ARC A3: A METHOD FOR EVALUATING TERM EXTRACTING TOOLS AND/OR SEMANTIC RELATIONS BETWEEN TERMS FROM CORPORA

Christophe Jouis[(1)] & ARC A3[(2)]

[(1)] CAVI − Censier
University Paris Sorbonne Nouvelle − Paris III

13, rue Santeuil, F−75251 Paris CEDEX, FRANCE
Phone: +33 (0) 1 45 87 42 74
Fax: +33 (0) 1 45 87 41 73
E−mail: Christophe.Jouis@univ−paris3.fr

[(2)] ARC A3, the whole of the participants who submit a software:

Bourrigault Didier (ERSS, Toulouse, France),
Bruandet Marie−France (CLIPS−IMAG, Grenoble, France), Chevallet Jean−Pierre (CLIPS−IMAG, Grenoble, France), Memmi Daniel (LEIBNIZ−IMAG, Grenoble, France), Descles Jean−Pierre (LALIC, CAMS, EHESS, France)
Daille Béatrice (Univ. Nantes, IRIN, France),
Enguehard, Chantal (IRIN, Univ. Nantes, France),
Falhon, Martine (Xerox−, Meylan, France)
Le Priol, Florence (LALIC, CAMS, EHESS, France),
Toussaint, Yannick (LORIA, UMR 7503, Vandoeuvre−les−Nancy, France),
Meunier, Jean−Guy (LANCI,UQAM, Canada),
Perrin Patrick (Logos Corp., NJ 07866, U.S.A.),

E−mail: aupelf−a3@univ−lille3.fr

Abstract

This paper describes an ongoing project evaluating Natural Language Processing (NLP) systems[1]. The aim of this project is to test software capabilities in automatic or semi−automatic extraction of terminology from French corpora in order to build tools used in NLP applications. We are putting forward a strategy based on qualitative evaluation. The idea is to submit the results to specialists (i.e. field specialists, terminologists and/or knowledge engineers).

Building terminology (terms or concept names and the logic−semantic relations they hold) from extensive textual data is not a simple task when the designer has to examine a new field of knowledge. The designer may not be acquainted with the representation of the field, its structures and the articulations between its objects. To make the designer’s task easier, natural language processing systems can be of help particularly those dedicated to the identification of terms or concepts names related to a specific field of knowledge (construction of a reference terminology) and the logic−semantic relations they contain. These systems can be applied to the modeling and designing of the following types of systems : (1) The modeling of an object−oriented database design (static aspects: i.e. describing the structure), (2) Knowledge−based systems : modeling the hierarchies between classes and the relations between the objets concerned by a set of rules, (3) Modeling the conceptual design of a relational database (domains, relations, coherence maintenance), (4) Thesaurus construction (documentary databases, Information Retrieval, ...), (5) Terminological database construction, and so on.

The Natural Language Processing systems we are evaluating use various modules in order to identify terms or concept names and the logic−semantic relations they hold. The approaches involved in corpus analysis are either based on morpho−syntactic analysis, statistical analysis, semantic analysis, recent connectionist models or any combination of two or more of these approaches. Most of these systems need, in addition, a general language dictionary, a glossary of technical terms covering the relevant field, etc. The identification of terms is in fact an extraction of noun phrases corresponding to the concepts representing the field of knowledge. In their current state, these systems are mostly semi−automatic processing tools.

In this paper, we will examine the evaluation problem.

[1] The research we are conducting is sponsored by the "Association des Universites Francophones" (AUF) an international Organisation whose mission is to promote the dissemination of French as a scientific medium. Software submitted to this evaluation are conceived by French, Canadian and US research institutions (National Scientific Research Centre and Universities) and/or companies : CNRS (France), XEROX, and LOGOS Corporation among others.

1. Introduction: Description of the project and the first Experiments

1.1. Description of the Project

The ARC A3 is a Concerted Research Project sponsored by the AUF ("Association des Universites Francophones" − Association of the French−speaking Universities). This project endeavors to promote the development of corpus and evaluation procedures related to French. The results of this project should benefit to the progress of research in the field of natural language processing evaluation techniques in many aspects: to create measurement tools allowing objective comparisons between approaches, promote the development of existing systems, to constitute huge corpora for future evaluations and to fine−tune test procedures used in evaluation in order to allow a better visibility of the offer. The first phase of this project (1995−1999) was an exploratory phase. It allowed us to constitute a first series of corpus, to define protocols of tests and to carry out a first test evaluation. The second phase (2000−2001) will consist in applying the methods defined and tested during of the first phase. Approximately ten new participants are expected to take part in the second phase.

1.2. Conclusions of the first phase (1995−1999)

The eight systems, tested during the first phase, have various functionality and provide various outputs: ordered grammatical terms, grammatical networks, classes of terms, and graphs or semantic networks. These systems have been described in our previous works (see Mustafa El Hadi and Jouis, 1996a, 1996b & 1998; Béguin, Jouis and Mustafa El Hadi, 1997; Jouis and Mustafa El Hadi 1997). The following diagram sums up the question of systems comparison:

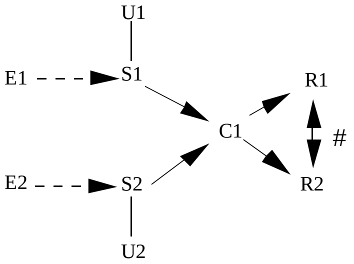

Figure 1: Systems comparison diagram

Two different systems S1 and S2 applied to the same corpus C1 can yield two comparable results. Nevertheless, if we take into consideration the purposes of S1 & S2, the nature of linguistic elements (E1 and E2) they use and the conditions of their use (end users) the comparison between the two systems would be rather difficult. For this reason, it is important to organize the system in such a way that their output can be comparable. Another problem has to do with the nature of the corpus provided as input (type of documents, fields of knowledge). These elements are not necessarily the same for S1 and S2. Given that, it is important to provide the systems with different types of corpus.

The second question deals with the criteria of choosing adequate corpuses. The elaboration of terminology includes the analysis of documents of different types and varying sizes (textbooks, technical manuals and specialized dictionaries, transcripts of interviews with field experts, etc.). The "ideal" corpora to efficiently identify terms and relations which can be considered as good descriptors providing efficient access to specialized information: automatic thesaurus construction, broadening of existing thesaurus, etc. should be representative of the field of knowledge considered. Three types of texts are needed (see Ahmed, 1993): instructional (textbooks, technical manuals, encyclopedic texts); informative (learned papers, advanced treatises, interview transcripts of experts, patent documents); and imaginative (popular science material, public information material i.e. advertisement about the goods and services of the subject domain). We need this variety because the terms in each type of corpus serve different purposes. This typology is used by a team at the University of Surrey to build term banks in about 10 different subject fields (see Ahmed 1993).

The evaluation was primarily qualitative. The same "experts" (i.e. archivists, terminologists, or specialists in the field) carried it out simultaneously for each system. The evaluation carried out was based on the analysis of the use of the output provided by the system. This analysis was based on two distinct applications: assisting in terminology construction and manual indexing.

As input, the systems were provided with textual data from SPIRALE, a periodical dealing with education and pedagogy issues. The size of each issue is about 200 pages. The following outlines the main difficulties we faced. First, too different systems are applied to the same textual data yielded results seemed initially incomparable.

Secondly, systems are initially conceived to meet varied needs (Indexing, Content analysis, Thesaurus construction, Knowledge acquisition, Information Retrieval, etc.). It is therefore very difficult to use the same criteria when evaluating their output.

We were thus obliged to make several distinctions.

1.2.1. We distinguished three but not clearly−cut distinct categories of tools:

− *Terms extraction Tools*: they are based on lexical, syntactic and statistical analysis to extract the most frequent terms (complex or not) in the texts.

− *Classifier Tools*: they build classes of terms, which regularly appear together in the texts. These systems use either a lexical, syntactic and statistic analysis or pure numerical approaches.

− *Semantic Relations extraction Tools*: These systems use a statistical approach or a linguistic approach by contextual exploration to establish relations between the terms.

1.2.2. Experts examined the system output and their use in two distinct applications:

We are trying to define a set of criteria for evaluating systems according to the following areas of performance:

(1) Indexing (information retrieval : recall and precision);

(2) Complete covering of a field.

1.2.3. We distinguished two categories of corpora:

(1) Homogeneous corpora (dealing with single field) and (2) heterogeneous corpora (dealing with different fields). Moreover, it is necessary, to consider polysemy. We were thus obliged to distinguish on one hand the literary texts and on the other hand the scientific or technical texts.

1.2.4. The electronic format and labeling problem of the textual files

The various systems tested do not run on the same Operating System. At the beginning, we had considered format RTF (Rich Text File) of Microsoft. This format is difficult to handle on the UNIX platforms. Finally, the simplest format that we chose is HTML, which offers many advantages:

a) It is only one universal language of description of text. For any Operating used (DOS, Windows 9X, MaOS, Linux, Unix, etc), it exists always a browser such Netscape to interpret it;

b) Systems which wish to take account of the position of the terms in the text (title, beginning of paragraph, italic, boldface, footnotes, etc.) can thus use this information;

c) For the systems which do not use information of position of the terms in the text, it is easy to build a module filtering mark−up;

d) Lastly, thanks to the hypertext links, it is possible, for each extracted result, to insert a link that makes it possible to the user to return to the context of extraction of the result.

2. Evaluation Problem

Comparing terms extractions tools and the logic−semantic relations they entertain with each other is not an easy task if we compare it to the work done within the framework of conventional algorithms comparisons (such as sorting algorithms, etc.). For those, we normally have as input a set of totally formalized and structured data (normally digital) at our disposal. Furthermore the expected results are usually defined beforehand (sorted value lists, etc). The purpose of ‘conventional’ evaluation process is not to compare results but rather to rank algorithms according to quantifiable criteria: processing time, required processing resources (usually measurement in required memory space).

2.1. TREC and ARC A3

With the difference of TREC (Text Retrieval Conference, see TREC, 1999), there are not directly applicable numerical methods (founded on statistical correlation between the outputs of the systems).

In our approach, the concept of evaluation is different even if we take into consideration required processing time, required systems resources (i.e. required linguistic tools, electronic dictionaries, etc.). The only relevant evaluation criterion is the quality of the resulting terminology. Therefore, we are going to be focusing mainly on the relevance of the terminology obtained. There are several problems of interpretations of measurements to classify the systems. Let us quote for example:

− Relevance of measurements independently of a relative existing reference terminology;

− Can we say that differences are significant? For example, it could be that a majority of systems produces poor results while a minority of systems produces good results. Interpretations could then be distorted. For a statistical evaluation, it seems necessary to introduce a reference terminology associated with the field. It would then be necessary to measure the variations of the various outputs compared to this terminology of reference.

2.2. However, what is the validity of a “terminology of reference”?

Specialized terminologists always carry out a terminology of reference. It is thus never a perfect result. Moreover, Two terminologists working on the same corpus can produce different results (i.e. a different hierarchy of terms). At best, the two results obtained can be equivalent in the point of view of their use. We can say whereas they are equal in intention but different (extensionally). In addition, a terminology of reference cannot be used to measure the results. They must be regarded as possible results.

For this reason, we introduce into our evaluation the existing terminology. Moreover, we will ask certain " specialists " in the field of the corpus to make their own terminology at the same time. All these manual results will be presented to the appraisers under an electronic format without specifying that it is not a question of automatic results.

The exploratory phase is finished. We will describe now to the second phase, which will have to be operational.

3. Method selected

In order to operate a statistical evaluation, we will introduce a set of “reference terminology”, built manually, and associated with the field of the corpus. We will measure then the variations of the various outputs compared to these reference terminologies. These manual thesauruses must be considered as a possible results as well as the others result. The thesaurus will thus be evaluated like the other outputs. The specialists should have no knowledge of the existence of manual thesaurus. In other words, manual thesaurus must be considered as a "relative" reference frames, taken as a possible result as well as the others.

The presentation of the results is a very important part of the judgement of the specialists. It is thus necessary to impose a standardized presentation (by category of system)

3.1. Covering of the field: standardized presentation of the results for qualitative and quantitative evaluation

Results should be displayed in a Relational data base table. The standard format is in the following:

T	A	B	C	E	T	T2	R

Table 1: standardized presentation (field)

The meaning of each column is as follows:

T = column of the terms, classes (set of terms) or triads [term/relation/term] entered under our data base by the evaluated tools. Pointers to the sentences where they appear in the text are inserted in order to allow the return to the text. For the extractors of terms category, the "grammatical leading term" considered is attached.

A = very good

B = acceptable

C = acceptable after modification (the specialist can then propose a modification in the column T1)

D = not absurdity but not very characteristic

E = artifact

T1 = term modified (by the specialist)

T2 = number of systems having found the same term

R = possible notices

The specialist, for each entry, tags one of exclusive boxes A, B, C, D or E, and if required fills the boxes T1 and R. The box T2 is filled automatically by our Data Base interface. In order to obtain quantitative results, it will then be enough, for each column, to count the number of tagged boxes.

3.2. 2.2 Indexing: Standardized Presentation of the results for qualitative and quantitative evaluation

In the same way that for the cover of the field, each result will be presented in the shape of a table standardized to the format below:

T	A	B	C	E	F

Table 2: standardized presentation (indexing)

The meaning of each column is as follows:

T = column of the terms, classes (set of terms) or triads [term/relation/term] entered under the database by the evaluated tools. Hypertext Pointers to the sentences where they appear in the text are inserted in order to allow the return to the text. For the extractors of terms category, the "lead−in term" considered is attached.

A = belongs to the Relative Thesaurus of Reference (RTHR).

B = does not belong to the RTHR, but returns there by synonymy.

C = does not belong to the RTHR, but should appear (just as it is).

D = does not belong to the RTHR, but should appear in it under another descriptor.

E = measurement of the noise: useless term

F = silence measures: term present in the RTHR but not extracted by the system[2].

The specialist, for each entry, tags boxes A, B, C, D, E, or F. In order to obtain quantitative results, it will then be enough, for each column, to count the number of tagged boxes.

4. Outlines

The second corpus selected is a set of technical or scientific papers in the field of biotechnology offered by INRA ("Institut National de Recherche en Agronomie" − French National Institute of Research in Agronomy). An official evaluation campaign should start by the end of April 2000. More systems will be tested and corpora of varying sizes will be used. We will of course integrate the conclusions to the further course of our study.

Acknowledgements

The authors gratefully acknowledge the financial assistance provided by the Association des Universités Francophones (AUF) − in funding the general research project within which this paper was written. We also thank the participants for the ARC A3 for their constructive remarks.

5. References

Ahmad, K. (1993). Terminology and Knowledge acquisition: A text Based approach, *Terminology an Knowledge Engineering* (TKE' 93). Frankfurt/Main: INDEX Verlag. 56−70.

Beguin, A, Jouis, C., Mustafa, W, (1997) : "Evaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus", Actes des Premières Journées Scientifiques et

[2] Taking into account the "relative" value of the THR, it possible that the column F cannot be measured.

Techniques, (JST’97), FRANCIL, AUPELF−UREF, Avignon, avril 1997, pp. 419−426

Bourigault, D. (1994). Extraction et structuration automatiques de terminologie pour l’aide à l’acquisition des connaissances à partir de textes. In *Actes du 9ème Congrès Reconnaissance des Formes et Intelligence Artificielle,* Paris, pp 397−−408.

Bruandet, M.−F. (1989). Outline of a knowledge base model for an Intelligent Information Retrieval System; In *Information Processing & Management,* Vol. 25, N° 3, 1989

Coret, A., Kremer, P., Landi, B, Schibler, D., Schmitt, L., Viscogliosi, N. (1997), "Acces à l’information textuelle en français : Le cycle opératoire Amaryllis", *In Actes des Premières Journées Scientifiques et Techniques (JST’97),* FRANCIL, AUPELF−UREF, Avignon, avril 1997, pp. 5−8

Daille B.,(1996) "ACABIT : une maquette d’aide à construction automatique de banques terminologiques monolingues ou bilingues", in A. Clas, P. Thoiron et H. Béjoint (eds) *Lexicomatique et Dictionnairiques,* p.123−136, FMA, Beyrouth, 1996.

Daille, B.(1996) "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology", in P. Resnik et J. Klavans (eds.) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, p. 49−66, MIT Press, Cambridge, MA, USA, 1996.

Descles, J.−P. (1990). *Langages applicatifs, langues naturelles et cognition.* Hermès eds., Paris, France.

EAGLES (1996): Evaluation of Natural Language Processing Systems FINAL REPORT [On−line] Available: http://issco−www.unige.ch/projects/ewg96/ewg96.html>

Enguehard, C., "Acquisition de terminologie à partir de gros corpus", Informatique & Langue Naturelle, ILN’93, Nantes, p.373−384, décembre 1993.

Enguehard, C., Pantéra, L., "Automatic Natural Acquisition of a Terminology", Journal of quantitative linguistics, vol.2, n°1, p.27−32, 1995.

ISO/IEC CD TR 9126−2 (1999) Information Technology −− Software Engineering −− Software products quality −− Part 2: External metrics (Ed. 1)

ISO/IEC CD TR 9126−3 (1999) Information Technology −− Software Engineering − Software products quality −− Part 3: Internal metrics (Ed. 1)

ISO/IEC DIS 9126−1 (1998): Information Technology −− Software quality characteristics and metrics −− Part 1: Quality characteristics and sub−characteristics (Ed. 1)

Jouis, C, Mustafa, W. (1997), " AUPELF Project : Term and Semantic Relation Extraction Tools. Evaluation Paradigms ", In *Proceedings of Workshop " Evalutation in Speech and Language Technology ",* University of Sheffied, June 17−18, Sheffield, UK, pp. 106−113

Mustafa W. & Jouis, C. (1998) " Terminology Extraction and acquisition from textual data: criteria for evaluating tools and methods " *In Proceedings of the First International Conference On Language Resources and Evaluation*, Granada (Spain) : 28−30 May 1998, organized by ELRA (European Language Resources Association). Granada: ELRA, Vol. 2, pp. 1175−1180

Mustafa, W. & Jouis, C. (1996), "Evaluating Natural Language Processing Systems as a Tool for Building Terminological Databases", in *Proceedings of the Fourth International ISKO Conference: Knowledge Organization and Change*, Library of Congress, Washington D.C., *Advances in Knowledge Organization*, Vol.5, INDEX Verlag, Frankfurt/Main, pp. 346−355

Mustafa, W., Jouis C. (1997), " Natural Language Processing−based Techniques and their Use in Data Modelling and Information Retrieval ", In *Proceedings of 6th International Study Conference on Classification Research, Knowledge Organization for Information Retrieval*, 16−19 June 1997, University College of London, London, FID/CR, & ISKO. The Hague : FID, pp. 157−161

Perrin, P. and F. Petry. (1999). “An Information−Theoretic Based Model for Large−Scale Contextual Text Processing”, *Information Sciences*, 116(2−4), 229−252, 1999.

Perrin, P. and F. Petry. (1998). “Lexical Contextual Relations for the Unsupervised Discovery of text Features”, chap 10, in “Feature Extraction, Construction and Selection: a Data Mining Perspective”, Liu & Motoda (Eds.), Kluwer, Boston, 1998.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Adisson−Wesley Publishing Company. *Terminology an Knowledge Engineering* (TKE’ 93). Frankfurt/Main: INDEX Verlag. 56−70.

Seffah, A., Meunier J.−G. (1995). ALADIN : Un atelier orienté objet pour l’analyse et la lecture de Textes assistée par ordinteur. In *International Conférencence On Statistics and Texts.* Rome

Sparck−Jones, K.. (ed.) (1981). " Retrieval Systems Test" In Sparck−Jones, K: *Information Retrieval Experiments,* (pp. 256−284). London: Butterworths.

Swanson, D.R. (1988). "Historical Note: Information Retrieval and the Future of an Illusion"*, Journal of the American Society for Information Science*, 39, 92−98.

Toussaint Y., Namer F., Daille, B., Jacquemin C., RoyautéJ. & Hathou N.(1998) “Une approche linguistique et statistique pour l’analyse de l’information en corpus”, In *TALN’98,* Paris, France, 1998.

TREC (1999) [On−line] Available: <http://trec.nist.gov/overview.html>