Mixing technologies for Intelligent Information Extraction

Thierry Poibeau
Mail: Thierry.Poibeau@lcr.thomson-csf.com

Thomson-CSF
Laboratoire Central de Recherches
Domaine de Corbeville
91404 Orsay cedex

Laboratoire d'Informatique de Paris-Nord
Institut Galilée, Université Paris-Nord
av. J.-B. Clément
93430 Villetaneuse

Abstract

This paper presents an Information Extraction (IE) system. This kind of system is intended to extract structured information from general texts. An evaluation is performed and the results are discussed. We show that, if IE is now an established technology, it suffers a number of limitations that prevent its dissemination through general public applications. To get over this obstacle, systems have to acquire domain-specific extraction patterns. This task makes IE being related to knowledge acquisition and corpus linguistics. We present the European ECRAN project and make some original proposals implicating Topic Detection, Lexical Tuning and Intelligent Interfaces in this framework. We show that an interaction is necessary between these different components and the end-user in order to incrementally increase the overall performances of the system.

1 Introduction

Information Extraction (IE) is a technology dedicated to the extraction of structured information from texts to fill pre-defined templates [Pazienza, 1997]. The American Message Understanding Conferences (MUC) provided a formidable framework for the development of research in this area. IE is known to have established a new linguistic architecture based on cascading automata and domain-specific knowledge [Appelt *et al.*, 1993]. Applications concern the extraction of information from newswires on terrorism events, job employment or TV programs for example.

However, even if IE seems to be now a relatively mature technology, we will see from our experience that it suffers a number of yet unsolved problems that limit its dissemination through industrial applications. Among these limitations, we can consider the fact that systems are not really portable from one domain to another. Additionally, systems analyze very homogeneous corpora and performances decrease rapidly if one gives to the system a wide variety of texts as input. It is necessary to find new approaches allowing people to glance through the texts and interact with the system to find rapidly the information they need even on new domains. This trend is characterized by several research programs, including Topic Detection to pre-filter the texts, Lexical Tuning to adapt the system resources to new domains and corpora, and Intelligent Interfaces to navigate through the text. We show that the system resources can be assimilated to domain-specific knowledge acquired from the corpus to feed the system.

This study presents the experiments made in the framework of the ECRAN project, a classical IE system. A precise evaluation of the system is presented. Then, some proposals are made to enhance the results by mixing different technologies into IE systems.

2 ECRAN: a Multilingual Information Extraction System

ECRAN[1] is a EU-funded project in lexically-driven Information Extraction that attempts to offer generic and portable IE systems. This kind of systems fills a predefined structured template from a text on a given domain [Pazienza, 1997].

2.1 A brief description of the system

In the framework of ECRAN, systems have been developed for French, English and Italian. Their various modules provide at first a very local analysis that is progressively extended to the near context in which appears an expression. All the modules for French are integrated through the platform for linguistic engineering GATE, from the University of Sheffield [Cunningham, 1996].

For example, from the following text relating a terrorism event in Spain:

[1] ECRAN is a European project from the Language Engineering framework (LE 2110).

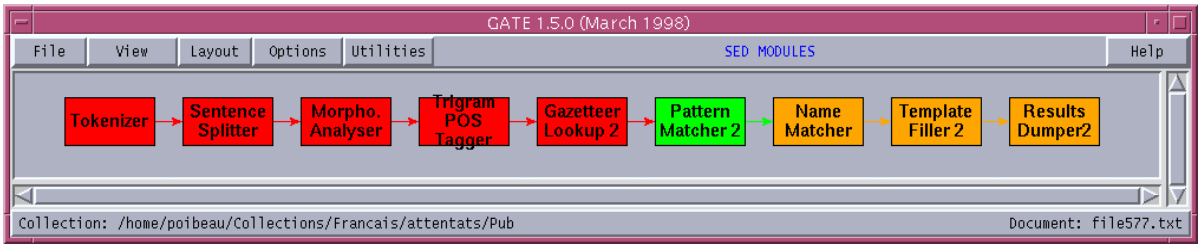

Figure 1: *the Extraction System for French*

22 août 1990, page 24

ESPAGNE un mort au cours d'un attentat à la voiture piégée au Pays Basque.

- Une personne a été tuée mardi 21 août en milieu de journée à Oyarzun (province basque de Guipuzcoa) au cours de l'explosion d'une voiture piégée dans le parking d'un hypermarché, a indiqué la police. (AFP.)[2]

a template will be produced containing the main information from the text:

Event date: 22 août 1990
Event location: Espagne, Pays Basque
Number of killed people: 1
Number of injured people: 0
Weapon: voiture priégée

The system can be divided into four parts: modules for shallow parsing, modules for named entity recognition (location, dates, etc), modules extracting information from the structure of the text and then modules to complete a result template.

- The shallow parsing stage produces a tagged text, with morphological and POS information. The tools producing these analyses have been developed outside from ECRAN: they are generic and offer common morpho-syntactic annotations over the document.

- The second stage is concerned with the recognition of pertinent information by means of a linguistic analysis. This stage allows to recognize the named entities (person names, organizations, locations and dates) and other related information (number of victims, weapon…). This step is achieved in applying a grammar of regular expressions, the extraction patterns, over the text. For example, with the expression `attentat à le <WEAPON>`, we are able to label *voiture piégée* as being the weapon of the terrorism event.

- Some information is also extracted from the structuration of the text. [Lacroix *et al.* 98] presents a system able to extract information from the structure of HTML pages. This kind of application will increase with the development of more structured document (like XML documents). Wrapper factories go one step beyond, in connecting together distant pieces of texts and in extracting information from poorly structured documents [Sahuguet and Azavant 98]. Given that the archives of *Le Monde* newspaper are formatted, we developed a wrapper to automatically extract information about the location and the date of the event. This non-linguistic extraction increases the quality of the result by providing 100% good results. It is also accurate when one thinks of the current development of structured text (HTML, XML) via the web and other corporate networks.

- The last stage links all these information together to produce a result template that presents a synthetic view of the information extracted from the text. Partial templates produced by different sentences are then merged to produce only one template per text. This merging is done according to some constraints on what can be unified or not. The results can then be stored in a database which exhibit knowledge extracted from the corpus.

2.2 Corpus and experiment description

An experiment has been done on articles dealing with terrorism events from the French newspaper *Le Monde*. We gave an example from this corpus in the section before. The IE system in French was run on a part of this corpus from *Le Monde* daily newspaper issued between January 1987 and December 1996. The full corpus is made of more than 5833 texts, containing more than 2 million words, totaling 13.5 million characters. About 1000 texts have been processed to elaborate the system and ensure its robustness. The evaluation was made on 50 new texts from the corpus (these texts have not been used during the elaboration of the extraction patterns).

Each text is considered to produce one template. The result is compared with the results of a hand-tagged collection of texts [Chinchor *et al.*, 1993]. The comparison may be equality, inclusion or fail. From these

[2] This is the translation of the text:

August 22th, 1990, page 24

SPAIN: one person killed in a booby-trapped car attack in the Basque Country

One person was killed on Tuesday August 21st during the day in Oyarzun (Basque province from Guipuzcoa) in the explosion of a booby-trapped car on the parking of a supermarket, indicated the police. (AFP)

indicators, we can define some equations to evaluate recall and precision in a way close to what is done for Information Retrieval [Van Rijsbergen, 1979].

The following table shows the results in terms of precision and recall. We differentiate the information obtained by linguistic analysis from the information obtained via the structure of the text.

	Structure of the text (1)	Linguistic Analysis (2)	Total
PRECISION	**1**	**0.89**	**0.94**
RECALL	**1**	**0.63**	**0.78**

Table 1: Evaluation results.

Column (1) reflects the scores concerning the quality of the information acquisition process from the structure of the text; column (2) the quality of the information acquisition process by linguistic analysis.

The date and the location of the event are extracted from the structure of the text and then are evaluated separately. Precision and recall is 1 because of the systematic structuration of these texts, which yields to systematically good results.

We evaluate then the information obtained by linguistic analysis. The system must give one and only one value for the weapon and the number of victims (injured and death persons). If the number of victims does not correspond to the exact value (PART), we do not consider it as good. Precision is 0.89. This good result is due to the fact that we are looking for a local information with very constraint extraction patterns. Recall is 0.63. Coverage of the system could be enhanced by the addition of new extraction patterns. Additionally, a larger syntactic analysis combined with the calculation of a logical form would allow parsing beyond modifiers and other complements and would lead to better results.

3 Some limitations of the current IE systems

The fact that MUC systems, developed in research laboratories, are still away from industrial applications in spite of good results is a problem we should have in mind [Wilks, 1997] [Grishman, 1997]. The appropriateness of the systems in comparison with user needs is a dimension that must be taken into account during evaluation. Neither our evaluation nor those made in the Message Understanding Conferences raised this aspect.

The named entity recognition is one of the generic tasks identified by the MUC Conferences [MUC-6, 1995]. This task is particularly useful for economic and strategic intelligence. In these domains, it is important to pinpoint names of persons and firms, dates and location. For some other domains or corpora, named entities are not so much important. For example, contrary to what was found in the MUC-5 corpus concerning terrorism events [1993], the corpus from *Le Monde* on the same subject has no personality as a victim. Victims are anonymous, sometimes named by their function (Prime Minister, policeman, civilian, …) and almost never by their name. During the experiment on *Le Monde* corpus, we gave up searching person names given that it was a pointless information for our corpus. It is then maybe necessary to question the architecture of MUC systems and adapt it, according to the task. For example, in our case, the module for named entity recognition is maybe not necessary to analyze terrorism events. On the contrary, we developed a named entity recognition system that can be distributed alone and corresponding to specific industrial needs (economic and strategic intelligence).

IE Systems must have as input very homogeneous corpora to give accurate results. Texts must have the same genre and must contain the information the system is looking for. During our experiment on the terrorism events, the corpus was established by selecting all the texts indexed by the `attentat` keyword in *Le Monde* archive. But then the corpus contains some texts that are not news relating terrorism event strictly speaking, but investigations, judgements, threats, … MUC systems try to extract information from the texts independently from their genre or characteristics. So, if one tries to give non-pertinent texts as input for a system, he will obtain, at best, no answer from the system and, at worse, some noise (information extracted by extraction pattern recognized accidentally). It is then necessary to elaborate, on the one hand, efficient filtering tools and, on the other hand, tools allowing to come back to the text to check the information proposed by the system. Some of these tools exist but their integration is not obvious.

The performances of the systems are such that people must have access to the texts to verify the extracted information. For the named entity task, results are over 90% and can be compared to the results obtained with a hand-tagged corpus. However, results decrease down to a precision of 70% on the most complex tasks (*relation template* and *template scenario*). These measures are good with regard to the complexity of the analysis they involve (anaphora resolution, complex expression extraction, syntactic and semantic transformations, parsing, …). On the other hand, if the extracted information is presented through a database, the context is lacking and information is then taken for granted. Search engines do not take risks: they propose a set of texts answering a request (generally ranked according to a score). However, classical search engines have shown their own limits: everybody knows the difficulty to find the good terms for a given request, obtain pertinent results and find the good information into the amount of possible pertinent documents. This task can only be achieved by a cooperation of different corpus-oriented techniques able to dynamically and interactively process important amount of texts.

4 Mixing technologies for IE

One of the main aims of the ECRAN project was to develop portable systems from one domain to another. We examine below several interesting techniques which

are needed for an efficient corpus-based IE system. The examples are taken from experiments made in the framework of ECRAN and other related research programs.

• *Efficient pre-filtering tools*. Current systems process very homogeneous corpora. These systems must now be integrated in general-purpose applications facing heterogeneous corpora. Some tools like Information Retrieval engines can provide consistent input to IE systems.

A tool providing homogeneous semantic classes has been developed for French. This tool takes as input a general dictionary and a sample of the corpus to generate high quality clusters based on an algorithm from A. Luk [1995].

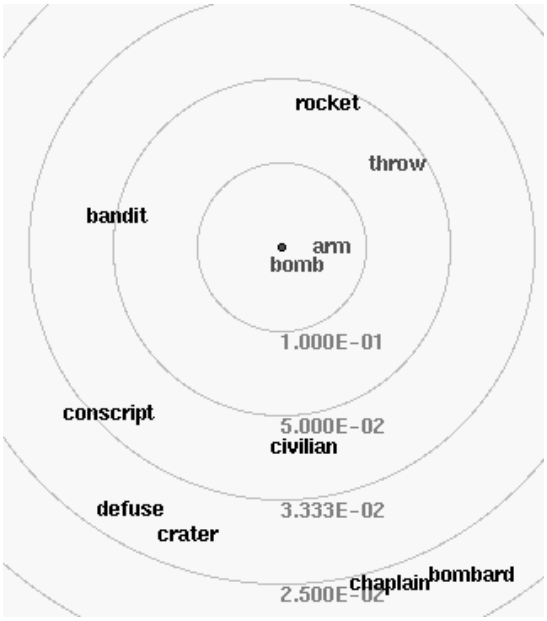

Figure 2: *a cluster generated automatically around the notion of* `bomb`

Figure 2 presents one of the clusters generated from the terrorism event corpus. These clusters allow to filter texts given some predefined subjects of interest for the end-user.

From this point of view, the Topic Detection and Tracking program gives an interesting framework, even if it concerns particular events rather than thematically related stories [Allan, 1998]. This program allows to cluster texts from their main topics to offer a homogeneous input to the IE system. The techniques developed also makes it possible to detect new subjects in an on-line stream of texts or retrospectively process the corpus.

• *Semi-automatic definition of extraction patterns*. To adapt an IE system to a new corpus, one has to customize the system resources as well as to create new template specifications and template-filling rules. In ECRAN the customization of system resources involves mainly the creation of new extraction patterns to cover the domain (see section 2.1.).

In this context, a pattern generator has been developed which generates semi-automatically domain-specific pattern rules [Karkaletsis *et al.* 97]. This tool is based on some measures concerning verb frequency and on specific clustering algorithms. It uses knowledge from outside resources like WordNet and the Longman Dictionary of Contemporary English (LDOCE), so that only an English version is available. Some syntactic schemas are acquired from the corpus and proposed to the end-user via an interface. A simple mechanism allows then to go from this corpus-based analysis to some more abstract rules which are able to recognize pertinent complex expressions through the text, including syntactic transformations. Experiments have been done on a financial corpus, showing that about 50% of the entries of the original dictionary could retrieved. The acquisition of new senses for a given verb is more difficult, simply because these senses can hardly be inferred from the text [Basili *et al.*, 98]. This case shows that lexical tuning and definition of extraction patterns cannot be fully automatic: an interaction with the user is still a necessity.

• *Hypertext consultation interface*. Interfaces of systems are often neglected because they are considered to be outside the scope of research. But recent developments have proven that experts need to have convivial interfaces in order to be able to validate the results produced by NLP systems [Benaki 97].

For IE, there is clearly a need for the end-user to go back to the text and verify the automatically extracted information. ECRAN has shown that, if people need, in certain strategic domains, to have rapid access to information, they also need to have nearly 100% good results. Since this objective is not realistic for most of current NLP systems, the solution is clearly in user-oriented tools. Markup languages such as SGML and now XML allow these links between information and texts. We developed hypertext interfaces in which the end-user can access the good information from the text itself. For example, clicking on a company name in a text makes appear a window in which the end-user finds useful information about this company, its benefits, its interests in other companies... Clicking on a button in the section “benefits” of this last window allows to go back to the texts speaking of the benefits of this company, etc.

We have done some experiments concerning economic intelligence: in this domain, people need to have a rapid access to company and person names, to dates and locations. Named entity recognition tools are then particularly useful. This study revealed that people do not need to extract information from texts as much as they need to have access at a glance to the main information contained in texts. Highlighting texts and linking them via a hypertext consultation interface is then a solution to allow the user have rapid access to information and check if a text is pertinent or not.

5 Conclusion

We have presented some limitations of current IE systems, from our experience in the ECRAN project. We have seen that, in spite of good results, systems must evolve. They must be integrated to other NLP tools providing intelligent access to text by an interaction with the end-user. In addition, they must be portable and adaptable to new domains and user requirements.

We examined some new trends that are essentially corpus-oriented. Traditional "MUC systems" were based on a very local analysis in reaction to understanding systems from the eighties. These systems were processing the whole text: this strategy was leading to local errors that were crucial for the task. However, more recent research, especially in the field of corpus linguistics, revealed that corpus analysis help to bring out the specific linguistic characteristics of a given corpus. This is particularly the fact in IE, where machine learning and stochastic methods are now widely used.

These technologies provide some means for an assisted acquisition of domain-specific knowledge, which was until recently a time-consuming task. Topic Detection is a means to extract pertinent pieces of text from a corpus, Lexical Tuning to adapt resources to the corpus and Intelligent Interfaces to control the results provided by the system. Thus, we should see, in the near future, systems based on IE techniques able to answer to very specific questions coming from the end-user.

6 References

[Allan *et al.*, 1998] James Allan, Jaime Carbonnel, George Doddington, Jonathan Yamron and Yiming Yang. Topic Detection and Pilot Study (Final report). *Proceedings of the DARPA broadcast news transcription and understanding workshop*, February 1998.

[Appelt *et al.*, 1993] Douglas E. Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyana and Mabri Tyson. FASTUS: a finite-state processor for information extraction from real-world text. *Proceedings of IJCAI'93*, pages 1172-1178, 1993. International Joint Committee on Artificial Intelligence.

[Benaki 97] Benaki, Vangelis Karkaletsis and Costas Spyropoulos. "Integrating User Modeling into Information Extraction: the UMIE Prototype". *Proceedings of the 6th International Conference on User Modeling (UM97)*, CISM No 383, Springer Wien New York, pp. 55-58, 1997.

[Basili *et al.*, 98] Roberto Basili, Roberta Catizone, Maria Tereza Pazienza, Mark Stevenson, Paola Velardi, Michele Vindigni and Yorick Wilks. An empirical approach to Lexical Tuning. *Workshop on Adapting lexical and corpus resources to sublanguages and applications*, Granada, Spain, May 1998.

[Chinchor *et al.*, 1993] Nancy Chinchor, Lynette Hirschman and David Lewis. Evaluating Message Understanding Systems: an analysis of the third Message Understanding Conference (MUC-3*), Computational Linguistics*, 19(3), 1993.

[Cunningham *et al.*, 1996] Hamish Cunningham, Kevin Humphreys, Rob Gaizauskas and Yorick Wilks. GATE: a general architecture for text engineering. *Proceedings of Computational Linguistics (COLING'96)*, Copenhagen, Denmark, August 1996.

[Grishman, 1997] Ralph Grishman. Information Extraction: techniques and challenges. *Information Extraction* (MT Pazienza ed.), Springer Verlag (Lecture Notes in computer Science), Heidelberg, Germany, 1997.

[Grishman and Sundheim, 1997] Ralph Grishman and Beth Sundheim. Message understanding conference-6, a brief history. *Proceedings of Computational Linguistics (COLING'96)*, Copenhagen, Denmark, August 1996.

[Karkaletsis *et al.*, 1997] Vangelis Karkaletsis, Costas Spyropoulos, and E. Benaki. Customising Information Extraction Templates according to Users Interests. *Proceedings of the International Workshop on Lexically Driven Information Extraction (LDIE'97)*, pp. 23-38, Frascati, Italy, July 1997.

[Lacroix et al. 98] Zoé Lacroix, Arnaud Sahuguet and Raman Chandrasekar. Information Extraction and Database Techniques: A User-Oriented Approach to Querying the Web. *Conference on Advanced Information Systems Engineering*, 1998.

[Luk, 1995] Alpha K. Luk. Statistical Sense Disambiguation with Relatively Small Corpora using Dictionary Definitions. *Proceedings of the 33rd Annual Meeting of the ACL*, July 1995.

[MUC-5, 1993] DARPA. *Proceedings Fifth Message Understanding Conference*. Morgan Kaufmann Publishers, San Francisco, 1995.

[MUC-6, 1995] DARPA. *Proceedings Sixth Message Understanding Conference*. Morgan Kaufmann Publishers, San Francisco, 1995.

[Pazienza, 1997] Maria Tereza Pazienza (éd.). *Information extraction (a multidisciplinary approach to an emerging information technology), International Summer School SCIE'97. Frascati, Italy, July 1997.* Springer Verlag (Lecture Notes in computer Science), Heidelberg, Germany, 1997.

[Sahuguet and Azavant] Arnaud Sahuguet and Fabien Azavant. W4F: a WysiWyg Web Wrapper Factory, *Technical report from the Penn Database Research Group*, University of Pennsylvania.

[Van Rijsbergen, 1979] CJ Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[Wilks, 1997] Yorick Wilks. Information Extraction as a core Language Technology. *Information Extraction* (MT Pazienza ed.), Springer Verlag (Lecture Notes in computer Science), Heidelberg, Germany, 1997.