Evolving and Validating Annotations in Web-based Collaborative Environments through Ontology Matching*

Angela Locoro
angela.locoro@unige.it

Viviana Mascardi
viviana.mascardi@unige.it

Anna Marina Scapolla
scapolla@unige.it

Università degli Studi di Genova
Genova, Italy

ABSTRACT

The research presented in this paper describes an automated approach for extracting concepts from annotated shared contents within a collaborative web environment, and matching them with domain ontologies. Feedback on the domain ontology suitability for the environment purposes is provided as a result of the automatic matching between the domain ontology and the free tags that users of the system employ. This approach will enable annotations and domain ontologies to evolve coherently with the real use of any social web environment. Experiments carried out on the Knowledge Practice Environment of the EU-funded project KP-Lab demonstrate the feasibility of our approach.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks; I.2.7 [**Natural Language Processing**]: Language parsing and understanding

General Terms

Algorithms

Keywords

Automated and Adaptive Annotation, Domain Ontologies Evolution, Collaborative Web Tools, Trialogical Learning.

1. INTRODUCTION

In the last decades, novel ways of empowering collaborative processes for creating or developing shared knowledge practices have been studied and developed [6]. One of the main results of the EU KP-Lab (Knowledge Practice Laboratory) project, `http://www.kp-lab.org`, is the Knowledge Practice Environment (KPE from now on), a collaborative web environment providing a large set of tools for reflecting on, making visible, and supporting the knowledge acquisition process. In KPE, each community of users can access a virtual workspace named "Shared Space" (SSP from now on) where Knowledge objects can be shared and evolved collectively. Users can annotate them either with natural language free tags (forming free tags vocabularies), or with concepts belonging to vocabularies developed by experts (forming a-priori domain vocabularies). Natural language processing and ontology matching techniques are the tools we exploit in our research for carrying out a systematic analysis of the intended usage of a collaborative web environment with respect to its actual usage, and for automatically adapting pre-defined domain vocabularies and ontologies to the actual needs of the users.

2. A METHODOLOGY FOR TESTING DOMAIN VS FREE TEXT VOCABULARIES

An overview of our procedure is graphically depicted in Figure 1. The process is divided into the four phases that have been described in [3], and that we recall here briefly:

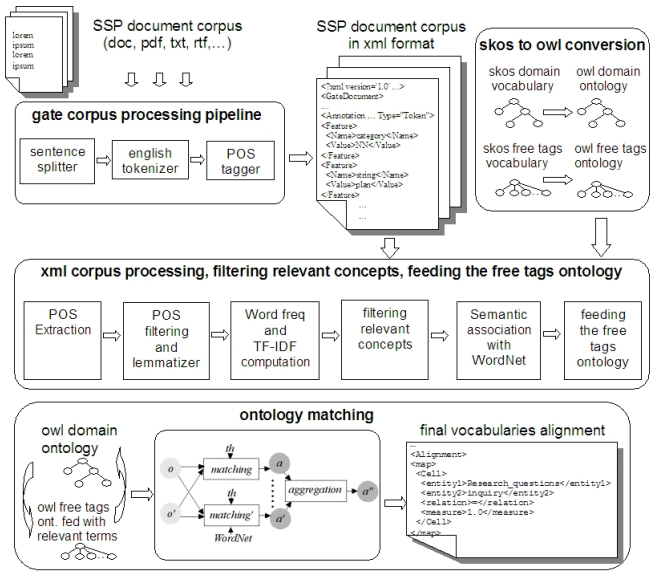

Figure 1: Our procedure.

1. **SKOS to OWL Vocabularies Conversion**, for compatibility with state-of-the-art ontology matching tools[1].

[1] `http://oaei.ontologymatching.org/2007/skos2owl.`

SAC'12 March 25-29, 2012, Riva del Garda, Italy.

2. **Corpus Processing**, carried out by first exploiting Gate (General Architecture for Text Engineering)[2] for normalizing words in documents, and then computing the TF-IDF (Term Frequency - Inverse Document Frequency) measure [4] for estimating the relevance of words in pieces of text.
3. **Concepts Discovery and Free Tags Vocabulary Feeding**, obtained by: filtering POS nouns categories and applying TF-IDF measure and a threshold to it; associating all those terms having a hyponym or hypernym relation with every other term through WordNet and creating an association list with such terms; creating an `owl:Class` with class name equal to the word and `rdfs:label` with the same name; creating an `owl:equivalentClass` for each synonym concept or a `rdfs:subClassOf` for each hypernym concept, taken from the above association list.
4. **Ontology Matching**: this phase takes the OWL domain ontology and the free tag ontology just fed with new terms from the shared space corpus, and runs five automatic ontology matching methods in parallel: substring, n-gram, SMOA, and WordNet by using the Alignment API[3], and our new method ExtendedWN (see Section 3.1). For each method it is possible to set a parametric threshold in $[0, 1]$ used for discarding correspondences with a lower confidence. We set the threshold to 0.5 for all of them. To obtain a final alignment we took the "partial" alignments and automatically aggregated them. After this first aggregation process, we applied our *SemAnchor* (see Section 3.2) method to such results, for each free tag ontology concept for which a mapping with a domain ontology concept was found. We then obtained a final alignment by applying again the aggregation between the *SemAnchor* results and the first aggregation set.

Comparative study: Text-2-Onto

In order to compare our approach to the one used in state-of-the art Ontology Learning tools we have used the Text-2-Onto tool as a plug-in of the Neon Toolkit[4]. We have created free tags vocabularies from the terms extracted with this tool to be compared to those created through our approach.

3. ONTOLOGY MATCHING BASED ON SEMANTIC RELATIONS

Based on [2] and the Alignment API, we designed and implemented two new matching methods.

3.1 The ExtendedWN approach

This method extends the *JWNLAlignment* one of the Alignment API, and matches two ontologies via WordNet, finding correspondences between entities e and e' if they satisfy one of these conditions:

– e and e' are synonyms
– e is a hypernym of e' or vice versa
– e is a hyponym of e' or vice versa

Our method changes the score measure (which is no more based on the longest common substring but is equality - two concepts must be identical) and the search space between e and e' that becomes that of their WordNet synonyms and, for each of them, their hypernyms and hyponyms.

`html`
[2]`http://gate.ac.uk/`
[3]`http://alignapi.gforge.inria.fr/`
[4]Version 1.2.3 available at `http://neon-toolkit.org/wiki/Download/1.2.3`.

SSP	Tot T	Tot C		$RC_{tf\text{-}idf}$	RC_{t2o}
		tf-idf	t2o	tf-idf	t2o
SSP1	32.660	1.634	1.569	606	540
SSP2	131.992	4.920	3.625	1.374	1.592
SSP3	6.930	685	544	244	384

Table 1: Results from SSPs corpus analysis.

3.2 The SemAnchor approach

This approach seeds from a set of mappings obtained by running different matching methods, treats them as "semantic anchors", and starting from them, extends the set of correspondences between concepts with all the synonym, hyponym and hypernym terms belonging to the ontology obtained from a document corpus.

Given an alignment that contains all the correspondences from an ontology o and o', for each concept c in ontology o, being c a concept from a free tag vocabulary:

– take the subset of all mappings for c inside the alignment with c' belonging to ontology o' and hence to a domain vocabulary;

– map each c' with each synonym, hypernym and hyponym concept of c that results from the ontology o through an association between c and terms belonging to o that are related to c through the WordNet relations just mentioned;

– aggregate the results with the seeds mapping in order to obtain a final alignment.

4. EXPERIMENTS AND RESULTS

Our experiments were run on the three most complete and representative SSPs provided by KPE, that we already used for the experiments described in [3]. They are SSP1, titled "The Bachelor Thesis SSP", SSP2, titled "The Learning Interaction SSP", and SSP3, titled "The Multimedia Project SSP".

SSP1 contains 10 documents and a domain ontology (DO), called `Bachelor.owl`, with 14 concepts. SSP2 contains 15 documents and a domain ontology (DO), called `PBL.owl`, with 47 concepts. SSP3 contains 6 documents and has the same domain ontology as SSP2. The free tag vocabularies of all the SSPs were initially empty.

The concept extraction procedure from each SSP corpus has resulted in the creation of a list of free tags whose ranking, thresholding, and representation in OWL led to one Free Tag Ontology (FTO) based on TF-IDF term weighting and threshold filtering ($FTO_{TF\text{-}IDF}$) and one FTO based on Text-2-Onto (FTO_{T2O}) tool according to the procedure briefly recalled in Section 2.

Table 1 shows the total number of tokens found in the corpus (Tot T), the total number of concepts extracted from the corpus after POS category filtering and lemmatisation (Tot C), the total number of Relevant Concepts (RC) after TF-IDF term weighting and threshold filtering ($RC_{TF\text{-}IDF}$), and the total number of Relevant Concepts after Text-2-Onto term extraction and threshold filtering (RC_{T2O}).

As it turns out that the free tags vocabularies were still empty at this stage, $RC_{TF\text{-}IDF}$ and RC_{T2O} also represent the total number of concepts in $FTO_{TF\text{-}IDF}$ and FTO_{T2O} respectively for each SSP.

In order to evaluate our approach, manual reference alignments were created and augmented by running our *SemAnchor* method on every correct mapping previously found by hand. The evaluation has then been conducted by two kinds of analyses:

Reference Alignment								
SSP	**Do**		**FTO$_{\text{tfidf}}$**		**Do**		**FTO$_{\text{t2o}}$**	
	♯	**Cov**	♯	**Cov**	♯	**Cov**	♯	**Cov**
SSP1	14	100%	333	55%	14	100%	156	29%
SSP2	44	94%	570	41%	33	70%	132	8%
SSP3	41	87%	150	61%	44	94%	132	34%

Table 2: Concepts covered by the reference alignments with respect to DO and FTO$_{\text{TF-IDF}}$ ontologies, and to DO and FTO$_{\text{T2O}}$ ontologies.

Analysis 1: coverage of DO and FTOs with respect to the reference alignment, for each SSP. In this analysis we counted the number of elements in the DO and FTOs that appear in at least one correspondence in the reference alignment. The results of this analysis are depicted in Table 2, where ♯ stands for the number of ontology concepts covered by the reference alignment, and **Cov** is the percentage of concepts covered by the reference alignment with respect to the total number of concepts in each ontology.

Table 2 shows that the concepts covered by reference alignments with respect to the DOs present some exceptions. In both SSP2 and SSP3, from 6 to 30% of the DOs concepts have no correspondence, in the reference alignment, with any of the FTOs concepts; SSP2 results show that only 8% of the FTO$_{\text{T2O}}$ free terms are covered by the reference alignment and a half of the coverage for FTO$_{\text{TF-IDF}}$ can be observed; for SSP3 61% on average coverage for FTO$_{\text{TF-IDF}}$ and 34% for FTO$_{\text{T2O}}$ represent the best coverage for a FTO$_{\text{T2O}}$ vocabulary by means of a reference alignment. A suggestion emerging from this analysis might be to do a first revision of the domain vocabulary in order to include relevant terms from the SSPs and to completely reflect their real contents and usage. The support given by our approach, that is the suggestion of new concepts, mined from these activities, can be used both for the domain modelling phase and in the future, when the practices will be more mature, to co-evolve the system knowledge with the community's knowledge.

Analysis 2: evaluation based on precision, recall, F-measure of ontology matching methods [1] used to align DO with FTO$_{\text{TF-IDF}}$ and FTO$_{\text{T2O}}$ respectively. This last analysis may be seen as a meta-evaluation of how reliable our approach was in evaluating a-priori vs evolving knowledge. In this analysis we compare each partial and final alignment obtained with the ontology matching procedure with the corresponding reference alignments. A synthesis of the experiments is reported in Table 3 showing the best results of the experiments, those where each SSP DO was matched with the corresponding FTO$_{\text{TF-IDF}}$ (from 1 to 3), and those where each SSP DO was matched with the corresponding FTO$_{\text{T2O}}$ (from 4 to 6).

The precision we obtained in our tests is higher than the recall, which is a typical result of any automatic ontology matching system as we already discussed in [5]. SSP1 and SSP2 are the domains where we obtained the best results. The reason for the poor results obtained in SSP3 is that the Multimedia Project domain contains much more technical and specialised terms than the Bachelor Thesis and the Learning Process ones. For this reason we think that most terms characterizing a multimedia project have not been, de facto, included in the domain vocabulary provided for SSP3. Still, tests for SSP2 reveal the lowest F-measure (20% in both the experiments). The FTO in SSP2 is the one with more concepts at all, quite a double of FTO$_{\text{TF-IDF}}$(SSP1), three times more than FTO$_{\text{T2O}}$(SSP1) and more than five times FTOs(SSP3). Large ontologies contain noise that impacts on the performance of automatic matching algorithm. This is particularly true for FTOs that are created in a fully automatic way and thus are inherently noisier than ontologies built by ontology engineers.

Test	**Best** P	**Best** R	**Best** F
1	0.75 WN-2	0.30 Final, SemAnch	**0.32** Final, SemAnch
2	**0.92** WN-2	**0.42** Final, SemAnch	0.20 Final, SemAnch
3	0.71 WN-2	0.31 Final	0.26 Final
4	**0.81** WN-2	**0.46** Final	**0.50** SemAnch
5	0.63 WN-2	0.43 Final	0.20 WN-2
6	0.69 N-gram	0.39 Final	0.30 SMOA

Table 3: Matching Domain Ontologies with FTO$_{\text{TF-IDF}}$ and FTO$_{\text{T2O}}$ in each SSP: best results.

The methods that give the best precision are the more sophisticated ones: in particular, the WordNet-based method that we developed, ExtendedWN, gave promising results.

Always obtaining the best recall by aggregating the results of all the matching methods adopted (aggregation that we name "Final alignment" and the intermediate one, called SemAnchor) is not surprising: the partial alignments contribute to discover correct correspondences that allow for the creation of a Final alignment whose correct correspondences are the union of all of them.

Based on the matching results obtained, we observe that our approach gave the best precision and, even if lower, a still comparable recall, whereas the Text-2-Onto approach gave the best F-measure. Based on that we may conclude that our concepts extraction approach is comparable to one of the most stable and mature state-of-the-art tools for ontology learning, and is worth working on.

5. REFERENCES

[1] H. H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *NODe 2002*, LNCS, pages 221–237, 2002.

[2] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, 2007.

[3] A. Locoro, V. Mascardi, and A-M. Scapolla. NLP and Ontology Matching - A Successful Combination for Trialogical Learning. In *ICAART 2010*, pages 253–258, 2010.

[4] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[5] V. Mascardi, A. Locoro, and P. Rosso. Automatic ontology matching via upper ontologies: A systematic evaluation. *IEEE Trans. Knowl. Data Eng.*, 22:609–623, 2010.

[6] S. Paavola and K. Hakkarainen. From meaning making to joint construction of knowledge practices and artefacts - a trialogical approach to CSCL. In *CSCL2009*, pages 83–92, 2009.