

Report on the joint FLaReNet – SILT workshop on lexicon-ontology relationship

The workshop on lexicon-ontology relationship has been organized as a joint event of the FLaReNet project (european) and the SILT project (U.S.A.) on September, 19th at the Istituto di Linguistica Computazionale “A. Zampolli” - CNR, in Pisa as part of the *5th International Conference on Generative Approaches to the Lexicon*. The workshop and panel discussions have seen the participation of some of the leading experts of the scientific communities involved the creation of Language Resources and Ontologies, such as Paul Buitelaar (speaker), Chu-Ren Huang (invited speaker), Massimo Poesio (invited speaker), Martha Palmer (panelist), Nicoletta Calzolari (chair), James Pustejovsky (panelist), Chumning Lee (invited speaker), Kiril Simov (invited speaker), Bernardo Magnini (panelist), Thierry Declerck (panelist), Aavid Joshi (panelist), Piek Vossen (invited speaker), Bolette Pedersen (panelist), Stelios Piperidis (discussant), Nuria Bel (discussant), Nancy Ide (discussant), Anna Rumshimsky (discussant). The aim of the workshop and of the lively debate which has followed is to address the attention of the two scientific communities on a set of open issues/questions whose answers could offer large benefits for both fields of study.

Conceptual knowledge, ontologies and lexical resources

The presence of ontologies has important effects for the development of well grounded language resources, in particular ontologies provide:

1. support for semantic annotation;
2. support for semantic search;
3. representation of application data;
4. help in the interaction with the users.

It is possible to identify two main types of ontologies: upper/core ontologies and domain ontologies. Domain ontologies are connected to the core/upper ontologies which describe more general and abstract concepts. Nevertheless, it is widely recognized that both upper/core ontologies and domain ones have limitations. In particular, the annotation of domain specific concepts is sparse in open domain corpora/text collections, while the annotation of core/upper ontological concepts is hard to be presented to the users. Some

solutions have been suggested during the debate. A first possible solution to overcome these limitations could be the introduction of a further layer of ontological representation, i.e. a middle-layer ontology, MLO. Such a level of representation could offer, on the one hand, good abstractions over the domain and, on the other hand, could result in more comprehensible representations for the end users and provide better granularities for semantic annotation. A different, but similar, solution is based on the distinction between rigid and non-rigid concepts. Ontologies should have in their structures and descriptions only the rigid concepts. This proposal is based on the claim that knowledge must be encoded into ontologies but ontologies cannot store all the knowledge of the world, in particular in the perspective of developing good applications for NLP. Rigid concepts are those which need to be present in ontological descriptions, while non-rigid ones can be “learned ” from document collections. As it appeared from the discussion, some differences are still existing in the communities on how to call the different layers in which the knowledge sources can be divided, but a common perspective on this issue is emerging and is consolidating.

Lexicalized ontologies or ontologized lexica?

One of the emerging issues which needs to be addressed, and where collaboration with other domain experts such as neuro-scientists and psychologists could be of utmost importance, is represented by the way in which *conceptual knowledge* is stored in human minds/brains. There seems to be evidence obtained from studies in different fields such as psychology, concept acquisition and neuroscience which seems to support the idea that conceptual knowledge is somehow encoded in the lexicon, though further research is needed. One of the compelling issues is the need of “empirically” grounded lexical resources and ontologies, where the term empirical is to be understood under the perspective of experiences reported by means of elicitations of human beings. So far, the models of lexica and of ontologies built in this way are too simplistic and the performance of applications trained on these studies is not constant: good performance in some tasks and bad results in other. Nevertheless, this new perspectives seem promising. A new result for the next years, connected also the notions of rigid and non-rigid concepts, could be represented by the development of empirical lexica and empirical ontologies in a two way relationships: empirical studies from neural science, psychology, natural language feed lexical and ontological constructions and viceversa.

Word meaning and world knowledge

The relationship between word meaning and world knowledge has been highly debated. Both scientific communities have developed stable methods to encode these two elements in their corresponding resources. Until recently, these two elements have had separate lives but the need to integrate lexical resources and ontologies is now a compelling issue. The two scientific communities agree on the vision that conceptual modeling and linguistic grounding are tied together but need to be best separated at a definitional and conceptual level. One of the main shortcomings of current ontologies is that no linguistic information is associated with the natural language terms used to describe the ontological concept. RDF and SKOS have improved the descriptive power and representation format of ontologies but do not provide a solution to the issue of the linguistic grounding problem of ontologies, though they do not have to. One of the major risk in this attempt to ground ontologies to the lexicon is the creation of ontologized lexicon or, worse, lexicalized ontologies, but the integration of the two domain is extremely necessary to boost the performance of applications and improve the Sematic Web. One of the most difficult aspect is represented by the fact that very often to concepts in an ontology do not correspond words, or single lexical items in the lexicon but phrases or sentences. Some of the desiderata which this integration process has to take into account are the following:

1. strict separation between the linguistic and conceptual levels of formalization and description;
2. lexica must support detailed information on the linguistic realization of the ontological concepts, ranging from part-of-speech and morphology to grammatical constructions;
3. both lexica and ontologies must support multilinguality

Some proposals on how to perform this integration process between lexica and ontologies are already available to the scientific communities. A necessary effort from both scientific communities on tackling this issue and start up a debate on possible solutions and elaboration of standards and best practices is required and of utmost importance.

