Information Extraction as a core language technology

What is IE?

YORICK WILKS

Information Extraction (IE) technology is now coming on to the market and is of great significance to information end- user industries of all kinds, especially finance companies, banks, publishers and governments. For instance, finance companies want to know facts of the following sort and on a large scale: what company take- overs happened in a given time span; they want widely scattered text information reduced to a simple data base. Lloyds of London need to know of daily ship sinkings throughout the world and pay large numbers of people to locate them in newspapers in a wide range of languages. All these are prima facie cases for what we are calling IE.

Computational linguistic techniques and theories are playing a strong role in this emerging technology (IE), not to be confused with the more mature technology of Information Retrieval (IR), which selects a relevant subset of documents from a larger set. IE extracts information from the actual text of documents, by computer and at high speed, and normally from publicly available electronic sources such as news wires. Any application of this technology is usually preceded by an IR phase, which selects a set of documents relevant to some query--normally a string of features or terms that appear in the documents. So, IE is interested in the structure of the texts, whereas one could say that, from an IR point of view, texts are just bags of unordered words.

You can contrast these two ways of envisaging text information and its usefulness by thinking about finding, from the World Wide Web, what TV programs you might want to watch in the next week: there is already a web site in operation with text descriptions of the programs on 25 or more British TV channels, more text than most people can survey easily at a single session. On this web site you can input the channels or genre (e.g. musicals, news etc.) that interest you and the periods when you are free to watch. You can also specify up to twelve words that can help locate programs for you, e.g. stars' or film directors' names. The web site has a conventional IR engine behind it, a standard boolean function of the words and genre/channel names you use. The results are already useful--and currently free--- and treat the program descriptions as no more than "bags of words".

Templates and their limitations

Now suppose you also wanted to know what programs your favourite TV critic liked: and suppose the web site also had access to the texts of recent newspapers. An IR system cannot answer that question because it requires searching review texts for films and seeing which one are described in favourable terms. Such a task would require IE and some notion of text structure. In fact, such a search for program evaluations is not a best

case for IE, and I mention it only because it is an example of the kind of leisure and entertainment application that will be so important in future informatics developments. To see that one only has to think of the contrast between the designed uses and the actual uses of the French Minitel! system--designed for phone number information but actually used largely as an adult dating service.

Some domains push out the limits of templatability, particularly any area with an evaluative component: e.g. one can search movie reviews for directors and actors--even for films where an individual appears in the non-standard role, such as Mel Gibson as a director, and that is a difficult task for an IR system----those are potentially matchable to templates but a much harder task is to decide if a movie review is positive or not. It is said that US Congressmen, who receive vast amounts of email that almost certainly cannot read, would welcome any IE system that could tell them simply, of each email message. The result of such a component could clearly be expressed as a template--what is unclear is how one could fill it in a reliable manner.

It is worth considering, in this context of vaguer forms of information, the extent to which a full text IR system can provide some apparently IE features. If one types to a well known and venerable IR system that has a windowing capacity "fat Scots psychologist", where the system has access to a year of indexed London Financial Times text and one can specify that say those words occur in that order within a window of 6 words---i.e. so very few other adjectives or nouns could intervene to break up that string. One will obtain citations of television program reviews including two of the prize - winning British TV series Cracker-- which is almost certainly what anyone in the UK would have wanted who used the quoted phrase. We must suppose a viewer who had a clear goal in mind but no access to the title or the star's name-- many users of such systems will be in that kind of position! This is still clearly an IR task but one whose boundary with IE is blurred because of the importance of syntagmic (NLP if you prefer) features imposed by the very length of the window chosen, as opposed to searching for ANY text with those three words.

A Brief History

Information extraction is a new technology not a new idea: as long ago as 1964 can be found papers with titles like "Text searching with templates" (REF), but these were ideas not backed by any computational power capable of carrying them out. The earliest effective IE work was undoubtedly that of Sager (1970) within a medical domain, and constituting a long- running project combining surface syntax analysis and the use of templates. The work depended on handcrafted structures and a narrow range of techniques, but was highly effective. More immediately striking was de Jong's FRUMP (1979) an adaptation of Schank's high-level script structures, or what we would now call scenarios (e.g. for a terrorist event), which attempted to fill the slots in such structures from the AP newswires. This work was prescient in many ways-- including the choice of terrorism as a topic--but it had the toy quality of its time and was never evaluated in any quantitative way, e.g. against standard information retrieval systems ability to select terrorism stories from the AP wire, since it was being used for routing as well as

extraction. FRUMP later became the basis for an early commercial system TRANS (REF). Other early work that was IE before it knew it was was Cowie's 1983 extraction of canonical structures (REF) from field-guide descriptions of plants and animals.

The first IE system to result from a complex commercially posed problem was Hayes et al.s JASPER system at Carnegie Group (1992?). Like the earlier systems it relied on high degree of handcrafting, like them too it had no access to major external linguistic resources: corpora, lexica or gazetteers, nor did it incorporate any learning algorithms. However it was benchmarked and evaluated in a serious manner, even if not within the US MUC and TIPSTER regimes.

The key roles of evaluation regimes

IE shares a key feature with an older, more mature, discipline, machine translation: both are associated with a strong evaluation methodology. This means that one can determine, by objective standards, how well a system is doing at a chosen task and whether it is getting better with time. It is not possible when writing of the brief history of IE as a technology to separate it from the ARPA-sponsored MUC competitions and the TIPSTER IE project, for it is this competitive and objective environment that have created they field by a kind of forcing..

IE as a subject and the current standards of evaluation and success were first surveyed in (Lehnert & Cowie 1996), and broadly one can say that the field grew very rapidly when ARPA, the US defence agency, funded competing research groups to pursue IE, and based initially round scenarios like terrorism events as reported in newspapers, a task where the funders wanted to replace the government analysts who read the newspapers for terrorist events and then filled templates: when, where, a terrorist event took place, how many casualties etc.. It was this activity that IE was designed to automate.

Empirical linguistics: corpora, modularity and architectures

However, the IE movement has also grown by exploiting, and joining, the recent trend towards a more empirical and text based computational linguistics, that is to say by putting less emphasis on linguistic theory and trying to derive structures and various levels of linguistic generalisation from the large volumes of text data that machines can now manipulate. This movement has many strands: the use not only of machine readable dictionaries, taken from publishers tapes and from which semantic information has been extracted on a large scale (REFS) but also resources like WordNet, a semantic thesaurus in network form, created explicitly as a basis for NLP. The empirical movement has also joined forces with a revival of machine learning (REF) so that large scale linguistic facts are not only derived from these sorts of resources but algorithms are used (both classic and novel) to derive structures at whatever level and then adapt them automatically to more data.

Two structural features have played roles in the astonishing growth of IE: first, that of modularity, the notion that computational linguistic tasks could be separated out and evaluated quite separately, and that the separation between the modules need not correspond only to classic divisions of linguistic levels, such as syntax and semantics.

Secondly, came the notion of an an architecture for NLP, and environment in which the modules just referred to could be combined in various ways to perform large scale tasks (e.g. IE but also MT or QA) and different modules for the same task could be systematically compared, or the same module could have its performance compared over different text types.

A conspicuous success has been a module usually called a part-of- speech tagger: a system that assigns one and only one part- of-speech symbol (like Proper noun, or Auxiliary verb) to a word in a running text and do so on the basis (usually) of statistical generalisations across very large bodies of text. Recent research has shown that a number of quite independent modules of analysis of this kind can be built up independently from data, usually very large electronic texts, rather than coming from either intuition or some dependence on other parts of a linguistic theory.

That these modules, can be independent--in their construction and optimisation--is something of a break with much traditional computational linguistics which tended to suggest that linguistic modules all depended on each other and could not be separated out in this way. The knowledge-based approach to NLP in artificial intelligence tended to reinforce this view: that even the simplest tasks, like tagging, could and did require the application of the highest level resources and inferences. That this is not in general true has been amazingly refreshing, and the relative independence of such modules at quite high levels of performance has made the construction of whole systems much simpler.

Such modules would usually be taken to include, as well as part of speech taggers: morphology analysis modules; modules to align texts sentence-by-sentence in different languages; syntax analysis modules, modules attaching word sense tags to words in texts to disambiguate them in context and so on. That these tasks can be done relatively independently is very surprising to those who believed them all contextually dependent sub-tasks within a larger theory. These modules have been combined in various ways to perform tasks like IE as well as more traditional ones like machine translation (MT). The modules can each be evaluated separately --but they are not in the end real human tasks that people actually do, as MT and IE are.

One can call the former "intermediate" tasks and the latter real or final tasks---and it is really only the latter that can be firmly evaluated against human needs ------by people who know what a translation, say, is and what it is for. The intermediate tasks are evaluated internally to improve performance but are only, in the end, stages on the way to some larger goal. Moreover, it is not possible to have quite the same level of confidence in them since what is, or is not, a correct syntactic structure for a sentence is clearly more dependent on one's commitments to a linguistic theory of some sort, and such matters are in constant dispute. What constitutes proper extraction of people's names

from texts, or a translation of it, can be assessed by many people with no such subjective commitments.

The MRD stream in the empirical movement: lexicon acquisition

The empirical movement, basing, as it does, linguistic claims on text data, has another stream, noted earlier: the use in language processing of large language dictionaries (of single languages and bilingual forms) that became available about ten years ago in electronic forms from publishers' tapes. These are not textual data in quite the sense above, since they are large sets of intuitions about meaning set out by teams of lexicographers or dictionary makers. Sometimes they are actually wrong, but they have nevertheless proved a useful resource for language processing by computer, and lexicons derived from them have played a role in actual working MT and IE systems (REF).

What such lexicons lack is a dynamic view of a language; they are inevitably fossilised intuitions. To use a well known example: dictionaries of English normally tell us that "television" is as a technology or a TV set, although it is mainly used now to mean the medium itself. Modern texts are thus out of step with dictionaries--even modern ones. It is this kind of evidence that shows that, for tasks like IE, lexicons must be adapted or "tuned" to the texts being analysed which has led to a new, more creative wave, in IE research: the need not just to use large textual and lexical resources, but to adapt them as automatically as possible, to enable them to adapt to new domains and corpora, which will mean dealing with obsolescence and with the specialised vocabulary of a domain not encountered before.

Moularity and the stereotypical IE system

Most of the modules listed above can be found, in some form, in virtually all the major IE systems currently competing in MUC. This is because, to a perhaps surprising degree, they share not only modules but assumptions about how the IE task with templates is to be done. In Cowie and Lehnert (1996) these asumptions were identified as:

* the power of shallow parsing (as against full syntax analysis) * the power of shallow knowledge (simple, low level knowledge such as gazetteer lists and what can be derived from the template fillers themselves) * mining the templates--effectively using filled templates to derive more shallow knowledge * use of the test corpora themselves for data,, including lexica, preferably with learning algorithms to tune at least some of the modules.

These assumptions are partially independent of the selection of modules, as can be seen from the Sheffield VIE system (see below) which has several conventional modules but does not embody these assumptions at all. However, the general structural similarity of the major systems is such that Hobbs (1995 REF) was able to describe, without satire, a Generic IE System which closely matches most major players. This genericity is in part

to be expected from an intensively competitive environment like MUC in which data and modules are also shared: the systems tend to all tune to each other and innovation is penalised. The benefits, as in horse breeding, say, are well known but there are penalties: the gene pool, as it were, becomes smaller and this may explain why outside entrants (from outside the trained group or gene pool: exogametes in fact) have done better than expected.

The European IE scene

As part of its large scale Language Engineering R & D program, the European Commission supports a considerable range of IE projects, built into either specific industrial-backed applications or as research projects: these include TREE (IE and MT for job applications across national boundaries), AVENTINUS (a classic IE application to police and terrorism issues for EU police forces); FACILE, ECRAN and SPARKLE (research projects combining IE and IR in different ways and with "tuning" or automatic adaptation aspects). ECRAN, for example, searches movie and financial databases and exploits the notion we mentioned of tuning a lexicon so as to have the right contents, senses and so on to deal with new domains and relations unseen before.

The EC funders have, on the whole, been less sympathetic to the benefits of MUC- style competition. They argue that it is wasteful of resources, though this ignores the fact that much of the MUC effort is "voluntary" in the sense of not directly funded by ARPA. In any case, the EU, like everyone else gets the benefit of published MUC-related research. But a more sophisticated defence could be given of the EC position, which is that ARPA's needs are basically defence driven and the ECs are commerce driven. Since all EC projects need industrial partnership and exploitation plans, then the market will ideally decide all issues, and there is no need of explicit competitions, nor could there be sufficient cooperation between the sponsoring companies to allow such diversion of effort and indeed openness. The paradox behind that view, of course, is that it is the US companies, driven by ARPAs competitive regime. which have also put the first IE products on the market.

So the EC has considered general NLP and IE assessment procedures and decided against them, but the French government is said to be actively considering some such competition in France, perhaps to compensate for the fact that French as a language is highly unlikely to interest ARPA in the forseeable future. The British government funds some IE programs (including VIE and the GATE architecture, see below) but has no incentive to promote competition since British teams can and do enter MUC, at their own cost.

Possible limitations of IE systems.

We have presnted IE here and as a new and vital technology, though in the following section we will also review ways in which it, like all language processing technologies, are being merged in original combinations to meet particular needs. But let us here, as an initial conclusion, review the possible limits to what we have described, quite apart from

quantitative limits which may be expected d to appear in MUC competitions as they have in IR and MT.

The main problem with IE is the degree to which knowledge IS template like, in the way history was once taught (but is no longer) as factual templates of kings, presidents, battles and dates. A major research issue is seeing how far the boundaries of templatability can be pushed out.

A closely related question is how far the construction and adaptation of templates to new domains (along with the adaptation of the associated lexical structures and lexicons) can be made practical and cost effective.

Both these questions can be brought under the heading of the adaptation of IE, to new domains, and to the needs of new users: research is going on not only on a‘semi-automatic techniques for deriving templates from corpora --by seeing if a corpus contains significant patterns of a template type--but doing this in conjunction with user modelling techniques, representing user needs, and uses of things like natural dialogue (as has been done in IR) to allow a user to express needs--such as particular patterns sought and so on--in a natural language.

IE as part of a cluster of information technologies

An important insight, even after accepting our argument that IE is a new, emergent technology, is that what may seem to be wholly separate information technologies are really not so: MT and IE, for example, are just two ways of producing information to meet people’s needs and can be combined in differing ways: for example, one could translate a document and then information extract from the result or vice-versa, which would mean just translating the contents of the resulting templates. Which of these one chose to do might depend on the relative strengths of the translation systems available: a simpler one might only be adequate to translate the contents of templates, and so on. This last observation emphasises that the product of an IE system--the filled templates--can be seen either as a compressed, or summarised, text itself, or as a form of data base (with the fillers of the template slots corresponding to conventional database fields). One can then imagine new, learning, techniques like data mining being done as a subsequent stage on the results of IE itself.

If we think along these lines we see that the first distinction of this paper, between traditional IR and the newer IE , is not totally clear everywhere but can itself become a question of degree. Suppose parsing systems that produce syntactic and logical representations were so good, as some now believe, that they could process huge corpora in an acceptably short time. One can then think of the traditional task of computer question answering in two quite different ways. The old way was to translate a question into a formalised language like SQL and use it to retrieve information from a database- as in "Tell me all the IBM executives over 40 earning under $50K a year". But with a full parser of large corpora one could now imagine transforming ing the

query to form an IE template and searching the WHOLE TEXT (not a data base) for all examples of such employees---both methods should produce exactly the same result starting from different information sources --- a text versus a formalised database.

What we have called an IE template can now be seen as a kind of frozen query that one can reuse many times on a corpus and is therefore only important when one wants stereotypical, repetitive, information back rather than the answer to one-off questions.

"Tell me the height of Everest?", as a question addressed to a formalised text corpus is then neither a IR nor IE but a perfectly reasonable single request for an answer. "Tell me about fungi", addressed to a text corpus with an IR system, will produce a set of relevant documents but no particular answer. Tell me what films my favourite movie critics likes, addressed to the right text corpus , is undoubtedly IE as we saw, and will produce an answer also. The needs and the resources available determine the techniques that are relevant, and those in turn determine what it is to answer a question as opposed to providing information in a broader sense.

A final issue is the future relationship of IE to text summarization, a functionality for which there is a clearly established need. In some sense the data base provided by an IE system can be sen either as pure data, for subsequent mining, or as a base from which a summary could be generated in a natural language, which could be the original language on which the IE operated or another one. This is IE as the basis of a text summarization system; however, there are quite other techniques already near market, (a typical one would be British Telecom's), which often rely on summaries derived from individual sentences pieced together from the original text, deemed significant sentences on broadly statistical grounds. [Jim, mention others???] This technique has the advantage of not needing templates, but the disadvantage that such sentences may or may not form a coherent text.

Moral

One moral from all this, and which is important to keep in mind with the advent of speech research products and the multimedia associated with the Web, is that most of our cultural, political and business patrimony is still bound up with texts, from manuals for machines, to entertainment news to newspapers themselves. The text world is vast and growing exponentially: one should never be seduced by multi-media fun into thinking that text and how to deal with it, how to extract its content, is going to go away.

References

Basili, R., Pazienza, M., & P. Velardi, (1993) Acquisition of selectional patterns in sub - languages. Machine Translation, 8.

Cunningham, H., Gaizauskas, R. & Y. Wilks, (1995) GATE: a general architecture for text extraction, University of Sheffield, Computer Science Dept. Technical memorandum.

Dorr, B. & D. Jones (1996) The role of word-sense disambiguation in lexical acquisition: predicting semantics from syntactic cues, Proc. COLING96.

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

Kilgarriff, A. (1993) Dictionary word-sense distinctions: an enquiry into their nature. Computers and the Humanities, 26.

Cowie, J. & W. Lehnert (1996) Information Extraction, in (Y. Wilks, ed.) Special NLP Issue of the Comm. ACM.

Levin, B. (1993) English verb classes and alternations, Chicago, IL.

Procter, P. et al. (1994) The Cambridge Language Survey Semantic Tagger. Technical Report, Cambridge University Press.

Pustejovsky, J. & Anick, P. (1988) On the semantic interpretation of nominals. Proc. COLING88.

Riloff, E. & J. Shoen (1995) Automatically acquiring conceptual patterns without an annotated corpus, Proc. Third Workshop on Very Large Corpora.

Wakao, T., Gaizauskas, R. & Wilks, Y.. (1996), Evaluation of an algorithm for the recognition and classification of proper names. Proc. COLING96.

Wilks, Y. (1978) Making preferences more active, Artificial Intelligence, 11.

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

Wilks, Y. (in press) Senses and Texts, Computational Linguistics.