Semi-automated ontology learning: the BOEMIE approach

Petasis George[1], Karkaletsis Vangelis[1], Krithara Anastasia[1], Paliouras Georgios[1], Spyropoulos D. Constantine[1]

[1] National Center of Scientific Research, NCSR Demokritos,
Athens, Greece
{petasis, vangelis, akrithara, paliourg,costass}@iit.demokritos.gr

Abstract. In this paper we describe a semi-automated approach for ontology learning. Exploiting an ontology-based multimodal information extraction system, the ontology learning subsystem accumulates documents that are insufficiently analysed and through clustering proposes new concepts, relations and interpretation rules to be added to the ontology.

Keywords: Ontology learning, Ontology population, Ontology enrichment

1 Introduction

In recent years, ontologies have become extremely popular as a means for representing machine-readable knowledge. Driven mainly by the movement towards a semantic web, which tries to attach machine-readable semantic information to all of the resources found in the "traditional" web, ontologies address the issue of representation for this semantic information. Having the infrastructure for representing knowledge, one can construct an ontology and semantically annotate all the desired web resources. Realizing the difficulty of designing the grand ontology of the world, research on the semantic web has focused on the development of domain or task-specific ontologies, which have started making their appearance in fairly large numbers. Three major approaches have been presented for acquiring domain or task-specific ontologies:

- By integrating existing ontologies.
- By constructing an ontology from scratch, or by extending (populating and enriching) an existing ontology, usually based on information extracted from a domain.
- By specialising a generic ontology, in order to adapt it to a specific domain.

Acquiring domain knowledge for constructing ontologies is a resource-demanding and time-consuming task. Thus, the automated or semi-automated construction, enrichment and adaptation of ontologies, is highly desired. The process of automated or semi-automated construction, enrichment and adaptation of ontologies is known as ontology learning. Ontology learning can be decomposed into six major subtasks:

- *Term identification*: Terms are the "symbols" that represent ontological concepts and relations, "lexicalising" them into objects of the real word.

- *Synonym identification*: Synonyms are sets of terms, representing the same real object or event.
- *Concept identification*: Concepts are the basic building blocks of an ontology, as they constitute the primitive elements through which the semantic model of the ontology is constructed.
- *Taxonomic relation identification*: Relations are semantic associations holding between two ontological concepts. Taxonomic relations organise concepts into a taxonomy.
- *Non-taxonomic relation identification*.
- *Rule acquisition*: Rules formalise constraints over the concepts and relations of an ontology.

The work in this paper presents an approach for learning an ontology by extending an existing one with the use of information extracted from a thematic domain. Assuming an ontology-driven content analysis system able to extract domain-specific information and semantically represent it in the form of an ontology, the proposed approach identifies possible additions to the ontology, in the form of new concepts, relations and rules. The strength of this approach is that it can identify and propose possible concepts and relations even when no initial ontology exists. In other words, it can be used to construct an ontology from scratch.

This paper is organised as follows: in section 2 the semi-automated approach for ontology learning is presented. The co-operation of the learning approach with an extraction system is discussed, and the organisation of a suitable system is briefly presented. Then, a more detailed approach is presented focusing on the task of learning new concepts. Finally, section 3 presents the results of an evaluation of the proposed approach with the following sections presenting some alternative approaches and concluding the document.

2 Ontology learning: a semi-automated approach

Assuming the existence of an ontology-driven information extraction system, the proposed learning approach is able to identify new concepts, relations and rules as candidates for extending the existing ontology that drives the information extraction system. These candidate additions are formed through similarity-based clustering of the extracted information. The candidate additions are shown to a domain expert through a suitable interface, and the expert is responsible for accepting, revising or rejecting them after examining the supporting evidence.

The proposed approach operates on the results of an ontology-driven information extraction engine, which must meet some requirements:

- To represent extracted information in terms of ontological concept instances.
- To extract relations between instances and represent them as instances of ontological relations.

In addition, an optional but useful characteristic is the adaptability of the information extraction engine to changes in the ontology. If the ontology is enriched with a new concept/relation/rule and the extraction engine can adapt to the change and extract instances of the new concept/relation, then the proposed learning approach will attempt to improve further the ontology, following for example a bootstrapping approach: the extracted information from the information extraction engine can is used to evolve the ontology, and through the evolved ontology the extraction of information is improved. The boostrapping process can continue until no more information can be extracted from the corpus.

2.1 Semantic Extraction

Such an extraction engine has been developed in the BOEMIE project. This engine implements a modular approach (Petridis et al. 2006) that comprises the following three level of abstraction:

- The low-level analysis, which includes a set of modality-specific (image, text, video, audio) content analysis tools.
- A modality-specific semantic interpretation engine.
- A fusion engine, which combines interpretations for each modality.

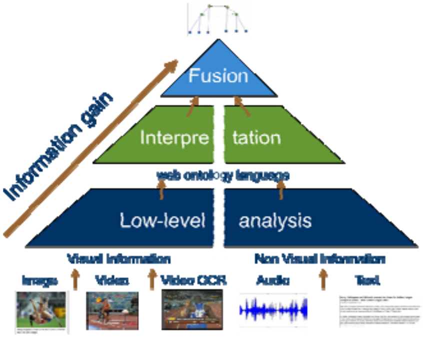

Figure 1: Semantic extraction from multimedia content.

The first two levels implement ontology-driven, modality-specific information extraction, while the last one fuses the information obtained from the previous levels of analysis.

Within BOEMIE, modality-specific analysis (levels 1 and 2) is performed as a two step process: the first step involves the identification of “primitive” concepts, as well as instances of relations between them. “Primitive” concepts, known as “mid-level” concepts (MLCs) in BOEMIE, are concepts whose instances can be directly identified in corpora of a specific modality. For example, in the textual modality the *name* or the *age* of a person is a mid-level or “primitive” concept, as instances of the concepts are associated directly with relevant text segments. The second one semantically interprets (with the help of reasoning rules (Espinosa et al., 2008)) the “primitive” concepts of each modality in combination, in order to create instances of “Composite” concepts that explain the event described in the document. “Composite” concepts are referred to as high-level concepts (HLCs) within the BOEMIE project, and cannot be directly identified in the content and thus associated with a content segment. For example, the concept *person* is a “composite” one, i.e., a concept that is defined as a composition of other concepts, such as person name, age, gender, etc. The fact that content analysis is separated from semantic interpretation, along with the fact that semantic interpretation is performed through reasoning using rules from the ontology, allows single-modality extraction to be adaptable to changes in the ontology.

Once a multimedia document has been decomposed into single-modality elements and each element has been analysed and semantically interpreted separately, the various interpretations must be fused into one or more alternative explanations of the multimedia document as a whole. This process is also performed through reasoning, using rules of the domain ontology.

This extracted information is given to the ontology learning module in the form of OWL ABoxes. These ABoxes typically include instances of MLCs, relations between them, HLCs, relations between them and possibly instances of the specific MLC “unknown” (when the low-level analysis could not classify a certain object).

According to the information contained in the ABoxes, the ontology learning system triggers either the ontology population or the ontology enrichment activity. The former refers to the activity of adding new individuals into the ontology. The latter refers to the extension of the ontology through the addition of new concepts and relations. The system identifies four different *evolution patterns*, each of them determining the characteristics of the ABox and defining the process to be performed over the ontology (figure 2). In this paper, the focus is on the ontology enrichment task. More information about the whole ontology evolution activity can be found in (Castano et al. 2008).

2.2 Ontology enrichment

In the case where the background knowledge is not sufficient to explain the extracted information from the processed documents, ontology enrichment is performed.

As explained above, ontology concepts in BOEMIE are separated into two main types, according to whether they can be identified directly in corpora (MLCs), or inferred through reasoning (HLCs). The BOEMIE ontology learning approach aims to discover both types of concepts. However, the discovery of MLCs is by its nature a modality-specific process, with significant dependencies upon the underlying information extraction system. Although the learning process is not very different from that of HLCs, it poses additional requirements to the extraction system and thus it cannot be presented in isolation from it. For this reason emphasis is given here to the discovery of high-level concepts.

The discovery of HLCs can be decomposed into the following tasks:

- Concept Learning: It includes two subtasks: *Object clustering*, which identifies similarities between unclassified objects and clusters them into potentially new concepts and relations. *Concept formation*, which examines the clustered elements in order to extract common information, such as concepts, properties and relations, and use this to form a new *concept.*
- Concept enhancement: This task is responsible for improving a candidate concept, through knowledge acquired from external knowledge sources, such as external domain ontologies or taxonomies.

- Concept definition: This task presents the definitions of the candidate concepts and relations to the ontology expert. An ontology expert is expected to revise and finally approve them in order to be assimilated into the ontology.
- Concept validation: This task performs consistency checking, by trying to detect possible inconsistencies caused by the addition of a new concept or relation to the ontology.
- Concept assimilation: The last task is responsible for performing the required changes to the ontology, in order to incorporate the newly formed concept/relation into the ontology.

From the above staged process the first two tasks constitute the core of concept learning and the focus of this paper.

We have identified and addressed three concept learning scenarios, according to the dependency of learned concepts to existing ones:

Specialising an existing concept. This scenario corresponds to the discovery of concepts that are specializations of existing ones. For example, assuming the domain of athletics and the existence of a concept covering all *jumping* sports in the domain ontology, this scenario examines whether a more specific concept describing a single sport (such as *high jump* or *pole vault*) can be learned.

Generalising a set of existing concepts. This scenario corresponds to the discovery of concepts that generalize existing ones. Continuing on the above example, if concepts representing specific sports like *pole vault* and *high jump* already exist in the ontology, this scenario examines whether similarities of these two sports can be captured and a more abstract concept such as *jumping* can be learned.

Learning a concept similar to an existing one. This scenario corresponds to the discovery of concepts that are similar to, but different from existing ones. An example of this learning scenario is to learn a concept for *high jump,* if a concept representing the sport of *pole vault* is already known. The difference of this scenario from the previous ones is that no assumption is made that the learned concepts have any hierarchical relations, i.e., subsume or subsumed by, existing ones. Usually the two similar concepts are sibling under the same parent concept, which may even not exist yet and need to be discovered.

2.3 Methodology

The above three scenarios are modeled after three basic tree operators, namely node splitting (for specializing an existing concept), node merging (for generalizing a set of existing concepts) and creation of new nodes (for learning a concept similar to an existing one). Each scenario identified as applicable may propose a set of possible proposals, which the domain expert is expected to filter in order to guide ontology evolution.

The first scenario (specializing an existing concept) is implemented using two operators: specialize by contextual evidence, and specialize by property evidence. The former examines cases that an HLC can be separated by the presence of another concept. For instance, sports involving the use of some equipment (i.e. throwing related sports). The implementation of this operator is based on co-occurrence statistics: examining HLCs with strong co-occurrence information with non-aggregated MLCs. The second operator examines cases where an HLC can be specialized due to its property/properties. For instance, athletes can be specialized by gender. This operator is implemented based on clustering property values, using standard clustering algorithms, on property values that have a textual representation (i.e. originate from the text modality).

The second scenario (generalizing a set of existing concepts) performs two operations: the identification of concepts that are "similar" and the identification of common properties from similar concepts. The first operation measures similarity between concepts, where the similarity is defined as the mean value of similarities among all individuals of the compared concepts. Concepts whose similarity is above a certain threshold (configurable by the domain expert with a default value of 0.8) are forwarded to the second operation, where the Least Common Subsumer is calculated and proposed as a possible abstraction to the domain expert.

The third scenario (learning a concept similar to an existing one) is implemented by two operators: extend by contextual evidence and reduce by property evidence. The first operator may propose to add a property and extend the original concept. Suppose for example, that the ontology contains an HLC that denotes a person, which does not have properties related to nationality, gender or age. Provided that there are instances of gender, age and nationality extracted by the semantics extraction toolkit, this operator may propose to add such a property and extend the original concept. This is implemented based on co-occurrence statistics: HLCs with strong co-occurrence information with non-aggregated HLCs/MLCs are examined as possible additions to any HLC. The second operator examines whether a property can be removed from an existing concept definition. An example of the latter is the case that an ontology contains an HLC that describes a sport like pole vault, but does not contain an HLC for high jump. The HLC representing pole vault may suggest the presence of a horizontal bar and a pole to completely represent the sport. However, instances representing high jump events are not expected to be associated with a pole, a situation that can be identified by this operator. The implementation of this operation is based on statistical information: if more than 30% of instances do not have a value for a specific property, this property is proposed for removal from the HLC under examination.

3 Evaluation

For the purposes of evaluation, a single evaluation scenario has been constructed which mainly concentrates on the second learning scenario, i.e., evaluating the ability to learn a generalised concept from a set of inter-related existing ones. According to this evaluation scenario an initial ontology is constructed manually. Then the goal of learning is to reconstruct the basic organisation, i.e., the most "important" concepts, of this ontology, from manually annotated corpora. The choice of this evaluation scenario is based on data availability. Annotations about a single sport only are needed in order to assess the performance of the system in this scenario, in contrast with the other two, where a larger manually annotated corpus involving two or more sports is needed.

3.1 Experimental setting

In order to perform the evaluation two resources are required: an ontology and corpus annotated with this ontology. For the purposes of this experiment, a small ontology has been manually constructed, covering the domain of athletics. This ontology contains only mid-level concepts and relations between MLCs. The concepts and relations contained in the ontology are shown in table 1[1].

As shown in the table, the ontology contains mid-level concepts acting as properties for four high-level concepts that are not present in the ontology: *athlete*, *sport*, *sport round* and *event*. In addition to these concepts, the ontology contains an extensive set of relations between these mid-level concepts. These relations connect concepts belonging to the same (absent) HLC, such as a relation between the name of an athlete and its age or performance, but also connect MLCs that "belong" to a different HLC, such as a relation between the name of an athlete and the name of an event.

Having constructed an initial ontology, the ontology was used to manually annotate a textual corpus, simulating the results of an ontology-based information extraction process driven by the ontology.

All instances of concepts found in the corpus and all possible relations between these instances were annotated, leading to the creation of an OWL ABox for each document in the corpus. The corpus contained 250 HTML web pages, from various sites belonging to official associations, like IAAF[2], EAA[3] and USATF[4]. The thematic domain of the collected corpus refers to the sport of high jump.

3.2 Clustering

The evaluation scenario seeks to reconstruct the ontology by learning new HLCs like *athlete*, *sport* or *event*, having as starting point instances of the mid-level concepts and the relations between them shown in table 1. This task is accomplished by exploiting similarities among instances, through clustering. Furthermore, in the concept enhancement step, the clustering results are filtered in order to form a proposal that an ontology expert should examine, possibly revise and finally approve.

Clustering is a form of machine learning and as such two questions arise when it is used: which clustering algorithm should be used, and how the data to be clustered can be more effectively represented in order to obtain the desired results. Regarding the clustering algorithm the Expectation Maximisation (EM) clustering algorithm (Ian et al. 2005), an generalization of *k*-means (Hartigan, 1975), was selected. Given a fixed

[1] Concepts are organised into "abstract" categories for presentation reasons, which causes some concepts to appear more than once in the table.

[2] International Association of Athletics Federations – http://www.iaaf.org/.

[3] European Athletics Association – http://www.european-athletics.org/.

[4] USA Track and Field – http://www.usatf.org/.

number *k* of clusters (desired or hypothesized), the *k*-means algorithm assign observations to those clusters so that the means across clusters are as different from each other as possible. The EM algorithm is a generalization of *k*-means that allows for soft clusters to be formed, by associating a probability to each data point to belong to each of the clusters, based on one or more probability distributions. EM also tries to maximize the overall likelihood of the data, given the (final) clusters. EM can be easily combined with *n*-fold cross validation in order to estimate an "optimal" number of clusters from the data and several implementations include this feature, among which the WEKA (Ian et al. 2005) implementation that has been used in this experiment.

Concepts	Relations
PersonName **Age** **Gender** **CountryName**	personNameToAge personNameToCountryName personNameToGender *personNameToPerformance, personNameToRanking*
Performance **Ranking**	performanceToRanking
SportsRoundName **Date**	sportsRoundNameToDate *sportsRoundNameToPerformance, sportsRoundNameToRanking* *sportsRoundNameToPersonName*
SportsName **Date** **CityName** **StadiumName**	sportsNameToCity sportsNameToStadiumName sportsNameToDate *sportsNameToPersonName, sportsNameToSportsRoundName* *sportsNameToPerformance, sportsNameToRanking*
SportsEventName **Date** **CityName** **CountryName**	sportsEventNameToCityName sportsEventNameToCountryName sportsEventNameToDate *sportsEventNameToSportsName, sportsEventNameToPersonName*

The representation scheme selected as input to clustering relies on the concepts and relations of each instance. A single feature vector is created for each concept instance found in the corpus, containing the following features:

- The concept of the instance.
- For each binary relation that has as subject the instance, the number of times this instance is related to other instances with this relation type.

Thus, the representation scheme contains only the concept of an instance and the relations the instance participates in as subject of the relation.

The desired result of the learning process is a concept proposal, which is created by filtering the clustering results. The desired characteristics of such a concept proposal include:

- The set of concepts the new concept "combines".
- The set of "internal" relations that are used to relate the concepts that are "combined" by the new concept.
- Possibly, a set of "external" relations that relate this new concept with other "composite" – HLCs.

The filtering has been implemented as a two phase process. During the first phase, an initial concept proposal is formed by collecting all concepts and binary relations found in all vectors of a cluster. This essentially collects the concept of all individuals in the cluster, along with all relations where each individual acts as a subject. Still within the first phase, all gathered relations are examined, and all concepts that appear as a subject of a relation are added to the concept proposal. By the end of the first phase, the concept proposal contains all concepts the proposed concept groups, and the set of relations between them. Furthermore, additional relations of the proposed concept with other proposed concepts may exist. The task of the second phase is to discover and eliminate such cases. This is achieved by examining all formed concept propositions to identify concepts that participate in more than one proposal. These concepts are eliminated from all proposals that do not contain a relation where this concept appears as a subject. In case a concept has been removed from a proposal, all relations that employ this concept as subject are marked as "external" ones. Finally, concept proposals that are subsumed by another concept proposal are eliminated.

3.3 Evaluation Results

All documents in the corpus were manually annotated with instances of concepts from the ontology, resulting in an OWL ABox for each document. Then, each ABox was processed creating a feature vector for each instance in the ABox. Vectors generated from all ABoxes were merged into a single training corpus, which has been processed by the EM algorithm. 10-fold cross validation was employed to the whole training corpus to identify the number of clusters that exist in the training data. The clustering results over the training data were filtered, in order to form the proposals of the new concepts, which were manually examined by an ontology expert. The concept proposals shown in table 2 were considered correct. The same experiment was conducted with various numbers of documents constituting the training corpus, while WEKA 3.5 implementations of both the EM algorithm and 10-fold cross validation were used for performing the experiments. The obtained results are shown in table 3.

Proposed Concept	**Combined Concepts**	**"Internal" Relations**	**"External" Relations**
Athlete	*PersonName, Age, Gender, CountryName*	*personNameToAge, personNameToCountryName, personNameTo\Gender*	*personNameToPerformance* *personNameToRanking*
Performance	*Performance, Ranking*	*performanceToRanking*	
Sport Round	*SportsRoundName, Date*	*sportsRoundNameToDate*	*sportsRoundNameToPerformance* *sportsRoundNameToRanking* *sportsRoundNameToPersonName*
Sport	*SportsName, Date, CityName, StadiumName*	*sportsNameToCity, sportsNameToStadiumName, sportsNameToDate*	*sportsNameToPersonName* *sportsNameToSportsRoundName* *sportsNameToPerformance* *sportsNameToRanking*
Event	*SportsEventName, Date, CityName,*	*sportsEventNameToCityName,*	*sportsEventNameToSportsName* *sportsEventNameToPersonName*

	CountryName	*sportsEventNameToCountryName,* *sportsEventNameToDate*

Table 1: The concept proposals considered as correct.

Corpus Size	# Proposed Concepts	Proposed Concepts	Correct proposals	% correct proposals
5	3	(Event+Sport Round), (Athlete+Performance), (Sport + Performance)	0	0%
10	5	(Athlete), (Performance), (Sport Round), (Sport), (Event)	5 (all)	100%
15	5	(Athlete), (Performance), (Sport Round), (Sport), (Event)	5 (all)	100%
25	5	(Athlete), (Performance), (Sport Round), (Sport), (Event)	5 (all)	100%
50	5	(Athlete), (Performance), (Sport Round), (Sport), (Event)	5 (all)	100%
80	3	(Event),(Athlete + Performance), (Sport + Sport Round + Performance)	1 (Event)	33%
100	4	(Event), (Athlete+Performance), (Sport + Performance), (Athlete+Sport+Sport Round+Performance)	1 (Event)	25%
175	5	(Event), (Athlete+Performance), (Sport + Performance), (Sport Round+Performance), (Athlete+Sport+Performance)	1 (Event)	20%
250	4	(Event), (Athlete+Performance), (Sport + Performance), (Athlete+Sport Round+Performance)	1 (Event)	25%

Table 2: The performance of the concept learning approach for variable training corpus size

The evaluation results clearly show that the proposed learning approach is able to make reasonable proposals for new concepts. The concept learning approach was particularly successful in relatively small corpora, and its proposals remained reasonable in larger corpora. When the training corpus contains more than 50 documents the *Performance* concept cannot be recognised easily and interferes with the recognition of the other concepts. For example, training with 100 documents learns the concept Athlete and Performance combined into a single concept, denoted as (*Athlete + Performance)* in table 3. The reason for this failure is two fold: the small size of the concept (it has only two MLCs and a single relation) and the limited occurrences of the single relation in the corpus. The number of documents that contain instances of the *performanceToRanking* relation, meaning that both performance and ranking are described in the document, is about 20% of the corpus.

When training with large corpora, the number of instances of this relation is much lower than the number of instances of the other six **ToRanking* and **ToPerformance* relations, misleading the clustering algorithm to attach the *Performance* and *Ranking* concepts to the concepts that refer to them more frequently through relations, like the *athlete, sport* and *sportRound* concepts.

Thus, the inability of the proposed approach to make accurate proposals can be attributed mostly to data sparseness and imbalanced distribution of the concepts and their relations.

4 Related Work

The recent success of distributed and dynamic infrastructures for knowledge sharing has increased the need of semi-automated or automated ontology evolution strategies (Haase and Sure, 2004; Klein and Noy, 2003). Overviews of some proposed approaches in this direction are presented in (Ding and Foo, 2002; Gómez-Pérez et al., 2004), even if limited concrete results have appeared in the literature. In most recent work, formal and logic-based approaches to ontology evolution are being proposed.

In (Haase and Stojanovic, 2005), the authors provide a formal model for handling the semantics of change phase embedded in the evolution process of an OWL ontology. The proposed formalization allows to define and preserve arbitrary consistency conditions (i.e., structural, logical, and user-defined). A six-phase evolution methodology has been implemented within the KAON (Oberle et al., 2004) infrastructure for business-oriented ontology management. It contains an algorithmic library that supports clustering, classification and other techniques. ASIUM (Faure et al., 1998) employs hierarchical clustering in order to learn concept hierarchies. TEXT-TO-ONTO (Maedche and Staab, 2001) uses a multi-strategy method which combines association rules, formal concept analysis and clustering. Also, HASTI (Shamsfard, 2003) learns using a combination of logical reasoning, linguistic analysis and heuristic methods.

The advantage of BOEMIE approach over the above methods is that it operates solely on the results of information extraction that have been augmented with the results of reasoning, as explained in the previous sections. For example, the method can construct an ontology for a different domain, given the extracted entities and relations of this new domain by the extraction engine. Also, the distinction made between “primitive” and “composite” concepts helps the information extraction process becoming more independent of the ontology structure.

5 Conclusions

In this paper an approach for ontology learning has been presented that is able to perform semi-automated learning of concepts. Relying on an ontology-based information extraction system, the proposed approach exploits similarities obtained through clustering from extracted data, to propose possible ontology extensions to a domain expert. Three learning scenarios have been briefly presented that can be supported by the proposed approach, one of which has been evaluated. Evaluation results have shown that the implemented system is able to perform reasonable suggestions to the ontology expert.

Acknowledges

The BOEMIE Project is funded by the 6th EU Framework Programme, under contract FP6-027538.

References

(Castano et al. 2008): S. Castano, I.S.E. Peraldi, A. Ferrara, V. Karkaletsis, A. Kaya, R. Möller, S. Montanelli, G. Petasis, and M. Wessel. Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation, September 2008.

(Ding and Foo, 2002): Ding, Y., Foo, S.: Ontology research and development. part 2 - a review of ontology mapping and evolving. Journal of Information Science 28 (2002) 375–388

(Espinosa et al., 2008): S. Espinosa Peraldi, A. Kaya, S. Melzer, and R. Möller. On Ontology Based Abduction for Text Interpretation. In A. Gelbukh, editor, *Proc. of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, number 4919 in LNCS, pages 194–205. Springer, 2008

(Faure et al., 1998): Faure, D., Nedellec C., and Rouveirol, C., Acquisition of Semantic Knowledge using Machine Learning Methods: The System ASIUM, Technical Report number ICS-TR-88-16, Laboratoire de Recherche en Informatique, Inference and Learning Group, Universite Paris Sud, 1998.

(Gómez-Pérez et al., 2004) : A. Gómez-Pérez, M. Fernandez-Lopez, and O. Corcho. Ontological Engineering. Advanced Information and Knowledge Processing. Springer-Verlag, 2004.

(Haase and Sure, 2004): Haase, P., Sure, Y. State-of-the-art on ontology evolution. SEKT informal deliverable 3.1.1.b, Institute AIFB, University of Karlsruhe (2004)

(Haase and Stojanovic, 2005): Haase, P., Stojanovic, L.: Consistent evolution of owl ontologies. In Gómez-Pérez, A., Euzenat, J., eds.: ESWC. Volume 3532 of Lecture Notes in Computer Science., Springer (2005) 182–197

(Hartigan, 1975): Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley.

(Ian et al. 2005): Ian H. Witten and Eibe Frank: "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

(Klein and Noy, 2003): Klein, M., Noy, N.A component -based framework for ontology evolution (2003) (Kim et al., 2002): Kim, S., Alani, H., Hall, W., Lewis, P., Millard, D., Shadbolt, N., Weal, M.: Artequakt: Generating tailored biographies from automatically annotated fragments from the web. In: Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM02), the 15th European Conference on Artificial Intelligence, (ECAI02), Lyon, France (2002) 1–6

(Maedche and Staab, 2001): A. Maedche, and S. Staab, Ontology learning for the Semantic Web, IEEE journal on Intelligent Systems, Vol. 16, No. 2, 72-79, 2001.

(Oberle et al., 2004): Oberle, D., Volz, R., Motik, B., Staab, S.: An extensible ontology software environment. In Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer (2004) 311–333

(Petridis et al. 2006): Petridis S., Tsapatsoulis N., Kosmopoulos D., Pratikakis Y., Gatos V., Perantonis S., Petasis G., Fragou P., Karkaletsis V., Biatov K., Seibert C., Espinosa S., Melzer S., Kaya A., Möller R.: "D2.1 Methodology for Semantics Extraction from Multimedia Content", BOEMIE Project Deliverable, 2006. Available from http://www.boemie.org.

(Shamsfard, 2003): M. Shamsfard, Designing the ontology learning Model, Prototyping in a Persian Text Understanding System, Ph.D. Dissertation, Computer Engineering Dept., AmirKabir University of Technology, Tehran, Iran, Jan. 2003.