Unhappy bedfellows: the relationship of AI and IR

Yorick Wilks

Part 1: AI and NLP in need of IR?

Introduction

Speaking of Artificial Intelligence in the past, one sometimes refers to "classical" or "traditional" AI, and the intended contrast with the present refers to the series of shocks that paradigm suffered from connectionism and neural nets to adaptive behaviour theories. The shock was not of the new, of course, because those theories were mostly improved versions of cybernetics which had preceded classical AI and been almost entirely obliterated by it. The classical AI period was logic or symbol-based but not entirely devoid of numbers, of course, for AI theories of vision flourished in close proximity to pattern-recognition research. Although, representational theories in computer vision sometimes achieved prominence (e.g. with Marr, 1981), nonetheless it was always, at bottom, an engineering subdiscipline with all that that entailed. But when faced with any attempt to introduce quantitative methods into classical core AI in the 70s, John McCarthy would always respond "But where do all these numbers **come from**?"

Now we know better where they come from, and nowhere have numbers been more prominent than in the field of Information Retrieval (IR), one of similar antiquity to AI, but with which it has until now rarely tangled intellectually, although on any broad definition of AI as "modelling intelligent human capacities", one might imagine that IR, like machine translation (MT), would be covered; yet neither has traditionally been seen as part of AI. On second thoughts perhaps, IR does not fall there under that definition simply because, before computers, humans were not in practice able to carry out the kinds of large-scale searches and comparisons operations on which IR rests. And even though IR often cohabits with Library Science, which grew out of card indexing in libraries, there is perhaps no true continuity between those subfields, in that IR consists of operations of indexing and retrieval that humans could not carry out in normal lifetimes.

Sparck Jones' case against AI

If any reader is beginning to wonder why I have even raised the question of the relationship of AI to IR, it is because Karen Sparck Jones (KSJ from now on), in a remarkable paper, has already done so (1999b) and argued that AI has much to learn from IR. In this paper my aim is to redress that balance a little and answer her general lines of argument. Her main target is AI researchers seen as what she calls "The Guardians of content". I shall set out her views and then contest them, arguing both in her own terms, and by analogy with the case of Machine Translation (MT) in particular, that the influence is perhaps in the other direction, and that is shown both by limitations on statistical methods that MT developments have shown in recent years, and by a curious reversal of terminology in IR that has taken place in the same period. However, the general purpose of this chapter will not be to redraw boundaries between these subfields, but will argue that subfields of NLP/AI are now increasingly hard to distinguish: not just MT, but Information Extraction (IE) and Question Answering (QA) are now beginning to form a general information processing functionality that is making many of these

arguments moot. The important questions in Sparck Jones resolve to one crucial question: what is the primitive level of language data? Her position on this is shown by the initial quotation below, after which come a set of quotations from two sources (1990, 1999b) that capture the essence of her views on the central issues:

(1) "One of these [simple, revolutionary IR] ideas is taking words as they stand" (2003)

(2) "The argument that AI is required to support the integrated information management system of the future "Is the heady vision of the individual user at his workstation in a whole range of activities calling on, and also creating, information objects of different sorts." (1990)

(3) "What might be called the intelligent library" (1990)

(4) "What therefore is needed to give effect to the vision is the internal provision of (hypertext) objects and links, and specifically in the strong form of an AI-type knowledge base and inference system" (1990)

(5) "The AI claim in its strongest form means that the knowledge base completely replaces the text base of the documents" (1990)

(6) "It is natural, therefore, if the system cannot be guaranteed to be able to use the knowledge base to answer questions on the documents of the form 'Does X do Y?' as opposed to questions of the form 'Are there documents about X doing Y?' to ask why we need a knowledge base" (1990)

(7) "The AI approach is fundamentally misconceived because it is based on the wrong general model, of IR as QA" (1990)

(8) "What reason can one have for supposing that the different [multimodal, YW] objects involved could be systematically related via a common knowledge base, and characterised in a manner *independent of ordinary language*" [YW's italics] (1990)

(9) "We should think therefore of having an access structure in the form of a network thrown over the underlying information objects" (1990)

(10) "When the key properties of document retrieval are recognised …… and the technologies that have been developed in the last forty years of IR research have important lessons for AI" (1999b)

(11) "A far more powerful AI system than any we can realistically foresee will not be able to ensure that answers it could give to questions extracted from the user's request would be appropriate" (1999b)

(12) "Classical document retrieval thus falls in the class of AI tasks that assist the human user but cannot, by definition, replace them" (1999b)

(13) This [IR] style of representation is the opposite of the classical AI type and has more in common with connectionist ones. (1999b)

(14) "The paper's case is that important tasks that can be labelled 'information management' are fundamentally inexact". (1999b)

(15) "Providing access to information could cover much more of AI than might be supposed". (1999b)

These quotations suffice to establish a complex position, and one should note in passing the prescience of quotations (2)(3)(4) and (10) in their vision of a system of information access something like the World Wide Web we now have. The quotations indicate three major claims in the papers from which they come, which I shall summarise as follows:

(A) Words are self-representing and cannot be replaced by any more primitive representation; all we, as technicians with computers, can add are sophisticated associations between them (quotations (1), (10) and (14)).

(B) Core AI-KR seeks to replace words, with their inevitable inexactness, with exact logical — or at least non-word based — representations. (quotations (5) (6) and (9))

(C) Human information needs are vague: we want relevant information, not answers to questions. In any case, AI-KR cannot answer questions. (quotations (7)(8)(11) and (12))

(D) The human reader/author relationship remains primary in the relationship, and is mediated by relevant documents. Anyway, systems based on association can do some kinds of (inexact) reasoning and could be used to retrieve relevant axioms in a KR system. (quotations (5)(13)(14) and (16)).

We should not see the issues here as simply ones of KSJ's critique (based on IR) of "core", traditional or symbolic AI, for her views connect directly to an internal interface within AI itself, one about which the subject has held an internal dialogue for many years, and in many of its subareas. The issue is that of the nature and necessity for structured symbolic representations, and their relationship to the data they claim to represent.

So, to take an example from NLP, Schank always held that Conceptual Dependency (CD) representations (1975) not only represented language strings but made the original dispensable, so that, for example, there need be no access to the source string in the process of machine translation after it had been represented by CD primitives; Charniak (1973) and I (1977) in our different ways, denied this and claimed that the surface string retained essential information not present in any representation. Schank's position here can be seen as exactly the type that KSJ is attacking, but it was not of course the only AI view.

But, more generally, the kind of AI view that KSJ had in her sights was the AI view that proclaimed the centrality and adequacy of knowledge representations, and their independence of whatever language would be used to describe what it is in the world they represent (that is the essence of her claims A and B). The key reference for the view she rejects would be McCarthy and Hayes (1969), and its extreme opposite, in machine vision at least, would be any view that has elements that could be termed Gibson (1968), one that insists on the primacy of data over any representation. The spirit of Chomsky, of course, hovers over the position, in language modelling at least, that asserts the primacy of a (correct) representation over any amount of data. Indeed, he produced a range of ingenious arguments as to why no amount of data could possibly produce the representations the brain has for language structure (1965), and those arguments continued to echo through the dispute, for example, between Fodor and Pollack (1990) as to whether or not nested representations could be derived by any form machine learning

from language data: Pollack claimed his connectionist RAAM system could do exactly that, and Fodor denied it.

Again, and now somewhat further from core AI, one can see the issue in Schvaneveldt's Pathfinder networks (1990) which he showed, in psychological experiments, could represent the expertise of fighter pilots in associationist networks of terms, a form very close to the data from which it was derived. This work was a direct challenge to the contemporary expert-systems movement for representing such expertise by means of high-level rules.

Some countervailing considerations from AI

It should be clear from the last paragraphs that KSJ is not targeting all of AI, which might well be taken to include IR on a broad definition, but a core of AI, basically the strong representationalist tradition, one usually (but not always, as in the case of Schank above) associated with the use of first order predicate calculus. And when one writes of a broad definition, it could only be one that does not restrict AI to the modelling of basic human functionalities, the notion behind Papert's original observation that AI could not and should not model superhuman faculties, ones that no person could have. In some sense, of course, classic IR is superhuman: there was no pre-existing human skill, as there was with seeing, talking or even chess playing that corresponded to the search through millions of words of text on the basis of indices. But if one took the view, by contrast, that theologians, lawyers and, later, literary scholars were able, albeit slowly, to search vast libraries of sources for relevant material, then on that view IR is just the optimisation of a human skill and not a superhuman activity. If one takes that view, IR is a proper part of AI, as traditionally conceived.

However, that being said, it may be too much a claim (D above) in the opposite direction to suggest, as KSJ does in a remark at the end of one of the papers cited, that core AI may need IR to search among the axioms of a formalised theory (1999b) in order to locate relevant axioms to compose a proof. It is certain that resolution, or any related proof program, draws in potential axioms based on the appearance of identical predicates in them (i.e. to those in the theorem to be proved). But it would be absurd to see that as a technique borrowed from or in any way indebted to IR; it is simply the obvious and only way to select those axioms that might plausibly take part in proofs.

A key claim of KSJ's (in (A) and especially (B) above) is the issue one might call *primitives*, where one can take that to be either the predicates of a logical representation, as in McCarthy and Hayes and most AI reasoning work, or the more linguistic primitives, present in Schank's CD work and my own under the name *preference semantics* (Wilks and Fass, 1992, Wilks et al., 1996). Her argument is that words remain their own best interpretation, and cannot be replaced by some other artificial coding that adequately represents their meaning. KSJ's relationship to this tradition is complex: her own thesis (Sparck Jones, 1966 and see Tait and Wilks, this volume) although containing what now are seen as IR clustering algorithms applied to a thesaurus, was intended, in her own words, to be a search for semantic primitives for MT. Moreover, she contributed to the definition and development of Cambridge Language Research Unit's own semantic interlingua NUDE (for "naked ideas"). That tradition has been retained in AI and computational linguistics, both as a basis for coding lexical systems (e.g. the work of Pustejovsky, 1995) and as another form of information to be established on an empirical

basis from corpora and can be seen in early work on the derivation of preferences from corpora by Resnik (1996), Grishman, (Grishman and Sterling, 1992), Lehnert (Riloff and Lehnert, 1993) and others. Work of this type certainly involves the exploitation of semantic redundancy, both qualitatively, in the early preference work cited above, and quantitatively, in the recent tradition of work on systematic Word Sense Disambiguation which makes use of statistical methods exploiting the redundancy already coded in thesauri and dictionaries. Unless KSJ really intends to claim that any method of language analysis exploiting statistics and redundancy (like those just cited) is really IR, then there is little basis to her claim that AI has a lot to learn from IR in this area, since it has its own traditions by now of statistical methodology and evaluation and, as I shall shown below, these came into AI/NLP from speech research pioneered by Jelinek, and indigenous work on machine learning, and not at all from IR.

Let us now turn to another of KSJ's major claims, (C above) that question-answering (QA) is not a real task meeting a real human need, but that the real task is the location of relevant documents, which is IR's classic function. First, one must be clear that there has never been any suggestion in mainstream AI that its techniques could perform the core IR task. To find relevant documents, as opposed to their content, one would have to invent IR, had it not existed; there simply is no choice. Information Extraction (IE), on the other hand, (Gaizauskas and Wilks, 1997) is a relatively recent content searching technique, usually with a representational non-statistical component, designed to access factual content directly, and that process usually assumes a prior application of IR to find relevant material to search. The application of an IR phase prior to IE in a sense confirms KSJ's "primacy of relevance", but also confirms the independence and viability of QA, which is nowadays seen as an extension of IE. IE, by seeking facts of specific forms, is always implicitly asking a question (i.e. What facts are there matching the following general form?).

However, recently Gaizauskas (2004) has questioned this conventional temporal primacy of IR in an IE application, and has done so by pointing out that the real answers to IE/QA questions are frequently to be found very far down the (relevance based) percentiles of returns from this prior IR phase. The reason for this is that if one asks, say, "What colour is the sky?" then, in the IR phase, the term "colour/color" is a very poor index term for relevant documents likely to contain the answer. In other words, "relevance" in the IR sense (and unboosted by augmentation with actual colour names in this case) is actually a poor guide to where answers to this question are to be found, and Gaizauskas uses this point to question the conventional relationship of IR and IE/QA.

One could, at this point, perhaps reverse KSJ's jibe at AI as the self-appointed "Guardians of Content" and suggest that IR may not be as much the "Guardian of relevance" as she assumes. But whatever is the case there, it seems pretty clear that wanting answers to questions is sometimes a real human need, even outside the world of TV quiz shows. The website Ask Jeeves seemed to meet some real need, even if it was not always successful, and QA has been seen as a traditional AI task, back to the classic book by Lehnert (1977). KSJ is, of course, correct that those traditional methods were not wholly successful and did not, as much early NLP did not, lead to regimes of evaluation and comparison. But that in no way reflects on the need for QA as a task.

In fact, of course, QA has now been revived as an evaluable technique (see below), as part of the general revival of empirical linguistics, and has been, as we noted already, a

development of existing IE techniques, combined in some implementations with more traditional abductive reasoning (Moldovan, 2001). The fact of its being an evaluable technique should have made it very hard for KSJ to dismiss QA as a task in the way she does, since she has gone so far elsewhere in identifying real NLP with evaluable techniques (Galliers and Sparck Jones, 1996).

Over a twenty year period, CQA has moved from a wholly-knowledge based technique (as in Lehnert's work) to where it now is, as fusion of statistical and knowledge-based techniques. Most, if not all, parts of NLP have made the same transition over that period, starting with apparently straightforward tasks like part-of-speech tagging (e.g. Garside, 1987) and rising up to semantic and conceptual areas like word-sense disambiguation (e.g. Stevenson and Wilks, 1999) and dialogue management (Churcher et al., 1997) in addition to QA. In the next section we shall return to the origin of this empirical wave in NLP and re-examine its sources, then claim that new and interesting evidence can be found there for the current relationship of AI and IR. In her paper, KSJ acknowledges the recent empirical movement in NLP and its closeness in many ways to IR techniques, but she does not actually claim the movement as an influence from IR. I shall argue in the next section that, on the contrary, the influence on NLP that brought in the empirical revolution was principally from speech research, and in part from traditional statistical AI (i.e. machine learning) but in no way from IR. On the contrary, the influences detectable are all *on IR from outside*.

Jelinek's revolution in Machine Translation and its relevance

A piece of recent NLP history that may not be familiar to AI researchers is, I believe, highly relevant here. Jelinek, Brown and others at IBM New York began to implement around 1988 a plan of research to import the statistical techniques that had been successful in Automatic Speech Processing (ASR) into NLP and into machine translation (MT) in particular. DARPA supported Jelinek's system CANDIDE (Brown and Cocke, 1989, Brown et al., 1990) at the same time as rival symbolic systems (such as PANGLOSS (Nirenburg et al., 1994) using more traditional methods.

The originality of CANDIDE was to employ none of the normal translation resources within an MT system (e.g. grammars, lexicons etc.) but only statistical functions trained on a very large bilingual corpus: 200 million words of the French-English parallel text from Hansard, the Canadian parliamentary proceedings. CANDIDE made use of a battery of statistical techniques that had loose relations to those used in ASR: alignment of the parallel text sentences, then of words between aligned French and English sentences, and n-gram models (or language models as they would now be called) of the two language separately, one of which was used to smooth the output. Perhaps the most remarkable achievement was that given 12 French output words so found (output sentences could not be longer than that) the generation algorithm could determine the unique best order (out of billions) for an output translation with a high degree of success. The CANDIDE team did not describe their work this way, but rather as machine learning from a corpus that, given what they called an "equation of MT" produced the most likely source sentence for any putative output sentence.

The CANDIDE results were at roughly the 50% level, of sentences translated correctly or acceptably in a test set held back from training. Given that the team had no access to what one might call "knowledge of French", this was a remarkable achievement and far higher

than most MT experts would have predicted, although CANDIDE never actually beat SYSTRAN, the standard and traditional symbolic MT system that is the world's most used system. At this point (about 1990) there was a very lively debate between what was then called the rationalist and empiricist approaches to MT, and Jelinek began a new program of trying to remedy what he saw as the main fault of his system by what would now be called a "hybrid" approach, one that was never fully developed because the IBM team dispersed.

The problem Jelinek saw is best called "data sparseness": his system's methods could not improve even applied to larger corpora of any reasonable size because language events are no rare. Word trigrams tend to be 85% novel in corpora of any conceivable size, an extraordinary figure. Jelinek therefore began a hybrid program to overcome this, which was to try to develop from scratch the standard NLP resources used in MT, such as grammars and lexicons, in the hope of using them to generalise across word or structure classes, so as to combat data sparseness. So, if the system knew elephants and dogs were in a class, then it could predict a trigram [X Y ELEPHANT] from having seen the trigram [X Y DOG] or vice versa.

It was this second, unfulfilled, program of Jelinek that, more than anything else, began the empiricist wave in NLP that still continues, even though the statistical work on learning part-of-speech tags actually began earlier at Lancaster under Leech (Garside, 1987). IBM bought the rights to this work and Jelinek then moved forward from alignment algorithms to grammar learning, and the rest is the historical movement we are still part of.

But it is vital to note consequences of this: first, that the influences brought to bear to create modern empirical, data-driven, NLP came from the ASR experience and machine learning algorithms, a traditional part of AI by then. They certainly did not come from IR, as KSJ might have expected given what she wrote. Moreover, and this has only recently been noticed, the research metaphors have now reversed, and techniques derived from Jelinek's work are now being introduced into IR under names like "MT approaches to IR" (Berger and Laferty, 2001, and see below) which is precisely a reversal of the direction of influence that KSJ argued for.

We shall mention some of this work in the next section, but we must draw a second moral here from Jelinek's experience with CANDIDE and one that bears directly on KSJ's claim that words are their own best representations (Claim A above). The empiricist program of recreating lexicons and grammars from corpora, begun by Jelinek and the topic of much NLP in the last 15 years, was started precisely because working with self-representations of words (e.g. n-grams) was inadequate because of their rarity in any possible data: 80% of word trigrams are novel, as we noted earlier under the term "data sparseness". Higher-level representations are designed to ameliorate this effect, and that remains the case whether those representations are a priori (like Wordnet, LDOCE or Roget's Thesaurus) or themselves derived from corpora.

KSJ could reply here that she did not intend to target such work in her critique of AI, but only core AI (logic or semantics based) that eliminates words as part of a representation, rather than adds higher level representation to the words. There can be no doubt that even very low-level representations, however obtained, when added to words can produce results that would be hard to imagine without them. A striking case is the use of part-of-

speech tags (like PROPERNOUN) where, given a word sense resource structured in the way Longmans LDOCE is, (Stevenson and Wilks, 1999) were able to show that those part of speech tags alone can resolve large-scale word sense ambiguity (called homographs in LDOCE) at the 92% level. Given such a simple tagging, almost all word sense ambiguity is trivially resolved against that particular structured resource, a result that could not conceivably be obtained without those low-level additional representations, which are not merely the words themselves, as KSJ expects.

Recent developments in IR

In this section, we draw attention to some recent developments in IR that suggest that KSJs characterisation of the relationship of IR to AI may not be altogether correct and may in some ways be the reverse of what is the case.

That reverse claim may also seem somewhat hyperbolic, in response to KSJ's original paper, and in truth there may be some more general movement at work in this whole area, one more general than either the influence of AI on IR or its opposite, namely that traditional functionalities in information processing are now harder to distinguish. This degree of interpenetration of techniques is such that it may be just as plausible (as claiming directional influence, as above) to say that MT, QA, IE, IR as well as summarisation and, perhaps a range of technologies associated with ontologies, lexicons, inference, the Semantic Web and aspects of Knowledge Management, are all becoming conflated in a science of information access. Without going into much detail, where might one look for immediate anecdotal evidence for that view?

Salton (1972) initiated CLIR (Cross-language Information Retrieval) using a thesaurus and a bilingual dictionary between languages, and more recent forms of the technique have used Machine-Readable Bilingual Dictionaries to bridge the language gap (Ballasteros and Croft, 1998), and Eurowordnet, a major NLP tool (Vossen, 1998), was designed explicitly for CLIR. CLIR is a task rather like MT but recall is more important and it is still useful at low rates of precision, which MT is not because people tend not to accept translations with alternatives on a large scale like "They decided to have {PITCH, TAR, FISH, FISHFOOD} for dinner".

(Gaizauskas and Wilks, 1997) describe a system of multilingual IE based on treating the templates themselves as a form of interlingua between the languages, and this is clearly a limited form of MT. (Gollins and Sanderson, 2001) have described a form of CLIR that brings back the old MT notion of a "pivot language" to bridge between one language and another, and where pivots can be chained in a parallel or sequential manner. Latvian-English and Latvian-Russian CLIR could probably reach any EU language from Latvian via multiple CLIR pivot retrievals (of sequential CLIR based on Russian-X or English-X). This IR usage differs from MT use, where a pivot was an interlingua, not a language and was used once, never iteratively. (Oh et al., 2000) report using a Japanese-Korean MT system to determine terminology in unknown language. (Gachot et al., 1998) report using an established, possibly the most established, MT system SYSTRAN as a basis for CLIR. (Wilks et al., 1996) report using Machine Readable Bilingual Dictionaries to construct ontological hierarchies (for IR or IE) in one language from an existing hierarchy in another language, using redundancy to cancel noise between the languages in a manner rather like Gollins and Sanderson.

All these developments indicate some forms of influence and interaction between traditionally separate techniques, but are more suggestive of a loss of borderlines between traditional functionalities. More recently, however, usage has grown in IR of referring to any technique related to Jelinek's IBM work as being a use of an "MT algorithm": this usage extends from the use of n-gram models under the name of "language models" (Ponte and Croft, 1998, Croft and Laferty, 2000), a usage that comes from speech research, to any use in IR of a technique like sentence alignment that was pioneered by the IBM MT work. An extended metaphor is at work here, one where IR is described as MT since it involves the retrieval of one string by means of another (Berger and Laferty, 1999). IR classically meant the retrieval of documents by queries, but the string-to-string version notion has now been extended by IR-researchers who have moved on to QA work where they describe an answer as a "translation" of its question (Berger, 2000). On this view questions and answers are like two "languages". In practice, this approach meant taking FAQ questions and their corresponding answers as training pairs.

The theoretical underpinning of all these researches is the matching of language models i.e. what is the most likely query given this answer, a question posed by analogy with Jelinek's "basic function of MT" that yielded the most probable source text given the translation. This sometimes sounds improbable, but is actually the same way up as theoretical science, namely that of proving the data from the theory, even though actually inferring the theory from the data, by abduction.

Preliminary conclusion

What point have we reached so far in our discussion? We have not detected influence of IR on AI/NLP, as KSJ predicted, but rather an intermingling of methodologies and the dissolution of borderlines between long-treasured application areas, like MT, IR, IE, QA etc. One can also discern a reverse move of MT/AI metaphors into IR itself, which the opposite direction of influence to that advocated by KSJ in her paper. Moreover, the statistical methodology of Jelinek's CANDIDE did revolutionise NLP, but that was an influence on NLP from speech research and its undoubted successes, not IR. The pure statistical methodology of CANDIDE was not in the end successful in its own terms, because it always failed to beat symbolic systems like SYSTRAN in open competition. What CANDIDE did, though, was to suggest a methodology by which data sparseness might be reduced by the recapitulation of symbolic entities (e.g. grammars, lexicons, semantic annotations etc.) in statistical, or rather machine learning, terms, a story not yet at an end. But that methodology did not come from IR, which had always tended to reject the need for such symbolic structures, however obtained e.g. in the on going, but basically negative, debate on whether or not Wordnet or any similar thesaurus, can improve IR.

In the second part of this paper, we shall return to, and focus on, this key hard issue, that of whether NLP, taken broadly to include both resources and techniques, can improve the performance of IR systems, again broadly construed. KSJ has taken a number of positions on this issue, from the agnostic to the mildly sceptical. This is a complex issue and one quite independent of her theme examined in this first part of the paper, namely that IR methods should play a larger role than they do in NLP and AI.

One interesting question to ask at the end of this initial discussion is: if GOFAI (good old fashioned AI) and its logic did not produce the results in NLP that had been hoped for,

and I agree with KSJ that it did not in its original form, then where did GOFAI go off to? The answer to which is that part of it has returned, replete with new claims about the nature of natural language, in the form of the Semantic Web (SW) movement (Berners-Lee et al., 2001). This is not the place for any full description of that development and its aims, but it incorporates aspects of the formal ontologies movement, which now can be taken to mean virtually all the content of classical AI Knowledge Representation, rather than any system of merely hierarchical relations, which is what the word "ontology" used to convey. More particularly, the Semantic Web movement envisages the (automatic) annotation of the texts of the World Wide Web with a hierarchy of annotations up to the semantic and logical, which is a claim virtually indistinguishable from the old GOFAI assumption that the true structure of language was its underlying logical form. Fortunately, SW comes in more than one form, some of which envisage statistical techniques, of the sort already discussed, as he basis of the assignment of semantic and logical annotations, but the underlying similarity to GOFAI is clear. One could also say that semantic annotation, so conceived, is the inverse of Information Extraction, done not at analysis time but, ultimately, at generation time without the writer being aware of this (since one cannot write and annotate at the same time). SW is as, it were, producer, rather than consumer, IE.

Two other aspects of the SW link it back directly to the goals of GOFAI: one is the rediscovery of a formal semantics to "justify" the SW. This is now taken to be expressed in terms of URIs (basic objects on the web), which are usually illustrated by means entities like lists of zip codes, with much indication of how such a notion will generalize to produce objects into which all web expressions can "bottom out". This concern, for non-linguistic objects as the ultimate reality, is of course classic GOFAI. Secondly, one can see this in KSJs terms with which we began this paper, namely her emphasis on the "primacy of words" and words standing for themselves, as it were: this aspect of the SW is exactly what KSJ meant by her "AI doesn't work in a world without semantic objects" (1990). In the SW, with its notion of universal annotation of web texts into both semantic primitives of undefined status and the ultimate URIs, one can see the new form of opposition to that view. The WWW was basically words—if we ignore pictures, tables and diagrams for the moment—but the vision of the SW is that of the words backed up by, or even replaced by, their underlying meanings expressed in some other way, including annotations and URIs. Indeed, the current SW formalism for underlying content, usually called RDF triples, is one very familiar indeed to those with memories of the history of AI: namely, John-LOVES-Mary, a form reminiscent at once of semantic nets (of arcs and nodes Woods et al., 1974), semantic templates (Wilks, 1964), or, after a movement of LOVES to the left, standard first order predicate logic. Only the last of these was full GOFAI, but all sought to escape the notion of words standing simply for themselves.

KSJs position here, an opposition to any kind of symbolic primitives standing behind words, has been a long held one, although at earlier periods (e.g. that of her thesis, see Tait and Wilks, this volume) she found such notions more congenial. One can also see the SW revival as again taking head on David Lewis' classic critique of what he called "markerese" (1972), an attack he aimed at the semantic markers of Fodor and Katz but which can be transferred to any project like the SW that makes use of "special languages", separate from natural languages, but not clearly grounded in any formal

semantics, which was what Lewis considered the only plausible grounding, though KSJ differs on this, of course.

It is not obvious, that the SW needs any of the systematic justifications on offer, from formal logic to URIs, to annotations to URIs: it may all urn out to be a practical matter of this huge structure providing a range of practical benefits to people wanting information. Critics like Ted Nelson (1997) still claim that the WWW is ill-founded and cannot benefit users, but all the practical evidence shows the reverse. Semantic annotation efforts are widespread, even outside the SW, and one might even cite recent work by Jelinek (Chelba and Jelinek, 1998), who is investigating systematic annotation to reduce the data sparseness that limited the effectiveness of his original statistical efforts at MT.

KSJ's position under discussion in this first part of the paper has been that words are just themselves, and we should not become confused (in seeking contentful information with the aid of computers) by notions like semantic objects, no matter what form they come in, formal, capitalized primitives or whatever. However, this does draw a firm line where there is not one: I have argued in many places—most recently against Sergei Nirenburg in (Nirenburg and Wilks, 2001)----that the symbols used in knowledge representations, ontologies etc., throughout the history of AI, have always appeared to be English words, often capitalized, and indeed are, in spite of the protests of their users, no more or less than English words. If anything else, they are slightly privileged English words, in that they are not drawn randomly from the whole vocabulary of the language. Knowledge representations, annotations etc. work as well as they do—and they do, and the history of machine translation using such notions as interlinguas is the clearest proof of that (1990)-----because it is possible to treat some words as more primitive than others and to obtain some benefits of data compression thereby, but these privileged entities do not thereby cease to be words, and are thus at risk, like all words of ambiguity and extension of sense. In (Nirenburg and Wilks, 2001) that was my key point of disagreement with my co-author Nirenburg who holds the same position as Carnap who began this line of constructivism in 1936 with Der Logische Aufbau der Welt, namely that words can have their meanings in formal systems controlled by fiat. I believe this is profoundly untrue and one of the major fissures below the structure of formal AI.

This observation bears on KSJs view of words in the following way: her position could be characterised as a democracy of words, all words are words from the point of view of their information status, however else they may differ. To this I would oppose the view above, that there is a natural aristocracy of words, those that are natural candidates for primitives in virtually all annotation systems e.g. ANIMATE, HUMAN, EXIST and CAUSE. The position of this chapter is not as far from KSJ's as appeared at the outset and we both remain opposed to those in AI who believe that things-like-words-in-formal-codings are no longer words.

KSJs position in the sources quoted remains basically pessimistic about any fully automated information process; this is seen most clearly in her belief that humans cannot be removed from the information process. There is a striking similarity between that and her former colleague Martin Kay's famous paper on human-aided machine translation and its inevitability, given the poor prospects for pure MT. I believe his pessimism was premature and that history has shown that simple MT has a clear and useful role if users adapt their expectations to what is available, and I hope the same will prove true in the topics covered so far in this paper.

Part 2: IR in need of AI and NLP?

In this section we turn the hard question, ignored in part 1 though long debated, as to whether or not the representational techniques, familiar in AI and NLP as both resources and the objects of algorithms, can improve the performance of classical statistical IR. The aim is go beyond the minimal satisfaction given by Croft's immortal phrase about IR "For any technique there is a collection where it will help".

Artificial Intelligence (AI), or at least non-Connectionist non-statistical AI, remains wedded to representations, their computational tractability and their explanatory power; and that normally means the representation of propositions in some more or less logical form. Classical Information Retrieval (IR), on the other hand, often characterised as a "bag of words" approach to text, consists of methods for locating document content independent of any particular explicit structure in the data. Mainstream IR is, if not dogmatically anti-representational (as are some statistical and neural net-related areas of AI and language processing), is at least not committed to any notion of representation beyond what is given by a set of index terms, or strings of index terms along with numbers themselves computed from text that may specify clusters, vectors or other derived structures.

This intellectual divide over representations and their function goes back at least to the Chomsky versus Skinner debate, which was always presented by Chomsky in terms of representationalists versus barbarians, but was in fact about simple and numerically-based structures versus slightly more complex ones.

Bizarre changes of allegiance took place during later struggles over the same issue, as when IBM created the machine translation (MT) system (CANDIDE, see Brown and Cocke, 1989), discussed earlier, based purely on text statistics and without any linguistic representations, which caused those on the representational side of the divide to cheer for the old-fashioned symbolic MT system SYSTRAN in its DARPA sponsored contests with CANDIDE, although those same researchers had spent whole careers dismissing the primitive representations that SYSTRAN contained. Nonetheless it was symbolic and representational and therefore on their side in this more fundamental debate! In those contests SYSTRAN always prevailed over CANDIDE for texts over which neither system had been trained, which may or may not have indirect implications for the issues under discussion here.

Winograd (1971) is often credited in AI with the first natural language processing system (NLP) firmly grounded in representations of world knowledge yet, after his thesis, he effectively abandoned that assumption and embraced a form of Maturana's autopoesis doctrine (see Winograd and Flores, 1986), a biologically-based anti-representationalist position that holds, roughly, that evolved creatures like us are unlikely to contain or manipulate representations. On such a view the Genetic Code is misnamed, which is a position with links back to the philosophy of Heidegger (whose philosophy Winograd began to teach at that period at Stanford in his NLP classes) as well as Wittgenstein's view that messages, representations and codes necessarily require intentionality, which is to say a sender, and the Genetic Code cannot have a sender. This insight spawned the speech act movement in linguistics and NLP, and also remains the basis of Searle's position that there cannot therefore be AI at all, as computers cannot have intentionality.

The same insight is behind Dennett's more recent view that evolution necessarily undermines AI, as it does so much else.

The debate within AI itself over representations, as within its philosophical and linguistic outstations, is complex and unresolved. The Connectionist/neural net movement of the 1980's brought some clarification of the issue into AI, partly because it came in both representationalist (localist) and non-representationalist (distributed) forms, which divided on precisely this issue. Matters were sometimes settled not by argument or experiment but by declarations of faith, as when Charniak said that whatever the successes of Connectionism, he didn't like it because it didn't give him any perspicuous representations with which to understand the phenomena of which AI treats.

Within psychology, or rather computational psychology, there have been a number of recent assaults on the symbolic reasoning paradigm of AI-influenced Cognitive Science, including areas such as rule-driven expertise which was an area where AI, in the form of Expert Systems, was thought to have had some practical success. In an interesting revival of classic associationist methods, Schvaneveldt developed an associative network methodology for the representation of expertise (1990), producing a network whose content is extracted directly from subjects' responses, and whose predictive power in classic expert systems environments is therefore a direct challenge to propositional-AI notions of human expertise and reasoning.

Within the main AI symbolic tradition, as I am defining it, it was simply inconceivable that a complex cognitive task, like controlling a fighter plane in real time, on the basis of input from a range of discrete sources of information from instruments, could be other than a matter for constraints and rules over coded expertise. There was no place there for a purely associative component based on numerical strengths of association or (importantly for Pathfinder networks) on an overall statistical measure of clustering that establishes the Pathfinder network from the subject-derived data in the first place.

The Pathfinder example is highly relevant here, not only for its direct challenge to a core area of classic AI, where it felt safe, as it were, but because the clustering behind Pathfinder networks was in fact very close, formally, to the clump theory behind the early IR work such as Sparck Jones (1966/1986) and others. Schvaneveldt and his associates later applied Pathfinder networks to commercial IR after applying them to lexical resources like LDOCE. There is thus a direct algorithmic link here between the associative methodology in IR and its application in an area that challenged AI directly in a core area. It is Schvaneveldt's results on knowledge elicitation by associative methods from groups like pilots, and the practical difference such structures make in training, that constitute their threat to propositionality here.

This is no unique example, of course: even in more classical AI one thinks of Pearl's long-held advocacy (1985) of weighted networks to model beliefs, which captured (as did fuzzy logic and assorted forms of Connectionism since) the universal intuition that beliefs have strengths, and that these seem continuous in nature and not merely one of a set of discrete strengths, and that it is very difficult indeed to combine any system expressing that intuition with central AI notions of logic-based machine reasoning.

Information Extraction (IE) as a task and the adaptivity problem.

In this chapter, I am taking IE as a paradigm of an information processing technology separate from IR; formally separate, at least, in that one returns documents or document parts, and the other linguistic or data-base structures. IE is a technique which, although still dependent on superficial linguistic methods of text analysis, is beginning to incorporate more of the inventory of AI techniques, particularly knowledge representation and reasoning, as well as, at the same time, finding that its rule-driven successes can be matched by machine learning techniques using only statistical methods (see below on named entities).

IE is an automatic method for locating facts for users in electronic documents (e.g. newspaper articles, news feeds, web pages, transcripts of broadcasts, etc.) and storing them in a data base for processing with techniques like data mining, or with off-the-shelf products like spreadsheets, summarisers and report generators. The historic application scenario for Information Extraction is a company that wants, say, the extraction of all ship sinkings, from public news wires in any language world-wide, and put into a single data base showing ship name, tonnage, date and place of loss etc. Lloyds of London had performed this particular task with human readers of the world's newspapers for a hundred years.

The key notion in IE is that of a “template”: a linguistic pattern, usually a set of attribute-value pairs, with the values being text strings. The templates are normally created manually by experts to capture the structure of the facts sought in a given domain, which IE systems then apply to text corpora with the aid of extraction rules that seek fillers in the corpus, given a set of syntactic, semantic and pragmatic constraints.

IE has already reached the level of success at which Information Retrieval and Machine Translation (on differing measures, of course) have proved commercially viable. By general agreement, the main barrier to wider use and commercialisation of IE is the relative inflexibility of its basic template concept: classic IE relies on the user having an already developed set of templates, as was the case with intelligence analysts in US Defense agencies from whose support the technology was largely developed. The intellectual and practical issue now is how to develop templates, their filler subparts (such as named entities or NEs), the rules for filling them, and associated knowledge structures, as rapidly as possible for new domains and genres.

IE as a modern language processing technology was developed largely in the US, but with strong development centres elsewhere (Cowie et al., 1993), (Grishman, 1997), (Hobbs, 1993), (Gaizauskas and Wilks, 1997). Over 25 systems, world wide, have participated in the DARPA-sponsored MUC and TIPSTER IE competitions, most of which have the same generic structure (as shown by Hobbs, 1993). Previously unreliable tasks of identifying template fillers such as names, dates, organizations, countries, and currencies automatically — often referred to as TE, or Template Element, tasks — have become extremely accurate (over 95% accuracy for the best systems). These core TE tasks were initially carried out with very large numbers of hand-crafted linguistic rules.

Adaptivity in the MUC development context has meant beating the one-month period in which competing centres adapted their system to new training data sets provided by DARPA; this period therefore provides a benchmark for human-only adaptivity of IE systems. Automating this phase for new domains and genres now constitutes the central

problem for the extension and acceptability of IE in the commercial world beyond the needs of the military sponsors who created it.

The problem is of interest in the context of this paper, to do with the relationship of AI and IR techniques, because attempts to reduce the problem have almost all taken the form of introducing another area of AI techniques into IE, namely that of machine learning, and which is statistical in nature, like IR but unlike core AI.

Previous work on ML and adaptive methods for IE

The application of Machine Learning methods to aid the IE task goes back to work on the learning of verb preferences in the Eighties by Grishman and Sterling (1992) and Lehnert (et al., 1992), as well as early work at MITRE on learning to find named expressions (NEs) (Bikel et al., 1997). Many of the developments since then have been a series of extensions to the work of Lehnert and Riloff on Autoslog (Riloff and Lehnert, 1993), the automatic induction of a lexicon for IE.

This tradition of work goes back to an AI notion that might be described as lexical tuning, that of adapting a lexicon automatically to new senses in texts, a notion discussed in (Wilks and Catizone, 1999) and going back to work like Wilks (1979) and Granger (1977) on detecting new preferences of words in texts and interpreting novel lexical items from context and stored knowledge. These notions are important, not only for IE in general but, in particular, as it adapts to traditional AI tasks like QA.

The Autoslog lexicon development work is also described as a method of learning extraction rules from <document, filled template> pairs, that is to say the rules (and associated type constraints) that assign the fillers to template slots from text. These rules are then sufficient to fill further templates from new documents. No conventional learning algorithm was used by Riloff and Lehnert but, since then, Soderland has extended this work by using a form of Muggleton's ILP (Inductive Logic Programming) system for the task, and Cardie (1997) has sought to extend it to areas like learning the determination of coreference links.

Grishman at NYU (Agichtein et al., 1998) and Morgan (Morgan et al., 1995) at Durham have done pioneering work using user interaction and definition to define usable templates, and Riloff (Riloff and Shoen, 1995) has attempted to use some version of user-feedback methods of Information Retrieval, including user-judgements of negative and positive <document, filled template> pairings.

Supervised template learning

Brill-style transformation-based learning methods are one of the few ML methods in NLP to have been applied above and beyond the part-of-speech tagging origins of virtually all ML in NLP. Brill's original application triggered only on POS tags; later (Brill, 1994) he added the possibility of lexical triggers. Since then the method has been extended successfully to e.g. speech act determination (Carberry, Samuel and Vijay-Shanker, 1998) and a Brill-style template learning application was designed by Vilain (1993).

A fast implementation based on the compilation of Brill-style rules to deterministic automata was developed at Mitsubishi labs (Roche and Schabes, 1995, Cunningham, 1999). The quality of the transformation rules learned depends on factors such as:

(1) The accuracy and quantity of the training data;

(2) The types of pattern available in the transformation rules;

(3) The feature set available used in the pattern side of the transformation rules.

The accepted wisdom of the ML community is that it is very hard to predict which learning algorithm will produce optimal performance, so it is advisable to experiment with a range of algorithms running on real data. There have as yet been no systematic comparisons between these initial efforts and other conventional machine learning algorithms applied to learning extraction rules for IE data structures (e.g. example-based systems such as TiMBL (Daelemans et al., 1998) and ILP (Muggleton, 1994). A quite separate approach has been that of Ciravegna (Ciravgna and Wilks, 1993) which has concentrated on the development of interfaces (ARMADILLO and MELITA) at which a user can indicate what taggings and fact structures he wishes to learn, and then have the underlying (but unseen) system itself take over the tagging and structuring from the user, who only withdraws from the interface when the success rate has reached an acceptable level.

Unsupervised template learning

We should also remember the possibility of unsupervised notion of template learning: in a Sheffield PhD thesis Collier (Collier, 1998) developed such a notion, one that can be thought of as yet another application of the early technique of Luhn (1957) to locate statistically significant words in a corpus and then use those to locate the sentences in which they occur as key sentences. This has been the basis of a range of summarisation algorithms and Collier proposed a form of it as a basis for unsupervised template induction, namely that those sentences, with corpus-significant verbs, would also contain sentences corresponding to templates, whether or not yet known as such to the user. Collier cannot be considered to have proved that such learning is effective, only that some prototype results can be obtained. This method is related, again via Luhn's original idea, to recent methods of text summarisation (e.g. the British Telecom web summariser entered in DARPA summarisation competitions) which are based on locating and linking text sentences containing the most significant words in a text, a very different notion of summarisation from that discussed below, which is derived from a template rather than giving rise to it.

Linguistic considerations in IR

Let us now quickly review the standard questions, some unsettled after 30 years, in the debate about the relevance of symbolic or linguistic (or AI taken broadly) considerations in the task of information retrieval.

Note too that, even in the form in which we shall discuss it, the issue is not one between high-level AI and linguistic techniques on the one hand, and IR statistical methods on the other. As the last section showed, the linguistic techniques normally used in areas like IE have in general been low-level, surface orientated, pattern-matching techniques, as opposed to more traditional concerns of AI and linguistics with logical and semantic representations. So much has this been the case that linguists have in general taken no notice at all of IE, deeming it a set of heuristics almost beneath notice, and contrary to all long held principles about the necessity for general rules of wide coverage. Most IE has

been a minute study of special cases and rules for particular words of a language, such as those involved in template elements (countries, dates, company names etc.).

Again, since IE has also made extensive use of statistical methods, directly and as applications of ML techniques, one cannot simply contrast statistical (in IR) with linguistic methods used in IE as Sparck Jones (1999a) does when discussing IR. That said, one should note that some IE systems that have performed well in MUC/TIPSTER – – Sheffield's old LaSIE system would be an example (Gaizauskas and Wilks, 1997) — did also make use of complex domain ontologies, and general rule-based parsers. Yet, in the data-driven computational linguistics movement in vogue at the moment, one much wider than IE proper, there is a goal of seeing how far complex and "intensional" phenomena of semantics and pragmatics (e.g. dialogue pragmatics as initiated in (Carberry et al., 1998)) can be treated by statistical methods.

A key high-level module within IE has been co-reference, a topic that linguists might doubt could ever fully succumb to purely data-driven methods since the data is so sparse and the need for inference methods seems so clear. One can cite classic examples like:

> {A Spanish priest} was charged here today with attempting to murder the Pope. {Juan Fernandez Krohn}, aged 32, was arrested after {a man armed with a bayonet} approached the Pope while he was saying prayers at Fatima on Wednesday night. According to the police, {Fernandez} told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope "looked furious" on hearing {the priest's} criticism of his handling of the church's affairs. If found guilty, {the Spaniard} faces a prison sentence of 15-20 years. (The London Times 15 May 1982, example due to Sergei Nirenburg)

This passage contains six different phrases {enclosed in curly brackets} referring to the same person, as any reader can see, but whose identity seems a priori to require much knowledge and inference about the ways in which individuals can be described.

There are three standard techniques in terms of which this infusion (of possible NLP techniques into IR) have been discussed, and I will mention them and then add a fourth.

i. Prior WSD (automatic word sense disambiguation) of documents by NLP techniques i.e. so that text words, or some designated subset of them, are tagged to particular senses.

ii. The use of thesauri in IR and NLP, the major intellectual and historical link between them.

iii. The prior analysis of queries and document indices so that their standard forms for retrieval reflect syntactic dependencies that could resolve classic ambiguities not of type (i) above.

Topic (i) is now mostly regarded as a diversion as regards our main focus of attention in this chapter; even though large- scale WSD is now an established technology at the 95% accuracy level (Stevenson and Wilks, 1999), there is no reason to believe it bears on this issue, largely because the methods for document relevance used by classic IR are in fact very close to some of the algorithms used for WSD as a separate task (in e.g. Yarowsky, 1992, 1995). IR may well not need a WSD cycle because it constitutes one as part of the retrieval process itself, certainly when using long queries as in TREC, although short web queries are a different matter, as we discuss below.

This issue has been clouded by the “one sense per discourse” claim of Yarowsky (1992, 1995), a claim that has been contested by Krovetz (1998) who has had had no difficulty showing that Yarowsky's figures (that a very high percentage of words occur in only one sense in any document) are wrong and that, outside Yarowsky’s chosen world of encyclopaedia articles, is not at all uncommon for words to appear in the same document bearing more than one sense on different occasions of use.

This dispute is not one about symbolic versus statistical methods for tasks, let alone AI versus IR. It is about a prior question as to whether there is any serious issue of sense ambiguity in texts to be solved at all, and by any method. In what follows I shall assume Krovetz has the best of this argument and that the WSD problem, when it is present, cannot be solved, as Yarowsky claimed in the one-sense-per-discourse paper, by assuming that only one act of sense resolution was necessary per text. Yarowsky’s claim, if true, would make it far more plausible that IR’s distributional methods were adequate for resolving the sense of component words in the act of retrieving documents, because sense ambiguity resolution would then be only at the document level, as Yarowsky’s claim makes clear.

If Krovetz is right, then sense ambiguity resolution is still a local matter within a document and one cannot have confidence that any word is univocal within a document, nor that a document-span process will resolve such ambiguity. Hence one will have less confidence that standard IR processes resolve such terms if they are crucial to the retrieval of a document. One will expect, a priori, that this will be one cause of lower precision in retrieval, and the performance of web engines confirms this anecdotally in the absence of any experiments going beyond Krovetz's own.

Let us now turn to (ii), the issue of thesauri: there is less in this link in modern times, although early work in both NLP and IR made use of a priori hand-crafted thesauri like Roget. Though there is still distinguished work in IR using thesauri in specialised domains, beyond their established use as user-browsing tools (e.g. Chiaramella and Nie, 1990), IR moved long ago towards augmenting retrieval with specialist, domain-dependent and empirically constructed thesauri, while Salton early on (1972) claimed that results with and without thesauri were much the same.

NLP has rediscovered thesauri at intervals, most recently with the empirical work on word-sense disambiguation referred to above, but has remained wedded to either Roget or more recent hand-crafted objects like WordNet (Miller, 1990). The objects that go under the term thesaurus in IR and AI/NLP are now rather different kinds of thing, although in work like Grefenstette and Hearst (1992) an established thesaurus like WordNet has been used to expand a massive lexicon for IR, again using techniques not very different from the NLP work in expanding IE lexicons referred to earlier.

Turning now to (iii), the use of syntactic dependencies in documents, their indices and queries, we enter a large and vexed area, in which a great deal of work has been done within IR (e.g. back to Smeaton and van Rijsbergen, 1988). There is no doubt that some web search engines routinely make use of such dependencies: take a case like

measurements of models

as opposed to

models of measurement

which might be expected to access different literatures, although the purely lexical content, or retrieval based only on single terms, might be expected to be the same. In fact they get 363 and 326 hits respectively in Netscape but the first 20 items have no common members. One might say that this case is of type (i), i.e. WSD, since the difference between them could be captured by, say, sense tagging "models" by the methods of (i), whereas in the difference between

the influence of X on Y

and (for given X and Y)

the influence of Y on X

one could not expect WSD to capture the difference, if any, if X and Y were 'climate' and 'evolution' respectively, even though these would then be quite different requests.

These are standard types of example and have been a focus of attention, both for those who believe in the role of NLP techniques in the service of IR (e.g. Strzalkowski and Vauthey, 1991), as well as those like Sparck Jones (1999a) who do not accept that such syntactically motivated indexing has given any concrete benefits not available by other, non-linguistic, means. Sparck Jones' paper is a contrast between what she call LMI (Linguistically Motivated Information Retrieval) and NLI (Non-Linguistically etc.), where the former covers the sorts of efforts described in this paper and the latter more 'standard' IR approaches. In effect, this difference always comes down to one of dependencies within, for example, a noun phrase so marked, either explicitly by syntax or by word distance windows. So, for example, to use her own principal example:

URBAN CENTRE REDEVELOPMENTS

could be structured (LMI-wise) as

REDEVELOPMENTS of [CENTRE of the sort URBAN]

or as a search for a window in full text as (NLI-wise)

[URBAN =0 CENTRE]<4 REDEVELOPMENTS

where the numbers refer to words that can intrude in a successful match.

The LMI structure would presumably be imposed on a query by a parser, and therefore only implicitly by a user, while the NLI window constraints would again presumably be imposed explicitly by the user, making the search. It is clear that current web engines use both these methods, with some of those using LMI methods derived them directly from DARPA-funded IE/IR work (e.g. NetOWL and TextWise). The job advertisements on the Google site show clearly that the basic division of methods at the basis of this chapter

have little meaning for the company, which sees itself as a major consumer of LMI/NLP methods in improving its search capacities.

Sparck Jones' conclusion is one of measured agnosticism about the core question of the need for NLP in IR: she cites cases where modest improvements have been found, and others where LMI systems' results are the same over similar terrain as NLI ones. She gives two grounds for hope to the LMIers: first, that most such results are over queries matched to abstracts, and one might argue that NLP/LMI would come into play more with access to full texts, where context effects might be on a greater scale. Secondly, she argues that some of the more negative results may have been because of the long queries supplied in TREC competitions, and that shorter more realistic and user-derived, web queries (which over 2.5 terms) might show a greater need for NLP. The development of Google, although proprietary, allows one to guess that this has in fact been the case in Internet searches.

On the other hand, she offers a general remark (and I paraphrase substantially here) that IR is after all a fairly coarse task and it may be not in principle optimisable by any techniques beyond certain limits, perhaps those we have already. Here the suggestion is that other, possibly more sophisticated, techniques should seek other information access tasks and leave IR as it is. This demarcation has distant analogies to one made within the word-sense discrimination research mentioned earlier, namely that it may not be possible to push figures much above where they now are, and therefore not possible to discriminate down to the word sense level, as oppose to the cruder homograph level, where current techniques work best, on the ground that anything "finer" is a quite different kind of job, and not a purely linguistic or statistical one, but rather one for future AI.

iv. The use of proposition-like objects as part of document indexing.

This is an additional notion, which, if sense can be given to it, would be a major revival of NLP techniques in aid of IR. It is an extension of the notion of (iii) above, which could be seen as an attempt to index documents by template relations, e.g. if one extracts and fills binary relation templates (X manufactures Y; X employs Y; X is located in Y) so that documents could be indexed by these facts in the hope that much more interesting searches could in principle be conducted (e.g. find all documents which talk about any company which manufactures drug X, where this would be a much more restricted set than all those which mention drug X).

One might then go on to ask whether documents could profitably be indexed by whole scenario templates in some interlingual predicate form (for matching against parsed queries) or even by some chain of such templates, of the kind extracted as a document summary by co-reference techniques (e.g. by Azzam et al., 1999).

Few notions are new, and the idea of applying semantic analysis to IR in some manner, so as to provide a complex structured (even propositional) index, go back to the earliest days of IR. In the 1960s researchers like Gardin (1965), Gross (1964) and Hutchins (1970) developed complex structures derived from MT, from logic or "text grammar" to aid the process of providing complex contentful indices for documents, entities of the order of magnitude of modern IE templates. Of course, there was no hardware or software to perform searches based on them, though the notion of what we would now call a full text search by such patterns so as to retrieve them go back at least to (Wilks,

1964, 1965) even though no real experiments could be carried out at that time. Gardin's ideas were not implemented in any form until (Bely et al., 1970), which was also inconclusive.

Mauldin (1991), within IR, implemented document search based on case-frame structures applied to queries (ones which cannot be formally distinguished from IE templates), and the indexing of texts by full, or scenario, templates appear in Pietrosanti and Graziadio (1997). The notion is surely a tempting one, and a natural extension of seeing templates as possible content summaries of the key idea in a text (Azzam et al., 1999). If a scenario template, or a chain of them, can be considered as a summary then it could equally well, one might think, be a candidate as a document index.

The problem will be, of course, as in work on text summarisation by such methods: what would cause one to believe that an a priori template could capture they key item of information in a document, at least without some separate and very convincing elicitation process that ensured that the template corresponded to some class of user needs, but this is an empirical question and one being separately evaluated by summarisation competitions.

Although this indexing-by-template idea is in some ways an old one, it has not been aired lately, and like so much in this area, has not been conclusively confirmed or refuted as an aid to retrieval. It may be time to revive it again with the aid of new hardware, architectures and techniques. After all, connectionism/neural nets was only an old idea revived with a new technical twist, and it had a ten year or more run in its latest revival. What seems clear at the moment is that, in the web and Metadata world, there is an urge to revive something along the lines of "get me what I mean, not what I say" (see Jeffrey, 1999). Long-serving IR practitioners will wince at this, but to many it must seem worth a try, since IE does have some measurable and exploitable successes to its name (especially Named Entity finding) and, so the bad syllogism might go, Metadata is data and IE produces data about texts, so IE can produce Metadata.

Question Answering within TREC

No matter what the limitation on crucial experiments so far, another place to look for evidence of the current of NLP/AI influence on IR might be the QA track within TREC since1999, already touched on above in connection with IRs influence on AI/NLP, or vice versa.

QA is one of the oldest and most traditional AI/NLP tasks (e.g. Green et al., 1961, Lehnert, 1977) but can hardly be considered solved by those structural methods. The conflation of the rival methodologies distinguished in this paper, can be clearly seen in the admitted possibility, in the TREC QA competition, of providing ranked answers, which fits precisely with the continuous notion of relevance coming from IR, but is quite counterintuitive to anyone taking a common sense view of questions and answers, on which that is impossible. It is a question master who provides a range of differently ranked answers on the classic QA TV shows, and the contestant who must make a unique choice (as opposed to re-presenting the proffered set!). That is what answering a question means; it does not mean “the height of St Pauls is one of [12, 300, 365, 508]feet”! A typical TREC question was “Who composed Eugene Onegin?” and the expected answer

was Tchiakowsky – which is not a ranking matter, and listing Gorbachev, Glazunov etc. is no help.

There were examples in the competition that brought out the methodological difference between AI/NLP one the one hand, and IR on the other, with crystal clarity: answers could be up to 250 bytes long, so if your text-derived answer was A, but wanting to submit 250 bytes of answer meant that you, inadvertently, could lengthen that answer rightwards in the text to include the form (A AND B), then your answer would become wrong in the very act of conforming to format. The anecdote is real, but nothing could better capture the absolute difference in the basic methodology of the approaches: one could say that AI, Linguistics and IR were respectively seeking propositions, sentences and byte-strings and there is no clear commensurability between the criteria for determining the three kinds of entities. More recently, Tait et al. (REF) have shown that if the queries are short (a crucial condition that separates off modern democratic and Google-based IR from the classic queries of specialists) then WSD techniques do improve performance.

Conclusion

One can make quite definite conclusions but no predictions, other than those based on hope. Of course, after 40 years, IR ought to have improved more than it has---its overall Precision/Recall figures are not very different from decades ago. Yet, as Sparck Jones has shown, there is no clear evidence that NLP has given more than marginal improvements to IR, which may be a permanent condition, or it maybe one that will change with full text search, and a different kind of user-derived query, and Google may be one place to watch for this technology to improve strongly. It may also be worth someone in the IE/LMI tradition trying out indexing-by-scenario templates for IR, since it is, in one form or another, an idea that goes back to the earliest days of IR and NLP, but remains untested.

It is important to remember as well that there is a deep cultural division in that AI remains, in part at least, agenda driven: in that certain methods are to be shown effective. IR, like all statistical methods in NLP as well, remains more result-driven, and the clearest proof of this is that (with the honourable exception of machine translation) all evaluation regimes have been introduced in connection with statistical methods, often over strong AI/linguistics resistance.

In IE proper, one can be moderately optimistic that fuller AI techniques using ontologies, knowledge representations and inference, will come to play a stronger role as the basic pattern matching and template element finding is subject to efficient machine learning. One may be moderately optimistic, too, that IE may be the technology vehicle with which old AI goals of adaptive, tuned, lexicons and knowledge bases can be pursued. IE may also be the only technique that will ever provide a substantial and consistent knowledge base from texts, as CYC (Lenat et al., 1986) has failed to do over twenty years. The traditional AI/QA task, now brought within TREC, may yield to a combination of IR and IE methods and it will be a fascinating struggle. The curious tale above, of the use of “translation” with IR and QA work, suggests that terms are very flexible at the moment and it may not be possible to continue to draw the traditional demarcations between IR and these close and merging NLP applications such as IE, MT and QA.

References

Agichtein, E., Grishman, R., Borthwick, A. and Sterling, J. (1998) Description of the named entity system as used in MUC-7. In Proc. of the MUC-7 Conference, NYU.

Azzam, S., Humphries, K., Gaizauskas, R. and Wilks, Y. (1997) Using a language-independent domain model for multilingual Information Extraction. Journal of Applied AI. 13.

Azzam, S., Humphreys, K. and Gaizauskas, R. (1999) Using coreference chains for text summarization. In Proc. ACL'99 Workshop on Coreference and its Applications, Maryland.

Ballasteros, L., and Croft, B. (1998) Statistical Methods for Cross Language Information Retieval. In Grefenstette (ed.) Cross Language Information Retrieval. Kluwer: Dordrecht.

Bely, N., Borillo, A., Virbel, J. and Siot-Decauville, N. (1970) Procedures d'analyse semantique appliquees a la documentation scientifique. Gauthier-Villars: Paris.

Berger, A. et al. (2000) Bridging the lexical chasm: statistical approaches to question answering. SIGIR'00.

Berger, A. and Lafferty, J. (1999) Information retrieval as statistical translation. SIGIR'99.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. Scientific American.

Bikel, D., Miller S., Schwartz, R. and Weischedel, R. (1997) Nymble: a High-Performance Learning Name-finder. In Proc. of the 5th conference on Applied Natural Language Processing (ANLP'97).

Brill, E. (1994) Some Advances in Transformation-Based Part of Speech Tagging. In Proc. of 12th National Conference on AI (AAAI'94), Seattle, Washington.

Brown, P.F. and Cocke, J. (1989) A Statistical Approach to Machine Translation, IBM Research Division, T.J. Watson Research Center, RC 14773.

Brown, P.F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R.L. and Roossin, P. (1990) A Statistical Approach to Machine Translation. Computational Linguistics, 16:2: 79-85

Bruce, R., and Guthrie, G. (1992) Genus disambiguation: a study in weighted preference. In Proc. COLING92, Nantes.

Carberry, S., Samuel, K. and Vijay-Shanker, K. (1998) Dialogue act tagging with transformation-based learning. In Proc. of COLING-ACL'98 Conference, vol. 2, pp. 1150-1156, Montreal, Canada.

Cardie, C. (1997) Empirical methods in information extraction. AI Magazine, 18(4), Special Issue on Empirical Natural Language Processing.

Cardie, C. and Lehnert, W. (1991) Preference Semantics and message Understanding. In Proc. DARPA Workshop on Spoeech and Language.

Charniak, E. (1973) Jack and Janet in search of a theory of knowledge. In Proc. IJCAI-73.

Charniak, E. (2001) Immediate-Head Parsing for Language Models. In Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01).

Chiaramella, Y. and Nie, J. (1990) A retrieval model based on an extended modal logic and its application to the RIME experimental approach. In Proc. of the 13th ACM International Conference on Research and Development in Information Retrieval (SIGIR'90), pp. 25-43.

Chomsky, N. (1965) Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Churcher, G., Atwell, E. and Souter, C. (1997) Dialogue Management Systems—a survey. Research Report, University of Leeds, Department of Computer Science.

Chelba, C. and Jelinek, F. (1998) Exploiting Syntactic Structures for Language and Modelling. In Proc. of ACL'98, Montreal, Canada.

Ciravegna, F. and Wilks, Y. (2003) Designing Adaptive Information Extraction for the Semantic Web in Amilcare, In S. Handschuh and S. Staab (eds). Annotation for the Semantic Web in the Series Frontiers in Artificial Intelligence and Applications by IOS Press, Amsterdam.

Colby, K.M. (1973) Simulation of Belief Systems, in Schank and Colby (eds.) Computer Models of Thought and Language, San Francisco, CA: W. H. Freeman.

Collier, R. (1998) Automatic Template Creation for Information Extraction. PhD thesis, University of Sheffield, Computer Science Dept., UK.

Cowie, J., Guthrie, L., Jin, W., Odgen, W., Pustejowsky, J., Wanf, R., Wakao, T., Waterman, S. and Wilks, Y. (1993) CRL/Brandeis: The Diderot System. In Proc. of Tipster Text Program (Phase I), Morgan Kaufmann.

Cunningham, H. (1999) JAPE – a Java Annotation Patterns Engine. Technical Report, Department of Computer Science, University of Sheffield.

Croft, W. and Lafferty, J. (eds.) (2000) Language Modelling for Information Retrieval. Kluwer: Dordrecht.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A. (1998) TiMBL: Tilburg memory based learner version 1.0. Technical report, ILK Technical Report 98-03.

Gachot, D., Lage, E., and Yang, J. (1998) The SYSTRAN NLP Browser: an application of MT technique in multilingual IR. In Greffenstette (ed.) Cross Language Information Retrieval, Kluwer: Dordrecht.

Gaizauskas, R. and Wilks, Y. (1997) Information Extraction: beyond document retrieval. Journal of Documentation.

Galliers, J. and Sparck Jones, K. (1996) Evaluating Natural Language Processing Systems. Lecture Notes in AI. Springer Verlag: berlin.

Gardin, J. (1965) Syntol. New Brunswick, NJ: Rutgers Graduate School of Library Science.

Garside, R. (1987) The CLAWS word-tagging system. In Garside, Leech and Sampson (eds.), The Computational Analysis of English. London and New York: Longman.

Gibson, J.J. (1968) The senses considered as a perceptual system. Allen and Unwin: London.

Gollins, T. and Sanderson, M. (2001) Improving Cross Language Information Retrieval with triangulated translation. SIGIR'01.

Granger, R. (1977) FOULUP: a program that figures out meanings of words from context. In Proc. 5th International Joint Conference on Artificial Intelligence (IJCAI'77).

Green, B., Wolf. A., Chomsky, C. and Laughery, K. (1961) BASEBALL, an automatic question answerer. In Proc. Western Joint Computer Conference 19, pp. 219-224

Grefenstette, G. and Hearst, M.A (1992) Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results. In Weir (ed.) Statistically-based natural language programming techniques. In Proc. AAAI Workshop, AAAI Press, Menlo Park, CA.

Grishman, R. (1997) Information extraction: Techniques and challenges. In M-T. Pazienza (ed.) In Proc. of the Summer School on Information Extraction (SCIE'97), LNCS/LNAI. Springer-Verlag.

Grishman, R. and Sterling, J. (1992) Generalizing automatically generated patterns. In Proc. of COLING'92.

Gross, M. (1964) On the equivalence of models of language used in the fields of mechanical translation and information retrieval. Information Storage and Retrieval. 2(1).

Hobbs, J.R. (1993) The generic information extraction system. In Proc. of 5th Message Understanding Conference (MUC-5), pp. 87-91. Morgan Kaufman.

Hutchins, W.J. (1970) Linguistic processes in the indexing and retrieval of documents. Linguistics, 61.

Jeffrey, K. (1999) What's next in databases? ERCIM News (www.ercim.org) 39.

Krovetz, R. (1998) More than one sense per discourse. NEC Princeton NJ Labs., Research Memorandum.

Lehnert, W. (1977) A Conceptual Theory of Question Answering. In Proc. of 5th IJCAI, Cambridge, MA. Los Altos: Kaufmann, pp. 158-164.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J. and Riloff, E. (1992) University of Massachusetts: Description of the CIRCUS system as used for MUC-4. In Proc. of the 4th Message Understanding Conference MUC-4, pp. 282-288. Morgan Kaufmann.

Lenat, D., Prakash, M. and Shepherd, M. (1986) CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks, The AI Magazine, 6(4).

Lewis, D. (1972) General Semantics. In Davidson and Harman (eds.) Semantics of natural language. Reidel: Dordrecht.

Luhn, H.P. (1957) A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1:309-317.

Marr, D. (1981) Artificial Intelligence: a personal view. In Haugeland, J. Mind Design. Cambridge, MA: MIT Press.

Mauldin, M. (1991) Retrieval performance in FERRET: a conceptual information retrieval system. SIGIR'91.

McCarthy, J. and Hayes, P. (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence, In Machine Intelligence 4. Edinburgh Univ. Press, Edinburgh.

Miller, G.A. (ed.) (1990) WordNet: An on-line Lexical Database, In International Journal of Lexicography, 3(4).

Moldovan, D. (2001) Question Answering Systems in Knowledge Management. In Proc. IEEE Intelligent Systems. 16.

Morgan, R., Garigliano, R., Callaghan, P., Poria, S., Smith, M., Urbanowicz, A., Collingham, R., Costantino, M. and Cooper, C. (1995) Description of the LOLITA System as used for MUC-6. In Proc. of the 6th Message Understanding Conference (MUC-6), pp. 71-86, San Francisco, Morgan Kaufmann.

Muggleton, S. (1994) Recent advances in inductive logic programming. In Proc. 7th Annual ACM Workshop on Computer Learning Theory, pp. 3-11. ACM Press, New York, NY.

Nelson, T. (1997) The future of information. ASCII Press: Tokyo.

Nirenburg, S., Frederking, R., Farwell, D., and Wilks, Y. (1994) Two types of adaptive MT environments. In Proc. COLING94, Kyoto.

Nirenburg, S. and Wilks, Y. (2001) What's in a symbol: ontology, representation and language. Journal of Experimental and Theoretical Artificial Intelligence (JETAI).

Oh, J., Chae, Y. and Choi, K. (2000) Japanese term extraction using a dictionary hierarchy and an MT system. Terminology 6.

Pearl, J. (1985) Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning, In Proc. of Cognitive Science Society (CSS-7).

Pollack, J. (1990) recursive Distributed Representations. Artificial Intelligence. 46.

Ponte, J. and Croft, B. (1998) A language modelling approach to Information Retrieval. In Proc. 21st ACM SIGIR, Melbourne.

Pustejovsky, J. (1995) The Generative Lexicon. Cambridge, MA: MIT Press.

Resnik, P. (1996) Selectional Constraints: information theoretic model. Cognition. 61.

Riloff, E. and Lehnert, W. (1993) Automated dictionary construction for information extraction from text. In Proc. of 9th IEEE Conference on Artificial Intelligence for Applications, pp. 93-99.

Riloff, E. and Shoen, J. (1995) Automatically acquiring conceptual patterns without an annotated corpus. In Proc. of 3rd Workshop on Very Large Corpora.

Roche, E. and Schabes, Y. (1995) Deterministic Part-of-Speech Tagging with Finite-State Transducers. Computational Linguistics, 21(2):227-254.

Salton, G. (1972) A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). Journal of the American Society of Information Science, 23(2).

Schank, R. (1975) Conceptual Information Processing, North Holland, Amsterdam.

Schvaneveldt, R. (ed.) (1990) Pathfinder Networks: Theory and Applications. Ablex, Norwood, NJ.

Smeaton, A. and van Rijsbergen, C. (1988) Experiments in incorporating syntactic processing of user queries into a document retrieval strategy. In Proc. 11th ACM SIGIR.

Sparck Jones, K. (1966/1986) Synonymy and Semantic Classification. Edinburgh University Press, Edinburgh.

Sparck Jones, K. (1990) Retrieving Information or Answering Questions? British library Annual research Lecture, London: British Library.

Sparck Jones, K. (1999a) What is the role of NLP in text retrieval. In Strzalkowski (ed.) Natural language Information Retrieval. Kluwer: New York.

Sparck Jones, K. (1999b) Information Retrieval and Artificial Intelligence. Artificial Intelligence Journal, vol. 114.

Sparck Jones, K. (2003) Document Retrieval: shallow data, deep theories, historical reflections and future directions. In Proc. 25th European IR Conference (ECIR03). Lecture Notes in Computer Science. Berlin: Springer. Pp.1-11.

Stevenson, M. and Wilks, Y. (1999) Combining Weak Knowledge Sources for Sense Disambiguation. In Proc. of the International Joint Conference for Artifical Intelligence (IJCAI'99)

Stevenson, M. and Gaizauskas, R. (2004) Recognition using annotated corpora. In proc. European Conference on Information Retrieval (ECIR04), Sunderland.

Strzalkowski, T. and Vauthey, B. (1991) Natural Language Processing in Automated Information Retrieval, PROTEUS Project Memorandum. Department of Computer Science, New York University.

Tait, J. (2004)

Vilain, M. (1993) Validation of terminological inference in an information extraction task. In Proc. of ARPA'93 Human Language Workshop.

Vossen, P. (ed.) (1998) Eurowordnet . Kluwer: Dordrecht.

Wilks, Y. (1964) Text Searching with Templates. Cambridge Language Research Unit Memo, ML 156.

Wilks, Y. (1965) The application of CLRU's method of semantic analysis to information retrieval. Cambridge Language Research Unit Memo, ML.173.

Wilks, Y. (1977) Good and Bad Arguments About Semantic Primitives. In Communication and Cognition, Vol. 10, No 3/4.

Wilks, Y. (1979) Frames, semantics and novelty. In Metzing (ed.) Frame Conceptions and Text Understanding. Berlin: de Gruyter.

Wilks, Y. (1990) "INterlinguas in Japanese Machine Translation" In Report of the National Science Foundation JTEC Panel, to assess Japanese Natural Language Processing (eds. Carbonell and Rich), National Science Foundation, Washington DC.

Wilks, Y. and Fass, D. (1992) Preference Semantics: a family history. In Computing and Mathematics with Applications, Vol. 23, No. 2.

Wilks, Y., Slator, B. and Guthrie, L. (1996) Electric Words: dictionaries, computers and meanings. Cambridge, MA: MIT Press.

Wilks, Y. and Catizone, R. (1999) Making information extraction more adaptive. In M-T. Pazienza (ed.) In Proc. of Information Extraction Workshop, Frascati.

Winograd, T. (1971) Understanding Natural Language, MIT Press, Cambridge, MA.

Winograd, T. and Flores, A. (1986) Understanding Computers and Cognition: A New Foundation for Design, Ablex: Norwood, NJ.

Woods, W., Kaplan, R. and Nash-Webber, B. (1974) The Lunar Sciences Natural Language Information System, Final Report 2378, Bolt, Beranek and Newman, Inc., Cambridge, MA.

Yarowsky, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proc. COLING'92, Nantes, France.

Yarowsky, D. (1995) Unsupervised word-sense disambiguation rivalling supervised methods. In Proc. of ACL'95.

Yngve, V.H. (1960) A model and an hypothesis for language structure. In Proc. of American Philosophical Society, Vol. 104, No. 5, pp. 444-466.

Acknowledgments:

The paper has benefited from discussions with Mark Sanderson, Rob Gaizauskas, Ted Dunning and others, but the errors are all mine of course