EVALUATING SEMANTIC CLASSES USED FOR ONTOLOGY BUILDING AND LEARNING FROM TEXTS

Sarra Ben Abbès, Haïfa Zargayouna and Adeline Nazarenko
LIPN - UMR 7030, Paris 13 University & CNRS, 99, Avenue Jean-Baptiste Clément, F-93430 Villetaneuse, France

Keywords: Evaluation, Semantic classes, Concepts, Terms, Lexical level, Structural level, Ontologies.

Abstract: A large effort has been devoted to the development of ontology building tools but it is still difficult to assess their strengths and limitations. Proposed evaluations are hardly reproducible and there is a lack of well-accepted protocols and data. In this paper, we propose to decompose the evaluation of ontology acquisition process into independent functionalities. We focus on the evaluation of semantic class acquisition considered as a main step in the ontology acquisition process. We propose an approach to automatically evaluate semantic classes of ontologies that offer lexical entries for concepts. It is based on the comparative paradigm (to a gold standard). Its main focus is to compare how similar the generated semantic classes are to the gold standard concerning the disposition of concepts frontiers. This comparison relies on the lexical level and on the hierarchical structure of the "gold" concepts. The propositions are implemented, two experiments are settled on different domains and prove that the measures give a more accurate information on quality of systems' performances.

1 INTRODUCTION

Ontologies are complex artifacts (composed of concepts, hierarchical relations and roles) which are built according to different points of views and purposes. In our work, we focus on ontology acquisition from texts which is generally a semi-automatic process that needs human validation in which evaluation is crucial. Given the complexity of ontological components, we propose to decompose the evaluation of ontology acquisition from texts into three separate evaluation tasks:

- Semantic class or class acquisition: the process gives as output a list of term clusters that are considered as semantic classes (draft concepts). A semantic class is a set of terms,
- Building concept hierarchies: the process aims to design an hierarchical structure of concepts,
- Role extraction: the process consists in identifying the semantic relations that hold between concepts (excluding hierarchical relations).

In this paper, we focus on the evaluation of *semantic classes acquisition* considered as a main step in the ontology acquisition process. We propose an automatic comparative approach that relies on the existence of a gold standard which is an ontology with lexical entries for concepts.

To compare semantic classes with the gold standard, we suppose that concepts of the gold standard (C) are also associated with one or several labels. In this paper, we called *semantic classes*, the outputs of acquisition tools and *concepts*, the conceptual entities of the gold standard.

We measure how similar the generated semantic classes are to the gold standard concerning the disposition of concepts frontiers. This comparison relies on the lexical level and the hierarchical structure of the "gold" concepts.

In the following paper, we present the evaluation protocol based on gradual measures that reflect the quality of outputs and take into account the specificity of the gold standard. Section 3 details the set of experiments done so far.

2 SEMANTIC CLASSES EVALUATION PROCESS

Our evaluation protocol takes as input: (1) the *semantic classes* as they are output by semantic class extraction systems or ontology building tools, and (2) *a gold standard* which is a lexicalized ontology repre-

Ben Abbès S., Zargayouna H. and Nazarenko A..
EVALUATING SEMANTIC CLASSES USED FOR ONTOLOGY BUILDING AND LEARNING FROM TEXTS.
DOI: 10.5220/0003661804450448
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 445-448
ISBN: 978-989-8425-80-5

sented as hierarchy of concepts. The evaluation process outputs a score which is a relevance measure of the semantic classes with respect to the gold standard ontology (see figure 1). Gold standard-based evaluation has been set up in some challenges such as OAEI[1] and in approaches dealing with ontologies evaluation (Brank et al., 2006; Zavitsanos et al., 2008).

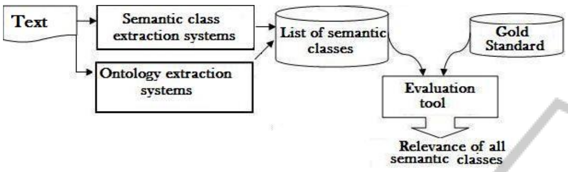
Text
Semantic class extraction systems
Ontology extraction systems
List of semantic classes
Gold Standard
Evaluation tool
Relevance of all semantic classes

Figure 1: Semantic classes evaluation process.

The process of the evaluation of semantic class acquisition is defined by three steps: anchoring, tuning and relevance computation.

2.1 Anchoring Step

The matching of a semantic class *SC* and reference concepts *C* is oriented and based on a lexical matching in which the list of semantic classes terms is compared with the concepts labels. A lexical matching is perfect when a *SC* has exactly the same terms as a *C*. However, a lexical matching is poor when a class has no common term with any concept of the gold standard. In practice, there is no exact matching between semantic classes and concepts and partial matching leads to three types of correspondence: 1 to 1, 1 to n, n to 1.

2.2 Tuning Step

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 its maximal correspondence with the gold standard: the output is tuned to the specific type and granularity of the chosen gold standard. This tuning is performed instead of considering several human judgments or revising the gold standard on the basis of the systems' outputs.
The **tuning process** takes three different matching cases into account:

- Some semantic classes and concepts stand in a 1 to 1 matching relationship. In that case, the output classes remain unchanged (no transformation).
- Some classes match several concepts of the gold standard (1 to n matching relation). In that case the output classes are split into several subclasses, each one corresponding to a different matched concept (splitting transformation, see figure 2-a).
- Several classes match the same concept (n to 1 matching relation). In that case the classes are merged into a larger one (merging transformation, see Fig. 2-b).

This tuning process is described in (Zargayouna and Nazarenko, 2010) and has been also applied to the evaluation of term extraction tools.

2.3 Relevance Computation Step

We choose to adapt well known classical measures, precision and recall, as they are generic and easy to interpret. However, these measures rely on a binary judgment of relevance. We want to take into account the gradual relevance. The overall precision and recall that we propose are computed in the basis of local relevances between tuned classes and gold concepts. These relevances are computed as follows:

- The relevance of non-transformed classes is based on the number of terms shared between the semantic class and the corresponding concept:

  $$\mathrm{P}(SC',C) = \frac{\text{number of relevant terms of class } SC'}{\text{number of terms of the class } SC'}$$
  $$\mathrm{R}(SC',C) = \frac{\text{number of relevant terms of class } SC'}{\text{number of terms of the concept of gold standard } C}$$
  $$relevance_{nt}(\mathrm{SC'}, \mathrm{C}) = \text{F-measure}(SC',C) = \frac{2 * \mathrm{P}(SC',C) * \mathrm{R}(SC',C)}{\mathrm{P}(SC',C) + \mathrm{R}(SC',C)}$$

  where SC' is an output class, C is the matching concept of the gold standard, P(SC',C) is the precision of SC' wrt. C and R(SC',C) is the recall of SC' wrt. C.
- **Merging Case (n to 1 Relation):** when many semantic classes are matched with only one concept of the gold standard, we propose to merge these classes into one semantic class SC. The relevance of SC is the average of F-measure of the different classes SC_i from which it is formed. It is computed as follows:

  $$relevance_{mt}(SC_i,C_s) = \frac{\sum_{i=1}^{|X|} F-measure(SC_i,C_s)}{|X|}$$

  where SC_i are semantic classes of the system, C_s is the concept of the gold standard which matches different SC_i and $|X|$ the number of anchored (or matched) classes of the output to C_s.
- **Splitting Case (1 to n Relation):** The relevance of the split classes depends on the relative position in the gold standard hierarchy of the matching concepts. If they are close to each other, the splitting process is a smaller transformation than

[1] Ontology Alignment Evaluation Initiative.

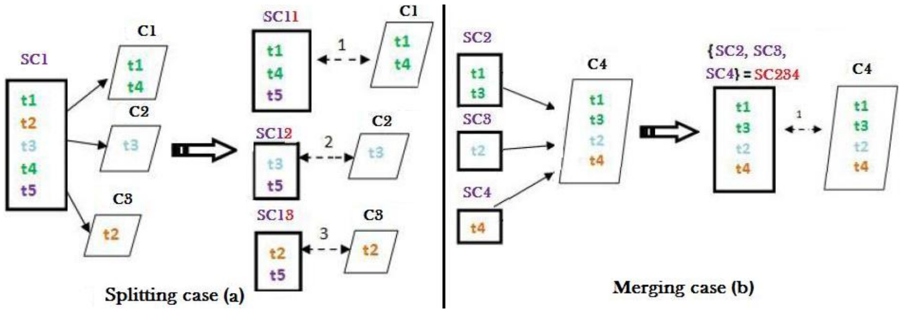

Figure 2: Splitting and Merging cases.

if they are far. The relevance of each split class depends on the relevance (F-measure) of the initial SC class from which it is derived and on the similarity of its matching concept with the concept that is considered as the pivot p of the splitting process:

$$relevance_{st}(SC_i',C) = F - measure(SC,C) \times Sim_{WP}(p,C)$$

where SC_i' is a class derived by splitting an initial SC class, C is a concept of the gold standard, p is the pivot concept of the gold standard which has the highest F-measure with the semantic classes derived from SC. Sim_{WP}(p,C) is a similarity measure between concepts p and C (Wu and Palmer, 1994).

These overall relevance measures are based on the lexical relevance of each class of the tuned output. The precision and recall of a system are computed as follows:

$$\mathrm{P} = \frac{\sum_{i=1}^{|S'|}\sum_{j=1}^{|GS|} relevance(SC_i',C_j)}{|S'|}\;;$$
$$\mathrm{R} = \frac{\sum_{i=1}^{|S'|}\sum_{j=1}^{|GS|} relevance(SC_i',C_j)}{|GS|}$$

where $|S'|$ and $|GS|$ are respectively the numbers of classes of the tuned output and concepts of the gold standard.

$$relevance(SC_i',C_j) = \begin{cases} relevance_{nt}(SC_i',C_j) & \text{if } SC_i' \text{ is not transformed} \\ relevance_{mt}(SC_i',C_j) & \text{if } SC_i' \text{ is obtained by a merging operation} \\ relevance_{st}(SC_i',C_j) & \text{if } SC_i' \text{ is obtained by a splitting operation} \end{cases}$$

(Maedche and Staab, 2002) proposed similarity measures between two ontologies that can be used for comparing an ontology to a gold standard. These measures take into account two levels: lexical (similarity measure *String Matching (SM)*) based on the edit distance and conceptual (*semantic cotopy measure (SC)*). However, the two ontologies are considered equal on quality. Our measures are adapted to the special context of evaluation: they are asymmetric i.e the similarity is oriented from the outputs to the gold standard. This enable to take into account the specific type and granularity of the chosen gold standard by tuning the systems' outputs.

3 EXPERIMENTS

Two experiments are done with two different domains: transport and astronomy. The goal of the first experiment is to check the behavior of the proposed measures. A small gold standard ontology had been built manually from the synsets of WordNet, in the transport domain. We artificially created 4 systems' outputs by removing terms or classes (silence) from the gold standard or by adding terms or classes (noise) to it. These outputs are based on the following data: (1) O_{nT} contains noisy terms, (2) O_{nC} contains additional noisy classes containing irrelevant labels, (3) O_{sT} has missing terms, and (4) O_{sC} has some missing classes. To test the limits of our measures, we considered additional artificial outputs that require a merging operation (O_m), a splitting operation (O_{sp}) and a combination of both splitting and merging cases (O_{spm}). The evaluation gives an information on the capacity of the evaluated systems to

Table 1: Evaluation of the tools outputs compared to the gold standard.

Outputs	Precision	Recall	F-measure	Ranking
GS	1	1	1	1
O_{sT}	0,74	0,74	0,74	5
O_{sC}	1	0,67	0,8	4
O_{nT}	0,84	0,84	0,84	3
O_{nC}	0,86	1	0,92	2

extract relevant domain terminologies and to regroup them into relevant classes. The second experiment have been carried out to show how far the proposed measures take into account the gold standard approximation and make an accurate evaluation comparing to the classical ones. An astronomy ontology (gold standard) has been built manually from texts of the astronomy domain using Terminae tool (Szulman et al., 2008). Associated to that ontology, two outputs have been provided by Formal Concept Analysis (FCA): a first output $O1$ has been settled (75 classes) by using a list of 24 initial terms/ basic concepts (V1). A second output $O2$ (111 classes) has been settled by using a set of terms (V2) that extends V1. There is no exact matching between the handcrafted clusters (O1 and O2) and the gold standard. Classical measures results are null (for O1 and O2, CP = CR = CFM = 0). This experiment proves that the proposed evaluation measures give a more accurate information on quality of systems' performances.

Table 2: Evaluation of the outputs compared to the gold standard *GS*.

Outputs	Precision	Recall	F-measure	Ranking
O_{spm}	0,54	0,54	0,54	1
O_m	0,2	0,2	0,2	3
O_{sp}	0,3	0,3	0,3	2

Table 3: Evaluation of the outputs obtained by *FCA* method: results.

Outputs	P	R	FM
O1	0,5	0,17	0,25
O2	0,37	0,2	0,26

4 CONCLUSIONS AND FUTURE WORK

We have decomposed the problem of ontology acquisition evaluation into different sub-problems. In this paper, we have focused on the evaluation of semantic classes acquisition. We proposed a protocol for comparative evaluation allowing the matching between semantic classes and the gold standard. Thus, the quality assessment on gold standards is controversial: there has to be a gold standard, quality has to be assumed, etc. From the same textual corpus, there is a multitude of acceptable solutions that vary from one expert to another. In order to take into account the variability of the gold standard we proposed to tune the systems' outputs. This enable to find the maximal correspondence with the gold standard. We also proposed to compute a gradual relevance, the aim is to detect differences which are due to errors from those that can be due to different conceptualisation choices.

Experiments have showed that the proposed evaluation measures gave higher values than traditional ones and a more accurate information on the quality of the performances of acquisition systems.

We have focused on the quality of the classification neglecting, at the lexical level, the correspondence between terms that may itself be only partial (a term in the semantic class can be a variant of another concept of the gold standard). We use a simple matching technique in order not to distort the evaluation. As future work, we aim at including in our measures, a terminological distance between labels as proposed in (Zargayouna and Nazarenko, 2010).

ACKNOWLEDGEMENTS

The authors would like to thank Thibault Mondary (LIPN) and Yue Ma (LIPN) for fruitful discussions. This work was partly realized as part of the Quaero Program, funded by OSEO, French State agency for innovation.

REFERENCES

Brank, J., Madenic, D., and Groblenik, M. (2006). Gold standard based ontology evaluation using instance assignment. In *Proc. of the 4th Workshop on Evaluating Ontologies for the Web (EON2006), Scotland.*

Maedche, A. and Staab, S. (2002). Measuring Similarity between Ontologies. In *Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Spain.*

Szulman, S., Aussenac-Gilles, N., and Despres, S. (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*. IOS press.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico.

Zargayouna, H. and Nazarenko, A. (2010). Evaluation of textual knowledge acquisition tools: a challenging task. In *LREC 2010*, Malta.

Zavitsanos, E., Paliouras, G., and Vouros, G. (2008). A distributional approach to evaluating ontology learning methods using a gold standard. In *Ontology Learning and Population Workshop (OLP 2008), European Conference on Artificial Intelligence (ECAI 2008).*