NLP AND ONTOLOGY MATCHING: A SUCCESSFUL COMBINATION FOR TRIALOGICAL LEARNING

Angela Locoro
DIBE, Biophysical and Electronic Engineering Department, University of Genoa, Via Opera Pia 11/A, Genova, Italy
angela.locoro@unige.it

Viviana Mascardi
DISI, Computer Science Department, University of Genoa, Via Dodecaneso 35, Genova, Italy
viviana.mascardi@unige.it

Anna Marina Scapolla
DIBE, Biophysical and Electronic Engineering Department, University of Genoa, Via Opera Pia 11/A, Genova, Italy
scapolla@unige.it

Keywords: Trialogical Learning, Natural Language Processing, Ontology Matching.

Abstract: Trialogical Learning refers to those forms of learning where learners are collaboratively developing, transforming, or creating shared objects of activity in a systematic fashion. In order to be really productive, systems supporting Trialogical Learning must rely on intelligent services to let knowledge co-evolve with social practices, in an automatic or semi-automatic way, according to the users' emerging needs and practical innovations. These requirements raise problems related to knowledge evolution, content retrieval and classification, dynamic suggestion of relationships among knowledge objects. In this paper, we propose to exploit Natural Language Processing and Ontology Matching techniques for facing the problems above. The Knowledge Practice Environment of the KP-Lab project has been used as a test bed for demonstrating the feasibility of our approach.

1 INTRODUCTION

In a Trialogical Learning (Paavola and Hakkarainen, 2005) environment, the collaborative knowledge creation process is characterized by the interaction through developing common, concrete objects (or artifacts) of activity, not just between people ("dialogical approach"), or within one's mind ("monological" approach)"[1].

The EU KP-Lab project (see next section) is developing software tools for the management of a shared environment able to represent the information flow across educational or professional communities, during their trialogical knowledge creation practices. Implementing such tools raises some main challenging problems, such as:

1. **Automatic and dynamic content classification**. One purpose of collaborative environments is to store and share the users' contents. Properly classifying contents is a key factor for efficiently retrieving them. Often, this process relies on pre-defined, static vocabularies that describe the environment's domain. However, since knowledge evolves, *the system must integrate content classification techniques that are both dynamic and automatic, because delegating the updates to human experts would be too expensive*.
2. **Understanding knowledge evolution**. As for the very nature of Trialogical Learning, knowledge evolves as users interact with the environment. *The initial system vocabularies (or taxonomies, or ontologies, depending on the system) should evolve in a (semi-)automatic way, in order to correctly and timely reflect the users' understanding and usage of the environment itself.*
3. **Automatic and dynamic suggestion of tags and relationships among knowledge objects**. Suppose a user wants to find the concepts related to a knowledge object, for example in order to tag or classify it, but he/she has no or little idea of where starting from. *The system should provide dynamically generated suggestions based on the knowledge and data currently stored in the system.*

Learning environments in commmon use today provide only limited support for knowledge creation processes and do not face the problems above.

While many core technologies of the Semantic Web infrastructure are already available, there is vast

[1] http://kplab.evtek.fi:8080/wiki/Wiki.jsp?page=TrialogicalLearning.

amount of work ahead in tuning them for the use of ordinary learners, instructors and professionals, especially with regard to stability, performance and usability.

The vision depicted in (Gruber, 2008) is a step forward the potential of combining the Web 2.0 perspective with the Semantic Web one. Consistently with this vision, we investigate how to combine Natural Language Processing and Ontology Matching techniques as we think that this kind of combination would be of great benefit for supporting Trialogical Learning. We tested the feasibility of our approach in the Knowledge Practice Environment of KP-Lab. Although at a prototypical stage, our approach seems promising for facing the challenges outlined before.

The paper is organized as follows: Section 2 gives an overview of the KP-Lab Project, Section 3 describes our approach whereas Section 4 presents the experiments conducted and the results obtained. Section 5 concludes by discussing related and future work.

2 THE KP-LAB PROJECT

The KP-Lab project[2] is an Integrated Project sponsored by the 6th EU Framework Programme in the Information Society Technologies, Technology-Enhanced Learning program. It aims at creating a learning system supporting trialogical learning in education and workplaces. The project promotes co-evolution of individual and organizational learning with technology through the development of a learning system based on technological, theoretical, pedagogical, and social innovations.

The main features of the learning system of KP-Lab, named the Knowledge Practice Environment (KPE), can be summarized as follows:
– shared working spaces (the domain specific work environments of the system, called "shared spaces", from now on abbreviated as SSPs) to organize activities around shared objects and to interact at personal and community levels;
– support to organize the community and to structure the learning process;
– support to reflective activities on the shared objects and the learning context, e.g. through resources annotation tasks with tags/concepts from the vocabularies of the SSP;
– awareness services to trace the knowledge evolution process that is embedded in the practices of the members of a community.

[2] http://kp-lab.org/.

The KPE system relies on the ontological representation of the SSP and its "knowledge artifacts", as well as the users' actions. The system model is defined in the TLO (Trialogical Learning Ontology) (Tzitzikas et al., 2007) described in OWL (Web Ontology Web Language[3]) and the system data format is that of RDF (Resource Description Framework[4]).

The KPE system allows the SSP users to tag the shared objects using structured SKOS (Simple Knowledge Organization System)[5] vocabularies. A free tags vocabulary is also present in each SSP and it contains free text terms that the users can create during the ongoing phase of knowledge elicitation.

Thus, in each SSP two types of vocabularies exist: the **domain vocabularies**, initially created by domain experts, and the **free tag vocabulary** that includes all the terms freely inserted by users. They can tag knowledge artifacts with terms selected from the domain vocabularies, or they can tag them with free text that enriches the free tag vocabulary. These vocabularies and the corpus of the SSP documents represent the basis over which we can combine natural language processing and ontology matching techniques to provide users with tools to classify contents, to reflect on their domain comprehension and to receive suggestions on tags and relationships between knowledge objects.

3 COMBINING NATURAL LANGUAGE PROCESSING AND ONTOLOGY MATCHING

To test the proposed approach we used real data from the SSPs created by different KP-Lab users communities. Each SSP deals with a specific domain and contains a set of documents (the SSP corpus), domain vocabularies and a free tag vocabulary. For each SSP we extract vocabulary concepts from the document corpus using Natural Language Processing techniques, we feed the free tag vocabulary with them and match it with the domain vocabulary, using ontology matching techniques. The procedure is graphically depicted in Figure 1. The process is divided into four phases.

Phase 1: SKOS to OWL Vocabularies Conversion. In the spirit of reusing ontology matching tools and methodologies, most of which operate on OWL ontologies, we defined a set of rules for translating

[3] http://www.w3.org/TR/owl-features/.
[4] http://www.w3.org/RDF/.
[5] http://www.w3.org/2004/02/skos/.

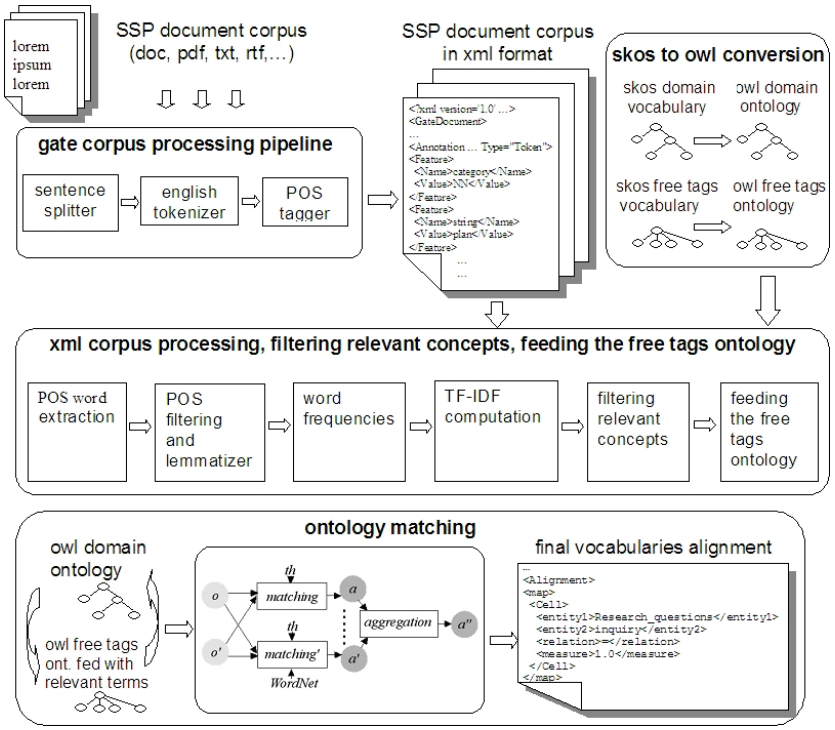

Figure 1: The procedure phases for concepts discovery and matching of domain and free tags ontologies.

SKOS vocabularies into OWL ontologies:

- *skos:Concept* becomes *owl:Class*
- *skos:prefLabel* becomes *rdfs:label*
- *skos:altLabel* becomes a new *owl:Class*, having as *owl:equivalentClass* the concept named with skos:preflabel (and viceversa)
- *skos:broader* becomes *rdfs:subClassOf*

Using the SKOS to OWL conversion rules we translated the domain vocabularies and the free tags vocabulary of each SSP used for our experiments (see Section 4). In the sequel we will use **domain ontology** to indicate the OWL translation of a domain vocabularies, and **free tag ontology** for the OWL translation of the free tags vocabulary.

Phase 2: Gate Corpus Processing Pipeline.
In order to process a SSP corpus we carried out the following steps:
1. extract all the documents from the SSP and save them into a local directory;
2. use the ANNIE component[6] of the Gate[7] tool (Cunningham et al., 2002) and for each SSP corpus load the documents into Gate, set and run the pipeline procedure with the Sentence Splitter, the English Tokenizer and the Part Of Speech (POS) Tagger;
3. save the XML version of the corpus obtained from steps 1 and 2 with all the annotations tags for the next elaboration phase.

Phase 3: Concepts Discovery and Free Tags Ontology Feeding.
We implemented a Java application for term discovery and ontology feeding, consisting of three modules.

Module 1: XML Corpus Processing for Concept Discovery. This module aims at processing the corpus, filtering POS and counting the word occurrences of each corpus document. It consists of three submodules:

- *Sub-module 1.1, XMLGateDocument* takes as input the XML files representing the SSP corpus with annotations and for each of them outputs a plain .txt file with POS and word information;
- *Sub-module 1.2, POSFilteringAndLemmatizer* takes as input the output files of the XMLGateDocument module, filters the POS according to the following POS categories referring to nouns: NN, NNP, NNPS, NNS, NP, NPS. Then it lemmatizes each word (meaning that for each word the

[6] http://gate.ac.uk/ie/annie.html.
[7] General Architecture for Text Engineering

module returns its canonical form - e.g. dogs becomes dog and so on) using WordNet 3.0 (Miller, 1995);

- *Sub-module 1.3, WordFrequencies* takes the files produced by the previous module and transforms them into lists of word lemmas and frequency counts (occurrences of the word in each document).

Module 2: Filtering Relevant Concepts with TF-IDF Measure. To retrieve key terms (relevant concepts) we use a standard weighting measure in the Information Extraction field, the TF-IDF (Term Frequency - Inverse Document Frequency (Spärck Jones, 1972)). The measure is an indicator of how relevant a term is for a document; too common terms or not relevant terms tend to be filtered out. This gives a chance to set a threshold under which only the document key terms with higher TF-IDF are selected. In the final procedure for extracting salient words we compute the final list of terms filtering out those with TF-IDF > 1.0.

Module 3: Feeding the Free Tags Ontology. The final list of terms, the output of Module 2, is used to feed the free tag ontology. The *FeedOntology* module, which integrates the OWL API[8] creates all the new concepts in the ontology. For each concept we create an owl:Class with class name equal to the concept name and an rdfs:label (for easy conversion to skos:PrefLabel) with the same name.

Phase 4: Ontology Matching.
An ontology matching process takes two ontologies o and o' and a set of resources r, and returns an alignment a (namely, a set of correspondences) between o and o'. A correspondence is of the form $< id, e, e', R, conf >$ where id is a (optional) unique identifier, e and e' are the entities (e.g. properties, classes, individuals) of o and o' respectively, R is a relation such as "equivalence", holding between the entities e and e', $conf$ is a confidence measure (typically in the $[0,1]$ range) holding for the correspondence between the entities e and e'. In our approach we consider only concepts as entities and equivalence as relation.

As depicted in Figure 1, the ontology matching phase takes each OWL domain ontology and the OWL free tags ontology just fed with new terms from the SSP corpus, and runs in parallel four automatic different ontology matching methods: substring, n-gram, SMOA, and one method based on WordNet.

To match ontologies we use the Alignment API[9] that implements all the above methods. For each method we set a parametric threshold in $[0,1]$ to 0.5 in order to discard correspondences with a confidence lower than it. In addition, we developed an aggregation function for aggregating the four alignments found by running the four methods. In case the same correspondence was found in more than one alignment, we keep the one with highest confidence measure.

4 EXPERIMENTS AND RESULTS

The KPE is still at a prototype level. During the past years field trials were conducted by some partners. Existing SSPs span different domains of activity. We selected three of them, the most complete and representative ones, to test our application. They are SSP1, titled "The Bachelor Thesis SSP", SSP2, titled "The Learning Interaction SSP", SSP3, titled "The Multimedia Project SSP".

SSP1 contains 10 documents (4 .txt, 2 .pdf and 4 .doc) and a domain ontology, called Bachelor.owl, with 14 concepts. SSP2 contains 15 documents (9 .pdf and 6 .doc) and a domain ontology, called PBL.owl, with 47 concepts. SSP3 contains 6 .doc documents and the same domain ontology as SSP2. All the three free tag vocabularies are empty.

The results of the corpus analysis phase are depicted in Table 1 where **tot T** stands for the total number of tokens found in the corpus, **tot C** stands for the total number of concepts after POS category filtering and lemmatization, and **tot RC** stands for the total number of Relevant Concepts after TF-IDF term weighting and threshold filtering. As the free tags vocabularies have no concepts at the beginning of the tests, tot RC represents the number of concepts of the three final free tags ontologies after they have been fed with relevant concepts from the corpus.

Corpus	Tot T	Tot C	Tot RC
SSP 1	32.660	1.634	606
SSP 2	131.992	4.920	1.374
SSP 3	6.930	685	244

Table 1: Results from SSPs corpus analysis.

The ontology matching phase between the domain ontologies and the free tag ontologies for each SSP resulted into three final alignments: the first (for SSP1) includes 128 correspondences, the second

[8] http://owlapi.sourceforge.net/.

[9] http://alignapi.gforge.inria.fr/.

(SSP2) amounts to 754 correspondences and the third (SSP3) has 203 correspondences.

A preliminary quantitative evaluation analysis is reported in Tables 2 and 3.

Table 2 reports the percentage of the concepts in each domain ontology having a correspondence with a concept in the free tag ontology (**% CDO** column) and the average number of correspondences found for each concept in the domain ontology (**Avg Corr.** column).

Ontology	SSP	% CDO	Avg Corr.
Bachelor	SSP1	100%	9
PBL	SSP2	100%	16
PBL	SSP3	98%	4

Table 2: Coverage and average correspondences per concept in the domain ontology.

Ontology	SSP	CFT $\in$ Corr.	% CFT
Bachelor	SSP1	119	20%
PBL	SSP2	623	45%
PBL	SSP3	140	57%

Table 3: Free Tags coverage.

Table 2 gives an indication on how well the domain ontology (and hence, the initial domain vocabulary) corresponds to the real content of the SSPs documents, from which the list of free tags (and hence, the free tags ontology) have been extracted. For example, in SSP2 all the concepts in PBL.owl have a correspondence with at least one concept (16 on average) extracted by the corpus of the documents, and in SSP3 98% of the concepts in the same ontology have been matched with at least one concept (4 on average) in the free tag ontology extracted from the documents. Bachelor.owl has 100% of its domain concepts covered by a corresponding concept (9 on average) in the free tags ontology extracted from the documents in SSP1.

The second analysis we carried out (Table 3) complements the previous one by showing how many concepts from the free tags ontology belong to at least one correspondence (**CFT $\in$ Corr.** column). By measuring the ratio between concepts in the free tag ontology that have a correspondence with at least one concept in the domain ontology and the total number of concepts in the free tag ontology (**% CFT** column) we obtain an indication of how many relevant terms appearing in the documents of a SSP also belong to the domain vocabulary.

Low values mean that many relevant terms that characterize, de facto, the SSP, have not been taken into consideration while designing and implementing the domain vocabulary. This might suggest to revise the domain vocabulary in order to include them, and better reflect the real content and usage of the SSP.

Table 3 shows that quite a half of the relevant concepts extracted from the corpora SSP2 and SSP3 were already present in the domain vocabularies and, hence, the ontology PBL.owl seems close to both corpora SSP2 and SSP3. Instead, only 20% of the relevant concepts extracted from SSP1 belong to Bachelor.owl.

For example, *Advancement*, *Argumentation*, *Degree*, *Tutor*, *Undergraduate* belong to the free tags ontology of SSP1 but correspond to no concept in Bachelor.owl.

If we go back to the challenges that we devised in the Introduction section, we notice that our approach may prove suitable for facing all of them:

1. **Automatic and dynamic content classification**. By extracting relevant concepts from a document and matching them to an existing ontology, we provide an effective and automatic means of classifying the document with respect to the domain ontology. Since the domain ontology may evolve, this activity can be carried out periodically in order to make the classification dynamic and always up-to-date.
2. **Understanding knowledge evolution**. The results reported in Table 3 suggest to revise Bachelor.owl because it does not stick any longer to the corpus of SSP1. The revision should be guided by the results of both the relevant term extraction and the ontology matching activities: useful concepts extracted form the corpus and not present in the ontology should be added to it. This activity could be carried out in an automatic way, for example by simply replacing the static domain ontology with the dynamically generated free tags ontology or by complementing the domain ontology with a subset of concepts in the free tags ontology. More sophisticated activities, such as maintaining *subClassOf* and *equivalentClass* relations consistent even when new concepts are added and old concepts are removed, require the supervision of a domain expert.
3. **Automatic and dynamic suggestion of tags and relationships among knowledge objects**. By extracting a list of free tags from the SSP corpus and matching them with existing concepts, we can give suggestions to users on how to tag the knowledge artifacts or on how to relate two artifacts, based on the current and actual content of the corpus. As this corpus evolves, suggestions will evolve too.

5 RELATED AND FUTURE WORK

The growth of interest in multidisciplinary researches such as those of social and semantic web systems is witnessed by the proliferation of works on domain ontology learning from texts and interoperability solutions for different vocabulary representation standards.

In (Velardi et al., 2007) a taxonomy learning system from web documents, called KMap, has been developed for achieving interoperability in enterprises environments. The system extracts knowledge through both automatic and manual steps, starting from web documents and using WordNet to infer relations among the extracted words and to retrieve words definitions (WordNet glosses), delegating both the taxonomy and the glosses evaluation to human experts validation procedures. In (Lae et al., 2008) an analysis of the characteristics of different tag vocabularies languages is carried out and mapping guidelines are provided. A federation of tagging ontologies is also suggested in order to define tags meaning and sharing tags from different sources. The work nearest to ours is (Zouaq and Nkambou, 2008) where a framework for learning domain ontologies in the educational field is presented. The paper depicts the TEXCOMON tool that 1) extracts knowledge from LOs (Learning Objects, a standard for educational resources representation) of a given domain; 2) builds concept maps from terms acquired from LOs; and 3) generates domain ontologies from these concept maps.

The originality of our approach with respect to the cited ones is that we reuse techniques developed in the ontology matching field in order to perform most of the challenging activities required within a Trialogical Learning system. This approach will allow us to take advantage of new ontology matching algorithms as they will appear, to obtain more and more sophisticated results almost for free. Similar considerations hold for the knowledge acquisition from texts: we use a general NLP tool that we will be able substitute with more sophisticated and efficient ones if it will be the case.

New application scenarios go in the direction of weaving the "Semantic Web joins the Social Web" paradigm. Some directions on how to analyse such paradigm are suggested in (Mika, 2007) and (Bateman et al., 2006). According to them, measures of associations can be mined from a unified analysis model coming from ontologies representing users and tags, knowledge artifacts and tags, knowledge artifacts and relationships between them, content tags and relationships. Knowledge patterns discovery, by means of semantic overlapping within different communities of practice working inside the system, also seems to be an interesting step towards the near future of knowledge practice environments. A systematic evaluation of the results of our approach from a qualitative perspective will start soon with pedagogical partners of KP-Lab project.

ACKNOWLEDGEMENTS

The 1st and 3rd authors were partly supported by the KPLab project, the 2nd author by the Italian project Iniziativa Software CINI-FinMeccanica.

REFERENCES

Bateman, S., Brooks, C., and McCalla, G. (2006). Collaborative tagging approaches for ontological metadata in adaptive e-learning systems. In *Proc. of SW-EL 2006*, pages 3–12.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th ACL Meeting*.

Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *J. Web Sem.*, 6(1):4–13.

Lae, K. H., Passant, A., Breslin, J., Scerri, S., and Decker, S. (2008). Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Proc. of IEEE-ICSC 2008*, pages 315–322.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *J. Web Sem.*, 5(1):5–15.

Miller, G. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Paavola, S. and Hakkarainen, K. (2005). The knowledge creation metaphor - an emergent epistemological approach to learning. *Science & Education*, 14:537–557.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *J. of Documentation*, 28(1):11–21.

Tzitzikas, Y., Christophides, V., Flouris, G., Kotzinos, D., Markkanen, H., Plexousakis, D., and Spyratos, N. (2007). Emergent knowledge artifacts for supporting trialogical e-learning. *J. of Web-Based Learning and Teaching Technologies*, 2(3):16–38.

Velardi, P., Cucchiarelli, A., and Petit, M. (2007). A taxonomy learning method and its application to characterize a scientific Web community. *IEEE Trans. on Knowledge and Data Eng.*, 19(2):180–191.

Zouaq, A. and Nkambou, R. (2008). Building domain ontologies from text for educational purposes. *IEEE Trans. on Learning Tech.*, 1(1):49–62.