Text Recognition using Collocations and Domain Codes[1]

T.G. Rose & L.J. Evett

Dept. of Computing, Nottingham Trent University, Nottingham, England

phone: 0602 418418 email tgr@uk.ac.trent.doc fax: 0602 484266

Keywords: handwriting recognition, OCR, collocation, sense tagging

Abstract: Text recognition systems require the use of contextual information in order to maximise the accuracy of their output. However, the acquisition of such knowledge for a realistically sized vocabulary presents a major problem. This paper describes methods for extracting contextual knowledge from text corpora, and demonstrates its contribution to the performance of handwriting recognition systems.

Introduction

Such is the visual ambiguity of handwriting that a number of possible interpretations may be made for any written word. Indeed, this is true of any text, but particularly handwritten text since the segmentation between the individual characters is often indistinct. 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 & Carpenter, 1987]. Analogously, computerised text recognition systems would be enhanced by using higher level knowledge. Character recognition techniques alone are insufficient to unambiguously identify the input, particularly that of handwritten data.

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

[1] The research reported in this paper was supported by the European Commission under the ESPRIT initiative.

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, such as machine-readable dictionaries [Rose & Evett, 1992] and text corpora. Corpora can be used to provide empirical information (such as collocations) concerning word use across a wide range of subject areas [Smadja, 1989]. A further source of information, known as *domain coding*, can be acquired either from a machine-readable dictionary or generated as a further product of corpus analysis. This paper is concerned with the acquisition of collocations and domain codes, and their contribution to text recognition systems.

Text Recognition Systems

Due to its greater inherent degree of ambiguity, handwritten text is seen as the main application of the following techniques. However, the methods may also be applied to OCR data, or indeed to any recognition system that produces word alternatives as its output (e.g. speech recognition systems). The system to which the current efforts are applied operates in the following way: 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. These characters are combined to produce candidate letter strings, which are checked against a list of acceptable words (as many as 71,000), and those strings not on the list are rejected from further processing. The remaining strings are then combined to produce possible phrases.

For example, consider the sentence "*this is a new savings account which you can open with one pound*" written as input to the system. This could produce the output shown in Figure 1 (in which the alternative candidates are shown in separate columns). The problem is now 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

Collocations

Introduction: 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 phrases "*to drink strong tea*" and "*to drive a powerful car*". They fulfil the same syntactic role, and both make a similar semantic modification to the subject. However, to interchange them (*"powerful tea"* & *"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 part of English word combinations.

An algorithm was developed to analyse a given corpus and transform the distributional patterns of the constituent words into a set of collocations. This algorithm was based on the work of Lancashire [1987], 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. For example, 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"*.

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, 1993] and the findings of other researchers (e.g. Jones & Sinclair, [1974]).

Investigation 1

1.1 Method: Test data consisted of fifteen documents each of 500 words, taken from separate domains, with alternative word candidates in each position as in the above example. Two types of collocation were investigated: (a) general, and (b) domain-specific. To this end, 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. 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.

1.2 Results: Table 1 shows the percentage of correct words identified by each collocation dictionary for each test document. N.B. - 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.

	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

Table 1: Percentage correct by domain for each collocation dictionary

1.3 Discussion: The average performances of the general and the domain-specific dictionaries are extremely close (they differ by only two per cent). This is somewhat surprising, since it would be reasonable to assume that domain-specific dictionaries would contain the most appropriate collocations for domain-specific documents. However, for 8 of the 15 documents, the general dictionary is more effective (by as much as 11.9% in one case).

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.

However, the 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. Furthermore, although the domain-specific corpora were all the same length, due to variations in the type:token ratio the resultant dictionaries varied greatly in size (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. This variation in performance is further reflected by the higher standard deviation of the specific dictionaries.

The performance level that could be expected from a random choice of candidates is 30.4% correct for this data. Clearly, the use of collocations represents a significant improvement on this baseline. Although the handwriting recogniser itself provides a ranking of the alternative candidates, its accuracy is variable (depending on the identity of the writer, the extent of training, the handwriting sample used, etc.) and it is clear that contextual information is needed to disambiguate many word positions [Powalka et al, 1993]. In this respect, collocations are just one of a number of sources of higher-level knowledge that may be independently applied to text recognition data. However, the question of how to combine these knowledge sources remains highly problematic, since it is unclear how much influence should be allocated to each of them. It is desirable therefore to measure their contribution in isolation before attempting to combine them within an integrated system [Evett et al, 1993].

Evidently, it would seem that "big is beautiful" when it comes to acquiring collocations from text corpora. 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 text recognition systems. However, it must be appreciated that for a real-time application such as handwriting recognition, processing and storage requirements constitute an overhead that must 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.

Domain Codes

Introduction: Domain codes are essentially labels that may be associated with words to describe the domain or subject area with which they are usually associated. The codes themselves can be organised as a simple series of subject areas, or as a hierarchy whereby specific domain codes imply inheritance of a more general subject area. Using them as an aid to recognition involves firstly determining the domain of the data, and then using the codes to favour those word candidates whose senses are appropriate to that domain.

A system of domain codes can be either created from scratch or obtained from a machine-readable dictionary (e.g. LDOCE). The first method is impractical due to the sheer size of the task; the second is derivative and produces a domain coding system that may not necessarily be the most suitable for a particular application. A third method, based on corpus analysis, has been developed that does not suffer from either of the above drawbacks. This method proceeds on a domain-by-domain (i.e. corpus-by-corpus) basis according to the following algorithm:

(1) Start with the raw domain corpus and reduce it to its uninflected root forms;
(2) Produce a type-frequency distribution for this corpus;
(3) Obtain a corresponding distribution from an undifferentiated (general) corpus;
(4) Normalise these frequency distributions so that each type's frequency is expressed as a proportion of the total number of tokens within that corpus;
(5) Calculate the comparative frequency of each type (i.e. its "distinctiveness");
(6) Select those words which have a distinctiveness of 3.0 or above, i.e. their frequency is at least three times greater in the domain corpus than in the general corpus (this threshold has been selected arbitrarily and needs to be investigated empirically);
(7) Normalise these comparative frequencies by expressing them as natural logarithms. The resultant file now contains those words distinctive to the domain, and a measure of their distinctiveness within that domain;
(8) Repeat steps (1)-(7) for all domains for which corpora are available.
(9) Merge the domain codes from each domain into a single file. This file now contains that section of the lexicon that displays domain-based behaviour, and identifies the domains with which each word is associated, with a measure of the strength of that association.

Essentially, the codes thus produced reflect the relative frequency of words within a domain-specific corpus compared to their frequency in an undifferentiated corpus. The domain code lexicon can never be exhaustive: their coverage can only be as complete as the corpora from which they are derived. However, the codes produced by this technique have a distinct advantage over those of LDOCE: they are *quantitative* rather than *qualitative*. Instead of just labelling words with a code to say whether they belong to a given domain or not (such distinctions are not always clear-cut), they also provide a measure of the strength of this association.

Investigation 2

2.1 Method: A set of domain codes was derived from LDOCE, and a further set (of comparable size) derived from a number of domain-based corpora according to the above algorithm. However, since a given word must display specialised domain-based behaviour to justify the possession of a domain code, high frequency words tend to be excluded. For this reason, using domain code information for text recognition tends to leave many word positions in the data unresolved. Consequently, it was preferable to apply each set of codes as a "supplement" to the collocational analysis technique (using the GCD). Test data consisted of the five documents that had produced the *poorest* performance in Investigation 1 (as they left the most room for improvement).

2.2 Results: Table 2 shows the percentage of correct words identified using (a) just collocations (i.e. the GCD), (b) collocations plus corpus-based codes, and (c) collocations plus LDOCE codes.

2.3 Discussion: Both the LDOCE domain codes [Walker & Amsler, 1986] and the corpus-based domain codes [Evett & Rose, Note 1] have been shown to be effective for topic identification. Consequently, if the domain of a document is known, it is reasonable to assume that such codes could contribute to recognition. However, their average contribution to the test documents is minimal. Since the corpus-based codes cover only ten broad subject areas, they may be simply not distinctive enough to identify the correct word.

Domain	Collocations	+ Corpus Codes	+LDOCE Codes
Engineering	70.3	75.2	70.3
Maths	70.5	68.8	74.1
Sociology	64.1	65.8	64.1
History	70.8	72.9	70.8
Finance	73.2	73.2	73.2
AVERAGE	69.8	71.2	70.5
STD. DEV.	3.03	3.40	3.85

Table 2: Percentage correct for each combination

Conversely, the LDOCE, with its elaborate coding system of 120 major subject areas and 212 subfields may be too fine, as the multitude of categories creates confusion rather than effective discrimination. It may transpire that an alternative coding system (possibly designed around an intermediate level of representation) may be optimal for recognition applications. The lack of coverage of both current sets of codes renders them ineffective for certain documents (e.g. *Finance*). The only successes seem to be in *Engineering* (with the corpus-based codes) and *Mathematics* (with the LDOCE codes). This result may reflect the propensity of these domains for using distinctive specialised terminology.

Summary

Investigation 1 has shown that collocations can be used to identify the correct words within text recognition data taken from a number of domains. Investigation 2 has demonstrated the limited improvement obtained by the use of two sets of domain codes.

Each of the techniques described above has been applied to handwriting recognition data. However, they are equally appropriate to other recognition applications, such as OCR or some speech systems. In the case of the former, it has been possible to test the various techniques using output from an existing system. The data source consisted of a collection of 22 scanned documents, and the OCR output consisted of the recognised characters that had been post-processed by a lexical lookup to identify word candidates. In most cases, the correct word had been uniquely identified, but for 38 word positions there was one or more alternative candidates. Collocation analysis identified the correct word for 31 of these, chose an incorrect candidate for 4 cases and left 3 positions unresolved. This represents a performance of 81.58% correct.

Evidently, there are a number of limitations to the above methods. Firstly, since processing takes place within an integrated recognition architecture (i.e. working in real time, with a pattern recogniser, lexical analyser and syntax analyser), computational overheads and memory requirements must be minimised wherever possible. For this reason, both collocation analysis and domain code analysis are based on lemmatised (root) forms rather than inflections. However, it is clear that some collocations only exist in particular inflected forms [Schuetze, forthcoming]. Consequently, it is intended to acquire inflected versions of the above collocation dictionaries and compare these with their lemmatised equivalents (using the same text 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, 1987]. 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 & Hanks, 1989]. Indeed, this is particularly relevant to a run-time recognition application, since data is usually input in one direction anyway (i.e. left-to-right). Consequently, the next phase of collocation acquisition will be to create a set of uni-directional collocations and compare them with their bi-directional equivalent. Finally, the collocation analysis makes no use of distance information. Clearly, some collocations are independent of distance, but there are others whose behaviour is highly distance dependent [Jones & Sinclair, 1974]. It is appropriate that future system development should exploit this constraint.

Likewise, there are ways in which domain code acquisition and analysis can be improved. For example, the acquisition algorithm has associated with it a number of parameters (e.g. the number of domains covered, the "specificity" of the domains, the optimal value for the threshold of distinctiveness); all of which need to be empirically investigated. Moreover, it is hoped that the current limitations concerning coverage will be eliminated by the availability of larger corpora. Consequently, their coverage may be such that domain code analysis becomes a viable aid to recognition in its own right, i.e. without needing supplementary collocation information. Part of this investigation will be the development of alternative coding systems, based on varying levels of domain-specificity.

Clearly, many of these extensions will involve the need to store and process a greater amount of data, which could compromise the efficiency of real-time applications such as handwriting recognition. It is suspected that the trade-off between performance and run-time efficiency will form the basis of further empirical investigation.

Note 1: **L.J. Evett & T.G. Rose** (1993) "*Automatic Document Topic Identification*", paper submitted to 2nd IAPR Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan.

References

K. Church & P. Hanks (1989) "*Word association norms, mutual information and lexicography*", Proc. 27th Meeting of the ACL, pp. 76-83.

L.J. Evett, T.G. Rose, F.G. Keenan & R.J. Whitrow (1993) "*Linguistic Contextual Constraints for Text Recognition*", ESPRIT Deliverable DLP 2.1, Project 5203.

S. Jones & J. Sinclair (1974) "*English Lexical Collocations*", Cahiers de Lexicologie, **24**, pp. 15-61.

M.A. Just & P.A. Carpenter (1987) "*The Psychology of Reading and Language Comprehension*", Allyn & Bacon Inc., Boston.

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

R.K. Powalka, N. Sherkat, L.J. Evett & R.J. Whitrow (forthcoming) "*Dynamic cursive script recognition: a hybrid approach to recognition*", Sixth International Conference on Handwriting and Drawing, Paris, July 1993.

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

T.G. Rose & L.J. Evett (1992) "*A Large Vocabulary Semantic Analyser for Handwriting Recognition*", AISB Quarterly, No. 80, Brighton, England.

H. Schuetze (forthcoming) "*Word space*", in S. Hanson, J. Cowan & C. Giles (Eds.) "Advances in Neural Information Processing Systems", San Mateo CA, Morgan Kaufmann.

F. Smadja (1989) "*Macrocoding the lexicon with co-occurrence knowledge*", Proc. 1st International Lexical Acquisition Workshop, Detroit, Michigan, pp.197-204.

D.E. Walker & R.A. Amsler (1986) "*The use of machine-readable dictionaries in sublanguage analysis*", in R. Grishman & R. Kittredge (Eds.) "Analyzing Language in Restricted Domains", LEA, Hillsdale, N.J.