Evaluation of Textual Knowledge Acquisition Tools: a Challenging Task

Haïfa Zargayouna, Adeline Nazarenko

LIPN
Université Paris 13 - CNRS (UMR 7030)
99, avenue Jean-Baptiste Clément -
F-93430 Villetaneuse, France
firstname.lastname@lipn.univ-paris13.fr

Abstract

A large effort has been devoted to the development of textual knowledge acquisition (KA) tools, but it is still difficult to assess the progress that has been made. The lack of well-accepted evaluation protocols and data hinder the comparison of existing tools and the analysis of their advantages and drawbacks. From our own experiments in evaluating terminology and ontology acquisition tools, it appeared that the difficulties and solutions are similar for both tasks. We propose a general approach for the evaluation of textual KA tools that can be instantiated in different ways for various tasks. In this paper, we highlight the major difficulties of KA evaluation, we then present our proposal for the evaluation of terminologies and ontologies acquisition tools and the associated experiments. The proposed protocols take into consideration the specificity of this type of evaluation.

1. Introduction

A large effort has been devoted to the development of textual knowledge acquisition (KA) tools, but it is still difficult to assess the progress that has been made. The results produced by these tools are difficult to compare, due to the heterogeneity of the proposed methods and of their goals.

Various experiments have been made to evaluate terminological and ontological tools, some took the form of evaluation challenges while others put focus on the application context.

Some challenges related to terminology have been set up (*e.g.* NTCIR[1] and CESART (Mustafa El Hadi et al., 2006)) but they did not have the popularity they deserved and were not renewed. Even if evaluation of ontology acquisition tool has its own workshop (EON[2]), no challenge has been organized and there is still no well-accepted evaluation protocol and data.

Application-based evaluation were carried out in order to evaluate the impact of the acquired knowledge in practice; *e.g.* for document indexing and retrieval (Névéol et al., 2006; Wacholder and Song, 2003; Köhler et al., 2006), automatic translation (Langlais and Carl, 2004), query expansion (Bhogal et al., 2007). Nonetheless none of the mentioned experiences gave a global idea of the impact of these semantic resources on the applications in which they were exploited.

These experiments show that in terminology as well as in ontology acquisition, it remains difficult to compare existing tools and to analyse their advantages and drawbacks.

From our own experiments in evaluating terminology and ontology acquisition tools, it appeared that the difficulties and solutions are similar for both tasks. We propose a unified approach for the evaluation of textual KA tools that can be instantiated in different ways for various tasks. The main originality of this approach lies in the way it takes into account the subjectivity of evaluation and the relativity of gold standards. The output of the systems is automatically tuned to the chosen gold standard instead of being compared to several human judgements as it can been done for the evaluation of machine translation.

In this paper, we highlight the major difficulties of KA evaluation, we then present a unified proposal for the evaluation of terminologies and ontologies acquisition tools and the associated experiments. The proposed protocols take into consideration the specificity of this type of evaluation.

2. Why are KA tools difficult to evaluate?

Various difficulties can explain the fact that no comprehensive and global framework has yet been proposed.

Complexity of artifacts The KA tasks themselves are difficult to delimit because their output are complex artifacts. For instance, terminology and ontology acquisition tasks are related as soon as one considers the terminological labels that are associated with ontological concepts. Even considered independently, a terminology and an ontology have several components (at least terms, variants and semantic relations for terminologies; concepts, hierarchies and roles for ontologies) which cannot be evaluated all together.

Heterogeneity of tools Even for a given KA task, there exists a wide variety of tools. For instance, a term extractor may produce twenty times as many terms as another for the same acquisition corpus. Some focus on the precision of the results while others favor the recall. Some extract only biword terms, some also consider more complex compounds. The same kind of heterogeneity can be observed for semantic class acquisition where the size and number of classes can vary from one system to another.

Gold standards variability It is difficult and unrealistic to establish a unique gold standard as the knowledge extracted depends on domains and applications. Even if textual corpora help to delimit the scope of interpretation, there is a multitude of acceptable solutions that vary from

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

[2] Evaluation of Ontologies for the Web

one expert to another. (Jorge Vivaldi and Lorente, 2008) reported a low agreement rate among human judgements and an important score variation from one domain to another. (Pazienza et al., 2005) relates that variability to the fuzziness of termhood definition.

Role of human interaction KA tools are often designed as assisting tools which output a draft model that can then be amended by a terminologist or a knowledge engineer rather than as fully automatic tools. In that context, the contribution of the tools is difficult to assess: serendipity can compensate a lack of precision in the results from an expert point of view.

Limitations of classic measures The quality of the output knowledge would be easy to evaluate if it relied on a binary judgement (relevant *vs.* irrelevant). Reality is more contrasted: a candidate term can be different from a standard one but nevertheless close to it and useful. A semantic class can be interesting even if it does not match exactly a standard concept. This gradual relevance is not captured by classical measures such as precision and recall.

3. Proposal: a unified approach

From our own experiments in evaluating terminology and ontology acquisition tools (Nazarenko and Zargayouna, 2009; Ben Abbès et al., 2010), it appeared that the difficulties and solutions are similar for both tasks. We present a unified approach for the evaluation of textual KA tools that can be instantiated in different ways for various tasks.

3.1. Task decomposition

To go beyond a black-box evaluation of terminological and ontological tools, it is important to evaluate their components independently. We consider that these tools are based on few elementary functionalities and that they should be evaluated along these various axes independently.
The basic functionalities of a *terminological tool* are the following in increasing complexity order:

- Term extraction: the system takes a corpus as input and outputs a list of mono or multi-word terms.
- Terminological variation calculus: the system takes a flat term list as input and outputs clusters of variant terms.
- Terminology structuring: the system takes a flat term list as input and proposes a semantic network according to the semantic relations extracted from a corpus.

Of course one can consider additional tasks such as term normalisation, extraction of specific types of semantic relations or term ranking as proposed by (Zhang et al., 2008).
Ontology acquisition task can also be decomposed. In (Ben Abbès et al., 2010) we propose to decompose it into three subtasks:

- Semantic class or concept acquisition: the system takes a corpus as input and outputs a list of (possibly overlapping) semantic classes. A semantic class is a set of terms.
- Building concept hierarchies: the system takes a corpus and a list of concepts as inputs and outputs a hierarchy of concepts.
- Role extraction: the system takes a corpus and a list of concepts and outputs a list of roles between concepts.

Evaluation can also take complementary subtasks into account: (i) evaluation of the ontology population as proposed by (Tao et al., 2009), (ii) evaluation of scalability, performance tradeoffs and persistance (Ma et al., 2006) or (iii) evaluation of formal properties such as inconsistency (Guarino and Welty, 2002).

Even if these functionalities correspond to abstract tasks, which results are generally not exploited in isolation, this decomposition helps to compare systems which rely on different methods and have heterogeneous goals. The following proposals concern the evaluation of two specific functionalities: term extraction and semantic class acquisition, which constitute the basis of terminology and ontology acquisition respectively.

3.2. Specific Measures

It is important to apply specific measures of scoring that take into account the complexity of the task to be evaluated, but these measures have to be generic and easy to interpret (Martin et al., 2004). The measures we propose can be used to compare the results of a system with a gold standard as well as to confront the results of a system before and after human validation (see figure 1). In the latter case, the validated output is then considered as a gold standard.

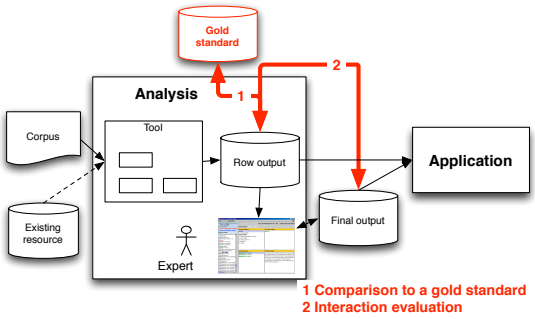

Figure 1: Two different different scenarios for comparative evaluation

For term extraction, the gold standard is a flat list of terms. For semantic class acquisition the gold standard is a hierarchy of concepts. A concept is denoted by an identifier and a list of terms that represents its lexical realisations[3].
Since no terminology or ontology can pretend to be a "good" gold standard, it would be artificial to compare directly the output of the systems with it. In order to avoid the scoring to be too dependent on the gold standard or a specific system behavior, the output is transformed to find

[3]We focus on two levels: (i) the lexical level that represent how terms convey meanings and (ii) the conceptual level that represent conceptual relations between terms (Maedche and Staab, 2002).

its maximal correspondance with the gold standard, which means that the output is tuned to the specific type and granularity of the chosen gold standard (see figure 2). This tuning is performed instead of considering several human judgements or revising the gold standard on the basis of the systems outputs.
We therefore propose to adapt precision and recall measures to take into account (i) the gradual relevance of KA judgements and (ii) the systems heterogeneity and gold standard variability.

4. Precision and recall measures

The adapted precision and recall formula are as follows:

$$A-precision = \frac{Rel(T(O), GS)}{|T(O)|} = \frac{\sum_{e'_o \in T(O)} rel_{GS}(e'_o)}{|T(O)|}$$

$$A-recall = \frac{Rel(T(O), GS)}{|GS|} = \frac{\sum_{e'_o \in T(O)} rel_{GS}(e'_o)}{|GS|}$$

where $Rel(T(O), GS)$ is the global relevance of the tuned output ($T(O)$) with respect to the gold standard (GS), $|T(O)|$ is the size of tuned output, $|GS|$ the size of the gold standard and $rel_{GS}(e'_o)$ the relevance of an element of the tuned output (e'_o) with respect to the gold standard. It is easy to verify that these measures correspond to traditional precision and recall when the relevance is binary and the output not tuned[4]. The following subsections explain how these measures can be computed.

4.1. Matching elements

The first step consists in defining a matching between the output elements ($e_o \in O$) to those of the gold standard ($e_{gs} \in GS$). The goal here is to find the best correspondence taking into account both exact and approximative matches.

Term extraction For term extraction, the matching is based on a terminological distance between the terms of the output list and those of gold standard. This distance d_t (Nazarenko and Zargayouna, 2009) is computed as the mean of string and complex term distances that is based on a normalized edit distance and takes into account word permutation.
The matching is a function that links each output term to its closest GS term e_{cgs}: $match(e_o, e_{cgs})$ holds if and only if $e_{cgs} = \arg\min_{e_{gs} \in GS} d_t(e_o, e_{gs})$ and if $d_t(e_o, e_{gs}) < \tau$, where τ is a threshold beyond which an output term is considered as noise.

Semantic class acquisition In that case, the matching gives a many-to-many relation that holds for each pair of output elements (semantic classes) and GS ones (concepts) sharing at least one term.

[4] $|T(O)| = |O|$ for unchanged outputs and $\sum_{e'_o \in T(O)} rel_{GS}(e'_o) = |O \cap GS|$ if $\forall e'_o, rel_{GS}(e'_o) \in \{0, 1\}$.

4.2. Output tuning

The output is tuned on the basis of the identified matching relations. The goal is to adapt the output results to the granularity and specificity of the gold standard.

Term extraction The output list of terms is clustered according to the matching relations identified above (see figure 3). All the output terms that match the same standard term are clustered and considered all together. The tuned

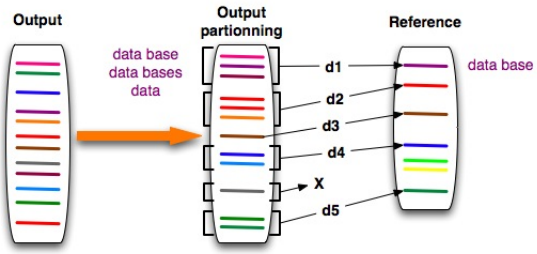

Figure 3: Term tuning

output $T(O)$ is such that any part e'_o of $T(O)$ either contains a set of terms of O that are close to the same term of GS and within a distance inferior to the threshold τ, or contains a single term that matches no term of GS:

$$e'_o = \begin{cases} \{e_1, e_2, ..., e_n\} \\ \text{if } (\exists e_{gs} \in GS)((\forall j \in [1, n])(match(e_j, e_{gs}) \\ \{e\} \text{ if } (\nexists e_{gs} \in GS)(match(e, e_{gs})\} \end{cases}$$

where $e \in O$ and $\forall j \in [1, n](e_j \in O)$

Semantic class acquisition For semantic class extraction, the tuning is based on the regrouping of output classes. Several cases are identified :

- In case of a 1:1 matching relation where the output class exactly matches one concept of the gold standard, no particular transformation is needed.
- In case of a 1:n matching relation where an output class matches different concepts of the gold standard, it is split into different classes (*splitting* transformation).
- In case of n:1 matching, several output classes match the same concept of the gold standard and they are merged (*merging* transformation).
- In case of a n:m matching relation, the splitting operation is performed before the merging one.

Figure 4 shows an example of splitting. The class $CS1$, which matches three different GS concepts ($CR1$, $CR2$ and $CR3$), is split into three subclasses, each of which sharing terms with a specific target concept. If there are terms in $CS1$ that do not appear in any concept of the GS, they are considered as noise. Since they cannot be assigned to any specific subclass, they are kept in all splitted subclasses. This is the case of term $t5$ in figure 4.
On the opposite, the classes that match the same GS concept are merged. Figure 5 shows an example of merging three output classes ($CS2$, $CS3$ and $CS4$) that share terms with a common GS concept $CR4$.

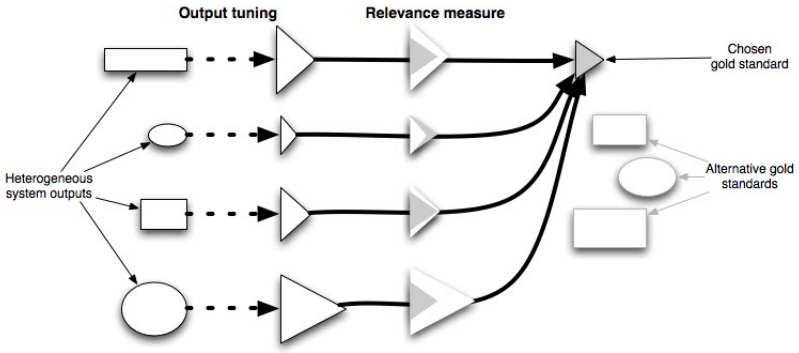

Figure 2: Output tuning. The tuning process is represented in an abstract way. The type and granularity of the gold standards are represented as various sizes and forms. The outputs of the systems are transformed so as to fit as much as possible to the chosen gold standard, before they are evaluated.

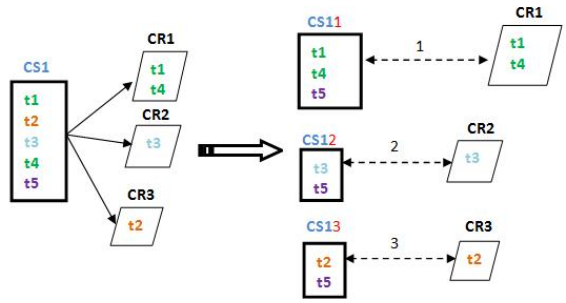

Figure 4: Splitting transformation

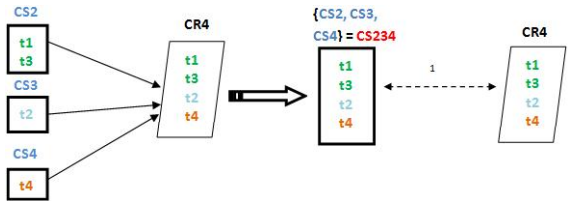

Figure 5: Merging transformation

4.3. Gradual relevance

The relevance of the output elements with respect to those of the gold standard is based on the matching and tuning steps. For each element of the tuned output ($e'_o \in T(O)$), a gradual relevance score is computed.

Term extraction For term extraction evaluation, $rel_{GS}(e'_o)$ is equal to the minimal distance of the terms of e'_o to the matching GS term:

$$rel_{GS}(e'_o) = \min_{e \in e'_o}(d_t(e, e_{gs}))$$

Semantic class acquisition For semantic class acquisition, the relevance value depends on the transformation step.
If the output class e_o is unchanged, that means that it matches only one concept. Its relevance is given by the F-measure computed on the basis of the terms that it shares with its matching concept.

$$rel_{GS}(e'_o) = fm(e_o, e_{gs}) = \frac{2*P(e_o,e_{gs})*R(e_o,e_{gs})}{P(e_o,e_{gs})+R(e_o,e_{gs})}$$

where $P(e_o, e_{gs})$ and $R(e_o, e_{gs})$ are precision and recall between output and GS elements. They are computed as follows:

$$P(e_o, e_{gs}) = \frac{\text{number of relevant terms of the output class } (e_o)}{\text{number of terms of the output class } (e_o)} \text{ and}$$
$$R(e_o, e_{gs}) = \frac{\text{number of relevant terms of the output class } (e_o)}{\text{number of terms of the GS concept } (e_{gs})}$$

If a class e'_o results from the merging of different output classes e_i, its relevance is the average of the relevance of the original ones (e_i):

$$rel_{GS}(e'_o) = \frac{\sum_{i=1}^{|X|} fm(e_i,e_{gs})}{|X|}$$

Where $e_1, e_2, ...e_{|X|}$ are the classes merged in e'_o and $|X|$ is their number.
Finally, a tuned output (e'_o) may result from a splitting operation applied on a class of the output (e_o) that matches many GS concepts (e_{gs}). In that case, we want to take into account the proximity relation of the matching GS concepts. If these concepts are close to each other, the relevance is better than if they are apart. The relevance of e'_o (the generated subclass) is computed as follows:

1. We select the GS concept (e_{gs})which has the maximal F-measure value ($fm(e_o, e_{gs})$) with the output class (e_o). This concept p is considered as central.

2. To compute $rel_{GS}(e'_o)$, the relevance values are weighted with a similarity value that expresses the proximity of various matching GS concepts with the central one (p). We use the measure of (Wu and Palmer, 1994) to compute a similarity between two concepts based on their distance:

   $$\text{Sim(p, } e_{gs}) = \frac{2*depth(C)}{depth_C(p)+depth_C(e_{gs})}$$

   Where C is the closest common ancestor of p and e_{gs}, $depth(X)$ et $depth_Y(X)$ are respectively the distance from X to the root of the ontology and the distance from X to the root by way of Y.

   $rel_{GS}(e'_o)$ is computed as follows:

$$rel_{GS}(e'_o) = fm(e'_o, e_{gs}) * Sim(p, e_{gs})$$

Where e_o is the element of the initial output from which e'_o is derived by splitting and e_{gs} is its matching concept.

5. Meta-evaluation

It is important to meta-evaluate the proposed measures before using them in real evaluation conditions, either challenges or benchmark comparisons. The meta-evaluation is needed to test the robustness of proposed measures and protocols. We have performed series of tests for term extraction evaluation, the tests exploited existing data sets and enabled to verify the adequacy of the protocol with initial specifications (Nazarenko and Zargayouna, 2009). The proposed measures give more precise evaluation of terminological extractors than traditional measures. Our first results in evaluation of semantic class acquisition are also promising.

The two following experiments show the behavior of the proposed measures in comparison with classical ones.

The first experiment is based on the following data: (i) an English corpus specialized 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[5].

	P	R	FM	AP	AR	FM
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, τ = 0.4 for terminological measures (TP, TR)

The second experiment is based on the following data: (i) a small English corpus dealing with volleyball and composed of 5,078 words, (ii) three ontologies built from this corpus by master students and respectively containing 64, 63 and 67 semantic classes and (iii) a gold standard (GS) of 64 concept that has been built manually[6].

	P	R	FM	AP	AR	FM
GS	1.0	1.0	1.0	1.0	1.0	1.0
O_1	0.4	0.4	0.4	0.83	0.47	0.60
O_2	0.46	0.45	0.45	0.84	0.47	0.60
O_3	0.34	0.36	0.34	0.81	0.37	0.51

Table 2: Results of evaluation of three ontologies

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

[6]The ontology acquisition was done by a PhD student using Terminae tool (Szulman et al., 2008). The ontology was validated by the authors.

Table 1 and table 2 present these evaluations. As expected, the adapted measures (AP and AR) follow the same curves as the classical ones (P and R), but they are higher, which proves that the proposed measures take into account the gold standard approximation. The main improvement is in precision values which is the most informative measure for acquisition task (Zhang et al., 2008). A difference of 10 points or more in F-measure (F) is also significant. We consider that these higher figures better reflect the users' feedback on the usefulness of the results. Actually in various cases, we noticed that the users consider that results help their manual acquisition task even when these results may be noisy.

6. Conclusion

The fact that knowledge acquisition tools are often assisting tools, the heterogeneity of acquisition tools, the relativity of gold standards and the complexity of terminologies and ontologies make the evaluation difficult. We propose to decompose the evaluation into independent tasks and set up a unified protocol with adapted measures that rely on gradual relevance and output tuning. This paper reports work done within the Quaero program[7]. Further work will consist in setting up internal Quaero evaluations. The first evaluation effort focused on term extraction and semantic class acquisition, the next step will consist on defining adequate protocols for the remaining KA tasks. We want to verify that the unified approach proposed based on the output tuning is more generally adequate to a wide range of KA tasks.

Acknowledgment

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

7. References

Sarra Ben Abbès, Haïfa Zargayouna, and Adeline Nazarenko. 2010. Évaluation de classes sémantiques pour la construction d'ontologies. In *Actes de la confrence Ingnierie des Connaissances (IC'2010)*, pages 12p, to be published.

Jagdev Bhogal, Andrew Macfarlane, and Peter Smith. 2007. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886.

Nicola Guarino and Christopher Welty. 2002. Evaluating ontological decisions with ontoclean. *Communication of the ACM*, 45(2):61–65.

Anna Joan Jorge Vivaldi and Mercè Lorente. 2008. Turning a term extractor into a new domain: first experiences. In N. Calzolari *et al.*, editor, *Proc. of LREC'08*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Jacob Köhler, Stehpan Philippi, Michael Specht, and Alexander Rüegg. 2006. Ontology based text indexing and querying for the semantic web. *Knowledge-Based Systems*, 19(8):744–754.

[7]http://www.quaero.org

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

Li Ma, Yang Yang, Zhaoming Qiuand Guotong Xieand Yue Pan, and Shengping Liu. 2006. Towards a complete owl ontology benchmark. In Springer Berlin / Heidelberg, editor, *The Semantic Web: Research and Applications*, volume 4011/2006, pages 125–139.

Alexander Maedche and Steffen Staab. 2002. Measuring similarity between ontologies. In *The European Conference on Knowledge Acquisition and Management (EKAW'02)*.

Alvin F. Martin, John S. Garofolo, Jonathan C. Fiscus, Audrey N. Le, David S. Pallett, Mark A. Przybocki, and Gregory A. Sanders. 2004. Nist language technology evaluation cookbook. In *Proc. of LREC'04)*.

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 and Haïfa Zargayouna. 2009. Evaluating term extraction. In *Recent Advances in Natural Language Processing (RANLP)*.

Aurélie Névéol, Kelly Zeng, and Olivier Bodenreider. 2006. Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for medline. In *Proc. of the AMIA Annual Symposium*, pages 589–593.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis, editor, *Knowledge Mining Series: Studies in Fuzziness and Soft Computing*. Springer Verlag.

Sylvie Szulman, Nathalie Aussenac Gilles, and Sylvie Despres. 2008. The Terminae Method and Platform for Ontology Engineering from Texts. In *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, page paru. IOS press, 02.

Jiao Tao, Li Ding, and Deborah L. McGuinness. 2009. Instance data evaluation for semantic web-based knowledge management systems. In *HICSS '09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Washington, DC, USA. IEEE Computer Society.

Nina Wacholder and Peng Song. 2003. Toward a task-based gold standard for evaluation of np chunks and technical terms. In *Proceedings of HTL-NAACL*, pages 189–196.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133–138.

Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.