AUTOMATIC IDENTIFICATION OF COHESION IN TEXTS: EXPLOITING THE LEXICAL ORGANISATION OF ROGET'S THESAURUS

A.C. Jobbins & L.J. Evett

Department of Computing, Nottingham Trent University
Burton Street, Nottingham NG1 4BU, England
e-mail: amj@doc.ntu.ac.uk

Abstract

The identification of semantic relations between words can be applied to the subsequent identification of cohesion in texts. Groups of cohesive portions of text can be used to identify document structure and sub-topic areas. One method of automatically identifying the semantic relations between words is the utilisation of an existing lexical knowledge source, such as Roget's Thesaurus. A technique has been developed that exploits the lexical organisation of the thesaurus and this has been applied to the identification of semantically related words and of cohesion in texts. The development of this technique is outlined and the results from experiments conducted to investigate the application of the technique are presented and discussed.

Introduction

The term text is used in linguistics to describe a passage of written words of any length that forms a unified whole. How can this notion of a unified whole be recognised as existing and therefore forming a text? Human readers are able to determine whether a specimen in their native language constitutes a text because a reader has the ability to distinguish between a series of unrelated sentences and a series of related sentences which would form a text. This ability is based on the reader's knowledge of language and of the world.

A coherent text would be about particular subject areas and this is reflected in the language used. For example, a text about *Sailing* would contain words associated with this subject area, such as *jibe*, *mast*, *sail* and *wake*. If some of the words in a text are associated with certain subject areas then such words would not only be related to those subject areas but also to each other within those subject areas. Therefore, a text would contain groups of related words, for example, the words given associated with the subject area of *Sailing* are also related in meaning to each other. This relationship of meanings can also be expressed as a semantic relation. From this it can be stated that a text contains groups of semantically related words and it could be hypothesised that these

semantic relationships provide some of the information to the reader that enables them to identify a text as constituting a unified whole. The identification of semantic relations between words in a text could facilitate the recognition of a text that does constitute a unified whole. Where semantic relations between words exist across a text this provides information about the continuity of the use of the language in that text and therefore, continuity of the same context.

Halliday and Hasan proposed a theory of cohesion which identifies a passage as a coherent text [1]. Any piece of written language that is functioning as a unity, constitutes a text. It will display a quality of consistency that is defined in its grammatical structure and the meanings of the words used. Cohesion distinguishes a text from a set of unconnected sentences and identifies a text as a unified whole. They defined cohesion as follows:

> "*The concept of cohesion is a semantic one; it refers to relations of meaning that exist within the text*"

Halliday and Hasan identified cohesion in texts by locating chains of semantically related words. Their theory of cohesion has been incorporated into various work to analyse natural language texts e.g. [2], [3], [4], [5], [6].

Knowledge Sources

To identify cohesion in texts a means of locating semantic relationships between words is required. To automatically identify semantic relationships between words an existing electronic lexical knowledge source could be used, for example, Chodorow used on-line dictionaries [7], Rose et al. utilised text corpora [8] and Amsler used lexical knowledge-bases [9]. To elicit semantic relations from a dictionary would require analysis of the words used within the definitions. Analysis of text corpora could be used to extract words that frequently co-occur together and which could then be deemed as demonstrating a semantic relationship. However, this method of corpus analysis is based on statistical derivation of words and those cases of words that are semantically related but do not happen to occur in conjunction with each other in the given corpus could not be collected. The option of developing a lexical knowledge-base which could contain semantically related words requires a means of acquiring such information to form the knowledge-base. A source of lexical information that has so far not been exploited in depth in its electronic format for the extraction of semantic relations is the electronic version of Roget's Thesaurus and it was proposed to investigate this source further. The technique developed provides an automated method of extracting semantic information from the thesaurus.

Roget's Thesaurus

The thesaurus was first conceived by Peter Mark Roget in 1806 and it was actually finished in 1852. In his introduction he described his thesaurus as being the "*converse*" of a dictionary. A dictionary explains the meanings of words, whereas a thesaurus, given an idea or meaning, aids in finding the words that best express that idea. The third edition electronic version of Roget's Thesaurus is composed of 990 sequentially numbered and named categories. There is a hierarchical structure both above and below this category level. There are two structure levels above the category level and the top-most consists of eight major classes where each class is further divided into a number of subclasses and within this level there are the 990 categories. Under each of the 990 categories there are groups of words that are associated with the category heading given. The words under the categories are grouped under five possible grammatical classifications, namely: noun, verb, adjective, adverb and preposition. The paragraphs within categories and grammatical classifications are further subdivided into semi-colon groups which contain words that are even more closely related. Some semi-colon groups may have cross-references or pointers indicating to other related categories or paragraphs, these are given by a numerical reference to the category number followed by the related title in brackets. Figure 1 gives an example of a paragraph extract within category 373 and the grammatical classification of noun, in the thesaurus.

H00373.03.03.04092.00.00.%H Female
P00373.03.03.04093.01.00.%P N.
100373.03.03.04094.02.00.%T female,feminine gender,she,her,-ess; femineity,feminality,muliebrity;femininity,femininеness,the eternal feminine;womanhood 134 (adultness);womanliness,girlishness; feminism,women's rights,Women's Lib (or) Liberation;matriarchy, gynarchy,gynocracy,regiment of women;womanishness,effeminacy, androgyny 163 (weakness);gynaecology,gyniatrics;obstetrics 167 (propagation).

Figure 1: Roget's Thesaurus Category Extract

The thesaurus contains a collection of words that are grouped by their relation in meaning. Those words grouped together have a semantic relationship with each other and this information could be used to identify semantic relationships between words. For example, a semantic relationship between two words could be assumed if they occurred within the same category in the thesaurus.

The work of Sedelow and Sedelow supports the use of Roget's Thesaurus, where they claimed it to be an adequate representation of human knowledge and of English semantic space [10]. They considered the issue of multilocality of words in the thesaurus and the disambiguation of homographs by the application of a general mathematical model of thesauri. They demonstrated

that it is possible to develop algorithms that can elicit semantic structures from the thesaurus and from manual experimentation tested the semantic organisation. From these results they concluded:

> "*...any assertions that the Thesaurus is a poor representation of English semantic organization would be ill-founded and, given the depth of analysis, would have to be regarded as counterfactual.*"

Morris & Hirst utilised Roget's International Thesaurus for the identification of lexical cohesion in a text as an indicator of text structure [4]. Using the thesaurus they devised a system of building lexical chains from the words of a text. These lexical chains occur due to a text being about particular subject areas and finding the structure of a text involves identifying areas of a text being about the same thing. They developed a method of determining whether two words demonstrated a cohesive link by using the information contained in the index of the thesaurus where they established five types of thesaural relations between words that constituted semantic relationships. This work involved the manual computation of lexical chains and a total of five texts were analysed [11].

Thesaural Connections

The application of the thesaurus for the identification of semantic relations between words required a means of determining what constitutes a valid semantic connection in the thesaurus between two words. For example, given words w^1 and w^2 how could the lexical organisation of the thesaurus be exploited to establish whether a semantic relation $\{w^1,w^2\}$ exists between them? Morris and Hirst identified five types of thesaural relations between words based on the index entries of Roget's Thesaurus [4]. For this approach four types of possible connections between words in the thesaurus were identified for the representation of semantic relations between words by considering the actual thesaural entries. This ensured the inclusion of all words located in the thesaurus, for example, those words that form part of a multi-word thesaurus entry may not be represented in an index entry. The connections that have been identified are considered between pairs of words and are outlined as follows:

(1) Same category connection is defined as a pair of words both occurring under the same category. Figure 2 gives an example of this connection type.

word [1]: *river*
word [2]: *tributary*

words [1] and [2] both occur under category 350

Figure 2: Same Category Connection

The words would be considered to be semantically related because they were found within the same category, where a category contains a group of associated words. This connection represents the strongest connection type of the four presented because the occurrence of words within the same category indicates they are highly related and therefore have been grouped within the same area of the thesaurus.

(2) Category to cross-reference connection occurs when a word has an associated cross-reference that points to the category number of another word. Figure 3 illustrates this connection type.

word [1]: *tide*
word [2]: *river*

word [1] occurs under category 350
word [2] has a cross-reference pointing to category 350

Figure 3: Category to Cross-Reference Connection

Cross-references occur at the end of semi-colon groups and point to other categories that closely relate to the current group of words. Therefore, the words contained under the group of words a cross-reference is pointing to are related to the current group of words that cross-reference is associated with.

(3) Cross-reference to category connection can be described as the inverse of the previous connection type given in (2). The cross-references associated with a word could be matched with the categories another word occurs under.

(4) Same cross-reference connection is defined as the cross-references of two words pointing to the same category number. Figure 4 gives an example of this connection type.

word [1]: *tide*
word [2]: *flood*

words [1] and [2] both have cross-references pointing to
category 350

Figure 4: Same Cross-Reference Connection

The association of a cross-reference with a group of words indicates that the category the cross-reference is pointing to contains words that are related to the current group. Therefore, if two groups of words both have the same cross-references associated with them this implies that the words within these two groups could also be related.

Semantic Relations

A semantic relation between two words could be predicted by the satisfaction of one or more of the four connection types identified in Roget's Thesaurus. The number of matches found between a pair of words for each of these connection types could be cumulated and this could provide a quantitative indication of the level of connectivity or semantic relatedness between the two words. However, the number of matches found between a pair of words would be influenced by the number of times those words appear in the thesaurus. For example, if a word had a high occurrence rate in the thesaurus, where it could appear under many different categories and could have many cross-references associated with it, this could distort the indications of connectivity. The probability of finding matches between words of a high occurrence would be greater than those of a low occurrence rate, due to the increased number of possible matches that could be made between these words. This could effect the accuracy of the assessment of the semantic relatedness between words, where a pair of words may have attained a high degree of matches simply because they had high rates of occurrences in the thesaurus and therefore, an increased probability of matches being found. Consequently, the number of matches found for each connection type between a pair of words were normalised. Figure 5 outlines the method of this normalisation process, where *n* is the number of matches found and *max* is the maximum number of matches that could have been made between a pair of words.

$$(n/max) \times 100$$

Figure 5: Normalisation of Number of Matches Found

Word Pairs Experiment

An experiment was conducted to determine whether the connections identified in Roget's Thesaurus could be successfully applied to the identification of semantic relations between words. This was carried out on a set of semantically related word pairs and on a corresponding set of unrelated word pairs.

Method: Forty word sets were used, each consisting of three words of between four to six characters long. The second member of each set was the primary associate of the first (a related word pair) and the third member of each set was a nonassociate of the first (an unrelated word pair). For example, for the word set: {*sweet,bitter,notice*}, *bitter* is an associate of *sweet* and *notice* is a nonassociate of *sweet*. The words and their associates selected were drawn from Postman and Keppel [12] and the nonassociates, acting as a control for each pair, were derived from an experiment conducted by Evett and Humphreys which investigated the type of information required for the lexical access of visual words [13]. The list of word sets used in this experiment are given at Appendix A.

Two approaches were taken where the same category connections were considered between pairs of words and then all four of the connection types identified were considered. For both sets of these results, to determine which pair of words in each word set (i.e. the related and unrelated word pairs) represented the strongest semantic relation, the number of matches attained were compared and the word pair that achieved the highest number of matches in each word set was selected. For example, if the word pair {*sweet,bitter*} attained a total of 20.5 matches and the word pair {*sweet,notice*} attained a total of 2.7 matches then the first word pair would be selected as representing the stronger semantic relation.

Results: Table 1 shows the results for the forty word sets, giving the overall percentage of related word pairs that attained a higher number of matches than the corresponding unrelated word pairs in each word set.

Connection Types Considered	Related word pairs more strongly semantically related
Same category	80
All four connections	87.5

Table 1: Percentage of Related Words Scoring Higher than Unrelated Words

Discussion: When considering only the same category connection 80% of the related word pairs attained a greater number of matches and therefore, were more strongly semantically related than the corresponding unrelated word pairs. When all four of the connection types were considered this result was improved upon where 87.5% of the related word pairs were correctly identified as representing a stronger semantic relation. When considering just the same category connection 40% of the unrelated word pairs failed to attain any matches and when considering all four connection types 35% of the unrelated word pairs failed to attain any matches. From these results

it can be observed that considering all four connection types yields a greater identification of semantic relatedness between a pair of words.

Cohesion in Texts

Halliday and Hasan's theory of cohesion proposed a classification of the semantic relations that exist between words within coherent texts [1]. It has been shown that the connections derived from Roget's Thesaurus can be utilised for the identification of semantic relations between words. The relations in the thesaurus represent many of the relations identified by Halliday and Hasan. The thesaurus method was extended to identify semantic relatedness or cohesion across an entire text. To assess the semantic relations found in an entire text, each word in a text was compared to every other word in that text. Therefore, if a text had *n* number of words then the total number of word pairs to be compared would be: *n*-1 + *n*-2 + *n*-3 +...+ *n*-*n*. The following algorithm, hereafter referred to as the cohesion algorithm, locates semantic relations between words across a text and provides an overall measure of cohesion for that text:

(1) Filter out the function words from the text[1];

(2) For each word in the text locate it in Roget's Thesaurus and extract the related information about categories and cross-references;

(3) Compare each word in the text to all the other words in the text and for each of these word pairs obtain the normalised number of matches found;

(4) For each word cumulate the total number of matches found and then calculate the average number of matches found for that word.

The average number of matches given for each word is used as an indication of the overall level of cohesiveness that word had with the rest of the words in the text. This figure ranges from 0 to 100 where the attainment of a 0 would indicate that word did not match with any other word in a text and 100 would indicate a successful match with every word in a text. The total number of matches found for every word in a text provides an overall measure of cohesion for that text.

1. For each of these documents the function words were removed leaving the remaining content word set. The function word set includes words such as *the*, *and*, *there*, etc., these words would be limited for the identification of semantic relations between words because of their generality of usage.

Cohesion Experiment

The experiments conducted by Halliday and Hasan to test their hypothesis of cohesion were manually executed [1]. The different aspects of cohesion were looked for in various texts and they subjectively determined whether cohesion existed within these texts. The cohesion algorithm developed produces a measure of cohesion and this measure provides a quantifiable level of cohesion in a text. This automatic technique will be consistent for all texts and not influenced by subjective decision-making about the existence of semantic relations between words.

Experiments were conducted that measured the amount of cohesion in a document and also measured the amount of cohesion in a control document, thereby providing a means to assess the success of this approach. The original document was a piece of coherent text and the corresponding generated control document represented a piece of 'incoherent' text. Fifty documents, each of at least 500 words in length, were selected at random from the Lancaster/Oslo/Bergen corpus [14]. For each of these fifty documents a control document was generated. The control documents were created with similar word characteristics to the corresponding original document. This was achieved by taking every word in the original document and randomly selecting from a lexicon a word of the same length and similar word frequency to create a control document.

To assess the level of cohesion in a text, pairs of words in that text must be compared. A text can be defined as being a piece of coherent language of any size and the comparison between word pairs could be done across an entire document or in smaller units within that document. Therefore, when determining cohesion, decisions need to be made about the search space adopted, for example, adjacent words, sentences or documents. Halliday and Hasan claimed that cohesion could exist within and across sentences:

> "*Since cohesive relations are not concerned with structure, they may be found just as well within a sentence as between sentences.*"

Two experiments were conducted where content word pairs were compared across entire documents and within sentence boundaries and the cohesion algorithm was applied using two approaches where only the same category connection was considered and all four category connection types were considered.

Method: The fifty original documents and the fifty control documents detailed above were used in this experiment. The cohesion algorithm was applied to each of these documents, extracting the number of matches found for each word in a document with all the other words in that document and this was also conducted at the sentence level. Every word in a document would have an associated measure of cohesion, i.e. the average number of matches found for that word. For each document an overall measure of cohesion was produced by calculating the average of the total number of matches found. To assess whether the original document attained a higher level of

cohesion than the corresponding control document, the measure of cohesion produced for each document was compared. The document with the highest measure of cohesion was selected as the document attaining the higher level of cohesion.

Results: Table 2 shows the overall results attained for the correct selection of the original documents at the document and sentence level of analysis with both approaches to the connection types considered.

Connection Types Considered	Document Level	Sentence Level
Same category	98	92
All four connections	100	94

Table 2: Percentage of Coherent Texts Achieving Higher Scores than Controls

Discussion: When considering only the same category connection and analysing at the sentence level 92% of the original documents were successfully identified as demonstrating a higher level of cohesion than the control documents. When considering all four connection types this result was improved upon, where the number of original documents successfully identified was 94%. Analysis at the document level is shown to be more successful than analysis at the sentence level, where the consideration of just the same category connection identified 98% of the original documents and when considering all four connection types a 100% success rate was attained, where all 50 of the original documents were identified as demonstrating a higher level of cohesion than the corresponding control documents.

The experiments applying the cohesion algorithm were conducted on a large sample size of documents and for one of the approaches taken it successfully identified cohesive texts over non-cohesive texts for every set of documents investigated. This success is strong evidence for the robustness of the approach taken. This is because there are bound to be many relations between the words in the control documents, since there are so many words (500) involved. However, the relations in the texts would be expected to be more consistent and the technique successfully reflects this.

Summary and Discussion

Roget's Thesaurus is a lexical tool for language construction and understanding. It has a hierarchical structure where words are grouped by meaning, according to their semantic relations and then by grammatical categorisation. It was hypothesised that these groups of semantically

related words could be used to automatically identify semantic relations between pairs of words. A method was developed which utilised the lexical organisation of the thesaurus to identify semantic relationships between words. An experiment was conducted which applied this algorithm to pairs of associated and non-associated word pairs, identifying those word pairs that demonstrated a semantic relationship between them. It was found that for 87.5% of the associated words pairs a semantic relationship was correctly predicted over the non-associated word pairs.

Halliday and Hasan proposed a theory of cohesion which described the manner in which a text is cohesive [1]. They tested their theory through manual analysis of texts, subjectively identifying the cohesive links that existed between words in the texts they examined. Although essentially cohesion is identifiable via a series of word pairs, the theory of cohesion was proposed for identifying cohesion within units of text, whether this textual unit is a sentence, paragraph or an entire document. The technique developed that employed Roget's Thesaurus for the identification of semantic relations between words was successfully applied to the identification of cohesion across units of text. A cohesion algorithm was developed and experiments were conducted to measure the amount of cohesion found across sentences and entire documents. To validate this measure of cohesion, the amounts of cohesion across control sentences and documents were also collected and the results compared. It was found that in every case analysis at the document level attained a measure of cohesion in the original document greater than for the amount attained in the corresponding control document. The results show that this technique successfully identifies a coherent text by its level of cohesion attained relative to a control text. By calculating the average of the measures of cohesion attained for each of the coherent texts analysed, a threshold measure of cohesion could be produced which could then be used for the application of this technique to previously unseen texts.

Semantic relations between words in a text can provide much information about that text, whether it is about the overall cohesiveness of that text or the subject area of that text. Locating groups of semantically related words could be used to extract sets of words that could represent particular subject areas. These groups of words could then be applied to the problem of text subject classification. Further to this application the identification of groups of semantically related words in a text could elicit information about the structure of a text. If semantically related words adhere to particular subject areas then the identification of groups of such words throughout a text could indicate the areas in that text where certain subject areas are covered. This could provide an outline of a text's structure, where sub-topic subject area changes could be identified. For example, the identification of semantically related groups of words could cluster in different parts of a text. These clusters may represent different subject areas within that text and this could reveal the overall structure of that text. Some work has been conducted on the identification of text structure although investigations have been carried out on only a few texts. Hearst used word repetitions to divide texts into sub-topic areas [6] and Morris and Hirst used thesaural relations to generate,

manually, chains of related words [4]. Work is currently being carried out to use the present measure of cohesion to identify text structure automatically.

References

[1] M.A.K. Halliday & R. Hasan (1976) '*Cohesion in English*', Longman Group

[2] E. Ventola (1987) 'The structure of social interaction: a systematic approach to the semiotics of service encounters', *Open Linguistic Series*, Frances Pinter Publishers

[3] G. Myers & T. Hartley (1990) 'Modelling lexical cohesion and focus in written texts: popular science articles and the naive reader' in U. Schmitz, R. Schütz & A. Kunz (Eds) '*Linguistic Approaches to Artificial Intelligence*', Verlag Peter Lang

[4] J. Morris & G. Hirst (1991) 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text', *Computational Linguistics*, Vol. 17, No. 1, pp. 21-48

[5] H. Kozima (1993) 'Text segmentation based on similarity between words', *Proceedings of the 31st Annual Meeting on the Association for Computational Linguistics*, pp. 286-288

[6] M.A. Hearst (1994) 'Multi-paragraph segmentation of expository texts', *Report No. UCB/ CSD 94/790*, University of California, Berkeley

[7] M.S. Chodorow, R.J. Byrd & G.E. Heidorn (1985) 'Extracting semantic hierarchies from a large on-line dictionary', *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Illinois, pp. 299-304

[8] T.G. Rose, L.J. Evett and A.C. Jobbins (1994) 'A context-based approach to text recognition', *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, pp. 219-227

[9] R.A. Amsler (1989) 'Research towards the development of a lexical knowledge base for natural language processing', *Proc. 1989 SIGIR Conf. Assoc. for Computing Machinery*, pp. 242-249

[10] S.Y. Sedelow & A. Sedelow (1986) 'Thesaural knowledge representation', *Proceedings, 2nd Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicography*, University of Waterloo

[11] J. Morris (1988) 'Lexical cohesion, the thesaurus, and the structure of text', *Technical report CSRI-219*, Department of Computer Science, University of Toronto

[12] L. Postman & G. Keppel (1970) '*Norms of Word Association*', Academic Press, New York

[13] L.J. Evett & G.W. Humphreys (1981) 'The use of abstract graphemic information in lexical access', *Quarterly Journal of Experimental Psychology*, 33A, pp. 325-350

[14] S. Johansson (1980) 'The LOB corpus of British-English texts: presentation and comments', *ALLC Journal*, 1

Appendix A

The following table gives the forty words sets used in the word pairs experiment, where the second word is an associate of the first word and the third word is a nonassociate of the first word.

First Word	Associate	Nonassociate
sweet	bitter	notice
butter	bread	class
smooth	rough	court
short	long	card
soft	hard	tray
chair	table	weeds
sand	dune	book
seeds	poppy	ruler
cats	dogs	pool
tree	forest	violin
never	always	tartan
cold	frost	point
pepper	salt	post
under	over	peel
thread	needle	wander
take	give	mask
apple	fruit	dress
band	brass	field
stars	moon	mind
nurse	doctor	bridge
bird	robin	class
dagger	cloak	tray

First Word	Associate	Nonassociate
light	dark	card
horse	pony	pool
white	black	court
fast	slow	book
fish	tuna	mind
plane	pilot	weeds
sleep	dream	ruler
fear	afraid	violin
round	square	tartan
nail	hammer	wander
water	bath	peel
sting	wasp	mask
face	nose	post
grass	green	point
church	priest	bridge
floor	carpet	notice
mother	child	field
lamb	sheep	dress