Matching Natural Language Data with Ontologies

Johannes Heinecke

Orange Labs, F-22307 Lannion cedex – France
johannes.heinecke@orange-ftgroup.com

Abstract. Ontologies and natural languages are complementary. Whereas ontologies are used to model knowledge formally, natural language is primarily used by users to communicate with ontology based systems. In order to transform information or queries in natural language into valid ontological expressions, the meaning of natural language entities have to be matched with the given ontologies. In contrast to pure ontology matching, the matching with natural language data poses some problems linked to their ambiguities (synonymy, homonymy/polysemy, redundancy, to name but a few).

1 Introduction, context and related work

In the context of the Semantic Web, the interfacing of ontological representation and natural language is an important issue. Since much information on the Web exists (only) in textual form, the usage of this information in ontology based tools is not possible unless these texts are made accessible or comprehensible by such tools. This means that texts and user queries have to be "translated" into an ontological representation language such as the W3C languages RDF/RDFS and OWL.

The need for the work described here came from the aceMedia project (*http://www.acemedia.org/*) [1,2] (cf. also [3]). In this project, the two tasks are

transforming textual annotations of multimedia contents into an ontological representation (based on an existing ontology) in order to make them available for a knowledge-base; and translating English and French user queries into an ontological query language (in our case SPARQL). The matching of linguistic data (lexicons, thesauri) with ontologies is similar but not identical to ontology matching or ontology alignment, i.e. trying to find corresponding classes of ontology *A* in ontology *B* [4]. Different methods of matching are discussed in detail by [5, p. 65]. Following this classification the present approach can be considered being terminological and linguistic, since we use relationships found in the lexicon (via a semantic thesaurus, [6]) and the taxonomic hierarchies of both, the lexical semantic data and the ontologies notably for the disambiguation of polysemous words. Similar work describe [7] and [8]. In contrast to their results we do not have a classification at hand.

2 Linguistic-ontology matching

Apart from the ontologies, the matching requires a complete lexicon of the language used to label or describe the ontological classes and properties (= entities). Our lexicon is lso linked to a semantic thesaurus. The ontologies, on the other hand, usually have non-ambiguous entity labels (like *http: //www.acemedia.org/ontos/tennis#Player*[1]) or a comment, explaining the entity. This is especially necessary if the entity labels are not self-explanatory like

[1] We shorten name spaces like *http://www.acemedia.org/ontos/tennis#* to "*tennis:*" etc.

tennis:C12 (a fortunately rare case). Further, the semantic thesaurus contains a thematic hierarchy of all semantic concepts to help disambiguation. These are grouped into 880 themes which in turn are organized in 80 domains. Domains are divided into about 10 macro-domains. The matching itself comprises several steps (cf. fig. 1). Apart from a (more or less manual) preparation in order to correct possible labeling errors in the ontologies, the other steps do not need any intervention: (a) extracting the "ontological context" of entities and assigning eventual reformulations of entity labels; (b) natural language processing passes: detecting meanings for classes using their ontological context (direct sub-classes); (c) and for properties using their ontological context (domain and range classes); (d) determining the application depending synonyms and co-hyponyms; (e) adding the ontological hierarchy to the semantic taxonomy; (f) creating semantic transformation rules for "complex class labels"[2]; (g) creating transformation rules for the creation of ontological representation (from semantic graphs. Synonyms (defined in our multilingual thesaurus, [9]) are all matched onto the same ontological class (e.g. "river", "stream", "creek" etc. → *holidays:River*). If a class has no sub-classes, we also match the co-hyponyms of the label to the class (e.g. in our case "car", "bus", "truck", "motorbike" ... → *general:Vehicle*. The resulting linguistic data is successfully used the aceMedia prototype, similarly produced data is used in an industrial application to create and access ontological based information from/via natural language.

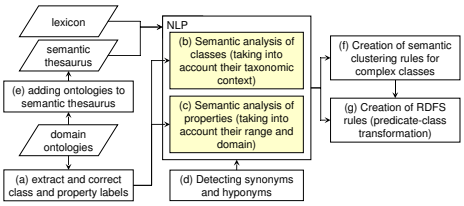

Fig. 1. linguistic-ontological matching

New perspectives are offered by structured semantic data which is getting more and more available. Databases like Wikipedia (especially the categorization schema used within) or RDF or ontology based information systems like DBpedia or freebase[3] (both initialized by Wikipedia contents) will help to improve the linking of natural languages and formally modeled ontologies.

References

1. Heinecke, J.: Génération automatique des représentations ontologiques. In: TALN. (2006) 502–511
2. Dasiopoulou, S., Heinecke, J., Saathoff, C., Strintzis, M.G.: Multimedia reasoning with natural language support. In: IEEE-ICSC. (2007) 413–420
3. Heinecke, J., Toumani, F.: A Natural Language Mediation System for E-Commerce applications. In: Workshop HLT for the Semantic Web and Web Services. ISWC. (2003) 39–50
4. Ehrig, M., Staab, S.: QOM - quick ontology mapping. In: ISWC. (2004) 683–697
5. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
6. Heinecke, J., Smits, G., Chardenon, C., Guimier De Neef, E., Maillebuau, E., Boualem, M.: TiLT : plate-forme pour le traitement automatique des langues naturelles. TAL **49:2** (2008)
7. Giunchiglia, F., Marchese, M., Zaihrayeu, I.: Encoding classifications into lightweight ontologies. Journal of Data Semantics VIII (2007) 57–81
8. Reiter, N., Hartung, M., Frank, A.: A resource-poor approach for linking ontology classes to Wikipedia articles. In: STEP. (2008) 381–387
9. Chagnoux, M., Heinecke, J.: Aligner ontologies et langues naturelles. gérer la synonymie. In: Plateforme AFIA, Grenoble (2007) 87–94

[2] Labels which use multi-word expressions like *tennis:ExhibitionMatch* instead of simple words.

[3] *http://dbpedia.org/*, *http://freebase.com/*