The Quæro Evaluation Initiative on Term Extraction

Thibault Mondary[1], **Adeline Nazarenko**[1], **Haïfa Zargayouna**[1], **Sabine Barreaux**[2]

[1] LIPN - UMR 7030 CNRS / Paris 13 University, France [2] INIST-CNRS, Vandoeuvre-lès-Nancy, France

Abstract

The Quæro program has organized a set of evaluations for terminology extraction systems in 2010 and 2011. Three objectives were targeted in this initiative: the first one was to evaluate the behavior and scalability of term extractors regarding the size of corpora, the second goal was to assess progress between different versions of the same systems, the last one was to measure the influence of corpus type. The protocol used during this initiative was a comparative analysis of 32 runs against a gold standard. Scores were computed using metrics that take into account gradual relevance. Systems produced by Quæro partners and publicly available systems were evaluated on pharmacology corpora composed of European Patents or abstracts of scientific articles, all in English. The gold standard was an unstructured version of the pharmacology thesaurus used by INIST-CNRS for indexing purposes. Most systems scaled with large corpora, contrasted differences were observed between different versions of the same systems and with better results on scientific articles than on patents. During the ongoing adjudication phase domain experts are enriching the thesaurus with terms found by several systems.

Keywords: Term extraction, Evaluation, Quæro, Pharmacology, Gradual relevance, Scalability

1. Introduction

Computational terminology is a twenty years old discipline that aims at building automatically or semi-automatically terminological resources from corpora. A lot of terminological tools have been developed, but despite the progress made, it remains difficult to get a clear idea of the maturity of this research field and to compare approaches. Some efforts have been made to set up an evaluation protocol adapted to the specificity of terminological tasks. The TEMREC task of the first NTCIR initiative (Kando et al., 1999) was the first initiative dedicated to the evaluation of term extraction, but it was not reconducted due to its lack of popularity. CoRReCT (Enguehard, 2003) has proposed an interesting dataset and protocol to evaluate term recognition in corpora (a task close to controlled indexing). CESART (Mustafa El Hadi et al., 2006) was the most comprehensive challenge: a gold standard list of terms and an acquisition corpus were chosen for a specific domain (medicine), the systems had to extract terms from the acquisition corpus and their results were compared against the gold standard using various relevance criteria.

Quæro is a program promoting research and industrial innovation on technologies related to the analysis and classification of multimedia and multilingual documents[1]. Partners technologies are yearly evaluated in internal or external initiatives. In this context, an internal Terminology Extraction evaluation was organized by LIPN in 2010 and reconducted in 2011.

[1] `http://www.quaero.org`

This paper presents the protocol and results obtained during these two evaluation initiatives.

2. The Task

Term extractors output lists of relevant terms from a domain specific corpus. The terms may be mono or polylexical units, *i.e.* made of one or several words. The goal of the evaluation was to measure the quality of the resulting lists of terms, through a comparison with a gold standard. We really focused on the specific task of term extraction without taking into account term ranking and variant clustering. This experiment has been done on the pharmacology domain: we proposed a set of corpora, every participant had to extract a list of relevant terms and submit it for evaluation. A participant could submit several runs and each run was evaluated against the gold standard.

Three objectives were targeted in this initiative. The first one was to evaluate the behavior and scalability of systems regarding the size of corpora (the ability of extractors to deal with large corpora and the stability of scores regarding the size of the acquisition corpora). The second goal was to assess the progress between different versions of the same systems. The last one was to measure the influence of corpus type on the performance of term extractors.

3. Corpora and Gold Standard

Two different types of pharmacology corpora were used. A first set of European patents was supplied by Jouve, an industrial Quæro partner. From it, three self-including corpora of growing size (called C1, C2 and

Peptide Extraction From Ionic Conjugates:
[0042] A 50 mg sample of an ionic molecular conjugate was mixed into 20 mls of methylene chloride. The mixture was sequentially extracted with 50 ml, 20 ml and 20 ml portions of 2N acetic acid. The acetic acid extracts were combined and analyzed for peptide content by high performance liquid chromatography(HPLC). Peptide analysis by HPLC is as follows. HPLC analysis were performed using a Waters model M-45 solvent delivery pump and an EM Science MACS 700 detector at wavelength 220 nm and 1.0 AUFS. Peptides were run using a Lichrospher (EM separations) C18, 100A, 5um, 25cm x 4.6 mm column and 30% acetonitrile/0.1% TFA as an isocratic eluent buffer.
[0043] Following are details (Table VI) of. . .

TABLE VI IN-VITRO ASSAY DATA

DAY OF ASSAY	PERCENT OF TOTAL PEPTIDE RELEASED		
	Example #8	Example #9	Example #10
1	5.5%	12.5%	11%
7	26.9%	21.3%	53%
14	55.2%	47.3%	55%

. . .

Figure 1: Excerpt of a patent

Voluntary exercise improves stress coping and lowers anxiety. Because of the role of GABA in these processes, we investigated changes in the central GABAergic system in rats with free access to a running wheel for 4 weeks. The control animals had no access to a running wheel. Using in-situ hybridisation histochemistry, we studied changes in gene expression of various GABA::(A) receptor subunits as well as the GABA-synthesising enzyme glutamic acid decarboxylase-67 (GAD67) in the forebrain. There were region-specific decreases in alpha 2, beta 3 and gamma 2 subunit mRNA expression and region-specific increases in beta 1 subunit expression. The alpha 5 and delta subunits, in the forebrain specifically associated with extrasynaptic GABA::(A) receptors in the hippocampus, showed differential increases in expression levels.
. . .

Figure 2: Excerpt of a typical abstract from CA

C3 in the following) were built, in order to test the systems' scalability. C1 contains 500,000 words, C2 1.5M words and C3 2.5M words. An excerpt of a patent is presented on Figure 1.

In order to measure the systems' behavior against another type of corpora of the same domain, an additional corpus (called CA) composed of abstracts of scientific articles in pharmacology was extracted from the PASCAL[2] database and provided by INIST-CNRS. CA weights 1.5M words (same size as C2), a typical abstract is presented on Figure 2.

The gold standard was an unstructured version of the pharmacology thesaurus[3] used by INIST-CNRS for indexing purposes. It consists of 76,466 English terms, mainly about general pharmacology, diseases and drugs. An excerpt is presented on Figure 3.

[2]The multidisciplinary bibliographical database produced by INIST-CNRS. PASCAL page on the official INIST site — `http://inist.fr/spip.php?article11`.

[3]Available on TermSciences, the multidisciplinary terminological portal developed by INIST-CNRS — `http://www.termsciences.fr`

5-HT3 Serotonine receptor
5-HT4 Serotonine receptor
5S-RNA
5s rrna
. . .
Bacillus subtilis ribonuclease
Bacterial lipopolysaccharide receptors
Connective tissue activating factor
. . .
Recombinant microorganism
Recombinant protein
Recombinant virus

Figure 3: Excerpt of the gold standard

4. Metrics

Scoring was performed using terminological precision, recall and F-Measure, introduced in (Nazarenko et al., 2009; Zargayouna and Nazarenko, 2010) and presented on Equations 1–3. The terminological precision and recall metrics take into account a gradual relevance. The systems' outputs are tuned to find their maximal correspondence with the gold standard, which means that the outputs are adapted to the terminological type and granularity of the gold standard.

$$TPrecision = \frac{\sum_{i \in T(O)} rel_{GS}(i)}{|T(O)|} \quad (1)$$

$$TRecall = \frac{\sum_{i \in T(O)} rel_{GS}(i)}{|GS|} \quad (2)$$

$$TFMeasure = \frac{2 \times TPrecision \times TRecall}{TPrecision + TRecall} \quad (3)$$

$|T(O)|$ is the size of tuned term extractor output $T(O)$, $|GS|$ the size of the gold standard. $rel_{GS}(i)$ is the relevance of a term i with respect to the gold standard. It is based on a terminological distance d_t and on a threshold τ on this distance:

$$rel_{GS}(i) = \begin{cases} 1 - \delta(i) & \text{if } \delta(i) < \tau \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\delta(i) = min_{e \in GS}(d_t(i, e))$. d_t is computed as the mean of a string and a complex term distances that is based on a normalized edit distance and takes into account word permutation.

LIPN developed a scoring tool called Termometer[4] which is available under GPLv3.

[4]`http://sourceforge.net/projects/termometerxd`

	Runs on	Comes from
Acabit 4.3	C1, C2, C3, CA	LINA
Fastr 2.04	C1	LIMSI
Jv1	C3	Jouve
YaTeA 0.5	C1, C2, C3, CA	LIPN
MIG-YaTeA	C1, C2, C3, CA	INRA/MIG
TermExtractor	C1, C2, C3	LCL
TermoStatWeb 2	C1	OLST
TermoStatWebV3-multi	C1, C2, C3, CA	OLST
TermoStatWebV3-nomulti	C1, C2, C3, CA	OLST
TS3-multi	C1, C2, C3	INRIA/TEXMEX
TS3-nomulti	C1, C2, C3	INRIA/TEXMEX

Table 1: Challengers, with Quæro partners highlighted

5. Participants and Systems

Three Quæro partners participated in the evaluation: INRA/MIG (academic partner), INRIA/TEXMEX (academic partner) and Jouve (industrial partner). Six publicly available term extractors, not developed within Quæro, were also tested: `Acabit` (Daille, 2003), `TermExtractor` (Sclano and Velardi, 2007), `TermoStatWeb`, `TermoStatWebV3` (Drouin, 2003; Drouin, 2006) and `YaTeA` (Aubin and Hamon, 2006). During the second evaluation initiative, one system became unavailable (`TermExtractor`) and another underwent a major revision (`TermoStatWeb`). The participating systems are presented on Table 1 and detailed bellow.

Acabit is a terminology extraction tool which takes as input a part-of-speech tagged corpus and proposes as output a ranked list of multi-word terms. Acabit is based on syntactical patterns and statistical filtering. It works on English and French, the tested version was 4.3[5].

Fastr is a multi-lingual tool for automatic indexing. In free indexing mode, it acts as a term extractor and extracts terms with their variations from a corpus. Fastr works on French and English, the tested version was 2.04[6].

Jv1 is developed by Jouve for patent classification. It produces nomimal or verbal phrases based on shallow parsing and statistical metrics. This system works on English, German and French.

YaTeA aims at extracting noun phrases that look like terms from a corpus. It is based on simple syntactic patterns and endogenous disambiguation. Exogenous disambiguation using external resources is also possible, but it was not used here. YaTeA works on English and French, the tested version was 0.5[7].

MIG-YaTeA is an enriched version of YaTeA developed by INRA/MIG. It uses two post-processing filters for cleaning and merging results. The first filter gets rid of extraction errors, incomplete terms, spelling errors and overgeneralized terms. The second filter merges terms sharing the same lemma. For this initiative, only one representative per cluster was output.

TermExtractor is an online term extractor designed to build ontologies. It is based on two entropy measures. The first one is used to select the terms which are consensually referred throughout the corpus documents, the second one is used to select only the terms which are relevant to the domain of interest. TermExtractor[8] works on English.

TermoStatWeb is an online term extractor based on TermoStat (Drouin, 2003), which basic principle is to compare the distribution of words between a specialized document (the corpus to be processed) and a large corpus of "general language", using different statistical measures, *e.g.* log likelihood, specifity or chi2. It works on English, Spanish, French and Italian. This version has been superseded by TermoStatWebV3 at the end of 2010.

[5] `http://www.bdaille.fr/index.php?option=com_content&task=blogcategory&id=5&Itemid=5`

[6] `http://perso.limsi.fr/Individu/jacquemi/FASTR`

[7] `http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5`

[8] `http://lcl2.di.uniroma1.it/termextractor`, available until early 2011

TermoStatWebV3 is the new version of TermoStatWeb[9]. Different runs were produced, some with only simple terms (`TermoStatWebV3-nomulti`) and other with both simple and multi-terms (`TermoStatWebV3-multi`).

TS3 is an experimental version of TermoStatWebV3 enriched by INRIA/TexMex with post-processing and filtering. Different runs were produced, some with only simple terms (`TS3-nomulti`) and other with both simple and multi-terms (`TS3-multi`).

6. Evaluation Results

Terminological scores have been computed for a total of 32 runs. Table 2 presents the number of returned terms. Huge differences can be observed. Although the corpus of abstracts (CA) contains 1.5MWords as C2, term extractors produce much more term candidates from CA than from C2.

Overall results are presented on Table 3, with best scores highlighted in bold font for each corpus. Two families of systems can be observed from the last two tables: those who favor precision by returning a small amount of correct terms (`Jv1`, `TermExtractor`, `TermoStatWebV3-multi` and `TermoStatWebV3-nomulti`) and those who favor recall (`Acabit`, `Fastr`, `MIG-YaTeA`, `TS3-multi`, `TS3-nomulti` and `YaTeA`). `TermoStatWeb` cannot be clearly categorized. This difference between systems can be explained by their design choices.

6.1. Scalability

Almost all systems scale with large corpora, *i.e.* provide results without crashing. However, one system has not been tested on C1 and C2 because it was designed to work with very large corpora.

For a first group of systems, precision decreases as the corpus size increases (it is an expected behavior, as more terms are able to be found, more terms might be erroneous). For the systems of a second group (`TermoStatWebV3` and `TS3`), however, the results (amount of terms, TPrecision, TRecall) do not increase monotonically with the corpus size. This is probably due to the statistical underlying filtering strategy.

6.2. Systems evolution

Evolution can be observed between the different versions of the same systems. As described before, `TS3` and `TermoStatWebV3` both derived from `TermoStatWeb`. `MIG-YaTeA` is an evolution of `YaTeA`.

Figure 4 presents the evolution of TPrecision, TRecall and TFMeasure between `TermoStatWeb` and `TS3-multi`. We observe improvements, mainly for TRecall and TFMeasure. Figure 5

	TPrecision	TRecall	TFMeasure
Acabit			
C1	60.31%	8.65%	15.13%
C2	56.26%	14.24%	22.72%
C3	53.62%	16.45%	25.18%
CA	54.06%	19.74%	28.92%
Fastr			
C1	54.20%	10.12%	17.06%
Jv1			
C3	74.05%	1.86%	3.62%
YaTeA			
C1	31.22%	10.09%	15.25%
C2	23.99%	16.96%	19.87%
C3	19.24%	19.80%	19.52%
CA	28.76%	23.78%	26.03%
MIG-YaTeA			
C1	57.30%	10.64%	17.94%
C2	52.09%	18.48%	27.28%
C3	48.85%	**22.13%**	**30.46%**
CA	55.08%	**26.57%**	**35.85%**
TermExtractor			
C1	78.81%	0.40%	0.79%
C2	76.16%	0.73%	1.44%
C3	75.82%	0.83%	1.64%
TermoStatWeb			
C1	67.05%	5.48%	10.14%
TermoStatWebV3-multi			
C1	76.93%	4.29%	8.13%
C2	76.32%	6.90%	12.66%
C3	77.98%	5.94%	11.04%
CA	83.46%	10.15%	18.10%
TermoStatWebV3-nomulti			
C1	**85.78%**	2.13%	4.15%
C2	**84.65%**	3.18%	6.14%
C3	**85.28%**	2.95%	5.69%
CA	**89.55%**	5.21%	9.84%
TS3-multi			
C1	70.63%	**12.25%**	**20.88%**
C2	53.66%	**19.43%**	**28.53%**
C3	66.41%	18.89%	29.41%
TS3-nomulti			
C1	82.84%	5.44%	10.21%
C2	77.32%	7.16%	13.10%
C3	82.46%	7.66%	14.01%

Table 3: Evaluation raw results

[9] `http://olst.ling.umontreal.ca/~drouinp/termostat_web`

	C1 (500k words)	C2 (1.5M words)	C3 (2.5M words)	CA (1.5M words)
Acabit	25,519	65,200	89,883	123,468
Fastr	36,932	failed	failed	failed
Jv1	-	-	2,129	-
YaTeA	47,499	125,245	184,729	206,733
MIG-YaTeA	25,634	64,823	91,907	119,137
TermExtractor	412	782	924	-
TermoStatWeb	9,490	failed	failed	-
TermoStatWebV3-multi	5,544	10,146	8,026	13,699
TermoStatWebV3-nomulti	2,166	3,417	3,081	5,056
TS3-multi	25,789	70,009	59,467	-
TS3-nomulti	6,899	10,056	10,563	-

Table 2: Number of returned terms

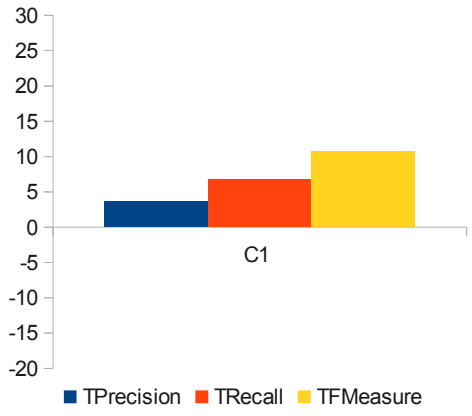

Figure 4: Increase of TPrecision, TRecall and TFMeasure from `TermoStatWeb` to `TS3-multi`

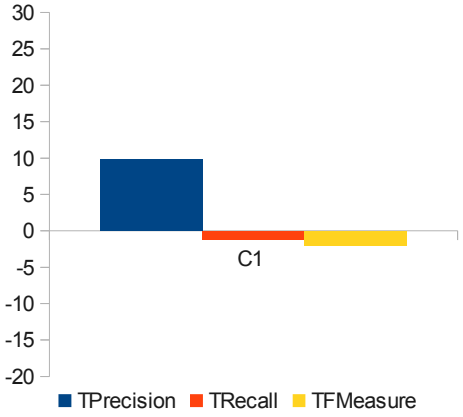

Figure 5: Increase of TPrecision, decrease of TRecall and TFMeasure from `TermoStatWeb` to `TermoStatWebV3-multi`

exhibits the evolution between `TermoStatWeb` and `TermoStatWebV3-multi`: in this case, improvements are more significant in terms of TPrecision.

Progression between `TS3-multi` and `TermoStatWebV3-multi` is presented on Figure 6. Significant improvements can be observed in terms of TPrecision, mostly for C2. However, TRecall notably decreases because `TermoStatWebV3` filters terms much more aggressively, focusing on precision.

The most significant improvement is obtained between `YaTeA` and `MIG-YaTeA`, as presented on Figure 7. TPrecision, TRecall and TFMeasure increase as the corpus grows.

6.3. Sensitivity to the corpus type

Regarding the sensitivity to the corpus type, results show that CA almost always provides better scores than C2 (although they both weight 1.5M words), both in term of TPrecision and in term of TRecall. The contrast between C2 and CA can be seen on Figure 8. These improvements can be explained because CA, composed of 7,030 abstracts where C2 contains only 106 patents, can potentially cover a larger field of knowledge.

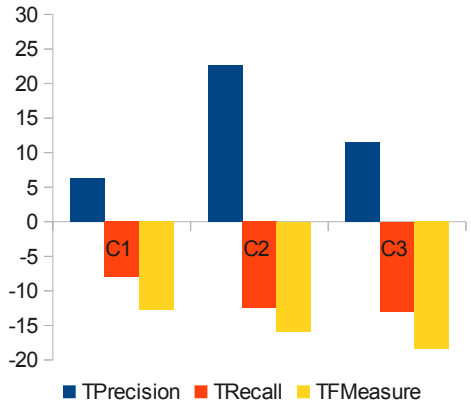

Figure 6: Increase of TPrecision, decrease of TRecall and TFMeasure from `TS3-multi` to `TermoStatWebV3-multi`

7. Conclusion

This paper presents the results of two consecutive initiatives on term extraction evaluation. The comparison of the results obtained during the two years was used to measure systems evolution. The protocol was easily reproduced for the second initiative. It provides inter-

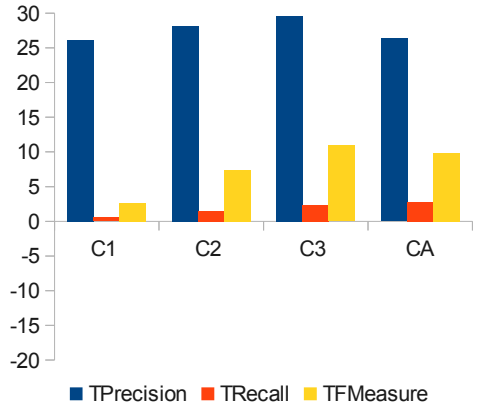

Figure 7: Increase of TPrecision, TRecall and TFMeasure from `YaTeA` to `MIG-YaTeA`

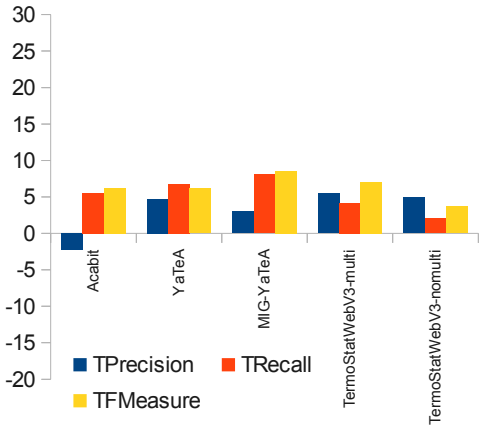

Figure 8: Impact of the corpus type: Evolution of TPrecision, TRecall and TFMeasure from `C2` to `CA`

esting keys to better characterize term extractors in their design choices, their scalability regarding corpus size and their robustness with respect to corpus type.

As the evaluation organizer, LIPN set up an adjudication phase after 2011 initiative, in order to get more reliable precision results. We actually noticed that relevant terms are missing in the gold standard, which potentially penalizes the systems. Tested term extractors have provided a total of more than one hundred thousand of distinct terms. Some of them are found by at least two systems and are not already present in the gold standard.

Some experts from INIST-CNRS are asked to check whether these terms are relevant or not. To ease this process, LIPN has developed a web based validation interface, which allows experts to view terms in context (the sentences where they appear). A screenshot of this interface is presented on Figure 9. For each term, experts have to decide if it is `GOOD`, `BAD` or `?`, a generic category for problematic terms. A free comment zone is available to indicate the correct form when needed.

These evaluations were also helpful for developers and for the gold standard provider. INIST-CNRS plans to exploit the results of the ongoing adjudication work to extend their thesaurus.

8. Acknowledgement

This work has been partially founded by OSEO under the Quæro program.

9. References

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing 5th International Conference on NLP, FinTAL 2006*, pages 380–387, Turku, Finland. Springer.

Béatrice Daille. 2003. Conceptual structuring through term variations. In F. Bond, A. Korhonen, D. MacCarthy, and A. Villacicencio, editors, *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 9–16.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Patrick Drouin. 2006. Termhood experiments: quantifying the relevance of candidate terms. *Modern Approaches to Terminological Theories and Applications*, 36:375–391.

Chantal Enguehard. 2003. Correct : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *Actes de la conférence Traitement Autmatique des Langues Naturelles (TALN) 2003*, pages 339–345, Batz-sur-Mer, France, Juin.

N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, and J. Adachi. 1999. The NTCIR workshop: the first evaluation workshop on japanese text retrieval and cross-lingual information retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages (1RAL'99)*.

Widad Mustafa El Hadi, Ismail Timimi, Marianne Dabbadie, Khalid Choukri, Olivier Hamon, and Yun-Chuang Chiao. 2006. Terminological resources acquisition tools: Toward a user-oriented evaluation model. In *Proc. of LREC'06*, pages 945–948, Genova, Italy, May.

Adeline Nazarenko, Haïfa Zargayouna, Olivier Hamon, and Jonathan Van Puymbrouck. 2009. Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique des Langues (TAL)*, 50(1 varia):257–281.

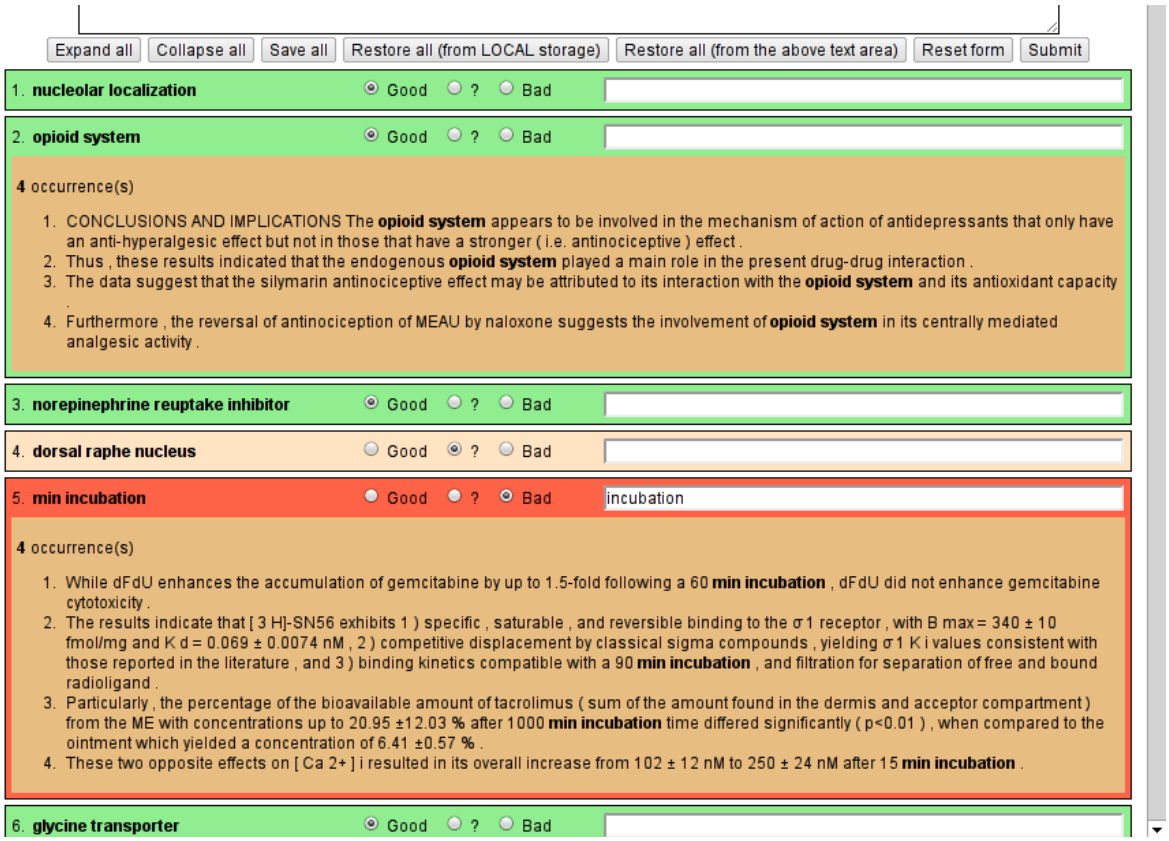

Figure 9: Screenshot of the validation interface

F. Sclano and P. Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. *Enterprise Interoperability II*, pages 287–290.

Haïfa Zargayouna and Adeline Nazarenko. 2010. Evaluation of Textual Knowledge Acquisition Tools: a Challenging Task. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 435–440, Valletta, Malte.