Chapter 6
A Survey of Domain Ontology Engineering: Methods and Tools

Amal Zouaq[1] and Roger Nkambou[2]

[1] Simon Fraser University Surrey, 13450 102 Ave. Surrey, BC V3T 5X3 Canada
azouaq@sfu.caa

[2] University of Québec at Montréal, 201 Du Président-Kennedy Avenue, PK 4150,
Montréal, QC, H2X 3Y7, Canada
Nkambou.roger@uqam.ca

Abstract. With the advent of the Semantic Web, the field of domain ontology engineering has gained more and more importance. This innovative field may have a big impact on computer-based education and will certainly contribute to its development. This chapter presents a survey on domain ontology engineering and especially domain ontology learning. The chapter focuses particularly on automatic methods for ontology learning. It summarizes the state of the art in natural language processing techniques and statistical and machine learning techniques for ontology extraction. It also explains how intelligent tutoring systems may benefit from this engineering and talks about the challenges that face the field.

6.1 Introduction

As described in chapter 2, the expert module is responsible for the learning content which indicates what can be taught by the ITS (the domain model). In this regard, some of the most important research issues that need to be addressed are: how the expert module can be effectively modeled, what kinds of knowledge representations are available and what kind of knowledge acquisition techniques are applicable. In fact, one of the main obstacles to ITSs development and wide dissemination is the cost of their knowledge base and particularly the cost of producing the domain model from scratch. Faced with these knowledge acquisition challenges, many attempts have been made to create automated methods for domain knowledge creation. However, these attempts have not been as successful as one would wish them to be. Moreover, these efforts have not led to reusable and standard methods and formalisms for knowledge base creation and update.

With the advent of the Semantic Web, new research avenues have been created, especially within the domain ontology engineering field. The research community now acknowledges the need to create domain ontologies in a (semi)automatic way. Representing knowledge using domain ontologies has two main advantages: first, their standard formalism makes it possible to share and reuse ontologies

R. Nkambou et al. (Eds.): Advances in Intelligent Tutoring Systems, SCI 308, pp. 103–119.
springerlink.com

between any ontology-friendly environments. Second, their formal structure makes it possible to work out how to obtain knowledge representations and figure out how to automatically extract ontological components in modular layers. This automatic ontological extraction is known as "**Ontology Learning**".

In general, the entire knowledge acquisition process is a tedious task, including such difficulties as building, reusing and propagating intelligent tutoring systems. In fact, we need to use standard representations to modularize the creation, evolution and maintenance of intelligent tutoring systems. It is also a necessary to provide explicit semantic relationships between the learning content concepts and to develop pedagogical activities that are built on this domain knowledge. With the advent of the Semantic Web, many questions have been raised about how the domain model could benefit from the Semantic Web languages and techniques. A successful integration of the Semantic Web and the ITS philosophy could ensure better reuse of ITS components and better sharing and engineering of domain knowledge. Because ontologies are the backbone of the Semantic Web, using them to represent domain and instructional knowledge can be an interesting avenue. These questions have been particularly high on the list with the rise of the *Educational Semantic Web* (Aroyo and Dicheva, 2004), which comes from the eLearning field and which has proposed the use of ontologies to index and structure the learning content. Intelligent tutoring systems have been slower in adopting the 'ontology' concept, especially for modeling domain knowledge but this is now an undeniable fact. Intelligent tutoring systems can benefit from ontology engineering because ontologies represent a standard way for modeling knowledge. They are expressed using formal and standard languages which facilitate sharing and reasoning. Moreover, there is a growing awareness within the ITS and eLearning communities of the importance of adopting common methods for domain knowledge acquisition and representation. As a result, ITS stands to benefit from the huge number of available eLearning resources. Similarly, eLearning systems will benefit from ITS domain modeling and reasoning. Finally, since ITSs are domain-dependent, it is important to develop easy and reusable knowledge acquisition tools and to integrate automatic methods for this acquisition and evolution. Ontology engineering can provide an answer to these needs and the following section introduces the reader to domain ontology engineering.

This chapter presents an overview of domain ontology engineering and focuses particularly on automatic methods for ontology learning, especially from texts. It is organized as follows. After the introduction, section 2 briefly explains the field of ontology engineering. Section provides an overview of the ontology learning process from text. Each task and component of this process is explained. We also present, for each task, the natural language processing (NLP) techniques and the statistical and machine learning techniques. Sections four and five, in addition to an ontology update task, briefly introduce the ontology learning process from other non text-based sources.. Section six highlights the more general challenges that face domain ontology engineering as well as more specific ITS-related ones. The entire chapter is summarized in the conclusion.

6.2 Ontology and Ontology Engineering

Before going into further detail, it is important to first define the notion of ontology. Very briefly, ontology is a formal specification of a conceptualization (in this case, a domain) and it includes the definition of classes, objects, properties, relationships and axioms. Ontologies are expressed using a formal language such as RDF or OWL and support automatic inference. Generally, ontologies involve a kind of consensus within a community, meaning that they formalize concepts that are generally accepted within this community. There are many kinds of ontologies such as upper-level ontologies, task ontologies and domain ontologies. We are especially interested here in domain ontologies.

As previously pointed out, the concept of a domain ontology as envisioned by the eLearning community is relatively new in the field of ITS. However, domain ontology engineering is a growing research area that has received much attention in other fields and it is the corner stone of the Semantic Web. Ontology engineering is a field that explores the methods and tools for handling the ontology lifecycle. It requires a general and domain-independent methodology that provides guidance for ontology building, refinement and evaluation (Guarino and Welty, 2002). The ontology life-cycle can be schematized in four main stages: the specification stage, the formalization stage, the maintenance stage, and the evaluation stage.

- *The specification stage* identifies the purpose and scope of the ontology. Generally, this relies a lot on domain experts and needs to define the competency questions that the ontology has to to answer. It is also dependent on the application that is going to be used by the ontology;
- *The formalization stage* produces a conceptual and formal model that meets the requirements of the specification stage;
- *The maintenance stage* keeps track of the ontology's updates and evolution, and checks its consistency;
- Finally, the *evaluation stage* analyzes the resulting ontology and checks if it meets the initial needs and has the desired features.

At this point, we are especially interested in the formalization stage and how it can benefit from automated methods for knowledge acquisition. In fact, the most common and successful techniques for domain engineering are generally manual and the best ITS authoring tools can help the expert formalize his knowledge but these tools are generally far from being part of an automated procedure (see chapter 18). It is therefore worthwhile to explicitly state the steps that can be automated to alleviate the task of human experts and the burden of knowledge acquisition. Ontology learning techniques have been adopted to achieve this goal (Aussenac-Gilles et al., 2000). These learning techniques can vary according to the degree of automation (semi-automatic, fully automatic), the ontological knowledge that has to be extracted (concepts, taxonomy, conceptual relationships, attributes, instances, axioms), the knowledge sources (texts, databases, xml documents, etc.) and finally the purpose (creating ontologies from scratch and/or updating existing ontologies).

6.3 Building Domain Ontologies from Texts

This section focuses on ontology learning from texts. At the specification stage, the knowledge engineer should prepare a corpus related to the domain of interest. Of course, this corpus has to be carefully chosen and should properly describe the domain. A number of sub-tasks have to be performed in order to learn a domain ontology including concepts, taxonomy, conceptual relationships, attributes, instances and axiom learning. Examples of systems that completethe ontology learning task include: Text-2-Onto (Cimiano and Volker, 2005a), TEXCOMON (Zouaq and Nkambou, 2009a; Zouaq and Nkambou, 2009b), OntoLearn (Velardi et al., 2005), and OntoGen (Fortuna et al., 2007). In the following sections, we highlight state-of-the-art knowledge extraction techniques used in each of the ontology learning sub-tasks. Each of these sub-tasks has an NLP-based technique and a statistical and machine learning technique.

6.3.1 Concept Extraction

The first task that has to be performed in ontology engineering is the identification of concepts. Concepts can be described as complex mental objects that are characterized by a number of features. Concept extraction refers to the identification of important domain classes.

Concepts are terms that are particularly important for the domain when using terminological approaches. These terms are generally extracted from the corpus as outlined by Buitelaar et al. (2005) who consider that a concept should have a linguistic realization. In this case, it is quite challenging to differentiate domain terms from non domain terms, especially when using statistical filtering. The identified terms—composed from single or several words—can then be either considered as specific concepts/classes or they can be classified according to broad classes already available in thesauri and vocabularies. Other approaches rely on clustering and machine learning as a way of learning semantic classes. In this case, a concept may have no corresponding term in the corpus. This is further explained in the following paragraphs.

6.3.1.1 NLP-Based Techniques

NLP-based techniques for concept learning consider terms as candidate concepts. These approaches rely on linguistic knowledge and use parsers and taggers to determine the syntactic roles of terms or to unveil linguistic patterns. Some works typically adopt a surface analysis by running a part-of-speech tagger over the corpus and identifying manually defined patterns (Sabou, 2005; Moldovan and Girju, 2001) while others use a deep-level analysis and use a NLP parser (Reinberger and Spyns, 2005; Zouaq and Nkambou, 2009a). In general, the syntactic analysis identifies the nominal phrases that may be important for the domain. For example, Zouaq and Nkambou (2009a) use dependency relationships indicating nominal phrases such *nominal subject*, *direct object* and *noun compound modifier* to detect these nominal phrases. Most of the time, there is also a list of manually defined seed words that triggers the ontology learning process. However, Zouaq and Nkambou (2009a) proposed the use of an automatic keyword extractor to help automate this task.

6.3.1.2 Statistical and Machine Learning Techniques

Usually, NLP-based approaches are not used alone and require statistical filtering. Statistical approaches consider all important terms in a domain as potential concepts and require quantitative metrics to measure the weight of a term. Such quantitative measurements include the popular TF*IDF (Salton and Buckley, 1988) and C-value/NC-value (Frantzi et al., 1998). The employed measurements can differ depending on the application.

Clustering techniques based on Harris' distributional hypothesis (Harris, 1954), can also be used to induce semantic classes (Almuraheb and Poesio, 2004; Lin and Pantel 2001). Here, a concept is considered as a cluster of related and similar terms. Harris' hypothesis, which is the basis of word space models, states that words that occur in similar contexts often share related meaning (Sahlgren 2006). Term similarity can be computed using collocations (Lin 1999), co-occurrences (Widdows and Dorow, 2002) and latent semantic analysis (Hearst and Schutze, 1993). For example, Lin and Pantel (2001) represent each word by a feature vector that corresponds to a context in which the word occurs. The features are specific dependency relationships coupled with their occurrence in the corpus. The obtained vectors are then used to calculate the similarity of different terms using measurements such as mutual information (Hindle, 1990; Lin, 1998) and to create clusters of similar terms. Comparable approaches include Formal Concept Analysis (such as the approach presented in (Cimiano, 2006)) and Latent Semantic Indexing algorithms (e.g. Fortuna et al., 2005). These approaches build attributes/values pairs that correspond to concepts.

Statistical approaches can also be used on top of NLP-based approaches to identify only relevant domain terms by comparing the distribution of terms between corpora (Navigli and Velardi, 2004). Another approach used by (Velardi et al., 2005) linguistically analyses WordNet glosses (textual description) in order to extract relevant information about a given concept and enrich its properties. This analysis can help detect synonyms and related words and can contribute to concept definition. In fact, concept learning requires not only that conceptual classes be identified but also makes it necessary to describe concepts through the identification of attributes, sub-classes and relationships. This is further explained in the following sections.

6.3.2 Attribute Extraction

Since concepts are characterized by a number of features, it is important to unveil the distinctive attributes or properties that define a concept. In his ontology, Guarino (1992) distinguishes between relational and non-relational attributes. Relational attributes include qualities and relational roles, while non-relational attributes include parts. Following Guarino (1992) and Pustejovsky (1995), Almuraheb and Poesio (2005) presented another scheme for classifying attributes into qualities, parts, related-objects, activities and related-agents.

In this chapter, attributes designate a data type property such as *id, name*, etc., in contrast with object properties which are considered as conceptual relationships—these are addressed in the Conceptual Relationships Extraction section.

6.3.2.1 NLP-Based Techniques

According to Poesio and Almuhareb (2005), the right meaning of attributes can be found by looking at Wood's linguistic interpretation (Wood, 1975): *Y is a value of the attribute A of X if it is possible to say that Y is an A of X (or the A of X).* If it is not possible to find a Y then A cannot be an attribute. In order to comply with this linguistic interpretation, linguistic patterns are also proposed for the detection of attributes. Following Woods (1975), Almuhareb and Poesio (2005) suggested the use of the following patterns in order to search for attributes of a concept C:

- "(a|an|the) * C (is|was)" (e.g.: *a red car is…*).
- "The * of the C (is|was)" (e.g.: *the color of the car is…*)
- "The C's * R" (e.g.: *The car's price is…*) where R is a restrictor such as "is" and the wildcard denotes an attribute.

Cimiano (2006) proposed another set of patterns for attribute extraction based on adjective modifiers and WordNet and presented a number of interesting patterns describing attributes and their range according to syntax (parts-of-speech).

6.3.2.2 Statistical and Machine Learning Techniques

As indicated earlier, natural-language processing techniques are generally coupled with statistical filtering and machine learning. Poesio and Almuraheb, (2005) proposed a supervised classifier for learning attributes based on morphological information, an attribute model, a question model, and an attributive-usage model. These models are used to differentiate types of(different kind of) attributes based on a specific classification scheme. In Poesio and Almuhareb (2008), the Web is used to extract concept descriptions. Another approach, proposed by Ravi and Pasca (2008), describes a weakly supervised classifier for learning attributes and values, based on a small set of examples.

6.3.3 Taxonomy Extraction

One of the most important tasks in knowledge engineering is the organization of knowledge into taxonomies which indicate generalization/specialization relationships between classes. These relationships enable inheritance between concepts and automated reasoning (Corcho & Gomez-Perez, 2000).

6.3.3.1 NLP-Based Techniques

The most common way of extracting taxonomical links is the use of specific lexico-syntactic patterns as proposed by Hearst (Hearst, 1992). In Pattern-based techniques, the text is scanned for instances of distinct lexico-syntactic patterns that indicate a taxonomical link. Patterns are usually expressed as regular expressions (Cimiano and Volker, 2005b) but they can also be represented by dependency relationships (Zouaq and Nkambou, 2009a; Lin and Pantel, 2001).

Since a domain corpus is sparse and because hierarchical patterns are rare in domain-specific corpora, many approaches extend the corpus by a search of taxonomical links in dedicated resources such as WordNet (Snow et al., 2004) or on the Web (Cimiano et al, 2004; Maedche and Staab, 2001) so as to increase their recall (Etzioni et al., 2004). A remedy for the burden of manually defining patterns is proposed by Snow et al. (2004) using a classifier for automatically learning hyponym (is-a) relations from text based on dependency paths and using WordNet.

Other linguistic approaches use the internal structure of multiple-word terms (nouns phrases) in order to deduce taxonomical links. For example, there is a taxonomical link between a term and the same term modified by an adjective (e.g.: an intelligent man is-a man). This approach is quite popular (Buitelaar et al., 2003) (Velardi et al., 2005; Zouaq and Nkambou, 2009a).

6.3.3.2 Statistical and Machine Learning Techniques

Statistical and machine learning approaches for taxonomy learning rely on Harris' distributional hypothesis, just as those used in concept learning. Hierarchical clustering algorithms are used to extract taxonomies from text and produce hierarchies of clusters. Maedche et al. (2002) describe the two main approaches that can be used to implement hierarchical clustering: the bottom-up approach which starts with individual objects and groups the most similar ones, and the top-down approach, where all the objects are divided into groups. This approach has been used in many works such as Bisson et al. (2000), Carabello (1999), and Faure and Nedellec (1998). Typically, as highlighted by Cimiano et al. (2004), a term t is a subclass of t2 if all the syntactic contexts in which t appears are also shared by t2. The syntactic contexts are used as feature vectors and a similarity measure is applied. For example, in order to compute the relation *is_a (t, t2)*, Cimiano et al. (2004) applied a directed Jaccard coefficient computing the number of common features divided by the number of features of term t.

Cimiano et al. (2004) propose also the use of multiple sources of evidence and techniques in order to learn hierarchical relationships. Similarly, Widdows (2003) proposes the use of unsupervised methods combining statistical and syntactic information to update an existing taxonomy with new terms.

6.3.4 Conceptual Relationships Extraction

Conceptual relations refer to any relationship between concepts aside from taxonomic relations. Specific conceptual relationships may include synonymy, part-of, possession, attribute-of, causality, as well as more general relationships referring to any labeled link between a source concept (the domain of the relation) and a destination concept (the range of the relation). In the following sections, we identify the different techniques used to describe specific relationships and generic relationships.

6.3.4.1 NLP-Based Techniques

In the information extraction community, conceptual relation extraction is known as template filling, frame filling, semantic role labeling or event extraction. In this case, it relies on lexico-semantic lexicons such as FrameNet (Baker et al., 1998) and VerbNet (Kipper et al., 2000) to extract particular relationships and to assign roles (such as Agent, Theme, etc.) to the arguments of the relation. Approaches based on frames include ASIUM (Faure and Nédellec, 1998) which enables an acquisition of relations between concepts based on triggering words. Another work related to roles is the identification of Qualia structures by Pustejovsky (1995). These qualia structures can help identify particular relationships as shown by Cimiano and Wenderoth (2005) who proposed a number of linguistic patterns indicating the different roles defined by Pustejovsky.

There is quite a lot of work on the use of linguistic patterns to unveil ontological relations from text. Following Hearst's work (Hearst, 1992) on taxonomic relations, different researchers created patterns for non-hierarchical relationships (Iwanska et al., 2000; Zouaq and Nkambou, 2009a), for part-of relations (Charniak and Berland, 1999; Van Hage et al., 2006) or causal relations (Girju et al., 2003). In fact, many works consider that ontological relationships are mostly represented by verbs and their arguments. In the same line of research, Navigli and Velardi (2004) use patterns expressed as regular expressions and restricted by syntactic and semantic constraints. Finally, WordNet can be used to extract synonyms, antonyms and other kinds of relationships. This also involves the detection of the right meaning of the term and thus the use of word meaning disambiguation algorithms.

6.3.4.2 Statistical and Machine Learning Techniques

Most of the work on relation extraction combines statistical analysis with more or less complex levels of linguistic analysis. For example, Zouaq and Nkambou (2009b) use typed dependencies to learn relationships and statistical measurements which in turn are used to determine whether or not the relationships should be included in the ontology.

Other machine learning techniques for learning qualia structures include the work of Claveau (2003) using inductive logic programming or Yamada and Baldwin (2004) whose work relies not only on lexico-syntactic patterns but also on a maximum entropy model classifier. Cimiano and Wenderoth (2007) developed an algorithm for generating a set of clues for each qualia role: download the snippets of the first 10 Google hits matching the generated clues, part-of-speech-tagging of the downloaded snippets, matching regular expressions conveying the qualia role of interest and finally weighting the returned qualia elements according to some measure.

An interesting technique for learning non labeled relationships is the use of association rule learning, where association rules are created from the co-occurrence of elements in the corpus. This technique has been adopted by the Text-to-Onto system (Maedche and Staab, 2001). However, these relationships should be manually labeled later and this task is not always easy for the ontology engineer.

6.3.5 Instance Extraction

Instance extraction, also known as Ontology Population (OP), is a classification task which aims at finding instances of concepts defined in an ontology. It is similar to Named Entity Recognition (e.g. Person, Location, Organization, etc.), which is often used in information extraction. Examples of systems especially devoted to instance extraction include WEB→KB (Craven et al., 2000) and Know-it-All (Etzioni et al., 2004).

6.3.5.1 NLP-Based Techniques

There are a number of approaches that use NLP-based techniques for ontology population. A pattern-based approach similar to the one presented in the taxonomy extraction section relies on Hearst patterns (Hearst, 92; Schlobach et al., 2004; Zouaq and Nkambou, 2009b; Etzioni et al., 2004) or on the structure of words (Velardi et al., 2005). These approaches try to find explicitly stated "is-a" relationships. Other linguistic approaches are based on the definition or the acquisition of rules. For example, the work of (Amardeilh et al., 2005) proposes the definition of acquisition rules that are activated once defined linguistic tags are found. These tags are mapped onto concepts, attributes and relationships from the ontology and help find instances of these elements.

6.3.5.2 Statistical and Machine Learning Techniques

There are supervised and weakly supervised techniques for ontology population (Tanev and Magnini, 2006). Among the weakly supervised techniques, Cimiano and Volker (2005b) used vector-feature similarity between each concept c and a term to be categorized t. Cimiano and Volker evaluated different context features (word windows, dependencies) and showed that syntactic features work better. Their algorithm assigned a concept to a given instance by computing the similarity of this instance feature vector and the concept feature vector. Tanev and Magnini (2006) used syntactic features extracted from dependency parse trees. Their algorithm required only a list of terms for each class under consideration as training data.

Supervised techniques for ontology population ensure higher accuracy. However, they require the manual construction of a training set, which is not scalable (Tanev and Magnini, 2006). An example of a supervised approach is the work of Fleischman (2001); Fleischman and Hovy (2002) which involved designing a machine learning algorithm for fine-grained Named Entity categorization. Web->KB (Craven et al., 2000) also relies on a set of training data, which consists of annotated regions of hypertext that represent instances of classes and relations, used to extract named entities. Based on the ontology and the training data, the system learns how to classify arbitrary Web pages and hyperlink paths.

6.3.6 Axioms Extraction

Axiom extraction represents one of the most difficult tasks of ontology learning. Axioms express necessary and sufficient conditions that are used to constrain the information contained in the ontology and to deduce new information (Shamsfard and Barforoush, 2003). Few systems have tackled the problem of axiom extraction. HASTI is a system that translates explicit axioms in conditional and quantified natural language sentences to logically formatted axioms in KIF (Shamsfard and Barforoush, 2002). LExO2 (Volker et al, 2008) is another initiative for transforming natural language sentences (definitions) into description logic axioms.

6.3.6.1 NLP-Based Techniques

Natural language techniques for axiom extraction rely on the syntactic transformation of natural language (definitions) into description logic axioms (Volker et al, 2008). This supposes the availability of such definitions. Volker et al (2008) also focus on learning a particular axiom which is disjointedness by a lexico-syntactic pattern used to detect enumerations. Their underlying assumption is that terms which are listed separately in an enumeration mostly denote disjointed classes. Zouaq and Nkambou (2009b) describe a pattern for defining equivalent classes. This pattern is based on the appositive grammatical relationship between two terms to indicate that these terms are similar and denote the same concept. Another interesting work is the Lin and Pantel (2001) approach (which uses of paths in dependency trees to learn similar relationships. This makes it possible to create inverse properties for these relationships, such as *X solves Y* and *Y is solved by X*.

6.3.6.2 Statistical and Machine Learning Techniques

To the best of our knowledge, there are very few machine learning approaches for learning axioms. A machine learning classification approach has also been used by Volker et al. (2008) to determine disjointedness of any two classes. They automatically extract lexical and logical features that provide a basis for learning disjointedness by taking into account the structure of the ontology, associated textual resources, and other types of data. The features are then used to build an overall classification model.

6.4 Ontology Learning from Non-text Sources

Ontology learning from non text sources involves the use of structured or semi-structured data as input to the learning process. Structured data refer to already defined knowledge models including database schemas or existing ontologies. Semi-structured data designates the use of some mixed structured data with free text such as Web pages, Wikipedia, dictionaries and XML documents. The existence of a structure helps direct the ontology learning process towards relevant parts of data.

6.4.1 NLP-Based Techniques

Most of the approaches for ontology learning from non text sources rely on linguistic techniques or on the underlying schema already available in the structure. As previously indicated, some researchers such as Cimiano and Staab (2004) have used the Web to deal with the data sparseness problem within domain corpora. In this case, the Web is used to retrieve and learn patterns. Other works rely on dictionaries such as WordNet and try to parse natural language definitions and WordNet Synsets. Examples of such works include OntoLearn (Navigli et al., 2004; Velardi et al., 2005; and Rigau et al., 1998). Others rely on thesauri as the knowledge source (Van Assem et al., 2004).

The approach of Volz et al. (2003) converts an xml schema into a domain ontology by translating non-terminal and terminal symbols into concepts and roles using a set of rules. Similarly, the work of Stojanovic et al. (2002) uses a rule mapping scheme to convert an xml schema or a relational database schema into a domain ontology. Finally, we can cite the work of Delteil et al. (2001) who created an ontology learning procedure from RDF annotations and Nyulas et al. (2007) who created a plug-in (for Protégé) for importing relational databases into an ontology editing environment.

6.4.2 Statistical and Machine Learning Techniques

The approach of Suryanto and Compton (2001) aims at learning an ontology from a rule knowledge base in the medical domain. Their algorithm creates a set of classes and uses statistical measures to determine the relationships between the classes (subsumption, similarity, mutual-exclusivity). The work of Jannink and Wiederhold (1999) extracts a graph structure from dictionaries and uses statistical filtering and the PageRank algorithm to determine important relationships and concepts. Another example is the work of Papatheodorou et al. (2002), who build taxonomies using cluster mining from xml or RDF domain repositories.

6.5 Ontology Update and Evolution

Despite the important number of initiatives for ontology learning, the results are not still completely satisfactory and the field has to gain more maturity. Moreover, the evolution of ontologies seems to be even less supported in the research community. In fact, enabling this evolution (semi)automatically is a key factor for the Semantic Web and this involves the ability to update an ontology with new concepts, relationships, properties and axioms, the ability to appropriately place a concept in the taxonomy and the ability to perform mapping and alignment between existing ontologies. Here again, we have only provided a brief and incomplete overview of the NLP-based and statistical and machine learning techniques. We also want to point out that we have notdealt with change, versioning and consistency management during the evolution process, as we prefer to refer the reader to Haase and Sure (2004) and Flouris et al. (2006) to gain more insight into this question. Other interesting questions not dealt with here include: ontology matching and alignment (Shvaiko and Euzenat, 2008).

Ontology evolution can target each component of the ontology learning process. From the NLP side, enriching an existing concept with new attributes and relationships has been done in the work of Velardi et al. (2005) by searching the concept in WordNet and reusing its Synsets to enrich the ontology. This involves word meaning disambiguation. For more on instance extraction, we refer the reader to section 3.6.

From the statistical and machine learning side, there has been attempts to add new concepts to the ontology taxonomy. Updating the ontology with a new concept involves placing it correctly in the hierarchy and retrieving appropriate parents. A number of categorization techniques have been used to augment an ontology with a new concept: the *k-nearest neighbor method (kNN),* the *category-based method* and the *centroid-based method* (Maedche et al., 2002). These methods use vector-based features for representing concepts based on co-occurrence and word windows. The new concept can then be placed in the hierarchy according to similarity metrics with existing concepts in the ontology. Maedche et al. (2002) provide a good review about these methods.

6.6 Current Challenges

There are many challenges that face the ontology engineering as well as computer-based educational communities which consider using ontologies for domain knowledge representation. These can be divided into general and ITS-specific challenges.

Despite the large number of available systems, there is still room for more development in ontology engineering. More importantly, a reusable framework has to be set up to make combining and comparing different extraction methods easier. In fact, there is a lack of reusable services for ontology learning, updating and evaluation. There is also a lack of a comprehensive framework that highlights the available methods for each subtask of the ontology learning process based on various criteria (corpus, task, etc.). Such a framework would provide informed choices for ontology learning. From my point of view, a service-oriented architecture is essential for a broader development and reuse of automatic methods for ontology learning.

One other issue relating to automatic methods for ontology learning is that these methods can produce inconsistent or duplicate entries and dealing with such inconsistencies is particularly challenging (Volker et al., 2008). Inconsistencies can arise not only from the methods used, but also from the input data, which may be too sparse or may contain contradictions. Volker et al. (2008) propose three alternatives: using a reasoning-supported process to guarantee that the learned ontologies are kept consistent over time; repairing consistencies following the production of an ontology; or setting up reasoning mechanisms that can deal with these inconsistencies.

Adding to the list of challenges is the limited support regarding several key aspects of ontology engineering, especially ontology evolution, reuse, merging, alignment and matching. These different areas still need to mature. It is especially important to make available an environment entirely dedicated to ontology engineering involving all the different aspects of the ontology lifecycle.

The general challenges of ontology engineering impact the specific ITS-related challenge of building a domain model. In fact, successful attempts to build an ITS domain model automatically have been limited (Suraweera et al., 2004). Ontology engineering can help satisfy this need and contribute to the wider adoption of Intelligent Tutoring Systems. Moreover, ontology engineering can contribute to building a bridge with the eLearning community by making eLearning resources the main material for building the ITS domain model (Zouaq and Nkambou, 2009a). Similarly, eLearning can benefit from this domain model for indexing learning resources and developing more "intelligent" techniques for training learners.

6.7 Conclusion

We have described the field of domain ontology engineering, focusing on ontology learning techniques and highlighting how intelligent tutoring systems may benefit from this ontology engineering. One of the main advantages of this engineering is that it can provide a solution to two issues: first, the difficulty of building an ITS domain model from scratch for each domain and second, the difficulty of sharing and reusing the available representations. As standard knowledge representations, ontologies can support the ITS community in producing ITS components more easily and at lower costs. However, this involves the availability of a unified framework for the entire ontology lifecycle, including ontology learning, evolution, alignment, matching and evaluation.

References

Almuhareb, A., Poesio, M.: Attribute-based and value-based clustering: an evaluation. In: Proc. of EMNLP, Barcelona (July 2004)

Almuhareb, A., Poesio, M.: Finding Concept Attributes in the Web. In: Proc. of the Corpus Linguistics Conference, Birmingham (July 2005)

Amardeilh, F., Laublet, P., Minel, J.-L.: Document annotation and ontology population from linguistic extractions. In: Proc. of the 3rd international conference on Knowledge capture, Banff, Alberta, Canada, pp. 161–168 (2005)

Aroyo, L., Dicheva, D.: The New Challenges for E-learning: The Educational Semantic Web. Educational Technology & Society 7(4), 59–69 (2004)

Aussenac-Gilles, N., Biebow, B., Szulman, S.: Revisiting Ontology Design: A Methodology Based on Corpus Analysis. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 172–188. Springer, Heidelberg (2000)

Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. of the COLING-ACL, Montreal, Quebec, Canada (1998)

Bisson, G., Nedellec, C., Canamero, L.: Designing clustering methods for ontology building – The Mo'K workbench. In: Proc. of the ECAI Ontology Learning Workshop, pp. 13–19 (2000)

Buitelaar, P., Olejnik, D., Sintek, M.: A protege plug-in for ontology extraction from text based on linguistic analysis. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870. Springer, Heidelberg (2003)

Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: An Overview. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications Series, vol. 123. IOS Press, Amsterdam (July 2005)

Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 120–126 (1999)

Charniak, E., Berland, M.: Finding parts in very large corpora. In: Proc. of the 37th Annual Meeting of the ACL, pp. 57–64 (1999)

Cimiano, P.: Ontology Learning Attributes and Relations. In: Ontology Learning and Population from Text, pp. 185–231. Springer, Heidelberg (2006)

Cimiano, P., Wenderoth, J.: Automatic Acquisition of Ranked Qualia Structures from the Web. In: Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Prague (2007)

Cimiano, P., Wenderoth, J.: Automatically Learning Qualia Structures from the Web. In: Proc. of the ACL Workshop on Deep Lexical Acquisition, pp. 28–37 (2005)

Cimiano, P., Völker, J.: Text2Onto–A Framework for Ontology Learning and Data-driven Change Discovery. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 227–238. Springer, Heidelberg (2005a)

Cimiano, P., Volker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Proc. of RANLP 2005, Borovets, Bulgaria, pp. 166–172 (2005b)

Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2004)

Cimiano, P., Staab, S.: Learning by googling. ACM SIGKDD Explorations 6(2), 24–33 (2004)

Claveau, V.: Acquisition automatique de lexiques sémantiques pour la recherche d'information. Thèse de doctorat, Université de Rennes-1, Rennes (2003)

Corcho, O., Gómez-Pérez, A.: A Roadmap to Ontology Specification Languages. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 80–96. Springer, Heidelberg (2000)

Craven, M., Di Pasquo, D., Freitag, D., McCallum, A., Mitchell, T.M., Nigam, K., Slattery, S.: Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence 1-2(118), 69–113 (2000)

Deiltel, A., Faron-Zucker, C., Dieng, R.: Learning ontologies from rdf annotations. In: Proc. of the IJCAI Workshop in Ontology Learning (2001)

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in KnowItAll (preliminary results). In: Proc. of the 13th World Wide Web Conference, pp. 100–109 (2004)

Faure, D., Nedellec, C.: A corpus-based conceptual clustering method for verb frames and ontology. In: Velardi, P. (ed.) Proc. of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications, pp. 5–12 (1998)

Frantzi, K., Ananiadou, S., Tsuji, J.: The c-value/nc-value method of automatic recognition for multi-word terms. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 585–604. Springer, Heidelberg (1998)

Fleischman, M.: Automated Subcategorization of Named Entities. In: 39th Annual Meeting of the ACL. In: Student Research Workshop, Toulouse, France (July 2001)

Fleischman, M., Hovy, E.H.: Fine Grained Classification of Named Entities. In: COLING 2002 (2002)

Flouris, G., Plexousakis, D., Antoniou, G.: Evolving Ontology Evolution. In: Proc. of the 32nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM-06), pp. 14–29 (2006) (invited Talk)

Fortuna, B., Grobelnik, M., Mladenic, D.: OntoGen: Semi-automatic Ontology Editor. In: HCI International 2007, Beijing (2007)

Fortuna, B., Mladevic, D., Grobelnik, M.: Visualization of Text Document Corpus. In: ACAI 2005 Summer School (2005)

Automatic Discovery of Part–Whole Relations. In: Proc. of the, Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, pp. 1–8. Association for Computational Linguistics, Morristown (2003)

Guarino, N.: Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge base. Data and Knowledge Engineering 8, 249–261 (1992)

Guarino, N., Welty, C.: Evaluating Ontological Decisions with OntoClean. Communications of the ACM 45(2), 61–65 (2002)

Haase, P., Sure, Y.: State-of-the-Art on Ontology Evolution, SEKT deliverable 3.1.1.b (2004), http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.b.pdf

Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)

Hearst, M.: Automatic Acquisition of Hyponyms from LargeText Corpora. In: Proc. of the Fourteenth International Conference on Computational Linguistics, Nantes, pp. 539–545 (1992)

Hearst, M., Schutze, H.: Customizing a lexicon to better suit a computational task. In: ACL SIGLEX Workshop, Columbus, Ohio (1993)

Hindle, D.: Noun classification from predicate-argument structures. In: Proc. of ACL 1990, Pittsburg, Pennsylvania, pp. 268–275 (1990)

Iwanska, L.M., Mata, N., Kruger, K.: Fully automatic acquisition of taxonomic knowledge from large corpora of texts. In: Natural Language Processing and Knowledge Processing, pp. 335–345. MIT/AAAI Press (2000)

Jannink, J., Wiederhold, G.: Ontology maintenance with an algebraic methodology: A case study. In: Proc. of AAAI workshop on Ontology Management (1999)

Kipper, K., Dang, H.D., Palmer, M.: Class-Based Construction of a Verb Lexicon. In: Proc. of AAAI-2000 Seventeenth National Conference on Artificial Intelligence, pp. 691–696 (2000)

Lin, D., Pantel, P.: Induction of semantic classes from natural language text. In: Proc. of SIGKDD 2001, San Francisco, CA, pp. 317–322 (2001)

Lin, D.: Automatic identification of non-compositional phrases. In: Proc. of ACL 1999, pp. 317–324 (1999)

Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proc. of COLING-ACL 1998, Montreal, Canada, pp. 768–774 (1998)

Maedche, A., Pekar, V., Staab, S.: Ontology Learning Part One—On Discovering Taxonomic Relations from the Web. Web Intelligence, pp. 301–322. Springer, Heidelberg (2002)

Maedche, A., Staab, S.: Ontology Learning for the Semantic Web. IEEE Intelligent Systems 16(2), 72–79 (2001)

Moldovan, D.I., Girju, R.C.: An interactive tool for the rapid development of knowledge bases. International Journal on Artificial Intelligence Tools (IJAIT) 10(1-2) (2001)

Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. In: Proc. of the 20th international conference on Computational Linguistics, Switzerland (2004)

Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. Computational Linguistics 30(2), 151–179 (2004)

Nyulas, C., O'Connor, M., Tu, S.: Datamaster–a plug-in for importing schemas and data from relational databases into protégé. In: Proc. of 10th Intl. Protégé Conference, Budapest (2007)

Papatheodorou, C., Vassiliou, A., Simon, B.: Discovery of Ontologies for Learning Resources Using Word-based Clustering. In: Proc. of ED-MEDIA 2002, AACE, Denver, USA (2002)

Poesio, M., Almuhareb, A.: Identifying Concept Attributes Using A Classifier. In: Proc. of the ACL Workshop on Deep Lexical Acquisition, pp. 18–27. Association for Computational Linguistics, Ann Arbor (2005)

Poesio, M., Almuhareb, A.: Extracting concept descriptions from the Web: The importance of attributes and values. In: Buitelaar, P., Cimiano, P. (eds.) Bridging the Gap between Text and Knowledge, pp. 29–44. IOS Press, Amsterdam (2008)

Polson, M.C., Richardson, J.J. (eds.): Foundations of Intelligent Tutoring Systems. L. Erlbaum Associates Inc., Mahwah (1988)

Pustejovsky, J.: The generative lexicon. MIT Press, Cambridge (1995)

Ravi, S., Pasca, M.: Using Structured Text for Large-Scale Attribute Extraction. In: Proc. of the 17th ACM Conference on Information and Knowledge Management, CIKM-2008 (2008)

Reinberger, M.-L., Spyns, P.: Unsupervised Text Mining for the learning of DOGMA-inspired Ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Applications and Evaluation. Advances in Artificial Intelligence, pp. 29–43. IOS Press, Amsterdam (2005)

Rigau, G., Rodríguez, H., Agirre, E.: Building Accurate Semantic Taxonomies from Monolingual MRDs. In: Proc. of the 17th International Conference on Computational Linguistics COLING-ACL 1998, Montreal, Quebec, Canada (1998)

Sabou, M.: Learning Web Service Ontologies: an Automatic Extraction Method and its Evaluation. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005)

Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. dissertation, Department of Linguistics, Stockholm University (2006)

Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 515–523 (1988)

Schlobach, S., Olsthoorn, M., de Rijke, M.: Type Checking in Open-Domain Question Answering. In: Proc. of European Conference on Artificial Intelligence, pp. 398–402. IOS Press, Amsterdam (2004)

Shamsfard, M., Barforoush, A.A.: The State of the Art in Ontology Learning: A Framework for Comparison. The Knowledge Engineering Review 18(4), 293–316 (2003)

Shamsfard, M., Barforoush, A.A.: An Introduction to HASTI: An Ontology Learning System. In: Proc. of 6th Conference on Artificial Intelligence and Soft Computing (ASC 2002), Banff, Alberta, Canada (2002)

Shvaiko, P., Euzenat, J.: Ten Challenges for Ontology Matching. In: Meersman, R., Tari, Z. (eds.) OTM 2008, Part I. LNCS, vol. 5331, pp. 1164–1182. Springer, Heidelberg (2008)

Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems (NIPS 2004), Vancouver, British Columbia (2004)

Stojanovic, L., Stojanovic, N., Volz, R.: Migrating data-intensive Web Sites into the Semantic Web. In: Proc. of the 17th ACM symposium on applied computing (SAC), pp. 1100–1107. ACM Press, New York (2002)

Suraweera, P., Mitrovic, A., Martin, B.: The role of domain ontology in knowledge acquisition for ITSs. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 207–216. Springer, Heidelberg (2004)

Suryanto, H., Compton, P.: Discovery of Ontologies from Knowledge Bases. In: Gil, Y., Musen, M., Shavlik, J. (eds.) Proc. of the First International Conference on Knowledge Capture, Victoria, British Columbia Canada, pp. 171–178. The Association for Computing Machinery, New York (2001)

Tanev, H., Magnini, B.: Weakly Supervised Approaches for Ontology Population. In: Proc. of EACL 2006, Trento, Italy, pp. 3–7 (2006)

Van Assem, M., Menken, M.R., Schreiber, G., Wielemaker, J., Wielinga, B.J.: A Method for Converting Thesauri to RDF/OWL. In: International Semantic Web Conference, pp. 17–31 (2004)

Van Hage, W.R., Kolb, H., Schreiber, G.: A Method for Learning-Part-Whole Relations. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 723–735. Springer, Heidelberg (2006)

Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of ontolearn, a methodology for automatic population of domain ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press, Amsterdam (2005)

Volker, J., Haase, P., Hitzler, P.: Learning Expressive Ontologies. In: Buitelaar, P., Cimiano, P. (eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Frontiers in Artificial Intelligence and Applications, vol. 167, pp. 45–69. IOS Press, Amsterdam (2008)

Volz, R., Oberle, D., Staab, S., Studer, R.: OntoLiFT Prototype. IST Project 2001-33052 WonderWeb Deliverable 11 (2003)

Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In: Proc. of HLT-NAACL, pp. 197–204 (2003)

Widdows, D., Dorow, B.: A graph model for unsupervised lexical acquisition. In: 19th International Conference on Computational Linguistics, Taipei, Taiwan, pp. 1093–1099 (2002)

Woods, W.A.: What's in a link: Foundations for semantic networks. In: Bobrow, D.G., Collins, A.M. (eds.) Representation and Understanding: Studies in Cognitive Science, pp. 35–82. Academic Press, New York (1975)

Yamada, I., Baldwin, T.: Automatic discovery of telic and agentive roles from corpus data. In: Proc. of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC), Tokyo, pp. 115–126 (2004)

Zouaq, A., Nkambou, R.: Enhancing Learning Objects with an Ontology-Based Memory. IEEE Transactions on Knowledge and Data Engineering 21(6), 881–893 (2009a)

Zouaq, A., Nkambou, R.: Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project. IEEE Transactions on Knowledge and Data Engineering 21(11), 1559–1572 (2009b)