Handwriting recognition using semantic information

Tony G. Rose, Lindsay J. Evett
Nottingham Trent University, England

ABSTRACT: Attempts to improve the performance of handwriting recognition systems have often involved the exploitation of linguistic constraints such as syntax or semantics. In either case, successful implementation requires the creation of a lexical database containing the relevant information. However, to create a database of semantic information from scratch for a realistically sized vocabulary is an enormous task - which is a major reason why so many semantic theories fail to "scale up" from the small, artificial domains in which they were developed. A better approach is to use existing sources of semantic information, such as machine-readable dictionaries (from which definitions may be extracted) and text corpora (from which collocations may be derived). This paper describes the development of techniques that use such resources to improve the performance of handwriting recognition systems.

KEY WORDS: handwriting recognition, semantic information, machine-readable dictionaries, text corpora, collocations.

1. Introduction

Such is the visual ambiguity of handwriting that a number of possible interpretations may be made for any written word. Human readers cope with this by making selective use of visual cues and using an *understanding* of the text to compensate for any degradation or ambiguity within the visual stimulus. Word images occur within a meaningful context, and human readers are able to exploit the syntactic and semantic constraints of the textual material [JUST and CARPENTER, 87]. Analogously, computer-based handwriting recognition systems would be enhanced by using higher level knowledge, since character recognition techniques alone are insufficient to unambiguously identify the input.

Ideally, this higher-level knowledge would be acquired by the creation of a lexical database that contains all the relevant information. However, to create a database of such information "from scratch" for a realistically sized vocabulary is an enormous task - which is a major reason why so many theories of language processing fail to "scale up" from the small, artificial domains in which they were developed. An alternative approach is to exploit existing sources of information. For example, machine-readable versions of many popular dictionaries are now available, and the definitions contained therein provide semantic information for a large vocabulary of words. Similarly, large bodies of text (known as *text corpora*) can be used to provide empirical information concerning word usage across a range of subject areas. These resources constitute readily available sources of semantic information. This paper is concerned with extracting that information and applying it to the problem of computer-based handwriting recognition.

2. Handwriting recognition systems

The system to which the current efforts are applied operates in the following way (see Figure 2). Input is written on a data pad using an electronic pen, and data is captured dynamically in the form of x-y co-ordinates. The co-ordinates are translated into a set of vector codes that are then matched against a database to produce candidate characters for the input, in the form of a character lattice. These characters are combined to produce candidate letter strings that are then checked against the system's lexicon (containing as many as 71,000 words). Those strings not on the list are rejected from further processing. The remaining strings are then combined to form a word lattice that is passed forward for further linguistic analysis, in which the candidate words are ranked according to their syntactic and semantic plausibility.

For example, consider the sentence "this is *a new savings account which you can open with one pound*" written as input to the system. The output from the lexical analyser could appear as in Figure 1, in which the alternative candidates are shown in separate columns. The problem addressed by the present paper is to select from these alternatives those words that are most likely to be correct.

```
this is a hen savings gallant which you can open with one round
tail      new         account       boy car oxen pick ore pound
tall      see         accept        nos oar oven lick due found
trio                                    our           bra hound
```

Figure 1. *Typical output from a handwriting recognition system*

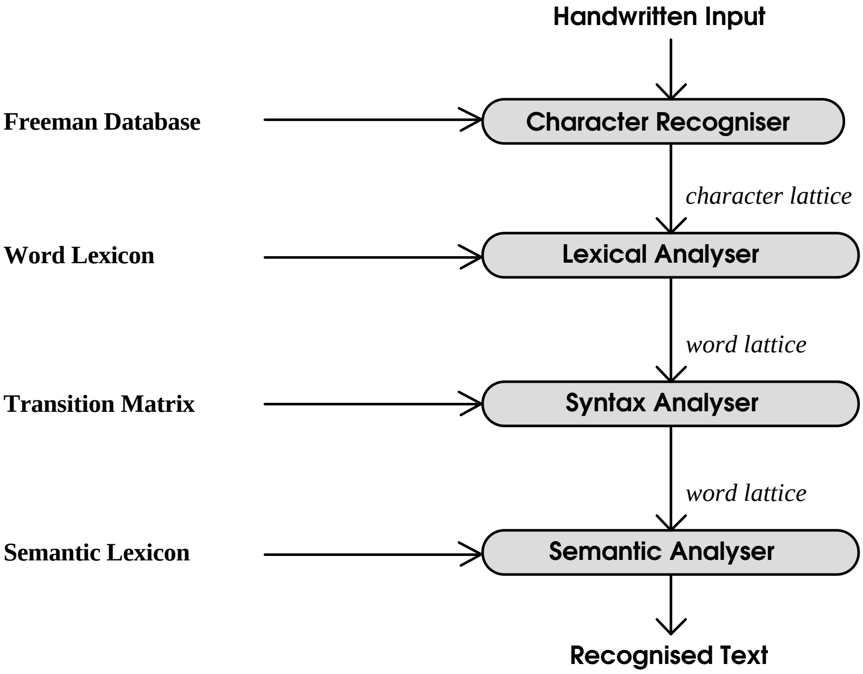

Figure 2. *System Overview*

3. Using machine-readable dictionaries

A technique has been developed that can detect semantic relations between words by comparing their dictionary definitions [ROSE and EVETT, 92]. The technique proceeds by accessing the definition of each candidate word, and counting the "overlap" (i.e. the number of words in common) with the definitions of each of its neighbours. Once a complete sentence has been processed in this manner, the candidates with the highest "overlap" in each position are deemed to be the correct words.

An investigation was set up to evaluate this technique. Test data consisted of fifteen documents each of 500 words, taken from different domains, with alternative word candidates in each position as in the above example. Three different dictionaries were used in turn as the source of definitions: **(i)** The Collins English Dictionary (CED), **(ii)** Longman's Dictionary of Contemporary English (LDOCE), and **(iii)** The Oxford Advanced Learner's Dictionary of Current English (OALD).

The performance of the technique was assessed by measuring the percentage of times it identified the correct word from the alternative candidates. However, since this data concerns only word positions in which there were two or more "competing" candidates, it does NOT directly reflect the overall (system) recognition rate. Table 1 shows the average percentage of correct words identified by each dictionary, and the standard deviation between domains.

In each of the dictionaries, the purpose of the definition is to provide a precise, sense-based statement of the meaning of each word. However, dictionaries are a human resource and as such reflect the subjective influence and styles of their respective lexicographic creators. Furthermore, dictionaries themselves are often designed to accommodate the needs of a particular subset of the possible readership (e.g. second language learners) and the contents of the definitions will vary according to the perceived needs of that target readership.

Table 1. *Percentage correct by domain for each dictionary*

	CED	**LDOCE**	**OALD**
computing	79.6	71.9	69.9
energy	66.7	70.1	74.1
engineering	64.7	57.9	59.4
business	69.9	74.3	68.4
employment	62.9	70.8	61.3
finance	66.7	73.1	68.7
biology	69.2	72.4	72.3
chemistry	76.0	76.9	71.4
maths	67.4	62.9	56.9
education	63.8	63.8	59.1
medicine	65.8	67.9	63.2
sociology	73.1	69.6	70.0
economics	69.2	74.8	67.8
history	63.6	63.5	67.4
politics	66.7	76.9	78.2
Average	**68.3**	**69.8**	**67.2**
Std. Dev.	**4.57**	**5.41**	**5.86**

Evidently, the LDOCE outperforms the CED and the OALD. The main reason for this may be related to the manner in which LDOCE definitions are constructed. It is claimed that the entries within LDOCE are defined using a controlled vocabulary of about 2,000 words, and that the entries have a simple and regular syntax [BOGURAEV and BRISCOE, 88]. This has the effect of reducing the *entropy* of the definitions, by cutting down on the randomness with which their constituent words are chosen. This increases the chance of successful overlaps, since the probability of two semantically related words being defined using common terms is now proportionately increased. This reduction in the "noise" within definitions means that where semantic relations are present, the definitional overlap technique is more likely to detect them.

Performance across domains, is however, highly variable, with no obvious pattern emerging. The CED is the most consistent, with 12 of the 15 scores being in the 60-70% range. The LDOCE shows more variability, with 9 scores in the 70-80% range, and one particularly low score (57.9%, for engineering). The OALD shows the most variability, with 5 scores in the 70-80% range and 3 in the 50-60% range. Given that the eventual needs of a realistic implementation may be biased towards commercial applications, the results for the *Business*, *Employment* and *Finance* documents are particularly significant. The LDOCE scores consistently in the 70-80% range for these documents, whilst the CED and OALD are both consistently in the 60-70% range. It may be inferred from these results that LDOCE would be the preferred choice of machine-readable dictionary.

In this investigation it was assumed that semantic relationships actually exist between words in ordinary sentences. However, it is possible that the definitional overlap effect may be due to factors other than this. A further experiment was designed to test whether such semantic relationships exist in the text used as test data. In this experiment, pairs of words that had shown a strong overlap were selected from within the sentences. A number of subjects (twenty-five) judged these to be semantically related compared to a control group of candidate pairs that had not shown a strong overlap. This result proved to be statistically significant ($p < 0.001$) using the Mann-Whitney U test [DOWNIE and HEATH, 70]. It therefore supports the assumption that words within ordinary sentences exhibit genuine semantic relationships, and these can be identified by the definitional overlap process.

4. Using text corpora

There are certain classes of English word combinations that cannot be explained purely by existing syntactic or semantic theories. For example, consider the use of "*strong*" and "*powerful*" in the following phrases:

"to drink strong tea"
"to drive a powerful car"

Both fulfil the same syntactic role, and both make a similar semantic modification to the subject. However, to interchange them ("*powerful tea*" and "*strong car*") would undoubtedly be judged anomalous by most English speakers. These predisposed combinations are called co-occurrence relations or collocations, and account for a large proportion of English word combinations [SMADJA, 89].

An algorithm was developed to analyse any given text corpus and transform the distributional patterns of the constituent words into a set of collocations. This algorithm was based on the work of LANCASHIRE [87], although modifications were made to reformat the output as a sorted, lemmatised, dictionary-like structure. This information could now be used to measure the plausibility of individual collocations in data such as the above, and thereby identify the correct word candidates. Using the example shown in Figure 1, the word "*savings*" should collocate more strongly with "*account*" than with "*gallant*" or "*accept*", and "*account*" should collocate more strongly with "*open*" than with "*oxen*" or "*oven*", and so on.

The collocation analysis technique proceeds by comparing the "neighbourhoods" of each word candidate (up to a distance of four words) with their likely collocates (as defined by the results of corpus analysis). Each candidate is assigned a score according to the overlap between its neighbourhood and its list of likely collocates. Once a complete sentence has been processed in this manner, the candidates with the highest scores in each position are deemed to be the correct words. The "window size" of four words reflects both the results of empirical investigation [ROSE, 93] and the findings of other researchers (e.g. [JONES and SINCLAIR, 74]).

A further experiment was set up, using the same test data as before. Two types of collocation were investigated: **(a)** general, and **(b)** domain-specific. Consequently, it was necessary to create a number of "collocation dictionaries". The first of these was the *General Collocation Dictionary* (GCD), which was derived from 5 million words of text, taken from all subject areas within the Longman Corpus [SUMMERS, 91]. The remainder were domain-specific collocation dictionaries, derived from 500,000-word domain-specific corpora. No part of any test document had been included in the corpora used for the creation of any collocation dictionary. For each of the fifteen documents, the collocations were analysed, once using the GCD and once using the appropriate domain-specific dictionary.

Table 2 shows the percentage of correct words identified by each collocation dictionary for each test document. As before, since this data only concerns word positions in which there were two or more "competing" candidates, it does NOT directly reflect the overall (system) recognition rate.

The average performances of the general and the domain-specific dictionaries are extremely close, with the domain-specific being slightly superior (by 2.0%). However, for 8 of the 15 documents, the general collocations are more effective (by as much as 11.9% in one case). This is somewhat surprising, since it would be reasonable to assume that domain-specific corpora would contain the most appropriate collocations for domain-specific documents.

Table 2. *Percentage correct by domain for each collocation dictionary*

	GENERAL	SPECIFIC
Computing	84.7	82.9
Energy	76.3	66.7
Engineering	70.3	68.4
Business	79.5	75.3
Employment	73.4	61.5
Finance	73.2	63.6
Biology	75.2	77.3
Chemistry	83.8	83.0
Maths	70.5	63.9
Education	68.7	88.7
Medicine	69.1	83.6
Sociology	64.1	73.1
Economics	83.6	94.4
History	70.8	80.0
Politics	77.4	88.6
Average	**74.7**	**76.7**
Std. Dev.	**5.95**	**9.95**

Explanations for this inevitably concern (a) the content of the textual material used as data, and (b) the content of the collocation dictionaries. Evidently, any given document will consist of a variety of language structures, some of which will be general (i.e. not exclusively associated with any particular domain) and some domain-specific (i.e. with restrictions on word senses, etc.). This ratio of "general" to "specific" material will vary between documents and domains, such that a high proportion of "general" material may render the use of a domain-specific collocation dictionary less appropriate, and vice-versa.

Secondly, the domain-specific dictionaries were derived from smaller corpora than the GCD and therefore contained fewer entries: 5,545 (on average) compared to 12,475 in the GCD. In particular, although the domain-specific corpora were all roughly the same size, due to variations in the type:token ratio the resultant collocation dictionaries varied greatly: from 3,960 entries to 7,748 entries. Indeed, this variation in size very closely matches their performance: those larger than average tend to do better than the GCD, and those smaller tend to do worse. The variation in performance is further reflected by the higher standard deviation of the specific dictionaries.

Evidently, it would seem that the number of entries is an important consideration in the creation of any collocation dictionary. The analysis of a single domain may be fruitful only if the size and type:token ratio of the domain corpus are such that collocates for a sufficiently wide variety of types can be acquired. A more reliable approach is to analyse as large and varied a corpus as possible to maximise the coverage of the resultant dictionary. Additionally, good coverage is required to process all the *alternative* candidates produced by the lexical analyser. However, it must be appreciated that for real-time handwriting recognition applications, processing and storage requirements constitute an overhead that should be minimised. Consequently, if the implementation is restricted to a single domain, then a specific dictionary may represent the best compromise between performance and efficiency.

5. Summary

A number of techniques have been developed that use existing sources of semantic information to improve the performance of handwriting recognition systems. When presented with multiple word candidates, the best of the machine-readable dictionaries identified the correct word in 69.8% of cases (on average). General collocations extracted from a 5 million-word corpus identified the correct word in 74.7% of cases. The use of domain-specific collocations increased this figure to 76.7%.

The performance level that could be expected given a random choice of candidates is 30.4% correct for this data. Clearly, the use of dictionary definitions and collocations represents a significant improvement on this baseline. Although the character recogniser itself provides a ranking that may be associated with each candidate in the word lattice, its accuracy is variable (depending on the identity of the writer, the extent of training, the handwriting sample used, etc.) and contextual information is still needed to disambiguate many word positions [EVETT *et al.*, 93]. Studies have shown that the application of semantic information can substantially improve the overall system performance, but the extent of the improvement is highly dependent on the quality of the output from the earlier stages in the recognition process [ROSE, 93].

Collocations are just one of a number of sources of higher-level knowledge that may be independently applied to handwriting recognition data. However, the question of how to combine these knowledge sources remains highly problematic, and it is unclear how much influence should be allocated to each of them. Steps towards a solution often begin with a consideration of the interface between the individual modules. In the present system, the semantic analyser has been designed to take input in the form of a lattice of word candidates, and to output that word lattice with associated scores. Consequently, the semantic analyser can be applied to any system within which word lattices are produced: handwriting, OCR, possibly even speech systems. This is true also of the syntax analyser, so these two modules can be run independently, in parallel if necessary, producing their own sets of results. The lexical analyser has been designed to accept input in the form of a character lattice, so this can work with any recogniser that produces output in this format.

Indeed, these modules have also been used in the design of an integrated OCR system. Given a TIFF file as the system input, the data flow and processes were organised as in Figure 3, using a network of transputers [SHERKAT *et al.*, 93]. The "voter" constitutes a module in which results are combined and a unique solution identified. The architecture is such that processing begins in each module as soon as data becomes available, and partial results flow along pipelines between the modules so that all may work simultaneously whenever possible.

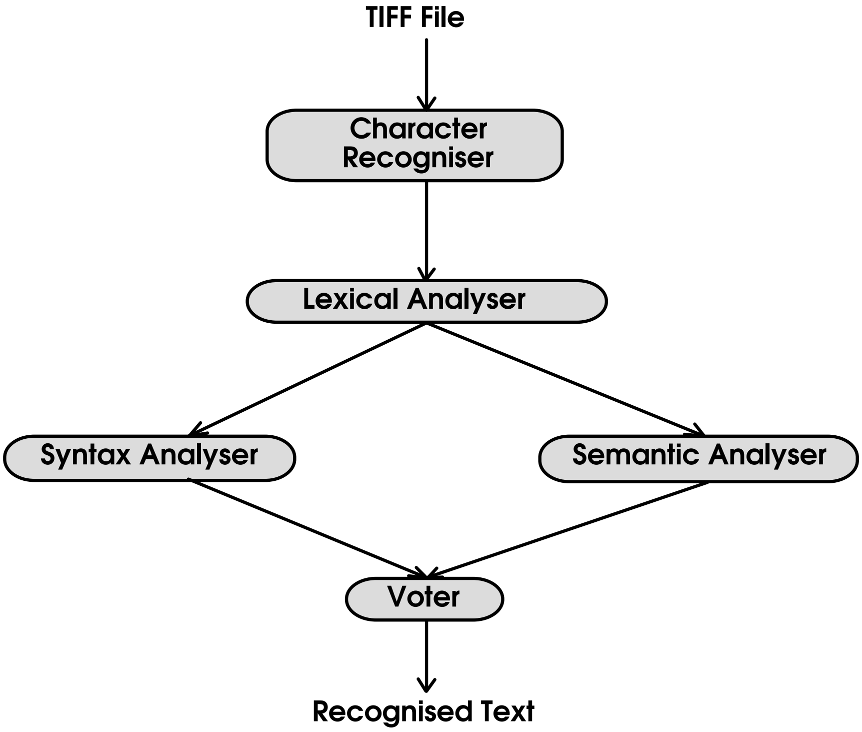

Figure 3. *An OCR system design*

Evidently, there are a number of limitations to the semantic analysis techniques described above. Firstly, since processing takes place within an integrated recognition architecture, computational overheads and memory requirements have been minimised wherever possible. For this reason, both definitional overlap and collocation analysis are based on lemmatised (root) forms rather than inflections. However, it is clear that some collocations only exist in particular inflected forms

[SCHUETZE, 93]. Consequently, it is intended to acquire inflected versions of the above collocation dictionaries and compare these with their lemmatised equivalents (using the same handwriting recognition data).

Secondly, the collocation analysis makes no use of function words (again to minimise processing overheads). However, these are an essential part of a number of important linguistic phenomena such as phrasal verbs [SINCLAIR, 87]. It is intended therefore to incorporate such information into future acquisition methods, and compare the results with the "content-word only" predecessors. Thirdly, no use is made of word order information. However, linear precedence has been shown to be a significant factor affecting the manner in which words associate with each other [CHURCH and HANKS, 89]. This is particularly relevant to a run-time recognition application, since data is usually input in one direction anyway (i.e. left-to-right). Indeed, the results of more recent studies provide further evidence for this, showing that word order constitutes a significant constraint that should be fully exploited by text recognition systems [ROSE *et al.*, 94].

Fourthly, the collocation analysis makes no use of distance information. Some collocations may be independent of distance, but there are others whose behaviour is highly distance dependent [JONES and SINCLAIR, 74]. It is appropriate that future system development should exploit this constraint. Finally, the acquisition of collocational information remains somewhat problematic. Whereas definitional information can be obtained from LDOCE for some 55,000 headwords, the acquisition of a similar number of collocational entries would require the processing of an immense corpus. Consequently, such a strategy would involve the need to store and process a much greater quantity of data. However, advances in hardware technology and parallelism should reduce the significance of such overheads.

The techniques described in this paper have been developed to disambiguate word lattices produced by an on-line handwriting recognition system. The algorithms have been coded in C and optimised for efficiency such that computation times are well within the limits required by real-time applications. They can also be run off-line, and used for other recognition applications: when applied to output from an OCR system, collocation analysis (using the general collocation dictionary) produced a performance of approximately 82% correct [ROSE and EVETT, 93]. It is envisaged that the compilation of larger, more sophisticated collocation dictionaries will form the basis of further studies.

References

BOGURAEV B., BRISCOE E., (Eds.), 1988, *Computational Lexicography for Natural Language Processing.* Longman, London.

CHURCH K., HANKS P., 1989, Word association norms, mutual information and lexicography. *Proc. 27th Meeting of the ACL,* pp. 76-83.

DOWNIE N., HEATH R., 1970, *Basic Statistical Methods.* Harper and Row, New York.

EVETT L.J., ROSE T.G., KEENAN F.G., WHITROW R.J., 1993, Linguistic Contextual Constraints for Text Recognition. *ESPRIT Deliverable DLP 2.1*, Project 5203.

JONES S., SINCLAIR J., 1974, English Lexical Collocations. *Cahiers de Lexicologie*, 24, pp. 15-61.

JUST M.A., CARPENTER P.A., 1987, *The Psychology of Reading and Language Comprehension*. Allyn and Bacon Inc., Boston.

LANCASHIRE I., 1987, Using a Textbase for English-language research. *Proc. 3rd Ann. Conf. of the UWC for the New Oxford English Dictionary*, Waterloo.

ROSE T.G, 1993, Large Vocabulary Semantic Analysis for Text Recognition. *Unpublished PhD thesis*, Dept. of Computing, Nottingham Trent University.

ROSE T.G., EVETT L.J., 1992, A large vocabulary semantic analyser for handwriting recognition. *AISB Quarterly*, No. 80, pp 34-39.

ROSE T.G., EVETT L.J., 1993, Text recognition using collocations and domain codes. *Proc. of the Workshop on Very Large Corpora*, Ohio State University, pp. 65-73.

ROSE T.G., EVETT L.J., JOBBINS A.C., 1994, A context-based approach to text recognition. *Proc. Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada.

SCHUETZE H., 1993, Word space. *Advances in Neural Information Processing Systems*, HANSON S., COWAN J., GILES C. (Eds.) , San Mateo CA, Morgan Kaufman.

SHERKAT N., POWALKA R., WHITROW R., 1993, A parallel engine for real time handwriting and optical character recognition. *Proc. JET POSTE 93 - The 1st European Conference on Postal Technology*, Nantes, France.

SINCLAIR J., 1987, *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins, Glasgow.

SMADJA F., 1989, Macrocoding the lexicon with co-occurrence knowledge. *Proc. 1st Int. Lexical Acquisition Workshop*, Detroit, Michigan, pp.197-204.

SUMMERS D., 1991, Longman/Lancaster English language corpus: criteria and design. *Technical Report*, Longman Publishers.