Natural-language processing and automatic indexing: a reply

Kevin P. Jones

Korycinski and Newell, in the previous issue of The Indexer, have produced a useful review of natural-language processing by computer, and have mentioned some problems of automating the indexing process.1 Nevertheless, their views must be approached with certain reservations.

Firstly, like many others working on the edge of the information retrieval scene, they appear to regard information retrieval (IR) systems as being little more than mechanized card indexes to bibliographical references: this is no longer the case, and possibly never was. An increasing number of IR systems are like the book index: i.e., an integral part of the text to which they relate. In some cases, the normal mode of access to such texts is via the index, and the browsing mode occupies a subsidiary role. Like hypertext, some modern IR techniques offer an alternative to indexes as previously perceived.

Secondly, little is offered to show how human indexing is performed. Unfortunately, the literature on the core activity (i.e., extraction of words/concepts from texts) is sparse. Perhaps this is a potential application for expert systems, but human indexers appear reluctant to analyse how they perform this activity. They discuss the form of indexes at great length, however, which may explain why greater progress has been made on automating this aspect of index construction. Perhaps the slow progress towards computerizing the core activity reflects Cleverdon’s strictures on inter-indexer consistency.2 Human beings also easily miss the obvious—even within relatively controlled environments.3

Human indexing is, almost certainly, partially rooted in statistical analysis, the type of statistical approach which enables innumerate shepherds to know the strength of their flocks. We all know whether the cake is being cut fairly!

Very little human indexing effort is wasted on disambiguating such linguistic delights as ‘the arrow of time’—the task is too mammoth for this. Instead, introspection based on my own non-computational indexing activity suggests that (if all stoplist/function words are eliminated) the following are/are not candidate index entries:

1. Words to reject (or almost entirely so) are words which feature throughout the text. That is, the overall text is about something that these words portray: for instance, the word RUBBER is unlikely to form many index entries in a book about rubber.
2. Words which are found significantly more often in a particular section of text than in the text as a whole (a chapter about oxidation in a book about rubber would be expected to contain the repeated use of key words such as OXIDATION, DEGRADATION, DETERIORATION, OXYGEN and OZONE: these would be obvious candidate index entries).
3. Words which are very improbable, either on an absolute basis, or within the context of the specific text, and which are not being used metaphorically.
4. Words indicative of a particular theme: e.g., locations in works relating to topography; names of musical scores or literary works within their appropriate contexts, etc.

Dillon4 appears to have been on the track of (1) and (2) with his statistical measures, but (3) is the province of Fairthorne’s (mythical) teenage girls5 (the words that they cannot understand), or of the expert indexer. It is the ability to trace the unexpected that makes an index or an information-retrieval system ‘good’ from the user’s standpoint. Taken to the ultimate, it would appear that a concordance (or the ability to search an entire machine-readable text) should afford this ability. Concordances are simple to produce by computer (in fact, some of the earliest computational linguistics was applied to concordance construction). Perhaps too little effort has been expended on attempting to convert the simple-to-produce concordance into an index.

In no way would I wish to restrict the potential applicability of computational linguistics to automating the indexing process, but it would seem that much of what has been attempted so far in computational linguistics is either too complex or too restricted. A much more pragmatic approach is probably needed. In large texts, idiomatic expressions (such as RAINING CATS AND DOGS) are likely to be encountered: dictionaries of such expressions (in machine-readable form) are simple to create, and would be simple to consult for a mechanized indexing system. More effort on the production of such tools might be more profitable than that devoted to theoretical linguistics. It may be observed that expert systems appear to favour pragmatism.
The higher theory of indexing

_The Journal of Documentation_ 46 (2), June 1990, includes a 21-page article by Bernd Frohmann, ‘Rules of indexing: a critique of mentalism in information retrieval theory’. After a brisk, five-word definition of indexing as ‘representing documents for information retrieval’, he gives a practical outline of the process, with a welcome emphasis on the devising of the content of an index rather than the structure and arrangement on which attention is so often focused:

Indexing is generally taken to consist of at least two distinct operations. The first involves either the implicit or explicit representation of a document by an indexing phrase. The second involves the translation of the terms of the indexing phrase into the lexicon of a controlled indexing vocabulary, with due regard to the semantics and syntax of the indexing language... the first [step] continues to be lamented as an intellectual operation fundamental to indexing yet so far resistant to analysis.

He goes on to consider the ‘intellectual operation’ of the indexer; the nature of thought and language; mental rules and their construction; ‘unruly indexing’; knowledge and truth; Wittgenstein’s criticism; theoretical linguistics; intertextuality; the purposes of text retrieval; the identification of needs. ‘Is the retrieval of truth a desirable (or even feasible) retrieval practice?’ he speculates. We recommend this article for indexers wishing for a profound and philosophical view of their occupation.

Charging for an index? Disgraceful!

‘When his father’s [Major Matthew Connolly] health grew worse towards the end of the war, [Cyril] Connolly became more tolerant and sympathetic. Moreover, his [father’s] endearing qualities became more apparent as he aged, as did the similarities between father and son. After all, [Major Connolly’s] passion for collecting shells and filing them away in boxes was not that far removed from collecting literary contributions and arranging them neatly in magazines. Indeed, in the early Forties he had applied his skills to the task of compiling the semi-annual index for _Horizon_ [edited by Cyril Connolly] with his usual efficiency. In the letter that accompanied his finished work, he wrote: ‘Herewith Index to Vol. III, which I suppose will come in useful to you. I suggest that _Horizon_ give me for it an Honorarium of 5/- [five shillings—in today’s UK currency, 25 p]. The British Museum values my time at 4/- an hour, and as this has taken me exactly 6 hours to make out, (4 x 6 = 24/-) I imagine it is cheap at 5/-.’ It is not easy to feel affection towards someone with a mind like that but as the war brought him closer to his father, Connolly did his best to be a good son.’


Thanks for permission to quote the extract above to Hamish Hamilton Ltd. and Harper & Row Inc., New York.

Medical book reviews

In the _Bulletin of the Medical Library Association_ 74 (1), 1986, Kurt Kroenke reported a survey of 480 book reviews in medical journals conducted through six months in 1983. The survey identified 25 features of reviews in three categories: style, content (general and specific), and extrinsic characteristics; and tabulated and discussed the findings. Only 14% of the reviews considered the indexes of the books among the features of content addressed.

Indexers on wheels

The days when a good engine sufficed to make a good car—the days when any Alfa Romeo could be relied upon to be, therefore, a good car—have long gone. A sweet gearbox, generous roadholding, predictable handling, supple ride, and enough other static and dynamic virtues to require a professional indexer are expected nowadays. The Alfa Romeo 164 has them all, except one.