Tools for Extracting and Structuring Knowledge from Texts.

Authors: Antoine Ogonowski* (Antoine.Ogonowski@erli.gsi.fr), Marie Luce Herviou*** (Marie-Luce.Herviou@der.edf.fr) and Eva Dauphin** (sylvie.regnier@siege.aerospatiale.fr)

The authors wish to thank M. Bernard*, G Clémencin*, S. Lacep*, R. Leblond**, MG. Monteil***, G. Morize* and B. Normier* for their collaboration while working on this project and precious advice for the preparation of this note.

* GSI-Erli 1 pl. des Marseillais , 94227 Charenton Cedex FRANCE
** AEROSPATIALE CCR 12 rue Pasteur, BP76, 92152 Suresnes Cedex FRANCE
*** Electricité de France (EDF), Research Center, 1 Av du Général de Gaulle, 92141 Clamart FRANCE

Abstract : We demonstrate an approach and an accompanying UNIX toolbox for performing various kinds of Knowledge Extractions and Structuring. The goal is to "practically" enhance the productivity while constructing resources for NLP systems on the basis of large corpora of technical texts,. Users are lexicon/grammar builders, terminologists and knowledge engineers. We stay open to already explored methods in this or neighbouring activities but put a greater stress on the use of linguistic knowledge. The originality of the work presented here lies in the scope of applications addressed and in the degree of use of linguistic knowledge.

1. Introduction

Since NLP has started moving from toy problems to real applications one of the biggest difficulty has been Knowledge Acquisition (KA) of different types (lexical, grammatical, domain and application specific). A lot of the needed information resides (in an implicit or explicit form) in texts that most of the time now exist in machine readable form.

It seems however too difficult to fully automate the KA process although steps have to be taken in that direction [WIL93]. Tools to help users deal with large corpora have been developed for some time, most of these however rely either on crude non linguistic approaches or mostly on statistic methods (eg [CAL90]). Credit should also be given to new approaches based on neural nets especially for dealing with Machine Readable Dictionaries (eg [IDE90]).

This project note illustrates a more linguistic and "open" approach (not ignoring the achievements of existing methods), basing itself on existing large electronic dictionaries compatible with the GENELEX model [GEN90].

The EUREKA project GENELEX (5 years, 39 MEcu, 250 man years) has produced public models encompassing morphological, syntactic and semantic knowledge in both monolingual (French, English and soon Italian and Portuguese) as well as bilingual contexts.

The tools that the authors want to discuss here are developed in the context of the closely related EUREKA project GRAAL[1] (acronym for Grammars that can be Reused to Automatically Analyse Languages). This project [GRA92] has the following objectives:

- the development of grammars that are easily maintainable and reusable (ie different types of NLP applications can be built on their basis)
- the development of tools (parsers, generators as well as workbenches for grammar construction, customisation and integration in specific application environments)
- delivery of industrial level applications.

This 4 year project is currently divided into several subprojects ("SPs") one of which is called "KES" (Knowledge Extraction and Structuring) and aims at the second and third types of GRAAL goals.

The three partners of this mentioned SP: EDF, AEROSPATIALE and GSI-Erli have built a modular extensible toolbox that should cover most of the needs that may occur in "any" knowledge extraction process and now validate the performance of the toolbox on several applications.

For the partners, Knowledge Extraction covers needs arising in various types of applications ranging from "terminology construction and enrichment" (problem largely studied these years [TER90], [LEX92] ...), "extension of lexicons coverage", through "grammar development" up to "construction of Large Knowledge Bases" for AI systems, or for technology assessment survey purposes. This means that potential users of the toolbox range from terminology experts, lexicon and grammar writers to knowledge engineers.

Two languages are currently considered: French and English, but the tools developed should be easily adaptable to other languages

[1] This 23MEcu, 150 man years project is conducted by an international consortium currently gathering in France: GSI-Erli (project leader), EDF, Aérospatiale and IRIT; in Italy: IRST, Centro Ricerche FIAT; in Switzerland: ISSCO; in Greece: ILSP; in Finland: Lingsoft, Nokia; in Portugal: ILTEC.

2. The Approach

Rather than developing a new automatic KA theory we have opted for a "practical" approach i.e. a set of tools that can assist the user in a bootstrapping process.

2.1 Principles: Our platform integrates all the resources and processes allowing to proceed from raw texts to a structured set of knowledge items (taken to mean words, terms, concepts, links, rules etc.) extracted from these texts.

Partners believe that the future industrial tools are to use much more linguistic knowledge than the tools currently available on the market (eg [SAT92]). Our goal is not to be 100% exact at the different stages of processing but to help the user rapidly explore various hypotheses.

2.2 Phases: Three main phases organise the KES process: "*Corpus Characterisation*", "*Extraction*" and "*Structuring*".

The first step takes as input text in a "KES" SGML format and performs a linguistic tagging of these texts (for more details see section 3.1.1).

The "extraction" and "structuring" phases are the real core of the KES process: implemented as cooperative processes (rather than purely sequential operations) they allow the manipulation of information found in the results of the previous stage, in the input texts or in lexicons, according to different criteria:

- linguistic information (morpho-syntactic tags, syntactic properties, thematic roles ...),
- statistical considerations (frequencies, weights...),
- "factual data" (eg. typographical structure indicators such as "title", or "lists" markups...)

This in order to select, extract, group items of information and link them together.(c.f. section 3.1.2. for details of the process).

The main idea is to manipulate "properties" added to words, terms or texts (see the SATO approach) like tags, statistical information, links, ... ; our novel contribution is to use linguistic information in all steps to add or control these properties (we can use more information than [ANI90]) while staying open to different modelling choices.

Furthermore, one of our constant concerns is to establish well-defined and standardised exchange formats (SGML DTDs) between the different steps ensuring modularity and simplifying data import/export from/to application databases or tools manipulating textual data.

3. The Tools

Our tools are developed in a modular way in C++, based on standards like OSF/MOTIF, SGML and run under the SUN OS UNIX operating system.

3.1 Current State: Two groups of tools compose the current toolbox. GCE [GCE93] - the first one, implements (in batch mode) a parameterised corpus characterisation and a first extraction of "interesting" items. EAEKES the second tool is much more interactive and accounts for the more domain specific part of extraction as well as for knowledge structuring and validation.

3.1.1 Corpus Characterisation + preliminary extraction

GCE (Graal Corpus Exploration) has been developed by the partners in a previous phase of GRAAL and is a set of software tools that ran in batch on a corpus perform a morpho-syntactic analysis (pattern-matching approach), and produce structured data representing :

- lists of tagged words (GENELEX categories),
- predictions on the categories of unknown strings ("date", "numeral", proper noun...) based on "morpho-graphic" patterns & context,
- lists of syntactic groups (Noun Phrases that appear to be potential terms of the domain, verbal forms...),
- various statistical information (ranging from frequencies of particular punctuations to frequencies of syntactic patterns),

Thus the tool can produce several representations of the corpus (eg: lemmatised, with various levels of tags etc ..).

GCE uses for its purpose large GENELEX lexicons (French 55 000 simple words and 18 000 compound words, English 40 000 words) and a constraint grammar like approach.

Because GCE performs a bottom-up analysis using a large coverage lexicon and makes lexical category predictions on unknown words the results are usually very satisfactory and constitute a valuable starting point for the subsequent phases, even for texts in very technical domains.

3.1.2 Extraction and Structuring

Here the implementing software tool called EAEKES is based on GSI-Erli's AlethSAC software (based in turn on GSI-Erli's experience in the E.C. A.I.M. project Menelas).

EAEKES' main goal is to allow for both interactive and batch knowledge extraction and characterisation. It automatically bases itself on the GCE results.

The most basic operation consists in manually creating domains of information and manually (either by typing them in, or by mouse selection in source text) inserting items[2] into them.

The tool in this mode of operation allows the navigation between items (on the basis of the links between them) and domains of information in a somewhat "hypertext" style (mouse clicking).

The user can also interactively change both terms and domains inter-relationship (in a cut/paste way) automatically maintaining inverse links.

[2] items can be made up of parts of words, words, phrases, or even disjoint text elements.

The second mode of operation offers the possibility to describe selection patterns that are then applied on the corpus in a batch mode.

The selection patterns are coupled with a description of actions that are to be performed forming together "KES rules".

The actions can among other perform "parsing like operations" by using a type of chart like representation of the analysed text.

Most often however users will perform actions that extract identified parts of text and assign it some characteristics and or link them to other already extracted items.

The reader will find below examples of patterns that can be specified and examples of actions performed with matching items.

The types of patterns can be:

- morphological - for example: "all words beginning with "aqu" or containing the infix "hydro" are to be placed in the domain "water"[3],
- simple contextual patterns - eg "all words that are not adverbs and immediately precede verbs related to the verb "to flow" are to be characterised as nouns denoting liquefied bodies[4]; Note that the type of relations that are to hold between verbs can be what is found in a rich "GENELEX dictionary", but can also be user defined criteria .
- syntactic - example: all NP heads following a form of "obtained by" are to be placed in the domains "methods"; all phrases of the type "all <NP-head>'like'<enumeration-heads>" describe an 'is_a' relationship between the NP head and each one of the enumeration heads (ex: "the data processing methods like automatic classification, formal links,...")[5],
- combinations of the above[6] types.

The above mentioned types of rules are to be provided by the user, this however is a task too difficult for some users that are not linguists or knowledge engineers. Therefore the toolbox provides a library of basic rules that can either be used as such or serve as starting patterns that users may refine and adapt.

Whether the extraction is made by "hand" or by rules it can be performed on any of the forms output by GCE. It is thus possible to combine forms as they appeared in the source text with results of lemmatisation, taking into account frequency data, logical markups or co-occurrences. The extraction process can be made "information sensitive" i.e. the selection patterns can be made to check whether a knowledge item is not already classified somewhere (by another rule, by the user or in an external source[7]) ; thus, it is possible to use all the information available on an item, coming from the original corpus or external resources[7].

Therefore information predicted by the rules can be used in other rules thus achieving a bootstrap type of effect. Facilities are available to keep track of the dependencies between hypotheses, the user can interactively explore retrieval of hypotheses and see the effect on the extracted knowledge.

This extracted knowledge (set of knowledge items) can be interactively checked and cleaned up.

Once checked the knowledge items can be structured: various types of links can be made, domains can be divided into subdomains and items dispatched into them[8].

Several ways are available to accomplish this task:

- Manually selecting and moving items using the mouse.
- Rules similar to the extraction patterns can also be applied on the extracted set of knowledge items (eg: all items placed in the domain of "energy production" that begin with the strings "atom" ,"nucl" or contain the word "fission" are to be moved into the subdomain "energy production by nuclear means").[9]

These structuring rules can also establish links between items, it is therefore possible to perform actions like: check for "inclusion" of item in another one and if positive link them with an "generic" link. Because both the possibilities of the rule language as well as the availability of large stocks of linguistic knowledge the previously mentioned "inclusion tests" can range from simple character string matching to testing for

---

3 various application domains offer degrees of such regularities- some applications in chemistry being perhaps the most illustrative (the above "hydro" would probably assign a different domain in chemistry).

4 presuming that we are dealing with a technical text.

5 regularities like these have been observed in technical texts (eg cf [ROU92]).

6 users with different skills write different types of rules. For instance a Knowledge engineer usually does not use the notion of a syntagmatic head.

7 note that such an external source is the GENELEX compatible dictionary but may also be a thesaurus that the user is trying to enrich, but could also be an ontology in the context of Expert System construction (cf [MIZ93]). The toolbox's underlying data model can in a "meta model way" host a large variety of resources.

8 the user can have two modes of operation either an unconstrained where any "domain" or link can be created or a "model" guided mode in which the "administrator" user has to specify the links used, the types of domains, the types of items and specify for each the possible interrelationships.

9 note that such rules can implement some simple forms of generalisation strategy.

example the variation of the prepositions used in a corresponding position in several terms.

Note that established links can also be tested in the rules and for example it is possible to detect "shortcut links" in hierarchies of items. These identified links can then be presented to the user for further operations.

The standard type of result display is in a workview which can handle lists of items and lists of links between items. Some graphical manipulation is also possible. (cf figures)

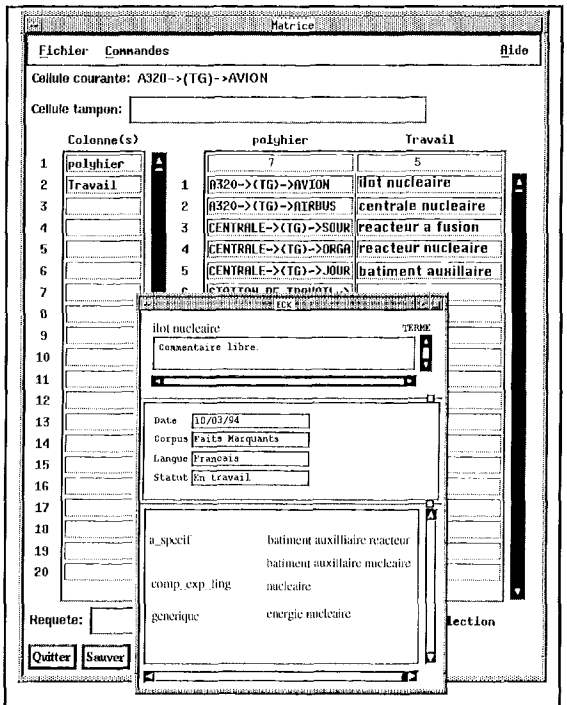

Fig.1: textual mode

Here the user works in a navigation mode; the window in the foreground of this view corresponds to a to the knowledge item "ilot nucléaire" of type "term". Through a pull-down menu one can view all occurrences of the term. The "spread sheet" like window ("work view") in the background allows to manage sets of links (cf first column), knowledge items (second columns) and attributes of knowledge items (not shown here).

Note that the steps during which the user writes different extraction rules and visualises their results, can also be used in grammar engineering tasks. There exists in fact a third viewing mode (not shown here) in which the user may see the places in his corpus where a rule applies. Nothing prohibits the rules from being "normal" parsing rules, that the user wants to explore.

This type of use has however been reserved for a subsequent phase of the SP - after July 1994 -and thus has not been fully explored yet.

3.2 Outline of an example session

We will illustrate the working of our toolbox in the context of an application whose goal is to enrich an existing thesaurus.

1. A large corpus (10 MBs) of texts in the domain of informatics is batch processed by GCE, yielding lists of nouns, adjectives, potential terms, unknown words and statistics...

2. Using a spell checker called from within the EAEKES system, the user eliminates unknown words that in fact were misspelled technical terms.

3. The remaining unknown words are studied in context thanks to the retrieval of the sentence where they occurred. The pertinent ones (very technical terms, domain specific proper nouns like "Unix", "Salton" ...) are kept.

4. The most frequent nouns and noun phrases are observed. Extraction rules allow to extract the most "productive" NP heads allowing to build "domains" such as : "machines" (d1), methods (d2), languages (d3)... This extraction is made "information sensitive" : the existing thesaurus is used to help defining these domains. Then, the NPs based on these heads are dispatched: "Unix machines", "IBM machines" in d1, "statistic methods" in d2, "C programming language" in d3, etc ...

5. With rules using syntactic or semantic information found in the GENELEX dictionary (for example, synonyms of "method" and "processing") and using contextual patterns (eg variants of the form "...methods such as ...") other items are dispatched: in d2 we will then find items like "document classification"; "data processing", "textual data processing", "formal links", "Salton theory" ...).

6. Graphical facilities allow to establish links between the different items: eg "isa_links" between "data processing" and "textual data processing", between "textual data processing" and "document classification" etc...

7. The structured results are then exported (encoded according to an SGML DTD) in order to be recovered by a terminology management system which will allow their integration in the original thesaurus.

4. Where are we?

The set of tools described here is a prototype and further work is planned in both the LRE project TRANSTERM and the continuation of this GRAAL subproject.

5. Conclusions

We have presented an approach supported by a toolbox corresponding to the aims of industrial actors in the field of NLP. The objectives targeted are an increase in the productivity of people manipulating large corpora. Rather than introducing a new theory of automatic KA we

have presented a "practical" approach allowing the combination of automatic and "hand" methods which can be based on large generic repositories of knowledge, working in a bootstrap type of cooperation.

References

[ANI90] "An Application of Lexical Semantics to Knowledge Acquisition from Corpora"; P. Anick and J. Pustejovsky in Proc. of Coling 1990.

[CAL90] "Acquisition of Lexical Information from a Large Textual Italian Corpus", N. Calzolari and R. Bindi in Proc. of Coling 1990.

[GCE93] "*Etude de corpus: un préalable nécessaire pour l'adaptation des systèmes de TA aux besoins des utilisateurs*", E. Dauphin, in proceedings of "Troisièmes Journées Scientifiques TA, TAO, Traductique", to be published as an "UREF" publication Université de Montréal.

[GEN90] "*GENELEX project : EUREKA for linguistic engineering*", B. Normier and M. Nossin in Proc of International workshop on electronic dictionaries 1990.

[GRA92] "*Outline Eureka GRAAL*";Coling 1992 (International Project Presentations Volume).

[IDE90] "Very Large Neural Networks for Word-Sense Disambiguation" N. Ide and J. Véronis in Proc. of ECAI90 1990.

[LEX92] "*Surface grammatical Analysis for the extraction of terminological noun phrases*", Bourigault Didier in Proc of Coling 1992.

[MIZ93] "*Knowledge Acquisition and Ontology" Riichiro* Mizoguchi, in Proc. of KB&KS93 - International Conference on Building and Sharing of Very Large-Scale Knowledge Bases 1993.

[ROU92] "Elaboration de Techniques d'analyse adaptées à la construction de bases de connaissances" F. Rousselot and B. Migault Ensais in Proc. of Coling 1992.

[SAT92] : "*L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistée par ordinateur : applications possibles avec SATO*", S. Bertrand-Gastaldy, G. Pagola, "Documentation et bibliothèques", Avril-Juin 1992.

[TER90] "*Termino v.1.0 : rapport de recherche "Novembre* 1990, par le groupe RDLC (Recherche et Developpement en Linguistique Computationnelle), Centre d'Analyse de Textes par Ordinateur, Université du Québec, Montréal.

[WIL93] "*Towards Automated Knowledge Acquisition"* Yorrick Wilks and Sergei Nirenburg in Proc. of KB&KS 93 - International Conference on Building and Sharing of Very Large-Scale Knowledge Bases 1993.

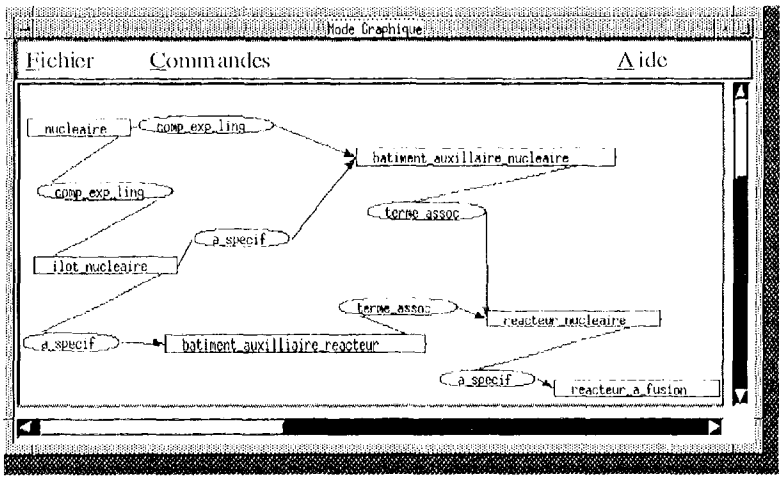

Fig2 Graphical Mode:

All objects and links can be displaced and updated using the mouse and keyboard.
Here the user working on nuclear emergency manuals has chosen to display part of the extracted *linguistic composition links*: "nucléaire" "occurs in" "ilot nucléaire" and "batiment auxilliaire nucléaire", but also *thesaurus like links* : "ilot nucleaire" appears to be a generic term of "batiment auxilliare reacteur".