EVALDA-CESART Project: Terminological Resources Acquisition Tools Evaluation Campaign

Widad MUSTAFA EL HADI, Ismail TIMIMI, Marianne DABBADIE

Equipe ILSI (Ingénierie Linguistique pour les systèmes d'Information)

IDIST/CERSATES UMR 8529

Université Charles De Gaulle Lille 3

BP 149

F-59653 Villeneuve D'Ascq, France

TEL +33 (0) 3 20 41 68 15

FAX +33 (0) 3 20 41 63 79

mustafa@univ-lille3.fr, timimi@univ-lille3.fr, dabbadie@univ-lille3.fr

Abstract

This paper describes the ongoing evaluation work in CESART research project supported by the French Ministry of Research and Technology[1] and coordinated by the University of Lille 3 and ELDA. The project deals with the evaluation of term and semantic relation extraction from corpora in French. CESART logically follows on the evaluation project achieved within the framework of the Concerted Research Project ARC A3[2] supported by the AUF, former Aupelf-Uref. This article sets the context, briefly mention the project objectives and reports on the suggested evaluation protocol.

1. Introduction

Terminology plays a major role in information processing and management and in specialized communication. Its role has been enhanced by the spread of automation and by the availability of electronic corpora. These two factors have had a massive impact on many different applications: systematic terminology, building, natural-language interface design, lexical units management for specific use in some sub-languages and technical writing, thesaurus construction, translation and indexing as well as the recent growth of cross-language information retrieval (CLIR).

The paper will describe the ongoing work we are undertaking within the framework of the Technolonague-Evalda evaluation platform, a joint venture between the French Ministry of Research and Technology and ELRA. The project is dealing with the evaluation of terminology resources acquisition tools. Eight participants (see below for more details), both from public institutions and industrial corporations were involved in this project and were responsible for producing corpora suitable for extraction tasks and elaborating a protocol in order to evaluate objectively terminology resources acquisition tools.. This expression covers respectively, term extractors, classifiers and semantic relation extractors, ontology editing, validating tools, concordances generators, …

The paper will also report on the suggested evaluation protocol in discussion within the framework of the project.

2. CESART organization

CESART brings together four kinds of actors: two coordinators (the University of Lille 3 and ELDA) who play an organizational role; corpora providers; participants to the test and two scientific advisors. The organizing teams in cooperation with the discussion group made up of representatives of each participating team and the scientific advisors are supposed to co-operate in defining a methodology for evaluating the systems.

1.1 2.1. The CESART project objectives

The project logically follows on the evaluation project achieved within the framework of the ARC A3 project. We here briefly mention the project objectives and set the context. The ARC A3 evaluation campaign has been one of the first attempts to define an evaluation protocol for terminology resources acquisition tools. This work has allowed pointing out the theoretical and methodological difficulties that led us to suggest some solutions. In order to improve future evaluation protocols, it is essential to include some elements in their definition. In defining CESART evaluation protocol we be focusing on the following points: -the limits observed during the last evaluation ARC A3 campaign will be taken into consideration: new metrics will be introduced especially for semantic relation extraction; integration of new functionalities for the validation of the results

[1] http://www.technolangue.net

[2] The ARC A3 is a project of the ILEC group coordinated and founded by AUF 1996-2000. The project aim was to test software capabilities in term and semantic relation extraction from corpora in French (cf. Béguin, *et al*., 2000; Jouis *et al*., 1997; Mustafa El Hadi *et al*., 1998; 2001).

provided by the systems; it is essential to have an interface in order to manipulate and interpret the results (validating term, relations and classes). This type of interface can dramatically facilitate the interaction with the evaluators and the end-user of these tools. To this end we can use and/or customize WorldTreck of EDF.

2.2. Participating systems

The systems participating to the CESART campaign are: *IDE/XTS* (TEMIS), *Lexter* (EDF), *SeekJava* (LaLICC, Paris 4), *Synoterm* (LIPN, Paris 13), *Termic* (CEA), *Terminae* (LIPN, Paris 13), *Termos* (RALI), *TermWatch* (ERSICOM –Lyon 3 et LITA-Univ. de Metz), *WorldTrek* (EDF R&D). Beyond the difference of their theoretical models and architectures, the systems are divided into two categories[3]:

Term extractors :

4 systems are being tested: *IDE/XTS, Lexter, Termic, Termos*

Relations extractors :

4 **syntactic relations extractors**: *Lexter, Termos, Termic et TermWatch*[4]

3 **semantic relations extractors**: *IDE/XTS, Seek-Java et SynoTerm.*

Terminae is considered as an ontology editor.

2.3. CESART evaluation protocol

The evaluation of terminology extraction tools usually takes place within the framework of methodologies that comprise an evaluation procedure founded most of the time on a «black-box» approach, test data, metrics and interpretation procedures.

2.3.1. Test collection

The evaluation procedure consists in making the system work with a set of test data (corpora and test referential) and in measuring the results using pre-determined metrics. In the absence of a referential or automated methods, human expertise is solicited.

2.3.1. 1. The corpus

The corpus used to carry out a terminology resources acquisition task, must be homogenous and domain representative. Corpus selection[5] must comply with two relevance criteria: it has to be domain specific (domain reference documents) when the task consists in a systematic terminology elaboration, and must be representative of the documents used in the final application (documents retrieved by a documentation search tool). Document retrieval has to be carried out with the help of domain specialists, according to the application tested. The corpus must be voluminous as well, in order to meet the requirements of systems based on statistics although only one extract (representing a rational sample) will actually be used for systems evaluation. The sample will remain confidential and only organizers and experts will have knowledge of these test data. For the first run we have privileged a corpus related to medicine for the already mentioned reasons and other reasons explained hereafter. Complementary resources will be put at participants' disposal according to their specific needs. The tools based on statistics, for example, will be supplied with a learning corpus, similar to the evaluation corpus. The similarity refers to length, format and thematic features.

2.3.1. 2. Test referentials

This type of referential can be extracted from a specialized dictionary, a thesaurus or a recognized list representative of the test corpus. It can be built *ex nihilo* from a corpus read by experts (e.g. Mustafa el Hadi *et al.*, 2001a) In this case, word lists are extracted from already existing referentials, but they have very little in common with the extracted terms. It is therefore necessary to have them completed by specific domain specialists. The use of referentials seems to be a more efficient evaluation procedure than human expertise, as far as, even more than human expertise, linguistic alignment guarantees the reproducibility of the experience and thus the attainment of objective results (Daille, 2002). Moreover, it offers a possibility to carry out horizontal evaluations thanks to the reproducibility of the protocol. In general, human expertise together with the gathering of test material are of an important cost that must be taken into account when defining the protocol. It is also necessary to use a list called *matching referential*. It serves as a basis to work out precision and recall after matching extractor results with the term list. It is a list related to the domain specific corpus. It can be structured (like a thesaurus) or not structured (flat list). The list is not given to the participants. The choice of the medical domain for the corpus will allow us to make this type of resource easily available.

2.3.1. 3. Human expertise

In the absence of standard evaluation procedures that can be applied to any linguistic tool and automated, evaluations almost always rely on human expertise. Human judgment constitutes a kind of referential that unfortunately is not

[3] Note that a system can belong to more than one category.

[4] *TermWatch* is not necessarily a relation extractor but rather a tool designed for domain specific text mining with the aim of extracting information that can be used in an information watch framework.

[5] For a recent detailed research on corpora for terminological resources acquisition see (Condamines (2003).

reproducible. On the other hand, this method does not allow the evaluation of silence, except when the protocol specifies that experts have to mark up missing terms, not extracted by the systems (see Daille, 2002).

We remain convinced, however, that "the different types of terminological resources or ontologies are distinguished according to their potential usage, i.e. the type of application these resources are used with. According to the application, they comply with different conceptualizations or requirements". Bourigault *et a*l. (2002b). Any evaluation protocol of this type of tools should take into account this specificity.

2.3.1. 4. Metrics

The metrics mainly used are the traditional precision and recall information retrieval systems performance metrics. These measures have their shortcomings and in particular for the evaluation of some linguistic processing tools as stated by Chaudiron (2001). They do not allow to make a distinction between softwares or to conclude on their usage value within the foreseen application fields.

3. CESART evaluation model

A black-box evaluation for the first run has been decided with a particular attention to adequacy for the control tasks i.e., construction of reference tools (specialized dictionaries), translation, indexing, scientific watch. We have to consider the user's dimension (evaluation of use). In other words, the adequacy of the tools in performing the mentioned tasks should be assessed in relation to a specific user need.

- Even if this approach may be criticized for its subjective side, end-users prefer it because of its usefulness when comparing two or more systems which differ in all their parameter settings.

- A black-box evaluation is more oriented towards system's end-user when compared to a glass-box evaluation which is obviously a developer oriented approach and not an end-user one.

3.1. A User-oriented evaluation

The fact of taking into account the user needs is probably the most important aspect of the CESART project when compared to ARC A3. In the field of terminological resources acquisition, some authors have put forward the necessity to take into account the various classes of application (Estopa, 2001), (Bourigault *et al*., 2001a, 2001b). We remain convinced that the scheduled application should be the very basis of tool design.

As already mentioned, any extraction or terminological resource elaboration should be considered in its context of use. In this case we need to see to what extent these tools are adapted to the aims they were designed for. On the other hand, it would be interesting to measure user satisfaction when using these tools. It would then be interesting to ground our work on the linguistic applications engineering model, based on a linguistic and informational ergonomy as proposed by (Chaudiron 2001). The author accounts for a number of innovative models in this paradigm among which the user needs modeling and defends the *Usability* concept in information system evaluations. The most important aspects are the measurement of the usefulness, the adequacy and the user satisfaction of a software component.

3.2. Defining a protocol per tool category

If we take into account the diversity of the terminology acquisition tools and our experience with ARC A3, it is not possible to rely only on one protocol for the evaluation of all available tools[6].. In the case of term extractors, the protocol founded on information retrieval system performance was both consistent and adapted to the evaluation of the three control tasks (systematic terminology extraction, translation and indexing). Applying it to the evaluation of a semantic relation extractor, though, was inadequate. It is possible on the one hand, to identify noise (i.e. precision) as long as one has a good knowledge of the domain and of possible semantic relations.

It might be useful though, beforehand, to define what is a correct semantic relation, which is not always simple. On the other hand, it appears difficult to work out recall figures (i.e. silence) if one does not know what semantic model has been implemented and whether there exists any gold standard for a determined semantic model and/or application. Finally if the model is not known one can wonder whether all identifiable relations are relevant for the considered application.

Another solution emerges from the qualitative paradigm founded on human expertise and that consists in « comparing semantic relations established by an extractor to a manual work on a part of a text » (L'Homme, 2001a). These proposals are merely dependent on a human expertise that would combine domain specific knowledge together with semantic models, but it is difficult however to gather these conditions. It would be more relevant to initiate collaboration between specific domain specialists in order to validate this type of resource.

3.3. Models to be tested

[6] We encountered much difficulty when trying to apply the protocole to semantic relations extractors and to the *CONTERM* classifier.

Models based on gain of time: it is an experience that takes into account the time used to build a terminological resource. (Bourigault *et al*. 2002a) relates an experience of the evaluation of an ontology in the field of surgery reanimation (Lemoigno, 2002 quoted by (Bourigault *et al*. 2002a).. The referential used was extracted from a thesaurus used for this specific domain. The time it took a doctor to build the ontology that accounts for the domain covered by the thesaurus was estimated to 50 hours for a 2000 concept ontology.

Moreover, we have noticed that some problems that rose during the ARC A3 campaign also arise with the CESART campaign. Let us list some of these problems: (i) semantic relations extractors are all different and hardly comparable products; (ii) the variety of the extracted relations; (iii) the very different implemented semantic models; (iv) the difference between the aims and the applications concerned; (v) as the terminology resources acquisition tools cannot be evaluated using the same protocol as the one used to evaluate term extractors, it is important to consider using the adequate protocol to test this kind of software; (vi) the relations are related to professional practices and different objectives; (vii) all applications are different and their structure therefore influences the type of extracted relations and outputs.

4. Conclusion

We have thus managed to realize the difficulty to set a common protocol due to the large variety of tools. This difficulty will be particularly acute for the evaluation of NLP-based ontologies that use a broad set of the terminology resources acquisition tools. We are facing hardly compatible choices: How is it possible to find a protocol that applies to all extraction, terminology resources and ontology building tools? Is it necessary to pre-determine groups of softwares for not being able to unify protocols? Or shall we then have to switch from the evaluation of a technology to the evaluation of applications as Chaudiron (2001) suggests[7]? The question remains unanswered.

2 References

Béguin, A., Jouis, Ch., Mustafa Elhadi, W. (2000). Evaluation d'outils d'aide à l'extraction et à la construction automatiques de termes et de relations sémantiques. In Chibout, K., Mariani, J., Masson, N., Neel, F. éds., (2000). Ressources et évaluation en ingénierie de la langue, Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF), pp 161-179.

Bourigault D., Lame G. (2002a). « Analyse distributionnelle et structuration de terminologie : application à la construction d'une ontologie documentaire du droit », Structuration de terminologieTAL, Nazarenko A., Hamon T., Vol. 43, n°1, 2002, Paris, Hermès, p. 128-150.

Bourigault, D., Aussenac-Gilles N., Charlet J. (2002b). « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas » Revue d'Intelligence Artificielle, Vol. X, n° X/ 2002.

Chaudiron, S. (2001). *L'évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigmes*, Mémoire pour l'habilitation à diriger des recherches en sciences de l'information, présenté devant l'Université de Paris 10, Paris, novembre 200.

Condamines A. (2003., *Sémantique et corpus spécialisés : constitution de bases de connaissances terminologiques*. Mémoire d'habilitation à diriger des recherches. Université Toulouse Le Mirail, Juin 2003.

Daille B. (2002). *Découvertes linguistiques en corpus*, Mémoire d'HDR, Université de Nantes, janvier 2002.

Ibekwe-Sanjuan F., Sanjuan E. (1997). «Term Watch: variations terminologiques et veille scientifique », à paraître dans Actes du 4ème Colloque ISKO-France, Grenoble, juillet 2003.

Jouis, C., Mustafa El Hadi, W. (1997). "AUPELF Project: Term and Semantic Relation Extraction Tools. Evaluation Paradigms, In Proc. of the Speech and Language Technology Club Workshop "Evaluation in Speech and Language Technology", Univ. of Sheffied, June 17-18, Sheffield, UK, pp. 106-113.

L'Homme M.-C. (2001a). « L'impact des nouvelles technologies sur la gestion terminologique », www.onterm.gov.on.ca/onterm/iso/proceedings.html).

L'Homme M.-C. (2001b). « Évaluation d'outils d'aide à la construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus ARC-A3, Rapport final, Montréal, mars 2001.

Mustafa El Hadi, W. & Jouis, C. (1998). Terminology Extraction and Acquisition from Textual Data: Criteria for Evaluating Tools and Method, Proceedings of the First International Conference on Language Resources and Evaluation, Grenada, Spain may 1998, pp. 11750-1178.

Mustafa El Hadi, W., Timimi, I. Béguin, and A. Debrito. M. (2001). The ARC A3 Project: Terminology Acquisition Tools: Evaluation Method and Tasks. In Evaluation Methodologies for Language ad Dialogue Systems Workshop, ACL/EACL, Toulouse, 6-7 July 2001, pp. 41-50.

Mustafa El Hadi, W. (2004). « Evaluation d'outils d'acquisition de ressources terminologiques » à paraître dans Chaudiron, S. (sous la dir. de), L'Evaluation des Systèmes de traitement de l'information, Paris, Hermès, 2004.

[7] See (Chaudiron 2001) for more details on the paradigm shift