Paul Holmes-Higgin & Khurshid Ahmad, University of Surrey

Assembling and Viewing a Corpus of Texts: Self-organisation, Logical Deduction and Spreading Activation as Metaphors

Abstract

Lexicographical and terminological work is increasingly dependent on the analysis of texts, particularly texts organised in a corpus and being made available through computer systems. We argue that the developments in corpus linguistics, artificial intelligence, connection sciences, and lexicography and terminology, can be conjoined together to analyse the various facets of a text. In particular, the users of corpora will be allowed to explore the 'family resemblance' of such texts with other texts. This will help in the creation of a user-defined corpus of texts that belong to a *family*, all having their own idiosyncrasies but all sharing something through a common 'genetic' pool. Our approach, a strictly computational account of corpus organisation and usage, will help corpus builders and end-users to incorporate as much as is known about the texts in general and whatever is known about the contents in the description of texts for storage and for retrieval. We demonstrate our interdisciplinary approach by describing how texts in a computer-based corpus can be (a) represented by using knowledge representation formalisms, such as *frames*, (b) automatically classified by using *self-organising artificial neural networks*, and (c) managed by using a hybrid representation scheme wherein interactive activation and competition networks are used in conjunction with frames and deductive data bases.

1. Introduction

Corpora of texts are increasingly being used to investigate a range of literary and linguistic phenomena: from authorship attribution to genre analysis; from lexicographic evidence to language change; from the study of dialects to syntactic and semantic analysis; from optical character recognition studies to language development and second language acquisition. Corpora of texts can also be used to investigate trends in science and technology, particularly through the analysis of texts produced in the different parts of the world and then relating the tokens in the text, including terminology usage, authorship and institutional aspects, to major developments in scientific methods, product developments and innovation management.

The advent of the World-Wide Web (WWW) comprising texts, images and sounds from many countries and in many languages, has added and will continue to add to the amount of available texts. This is reflected in the availability of the network 'crawlers', programs designed to access texts across the WWW based on keyword search. It is, for example, possible to collect literally hundreds of texts in the languages of numerically-determined minorities in a range of disciplines, genres and so on.

Any large scale organisation of data involves the development of categorisation formalisms and cataloguing schemes that will, in the first instance, distinguish between the physical storage of data, for example, which files, what file directory structure, what compression algorithms are to be used and so on, and the logical organisation of data, that is the organisation of data based on a model of the enterprise from which the data originates. The terms 'physical' and 'logical' organisation are computer science terms and the concept underpinning the two terms is that no matter how the data is stored the user is always able to access, organise and analyse the entire data he or she wants.

It appears that information scientists focus on the lexical-semantic aspects, lexicographers on the lexical and pragmatic aspects, and that corpus linguists focus on the semantic and pragmatic aspects. However, this does not mean that the lexicographers ignore the semantic aspects and the information scientists the pragmatic, rather the assumptions regarding these attributes are not articulated and made explicit for building a categorisation scheme that will be put into operation on a computer system. The approach discussed in this paper looks at a method for automatically categorising (or representing) text knowledge: knowledge about the text or knowledge in the text. Once the method is put into operation, one can use it for retrieving texts from the corpus on the basis of similarities and differences in lexical, semantic and pragmatic attributes of a set of texts.

The 'register variation' in the so-called balanced corpora, that is corpora that comprise a wide-range of styles and varieties according to a number of criteria like the lengths of the text in each of the registers, can be quantified in terms of differences grammatical features. In practical terms such differences are essential for part-of-speech taggers and syntactic parsers, in that the 'probabilities associated with grammatically ambiguous forms are often markedly different across registers' (Biber, 1994:179). This may also have an impact on corpus-based lexicography, in that it will be very important for a lexicographer to be aware of the type of texts he or she wishes to use in order to use quantitative data related to word usage information. Meyer (1991) has noted that different text types, such as learned texts and fiction, can be characterised by differences in the use of appositions expressed through reference, synonymy, attribution and hyponymy.

However, a number of other studies, particularly on the acquisition of semantic knowledge from texts, appear not to be too concerned about register variation. Consider, for instance, Pustejovsky, Bergler and Anick's (1995) work on 'sublanguage' vocabulary, based exclusively on specialist language texts. Smadja's work on extracting collocation patterns from texts is based on news-wire texts and stock market reports (1994). Like Pustejovsky and colleagues, Smadja also uses corpus tagging as an important preliminary for his analysis.

The debate in corpus linguistics literature on 'balanced' versus 'open' corpora, and 'representative' versus 'randomly' selected corpora, is of significant import: if the register-variation, vocabulary-variation, and the variation of other lexical, semantic and pragmatic attributes, has some bearing on the results of a corpus user's analysis, then it is essential that the user is capable of including or excluding some or all parts of a text within a corpus and across corpora. This will enable the user to quantify his or her results of the analysis of the behaviour of words and sentences in terms of the above mentioned variations.

Our approach suggests that a corpus can be defined at a meta-level and can be organised in a distributed manner, possibly on many computer systems distributed across the globe. The task of a lexicographer, or a terminologist, involved in extracting subcorpora would be to use family resemblances between the parameters describing texts in a corpus as the basis for its organisation. Extraction of a subcorpus would then require the specification of a *prototype* text against which similar texts may be adjudged.

1.1 Notes from corpus linguistics literature

Corpus linguists use a variety of different schemes for cataloguing texts: these schemes are based on conventions defined by the original disciplines of literary criticism, stylistic analysis, text linguistics, sociolinguistics, information science and so on. All the information the computer has is highly encrypted and represented thus on the computer. Of course, the term *representation* itself has been defined as a set of lexical, syntactic and semantic conventions for organising knowledge. The computer has a description, often called a data model, which is the result of analysis by a team of computer scientists and corpus builders. These descriptions are used as a basis of a classification system that is then used to *label* individual text files: the labels 'informative' and 'imaginative' are usually the superordinates of learned/popular and fiction/non-fiction labelling, respectively. This kind of tagging is hand-coded in that most of these labels are specified by the corpus builder to name files, or to create relation tables or hierarchical trees for storing the files in directories.

One solution to such a manual labelling exercise is that the user may be allowed to re-configure the labels and to create his or her own hierarchies of texts. This enables the user to view the corpus selectively. We have discussed this kind of approach previously and argued for a *virtual* corpus rather than a physically hard-coded corpus handed down by corpus linguists (Holmes-Higgin, Ahmad and Abidi, 1994). However, we believe that such an approach, despite its utilitarian appeal, does not bring to fore the major issue: the *representation* of texts in text corpora on the basis of a set of conventions that help put into operation a collection of primitives, both specific and generic, that will enable a computer system to automatically organise a multi-dimensional space (cf. Biber 1994) in which these texts can be stored and subsequently retrieved.

We suggest three new ways of representing texts in text corpora: a *frame-based* approach inspired by knowledge representation formalisms in conventional artificial intelligence, *self-organised* corpora based on connectionist notions of unsupervised pattern classification, and a corpus organisation that is a hybrid of the classical AI formalisms and the connectionist formalism.

2. 'Frame-based' organisation

The first approach involves the use of a *frame-based* system: each text is represented by a frame, the slots of each of the frames are named after a primitive, and the fillers are the values that can be assigned to the slots. Some of the slots can be filled in by default, others may be filled in by the users, and still others can be computed by a set of programs that can operate semi-autonomously – *demons* – that can add, delete and deduce values if needed. The frames can be linked to each other, and indeed some slots can themselves be attached to other frames.

The frame-based system, originally due to Marvin Minsky and whose logical properties were specified by Patrick Hayes, can be used to infer new facts from old data, can inherit data from other frames, and can delegate data to still others. The frame system used in our study, MARVIN, was developed by Holmes-Higgin (1990). A graphical example of this representation is shown in Figure 1, with two instances of texts and one instance of a publisher and an author. The language of 'text (1)' can be inherited from the native language of the text's author. In 'text (2)' this value is overridden with an explicit value (this might lead us to be able to deduce that the text is either a translation or that the author is bilingual).

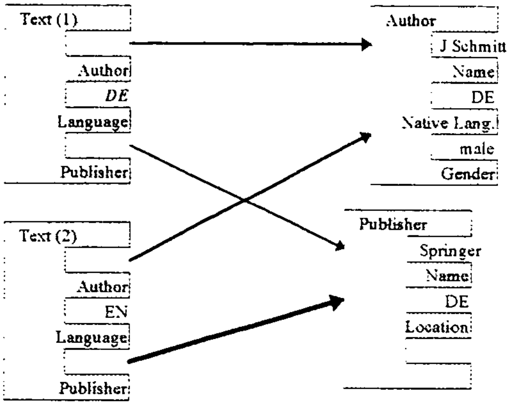

Figure 1: Representation of text descriptors as MARVIN objects with properties and relationships to other objects. The bold arrow shows a deducible relationship, while the italic text shows an inherited slot value.

More complex relationships between frames can be defined using demons, which allow the specification of heuristics that may be used to make inferences from known facts about the texts. With the MARVIN system, queries can be made such as requesting the language of a particular text, which may, for example, be determined through an if-needed demon when no explicit value is given. For the example above, a demon can be defined to deduce that the publisher for 'text (2)' is the same as the publisher of other texts by the same author.

3. Self-organised corpora

There are a number of criticisms of classical AI. Amongst one, put forward by a number of authors, is that the knowledge representation formalisms – or the so-called *schemata* – have too rigid a structure to represent real world objects and events: the formalisms represent knowledge at a macroscopic level and that these formalisms are too coarse-grained to represent individual features of objects and events they claim to represent. Above all, there is the criticism that any formalism that claims to simulate aspects of human cognitive behaviour, like the frame-based formalism, must have some capacity to learn in an autonomous manner.

This brings us to our second method of representing text in a corpus, based on 'family resemblance' (Wittgenstein, 1953). We have used Kohonen *feature maps*, a connectionist formalism that can not only represent, no matter in howsoever a limited fashion, objects in the real-world, but also learns to represent. Neural computing systems are able to compute through the use of a (potentially large) number of simple processing elements, the so-called *neurons* organised in a network, and furthermore, they can also learn to compute.

Kohonen feature maps are used in the categorisation of complex objects that need a number of primitives to effectively represent them. A set of these primitives is called a *feature vector*, a vector that may have any number of components. Kohonen maps comprise an algorithm that maps these components onto a two-dimensional surface. This surface is effectively a network of neurons, or a feature map. The algorithm helps in the computation of the 'distance' between the feature vectors of each of the objects: proximate neurons can be deemed to belong to a category. Self-organisation has also been explored in the Information Retrieval literature, with both neural network approaches (Gersho & Reiter, 1990; Lin, Soergel & Marchionini, 1991) and statistical algorithms (Faieta & Lumer, 1994).

Initially, each of the real world objects is assigned a feature vector and randomly mapped onto the map: in our case the objects are (full) texts. The distances between the vectors are then computed over a number of cycles. If at the end of a number of cycles the distances do not change much beyond a given statistical threshold, then the network organises the objects in such a way as to reveal 'hidden' categories. Note, here there is no *tutor* to tell the system as to how accurate the categories were at the end of each cycle. This unsupervised mode of learning was first put forward by Teuvo Kohonen, hence the term Kohonen feature maps. Essentially, the algorithm used in building a feature map involves a layer of *adaptive artificial neurons* that gradually develop into an array of feature detectors that are spatially organised such that, when stimulated by the presentation of a feature vector, one and only one adaptive neuron gets activated: the location of the excited neuron becomes indicative of statistically important features of the input stimuli.

Any text can be described by a number of text external features, such as its author, title, text type, publisher, publication location and date, and so forth. These we regard as the defining features of a text. Each text can also be distinguished by the frequency distribution of words in the text. For instance, the first 50 most frequent nominals will give us some indications about the physical and abstract entities and events being discussed in a scientific or technical text (see Ahmad, 1995 for details). This set of words we regard as the individual features of a text. A feature vector describing a text can be construed to comprise two parts: the individual and defining parts.

We have considered the classification, and subsequent retrieval of texts, from a specialist text corpora and used Kohonen maps to set up the classification. To this end, we have created a 62 component feature vector. The first 50 are the individual features and the rest are the defining features. The individual features were encoded as a 50-digit binary number comprising '1' and/or '0'. We have also specified 12 defining features: text type, word count, publisher's location, publication date, authors origins and so on (Table 1). Each feature was assigned a binary code according to the possible values the component was intended to represent.

Text Type	Code (3)	Word Count	Code (3)	Publisher Location	Code (3)	Publication Date	Code (3)	Author's Origins	Code (3)
unknown	000	unknown	000	unknown	000	unknown	000	unknown	000
advertisement	001	1 -100	001	UK	001	before 1915	001	African	001
journal	010	101 – 500	010	rest of Europe	010	1916 – 1930	010	Afro-Carib.	010
book	011	501 – 1000	011	USA	011	1931 – 1945	011	Asian	011
manual	100	1001 – 10000	100	N/S America	100	1946 – 1960	100	British	100
newspaper	101	10001 – 100000	101	Asia	101	1961 – 1975	101	European	101
official	110	100001 – 1m	110	Africa	110	1976 – 1990	110	Chinese	110
other	111	1m+	111	Australasia	111	after 1990	111	other	111
Author's Gender	Code (2)	Page Count (2)		Author's Native Lang.	Code (4)				
unknown	00	unknown	00	unknown	0000	Am. Eng.	0100	Italian	1000
male	01	<= 100	01	Catalan	0001	Br. Eng.	0101	Spanish	1001
female	10	<= 1000	10	Danish	0010	French	0110	Welsh	1010
		<= 10000	11	Dutch	0011	German	0111	other	1111

Table 1. A set of defining feature vectors for texts.

For gender the component was represented by a two-digit binary number, but for author's origins a three digit binary number was used, and for his

For gender the component was represented by a two-digit binary number, but for author's origins a three digit binary number was used, and for his or her native language a 4-digit binary number was used. (Subject field and the language of the text was also assigned 5-digit and 4-digit binary code, although these are not shown for reasons of brevity). All told, defining features require a 32 digit binary number.

Using this method, we have been able to successfully classify texts in a specialist domain, namely automotive engineering. A total of 79 texts were used in the training cycle with each text given a 62 component, 82-digit binary encoded feature vector. Within the automotive texts there were four subdomains (anti-lock braking systems, four-wheel drive, catalytic converters and miscellaneous), however this information was not encoded in the feature vectors. The defining features were derived automatically using System Quirk, a terminology and lexicography management system (Ahmad and Holmes-Higgin, 1995), to compute the frequency of the 50 most frequent nominals in a corpus of over 300,000 words distributed over 130 texts and to compute the presence or absence of these 50 words in each of the 79 texts used for training the feature map.

The trained Kohonen feature map of these 79 texts clearly showed four basic clusters, each cluster corresponding to the four subdomains: this reflects the dominance of the individual feature vectors that determined the clusterings. However, the proximity of the texts also appeared to depend upon the defining feature vectors. Furthermore, when the feature vector of another text, on which the map had not been trained, was presented to the Kohonen network, it appeared to successfully classify this text in the correct subdomain cluster. Figure 2 sums up the training and recognition processes, with the user being able to provide some control over the classification through the specification of the linguistic metrics to be used.

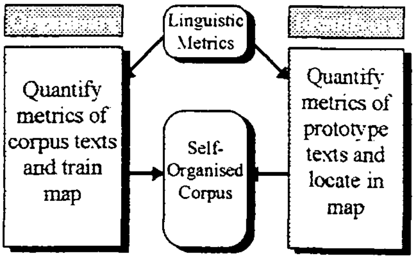
Linguistic Metrics
Quantify metrics of corpus texts and train map
Self-Organised Corpus
Quantify metrics of prototype texts and locate in map

Figure 2. Storing and retrieving from a self-organising corpus.

4. Hybrid approach to corpus organisation

There are a number of limitations associated with the classical AI representation formalisms, like frames, semantic networks and predicate logic, on the one hand, and on the other hand, other criticisms exist for neural network-based representations of knowledge. The AI formalisms lack flexible reasoning, neural plausibility and self-organisation considerations, and the neural network type formalisms are flexible, plausible and some are self-organising, but on the whole lack structured representation, cannot easily process hierarchical data and generally do not generate output that can be easily visualised.

From the two experiments mentioned above, it became clear to us that the AI knowledge representation formalisms can be used to represent the individual features well, and that the defining features can be represented well by neural networks. We were thus motivated to propose a hybrid representation.

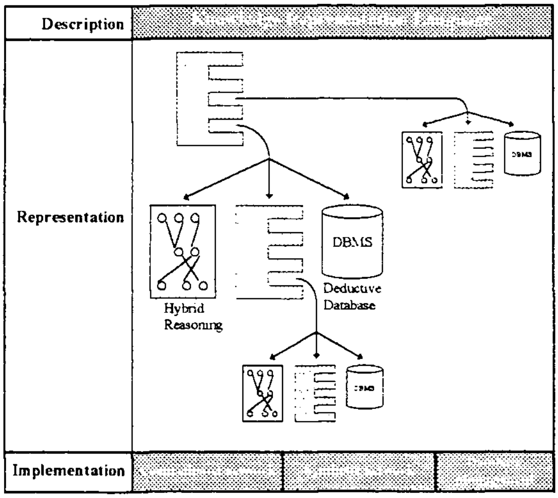

Figure 3: A symbolic description of a concept can lead to a number of combinations of representation and processing paradigms.

The MARVIN system was adapted such that it can also access neural networks and relational databases, and it could provide communication between frames, neural networks and relational databases. This complex system was called μMARVIN (Holmes-Higgin, 1995). To meet the goal of embedding more intelligent behaviour in the text identification process, an extended browsing system architecture based on part of System Quirk has been developed. The extended architecture directs object queries to μMARVIN rather than directly to a database, and thus provides a principled environment for describing and managing the text knowledge.

5. Conclusions

In our experiments we have, in effect, posited that there is some idealised representation scheme that may be used to describe any text within any collection, which, to be of some benefit, is an abstraction or reduction of the text. While there may seem to be some circularity in way in which we have defined our experimental representations, this is an essential ingredient for a model of knowledge discovery. The framework of such a model for discovering text knowledge can be found in Holmes-Higgin (1995). The purpose of our experiments was to explore this idealised representation of text collections from different perspectives: through explicit engineering, describing the texts to some predefined specification; and through implicit self-organisation, where the content of the text collection dictates the classification of the texts. These two approaches represent either end of a scale of hand-coded to automatic descriptions of text collections, and we have proposed three experiments that explore this continuum, as shown summarised in Figure 4.

Much of the above discussion relates to the organisation of corpora on computer system. Our experiments were designed to emphasise the point that lexica can be so organised such that these collections of text can be viewed from a range of perspectives. Such a view will be indispensable for lexicographers and terminologists as they work with increasingly large number of texts, particularly when they can themselves access the texts through the communications networks.

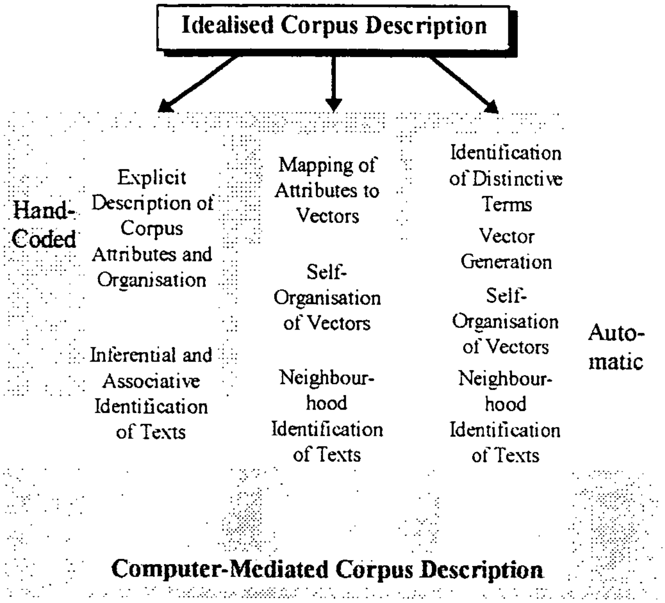
Idealised Corpus Description
Hand-Coded
Explicit Description of Corpus Attributes and Organisation
Inferential and Associative Identification of Texts
Mapping of Attributes to Vectors
Self-Organisation of Vectors
Neighbourhood Identification of Texts
Identification of Distinctive Terms
Vector Generation
Self-Organisation of Vectors
Neighbourhood Identification of Texts
Automatic
Computer-Mediated Corpus Description

Figure 4: Towards a hybrid computer-mediated corpus description.

Lexica that are prepared for different audiences in mind will frequently require selective views of the corpora, that is user-defined subcorpora, that are used for finding entries, their elaborations and usage. Last but by no means least, it is also becoming clearer that there is an urgent need to collate the various views about texts as pre-theoretical and theoretical notions, and to build text analysis systems on the basis of this collation.

References

Ahmad, K. 1995. “Pragmatics of Specialist Terms: The Acquisition and Representation of Terminlogy”, in: P. Steffens (ed.) *Machine Translation and the Lexicon*. Heidelberg, Springer, pp. 51–76.

Ahmad, K. & Holmes-Higgin, P. 1995. “System Quirk: a unified approach to text and terminology”, in: H. Picht & G. Budin (eds.) *TAMA ’94 Proceedings, Third TermNet Symposium, Terminology in Advanced Microcomputer Applications – Recent Advances and User Reports*. Vienna, TermNet, pp. 181–194.

Biber, D. 1994. "Using Register-Diversified Corpora for General Language Studies" in: S. Armstrong (ed.) *Using Large Corpora.* London, The MIT Press, pp. 179–201.

Faieta, B. & Lumer, E. 1994. "Exploratory Database Analysis via Self-Organization" in: *Proceedings Intelligent Multimedia Information Retrieval Systems and Management, RIAO 94. October, 1994, New York*, pp. 570–584.

Gersho, M. & Reiter, R. 1992. "Information retrieval using self-organizing and heteroassociative supervied neural networks" in: *Proceedings of the Int. Neural Network Conference, Paris, 1992*, pp. 361–364.

Holmes-Higgin, P.R. 1990. *MARVIN User Guide and Reference Manual.* Technical Report. Dept. of Mathematical and Computing Sciences Department, University of Surrey. Guildford.

Holmes-Higgin, P.R. 1995. *Text Knowledge: the Quirk Experiments.* PhD. Thesis. Dept. of Mathematical and Computing Sciences, University of Surrey. Guildford.

Holmes-Higgin, P., Ahmad, K. & Abidi, S. R. 1994. "A description of texts in a corpus: 'virtual' and 'real' corpora" in: *Proceedings EURALEX-94.* Amsterdam, pp. 390–402.

Kohonen, T. 1990. *Self Organisation and Associative Memory.* London, Springer-Verlag.

Lin, X., Soergel, D., & Marchionini, G. 1991. "A Self-Organising Map for Information Retrieval", in: A. Bookstein et al. (eds.). *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Developments in Information Retrieval. Chicago, Illinois, Oct. 13–16*, pp. 262–269.

Meyer, C.F. 1991. "A Corpus-based study of apposition in English", in: K. Ajmer & B. Altenberg. *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London, Longman, pp. 166–181.

Pustejovsky, J., Bergler, S., & Anick, P. 1994. "Lexical Semantic Techniques for Corpus Analysis", in: S. Armstrong (ed.) *Using Large Corpora.* London, MIT Press, pp. 291–318.

Smadja, F. 1994. "Retrieving Collocations from Text: Xtract", in: S. Armstrong (ed.) *Using Large Corpora.* London, MIT Press, pp. 143–178.

Wittgenstein, L. 1953. *Philosophical Investigations.* Oxford, Basil Blackwell.