POSTPROCESSING FOR OCR: CORRECTING ERRORS USING SEMANTIC RELATIONS

A.C. Jobbins, G. Raza, L.J. Evett[1] & N. Sherkat

Department of Computing, Nottingham Trent University
Burton Street, Nottingham NG1 4BU, England
telephone: +44 (0)115 948 6018
fax: +44 (0)115 948 6518
e-mail: lje@doc.ntu.ac.uk

Abstract

Semantic relations between words can be used to aid selection from alternative candidate words output from an Optical Character Recognition (OCR) system in order to improve the overall recognition rate. One method of automatically identifying the semantic relations between words is by using an existing knowledge source, Roget's Thesaurus. A technique has been developed which exploits the lexical organisation of the thesaurus and identifies semantic relations. The development of this technique is outlined and the results from its application to OCR output is presented and discussed.

1 Introduction

Most work in OCR has been concerned with physical pattern recognition. The output of an OCR system consists of recognised words. In some cases there are alternative candidates of the sample word suggested. When the original word has been recognised and there are no other alternatives words this is recorded as a correct result. However, when there are alternative words then higher-level linguistic information can be used to determine which is the correct word.

1. To whom correspondence should be addressed.

It would be expected that words within the same text and related to the same subject areas of that text would also be related in meaning to each other. When an OCR system produces alternative candidate words at the same word position semantic information can be applied to determine which of these alternative words is most likely to be correct. For example, consider the following phrase which is output from an OCR system:

... many employers assume that women in general have lower income needs ...
tower

There are two word alternatives given at the ninth word position, words *lower* and *tower.* Semantic information can be employed to determine which of these alternative words is more likely to be the correct word. In this example, a semantic relationship exists between the words *lower* and *income*, whereas the word *tower* does not demonstrate a semantic relationship with any of the words in the phrase. Semantic information can be used to bias alternative words given for recognition. However, to apply semantic information to the output of an OCR system some source of such information is required.

2 Automatic Identification of Semantic Relations

To automatically identify semantic relationships between words an existing electronic lexical knowledge source can be used. For example, Chodorow has used on-line dictionaries [1], Rose et al. utilised text corpora [2] and Amsler used lexical knowledge-bases [3]. 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 contains 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. The thesaurus

contains explicit links between words, unlike the dictionary, and has more reliable coverage than co-occurrence information.

2.1 Roget's Thesaurus

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 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: noun, verb, adjective, adverb and preposition. These classifications are further subdivided into more closely related groups of words. Some groups of words have cross-references associated with them that point to other closely related groups of words. Figure 1 gives an example of an extract within category 373 and the grammatical classification of noun, in the thesaurus. The cross-references are given by a numerical reference to the category number followed by the title given in brackets.

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 relations 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 [4]. 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.*"

3 Developing a Post-Processing Technique

3.1 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 [5]. 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.

3.2 Quantitative 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

3.3 Relations Algorithm

The following algorithm, hereafter referred to as the Relations Algorithm, locates semantic relations between words across a text and for each word in that text an associated score of its degree of relatedness to the rest of that text is calculated.

(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 relatedness 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.

1. For every document processed the function words are 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.

4 Experiment

4.1 Method

Five files, each of at least 500 words in length, were selected at random from the Lancaster/Oslo/Bergen corpus [6]. These were scanned in using DeskScan software at 300 dots per inch and stored as tiff files. The words in these files were recognised using a word shape recogniser which uses multiple independent features and dictionary look-up [7]. For every sample word a number of alternative words from the dictionary were found based on features. These words were ranked in descending order according to the number of features matched. Those words with the highest number of features matched were biased for recognition. The output from this OCR system produced zero (i.e. word was not recognised), one or several alternative words at each word position. Figure 6 gives the output from the OCR system for the phrase: ... *women in general have lower income needs* ..., where the alternative candidate words are listed at each word position.

```
women Women
in In
general
have
lower tower Tower
income Income
needs
```

Figure 6: OCR System Output

Table 1 gives the statistics for the types of OCR errors that occurred for the content words within each of the five test files. In some cases the original word was not recognised by the OCR system and in other cases the original word was recognised but there were other alternative words suggested at the same word position. For all other cases the original word was found and there were no alternative words suggested, therefore, the OCR system successfully recognised the original word.

	Number of Words Not Recognised by OCR System	Number of Words with Alternatives	Total Number of OCR Errors
file #1	4	7	11
file #2	3	12	15
file #3	4	3	7
file #4	4	10	14
file #5	10	11	21

Table 1: OCR Errors for Content Words

The output from the OCR system for each of the five test files was input to the Relations Algorithm[1]. This produced a score for each word indicating its measure of relatedness to the rest of the text. For those word positions where there were alternative words, the word with the highest score was biased for recognition. Figure 7 shows an example of the output from the application of this post-processing technique, where the words are given followed by their score attained (the presence of function words are indicated by an F).

```
women 0.6142
F
general 4.4714
F
lower 3.6172 tower 1.9408
income 2.4903
needs 0.2843
```

Figure 7: OCR Output with Post-Processing

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. The Relations Algorithm was applied at two levels of analysis, at the sentence level and the document level. For the sentence level semantic relations were considered between words within the same

1. Output was put in lower case which removed instances of alternative candidate words that arose due to case differences, for example, alternative candidates: *women* and *Women* would be reduced to the word *women*.

sentence and for the document level semantic relations were considered between words across an entire document.

4.2 Results

Table 2 gives the results of applying the Relations Algorithm to the content words of the OCR output for both the sentence level and document level of analysis. These results are based on the number of words correctly selected for those words that had alternative candidates and the correct word was present. For example, at the document level of analysis for test file number 4 the correct words were selected for all the words that had alternative candidates (i.e. a result of 100%). The average recognition rate for the sentence level of analysis was 78.6% and this was improved upon by considering analysis at the document level where a recognition rate of 82.5% was attained.

	Sentence Level	Document Level
file #1	71.4%	71.4%
file #2	83.3%	83.3%
file #3	66.7%	66.7%
file #4	90%	100%
file #5	81.8%	90.9%
Average	78.6%	82.5%

Table 2: Percentage of Correct Content Words Found

Table 3 gives the overall results for the OCR system and the results following the application of the post-processing technique where the document level of analysis was employed. These results are given for all the content words in the test files, regardless of whether the original word was recognised.

	OCR Result	Result with Post-Processing	Maximum Result Possible
file #1	95.6%	97.6%	98.4%
file #2	94.9%	98.5%	98.9%
file #3	97.4%	98.1%	98.5%
file #4	94.9%	98.6%	98.6%
file #5	91.9%	95.8%	96.1%
Average	94.9%	97.7%	98.1%

Table 3: Recognition Rates for Content Words

4.3 Discussion

Overall the OCR system attained an average recognition rate of 94.9% for the five test files, for the content words. The application of the post-processing technique improved this recognition rate by a further 2.8% to 97.7%. For each of the five test files, the application of the post-processing technique improved upon the recognition rate of the OCR system.

For each of the five test files there were words that the OCR system failed to recognise, therefore, this introduced an error rate that would be propagated down to subsequent processing stages. This error rate prevents any post-processing of the OCR output producing a 100% recognition rate. For each of the five test files the maximum result possible was calculated (i.e. by considering the number of words that were not recognised by the OCR system). The results of the post-processing technique nearly attain the maximum possible result in every case and for test file number 4 the best recognition rate possible is achieved of 98.6%. Overall the post-processing technique was only 0.4% short of attaining the maximum possible result, whereas the output from the OCR system was 3.2% short of this target.

5 Conclusions

Those words which the post-processing technique failed to recognise may be able to be recognised by an alternative technique which employs higher-level information. Analysis of these instances revealed that when an incorrect word was selected it tended to have a low score compared to the next best score, whereas, when a correct word was selected its score tended to be comparatively higher than the next best scoring word. This points to a possible method of error detection where the results of this post-processing technique could be deemed correct if a sufficiently reliable score (a threshold measure would have to be applied here) was attained. For those words where such a score was not attained another post-processing technique could then be applied.

Taking into consideration the error rate of the OCR system (i.e. those words not recognised) any post-processing techniques have only a slight room for improvement. Without the presence of the correct word in the OCR output the problem of attaining a 100% recognition rate would be inherently difficult. However, there may be possibilities for the development of a technique that could predict words, based on semantic information.

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

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

[7] G. Raza (1995) ‘Algorithms for the Recognition of Poor Quality Documents’, *Unpublished Transfer Report*, Nottingham Trent University