Measuring Topic Homogeneity and its Application to Dictionary-based Word Sense Disambiguation

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

The use of topical features is abundant in Natural Language Processing (NLP), a major example being in dictionary-based Word Sense Disambiguation (WSD). Yet previous research does not attempt to measure the level of topic cohesion in documents, despite assertions of its effects. This paper introduces a quantitative measure of *Topic Homogeneity* using a range of NLP resources and not requiring prior knowledge of correct senses. Evaluation is performed firstly by using the WordNet::Domains package to create word-sets with varying levels of homogeneity and comparing our results with those expected. Additionally, to evaluate each measure's potential value, the homogeneity results are correlated against those of 3 co-occurrence/dictionary-based WSD techniques, tested on 1040 Semcor and SENSEVAL sub-documents. Many low-moderate correlations are found to exist with several in the moderate range (above .40). These correlations surpass polysemy and sense-entropy, the 2 most cited factors affecting WSD. Finally, a combined homogeneity measure achieves correlations of up to .52.

1 Introduction

Topical features in NLP consist of unordered bags of words, often the context of a target word or phrase. In WSD for example, the word *bank* in the sentence: 'If you're OK being tied to one *bank*, you can get all your *financial* products there.' might be assigned its monetary sense, based on the occurrence of the term *financial*. Often referred to as 'topical features', these are an important part of many NLP methods, such as WSD (Yarowsky 1995) and Topic Area Detection (TAD) (Hearst 1997). Furthermore, in the SENSEVAL WSD competitions they are included in the highest performing systems.

We assert that the effectiveness of topical features in NLP depends upon the level of topic homogeneity in the text. To illustrate two extremes: the disambiguation of the word *bank* might be more difficult if occurring in i) a work of fiction describing a series of activities which includes the phrases: 'stroll along the river' and 'pick up her cheque book'; than in ii) a news report on a bank in financial difficulty (a topically homogeneous text).

This paper contributes a set of unsupervised Topic Homogeneity measures requiring no knowledge of correct senses. A variety of NLP resources are utilized and a set of evaluation methods devised, providing useful results. The paper is structured as follows: Section 2 outlines related work; Section 3 describes the experiments focusing on the resources used and their associated homogeneity measures; in Section 4 three evaluation methods are described, including a WSD task-based evaluation; Conclusions and future work are presented in Section 5.

2 Related Work

TAD research (Hearst 1997) has revealed that word patterns within texts can be used to locate topic areas. Salton and Allan (1994) distinguish between homogenous texts, where the topic of the text might be ascertained from a small number of paragraphs, and heterogeneous texts, where topic areas change considerably. Unfortunately, this research only detects topic homogeneity at inter-paragraph level. We assert that

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 273–280
Manchester, August 2008

the strength of relationships between the *words* of a text can vary from one text to another and that this is likely to affect the usefulness of topic features in NLP tasks. Caracciolo et-al (2004) also indicate that documents can vary in topical structure, mentioning text homogeneity as an important feature but they do not evaluate these findings explicitly.

Lexical cohesion (Halliday and Hasan 1976) is the analysis of the way the phrases and sentences of a text adhere to form a unified, comprehensible whole. Morris and Hirst (1991) describe a set of relationships between word-pairs, which include their likelihood to co-occur. These are used to create Lexical chains, which are sequences of related words in a document. The 2 main reasons for creating these chains are that they are 'an aid in the resolution of ambiguity' and they determine 'coherence and discourse structure'. We propose the use of lexical chains, alongside other methods, to measure document homogeneity.

Measuring the semantic relatedness between words is an important area in NLP, and has been used in areas such as WSD, Lexical Chaining and Malapropism Detection. Budanitsky & Hirst (2006) evaluate 5 such measures, based on lexically assigned similarities, and conclude that an area of future research should be the capture of 'non-classical' relationships not found in dictionaries, such as distributional similarity. Weeds and Weir (2006) evaluate distributional similarities using document retrieval from the BNC. Chen et-al (2006) and Bollegala et-al (2007) use web search-engine results. All of the reported similarity measures are between two words (or texts) only. We extend these to calculate the topic homogeneity of groups containing up to 10 words.

A small number of unsupervised WSD methods exist which take topic areas and/or document types into account (eg Gliozzo et-al 2004, McCarthy et-al 2007), but these work at corpus level and do not measure the level of topical homogeneity in each text. An exception is the WSD work by Navigli and Velardi (2005) who report differing results for 3 types of text: *unfocussed*, *mid-technical* (eg finance articles) and *overly technical* (eg computer networks).

3 Experiments

We propose the creation of a quantifiable measure of text homogeneity in order to predict the effectiveness of using co-occurrence features in WSD. In this initial set of experiments, documents are divided into simple ~50 content-word blocks, regardless of topic boundaries, to maximize the range of homogeneity levels in the texts and to nullify the effects of text length. The long-term objective is for documents to be divided into topic areas at the pre-processing stage, using a TAD algorithm. We also propose to take a single homogeneity measure of each entire sub-text, as opposed to using a sliding window approach, as the latter would be over-complex and comparable with a similarity-based WSD process.

3.1 Input Texts

The documents of Semcor[2] and SENSEVAL (2 & 3) are used in the experiments. Each text is divided into sub-documents of approximately 50 content-words[3]. All documents are converted into Semcor 2.1 format.

From the resulting sub-documents, 1040 random Semcor texts and the entire set of 73 SENSEVAL 2 and 3 texts are used in the experiments.

3.2 Text Pre-Processing

Non-topic words are found to have a negative effect on NLP methods using topic features, such as WSD (eg Yarowsky 1993, Leacock et-al 1998), and are therefore excluded from our topic homogeneity experiments. Unfortunately, no precise definition of non-topic words exists. Under ideal conditions, where correct senses are known, we define non-topic words as word-senses appearing in over 25% of all Semcor texts and/or being marked as *factotum* in the WordNet Domains 3.2 package (Magnini and Cavaglia 2000).

As the experiments described in this work assume no such prior knowledge, an approximation of the criteria used above is made. A J48 (pruned) decision tree is used to decide whether each word is non-topical. The input attributes were the PoS, sense-count, Semcor distribution (*SenseEntropy*) of its possible senses, whether all of the word's senses are factotum, and the percent of Semcor documents containing that word. The training and test data was the entire set of Semcor and SENSEVAL 2 and 3 English all-word task data and the minimum node size was set to 4000 instances to minimize the tree size and prevent over-training. Using a 10-fold cross-validation test mode the tree obtains 83% accura-

[2] Using the Brown-1 and Brown-2 documents only

[3] Splits are made as near as possible to the 50 content-word length whilst keeping sentences intact.

cy. The learned filter for selecting non-topical nouns and verbs is:

(All-Senses = Factotum) || (Corpus-Hit-Percent >25.0%) ||
((Sense_Count > 1) && (PoS = v)) ||
((Sense_Count > 3) && (*SenseEntropy* > 0.5668))

Upon entry into the system, each sub-document has all such words labeled as non-topical. The remaining words are labeled as topic-content. The confusion matrix output is shown in Table 1.

Classified As →	**OTHER**	**NON-TOPIC**
OTHER	15017	5768
NON-TOPIC	4278	34292

Table 1: J48 Confusion Matrix

In addition, only nouns and verbs are considered in the experiments as these word types are considered most likely to contain topical information.

3.3 Homogeneity Measures

Five homogeneity measures have been created that cover a broad range of techniques for embodying topic-area information in natural language texts. This is to facilitate comparisons between different techniques and if such a variety of aspects is captured, it improves the likelihood of a successful combination of the methods to produce an optimised measure. Each takes a full pre-processed sub-document as input and outputs a single score.

Word Entropy

It is possible to capture topic homogeneity by using simple measures that require minimal reliance on external resources. Word entropy is considered as having the potential to reflect topical cohesion.

To measure *WordEntropy*, the frequency of each topic-content lemma of an input document *d* is obtained, and *Entropy(d)* is measured using this set of frequencies, as follows:

$$- \sum_{i=1..n} p(x_i) \log_2 p(x_i) \quad [1]$$

Where *n* is the number of different topic content lemmas in *d*, and $p(x_i)$ is calculated as

$$\text{frequency}(\text{lemma}_i)/\sum_{j=1..n}\text{frequency}(\text{lemma}_j) \quad [2]$$

As *Entropy(d)* is affected by *n* and *n* varies from one document to another, *Entropy(d)* is normalised by dividing it by the maximum possible Entropy calculation for *d*, that is if all lemmas had equal frequencies.

WordNet Similarities

WordNet::Similarities 1.04 (Pedersen et-al 2004) is publicly available software which uses aspects of WordNet to 'measure the semantic similarity and relatedness between a pair of concepts'. The package can measure similarities between lemma pairs, where no knowledge of the correct sense or PoS is required. These similarities can be easily adapted to assist with the measurement of document homogeneity, by comparing similarities between sets of word-PoS pairs in the document.

Three WordNet Similarities homogeneity measures ($AvgSim_{Measure}$) are calculated for each document as follows: **Step 1**: Order the topic-content lemmas of the input text firstly in descending order of frequency, and then by first appearance in the text. **Step 2**: Take the first *n* lemmas from this list (where *n* is all lemmas up to a maximum of 10) and add to *FreqLemmas*. **Step 3**: Calculate the mean of all of the similarity measures between each pair of lemmas in *FreqLemmas*. *AvgSim* can be defined as:

$$\text{Mean}(\sum_{i=1..n} \sum_{j=i+1..n} Sim_{Measure}(\text{lemma}_i, \text{lemma}_j)) \quad [3]$$

Where $Sim_{Measure}(\text{lemma}_i, \text{lemma}_j)$ is the WordNet::Similarity calculation between lemma_i and lemma_j, where all allowable PoS combinations for the two lemmas when using the selected similarity measure are included.

The WordNet Similarity measures selected for use are Lesk, JCN[4], and Lch (see Pedersen et-al (2004) and Patwardhan et-al (2003) for details), as each measure represents one of the 3 main algorithm types available: WordNet Gloss overlaps, information content of the least common subsumer and path lengths respectively.

Yahoo Internet Searches

The web as a corpus has been successfully used for many areas in NLP such as WSD (Mihalcea and Moldovan 1999), obtaining frequencies for bigrams (Keller and Lapata 2003) and measuring word similarity (Bollegala et-al 2007). Such reliance on Web search-engine results does come with caveats, the most important in this context being that reported hit counts are not always reliable, mostly due to the counting of duplicate documents. (Kilgarriff 2007).

Using web-searches as part of the homogeneity measure is considered important to our experiments, as it provides up-to-date information on word co-occurrence frequencies in the largest available collection of English language docu-

[4] TheBNC information-content file is loaded.

ments. In addition, it is a measure that does not rely on WordNet. It is therefore necessary to produce a web-based homogeneity measure that limits the effects of inaccurate hit counts.

The *SearchYahoo* homogeneity measure is calculated for each document *d* as follows:
Steps 1 and 2: Perform steps 1 and 2 described above (WordNet Similarities). **Step 3**: Using an internet search-engine, obtain the hit counts of each member of *Freqlemmas*. **Step 4**: Order the *resulting Frequlemmas* list of *n* lemma/hit-counts combinations in descending order of hit-counts and save this list to *IndivHitsDesc*. **Step 5**: For each lemma of *IndivHitsDesc*, save to *CombiHitsDesc* preserving the ordering. **Step 6**: For each member of *CombiHitsDesc*: $CombiHitsDesc_i$, 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 7**: Calculate the gradients of the best-fit lines for the hit-counts of *IndivHitsDesc* and *CombiHitsDesc*: creating *gradIndiv* and *gradCombi* respectively. **Step 8**: *SearchYahoo* is calculated for *d* as *gradIndiv* minus *gradCombi*.

As *SearchYahoo* is taken as the difference between the two descending gradients, the measure is more likely to reveal the effects 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.

A web-service enabled search-engine was required to create a fully automated process. The Google search-engine hit-counts were less suitable, as they did not always decline as the number of query terms increased. This is perhaps because of the way in which Google combines the results of several search-engine hubs. The Yahoo web-services were therefore selected, as these produced the necessary declines for the measure to work.

Further evaluation of the Yahoo Internet searching homogeneity measure is presented in Gledson and Keane (2008), along with comparisons with similar methods using the Google and Windows LiveSearch web-services.

WordNet Domains

Magnini et-al (2002) describe the WordNet Domains[5] (Magnini and Cavaglia 2000) package as:

> 'an extension of WordNet in which synsets have been annotated with one or more domain labels, selected from a hierarchically organized set of about two hundred labels'.
> (Magnini et-al 2002 p.361)

They describe a domain (eg 'Politics') as 'a set of words between which there are strong semantic relations'. This resource is useful for measuring topic homogeneity, as it stores topic area information for word-senses directly, and complements the other measures, thus contributing to a diverse set of measures.

Two WordNet Domains homogeneity measures are calculated: *DomEntropy* and *DomTop3Percent*. These are calculated for each input document *d* as follows:

Step 1: Add each topic-content lemma of *d* to the list *TopicContents*. **Step 2**: for each WordNet sense of each topic-content lemma in *TopicContents*, find all associated domains using the Domains package, and add these to a *DomainCounts* list. This list contains each distinct domain *dom* present in the document, each with its associated count of the number of times it occurs in *TopicContents*: $freq(dom_i)$. **Step 3**: Calculate *DomEntropy* using the equation [1] above, where *n* is the number of items in *DomainCounts*, and

$$p(x_i) = freq(dom_i) \,/\, \sum_{j=1..n} freq(dom_j) \qquad [4]$$

Step 4: Calculate *DomTop3Percent* as follows:

$$100 \left(\sum_{i=1..3} freq(dom_i) \,/\, \sum_{j=1..n} freq(dom_j)\right) \qquad [5]$$

Lexical Chaining

'Lexical chains are defined as clusters of semantically related words' (Doran et-al 2004). These words are usually related by electronic dictionaries such as WordNet or Roget's Thesaurus, and chains are created using a natural language text as input.

The lexical chaining method used in our experiments is a greedy version of the algorithm described in Ecran and Cicekli (2007). Their method uses the WordNet dictionary and calculates all possible chains in the text. We adopt a greedy approach to chaining, as it is only necessary to get an overall estimate of the levels of topic homogeneity within the text, rather than producing

[5] We use version 3.2 released Feb 2007

lists of keywords or document summaries. The *LexChain* homogeneity measure is calculated for the input document *d* as follows:

Step 1: Add each topic-content noun occurrence in d to the list *UnusedNouns*. **Step 2**: For each item in *UnusedNouns*, find all other items in *UnusedNouns* that it is related to and add them to its corresponding *RelatedNouns* list. Each item of *RelatedNouns* is mapped to a score (*relScore*) using the following system (Ercan and Cicekli 2007): 10 points are awarded if the word-senses have identical lemmas or belong to the same WordNet 2.1 synset. 7 points are awarded if it is a hyponymy relationship and 4 points are awarded if it is a holonymy relationship. **Step 3**: Create chains: Iterate through *UnusedNouns* recursively, adding all related senses to the first chain, until no further linked nouns can be found. As each new node (*UnusedNouns* item) is added to the chain, remove it from *UnusedNouns*. Continue creating further chains, until no more related nouns can be found. **Step 4**: Calculate the *chainScore* of each chain by adding together all of the *relScores* contained for each sense, at each node. **Step 5**: Set *LexChain* as the *ChainScore* of the highest scoring chain.

3.4 Adjusting for Polysemy and Skew

Each of the homogeneity measures (except *WordEntropy*) has the potential to be affected by the average polysemy and sense skews of the document. The effects are measured statistically using linear regression and the resulting line of best fit equation is used to reverse them.

To calculate the adjustments for each measure, the effects of polysemy and skew must be approximated. This is achieved by applying linear regression[6] over the entire result set. The homogeneity measure is entered as the dependant variable and the appropriate[7] average polysemy and skew measures (per doc) are input as independent variables. If the homogeneity measure is affected by either (or both) of the independent variables and the effect is statistically significant, a line of best fit equation is output representing the gradient of the effect caused by those variable(s). The appropriate homogeneity measure for each input document is adjusted by subtracting the co-efficient of the gradient multiplied by the appropriate variable(s): polysemy and/or skew.

[6] Using SPSS 13.0 statistical software package.

[7] Avg. document polysemy/skews are only calculated for lemmas incl. in homogeneity measures.

4 Evaluation and Discussion

It is anticipated that the main users of a set of topic homogeneity measures are other NLP techniques. They are, therefore, best measured in terms of the actual results of the processes they are intended to improve. Human judgments can be subjective (Doran et-al 2004) and are therefore deemed inappropriate for the evaluation of this task.

Three methods are used to evaluate the homogeneity measures. Firstly, each measure is compared with its equivalent where only correct senses are used. Secondly, the WordNet::Domains (version 3.2) hierarchy (Magnini and Cavaglia 2000) is used to generate sets of words with varying levels of topic homogeneity. Each set is then tested using the proposed measures, and the results compared with those expected. Finally, the usefulness of each measure is tested by evaluating their ability to indicate the likely outcome of several co-occurrence/dictionary-based WSD measures. In the WSD literature, the main non-topic related variables reported as affecting WSD results are polysemy and skew, so these two measures will be used as the baselines.

4.1 All Senses vs. Correct Senses

Table 2 show the set of measures correlated against their correct-sense complements, where only correct senses are utilized. Most of the results are in the moderate range (.40-.60) with $AvgSim_{LESK}$ and *LexChains* achieving correlations in the high and very-high ranges. Only the SENSEVAL *DomEntropy* result falls below the moderate level, indicating that homogeneity information is, in general, not negated by incorrect-sense noise.

Measure	Correlation (Pearson's R)	
	Semcor	SENSEVAL
$AvgSim_{LESK}$	0.814**	0.834**
$AvgSim_{JCN}$	0.491**	0.610**
$AvgSim_{LCH}$	0.569**	0.474**
DomEntropy	0.575**	0.371**
DomTop3%	0.535**	0.444**
LexChain	0.659**	0.967**

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

Table 2: Correlation with correct-sense equivalents.

4.2 WordNet::Domains

The WordNet Domains package (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.

For our experiments, we have 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 sets contain 10 words each. We then assign an expected score to each set, equal to its domain level.

Measure	Correlation (Pearson's R)	
	All	Extreme
SearchYahoo	0.46**	0.80**
$AvgSim_{LESK}$	0.23**	0.15*
$AvgSim_{JCN}$	0.47**	0.45**
$AvgSim_{LCH}$	0.35**	0.25**
DomEntropy	0.70**	0.75**
DomTop3%	0.75**	0.83**
LexChain	0.42**	0.40**

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

*Significant at the .05 level (2-tailed)

Table 3: Correlation with expected scores for *WordNet::Domains* selected sets

The first column of results on Table 3 represents the correlations with expected results for all 492 word-sets. The high WordNet::Domains results (*DomEntropy* and *DomTop3%*) probably reflect the fact that they are produced using the same resource as the creation of the test sets. On the other hand, knowledge of correct senses is not required for the homogeneity measures and these scores indicate that they are capable nonetheless of capturing topic homogeneity. The *SearchYahoo*, $AvgSims_{JCN}$ and *LexChain* methods all produce promising results with correlations in the moderate range (0.40 to 0.59) and again indicate that they can capture topic homogeneity.

To indicate whether the measures are more capable of distinguishing between extreme levels of homogeneity, we repeated the above tests, but included only those sets of level-0 and level-3. The results displayed in the final column of Table 3 and provide evidence that this might be the case for the WordNet::Domains measures and *SearchYahoo*, as the correlations are significantly higher for these more extreme test sets.

4.3 Dictionary-based WSD

The three WSD algorithms selected for evaluation are the technique described in Mihalcea and Moldovan (2000) and two WordNet Similarities measures: Lesk and JCN, adapted for WSD as described in Patwardhan et-al (2004), in which they are found to achieve the best performances. Each WSD technique uses the entire document as the context for each target word. The method of Mihalcea and Moldovan (2000) is included as it incorporates several techniques, all complementary to our overall set of evaluation methods. For our experiments it is split into three: *Mih-ALL*: covering all 8 procedures, including one that relies on co-location information; *Mih-4*: utilising 'procedure-4', which involves the use of noun co-occurrence and WordNet hyponymy data; and *Mih-5-8*: using procedures 5 to 8, which involves synonymy and hyponymy. The results are adjusted to remove the effects of polysemy and *SenseEntropy* (section 3.4).

In Table 4, the fine-grained WSD accuracy results (for topic-content words) are compared to those of the homogeneity measures, (including the correct-sense measures which set the high-standard benchmark). As a baseline, non-adjusted WSD accuracies are compared with the average *polysemy* and average *Sense Entropy* of each document.

All of the 'all-senses' results, except *DomainEntropy,* are statistically significant and achieve at least low-moderate correlations with one or more of the WSD measures. All of the measures outperform the baseline correlations for most of the WSD algorithms displayed.

A *COMBINED* measure is calculated for the all-sense and the correct-sense sets of measures respectively. The measures included (based on their individual performances and maintaining maximum diversity) are $AvgSims_{JCN}$, *WordEntropy*, *DomTop3Percent*, *LexChain* and *SearchYahoo*. Each of these result sets are ordered by homogeneity score (most homogeneous

		SENSEVAL 2 & 3					SEMCOR				
		Mih ALL	Mih 4	Mih 5-8	LESK	JCN	Mih ALL	Mih 4	Mih 5-8	LESK	JCN
ALL Senses	*Word Entropy*				**-.325**	**-.301**	**-.419**	**-.442**	**-.206**		**-.286**
	AvgSim$_{LESK}$				**.500**		**.138**	**.132**	**.211**	**.121**	**.097**
	AvgSim$_{JCN}$	.286	.276	**.357**	.233		**.255**	**.231**	**.284**		**.081**
	AvgSim$_{LCH}$	.271	*.211*	*.224*	*.202*		**.254**	**.248**	**.285**		**.104**
	DomEntropy	*-.193*	*-.205*				**-.163**	**-.157**			**-.091**
	DomTop3%	**.308**	.281				**.224**	**.215**	**.145**		**.108**
	LexChain				**.344**		**.328**	**.345**	**.296**		**.095**
	SearchYahoo	-.232	**-.290**	**-.317**	**-.295**		**-.105**	**-.116**		**-.085**	**-.089**
	COMBINED	**.382**	**.421**	.270	*.243*	*.171*	**.521**	**.517**	**.256**		**.350**
CORRECT Senses	*AvgSim$_{LESK}$*	**.461**	**.467**		**.630**	*.189*	**.186**	**.174**	**.237**	**.171**	**.115**
	AvgSim$_{JCN}$	.257		.272	**.357**	**.325**	**.218**	**.162**	**.234**		**.210**
	AvgSim$_{LCH}$	**.310**		**.483**	*.187*		**.198**	**.181**	**.247**		**.122**
	DomEntropy	**-.340**	-.295				**-.334**	**-.335**	**-.106**		**-.270**
	DomTop3%	**.376**	**.353**			*.214*	**.398**	**.393**	**.144**		**.316**
	LexChain				**.372**	**.297**	**.426**	**.440**	**.164**	**.112**	**.364**
	COMBINED	**.398**	**.363**		**.398**	**.379**	**.519**	**.494**	**.294**		**.434**
Baseline	*AvgPolysemy*	*-.100*	*-.030*	-.234	(.252)	*(.113)*	**(.146)**	**(.169)**	*-.004*	*.032*	**-.144**
	AvgSenseEntropy	*.031*	*-.044*	*.143*	**-.307**	*-.019*	-.073	-.064	-.083	*.041*	**-.141**

Results in bold: significant at the 0.01 level (2-tailed); Results in non-bold: significant at the 0.05 level (2-tailed)
Italicised results: not significant at the 0.05 level but considered to be of interest
Bracketed results: inverse to expectations
All WSD results are adjusted for polysemy / *SenseEntropy*, with the exception of where compared with baselines.

Table 4: Topic-content word WSD accuracy vs. Homogeneity: Correlations

first) and banded into 5 groups making 4 cut-points at equal percentiles and numbering them from 5 down to 1 respectively. The combined measure for each document is the sum of all such scores. These measures often outperform all of the individual methods and achieve correlations of up to .52 in Semcor, the largest of the datasets.

5 Conclusions and Further Work

This paper presents a first attempt to measure *Topic Homogeneity* using a variety of NLP resources. A set of 5 unsupervised homogeneity measures are presented that require no prior knowledge of correct senses and which exhibit moderate to high degrees of correlation with their correct-sense-only equivalents. When used to measure word-sets created using the WordNet::Domains package and which have varying levels of homogeneity, they are found to correlate well with expected results, further supporting our conjecture that they represent topical homogeneity information. Finally, when compared with WSD topic-content word accuracies, the effect of topic homogeneity is shown to surpass that of polysemy and sense-entropy, which have been recognized previously as having an influence on such results. By combining these measures, correlations are improved further, often outperforming the individual methods and achieving up to .52 for over 1000 random Semcor sub-documents, again indicating their potential importance. Correlations in SENSEVAL are often higher, but due to the lower number of documents, it is more difficult to obtain statistically significant results.

Our results provide evidence that improvements could be made to WSD and other NLP methods which utilise topic features, by adapting the algorithms used depending on the level of topic cohesion of the input text. For example, window-sizes for obtaining contextual data might be expanded or reduced, based on the homogeneity level of the target text. Furthermore, non-topical features such as collocation and grammatical cues might be given more emphasis when disambiguating heterogeneous documents.

Further work includes testing the measures on other NLP tasks. A machine learning approach might also be used to further optimize the combination of homogeneity measures. Finally, it is intended that our approach should eventually be combined with a TAD method to improve WSD results.

References

Bollegala, Danushka, Yutaka Matuo and Mitsuru Ishizuka, 2007. Measuring Semantic Similarity between Words Using Web Search Engines. *In Procs World Wide Web Conference*, Banff, Alberta.

Caracciolo, Caterina, 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

Doran, William, Nicola Stokes, Joe Carthy and John Dunnion, 2004. Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization, *LNCS 2945*, pp 627-635, CICLing 2004

Ercan, Gonenc and Ilyas Cicekli, 2007. Using Lexical Chains for Keyword extraction, *Information Processing and Management*, 43(2007), pp 1705-1714.

Gledson, Ann and John Keane, 2008. Using web-search results to measure word-group similarity. *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*

Halliday, Michael and Ruqaiya Hasan, 1976. Cohesion in English, *Longman Group*

Hearst , Marti, 1997. Text Tiling: segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23(1), pp 33-64

Keller, Frank and Mirella Lapata, 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

Leacock, Claudia, Martin Choderow and George A. Miller, 1998. Using Corpus Statistics and Wordnet Relations for Sense Identification, *Computational Linguistics*, 24(1), pp 147-165

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

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

Magnini, Bernardo, Carlo Strapparava, Giovanni Pezzulo and Alfio Gliozzo, 2002. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4), pp 359-373

Mihalcea, Rada and Dan Moldovan, 1999. A method for word sense disambiguation of unrestricted txt. *In Procs. 37th Meeting of ACL*, pp 152-158

Mihalcea, Rada and Dan Moldovan, 2000. An Iterative Approach to Word Sense Disambiguation, *In Proc. FLAIRS 2000*, pp. 219-223, Orlando

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

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, Siddharth, Satanjeev Banerjee and Ted Pedersen, 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation, *LNCS 2588*, pp 241-257, CICLing 2003

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

Salton, Gerard and James Allan, 1994. Automatic text decomposition and structuring, *In Procs. RIAO International Conference*, New York.

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

Yarowsky, David, (1995), Unsupervised word sense disambiguation rivaling supervised methods. *In Procs 33rd Annual Meeting of the Association for Computational Linguistics (ACL) 1995*, pp 189-196