What exactly should we look to AI, and NLP especially, for?

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK
sparckjones@cl.cam.ac.uk

Working notes,1990 AAAI Spring Symposium on Text-Based Intelligent Systems, appeared as
Text-based intelligent systems: current research in text analysis, information extraction, and retrieval,
Ed, P.S.Jacobs, Report 90CRD198, General Electric R & D Csntre, Schenectady NY, 1990, 33-37.

Text is now, and will remain, an essential embodiment of information. It is a logical error, not a practical folly, to suppose that in general we can replace a mass of text by an equivalent AI-type propositional knowledge base with accompanying inference engine for search: getting information mostly, because properly, means seeing what someone has said about something.

What, therefore, could AI (and especially NLP) techniques contribute to text-based information handling? Specifically, would they be for searching, or for indexing?

Why should we use AI?

It is important to distinguish the broad and narrow cases. In the broad case there are a lot of file texts, and imprecise requests for information, from many users (as in ordinary document retrieval). In the narrow case there is little (operational) text, and precise requests. In the broad case, with a lot of texts, efficient access presupposes reduced, selective or condensing, summary descriptions of the file items, either explicitly provided or implicitly supplied by search requests, which are much briefer than the file items. Index descriptions are therefore necessarily only rough guides to their sources.

The strong AI position is that we need AI techniques for searching. That is we need an access structure in the form of a knowledge base providing a solid characterisation of the domain motivated by the content of the underlying documents to which it points. While this knowledge base should in principle be automatically derived from the documents, the NLU system to do this is at least light years away, so in practice the base will be constructed by hand. Its function is to support request processing, by fully interpreting the request to determine its real meaning and hence to arrive, by as much inference as necessary, at the matching proposition which points outwards to the right amplifying documents.

There is a substantial problem here about the precise relationship betwen the knowledge base and the documents. But the real problem is with the tacit assumption that a request is really a question, and so can be reasoned about in a well-defined way: request interpretation is question answering, using the knowledge base to identify the precise propositional link with a document expanding on the proposition. But many requests are just not like this at all: they are vague topic references. So there is a mismatch, in the strong model, between the nature of the search apparatus and the function it is in fact required to serve.

The moderate AI position, responding to these problems, is that we indeed need a knowledge base, but a much looser one like a large mass of linked frames, for instance, with weaker,

primarily terminological inference like structured inheritance; a base of this sort could be more directly related to the source documents, even if still not automatically derived from them. But there is still the problem of the precise nature of the relationship between base and texts, and more importantly, still the tacit presumption that a request expresses a well-founded concept, is a sort of proto-question, which can be treated by processes like seeking slot filler values.

The weak AI position, adopted in an attempt to avoid these difficulties, is that we need, not an integrated knowledge base, but at least a set of text descriptions of the kind familiar in AI, like limited subject domain or case frames, each fairly straightforwardly (if not yet automatically) derived from a text. Descriptions like this are in fact essentially those which have been proposed, and are operationally utilised, by librarians and documentalists advocating complex types of indexing language or organising principles for complex subject descriptions, but which have hitherto been provided manually. Inference then applies in very limited ways to structure matching or in hierarchical term substitution, in a manner which is quite familiar from previous experience with indexing languages. It is hard to see, however, how this approach under the AI label will overcome the equally familiar difficulties for effective retrieval, given intrinsically vague requests, associated with highly and rigidly structured index descriptions.

The argument (further developed in Sparck Jones 1990) is therefore that insofar as AI might be thought to be about searching, it is not clear how relevant it is to the general case. The position may be quite different for the special case: given a well-defined context, it may be quite appropriate to adopt even a relatively strong approach. Every application has to judged on its specific merits. But it is certainly not enough just to assume that requests for documents are only regrettably inferior substitutes for the questions we would like answered directly. Yet accepting the legitimacy of requests for documents means having to accept the importance of the actual document text in its own right as an information source, even if at the same time this means we have to acknowledge the intrinsic uncertainty of retrieval stemming from the user's own necessary prior ignorance, from the necessary loss of information with reduced descriptions, and from the necessary variety of linguistic expression.

How should we use NLP?

The alternative role for AI, and more particularly NLP, in the general case is thus in indexing, as a means of providing better Using a knowledge base as an expert intermediary to help the user, interactively, to refine his request is a quite separate matter. descriptions of the weak conventional sort generically represented by 'coordinated terms', and in particular of providing individual compound terms. Given imprecise requests, indirect access, and inconsistent expression in the base document or request texts, all one can look for is good natural language clue words and good relations between them for explicit or implicit coordination, and also lots of words and relations, to obtain the natural language associative network with heavy redundancy that is the best response to fundamental uncertainty.

The essential questions to be asked about the role of NLP for indexing in the general case are therefore whether full processing (i.e. complete sentence analysis) is necessary to identify the 'good elements', and consequently if it is, how at least the necessary semantic, if not also the pragmatic, processing needed to control syntax is to be achieved without unacceptably heavy lexical work. What is the minimum semantic apparatus required to constrain syntax sufficiently to get good enough terms? Thus the question is not only whether we should

aim at linguistically-oriented rather than AI-style representations, but also whether we can obtain useful linguistically-oriented representations without having to rely, in processing to achieve this, on a domain knowledge base. The motivation for this is primarily practical, in terms both of feasibility and economy, for other than very limited retrieval environments, but there is also an issue of principle about properly fitting means to ends, and so operating linguistically throughout.

Questions about the role of NLP are also questions which have to be asked separately about document processing and request processing, especially with document processing from full texts rather than short surrogates like abstracts. The problem is not just that document processing, even assuming full NLP is feasible (which it currently isn't), is much more costly because of the bulk of material and much more effort because of the greater textual variety of documents. There are the crucial issues with documents of conflation - deriving single descriptors from variant forms, and of reduction - deriving short descriptions from long texts. The motivation for doing NLP is not primarily to identify good words, which can be done well enough statistically, but to identify good phrases, i.e. well-founded complex terms. It then becomes much more difficult to conflate non-identical terms into single descriptors either at the straightforward lexical level or to achieve some deeper common concept normalisation, or, when they overlap with others, to select some terms for the description because they are more important. It is the tradeoff between term complexity and the ease of determining what concepts are shared and are important that matters.

With request processing on the other hand, even given quite long need statements, these issues do not really arise, or at any rate arise so sharply. But if processing is confined to requests it is necessary to provide search specifications suited to In principle one could ask for good natural language inputs to controlled index language terms, but this would need more certainty about legitimate mapping and so is more problematic, quite apart from the doubtful value of controlled index languages. This is true whether the terms' internal relationships are permanently fixed or are modifiable as in Syntol. matching against unprocessed document texts. Thus whatever stages of intermediate interpretation and representation the input request or need statement goes through, the end result for searching has to be in surface language. This implies, as the document (title, abstract or full) texts are their own representations and have not been explicitly normalised in any way, that alternative surface language forms for terms have to be supplied for matching. This is perfectly feasible, as Sparck Jones and Tait 1984 shows, but has consequences in the way of vulgar search effort, and still depends on normalising principles as a base for generating alternative surface expressions. These principles may, however, be purely linguistic ones operating at the level of grammar and lexicon, and not involve any domain modelling.

It is unfortunately completely unclear, even for the second 'easier' case where language processing is limited to requests, whether the desired payoff for retrieval performance from sophisticated, relatively full NLP will be achieved. The experiments done so far have been far too limited. In particular, there is the very important question of how far indexing sophistication matters compared with the quality of the initial request statement derived from the user's information need (these are logically separable though in practice improving the formulation of the request is often mixed up with improving the search specification for that request). It is at least possible that the real payoff in retrieval comes from working on identifying the user's motivating need and on formulating an appropriate request to meet it, rather than on representing this request in a fancy way. There is certainly some experimental evidence from older-fashioned contexts (e.g. Saracevic et al 1968) that input request quality

is important, and this is a point which has clearly to be checked for new contexts.

There is still the difficulty in the general case, that even with good statements or requests, and good forms of representation, we are saddled with the approximation that comes from different perspectives on information content. This is why proper experiments are crucial. The argument is that NLP will give better results than statistical or semi-statistical methods; but if everything is bound to be approximate, what gains is it rational to expect from more sophisticated, but still relatively shallow, techniques? Performance prediction has proved a dangerous game in information retrieval: it is as essential to conduct proper evaluation experiments to justify the claims for NLP in the general case, as it is to justify those for more thorough AI in the special ones. These experiments will have, moreover, not only to supply evidence on the relative contributions of indexing method and request quality, but also on the contribution powerful interface facilities and resources can make to request quality. It is at least possible that 'old- fashioned' approaches using associative information in whatever forms it can be obtained (and hence including manually- constructed aids as well as automatically-obtained ones), to soup up crude word-based indexing may be as effective as more sophisticated NLP applied to requests or even documents. There is certainly no substitute for properly-conducted retrieval experiments, however hard these experiments are to do (cf Sparck Jones 1981), to determine performance levels and device tradeoffs.

References

Saracevic, T, et al *An inquiry into testing of information retrieval systems—, Final Report, Centre for Documentation and Communication Research, Case Western University, 1968.*

Sparck Jones, K. (ed) Information retrieval experiment*, London: Butterworths, 1981.*

Sparck Jones, K. and Tait, J.I. 'Automatic search term variant generation', Journal of Documentation *40, 1984, 50-66.*

Sparck Jones, K. 'Retrieving information or answering questions?' (British Library Annual Research Lecture 1989), London: The British Library, 1990.