Terminological Resources Acquisition Tools: Toward a User-oriented Evaluation Model

Widad Mustafa El Hadi[1], Ismail Timimi[1], Marianne Dabbadie[1], Khalid Choukri[2], Olivier Hamon[2,3], Yun-Chuang Chiao[2]

1. IDIST / CESARTES - University of Lille 3 - rue du Barreau BP 149 - 59653 Villeneuve d'Ascq Cedex – France
2. ELDA - 55-57, rue Brillat Savarin, 75013 Paris – France
3. LIPN UMR 7030 - Université Paris 13 & CNRS – 99 av. J.-B. Clément, 93430 Villetaneuse – France

E-mail: widad.mustafa@univ-lille3.fr, ismail.timimi@univ-lille3.fr, dabbadie@univ-lille3.fr, choukri@elda.org, hamon@elda.org, chiao@elda.org

Abstract

This paper describes the work achieved in CESART (Campagne d'Evaluation des Systèmes d'Acquisition des Ressources Terminologiques) evaluation project supported by the French Ministry of Research and Technology1 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 A32 supported by the AUF, former Aupelf-Uref. This paper sets the context, briefly mentions the project objectives, reports on the adopted evaluation protocol, describes the evaluation tasks and finally provides the results of the official evaluation campaign.

1. Introduction

The CESART project deals with the evaluation of terminological resources acquisition tools. Five participants, both from public institutions and industrial corporations were involved in this project and were responsible with the organizer for producing corpora suitable for extraction tasks and elaborating a protocol in order to evaluate objectively terminology acquisition tools. This expression covers respectively, term extractors, classifiers and semantic relation extractors. The paper reports on the evaluation protocol, the official campaign and the results.

2. Participating systems

Five participants, both from public institutions - CEA, University of Paris 13 and University of Montreal - and industrial corporations - EDF and TEMIS - were involved in this project. Beyond the difference of their theoretical models and architectures, the systems are divided into two categories according to their outputs: a) Term extractors and b) semantic relations extractors.

3. Evaluation protocol

A black-box evaluation has been defined with a particular attention to adequacy for the control tasks i.e. construction/enrichment/updating of the reference and for indexing. 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. Taking into account the diversity of tools tested and our experience with ARC A3 project, we tried to adapt so far as we can the protocol to each category of the tested tools.

3.1. Test corpora

Two corpora for testing the systems have been provided:

- a corpus of medical texts. This corpus is gathered from *Health Canada* Web site (http:// www.hc-sc.gc.ca).

- a corpus dealing with educational sciences, *SPIRALE* journal.

A rational sample of the whole corpus has been used for evaluating the systems as voluminous corpora are too restrictive. Before the official test the sample remained confidential and only organizers and experts had knowledge of the data domain. When we could not do otherwise, complementary resources were provided to participants according to their specific needs. For instance, the semantic relation extractor was supplied with one output of the term extraction task.

Table 1 gives an idea of the corpora constituting elements and size.

Corpus	# documents	# segments	# words
CISMEF	7 514	255 161	9M
SPIRAL	149	12 109	535K

Table 1 – Data Resources

3.2. Test material

Test material 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. In the CESART project, the experts from the *CISMEF* team (Catalogue and Index of French-language Health Internet resources) of Rouen University Hospital (http://www.chu-rouen.fr/cismef) use the *MeSH* thesaurus (http://www.nlm.nih.gov/mesh/) as a reference for the evaluation. For the corpus of the education domain, *Motbis* (http://www.cndp.fr/motbis), a thesaurus

[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. Mustafa El Hadi *et al*., 1998; 2001).

largely used in the educational circles is used to compare the systems output with the thesaurus elements.

4. Control Tasks

4.1. Task One: Term Extraction for Creating Terminological Resources

For each of the two corpora described above, participants should return a ranked list of terms according to their own criteria of relevance. The output list should consist of the following components:

- canonic form,
- ranked relevance,
- variations,
- frequency,
- contexts of the extracted term within the corpus.

All the data are XML and UTF-8 encoded, and provided in DOS and UNIX format. For the input and output data, the systems have to refer to an exact DTD. Four systems submitted runs to the evaluation.

4.1.1. Evaluation Measures

An automatic procedure is first applied by comparing the output of a system with a reference list generated from an existing thesaurus of the concerned domain. All matched terms are considered as relevant. The non-matched terms are then submitted to experts for the manual evaluation. For the manual relevance assessment, experts judge the relevance of the extracted terms in order to establish a thesaurus of the domain. Experts should rank the terms according the following criteria:

- C0: the extracted term match exactly with an entry of the thesaurus,
- C1: the term is not in the thesaurus but is assessed as relevant,
- C2: at least two components of the term are present in the thesaurus
- C3: one component of the term is present in the thesaurus
- C4: any component of the term is present in the thesaurus

4.1.2. Results

We report in this section some of the results of the two corpora. We will first give the statistical results of the output of the participant systems and then the manual evaluation we conduct. Evaluation is in progress therefore we will be glad to get more deeply into it during the oral presentation.

4.1.2.1 Statistical results of candidate term extraction

We decided to limit the sample for the evaluation to the first 10,000 ranked candidate terms. It is however important to report on the over productivity of the participant systems concerning candidate term extraction.

System	*CISMeF* corpus	*SPIRAL* corpus
Sys 1	10,000	10,000
Sys 2	108,074	60,695
Sys 3	26,053	3,447
Sys 4	286,018	41,377

Table 2: Output statistics on both two corpora

The results show a significant discrepancy between the four systems if we take into consideration the total number of the extracted candidate terms.

4.1.2.3. Manual Evaluation

Here we present the manual evaluation only on the medical corpus. Each term returned by systems is assessed according to the criteria described in section 4.1.1.

Three evaluators assessed the first 100 terms of each system output in order to calculate the correlation between the different judges. The judges' profiles are in the following: evaluators 1 and 2 are archivists in the CHU de Rouen (the Rouen Hospital University), evaluator 3 is a doctor of medicine and linguistics having competence in the field of documentation.

We compute the Pearson's correlation coefficient which is 95% between both archivists, 78% between the first archivist and the doctor and finally 80% between the second archivist and the doctor. Thus the archivists' assessments are close whereas the doctor assessed differently the systems. The first archivist has done all the manual assessment of the evaluation.

The assessments proceeded as follows: output terms were automatically matched with the *MeSH* thesaurus (according to the criterion C0), the evaluator then validated the matched terms and pursued the evaluation according to the other criteria.

Table 3 presents the results we obtained on the first 1,000 ranked terms regarding the best variation of a term. It shows the cumulative precision on the terms. Thus the C1 columns present the system's performance according to the criterion C1, the C2 columns the ones according to the criteria C1 and C2, etc.

System	C1	C2	C3	C4
Sys 1	10.6	34.3	47.3	52.1
Sys 2	28.8	34.1	35.7	38.5
Sys 3	8.5	14.6	20.7	36.1
Sys 4	0.4	3.4	10.3	29.0

Table 3: results on the best variation of terms

Table 4 presents the same results but this time according to all the variations that the systems returned.

System	C1	C2	C3	C4
Sys 1	10.8	38.3	50.6	56.0
Sys 2	24.5	37.2	39.2	44.5
Sys 3	8.5	14.6	20.7	36.1
Sys 4	0.7	5.3	11.7	30.3

Table 4: results on the terms with all the variations

Results of both tables are similar. Anyway the scores are not as good as expected.

As shown in the tables, the performance of System 2 are homogeneous and the best when using the criterion C1. While System 1 presents a worse performance regarding the same criterion C1, but shows a significant improvement when considering the three other criteria. If we take all the criteria in consideration, System 1 leads the best results among all the other systems. System 3 has worse results than the two first, but shows an important improvement when using all criteria to assess the outputs.

System 4 shows the worst performance among the three other systems particularly with the criteria C1, even if an improvement of the results can be observed when taking into account all the criteria of the relevance assessment.

When comparing the results shown in tables 3 and 4, we find that there is no significant difference although systems lead slightly better performance when all the variations of the extracted candidate terms are evaluated, except for System 3 which only returned candidates terms without variations. We assume that system's performance might not depend on the number of variations of a term which can be extracted since only one variation of a term is enough for its relevance assessment. Thus it should be sufficient to evaluate the terms on the canonic form, even if it is not so easy to proceed for the experts.

4.2. Task Two: Controlled Indexing and Thesaurus Enrichment

The same corpora were provided to the participant systems with the thesaurus of the concerned domain. Two evaluations were planned:

- Automatic evaluation: matching the systems output with a list of descriptors drawn from the existing thesaurus (recall/precision measures are applied)
- Manual evaluation: experts judge the relevance of the newly extracted descriptors (their contribution to enrich and update the thesaurus). Experts should rank the descriptors according to three scales of measurements: a) good term or descriptor; b) acceptable; c) discarded).

Unfortunately no system participated to this task.

4.3. Task Three: Semantic Relations Extraction

According to the protocol, the system would be provided by the corpus and a list of terms used as "focal" terms. These terms are supposed to function as probes which can help in extracting other terms with which they hold semantic relations. The system should then return a list of synonymic relations within the corpora.

Only one system participated to the task number three. As the system used a general dictionary to extract the synonyms from the corpus, its designer agreed to adapt the procedure and the extracted candidates synonyms were matched with a sample list of synonyms extracted from the corpus using the domain-specific thesaurus.

Whatever the type of evaluation performed (automatic evaluation by matching the output with the reference or manual evaluation with human assessment), it is difficult to find a minimum of a convergence between the system's output and the reference.

Anyway the third expert of the first task evaluated manually a sample of this output with the both criteria: "validated" or "non-validated". Only some relations suggested by the system were assessed as valid and estimated relevant by the expert. Consequently, in the context of an over productivity of results, the metrics used become in this case slightly insignificant. It is normal to note a high rate of noise and a low rate of silence.

5. Discussion

Our two evaluation experiences brought us to consider the meta-evaluation, an interesting and emerging theme in this relatively young field. A lot of questions have been raised amongst evaluation campaign organizers and participants to the tests. In some evaluation campaigns where the protocol is totally based on an agreement between the organizers and the participants (this is the case of ARC A3 and CESART), can the results be reliable and significant indicators of systems' performance? The answer is definitely no. The quality of the results will depend on the different system modules, resources added to enrich the system while processing a task, size and genre of the corpora, etc.

In addition, and since it is vital to harmonize input processing conditions for the evaluated systems system designers are sometimes forced to accept some adaptation in order to abide by the terms of the protocol. To stick to the adopted protocol can have either a negative or a positive impact on a system. Examples can be drawn from the ARC A3 protocol and our final campaign (Mustafa el Hadi et al., 2001).

As far as human expertise is concerned, is it possible to discard the subjectivity of judges and their competence level and merely rely on average calculation in order to rank the systems? To the evaluators' subjectivity, which is firmly linked to their degree of tolerance, we can also add other factors that can have an impact on their appreciations, such as the cognitive overload generated by the huge quantity of the evaluated data, their competence in manipulating computer tools, etc.

Within this evaluation campaign we measured the correlation between the three judges on a sample. If the correlation between the two archivists experts in the medical filed (correlation of 95%) can prove that there is no impact linked to the subjectivity of theses two experts it is not the case for the assessment of the medical doctor whose results are different from the two archivists (correlations of 78% and 80%).

Another question can be raised, if the human judge refers to his own knowledge in order to asses the systems' output and use it as referential data, which can spontaneously adapt and grow, can we compare this particular case to a pre-established textual referential in order to automatically match the results to his or her personal-knowledge-referential? Automatic evaluation can be a value-added one when it confirms human judges' appreciations but how can we assess it when it refutes them? (Timimi, 2006).

One of the assets of our evaluation campaign is to survey evaluators' usage practices. Our evaluators and this applies to the two corpora, can be the potential users of the systems expertise tools. In addition to the set of criteria and recommendations we designed for the evaluators we asked them to give us a feedback in an opinion-poll-fashion on their general appreciations and the unspoken aspects of their task. The objective is to give us an idea about how the experts deal with the results (considered as particular documents). The examination of the evaluators' comments is underway and the results cans be helpful for the overall study of the evaluation paradigm.

6. Towards a User-oriented evaluation

Taking into account the user needs is probably the most important aspect of the CESART project when compared to previous campaign for testing this type of tools. Concerning the medical corpus the Rouen Hospital

University team is assessing the extracted terminology for two tasks (see above). The idea was to measure the adequacy of the tools to their daily work, which is free indexing, enriching the *CISMeF* database and the related thesaurus. The objective of *CISMeF* is to describe and index the main French- language health resources to assist health professionals and consumers in their search for electronic information available on the Internet

As for the second corpus, specialists in educational science are testing the adequacy of the tools in their daily work, i.e. updating and enriching the terminology for CNDP (*Centre National de Documentation Pédagogique*) and the related indexing tools.

In these two use cases the idea to assess to what extent the tested tools are adapted to accomplishing theses tasks though they are not really designed for them. They are considered as generic tools. It would be however interesting to measure users satisfaction when using these tools (for more details on this question see Chaudiron 2001, 2004).

It would have been more interesting to assess the systems on more tasks corresponding to more use-cases (Oudshoorn & Pinch, 2003) but the main problem is that the tools we evaluated are generic tools. The four term extractors we assed are merely geared towards producing candidate terms which should ultimately be assessed by human interventions. We therefore limited the tasks and the use-cases to free indexing and reference tools enrichment. Many systems which could have been an asset in this campaign withdrew since they were "orphans", that means they were the only representatives of their category and the idea was to compare them to others on the same task. As an example we can mention TermWatch (Ibekwe-San-Juan, 2004) geared to strategic and scientific watch.

7. Conclusion

CESART provided us with an awareness of the state-of-the-art in the field of terminology acquisition tools. Considering the results, we cans say that term extractors have reached a certain scientific maturity in spite of still remaining draw backs but semantic relation extraction tools have not yet reached the stage of scientific maturity; there is still a long way to go. Hypotheses are still to be tested for this type of tools. Regarding the implemented evaluation protocol, if the qualitative approach offers the easiest form of systems evaluation it nevertheless retains two major drawbacks: (i) it makes up for a very boring job when there are too many results (ii) judgments can easily be slanted by the subjective approach of the expert.

Automatic matching concurred with human experience which notices that the systems produce many "noisy" terms, terms not existing in the *MeSH* or in the CNDP reference tools. Hence the interest of some of these "noisy" terms for enriching and updating reference lists and terminology data bases.

Acknowledgements

The authors gratefully acknowledge the financial assistance provided by *The French Ministry of Research* in funding the general research project within which this paper was written.

We would like to thank all the participants of the CESART campaign, CEA, EDF, TEMIS, the University of Montreal and the University of Paris 13. We are grateful to Pierre Zweigenbaum for his help in developing the protocol and providing the medical corpus. We are also very grateful to all the evaluators for their participation to the project.

References

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

Chaudiron, S. (dir) (2004). Evaluation des systèmes de traitement de l'information, coll. *Traités des sciences et des techniques de l'information*, Hermès, 2004.

Ibekwe-SanJuan F., SanJuan E. (2004). Mining textual data through term variant clustering: the Termwatch system. *Proceedings of the Conference Recherche d'Information assistée par ordinateur (RIAO).* Avignon", 2004, 487-503.

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, 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., Timimi, I., Dabbadie, M. (2004). CESART : Campagne d'Evaluation des Systèmes d'Acquisition de Ressources Terminologiques. In *Actes du Colloque LREC' 2004*, pp. 515-518.

Mustafa El Hadi, W. (2004). Acquisition de ressources terminologiques. In: *Evaluation des Systèmes de traitement de l'information,* sous la direction de Stéphane Chaudiron, Hermès, pp. 149-169.

Oudshoorn, N. and Pinch, T. (2003). *How users matter. The co-construction of users and technologies*, Cambridge, MA: MIT press.

Timimi, I. (2006). Outils de terminologie : nouvelles métriques, nouvelles pratiques. In *Proceedings of JADT*, 2006 (à paraître).