IR and AI: traditions of representation and anti-representation in information processing

Yorick Wilks
University of Sheffield

Introduction

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), 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 figures 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 a machine translation (MT) system (CANDIDE, see Brown and Cocke, 1989) 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 cannot contain or manipulate representations. On such a 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, an insight that spawned the speech act movement in linguistics and NLP, and 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 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 revived Connectionist/neural net movement of the

1980's brought some clarification of the issue within 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 (Pathfinder, 1990)--producing a network whose content is extracted directly from subjects' responses--and whose predictive powers 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, given input of 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 his Pathfinder networks) on an overall statistical measure of clustering that establishes the Pathfinder network from the subject-derived data in the first place. Its challenge to traditional AI can be guaged from John McCarthy's classic response to any attempt to introduce statistical notions into 1970's AI: "Where do all these numbers COME FROM?".

The Pathfinder example is highly relevant here, not only for its direct challenge to a core area of old 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 the same 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 took AI on directly in a core area.

It is Schvaneveldt's results on knowledge elicitation by these methods from groups like pilots, and the practical difference such structure make in training, that constitute their threat to propositionality here.

This is no unique example of course: even in older AI one thinks of Judea Pearl's long 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 this intuition with central AI notions of machine reasoning.

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

In this paper, I am taking IE as a paradigm, naive though it still is, 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, although one must always bear in mind that virtually all IE search rests on a prior IR retrieval of relevant documents or paragraphs.

IE is a technique which, although still dependent of 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 some of its rule-driven successes can be matched by new 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 commercialization 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 Defence agencies from where the technology was largely developed. Yet this is not always the case in areas appropriate for the commercial deployment of IE. 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 recent DARPA-sponsored MUC and TIPSTER IE competitions, most of which have a generic structure (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 have been 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 adapt 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 and 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 those of machine learning, and which are statistical in nature, like IR but unlike core AI. However, most of those used have been supervised techniques, which do tend to assume the need for some form of human-assignable representations.

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). The most interesting 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 (1978) and Granger (1977) on detecting new preferences of words in texts and interpreting novel lexical items from context and stored knowledge. This notion is important, not only for IE in general but in particular as it adapts to traditional AI tasks like Question Answering, now also coming within the IR remit (see below on TREC). There are also strong similarities between these forms of lexicon development and tuning done within AI/NLP and recent activities by e.g. Grefenstette and Hearst (see below) on building massive IR lexicons from texts.

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. Muggleton's (et al., 1997) learning system at York has provided very good evaluated figures indeed (in world wide terms) in learning part of speech tagging and is being extended to grammar learning. Muggleton also has experimented with user interaction with a system that creates semantic networks of the articles and the relevant templates, although so far its published successes have been in areas like Part-of- Speech tagging that are not inherently structural (in the way template learning arguably is).

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 machine learning 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).

Unsupervised template learning

We should 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 old technique of Luhn (1957) to locate statistically significant words in a corpus and 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 for over 30 years, in the debate about the relevance of symbolic or linguistic (or AI taken broadly) considerations in the task of information retrieval. Note immediately that this is not the reverse question--touched on in the historical review above----as to the relevance of IR-type methods for traditional NLP processing tasks, like machine translation and lexical structure, and which in its wider form concern the relevance of statistical methods to NLP in general.

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 for, 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 part of ML techniques, one cannot simply contrast statistical with linguistic methods in IE as Sparck Jones (1999a) does when discussing IR. In this connection, one must mention one of the most recent successes of purely statistical methods in IE, namely the BBN trained named entity finder, which is wholly statistical and producing results comparable with the large sets of grammatical rules just mentioned.

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 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 in (Carberry et al.)) can be treated by statistical methods.

A key high-level IE task at the moment is co-reference, a topic that one 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 which seem a priori to require knowledge and inference about the ways in which individuals can be described.

There are three standard techniques in terms of which this possible infusion (of possible NLP techniques into IR) have been discussed, and I will 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 in the form of a document to be retrieved by IR engines.

ii. The use of thesauri in IR and NLP as an 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 a diversion as regards our main focus of attention in this paper; 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 the issue to hand, 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 has one as part of the retrieval process itself, certainly when using long queries as in TREC.

This issue has been clouded by the "one sense per discourse" claim of Yarowsky (1992, 1995), which has been hotly 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 just wrong and that, outside Yarowsky's chosen world of encyclopedia articles, is not at all uncommon for words to appear in the same document bearing more than one sense on different occasions of use.

Note that this dispute is not one about symbolic versus statistical methods for task, 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 distributional methods were adequate for resolving the sense of component words in the act of retrieving documents, because sense ambiguity resolution is 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 and 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 Hearst and Grefenstette (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 and large and vexed area, in which a great deal of IR work has been done within IR (e.g. 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

where these might be taken 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 the focus of attention both of 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 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. 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 paper 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 achievements 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 and, secondly, 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, queries 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 the world of 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 one not in principle optimisable by any techniques beyond certain limits, perhaps those we have already; the suggestion being 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, if anything.

Sparck Jones developed these views further in (1999b), and as far as to call on AI in general to adopt more IR-like methodologies. In so far as that means evaluation techniques, no one could possibly disagree but, curiously, in the area under discussion one might even say the battle has gone the other way. Since about the time of her own paper, have come a stream of papers, starting with Berger and Lafferty (1999) with titles like "Information Retrieval as Statistical Translation" in which, in a curious and inverted sense, machine translation is being taken as a desirable model of IR to conform to, which is certain the reverse of the shift Sparck Jones wanted.

As always, things are not as radical as they seem: the genesis of the Berger and Lafferty work, under the broad heading of "language models in IR", was the IBM statistical translation work at IBM under Jelinek referred to earlier, and the word "translation" in the Berger/Lafferty title refers in a sense to any technique which considers two strings in relationship to each other as, indifferently, translations of each other or retrievals of each other. This is undoubtedly a touch of sleight of hand here, since translation is normally considered a symmetrical relationship, but retrieval is not, since documents are not normally considered as retrieving queries.

However, and quibbles aside, this approach has now even suggested considering question-answering as a form of translation (as it been seen as a form of IR for quite a while) and in that case the asymmetry is not so striking (see berger et al., 2000). All this work remains statistical, as indeed is so much of NLP and AI these days, but there is clearly an element here of NLP techniques, however construed, being applied to IR (rather than the reverse) even if much of this is achieved by a widening or redefinition of the core IR task itself. It is worth noting in that connection that the development of IR as cross-language task has also, and inevitably, increased the role of NLP techniques, and made access to NLP resources, such as lexicons seems natural and obvious, as is shown in work like Gollins and Sanderson's (2001) cross-language retrieval by what they call "language triangulation."

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

This is a 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 ideas 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 indexes for documents, 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 appears 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 the summarisation work by such methods, what would cause one to believe that an a priori template would or 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 has 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 goes, 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 new Question-Answering track within TREC 1999, already touched on above in connection with IRs influence on AI?NLP, or vice versa.

Question answering 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, or confusion, 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 Gorbachev, Glazunov etc. are no help.

There were examples in this 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), in which case 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 ontology 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.

For this first TREC question answering competition, comparative results across sites will undoubtedly be forthcoming soon though this will be taken as only bench marking for subsequent years, and not the settling of this deep ideological divide, one which web entrepreneurs cheerfully ignore, taking their techniques from wherever they can. But I suggest, for anyone interested in the issue, this TREC track is the one to watch. There is an interesting parallel, by the way, for this byte-approach to Q/A: in 1996 CMU entered into the Loebner Turing competition, a computer discussant that answered questions, and produced all responses, by a closest match word-window algorithm into a large newspaper corpus. It didn't do very well, but not nearly as disastrously as those in the LMI school would have predicted and wanted! One should remember too that very early attempts were made, within the IR tradition, (by O'Connor and Miller) to use retrieval of micro-texts as a form of QA.

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 adds more than marginal improvements to IR, which may be a permanent condition, or 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 striingly. 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 te 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 of ontology, knowledge representation 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 IR, IE methods or some mixture of the two, but it will be a fascinating struggle. The curious tale above, of the use of translation with IR and QA work suggests that ters are very flexible at the moment and it may not be possible to continue to draw the traditional demarcations between these close and merging NLP applications IE, MT, QA and so on.

REFERENCES

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

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

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., Lafferty, J. (1999) Information retrieval as statistical translation. SIGIR 1999.

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

Bikel D., Miller S., Schwartz R., and Weischedel R. (1997) Nymble: a High-Performance Learning Name-finder. In Proceedings of the Fifth conference on Applied Natural Language Processing.

Brill E. (1994) Some Advances in Transformation-Based Part of Speech Tagging. In Proceedings of the Twelfth National Conference on AI (AAAI-94), Seattle, Washington.

Brown, P. F. and Cocke, John, (1989) A Statistical Approach to Machine Translation, IBM

Research Division, T.J. Watson Research Center, RC 14773.

Carberry S., Samuel K. and Vijay-Shanker K. (1998) Dialogue act tagging with transformation-based learning. In Proceedings of the COLING-ACL 1998 Conference, volume 2, pages 1150--1156, Montreal, Canada, 1998.

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

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

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 Proceedings 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.

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.

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

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

Gollins, T., Sanderson, M. (2001) Improving Cross Language Information Retrieval with triangulated translation. SIGIR 2001.

Granger, R. (1977) FOULUP: a program that figures out meanings of words from context. Proc. Fifth Joint Internat. Conf. on AI.

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

Grefenstette, G., 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, Proc. AAAI Workshop, AAAI Press, Menlo Park, CA.

Grishman R. (1997) Information extraction: Techniques and challenges. In M-T. Pazienza, editor, Proceedings of the Summer School on Information Extraction (SCIE-97), LNCS/LNAI. Springer-Verlag.

Grishman R. and Sterling J. (1992) Generalizing automatically generated patterns. In Proceedings 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 Proceedings of the Fifth Message Understanding Conference (MUC-5), pages 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.

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

Lehnert, W. (1977) A Conceptual Theory of Question Answering. Proc. Fifth IJCAI, Cambridge, MA. Los Altos: Kaufmann, 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 Proceedings of the Fourth Message Understanding Conference MUC-4, pages 282--288. Morgan Kaufmann.

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

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

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

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 Proceedings of the Sixth Message Understanding Conference (MUC-6), pages 71--86, San Francisco, Morgan Kaufmann.

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

Muggleton S., Cussens J., Page D., and Srinivasan A. (1997) Using inductive logic programming for natural language processing. In Proceedings of in ECML, pages 25--34, Prague. Springer-Verlag. Workshop Notes on Empirical Learning of Natural Language Tasks.

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

Pietrosanti, E., and Graziadio, B. (1997) Extracting Information for Business Needs. Unicom Seminar on Information Extraction, London, March.

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

Riloff E. and Shoen J. (1995) Automatically acquiring conceptual patterns without an annotated corpus. In Proceedings of the Third 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).

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. Proc. 11th ACM SIGIR.

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

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.

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

Vilain M. (1993) Validation of terminological inference in an information extraction task. In Proceedings of the 1993 ARPA Human Language Workshop.

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

Wilks, Y., (1998) Senses and Texts, In N. Ide (ed.) special issue of Computers and the Humanities.

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

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

Winograd, T. (1971) Understanding Natural language.

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

Yarowsky, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proc. COLING92, Nantes, France.

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

Acknowledgments

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