Semantic Information from Roget's Thesaurus: Applied to the Correction of Cursive Script Recognition Output

A.C. Jobbins and L.J. Evett[1]
Dept. of Computing, Nottingham Trent University
Nottingham NG1 4BU, UK

Abstract

Compared to other electronic lexical resources available, thesauri have yet to be exploited in such depth in text processing systems. This paper describes a technique which identifies semantic relations using Roget's Thesaurus. This technique generates a *relation weight* which provides a measure of the relation between word pairs. In one experiment it is shown that relation weights can be used to distinguish semantically related words from unrelated words. A practical application of this technique is the correction of cursive script recognition errors. Typical errors include unrecognised words or multiple candidate selections for the same word position. Semantic information can be used to select between multiple candidates. Another experiment uses relation weights to locate the target word from a number of alternative candidates.

1. Introduction

The current trend for the development of text recognition systems has focused on the pattern recognition stage. However, accurate text recognition is not possible based on image features alone. There will always be cases of visual ambiguity where more than one candidate word is selected. To resolve these cases, contextual information is required such as the transitional probabilities of words or the identification of semantic relations between words. For text recognition systems to attain the same level of recognition performance as human readers, diverse knowledge sources must be considered. The need is recognised for integrating different knowledge sources for successful recognition [1] but few systems actually utilise linguistic information beyond the word level. The relatively poor performance rates of current text recognition systems, in comparison to human reading abilities, is attributable to the lack of integration in such systems [2]. Text recognition approaches based on contextual cues include the consideration of word level features [3], syntactic information [4,5] and semantic information. Semantic information has been derived from machine readable dictionaries (MRDs) [6] and large text-based corpora [7]. Another lexical resource which can be used to extract semantic information, is a thesaurus.

Processing of text for real world applications requires lexicons which provide rich information about morphology, syntax and semantics [8]. Lexicons can be constructed from existing lexical resources. A thesaurus is a unique source of information based on completely different organisational principles. Therefore, it is likely to contain semantic information additional to that captured in MRDs and corpora. Unlike a dictionary, which explains the meaning of words, a thesaurus groups words that express similar meanings or ideas. Electronic thesauri currently available include Collins Electronic English Dictionary and Thesaurus consisting of 275,000 synonyms; Random House Webster's Electronic Dictionary and Thesaurus consisting of 180,000 entries; and Roget's Thesaurus (hereafter RT) also containing approximately 180,000 entries. RT is probably the most well-known thesaurus for the English language and is a well-established writing tool (over 30 million copies have been sold [9]). The use of RT as a lexical resource for semantic information is supported by three points. Firstly, RT is organised into groups of semantically related words, representing a ready-made knowledge base of linguistic information. Secondly, it is comprised of object words [10] which constitute words used in everyday language. Therefore, it provides lexical coverage of frequently used words. Thirdly, the application of RT for text processing is supported by the work of other researchers in the field [11,12,13]. Semantic relations identified in RT compliments the existing sources of semantic information already applied to text recognition.

2. Relation Weights

The following section describes a technique which quantifies the amount of semantic relation between words based on the entries in RT. The third edition electronic version of RT is comprised of 990 sequentially numbered and labelled categories. In each of these categories, closely related words are grouped under paragraph groups and then semi-colon groups. Within the semi-colon groups there may be cross-references which point to other related categories in the

1. Contact the authors at ajobbins@resumix.com and lje@doc.ntu.ac.uk

thesaurus. Words grouped together in the thesaurus are semantically related. For example, a semantic relationship between two words can be assumed if they occur in the same category. Morris and Hirst [12] utilised Roget's International Thesaurus for the detection of lexical cohesion as an indicator of text structure. They identified five possible links in the thesaurus index to manually locate chains of semantically related pairwise words in text. Morris and Hirst's method treats word pairs as equally related; the strength of relation is not considered. In the present work, a measure of the amount of semantic relation between words is calculated. Four types of connections in the thesaurus are identified and automatically searched for in the thesaurus entries.

RT's format is not conducive for automatic extraction of information. Each paragraph is represented as one record containing a continuous string of words. Furthermore, RT contains abbreviations and annotations that are intended for human perusal but present problems for automatic processing. A paragraph excerpt from category 274 labelled "Vehicle," is given in Figure 1. It shows some of the typographical features employed (e.g. ">1e" encodes é).

```
100274.04.02.03055.13.11.%T automobile,ho
rseless carriage,car,motor car;motor,auto
;limousine,gas guzzler;saloon,two-door s.
,four-d. s.;tourer,roadster,runabout,bugg
y;hard-top,soft t.,convertible;coup>1e …
```

Figure 1: Paragraph Excerpt from Roget's Thesaurus

A method was developed to generate an alphabetically-ordered index of RT. This index is referred to as the Thesaurus Lexicon (TLex). The organisation of TLex optimises both the size and the search space of the thesaurus. Each record in TLex consists of a headword, followed by a list of the category numbers in which it appears (preceded by tag "ht") and any associated cross-references (preceded by tag "cr"). An excerpt from TLex is given in Figure 2 showing the entries for the headwords *automobile*, *convertible* and *limousine*.

```
automobile ht 274
convertible ht 13 28 147 151 274 514 673 cr
28 143 640
limousine ht 274
```

Figure 2: Excerpt from TLex Showing Three Records

A semantic relation between two words can be predicted by the satisfaction of one or more of four connection types in TLex. These connections are as follows:

[1] Same category connection (ht^ht) is defined as a pair of headwords occurring in the same category;

[2] Category to cross-reference connection (ht^cr) is defined as a headword occurring in a category that is pointed to by another headword's cross-reference;

[3] Cross-reference to category connection (cr^ht) is defined as a headword having a cross-reference that points to the category of another headword;

[4] Same cross-reference connection (cr^cr) is defined as the cross-references of two headwords pointing to the same category.

The calculation of a *relation weight* quantifies the amount of semantic relation two words demonstrate. This is based on the total number of connections made normalised by the total number of connections that could be made. The calculation of a relation weight for one connection type is expressed as

$$RW\left(w_i, w_j\right) = \frac{C\left(w_i, w_j\right)}{C_M\left(w_i, w_j\right)} \times 100$$

where $RW(w_i,w_j)$ is the relation weight for words w_i and w_j. $C(w_i,w_j)$ is the total number of connections made between words w_i and w_j. $C_M(w_i,w_j)$ is the maximum number of connections that could have been made between these words. A relation weight is calculated for each connection type. It ranges from 0, which indicates no semantic relation, to 100, which indicates the strongest possible semantic relation.

3. Identifying Semantic Relations

An experiment was conducted which compared the relation weights of semantically related pairwise words to unrelated pairwise words. The objective of investigating the relation weights was to determine whether higher weights are assigned to the related word pairs.

Method: Forty word sets were used as test data, each consisting of three words between four and six characters long (e.g. {*butter,bread,class*}). The second word in each set is a primary associate of the first word (e.g. {*butter,bread*}). The associate word pairs were selected from association norms [14]. The third word of each set is a nonassociate of the first word (e.g. {*butter,class*}). The nonassociate word

Associate Relation Weight Compared to Nonassociate Relation Weight	Connections				
	ht^ht	ht^cr	cr^ht	cr^cr	All
Associate Weight > Nonassociate Weight	32	28	28	20	35
Associate Weight < Nonassociate Weight	6	7	7	6	4
Associate Weight = Nonassociate Weight	2	5	5	14	1

Table 1: Comparison of Relation Weights for Associate and Nonassociate Word Pairs

pairs acted as controls. They were selected to balance the corresponding associate pairs in terms of word frequency and word length [15]. To generate the relation weights, connections were considered both individually and in combination. Weights were combined by addition and then re-normalised from 0 to 100. To determine which word pair in each set demonstrated the strongest semantic relation, their relation weights were compared. For example, associate pair {*church,priest*} attained a relation weight of 81.25 which indicates a strong semantic relation. The nonassociate pair {*church,bridge*} in the same word set attained a relation weight of 1.19 which implies a weak semantic relation.

Results: Table 1 shows the results for each connection and the combination of all connections. The difference between the relation weights for the associate and nonassociate word pairs was compared using a two-sample *t* test. This showed that the associate word pairs achieved significantly higher weights than the nonassociates ($t = 4.88$, $p < 0.001$, $df = 39$). This result provides evidence that the relation weight is a reliable indicator of a semantic relation between word pairs.

Discussion: When considering the ht^ht connection, 32 out of the 40 associate word pairs attained a higher relation weight than the corresponding nonassociates. For both ht^cr and cr^ht connections, 28 of the associates were more highly weighted. The results for these two connection types are identical because they are directionally associated. For the cr^cr connection only 20 of the associate word pairs attained a higher weight. The combination of all connections attained the best correct rate with 35 out of 40 associates achieving higher weights than the nonassociates. Five associates failed to score higher than the corresponding nonassociates. One of these associate pairs {*thread,needle*} may have failed because the corresponding nonassociate pair {*thread,wander*} could actually be considered related. For another of the failed pairs the outcome was a tie. Both associate pair {*seeds,poppy*} and nonassociate pair {*seeds,ruler*} were weighted at 0. The combination of all four connection types identified in TLex can be used to generate a relation weight between pairwise words. This relation weight reliably identifies semantically related words.

4. Correcting Recognition Errors

A problem for cursive script recognition is the variability between samples with regard to size, shape and slope. There are many variations on the stylisation of individual characters that can be found between writers. Even when written by the same writer, different instances of the same word can vary considerably. Recognition of cursive script can produce many errors because of the problem of determining the correct character segmentation of a word. Consequently, multiple characters are recognised which combine to produce multiple candidate words at each word position. Relation weights can be used to select the most likely correct candidate.

Method: Twenty texts of cursive script recognition output[1] were used as test data. The texts consisted of approximately 500 words each and were taken from five different subject areas (four from each). In this test data, 77.1% of the word positions had multiple candidates. Relation weights were calculated for each candidate word in the texts using a combination of all connection types. The relation weight of word *w* with text *T* can be calculated as

$$RW(w, T) = \frac{\sum_{i=1}^{n-1} \frac{C(w, w_i)}{C_M(w, w_i)} \times 100}{n-1}$$

where $RT(w,T)$ is the relation weight of word *w* compared to all other words in text *T*. This algorithm runs in order $O(n^2)$ time.

1. A large sample of recognition test data was required, so the output was generated by a simulator program [16]. This program generates a number of lexical strings for each word in a text. Lexical strings are derived from a confusion matrix of commonly mis-recognised characters in a cursive script recognition system

Subject Area of Text Group	Text Level		Sentence Level	
	Correct (%)	Tie (%)	Correct (%)	Tie (%)
Applied Science	64.2	6.5	61.4	10.2
Commerce	69.9	3.5	68.1	4.8
Pure Science	61.5	5.5	58.8	9.8
Social Science	65.3	4.9	63.6	6.8
World Affairs	71.6	2.7	71.1	5.4
Average	66.3	4.7	64.5	7.4
Total (Baseline: 30.8%)	71		71.9	
Standard Deviation	6.7		5.5	

Table 2: Cursive Script Recognition Rates for Text Level Compared to Sentence Level

Two types of context in text that effect word meaning comprehension for human readers have been determined: global context and local context [17]. *Global context* is the text in which the target word is embedded and *local context* is the immediately surrounding sentence or phrase. Hence, it can be assumed that semantic relations between words can occur both across entire texts and within sentence boundaries. Experiments were conducted at both the text level, where T represents an entire text, and at the sentence level, where T represents a sentence. Therefore, connections were considered between words at the supra-sentential and sentential levels. Candidates were ranked according to their relation weights. The highest weighted candidate was ranked in first place, the second highest weighted candidate was ranked in second place, and so on. For each text, a baseline recognition rate was calculated which represents the chance percentage of ranking the target word at the first position[1].

Results: Table 2 gives the results for both the text level and sentence level analysis. The results are given for groups of four texts belonging to the same subject area. The table shows the percentage of target words that were correctly ranked in first place for those cases where there were two or more alternative candidates. The percentage of target words that were jointly ranked in first place with at least one other word (i.e. they attained the same top ranking relation weight) is shown as a tie.

Discussion: In each case, the percentage of correct words selected was significantly higher than the baseline percentage. For each subject area the text level attained the highest correct rate, with an average of 66.3% of the target words ranked top. The majority of other target words were ranked in second place. More ties were found at the sentence level which reduced its recognition rate. Many ties occurred because of non-scoring words. For instance, if all candidates for a word position fail to form connections (i.e. have relation weights of 0) they are all tied at first place. More words failed to form connections at the sentence level because less connections are found between words within the same sentence than words across an entire text. The resolution of tied words can increase the recognition rate. At the sentence level the combination of correct and tied words attained an average recognition rate of 71.9%. Words can also fail to score because they do not appear in TLex. Therefore, they cannot form connections with other words. Omissions in TLex indicates that those words did not originally appear in RT. An investigation into the lexical coverage of RT, when compared to a 160,000 word sample from the Lancaster-Oslo/Bergen Corpus of British English [18], found it to be 92%. The majority of omissions were attributed to word inflections and proper nouns.

Using relation weights, the correct target words are selected significantly higher than chance and comparable to the rates obtained when using other semantic techniques, such as definitional overlap [6] and collocations [7]. Integration of different techniques can improve text recognition rates [2]. Relation weights are based on TLex which is generated from RT. In future work, this technique could be integrated with semantic information derived from other lexical resources, such as corpora and MRDs.

1. The chance of selecting the target word from the multiple candidates was calculated by: $\frac{1}{n} \times 100$ where n is the total number of candidates.

5. Summary

RT was investigated as a potential source of semantic information for the resolution of cursive script recognition errors. An alphabetically-ordered index of RT was created, referred to as the Thesaurus Lexicon (TLex). This organisation reduced both the size and the search space of the thesaurus. A semantic relation between two words can be predicted by the satisfaction of one or more of four connections located in TLex. A measure of this semantic relation is given by a relation weight. An experiment was conducted which calculated the relation weights for associate and nonassociate words pairs. It was found that in 35 out of 40 cases the associate pairs were weighted significantly higher than the corresponding nonassociate pairs. The identification of semantically related words is useful for automatic text processing applications, such as the correction of text recognition errors.

Text recognition provides an alternative mode of communication with a computer. Recognition can be of an existing paper document or of handwriting as it is being written. A text recognition system initially conducts pattern recognition on the target text where recognised characters or words are output. The combination of recognised characters can produce a number of candidates at each word position. The consideration of contextual cues, such as semantic relations between words, can be used to select those candidates that are most likely correct. Relation weights were successfully used to resolve cursive script recognition errors. Candidates that attained the highest relation weights were biased for recognition. Connections between words were considered at both the supra-sentential and sentential levels. The target word was selected an average of 66.3% of the time when considering connections across an entire text. The inclusion of tied words at the sentence level achieved an average recognition rate of 71.9%.

Human reading ability outperforms text recognition systems. For automatic techniques to achieve the level of human competence in cursive script recognition the integration of knowledge sources is required. Additional semantic information to that captured by the thesaurus can be obtained from corpora and machine readable dictionaries and integrated with the relation weights technique.

References

[1] S.N. Srihari, J.J. Hull & R. Choudhari, "Integrating diverse knowledge sources in text recognition," *Association for Computing Machinery Transactions on Office Information Systems*, 1(1), pp. 68-87, 1983

[2] J.J. Hull, "A computational theory and algorithm for fluent reading," *Proceedings of the 3rd IEEE Conference on Artificial Intelligence*, pp. 176-181, 1987

[3] G.J. Bellaby & L.J. Evett (1994) "The integration of knowledge sources for script recognition," *4th International Workshop on Frontiers of Handwriting Recognition*, Taiwan, 1994

[4] R. Srihari & C.M. Baltus, "Incorporating syntactic constraints in recognizing handwritten sentences," *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambery, France, pp. 1262, 1993

[5] F.G. Keenan & L.J. Evett, "Applying syntactic information to text recognition," in L.J. Evett & T.G. Rose (eds) *Computational Linguistics for Speech and Handwriting Recognition*, AISB Workshop series, Leeds, England, 1994

[6] T.G. Rose & L.J. Evett, "Semantic analysis for large vocabulary cursive script recognition," *Proceedings of the 2nd IAPR Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, 1993

[7] T.G. Rose, L.J. Evett and A.C. Jobbins, "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, 1994

[8] E.J. Briscoe, "Lexical issues in natural language processing," in E. Klein & F. Veltman (eds) *Natural Language and Speech*, Springer-Verlag, pp. 39-68, 1992

[9] P. Howard, "On English English," *Verbatim*, winter: 5, 1979-80

[10] B. Russell, *An Inquiry into Meaning and Truth*, Allen & Unwin, London, 1940

[11] S.Y. Sedelow & A. Sedelow, "Thesaural knowledge representation," *Proceedings of the 2nd Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicography*, University of Waterloo, pp. 29-43, 1986

[12] J. Morris & G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational Linguistics*, 17(1), pp. 21-48, 1991

[13] D. Yarowsky, "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora," *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, pp. 454-460, 1992

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

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

[16] F.G. Keenan, “Large Vocabulary Syntactic Analysis for Text Recognition,” Unpublished Ph.D. Thesis, Nottingham Trent University, 1992

[17] P.B. Gough, J.A. Alford & P. Holley-Wilcox, “Words and context,” in O.J.L. Tzeng & H. Singer (eds) *Perception of Print Reading Research in Experimental Psychology*, Lawrence Erlbaum Associates, New Jersey, pp. 85-102, 1981

[18] S. Johansson, “The LOB corpus of British-English texts: Presentation and comments,” *ALLC Journal*, 1, 1980