarities. Section 5 outlines the conclusions drawn from these experiments and Section 6 discusses further work.

2 Related Work

The most commonly used similarity measures are based on the WordNet lexical database (eg Budanitsky and Hirst 2006, Hughes and Ramage 2007) and a number of such measures have been made publicly available (Pedersen et-al 2004). The problem with such methods is that they are confined to measuring words in terms of the lexical connections manually assigned to them. Language is evolving continuously and the continual maintenance and updating of such databases is highly labour intensive, therefore, such lexical resources can never be fully up-to-date. In addition, the lexical connections made between words and concepts do not cover all possible relations between words. An important relationship between words is distributional similarity and Budanitsky and Hirst (2006) conclude that the capture of these 'non-classical' relationships is an important area of future research.

Weeds and Weir (2006) use a 'co-occurrence retrieval method to compute word relatedness, which is described as analogous to the precision and recall metrics of document retrieval. They observe the 'plausibility' of substituting word-1 for word-2 within verb/object-noun grammatical relationships. This method is restricted to document retrieval from the BNC corpus, due to the pre-requisite that words are part-of-speech tagged and that some grammatical parsing is performed. The similarity measure that we have developed is simpler and does not rely on preprocessing of texts. This distributional similarity measure calculates the propensity of words to appear together, regardless of part-of-speech or grammatical functions.

We assert that the world-wide-web can be used to capture distributional similarity, and is less likely to suffer from problems of coverage, found in smaller corpora. The web as a corpus has been successfully used for many areas in NLP (Kilgarriff and Grefenstette 2003) such as WSD (Mihalcea and Moldovan 1999), obtaining frequencies for bigrams (Keller and Lapata 2003) and noun compound bracketing (Nakov and Hearst 2005). Such reliance on web search-engine results does come with caveats, the most important (in this context) being that the reported hit counts may not be entirely trustworthy (Kilgarriff 2007).

Strube and Ponzetto's (2006) use the Wikipedia database, which includes a taxonomy of categories, and they adapt 'well established semantic relatedness measures originally developed for WordNet'. They achieve a correlation coefficient of 0.48 with human judgments, which is stated as being higher than a Google-only and WordNet-only based measure for the largest of their test datasets (the 353 word-pairs of 353-TC)

Chen et-al (2006) use the snippet's returned from web-searches, in order to perform word similarity measurement that captures 'new usages' of ever evolving, 'live' languages. A double checking model is utilized which combines the number of occurrences of the first word in the snippets of the second word, and vice-versa. This work achieves a correlation coefficient of 0.85 with the Miller and Charles (1998) dataset of 28 word-pairs, but to achieve the best results, 600-700 snippets are required for each word-pair, requiring extra text-processing and searching.

The work most similar to ours is the set of page-counting techniques of Bollegala et-al (2007). They combine the use of web-search page-counts with the analysis of the returned text snippets and achieve impressive correlations with human judgment (0.834 with the Miller and Charles dataset). The text-snippet analysis is more complex than that of Chen et-al (2006), as the optimal number of snippets is over 1000, and their process involves complex pattern matching, which may not scale well to measuring word-groups similarity. Their page counting heuristics utilize 4 popular co-occurrence methods: Jaccard, Overlap (Simpson), Dice and PMI (Pointwise mutual information) but again, these techniques are not designed to scale up to larger numbers of input words.

Mihalcea et-al (2006) and Gabrilovich and Markovitch (2007) achieve good results when comparing texts, the former utilizing the inverse-document frequency heuristic and the latter indexing the entire Wikipedia database and comparing vector representations of the two inputs.

None of the above work is adapted to be used on single groups of words as a measure of topic cohesion. They could be adapted by combining similarity results between many pairs, but this might similarly have a high computational cost. In addition, no comparisons of different web-search engines are made when using web-counts.

As the use of web-counts is precarious (Kigarriff 2007), these types of comparisons are of high practical value.

3 Similarity Measure

The proposed measure can be used with any web-service enabled search-engine and in our experiments we compare three such search-engines: Yahoo *[Yahoo]*, Windows Live Search *[Live-Search]* and Google *[Google]*. The similarity measure $WebSim_{[search\text{-}engine]}$ is calculated for each document d as follows:

Step 1: Add the required set of *n* lemmas to the *Lemmas* list.

Step 2: Using an internet *search-engine*, obtain the hit counts of each member of *Lemmas*.

Step 3: Order the resulting list of n lemma/hit-counts combinations in descending order of hit-counts and save lemma/hit combinations to *IndivHitsDesc*.

Step 4: For each lemma of *IndivHitsDesc*, save to *CombiHitsDesc* preserving the ordering.

Step 5: For each member of *CombiHitsDesc*: *CombiHitsDesci*, obtain the hit counts of the associated lemma, along with the concatenated lemmas of all preceding list members of *CombiHitsDesc* (*CombiHitsDesc*[0] to *CombiHitsDesc*[i-1]). This list of lemmas are concatenated together using ' AND ' as the delimiter.

Step 6: Calculate the gradients of the best-fit lines for the hit-counts of *IndivHitsDesc* and *CombiHitsDesc*: creating *gradIndiv* and *gradCombi* respectively.

Step 7: $WebSim_{[search\text{-}engine]}$ is calculated for d as *gradIndiv* minus *gradCombi*.

As $WebSim_{[search\text{-}engine]}$ is taken as the difference between the two descending gradients, the measure is more likely to reveal the affects of the probability of the set of lemmas co-occurring in the same documents, rather than by influences such as duplicate documents. If the decline in hit-counts from *IndivHitsDesc*[i-1] to *IndivHitsDesc*[i] is high, then the decline in the number of hits from *CombiHitsDesc*[i-1] to *CombiHitsDesc*[i] is also expected to be higher, and the converse, for lower drops is also expected. Deviations from these expectations are reflected in the final homogeneity measure and are assumed to be caused by the likelihood of lemmas co-occurring together in internet texts.

Search-engines are required that publish a set of web-services for a fully automated process. The Google, Yahoo and Windows Live Search search-engines have been selected and the results of each measure are compared. In response to important problems highlighted by Kilgarriff (2007) relating to the use of web counts in NLP experiments: firstly, a measure is proposed between words that does not require the pre-annotation of part-of-speech information and does not rely on query syntax / meta-language. Secondly, our measure relies on the use of web-search *page* counts (as opposed to word instance counts) as we are measuring the likelihood of co-occurrence in the same text. Finally, measures are taken to try to avoid the problem of arbitrary search-engine counts. For example, each measure is the result of a comparison between the decline rates of 2 sets of hit counts and the full set of queries for each input text are taken within a 20 second interval (for groups of 10 words). In addition, for the Google and Yahoo measures the web-service request parameter includes options to avoid the return of the same web-page (Google: 'filter_alike_results'; Yahoo: 'AllowSimilar').

4 Evaluation

Four methods of evaluation are used to verify that the similarity measure is capable of measuring similarity between words. These methods are selected to be as varied as possible, to provide a fuller understanding of the usefulness of our technique and to verify that it is working as expected.

4.1 Word-pairs

The results of Rubenstein and Goodenough (1965) (65 word-pairs), Miller and Charles (1998) (30 word-pair subset of Rubenstein and Goodenough 1965) and Finkelstein et-al (2002) (353 word-pairs) are used to evaluate the ability of the proposed method to calculate the similarity between word-pairs. These results sets list the similarity scores of the word-pairs as assigned by humans. Although this does provide a useful benchmark, extremely high, or perfect correlations are unrealistic, as those involved in the experiments were asked to think about the words association in terms of lexical relations such as synonymy as opposed to the broader idea of distributional similarity. Nevertheless, in conjunction with other evaluation techniques, these comparisons can still be useful, as some correlation would be expected if our measure was functioning as required.

In addition, our results are compared with the page-count based similarity scores of Bollegala

Measure	Correlation (Pearson's R)
$WebSim_{[YAHOO]}$	-0.57**
$WebSim_{[LIVE\text{-}SEARCH]}$	**-0.60****
$WebSim_{[GOOGLE]}$	-0.43**
Google-Jaccard (Strube & Ponzetto 2006)	*0.41*
pl (path-lengths) (Strube & Ponzetto 2006)	*0.56*
wup (Strube & Ponzetto 2006)	*0.52*
lch (Strube & Ponzetto 2006)	*0.54*

** Significant at the 0.01 level (2-tailed)

Italics: Statistical significance not specified in Strube and Ponzetto (2006)

Table 1: Correlation with Human Ratings on Rubenstein-Goodenough dataset

Measure	Correlation (Pearson's R)
$WebSim_{[YAHOO]}$	**-0.55****
$WebSim_{[LIVE\text{-}SEARCH]}$	-0.53**
$WebSim_{[GOOGLE]}$	-0.39*
Jaccard (Bollegala et-al 2007)	*0.26*
Dice (Bollegala et-al 2007)	*0.27*
Overlap (Bollegala et-al 2007)	*0.38*
PMI (Bollegala et-al 2007)	***0.55***
pl (Strube & Ponzetto 2006)	*0.49*

** Significant at the 0.01 level (2-tailed)

* Significant at the 0.05 level (2-tailed)

Italics: Statistical significance not specified in Bollegala et-al (2007) or Strube and Ponzetto (2006)

Table 2: Correlation with Human Ratings on Miller-Charles' dataset

Measure	Correlation (Pearson's R)
$WebSim_{[YAHOO]}$	-0.37**
$WebSim_{[LIVE\text{-}SEARCH]}$	-0.40**
$WebSim_{[GOOGLE]}$	-0.119*
Google-Jaccard (Strube & Ponzetto 2006)	*0.18*
wup (Strube & Ponzetto 2006)	***0.48***
lch (Strube & Ponzetto 2006)	***0.48***

** Significant at the 0.01 level (2-tailed)

* Significant at the 0.05 level (2-tailed)

Italics: Statistical significance not specified in Strube and Ponzetto (2006)

Table 3: Correlation with Human Ratings on 353-TC dataset

et-al (2007) and the best performing Wikipedia-based measures of Strube and Ponzetto (2006). The former gauge similarity using a number of popular co-occurrence measures adapted for web-search page-counts and their method is considered closest to our approach. The individual results for each word-pair are shown in Tables 1, 2 and 3. They indicate that moderate correlations exists, particularly in the Rubenstein and Goodenough set. $WebSim_{[YAHOO]}$ and $WebSim_{[LIVE\ SEARCH]}$ significantly outperform the $WebSim_{[GOOGLE]}$ method. This may be due to Google's results being more erratic. Google's returned counts were sometimes found to increase as extra 'AND' clauses were added to the query string. This is perhaps because of the way in which Google combines the results of several search-engine hubs. (This was accommodated to a degree by setting any negative scores to zero.) All three of the methods were comparable with the results of Bollegala et-al (2007), with the $WebSim_{[LIVE\ SEARCH]}$ measure performing at the highest levels and $WebSim_{[YAHOO]}$ producing the highest of all measures on the Miller and Charles (1998) dataset. (Unfortunately, Bollegala et-al (2007) do not compare their results with the Rubenstein-Goodenough and 353-TC dataset.) In the largest (353-TC) dataset, the $WebSim_{[YAHOO]}$ and $WebSim_{[LIVE\ SEARCH]}$ results were found to be, lower, but comparable with the best of Strube and Ponzetto's (2006) Wikipedia-based results and significantly higher than their Jaccard measure, adapted to use Google web-search page-counts.

Set	Words
A	law, crime, perpetrator, prison, sentence, judge, jury, police, justice, criminal
B	astrology, “star sign", Libra, Gemini, Sagittarius, zodiac, constellation, Leo, Scorpio, birthday
C	ocean, "hydrothermal vent", fauna, botanical, biogeography, tube-worm, Atlantic, species, biology, habitat
D	economy, credit, bank, inflation, "interest rate", finance, capital, expenditure, market, profit
E	health, diet, swimming, fitness, exercise, "heart disease", stroke, fruit, jogging, work
F	football, stadium, striker, trophy, match, referee, pitch, crowd, manager, kick
G	education, exam, teacher, timetable, classroom, pupils, homework, results, parents, teenager
H	country, election, president, vote, leader, population, opposition, party, government, count
J	computer, "hard drive", keyboard, connection, monitor, RAM, speaker, "mother board", CPU, internet
K	"film star", drugs, money, millionaire, fame, music, actor, debut, beauty, paparazzi
M	"house prices", monthly, mortgage, home, borrowing, "buy-to-let", figures, homeowners, lending, trend
N	inmates, overcrowding, prisoners, custody, release, warden, cell, bars, violence, detention

Table 4: Manually selected test sets

4.2 Word-groups

In order to indicate whether the proposed measure is capable of measuring similarity amongst a set of words, a broad range of evaluation methods is required. No human evaluations exist, and to produce such a data-set would be difficult due to the larger quantity of data to evaluate. The following three methods are used:

Manual Selection of Word-groups

The manual selection of word-groups, for testing the proposed measure is an important part of its evaluation as it is possible to construct word-sets of varying similarity, so that expected results can be predicted and then compared with actual results. In addition, the datasets created can be easily manipulated to range from highly homogenous / similar sets of words to extremely heterogeneous – where words are highly unlikely to appear together. The latter is achieved by systematically merging the groups together, until a complete mixture of words from different word-sets is produced.

The method used to compile the words groups is as follows: firstly, groups of words with a known propensity to appear together were selected by browsing popular internet web-sites (see Table 4 for the words contained in each set). Secondly for each of these original sets, a series of 5 measures is taken, the first with all words from the original set (to illustrate, this might be represented as AAAAAAAAAA – where each letter represent a word from the set shown), this therefore is the most homogeneous group. Then two words from this set are replaced with two from a new set (eg B) (AAAAAAAABB). Then a further two words (again originally from set A) are replaced with two more words from another new set (eg from Set C) (AAAAAABBCC), and so on, until the final set of 10 words to be measured consists of 5 pairs of words, each from 5 different sets (AABBCCDDEE). These steps were first performed for sets A, B, C, D and E and then for sets F, G, H, J and K respectively. The results were compared (using Pearson’s correlation) against the expected results. For example: AAAAAAAAAA = 10 points, AAAAAAAABB = 8 points, AAAAAABBCC = 6 points, AAAABBCCDD = 4 points and AABBCCDDEE = 2 points.

On the whole, the word-groups contained in each of these two sets of sets are considered to be heterogeneous (eg A is dissimilar to B, C, D and E etc). To introduce more ‘blurring’ of these categories, a third set of sets was measured, consisting of the word-sets A, D, H, M and N. It was considered more conceivable that the words in these sets could be found in the same documents. The expected results for this set of measures were modified slightly to become: AAAAAAAAAA = 8 points, AAAAAAAABB = 7 points, AAAAAABBCC = 6 points, AAAABBCCDD = 5 points and AABBCCDDEE = 4 points. This was done to reflect the fact that the differences between the highest and lowest were expected to be less.

	Correlation (Pearson’s R)	
Measure	Heterogeneous Sets only: ABCDE FGHJK	Homogenous Sets Included: ABCDE FGHJK +ADHMN
$WebSim_{[YAHOO]}$	**-.80****	**-.68****
$WebSim_{[LIVE\text{-}SEARCH]}$	-.65**	-.57**
$WebSim_{[GOOGLE]}$	-.70**	-.64**

**Significant at the .01 level (2-tailed)

Table 5: Correlation with expected scores for *manually* selected sets

As illustrated in Table 5, the 'Heterogeneous Sets' group similarity scores were found to correlate very highly with those expected, with the $WebSim_{[YAHOO]}$ measure achieving .80. The $WebSim_{[GOOGLE]}$ measure was also found to improve when tested on groups of words.

The measures were found to perform less well when the third set of word-sets, containing more closely related members, was introduced (Table 5, final column). This indicates that the measures are better at distinguishing between highly homogeneous and highly heterogeneous word-sets, but appear less proficient at distinguishing between word-sets with moderate levels of topic homogeneity.

WordNet::Domains Selection of Word-groups

The WordNet Domains package[2] (Magnini and Cavaglia 2000) assigns domains to each sense of the WordNet electronic dictionary. Therefore, for each domain a relevant list of words can be extracted. The domains are arranged hierarchically, allowing sets of words with a varied degree of topic homogeneity to be selected. For example, for a highly heterogeneous set, 10 words can be selected from any domain, including factotum (level-0: the non-topic related category). For a slightly less heterogeneous set, words might be selected randomly from a level-1 category (eg 'Applied_Science'), and any of the categories it subsumes (eg Agriculture, Architecture, Buildings etc). The levels range from level-0 (factotum) to level-4; we merge levels 3 and 4 as level-4 domains are relatively few and are viewed as similar to level-3. This combined set is henceforth known as level-3.

Measure	Correlation (Pearson's R)	
	All	Extreme Only
$WebSim_{[YAHOO]}$	**-0.46****	**-0.80****
$WebSim_{[LIVE-SEARCH]}$	-0.06	-0.23**
$WebSim_{[GOOGLE]}$	-0.42**	-0.71**

**Significant at the .01 level (2-tailed)

Table 6: Correlation with expected scores for *WordNet::Domains* selected sets

For our experiments, we collected 2 random samples of 10 words for every WordNet domain (167 domains) and then increased the number of sets from level-0 to level-2 domains, to make the number of sets from each level more similar. The final level counts are: levels 0 to 2 have 100 word-sets each and level 3 has 192 word-sets. The word-sets contain 10 words each. We then assign an expected score to each set, equal to its domain level.

Table 6 displays the resulting correlation between the *WebSim* scores and the expected scores (column: 'All'). In the previous section, we observed that the measures are less competent at distinguishing between moderate levels of homogeneity, and the WordNet::Domain test sets contain many sets which could be described as having moderate homogeneity. To further test this, we repeated the above WordNet::Domain tests, but included only those sets of level-0 and level-3. The results displayed in the final column of Table 6 and provide further evidence that this might be the case, as the correlations are significantly higher for these more extreme test sets.

SENSEVAL 2 & 3 Data

The selection of the most frequent words from natural language documents is another important part of our evaluation, as it is representative of real-world data-sets. As it is anticipated that our results could be of use for WSD, we opted to measure the topic homogeneity of a publicly available set of documents from a well established WSD competition. The documents of the SENSEVAL 2 & 3[3] English all-words WSD task were divided into 73 sub-documents, each containing approximately 50 content-words. The stop-words and non-topical[4] content words of each sub-document were then removed and the remaining words lemmatised and aggregated by lemma. This list of lemmas was then arranged in descending order of the number of occurrences in the sub-document. The top-10 most frequent lemmas were selected as input to the similarity measure. The results for each set along with descriptions of the documents used are displayed in Table 7.

The scientific documents (d01) could be viewed as the most homogeneous, with the earthquake accounts (d002) and the work of fiction (d000) being considered the most heterogeneous. With this in mind, it is evident in Table 7 that the $WebSim_{[MSN]}$ measure performed the least well and other two methods performed as expected, and with identical rankings.

Further analysis is required to compare these results with the WSD results for the same

[2] We use version 3.2 released Feb 2007

[3] See http://www.senseval.org/

[4] Non-topical words are those that are found in over 25% of Semcor documents or have their correct sense(s) belonging to the 'Factotum' category in the WordNet Domains package by Magnini and Cavaglia (2000).

Text	Description	Average *WebSim* [YAHOO]	Average *WebSim* [GOOGLE]	Average *WebSim* [MSN]
d00	**Change-Ringing**: – History of Campanology, Churches, Social History Typical set = {bell, church, tower, "change ringing", English, peculiarity, rest, stone, faithful, evensong}	1.52 (4)	1.10 (4)	1.11 (5)
d01	**Cancer Cell Research**: Typical Set = {cancer, gene, parent, protein, cell, growth, "suppressor gene", individual, scientist, beneficiary}	*1.19 (1)*	*.93 (1)*	.89 (2)
d02	**Education**: Typical Set = {child, belief, nature, parent, education, medium, politician, "elementary school", creativity}	1.36 (3)	1.05 (3)	1.03 (3)
d000	**Fiction**: Typical set = {stranger, drinking, occasion, street, "moving van", intersection, brake, guy, mouth, truck}	1.55 (5)	1.18 (5)	1.03 (3)
d001	**US Presidential Elections**: Typical set = {district, candidate, half-century, election, percentage, president, vote, hopeful, reminder, ticket}	1.29 (2)	.96 (2)	*.85 (1)*
d002	**California Earthquake** – First hand accounts / stories Typical Set = {computer, quake, subscriber, earthquake, resident, solace, "personal computer", hundred, "check in", "bulletin board"}	**1.63 (6)**	**1.32 (6)**	**1.13 (6)**

Rankings for each measure are shown in brackets

Table 7: Average group similarity measures of the SENSEVAL 2 and 3 datasets

documents, as performed by Gledson and Keane (2008), and this is highlighted as part of the further work.

5 Conclusions

We proposed a simple web-based similarity measure which relies on page-counts only, can be utilized to measure the similarity of entire sets of words in addition to word-pairs and can use any web-service enabled search engine. When used to measure similarities between two words, our technique is comparable with other state-of-the art web-search page-counting techniques (and outperforms most). The measure is found to correlate to a moderate level (the highest being .60 correlation) with human judgments. When used to measure similarities between sets of 10 words, the results are similarly encouraging and show the expected variations for word-groups with different levels of homogeneity. Where word-groups are manually created, with known expectations of the similarities between each word-set, correlations with these expected results are as high as .80. Noticeable differences between the performances of each of the 3 search-engines, for each evaluation method, are evident. Google performs poorly for the word-pair similarity measures and Yahoo and Live Search both perform substantially better. For the word-set comparisons, Google performance improves (perhaps as the erratic single counts are stabilised as larger sets of words are used), but Yahoo is again superior and MSN performs much less well. Overall, our results indicate that the Yahoo measure is the most consistent and reliable.

6 Further Work

The group similarity measure calculates the propensity of words to co-occur with one-another, which can be described as a measure of the topic homogeneity of the set of input words. Gledson and Keane (2008) propose the use of further measures of topic homogeneity using a variety of available resources. These measures are compared with the results of WSD approaches reliant on topic features and low to moderate correlations are found to exist. It is also proposed that useful correlations with other NLP techniques utilising topical features (such as TAD and Malapropism Detection) might also exist.

The word-group similarity measures performed better for extreme levels of topic homogeneity. The measures must be improved to enable them to distinguish between moderate homogeneity levels. This may be achieved by combining our simple measure with other word similarity/relatedness techniques such as the use of WordNet Domains, or Lexical Chaining.

It is expected that polysemy counts of words influence the outcome of these experiments, especially for the word-pairs which have less data and are more susceptible to erratic counts. Results might be improved by measuring and offsetting these effects.

In addition, an upper limit of word-set cardinality should be determined, which is the maximum number of input words that can be measured. Further testing is necessary using a range of set cardinalities, to obtain optimal values.

References

Bollegala, D., Matsuo, Y., Ishizuka, M., 2007. Measuring Semantic Similarity between Words Using Web Search Engines. In Proceedings of World-Wide-Web Conference 2007 (Track: Semantic Web), Banff, Alberta, Canada. pp. 757-766.

Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* , 32(1), pp. 13--47.

Caracciolo, C., Willem van Hage and Maarten de Rijke, 2004. Towards Topic Driven Access to Full Text Documents, in *Research and Advanced Technology for Digital Libraries, LNCS*, 3232, pp 495-500

Chen, Hsin-Hsi, Ming-Shun Lin and Yu-Chuan Wei, 2006. Novel Association Measures Using Web Search with Double Checking. In *Proc.s 21st Intl. Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 1009-1016

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, 2002. Placing Search in Context: The Concept Revisited**,** *ACM Transactions on Information Systems*, 20(1), pp. 116-131, January

Gabrilovich, Evgeniy and Shaul Markovitch, 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *In Proc.s Intl. Joint Conference on Artificial Intelligence 2007*, Hyderabad, India.

Gledson, Ann and John Keane. 2008. Measuring topic homogeneity and its application to word sense disambiguation. *In Procs 22nd Intl Conference on Computational Linguistics (COLING)*, Manchester. (To Appear)

Gliozzo, Alfio, Carlo Strapparava and Ido Dagan, 2004. Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation, *Computer Speech and Language*

Hughes, T. and D. Ramage, 2007. Lexical Semantic Relatedness with Random Graph Walks, *In Proc.s of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 581-589, Prague.

Keller, F. and Lapata, M., 2003. Using the Web to Obtain Frequencies for Unseen Bigrams, *Computational Linguistics*, 29(3)

Kilgarriff, Adam, 2007. Googleology is Bad Science. Computational Linguistics, 33(1), pp. 147—151.

Kilgarriff, A and Grefenstette, G., 2003. Web as Corpus, In Introduction to the special issue on the web as corpus, *Computational Linguistics*, 29(3), pp. 333--347

Magnini, B. and Cavaglià, Gabriela. 2000. Integrating Subject Field Codes into WordNet. *In Proceedings of LREC-2000,* Athens, Greece, 2000, pp 1413-1418.

McCarthy, Diana, Rob Koeling, Julie Weeds and John Carroll, 2007. Unsupervised Acquisition of Predominant Word Senses, *Computational Linguistics*, 33(4), pp. 553-590.

Mihalcea, R and Moldovan, D., 1999. A method for word sense disambiguation of unrestricted txt. *In Proceedings of the 37th Meeting of ACL*, pp 152-158

Mihalcea, Rada, Courtney Corley and Carlo Strapparava, 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity, *In Proc.s of American Association for Artificial Intelligence 2006*

Miller, G. and Charles, W., 1998. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), pp. 1--28.

Morris, J. and Hirst, G., 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, 17(1), pp. 21-48

Nakov, P. and Hearst, M. 2005. Search Engine Statistics Beyond the n-gram: Application to Noun Compound Bracketing, *In Proceedings of the 9th Conference on CoNLL*, pp. 17--24, Ann Arbor, June

Navigli, Roberto, Paola Velardi, 2005. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7), pp. 1075-1086, July.

Patwardhan, S., Banerjee, S. and Pedersen, T., (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation, *LNCS 2588 - CICLing 2003*, pp. 241--257.

Pedersen, Ted, Siddharth Patwardhan and Jason Michelizzi, 2004. WordNet::Similarity – Measuring the Relatedness of Concepts, *In Proc.s 19th National Conference on Artificial Intelligence*.

Rubenstein, H. and Goodenough, J., 1965. Contextual correlates of synonymy. Communications of the ACM, 8, pp. 627--633.

Strube, Michael and Paolo Ponzetto, 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia, *In Proc.s of American Association for Artificial Intelligence 2006*

Weeds, Julie and, David Weir, 2006. Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, 31(4), pp 433-475.