Research Memoranda CS-01-13
To be published in IEEE Intelligent Systems and Their Applications, November 2001 (Trends and Controversies)

Challenges in Information Extraction from Text for Knowledge Management

Fabio Ciravegna
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, S1 4DP, Sheffield, UK.
F.Ciravegna@dcs.shef.ac.uk

Nowadays large part of knowledge is stored in an unstructured textual format. Texts cannot be queried in simple ways and therefore the contained knowledge can neither be used by automatic systems, nor be easily managed by humans. The traditional process of manual knowledge identification and extraction by knowledge engineers used in KM is a complex time consuming process, as it requires a great deal of manual input. An example of such process is the collection of interviews to experts ('protocols') and their analysis by knowledge engineers in order to codify, model and extract the knowledge of an expert in a particular domain. In this context, Information Extraction from texts (IE) is one of the most promising areas of HLT for Knowledge Management (KM) applications. IE is an automatic method for locating important facts in electronic documents, e.g. information highlighting for document enrichment or for information storing for further use (such as populating an ontology with instances). IE as defined above is the perfect support for Knowledge Identification and Extraction as it can – for example - provide support in protocol analysis either in an automatic way (unsupervised extraction of information) or semi-automatic way (e.g. helping knowledge engineers locating the important facts in protocols, via information highlighting).

It is widely agreed that the main barrier to the use of IE is the difficulty in adapting IE systems to new scenarios and tasks, as most of the current technology still requires the intervention of IE experts. This makes IE a technology difficult to apply, because personnel skilled in IE are difficult to find in industry, especially in small medium enterprises [Ciravegna 2000]. A main challenge for IE for the next years is to enable people with knowledge of Artificial Intelligence (e.g. knowledge engineers) but no or scarce preparation on IE and Computational Linguistics to build new applications/cover new domains. This is particularly important for KM: IE is just one of the many technologies to be used in building complex applications: wider acceptance of IE will come only when IE tools will not require any specific skill apart from notions of KM.

A number of Machine Learning based tools and methodologies are emerging [Freitag 1999, Ciravegna 2001], but the road to fully adaptable and effective IE systems is still long. In this paper, I will focus on two main challenges for adaptivity in IE for KM that in my opinion are paramount in the current scenario:

1. Automatic adaptation to different text types
2. Human-centred issues in copying with real users.

Adaptivity to Different Text Types

Porting IE systems means coping with four main tasks:

1. Adapting to the new domain information: implementing system resources such as lexica, knowledge bases, etc. and designing new templates, so that the system is able to manipulate domain-specific concepts;

2. Adapting to different sublanguages features: modifying grammars and lexical so to enable the system to cope with specific linguistic constructions that are typical of the application/domain;
3. Adapting to different text genres: specific text genres (e.g. medical abstracts, scientific papers, police reports) may have their own lexis, grammar, discourse structure.
4. Adapting to different types: Web-based documents can radically differ from newspaper-like texts: we need to be able to adapt to different situations

Most of the literature on IE has focused so far on issues 1, 2 and 3 above, with limited attention to text types, having focused mainly on free newspaper-like texts [Cardie 1997]. This is a serious limitation for portability, especially for KM, where an increase in the use of inter/intranet technologies has moved the focus from a free texts only scenarios (based on e.g. reports and protocols) to more composite scenarios including (semi)structured texts (e.g. highly structured web pages as produced by data bases). In classical Natural Language Processing (NLP) adapting to new text types has been generally considered as a task of porting across different types of free texts. The use of IE for KM requires an extension of the concept of text types to new, unexplored dimensions. As a matter of fact linguistically-based methodologies used for free texts can be difficult to apply or even ineffective on highly structured texts such as web pages produced by data bases. They are not able to cope with the variety of extralinguistic structures (e.g. HTML tags, document formatting, and stereotypical language) that are used to convey information in such documents. On the other hand, wrapper-like algorithm designed for highly structured HTML pages are largely ineffective on unstructured texts (e.g. free texts). This is because such methodologies make scarce (or no) use of NLP, tending to avoid any generalization over the flat word sequence tending to be ineffective on free texts, for example because of data sparseness [Ciravegna 2001b].
The challenge is developing methodologies able to fill the gap between the two approaches in order to cope with different text types. This is particular important for KM with its composite Web-based scenarios, as Web pages can actually contain documents of any type and even a mix of text types, e.g., an HTML page can contain both free and structured texts at the same time. Work on this topic has just started. Wrapper induction systems based on lazy NLP [Ciravegna 2001b] try to learn the best (most reliable) level of language analysis useful (or effective) for a specific IE task by mixing deep linguistic and shallow strategies. The learner starts inducing rules that make no use of linguistic information, like in wrapper-like systems. Then it progressively adds linguistic information to its rules, stopping when the use of NLP information becomes unreliable or ineffective. Linguistic information is provided by generic NLP modules and resources defined once for all and not to be modified by users to specific application needs. Pragmatically, the measure of reliability here is not linguistic correctness (immeasurable by incompetent users), but effectiveness in extracting information using linguistic information as opposed to using shallower approaches. Unlike previous approaches where different algorithm versions with different linguistic competence are tested in parallel and the most effective is chosen [Soderland 2000], lazy NLP-based learners learn which is the best strategy for each information/context separately. For example they may decide that using parsing is the best strategy for recognising the speaker in a specific application on seminar announcements, but not the best strategy to spot the seminar location or starting time. This is very promising for analysing documents with mixed genres, e.g. web pages

containing both structured and unstructured material, quite a common situation in web documents.

Coping with Non IE Experts

The second main task in adaptive IE concerns human-computer interaction during application development. Non-expert users need to be supported during the whole adaptation process so to maximize effectiveness and appropriateness of the final application. A typical IE application lifecycle is composed of: scenario design, system adaptation, results validation and application delivery [Ciravegna and Petrelli 2001].

Scenario design is the task of defining the information to extract. Large part of potential users need specific support, as they may find difficult to manipulate IE-related concepts, such as templates. Moreover there may be a gap between what information the user needs, what information the texts contain and what the system can actually extract. It is very important to help users recognize such discrepancies, forcing them into the right paradigm of scenario design. Highlighting information in different colors is generally a good way to do it. Tagging-based interfaces, such as Mitre's Alembic, have proven to be quite effective and have become a standard in adaptive IE.

Selecting the corpus to be tagged for training is also a delicate issue. Non-linguistically aware users tend to focus on text content rather than on linguistic variety. Unfortunately IE systems learn from both. Provided corpora may be unbalanced wrt types or genres (e.g. emails could be underrepresented wrt free texts), or even show peculiar regularities due to wrong selection criteria. For example in designing an application on IE from professional resumees, our user selected the corpus by using the names of Us cities as keywords. When the system was tested in the real world environment it became clear that most of the resumees were actually originating from Europe, where addresses, titles of study and even text style can be very different from the American ones. The resulting system was therefore largely ineffective and left the user dissatisfied with the final application. A number of methodologies can be used to validate the training corpus wrt a (hopefully big) untagged corpus. One possible validation concerns the formal comparison of training and untagged corpus. [Kilgarriff 2001] proposes heuristics for discovering differences in text types among corpora. Average text length, distributions of HTML tags and hyperlinks in web pages, average frequency of lexical classes in texts (e.g. nouns), etc. can be relevant indicators of corpus representativeness and can be used to warn inexperienced users that some training corpora can be not representative enough of the whole corpus. Even the detection of an excess of regularity in the training corpus can be a good indicator of an unbalanced corpus selection, e.g. if a high percentage of fillers for some slots is the same string.

With a corpus reasonable in size and quality, the IE system can then be trained. Unfortunately, even the best algorithm is unlikely to provide optimised results for specific use. This is because a 100% accurate system is out of reach of the current IE technology, and therefore there is the necessity of balancing recall (i.e. ability to retrieve information when present) and precision (the ratio of correct information on the total of information extracted) so to produce the optimal results for the task and users at hand. Different uses will require different types of results (e.g., higher recall in some cases, higher precision in others). Users must be enabled to evaluate results from both a quantitative and qualitative point of view, and to change the system behaviour if necessary. Most of the current technology provides satisfying results for

results inspection: tools such as the MUC scorer [Douthat98] allow users to understand the system effectiveness in details. The challenging step is now to enable users to change the system behaviour. In case of occasional or inexperienced users, the issue arises of avoiding the use of technical or numerical concepts (such as precision and recall). This requires the ability from the IE system of bridging the user's qualitative vision ("you are not capturing enough information") with the numerical concepts the learner is able to manipulate (e.g. moving error thresholds in order to obtain higher recall).

When the application is tuned to the specific user needs, it can be delivered and used in the application environment. Corpus monitoring should be enabled even after delivery, though. One of the risks in highly changing environments such as the Internet is that information (e.g. web pages) can change format in a very short time, and the system must be able to detect such changes [Muslea 2000]. The same techniques mentioned above for testing corpus representativeness can be use to identify changes in the information structure or test type.

For further discussion on human-centred issues, see [Ciravegna and Petrelli 2001].

Conclusion

Adaptive IE is already providing useful results and technology for KM. Fully integrated user-driven solutions are still to come, but current research results are promising. In this essay I have discussed two of the major challenges for IE for the next years, namely adaptivity to text genres and human-centred issues that are paramount for an effective use of IE for KM purposes.

Acknowledgement

My work on adaptive IE is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01). AKT comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University (www.aktors.org).

Bibliography

[Cardie 1997] Claire Cardie, `Empirical methods in information extraction', *AI Journal*, 18(4), 65-79, 1997.

[Ciravegna *et al.* 2000] Fabio Ciravegna, Alberto Lavelli, and Giorgio Satta, `Bringing information extraction out of the labs: the Pinocchio Environment', in *ECAI2000, Proc. of the 14th European Conference on Artificial Intelligence*, ed., W. Horn, Amsterdam, 2000. IOS Press.

[Ciravegna 2001a] Fabio Ciravegna: "Adaptive Information Extraction from Text by Rule Induction and Generalisation" in *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, August 2001."

[Ciravegna 2001b] Fabio Ciravegna: "$(LP)^2$, an Adaptive Algorithm for Information Extraction from Web-related Texts" in *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* held in conjunction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001

[Ciravegna and Petrelli 2001] Fabio Ciravegna and Daniela Petrelli: "User Involvement in Adaptive Information Extraction: Position Paper" in *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* held in conjunction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001

[Douthat 1998] Aaron Douthat, `The message understanding conference scoring software user's manual', in *the 7th Message Understanding Conf.*, www.muc.saic.com

[Freitag 1998] Dayne Freitag, `Information Extraction from HTML: Application of a general learning approach', *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[Kilgarriff 2001] A. Kilgarriff, 'Comparing Corpora', to appear in Corpus Linguistics.

[Muslea 2000] I. Muslea, S. Minton, and C. Knoblock, 'Co-testing: selective sampling with redundant views' in *Proc. of the 17th National Conference on Artificial Intelligence, AAAI-2000*.

[Soderland 1999] Steven Soderland, `Learning information extraction rules for semi-structured and free text', *Machine Learning*, (1), 1-44, 1999.

Dr Fabio Ciravegna is Senior Research Fellow at the Department of Computer Science of the University of Sheffield, UK. His research interest includes Classical and Adaptive Information Extraction from text (IE), with a particular focus on user-centered methodologies. He has been active in IE since 1988. Since then he was principal investigator and coordinator for Information Extraction at Fiat Research Center (Turin, Italy) from 1991 to 1993 and at ITC-Irst (Trento, Italy) from 1993 to 2000. His work on IE has been published in a number of international conferences such as EACL-1999, IJCAI-1999, ECAI-2000 and IJCAI2001. He co-organized two workshops on the use of Machine Learning for Information Extraction at ECAI-2000 and IJCAI2001. Contact him at F.Ciravegna@dcs.shef.ac.uk