Chapter VII
Lexical Enrichment of Biomedical Ontologies

Nils Reiter
Heidelberg University, Germany

Paul Buitelaar
DERI - NLP Unit, National University of Ireland Galway, UK

ABSTRACT

This chapter is concerned with lexical enrichment of ontologies, that is how to enrich a given ontology with lexical information derived from a semantic lexicon such as WordNet or other lexical resources. The authors present an approach towards the integration of both types of resources, in particular for the human anatomy domain as represented by the Foundational Model of Anatomy and for the molecular biology domain as represented by an ontology of biochemical substances. The chapter describes our approach on enriching these biomedical ontologies with information derived from WordNet and Wikipedia by matching ontology class labels to entries in WordNet and Wikipedia. In the first case the authors acquire WordNet synonyms for the ontology class label, whereas in the second case they acquire multilingual translations as provided by Wikipedia. A particular point of emphasis here is on selecting the appropriate interpretation of ambiguous ontology class labels through sense disambiguation, which we address by use of a simple algorithm that selects the most likely sense for an ambiguous term by statistical significance of co-occurring words in a domain corpus. Acquired synonyms and translations are added to the ontology by use of the LingInfo model, which provides an ontology-based lexicon model for the annotation of ontology classes with (multilingual) terms and their linguistic properties.

INTRODUCTION

As information systems become more and more open, i.e. by including web content, as well as more complex, e.g. by dynamically integrating web services for specific tasks, data and process integration becomes an ever more pressing need - in particular also in the context of biomedical information systems. A wide variety of data and processes must be integrated in a seamless way to provide the biomedical professional with fast and efficient access to the right information at the right time.

A promising approach to information integration is based on the use of ontologies that act as a formalized inter-lingua onto which various data sources as well as processes can be mapped. An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest as defined by Gruber (1993), where 'formal' implies that the ontology should be machine-readable and 'shared' that it is accepted by a community of stakeholders in the domain of interest. Ontologies represent the common knowledge of this community, allowing its members and associated automatic processes to easily exchange and integrate information as defined by this knowledge.

For instance, by mapping a database of patient radiology reports as well as publicly accessible scientific literature on related medical conditions onto the same ontological representation a service can be build that provides the biomedical professional with patient-specific information on up-to-date scientific research. Scenarios like these can however only work if data can be mapped to ontologies on a large-scale, which implies the automation of this process by automatic semantic annotation. As a large part of biomedical data is available only in textual form (e.g. scientific literature, diagnosis reports), such systems will need to have knowledge also of (multilingual) terminology in order to correctly map text data to ontologies.

This chapter is therefore concerned with the enrichment of ontologies with (multilingual) terminology. We describe an approach to enrich biomedical ontologies with WordNet (Fellbaum, 1998) synonyms for ontology class labels, as well as multilingual translations as provided by Wikipedia. A particular point of emphasis is on selecting the appropriate interpretation of ambiguous ontology class labels through sense disambiguation. Acquired synonyms and translations are added to the ontology by use of the LingInfo model, which provides an ontology-based lexicon model for the annotation of ontology classes with (multilingual) terms and their linguistic properties.

Related work to this chapter is on word sense disambiguation and specifically domain-specific word sense disambiguation as a central aspect of our algorithm lies in selecting the most likely sense for ambiguous labels on ontology classes. The work presented here is based directly on Buitelaar & Sacaleanu (2001) and similar approaches (McCarthy et al., 2004a; Koeling & McCarthy, 2007). Related to this work is the assignment of domain tags to WordNet synsets (Magnini & Cavaglia, 2000), which would obviously help in the automatic assignment of the most likely synset in a given domain – as shown in Magnini et al. (2001). An alternative to this idea is to simply extract that part of WordNet that is directly relevant to the domain of discourse (Cucchiarelli & Velardi, 1998; Navigli & Velardi, 2002).

However, more directly in line with our work on enriching a given ontology with lexical information derived from a semantic lexicon is presented in Pazienza and Stellato (2006). In contrast to Pazienza and Stellato (2006), the approach we present in this chapter uses a domain corpus as additional evidence for statistical significance of a synset.

Finally, some work on the definition of ontology-based lexicon models (Alexa et al., 2002; Gangemi et al., 2003; Buitelaar et al., 2006) is of