Evaluating term extraction

Adeline Nazarenko
Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 - Univ. Paris 13
99, avenue J.B. Clément
F-93430 Villetaneuse
nazarenko@lipn.univ-paris13.fr

Haïfa Zargayouna
Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 - Univ. Paris 13
99, avenue J.B. Clément
F-93430 Villetaneuse
haifa@lipn.univ-paris13.fr

Abstract

In contrast with other NLP tasks, only few and limited evaluation challenges have been carried out for terminology acquisition. It is nevertheless important to assess the progress made, the quality and limitations of terminological tools. This paper argues that it is possible to define evaluation protocols for tasks as complex as computational terminology. We focus on the core task of term extraction for which we propose evaluation metrics. We take into account the specificity of computational terminology, the complexity of its outputs, the application, the user's role and the absence of well-established gold standard.

Keywords

Term extraction, evaluation, terminological distance.

1 Introduction

Stemming from traditional terminology and natural language processing (NLP), computational terminology aims at building automatically or semi-automatically terminological resources from acquisition corpora. The growing needs in information management and localization make it more and more necessary to assist and automate terminological tasks.

A lot of terminological tools have been developed since the early research works of the 90s and many content management companies now rely on them [20]. However, despite the progress made, it remains difficult to get a clear idea of the maturity of computational terminology and to compare the proposed approaches. Unlike many other NLP fields, only few effort has been made to set up an evaluation protocol adapted to the specificity of terminological tasks.

We nevertheless argue that an evaluation is possible in computational terminology and that defining a clear and consensual evaluation protocol would benefit to the whole field. This paper focuses on technical aspects of evaluation putting aside the ergonomic and software aspects. We propose a comparative and application-independent evaluation protocol for monolingual term extraction as a first step towards more global and application-oriented evaluations.

Sections 2 and 3 review the first experiments that have been carried out for evaluating terminological tools and the difficulties that such evaluations raise. Sections 4 and 5 present our proposal: a protocol for evaluating term extractors and the specific metrics on which it relies. Section 6 describes experiments for meta-evaluating the proposed metrics.

2 State of the art

Various experiments have been made to evaluate terminological tools. Some were technologically oriented and took the form of evaluation challenges while others put focus on the application context.

2.1 Evaluation challenges

Traditionally, evaluation challenges aim at evaluating a set of systems on a specific task and for a common data set. The systems are compared to each other or wrt. a common data set. This enables the ranking of the systems for the specified task. The first evaluation challenges proposed interesting protocols.

The NTCIR[1] initiative was launched in 1999 and aimed at evaluating information retrieval and term recognition in Japanese [15]. The term recognition task (*TEMREC*) was decomposed into three subtasks: term extraction, key-word extraction and key-word roles analysis. The systems were evaluated on the basis of a standard set of terms. Unfortunately, this task was not very popular and it was eliminated in posterior NTCIR initiatives. [9] explains that TEMREC suffered from the limited number of participants and the absence of previous evaluation initiative for computational terminology.

CoRReCT proposed interesting data set and protocol [7]. The goal was to evaluate term recognition in corpora, a task that is close to controlled indexing. Participating systems took a corpus and a terminology as inputs and indexed the corpus with the terms of the input terminology and their variant forms. The incremental annotation of the corpus is an originality of CoRRecT.

CESART is the most complete challenge [14]. Three different tasks were planned (term extraction, controlled indexing and relation extraction) but only the first one gave rise to a real evaluation, due to the reduced number of participants. CESART nevertheless proposed an interesting protocol for term extraction. A gold standard list of terms and a corresponding acquisition corpus were chosen for a specific domain.

[1] http://research.nii.ac.jp/ntcir/

International Conference RANLP 2009 - Borovets, Bulgaria, pages 299–304

The systems were given the acquisition corpus as input. They extracted terms from the acquisition corpus and the resulting lists of terms were compared with the gold standard, using traditional precision and recall metrics. The originality of CESART was to consider term relevance on a 5-value scale rather than as a Boolean value. An adjudication phase allowed to add some missing relevant terms to the gold standard. Despite the small number of term extractors that participated in CESART, the challenge highlighted the heterogeneity of their results, especially regarding the length of the output lists of terms[2]. This reflects the diversity of the their methods and the differences in their underlying conceptions of what a terminology should be.

2.2 Application-based Evaluation

Smaller experiments have been carried out to evaluate the impact of terminological tools on parsing sublanguages [3], indexing and retrieving documents [6, 16, 19], building back-of-the-book indexes [1] or automatically translating specialised documents [12].

These experiments proposed original approaches to evaluate computation terminology without any terminological gold standard. The impact of terminologies is measured through the own application quality criteria.

Even if they reported very positive results, none of the mentioned experiences gave a global idea of the impact of terminological tools on an application. Coming to such a conclusion would require, for each application, to integrate various term extractors in various application systems, to really assess the impact of the first ones on the second ones and to compare various extracting methods independently on how they are integrated.

This is the reason why we consider that application oriented evaluations are more complex to set up than technological ones, which are addressed in this paper.

3 Evaluation Difficulties

Despite these first evaluation experiences, no comprehensive and global framework has yet been proposed for computational terminology as there exist for many other NLP fields. Beside economic factors, it seems that evaluating terminology acquisition raises some specific intrinsic difficulties.

Heterogeneity of terminology acquisition tools Computational terminology quickly developed in the 1990s and diversified into many subtasks at the end of the decade [4, 8, 5].

The first works focused on term extraction and a large variety of tools have been developed. Some of them rely on a morphological and syntactic analysis to identify the textual units that can be considered as terms. Others are based on statistics and word cooccurrences. Statistical extractors generally produce ranked lists of terms while linguistic ones output unordered ones. Depending on the extractor, focus is put on term well-formedness or on analysis robustness and result coverage. The results produced by these tools are therefore difficult to compare, which makes evaluation more complex.

Following term extraction, many additional terminological functionalities have been proposed. Terminology acquisition now covers a large variety of tools and this diversification also hinders evaluation. We argue that, for evaluation purposes, computational terminology must be split into several, clearly identified, independent and elementary tasks. We focus on the first one in this paper: term extraction.

Complexity of terminological resources The diversity of terminological tools reflects in the resulting terminological resources. Terms are often multi-word units that follow various variation rules in corpora. Terms can also be related to each others by morphological (*schedule/schedules*), synonymy (*plan/schedule*) or hyponymy (*time schedule/schedule*) relations. The quality of the resulting terminology cannot be measured with a single metric, as it is the case with word error rate in speech recognition, for instance. This also leads to decompose term acquisition into several tasks to be evaluated independently.

Gradual relevance The quality of extracted lists of terms would be easy to measure if terms were either relevant or irrelevant because one could rely on the well-known evaluation metrics such as recall and precision. Unfortunately, the underlying hypothesis that relevance is a binary value does not hold for term extraction: a term candidate can be different from a standard term but nevertheless close to it and interesting. In biology, it was reported that extractors often propose incomplete terms, such as *core rna*, which are nevertheless kept in a modified form (*core rna polymerase*) by terminologists [2]. CESART strategy to avoid this binary relevance constraint consisted in defining four levels of precision.

Gold standard variability A major difficulty comes from the fact that, even for a same domain and corpus, different terminologists will produce different terminologies, which reflects their different points of view, different terminological traditions and different choices regarding the granularity of the description. There is usually not a single gold standard but a large set of "acceptable solutions". To palliate the variability of the gold standards, we propose to tune the output of the terminology chosen as standard. This is an original way to free evaluation from the imperfection and relativity of the gold standard. It plays the same role as adjudication but it can be used on a larger scale and it is less costly.

Application role The application for which the terminology is designed also determines what must be evaluated. Although the role of application is important, we do not propose an application-oriented evaluation here. We focus on technological evaluation, which we consider as a first generic step towards more global evaluations.

Interaction Since terminology acquisition is seldom seen as a fully automatic process, terminological tools are generally assisting tools that integrate

[2] The extractors were evaluated on the basis of their first 10,000 terms but some systems outputted 20 times more.

the terminologist role in the resource building process. This also makes evaluation complex because it is difficult to distinguish what comes from the acquisition tools and what results from the terminologist's work. It is nevertheless interesting to compare the output of the system with its amended version. It gives and interesting feed-back for interactive tools.

4 Protocol

Despite those difficulties, we argue that it is possible to design a generic protocol to evaluate term extraction.

4.1 Comparative Protocol

We identified three scenarios for evaluating terminological tools and term extraction in particular. They all rely on comparison. The first scenario compares the output of a term acquisition tool with an independent gold standard. The second takes interaction into account and measures the effort required to turn the draft terminology into an acceptable one. This effort corresponds to the minimal number of elementary operations (term deletion, addition or modification) that allows to transform the draft terminology into the final one. The third scenario evaluates the terminology indirectly through an application (*e.g.* machine translation) using the application own quality criteria.

The comparative protocol that we propose can be used in the two first scenarios: the output of a system (O) is compared with a gold standard (R) that is either an external resource or the validated output[3].

4.2 Choice of the gold standard

Even if, as mentionned above, *gold standard* is variable. There are several ways to build it. One may reuse an existing terminology as in CESART but this terminology may be only partially related to the acquisition corpus. A costly solution consists in asking one or several terminologists to build that standard terminology out of the acquisition corpus. A third solution consists in having the terminologists validating the results of the systems. In that case, the gold standard is the fusion of the validated outputs of the systems, as in CoRReCT. This last solution is not exhaustive but it is less costly than the second one.

The metrics proposed in the following for term extraction are compatible with any of these gold standard building methods.

4.3 Sub-tasks Decomposition

As mentioned above, term extraction is not the only terminological task. Other tasks such as variation calculus, relation extraction, term normalisation must be considered as well, even if they are left apart here. As far as evaluation is concerned, those tasks must be considered as independently as possible. This is especially important for term extraction and variation calculus, which consists in clustering terms that are variant forms of each others. The term lists output by the systems must be considered as unstructured lists even if the systems propose some variation relations among the terms. Those systems have to be evaluated twice (for the term extraction in the first place and for the variation in a second phase) but the quality of one task must not affect the evaluation of the other.

[3] In that case, there is no need for a standard terminology but recall cannot be measured or only partially because validation consists in deleting rather than adding terms.

5 Metrics

As mentioned above, traditional metrics of precision and recall are not appropriate for term extraction evaluation. One problem is that term relevance is a gradual rather than a binary notion and that one cannot expect all extractors or terminologists to deliver ranked list of terms. This led us to stem relevance on a terminological distance. A second problem is that no terminological standard can be considered as a stable and unique gold standard. We propose to overcome that difficulty by tuning the system output to the granularity of the chosen gold standard in order to avoid arbitrarily favouring one system against another.

It is important however to have simple and well-known metrics [13], so we adapt traditional metrics of precision and recall rather than defining new ones[4].

The proposed metrics are implemented in a tool, called TermoMeter. It takes two unstructured term lists as input (the output O of a term extractor and the gold standard R without any hypothesis on the type, length and granularity of that gold standard) and it computes the terminological precision and recall (resp. TP and TR) of O wrt. R. For a perfect system ($O = R$), TermoMeter gives $TP = TR = 1$. A system that extracts in addition terms that are close to the gold standard is not penalised or only a little bit. TermoMeter gives $TP = TR = 0$ for systems that give only irrelevant terms ($S \cap R = \emptyset$).

Let us consider the cases where $R = \{$*data base*$\}$ and we have the following output lists: $O_1 = \{$*data base, data bases*$\}$, $O_2 = \{$*data bases*$\}$ et $O_3 = \{$*data base, table of content*$\}$. We expect O_1 and O_2 to be of similar quality: O_1 gives the term of R along with a second redundant but relevant one; O_2 gives a single term that is close but not exactly that one of R. O_3 has a lower quality since the extra term, which is to too far from the term of R, is considered as noise.

5.1 Term Distance

Several methods have been proposed to measure word distance. They rely on character string comparison, on Levenshtein distance and the minimal number of editing operations required to transform one word into another [17], or on morphological and linguistic transformation rules. However, since terms are generally multi-word units having their own variation rules, two independent levels must be considered in term distance: the word level (*base* vs *bases*) and the phrase level (*file system* vs *disk file system*). [18] proposed to

[4] Recall that:

$$precision = \frac{|O \cap R|}{|O|} \qquad recall = \frac{|O \cap R|}{|R|}$$

where $|O|$ is the number of elements retrieved by the system, $|O \cap R|$ is the number of relevant elements retrieved by the system, and $|R|$ is the number of elements in the gold standard.

exploit Levenshtein distance to compute distances on these two levels in an homogeneous way but our approach differs from this work because we avoid relying on external linguistic knowledge. This allows to keep computation simple, which is important if one considers the number of distances that must be computed in real evaluation conditions. To keep distance unbiased, it is also important to avoid using for evaluation the linguistic knowledge (such as term variation rules or POS-tagging) that may be used by extractors.

We define a first distance on character strings (d_s. It is a normalised Levenshtein distance, where the sum of the costs of the elementary editing operations (character insertion, deletion, substitution) required to transform a string into another is divided by the length of the longer string. The normalisation allows to compare the distances of various string pairs. If all the elementary operations have an equal cost of 1, TermoMeter gives the following distances :

$$d_s(base, bases) = 1/5 = 0.2$$
$$d_s(base, basement) = 4/8 = 0.5$$
$$d_s(base, relational) = 9/10 = 0.9$$

Of course, that distance does not always match with linguistic intuition but it must be evaluated globally through the evaluation of a list of terms (see Sec. 6) and not on individual pairs of words.

The distance on complex terms is based on the same principle, except that editing operations apply on words instead of characters. The distance between two complex terms (d_c) is the sum of the elementary operations (word insertion, deletion or substitution) required to transform a term into another. To search for the best word alignment that minimizes the global cost, TermoMeter relies on Hungarian algorithm [11]. The cost of an insertion or deletion is 1 while the cost of a substitution is equal to the normalised Levenshtein distance (d_s) of the corresponding words. This gives the following distances, for instance:

$$d_c(\textit{data base, data bases}) = 0.1$$
$$d_c(\textit{relational data base, data base}) = 0.33$$
$$d_c(\textit{relational data base, web site}) = 0.88$$

This measure allows to take into account word permutations that are frequent in term variation as for *expression of gene* and *gene expression*. Its drawback is that it relies on a necessarily arbitrary word segmentation method.

Finally the term distance (d_t) is defined as the mean of the distances on strings and on complex terms :

$$d_t(t_1, t_2) = (d_s(t_1, t_2) + d_c(t_1, t_2))/2$$

d_t is a normalised distance ranging between 0 and 1. The following examples show that the d_c factor allows for permutation but that the d_s factor both attenuates that effect and limits the impact of segmentation choices:

$$d_t(\textit{precise gene localization, precise localization of gene}) = (10/25 + 1/4)/2 = 0,34$$
$$d_t(\textit{porte folio, portefolios}) = (1/11 + 3/11)/2 = 0,4$$

This term distance is robust, quick to compute and easy to interpret. It does not require any external knowledge and is language independent. It takes a gradual relevance into account without considering the eventual variants that systems may propose and that must be evaluated on a separate task.

5.2 Gradual relevance

The terminological precision and recall metrics take that gradual relevance into account. The global relevance of a term list with respect to a gold standard is defined as the function $Pert(O, R)$ that verifies

$$|O \cap R| \leq Pert(O, R) \leq min(|0|, |R|)$$

and reflects the global distance between O and R. It is based on the term distances $d_t(e_o, e_r)$ between the terms of the output O and that of the gold standard R. The relevance of a term e_o of O is based on its distance to the closest term e_r of R:

$$pert_R(e_s) = \begin{cases} 1 - min_{e_r \in R}(d_t(e_o, e_r)) \\ \text{ if } min_{e_r \in R}(d_t(e_o, e_r)) < \tau \\ 0 \text{ otherwise} \end{cases}$$

where τ is a threshold such that if the distance between two terms is superior to τ, they are considered as totally different.

5.3 Output tuning

Since any terminology is a relative gold standard, it would be artificial to compare directly the output of the systems with it. It might favour the systems that would have made "by chance" the same granularity choice. The evaluation scores and system ranking might be too dependent on the gold standard. To avoid this problem, the output is transformed to find its maximal correspondence with the gold standard, which means that the output is tuned to the terminological type and granularity of the gold standard.

Since several output terms may correspond to the same standard term, they are considered all together. The precision and recall measures are not computed directly on O but on a partition of O that is defined relatively to R. This partition $\mathcal{P}(\mathrm{O})$ is such that any part p of $\mathcal{P}(\mathrm{O})$ either contains a set of terms of O that are close to the same term of R and with a distance inferior to the threshold τ, or contains a single term that matches with no term of R:

$$p = \begin{cases} \{e_1, e_2, ..., e_n\} \\ \text{ if } (\exists e_r \in R)((\forall i \in [1, n])(\forall e'_r \in R) \\ (d_t(e_i, e'_r) \geq d_t(e_i, e_r))(d_t(e_i, e_r) \leq \tau)) \\ \{e\} \text{ if } (\nexists e_r \in R)(d_t(e, e_r) \leq \tau)\} \end{cases}$$

$$\text{where } e \in O \text{ and } \forall i \in [1, n])(e_i \in O)$$

The relevance of a part p of $\mathcal{P}(\mathcal{O})$ wrt. R is defined as follows[5].

$$Pert_R(p) = max_{e \in p}(pert_R(e))$$

Terminological precision (TP) and recall (TR) are defined as follows:

$$TP = \frac{Pert(O, R)}{|\mathcal{P}(\mathcal{O})|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{O})} Pert_R(p)}{|\mathcal{P}(\mathcal{O})|}$$

[5] Note that the relevance is null if p contains only one term of O with a distance superior to τ to any term of R.

$$TR = \frac{Pert(O,R)}{|R|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{O})} Pert_R(p)}{|R|}$$

As expected, TermoMeter gives $TP = TR = 1$ for a perfect system, TP and TR tend towards 0 for systems that extract mainly irrelevant terms and TR decreases when the size of the gold standard increases wrt. that of the output.

6 Meta-evaluation of metrics

As it has been done for machine translation [10], it is important to meta-evaluate the proposed metrics before starting to use them in real evaluation conditions, either challenges or benchmark comparisons.

To achieve this meta-evaluation at low cost, existing independent data have been exploited. Two series of tests have been made on terminological results provided by MIG laboratory of INRA. These data sets are used to test the robustness of TermoMeter and its adequacy to initial specifications.

6.1 Terminological *vs.* usual metrics

The first experiment is based on the following data: (i) an English corpus specialised in Genomics and composed of 405,000 words, (ii) the outputs of three term extractors in which only frequent candidate terms (more than 20 occurrences) have been kept to alleviate the terminologist's work. The outputs of the systems S_1, S_2 and S_3 respectively contain 194, 307 and 456 candidate terms and (iii) a gold standard (GS) of 514 terms, which has been built by asking a terminologist to validate the outputs of the three extractors[6].

Table 1 presents the evaluation of these systems wrt. the gold standard. As expected, the terminological measures (TP and TR) follow the same curves as the classical ones (P and R), but they are higher, which proves that the terminological measures take into account the gold standard approximation. A difference of 10 points in F-measure (F) is significant.

	P	R	F	TP	TR	F
GS	1.0	1.0	1.0	1.0	1.0	1.0
S_1	0.71	0.42	0.52	0.95	0.48	0.63
S_2	0.77	0.68	0.72	0.94	0.70	0.80
S_3	0.76	0.28	0.40	0.95	0.34	0.50

Table 1: *Results of the output of three term extractors, $\tau = 0.4$ for terminological measures (TP, TR)*

The output partitioning leads to cluster mostly morphological (singular/plurial) or typographic (lower/upper cases) variants. The analysis of the incomplete terms (8.5% of the extracted terms as reported by the terminologist) shows that most of them are considered as close to the corresponding complete terms that the terminologist has added to the gold standard. For instance, *acid residue* is considered as close to *amino acid residues* with a distance of 0.35. In fact, four terms are clustered in the same output part associated to *amino acid residues*: acid residue (distance=0.35), amino acid residue (distance=0.05), acid residues (distance=0.29), amino acid residues (distance=0.0).

[6] The terminologist was allowed to supplement the incomplete terms.

6.2 Large scale experiment

The second experiment exploited the following data: (i) an English patent corpus in the agrobiotech domain, (ii) the raw output of an extractor (4,200 candidate terms extracted from the corpus) without any ranking or filtering. It contains almost and (iii) a gold standard of 1,988 terms resulting from the validation of the extractor output by two terminologists[7].

This experiment allows to analyse globally the behaviour of TermoMeter and its metrics on a large scale sample: comparing the output to the gold standard led to compute around $8 * 10^6$ term distances, which required only few minutes on a standard PC.

The results are presented on Figure 1[8]. It shows the correlation between τ (the threshold above which a candidate term is not clustered with others) and terminological precision (Fig. 1 a.). When $\tau = 0$, there is no clustering at all ($TP = precision$) but TP increases with τ. The threshold value has a direct impact on the size of the output partition (Fig. 1 b.): the higher the threshold is, the more numerous are the terms that are clustered and match with the gold standard. When τ has its maximal value, all the candidate terms match with the gold standard and the terminological precision cannot get higher. The shapes of the curve show that it should be possible to determine the threshold value automatically (between 0.4 and 0.5 in the present case).

The relative quality of the system output and the lists validated by a single terminologist have also been measured. Three output lists of terms have been considered: the raw system output (O_r) and the outputs validated by the two terminologists independently (V_1 and V_2). They all have been compared with the gold standard (V_{12}) that resulted from the join validation of the two same terminologists. Table 2 shows that the expected quality ranking of the three outputs is verified:

$$TP(S_b) < TP(V_1) < TP(V_{12})$$
$$TP(S_b) < TP(V_2) < TP(V_{12})$$

and that the first terminologist judgement is closer to the gold standard than that of the second one.

	S_b	V_1	V_2	V_{12}
TP	0.55	0.91	0.97	1.0

Table 2: *Terminological precision, $\tau = 0.4$*

7 Conclusion

After 15 years of research in computational terminology it is important to assess the maturity of terminological tools. Evaluating term extraction is a first step in that direction.

[7] Inter-annotator variations (11% of the candidate terms) have been solved through discussion.

[8] Only precision measures are presented. The gold standard cannot be used to measure recall since terminologists validated or deleted terms without adding any new one.

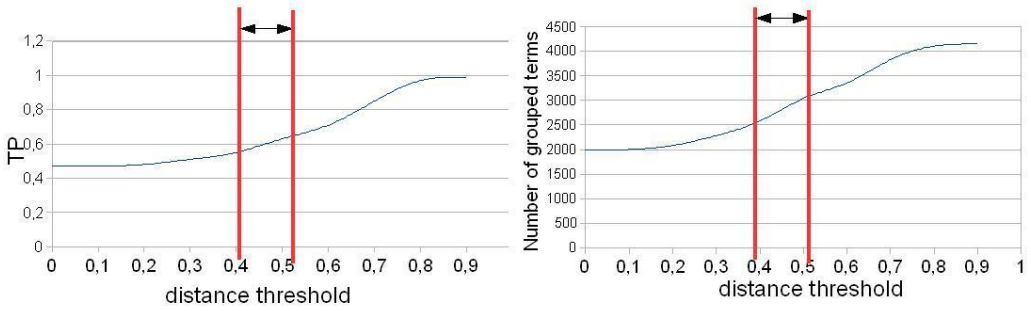

Fig. 1: *Curves of terminological precision (TP) (a) and the number of candidate terms matching with the gold standard (b) wrt. the threshold values.*

The fact that terminological tools are often assisting tools for terminologists and application dependent, the relativity of any standard terminology, the complexity of terminological resources make the evaluation difficult in computational terminology. However, this paper proposes an evaluation protocol and the associated metrics that take into account terminological specificity and nevertheless enable to set up comparative evaluations in a simple way. The proposed terminological measures differ from traditional precision and recall in two ways: they take into account the gradual relevance of terms and the relativity of the gold standard.

The first meta-evaluation experiments have shown that the TermoMeter tool globally behaves as expected on large scale term lists and that it gives a more precise evaluation of terminological extractors than traditional measures.

Further work will consist in setting up evaluation experiments for term extraction and to define adequate protocols for other terminological tasks such as term variation calculus and term relation extraction.

Acknowledgments. The authors would like to thank Olivier Hamon (ELDA-LIPN) and Jonathan van Puymbrouck (LIPN) for fruitful discussions and Sophie Aubin (MIG) for helpful material used in meta-evaluation. This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] T. Ait El Mekki and A. Nazarenko. An application-oriented terminology evaluation: the case dof back-of-the-book indexes. In R. C. *et al.*, editor, *Proceedings of the Workshop "Terminology Design: Quality Criteria and Evaluation Methods" (LREC-TerEval)*, pages 18–21, Genova, Italy, May 2006.

[2] S. Aubin. Comparaison de termes extraits par acabit, nomino, syntex de fréquence supérieures ou égales à 20. Livrable 3.2, Projet ExtraPloDocs, INRA-MIG, 2003.

[3] S. Aubin, A. Nazarenko, and C. Nédellec. Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 89–93, Borovets, Bulgaria, 2005.

[4] M. Cabré Castellv, R. Estopà, and J. Vivaldi Palatresi. Automatic term detection: A review of current systems. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme, editors, *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam, 2001.

[5] B. Daille, K. Kageura, H. Nakagawa, and L.-F. Chien, editors. *Terminology. Special issue on Recent Trends in Computational Terminology*, volume 10. John Benjamins, 2004.

[6] B. Daille, J. Royauté, and X. Polenco. Evaluation d'une plateforme d'indexation des termes complexes. *Traitement Automatique des Langues*, 41(2):396–422, 2000.

[7] C. Enguehard. Correct : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, pages 339–345, Nancy, 2003.

[8] C. Jacquemin and D. Bourigault. Term extraction and automatic indexing. In R. Mitkov, editor, *Handbook of Computational Linguistics*, chapter 19, pages 599–615. Oxford University press, Oxford, GB, 2003.

[9] K. Kageura, T. Fukushima, N. Kando, M. Okumura, S. Sekine, K. Kuriyama, K. Takeuchi, M. Yoshioka, T. Koyama, and H. Isahara. Ir/ie/summarisation evaluation projects in japan. In *LREC2000 Workshop on Using Evaluation within HLT Programs*, pages 19–22, 2000.

[10] P. Koehn, N. Bertoldi, O. Bojar, C. Callison-Burch, A. Constantin, B. Cowan, C. Dyer, M. Federico, E. Herbst, H. Hoang, C. Moran, W. Shen, and R. Zens. Factored translation models. In J. H. University, editor, *CLSP Summer Workshop Final Report WS-2006*, 2006.

[11] B. H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[12] P. Langlais and M. Carl. General-purpose statistical translation engine and domain specific texts: Would it work? *Terminology*, 10(1):131–152, 2004.

[13] A. F. Martin, J. S. Garofolo, J. C. Fiscus, A. N. Le, D. S. Palett, and M. A. P. ang Gregory A. Sanders. Nist language technology evaluation cookbook. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[14] W. Mustafa el Hadi, I. Timimi, M. Dabbadie, K. Choukri, O. Hamon, and Y. Chiao. Terminological resources acquisition tools: Toward a user-oriented evaluation model. In *Proceedings of the Language Resources and Evaluation Conference (LREC'06)*, pages 945–948, Genova, Italy, May 2006.

[15] National Center for Science Information Systems, editor. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

[16] A. Névéol, K. Zeng, and O. Bodenreider. Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for medline. In *Proceedings of the AMIA Annual Symposium*, pages 589–593, 2006.

[17] P. M. Sant. Levenshtein distance. in *Dictionary of Algorithms and Data Structures* [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology, April 2009 2004. (accessed 2 April 2009) Available from: http://www.itl.nist.gov/div897/sqg/dads/HTML/rootedtree.html.

[18] A. Tartier. *Analyse automatique de l'évolution terminologique : variations et distances*. PhD thesis, Université de Nantes, 2004.

[19] N. Wacholder and P. Song. Toward a task-based gold standard for evaluation of np chunks and technical terms. In *Proceedings of HTL-NAACL*.

[20] D. Zielinski and Y. R. Safar. t-survey 2005: An online survey on terminology extraction an terminology management. In *Proceedings of Translating and the Computer*.