Ontology and Lexicon: The Missing Link

Fadi Badra, Sylvie Despres, Rim Djedidi

Ontology and Lexicon: The Missing Link

Fadi Badra **Sylvie Despres** **Rim Djedidi**
LIM&BIO
Université Paris 13
UFR de Santé, Médecine et Biologie Humaine (SMBH) - Léonard de Vinci
74, rue Marcel Cachin 93017, Bobigny Cedex France
fadi@fadi.lautre.net sylvie.despres@univ-paris13.fr rim.jedidi@univ-paris13.fr

1 Introduction

Ontologies specify formally concepts and relations of a specific domain and their related constraints (axioms, rules, etc.). Lexica (or terminologies) define terms that refer to a domain as lexical entities associated to linguistic information (morpho-syntactic properties and linguistic relations between terms).

Ontology building from text methodologies deal with intermediate representation levels associated to the different available resources. These intermediate levels –including lexica and termino-ontological resources – contain rich linguistic information on the initial corpus that is lost in the formal representation of the final ontology. In fact, transition from lexical layer to ontological layer looks like a sleight of hand. It is driven implicitly as it is buried in ontologist's mind and no trace of the activity remains in the resulting ontology. This is the *missing link* between lexical and ontological layers. It is not possible to represent everything either in lexicon or in ontology. We need an interface that keeps the link between the two layers. And for that, we need to think about the format of such interface.

Several researches (Szulman et *al*., 2009; Tiscornia, 2006; Cimiano et *al*., 2011) have underlined the interest of preserving the link between lexical and ontological layers and articulating linguistic expression with the associated knowledge model. Moreover, emerging initiatives[1] are working on defining a representation model that ensures the interface between these two layers.

The paper is structured as follow: first, we summarize an experience feedback in nutrition domain through two use cases and then, we discuss raised points and conclude.

2 Experience Feedback in Nutrition Domain

Working on knowledge modelling and exploitation in nutrition domain and on recipe corpus analysis, we have been confronted to the need of articulating ontologies with their associated lexical components and thus, to the need of a representation model carrying out this articulation.

In this section, we summarize through two use cases, the raised issues and the alternatives adopted in this work.

2.1 Use case 1: a lexicon and an ontology for recipe search engine

In this use case, we have worked on improving a recipe search engine (Benamar, 2011). A lexicon has been built from recipe corpus to facilitate user query analysis. Then, an ontology has been developed to complete the bringing-in of the lexicon by providing reasoning capabilities to the search engine.

The approach adopted combines syntactic and semantic methods. A first set of domain nutrition knowledge and descriptive properties has been identified and exploited to help in selecting relevant recipes.

Recipe corpus analysis with NLP tools (Ogmios platform, TreeTagger, Yatea and SynoTerm) has produced a list of linguistic entities – associated to their lemma, grammatical category and synonyms – structured in an XML

[1] http://www.w3.org/community/ontolex/

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 16–18
Paris, 10 November 2011

lexicon. The lexicon is composed of terms related to ingredients, quantities, unities, kitchenware, preparation methods, etc. It brings a first level of enhancement to the search engine taking particularly into account term synonymy, and hyperonymy to distinguish terms designating specific notions from those designating general ones (as "monkfish" and "fish").

The aim of this work was also to provide results that are suitable to user research criteria, profile and preferences, and that give nutritional recommendations. As the lexicon is not enough to meet this purpose, we have developed an ontology that models the vocabulary associated to recipes and integrates nutritional and physiopathological knowledge.

This use case has confirmed the complementarity between lexicon and ontology and the importance of the link between them. Lexicon provides a linguistic base that allows user query analysis and facilitates extraction of recipe results. Ontology provides a semantic referent that enables reasoning mechanisms to enrich user query and extract the most suitable recipes.

To extend engine usage by including in the research recipes available on the web, it is necessary to translate imported recipes in a format that is compatible with the built ontology. Processing these recipes also needs to exploit the lexicon. XML format representing the built lexicon do not allow to fully exploit linguistic information that can be associated to ontology. A more expressive format would be more appropriate for this need.

The second use case presented in the following section, underlines the interest of using a rich lexicon representation format.

2.2 Use case 2: information extraction from recipe corpus guided by a lexicon and an ontology

A model as LEMON (McCrae et *al*., 2011(a)) provides a rich expressivity to lexicon representation associating to each lexical entry a lemma, a lexical form, components, and also a lexical sens ensuring the link with the associated ontological reference (Mccrae et *al*., 2011(b)).
Some ongoing work aims at comparing LEMON with conventional ontology lexicalization approaches in the context of ontology-based information extraction (IE).

In (Davis et *al*., 2011), LEMON was exploited to automatically generate lexical resources associated with a cooking ontology and the resulting ontology was used to semantically annotate a small text corpus of 4650 lines of cooking recipes. The study showed that the LEMON API can be easily wrapped as a resource in the open-source text analysis framework GATE (Cunningham et *al*., 2011) to write ontology-based gazetteers that exploit LEMON-generated lexical resources.

A first version of a LEMON gazetteer (called the LemonOntoGazetteer) has been implemented in GATE and its performance was compared with the state of the art (OntoRooGazeteer[2]) that is already available in GATE. While this work constitutes only a preliminary study, the first experiments were encouraging since the LemonOntoGazetteer matched 74% of the 798 annotations created by the OntoRootGazetteer.

More research is however needed in order to study more thoroughly the benefits of using LEMON for ontology lexicalization in the context of ontology-based IE. In particular, the lexical model generated by LEMON served only in a pre-processing phase to generate a list of entries to be used by a conventional list gazetteer.

Future work will include writing a full-blown LEMON Gazetteer for GATE that exploits LEMON as a lexicon model during the entity recognition phase. Besides, the performance of the LemonOntoGazetteer was only compared to the output of the OntoRootGazetteer. Running more thorough experiments will require to create a Gold Standard of cooking recipes semantically annotated by domain experts on which the performance of the LemonOntoGazetteer could be compared with other semantic annotation processes.

3 Discussion and Conclusion

Through these use cases, it appears necessary to have both a formal ontology as a semantic referent, and a lexicon as a rich linguistic base represented in an expressive format. We also need a representation format of linguistic information that preserves richness of the exploited terminological resources.

We might think to represent everything in the ontology (lexical and conceptual aspects) but this inevitably deprives lexical aspect and limits the evolution of domain lexicon. We might also

[2] http://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:ontoRootGaz

think to transform existing lexical resources by using standards as RDFS and SKOS, but this solution even if it could be applied to some resources, limits linguistic information associated to ontology (Tian, 2011; Mccrae et *al.*, 2011(b)). For some other resources, it is not even clear how to translate them automatically in these standards.

Complementarity between lexicon and ontology is thus obvious. To take advantage from it, we need to define a representation format that allows articulating lexical and ontological layers. It would be a first step to fill the *missing link*.

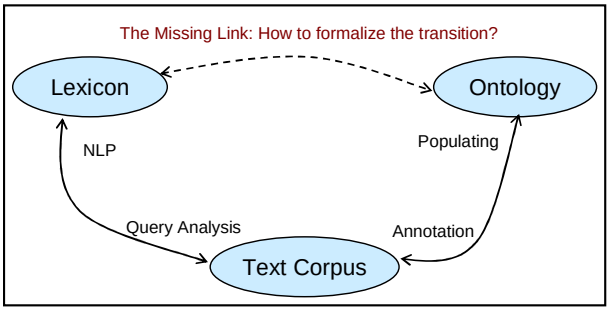

Figure 1. Ontology and Lexicon: The Missing Link

References

Benamar S. 2011. Toward integrating a knowledge layer to a research engine in nutrition domain (in French). Bioinformatics Master Dissertation. Paris 13 University.

Brian D., Badra F., Buitlelaar P., Handschuh S., Wunner T. 2011. Squeezing LEMON with GATE. The 10th International Semantic Web Conference ISWC 2011, Bonn Germany.

Cimiano P., Buitelaar P., McCrae J., Sintek M. 2011. LexInfo: A declarative model for the lexicon-ontology interface J. Web Sem. 9(1): 29-51

Cunningham, H., et *al*. 2011. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. ISBN 0956599311.

McCrae J., Aguado-de-Cea G., Buitelaar P.,Cimiano Ph., Declerck T., Gómez Pérez A. Gracia J., Hollink L., Montiel-Ponsoda E., Spohr D., Wunner T. 2011 (a). The Lemon cookbook. Monnet project.

McCrae J., Spohr D., Cimiano P. 2011(b). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete.

Szulman S., Charlet J., Aussenac-Gilles N., Nazarenko A., Sardet E., Teguiak H.V. 2009. DAFOE: an Ontology Building Platform From Text or Thesauri. In International Conference on Knowledge Engineering and Ontology Development (KEOD 2009).

Tian T. 2011. Identification And Analysis Of Lexical Resources In Nutrition Domain To Be Translated In SKOS (in French).

Tiscornia D. 2006. The LOIS Project: Lexical Ontologies For Legal Information Sharing. Proceedings of the V Legislative XML Workshop.