Ontology Population and Enrichment: State of the Art

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras,
Anastasia Krithara, and Elias Zavitsanos

Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos",
15310, Ag. Paraskevi, Attiki, Greece
{petasis,vangelis,paliourg,akrithara,izavits}@iit.demokritos.gr

Abstract. Ontology learning is the process of acquiring (constructing or integrating) an ontology (semi-) automatically. Being a knowledge acquisition task, it is a complex activity, which becomes even more complex in the context of the BOEMIE project[1], due to the management of multimedia resources and the multi-modal semantic interpretation that they require. The purpose of this chapter is to present a survey of the most relevant methods, techniques and tools used for the task of ontology learning. Adopting a practical perspective, an overview of the main activities involved in ontology learning is presented. This breakdown of the learning process is used as a basis for the comparative analysis of existing tools and approaches. The comparison is done along dimensions that emphasize the particular interests of the BOEMIE project. In this context, ontology learning in BOEMIE is treated and compared to the state of the art, explaining how BOEMIE addresses problems observed in existing systems and contributes to issues that are not frequently considered by existing approaches.

Keywords: Ontology learning, Ontology population, Ontology enrichment.

1 Introduction

In recent years, ontologies have become extremely popular as a means for representing machine-readable semantic knowledge. The rapid growth of the Web and the information overload problem that it has caused has triggered significant research in the development of practical information extraction solutions that process Web content. However, the difficulty of extracting information from the Web, which was produced mainly for visualising information, has driven the birth of the Semantic Web. The Semantic Web will contain many more resources than the Web and will attach machine-readable semantic information to these resources. The first steps towards that goal, addressed knowledge representation issues for this semantic information, with the development of ontologies. Realizing the difficulty of designing the grant ontology for the world [96], 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. Having provided an ontology for a specific domain, the next step is to annotate semantically related Web resources. If done manually, this

[1] The BOEMIE project is presented in chapter 1.

G. Paliouras et al. (Eds.): Multimedia Information Extraction, LNAI 6050, pp. 134–166, 2011.

process is very time-consuming and error-prone. Information extraction is a promising solution for automating the annotation process. However, it comes along with the aforementioned knowledge acquisition bottleneck and the need for learning.

At the same time, acquiring domain knowledge for ontologies is also 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 automatic or semi-automatic construction, enrichment and adaptation of ontologies is known as ontology learning [79]. From our perspective, ontology learning is a wide research area that includes work on ontology enrichment, inconsistency resolution and ontology population. Ontology enrichment is the task of extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology. Inconsistency resolution is the task of resolving inconsistencies that appear in an ontology with the view to acquire a consistent (sub)ontology. Ontology population, on the other hand, is the task of adding new instances of concepts to the ontology.

Despite the fact that it is an emerging field, a significant amount of research has been performed already, leading to a large number of proposed approaches and practical systems. A fairly complete overview of the work performed in the field until 2003 is presented in [45], as well as in [99]. An updated overview of the field is also presented in [24]. Ontology learning has also significant presence in the major AI conferences, with workshops such as "Ontologies and Texts" (OLT) (EKAW2000 [8], ECAI2002 [9]), and other important conferences (IJCAI2001 [76], ECAI2000 [105] and workshops ECAI2004-OLP [18], OLP2 [20] and ECAI2008-OLP3 [22]).

The purpose of this chapter is to present the state of the art in ontology learning, by presenting the major approaches and most important practical systems that appear in the literature. The BOEMIE project is compared to these systems throughout this chapter and the solutions it gives to the various problems faced by the others are discussed. Systems and approaches are categorised along significant dimensions, such as the ontology elements learned, the starting point, the learning approach and the final outcome. The task of ontology learning is presented in section 2, covering the most significant approaches found in the literature. In section 3, ontology population is presented, as well as some important ontology population tools, which are also compared. Section 4 discusses ontology enrichment and follows a comparative presentation of ontology enrichment tools. Ontology evaluation is presented in section 5, while section 6 concludes this document.

2 Ontology Learning Foundations

Ontologies are a means for sharing and re-using knowledge, a container for capturing semantic information of a particular domain. A widely accepted definition of ontology in information technology and AI community is that of "a formal explicit specification of a shared conceptualization" [44], where "formal implies that the ontology should be machine-readable and shared that it is accepted by a group or community" [19]. Additionally, in the case of a domain ontology, it is usually assumed that it conveys concepts and relations relevant to a particular task or the application domain, which is the case we are interested in.

Ontology learning is the process of acquiring (constructing or integrating) an ontology (semi-) automatically. The acquisition of ontologies can be performed through three major approaches:

- By integrating existing ontologies. The integration process tries to capture commonalities among ontologies that convey the same or similar domains, in order to derive a new ontology. Several methods have been proposed in the literature, such as:
    - the merging of ontologies to create a single coherent ontology,
    - the alignment of ontologies by establishing links between them and allowing them to reuse information from each another, and
    - the mapping of ontologies by finding correspondence among elements in the ontologies.
- By constructing an ontology from scratch or by extending (populating and enriching) an existing ontology, usually based on information extracted from domain-specific content.
- By specialising a generic ontology, in order to adapt it to a specific domain.

In this chapter we will concentrate on the last two approaches, the construction of new ontologies and the enrichment/specialisation of existing ontologies.

Research in ontology learning studies methods and techniques for the acquisition of an ontology, based on semantic information, extracted from domain-specific content. Being closely related to the field of knowledge acquisition, a significant amount of the work presented in the bibliography concentrates on the task of knowledge acquisition from text, through the re-use of widely adopted natural language processing and machine learning techniques. However, ontology learning is not simply a replication of existing work under a different name, as it adds novel aspects to the problem of knowledge acquisition [19]:

- Ontology learning combines research from knowledge representation, logic, philosophy, databases, machine learning, natural language processing, image/audio/video analysis, etc.
- Ontology learning in the context of the Semantic Web must deal with the massive and heterogeneous data of the World Wide Web and thus improve existing approaches for knowledge acquisition, which target mostly small and homogeneous data collections.
- Substantial effort is being put into the development of extensive and rigorous evaluation methods in order to evaluate ontology learning approaches on well-defined tasks with well-defined evaluation criteria.

Following [19], the ontology learning process can be decomposed into six layers, forming a "layer cake"[2] of increasingly complex subtasks, which can be seen in Fig. 1.

[2] Ontology learning "layer cake" has been originally formulated with terminology originating from the textual modality. However, since the "layer cake" is applicable to multiple modalities, the labels of the layers have been slightly extended to cater for multimodality.

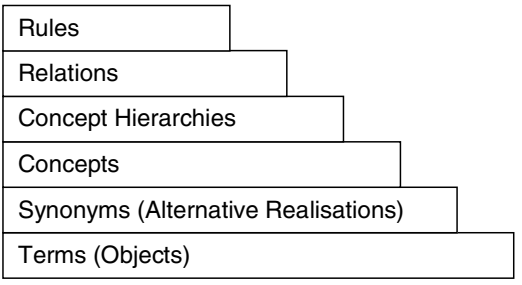

Fig. 1. Ontology learning "layer cake"

The main target of ontology learning is the definition of concepts and the relations between them. However, this implies substantial knowledge about the "symbols" that represent these concepts and relations and "instantiate" these into entities of the real word. We will use the notion of object or term to refer to these instances of concepts and relations, but it should be noted that we do not necessarily refer to the text modality: an object can be an audio, image or video segment that instantiates a concept or a relation in a corpus of the corresponding modality. Thus, in order to define new concepts/relations, the acquisition of knowledge about the objects that instantiate these concepts/relations in content is equally important. In addition to knowledge about objects/terms, object/term synonyms are also important: all terms that are synonyms (alternative realisations) refer to the same real object or event, and thus all materialise a single concept or relation. Failure to identify which terms/objects are synonyms may result in the introduction of redundant concepts or relations in an ontology, which in most cases is undesirable.

Among relations, one type is of particular importance to ontologies, namely hierarchical ones. These are the relations that realise the taxonomy backbone of an ontology, such as the subsumption relation (also referred as "is-a" relation in many cases). On the other hand, non-hierarchical relations are all relations that are not used in the formation of the concept hierarchy. Despite the fact that the relations are categorised into types, no type categorisation is performed at the concept level in the vast majority of the work presented in the literature.

Finally, an important aspect of an ontology is the ability to derive and make explicit facts that are implied by the knowledge in the ontology, mainly through reasoning. But for such derivations to occur, rules must be defined (and possibly acquired) to allow for such derivations. All of these aspects of ontology learning, related to things that can be learned, can be organised into the "layer cake" of Fig. 1 [19]. In the following subsections we are going to briefly present the state of the art for each layer of this "cake".

2.1 Object Identification

Object extraction (or identification) is a prerequisite for all aspects of ontology learning. An object is an instance of a recognisable entity in a multimedia corpus that conveys a single meaning within a domain (concept). A recognizable entity is something that can be recognized in multimedia corpora, such as words or phrases in textual

corpora, or areas in images. Since objects "materialise" a concept, objects found in a corpus usually represent candidate concepts that can enrich an ontology. Thus, the main objective is the identification of objects in a multimedia corpus that possibly convey concepts, which can be used for enriching an ontology. The object identification task can be decomposed into three subtasks [61]:

- Object recognition. This task is responsible for finding recognisable entities in the corpus that are objects.
- Object classification. This task assigns a semantic category to recognised objects. This categorization is important for the task of ontology learning, as these categories are often the concepts of the thematic domain.
- Object mapping. This task tries to link identified objects with relevant entities in other data sources, such as object libraries, vocabularies, lexica, thesauri and databases. A frequent use of this task is for exploiting similarities that potentially exist in the referred data sources, in order to identify clusters of objects that represent the same concept – synonyms/alternative realisations.

As object/term identification is an important task, not only for concept discovery for ontology learning but also for textual information extraction and retrieval, many approaches have been presented in the literature (mainly for the processing of textual corpora). Among the most successful ones are statistical methods, which usually measure the significance of each word with respect to other words in a corpus, based on word occurrence frequencies. TF/IDF [91] is often employed for this task [3, 30], possibly combined with other methods, such as latent semantic indexing [41] or taking into account co-occurrence information among phrases [43].

Clustering techniques also play an important role in object identification: recognizable entities can be clustered into groups based on various similarity measures, with each cluster being a possible object (consisting of synonyms). Approaches like [2, 37, 57] employ clustering techniques and other resources, such as the WWW and WordNet [38], to successfully extract terms. Additionally, both frequency and clustering-based approaches can be substantially enhanced through the use of natural language processing techniques, such as morphological analysis, part-of-speech tagging and syntactic analysis, as terms usually are noun phrases or obey specific part-of-speech patterns [47, 49]. Finally, morphological clues, such as prefixes and suffixes, can be very useful for some thematic domains: suffixes like "-fil" and "-itis" quite often mark terms in medical domains [50, 51].

Other methods use filters and heuristics. For example, Glossex [60] filters terminological candidates using lexical cohesion and a measure of domain relevance. It also uses some additional heuristics for extracting useful terms. TermExtractor [93] extracts a list of "syntactically plausible" terms and uses two entropy-based measures. The first metric, called Domain Consensus, is used to select only the terms which are used consistently throughout the corpus. The second one, Domain Relevance, is used to select only the terms that are relevant to the domain of interest. Finally, extracted terms are further filtered using Lexical Cohesion, which measures the degree of association of all the words in a terminological string.

2.2 Alternative Realization/Synonym Identification

Alternative realisations/synonyms are objects that refer to the same real object or event, variants in a corpus that can be thought to represent the same concept or relation. A significant amount of work has been performed mainly for text corpora, by exploiting resources such as WordNet [38]. Employing standard word sense disambiguation techniques [29, 64, 109] they seek to identify the most appropriate (WordNet) sense of each term, in order to collect synonyms associated with the sense. Other approaches try to locate term synonyms through clustering, mainly based on Harris's distributional hypothesis, according to which similar terms in meaning tend to share syntactic contexts [54, 68, 70, Hindle, 1990]. Related work is also performed in the field of information retrieval for term indexing, such as the family of Latent Semantic Indexing algorithms (LSI, LSA, PLSI, etc.), and the family of probabilistic topic models, e.g. Latent Dirichlet Allocation (LDA [12]). These methods apply dimensionality reduction techniques to reveal inherent relations between terms, in order to form clusters [63, 94]. Finally, more recent approaches extract synonyms by applying statistical approaches over the Web [10, 107]. For more information on such methods, the reader is referred to [19].

2.3 Concept Identification

Despite the fact that concepts are an important part of an ontology, what constitutes a concept is controversial. According to [19], concept formation should provide:

- An intentional definition of the concept.
- A set of concept instances.
- A set of realisations (i.e. terms).

Two types of intentional concept definition can be identified: informal and formal. An informal concept definition does not define a concept in terms of properties and relations between them, but in a more general, descriptive way, like for example a textual description or a concept gloss in a dictionary. Informal concept identification is quite rare, with only one approach appearing in the literature, the OntoLearn system [111], which associates WordNet glosses with domain specific concepts. Formal concept definition, on the other hand, builds on top of object and synonym identification, by formulating concepts as clusters of "related" objects. It exploits relations among objects that are discovered using approaches which will be described in the following two subsections. Basing the definition of a concept on a cluster of objects automatically provides the set of realisations of the new concept. The association of a set of instances with a concept is known as ontology population or ontology tagging, and it will be presented in greater detail in section 3.

2.4 Taxonomy Construction

An important part of an ontology is its taxonomy, or the hierarchy of concepts. Subsumption relations (also known as "is-a" or inclusion relations) provide a tree view of the ontology and determine inheritance between concepts. A popular approach for taxonomy discovery in textual domains is the use of lexico-syntactic patterns (such as

Hearst patterns [53]). According to this approach, syntactic elements (such as noun phrases) are combined with characteristic phrases to identify inclusion relations. Examples of such patterns can be the following ones (NP stands for noun phrase):

- NP such as NP, NP,..., and NP
- such NP as NP, NP,..., or NP
- NP, NP,..., and other NP
- NP, especially NP, NP,..., and NP
- NP is a NP

Several systems have been proposed based on simple variations of the above idea, such as [56, 57, 84]. More recent systems also employ pattern learning algorithms to automate pattern construction [1, 31, 103]. For non-textual domains, machine learning methods, such as hierarchical clustering, can be used. Further details on such approaches can be found in [115] and [19].

Yang and Callan [108], in a metric-based taxonomy induction framework, combine the strengths of pattern-based and clustering-based approaches. The framework incorporates lexico-syntactic patterns as one type of feature in a clustering framework. It integrates contextual, co-occurrence, syntactic dependency, lexical-syntactic patterns, and other features to learn an *ontology metric*, i.e. a score indicating semantic distance, for each pair of terms in a taxonomy; it then incrementally clusters terms based on their ontology metric scores.

Snow et al. [102] have presented an algorithm for inducing semantic taxonomies, which attempts to globally optimize the entire structure of the taxonomy. The model has the ability to integrate heterogeneous evidence from different classifiers, offering a solution to the key problem of choosing the correct word sense for a new hypernym.

A particularly interesting machine learning technique for hierarchy construction is the estimation of Probabilistic Topic Models that produce a hierarchical modelling of a particular collection. Among the well known models of this family is the hierarchical Latent Dirichlet Allocation (hLDA) [13], where each document is modeled as a set of topics across a specific path of the learned hierarchy from the root to a leaf. In addition, the models of the Pachinko Allocation family, like PAM [66], hPAM [83] and non-parametric PAM [67] deal with some of the problems of hLDA, such as the lack of multiple inheritance between topics at different levels of the hierarchy. Among the major benefits of methods that rely on such models is that the identification of topics, which serve as concepts in the ontology, and their taxonomic arrangement is performed simultaneously. In addition, these models do not require an initial ontology to start from. They construct a taxonomic backbone without any prior knowledge, but a collection of documents. In order to learn topic ontologies, probabilistic topic models have been applied in [117,118] and in [114].

2.5 Semantic Relation Extraction

Relations beyond the concept hierarchy (non-taxonomic relations) constitute also an important component of an ontology. Such relations can be extracted with approaches similar to the ones used for extracting taxonomic relations. In textual domains, where most of the existing work has focussed, lexico-syntactic patterns again play an important role. Verbs usually represent actions or relations between recognisable entities in

sentences. As a result, verbs are assumed to express relations between entities, which may be useful for enriching an ontology, provided that the involved entities can be associated with concepts of the ontology. Systems like the RelExt tool [95] use such patterns to identify related pairs of concepts. Additionally, semantic clustering of verbs has been reported to help in situations where extraction of specific relation types is desired [101]. Finally, association rule mining algorithms have been used for the acquisition of non-taxonomic relations for ontology enrichment [74, 75].

2.6 Ontology Rule Acquisition

Ontology rule acquisition is probably the least addressed aspect of ontology learning, as almost no work has been presented that acquires rules. An initial attempt to formulate the problem is presented in [69], where an unsupervised method for discovering inference rules from text is presented. Learned rules are of the following form "*X is author of Y* ≈ *X wrote Y*, *X solved Y* ≈ *X found a solution to Y*, and *X caused Y* ≈ *Y is triggered by X*" [69]. Also, Sangun et al., [92] proposed an ontology rule acquisition procedure using an ontology, which includes information about the rule components and its structure. The procedure comprises rule component identification and rule composition. They use stemming and semantic similarity for the former and a Graph Search method for the latter. Finally, in the field of inductive logic programming (ILP), which deals with the induction of first-order rules, some attempts have been made to address reasoning for the Semantic Web [71].

2.7 Comparative Analysis of Ontology Learning Tools

During the last decade, a large number of approaches and practical systems have been presented that try to automate ontology construction. The presented approaches are so diverse, and thus trying to classify existing systems along a single "dimension" will be at least incomplete. Thus, for this document a comparison framework similar to the one proposed in [99] will be adopted, where some important comparison "dimensions" are defined. Following [99], we will classify existing approaches/practical systems performing both ontology population as well as ontology enrichment, according to the following categorisation criteria:

- **Elements of the "layer cake" learned.** The elements of the "layer cake" that are learned provide a good view of the complexity and capabilities of an ontology learning system, through the ontological aspects learned by the system. It is desirable for a system to provide solutions to as much layers as possible.
- **Initial requirements.** Initial requirements, such as prior knowledge and type of required input for learning an ontology, clarify the starting point of an ontology learning system, the background knowledge and the resources available in order to help knowledge acquisition. In addition, the use of domain-depended resources affect directly the feasibility of a system, as it restricts its portability to new thematic domains.
- **Learning approach.** Of particular interest is also the approach an ontology learning tool adopts in order to extract knowledge, and whether this approach is specialised to the domain, e.g. an extraction engine based on manually

constructed patterns, or a more general one, e.g. based on machine learning or statistical methods. The learning approach adopted by a system usually affects other categorisation criteria, such as the initial requirements and of course the degree of automation, as the usage of machine learning methods usually reduces the degree of manual intervention of the domain expert during knowledge acquisition.

- **Degree of automation.** The degree that a system automates decisions is important, as it contributes to the plausibility of the system. A fully automated system is of course desirable, but it may not be always possible, especially with tasks related to ontology enrichment. But even in the case of semi-automated or cooperative systems, various degrees of automation can be identified. For example, the required knowledge expected by the expert: interaction through a domain expert may be more desirable than interaction through an ontology expert, who is expected to know both the thematic domain in addition to ontology engineering.
- **Consistency maintenance and redundancy elimination.** We are also interested in the outcome of the system and the knowledge representation structures used for storing the acquired information. Systems that do not enhance an ontology usually do not deal with aspects such as consistency maintenance or redundancy elimination. Maintaining the consistency of an ontology is crucial, as an ontology that contains conflicting information is of little use. Redundancy elimination on the other hand is not as crucial as consistency, i.e., redundancy cannot render an ontology useless, unless it also introduces contradictions. However, redundancy elimination can enhance the plausibility of an ontology by facilitating the process of querying the ontology, and at the same time by limiting the size (and complexity) of the ontology.
- **Domain portability.** An important aspect of an ontology learning system is whether it can be ported to other thematic domains or not. Systems that exhibit increased domain portability tend to explicitly define the required domain knowledge, whereas less portable system can contain domain specific knowledge in the internals of the system.
- **Corpora Modality.** It is desirable for a system to be able to process more than one modalities, as it can provide evidence of the ability of a system to accommodate and exploit diverse knowledge sources, fuse the extracted information and provide unified results that are valid across modalities.

2.8 A Procedural View of Ontology Learning

Based on our experience in the area from our involvement in several relevant projects, we consider that the task of ontology learning involves the subtasks of population, enrichment, and inconsistency resolution. Ontology population is the process of adding new instances of concepts/relations into an ontology, usually by locating the corresponding object/terms and synonyms in the corpus. Ontology enrichment is the process of extending an ontology with new concepts, relations and rules. Inconsistency resolution is responsible for remedying problems introduced by population and enrichment. In addition to these subtasks, ontology evaluation is also needed in order

to measure the plausibility of the learned ontology by evaluating the usefulness of the changes. Fig. 2 depicts a typical ontology learning process.

Very often, ontology learning is modelled as a bootstrapping process: an initial ontology is used as a basis for learning a new ontology, which in turn substitutes the initial one and the whole process restarts. In particular, an initial ontology is used to analyze and extract information from a corpus. The extracted information is used to evolve the ontology, and through the evolved ontology the extraction of information is improved. The bootstrapping process continues until no more information can be extracted from the corpus. Here we have to note that in every cycle the consistency of the ontology is checked and in the case of inconsistency, the changes are discarded. In the following section, the steps involved in ontology population will be described in more detail, along with a comparative analysis of the most important approaches and practical systems performing ontology population. The steps of ontology enrichment will be presented in section 4, along with a comparative analysis of the most important approaches and practical systems performing ontology enrichment. Finally, ontology evaluation will be presented in section 5.

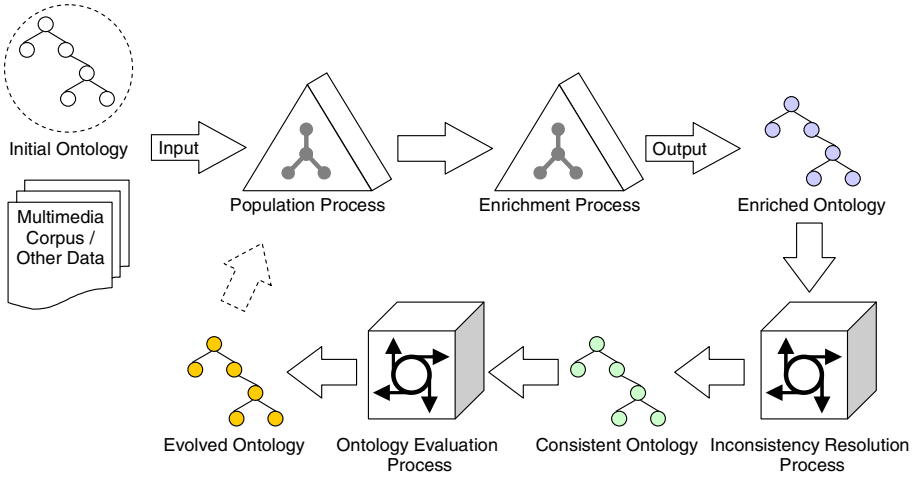

Fig. 2. The process of ontology learning

3 Ontology Population

Ontology population is the process of inserting concept and relation instances into an existing ontology. In a simplified view, an ontology can be thought of as a set of concepts, relations among the concepts and their instances. A concept instance is a realisation of the concept in the domain, e.g. the instantiation of the concept as a phrase in a textual corpus. The process of ontology population does not change the structure of an ontology, i.e., the concept hierarchy and non-taxonomic relations are not modified. What changes is the set of realisation (instances) of concepts and relations in the domain. A typical ontology population methodology is depicted in Fig. 3.

Ontology population requires an initial ontology that will be populated and an instance extraction engine. The extraction engine is responsible for locating instances (realisations) of concepts and relations in a multimedia corpus. A multimedia corpus is processed by the extraction engine, in order to locate concept/relation. The list of extracted concept/relation instances is subsequently used to populate the ontology.

Recalling the "layer cake" idea, the population process involves some of the layers presented in section 2. In particular, it deals with the acquisition of realisations (i.e. objects and alternative realisations/synonyms) of both concepts and relations. A typical approach is to use known realisations associated with concepts/relations which may have been identified during concept/relation formation, to locate the corresponding objects/synonyms in a corpus. This process is also known as lookup text extraction or prototype recognition in image analysis. The result is an annotated corpus, which can be used to construct more general instance extractors, using machine learning.

An interesting aspect of ontology population, which is not addressed adequately in the literature, is the handling of redundancy. The elimination of redundancy in the instance set requires entity disambiguation, which is the process of identifying instances that refer to the same real object or event. If an ontology is populated with an instance without checking if the real object or event represented by the instance already exists in the ontology, then redundant instances will be inserted. A worst case scenario is that redundant instances contain contradicting information, which may lead to an inconsistent ontology.

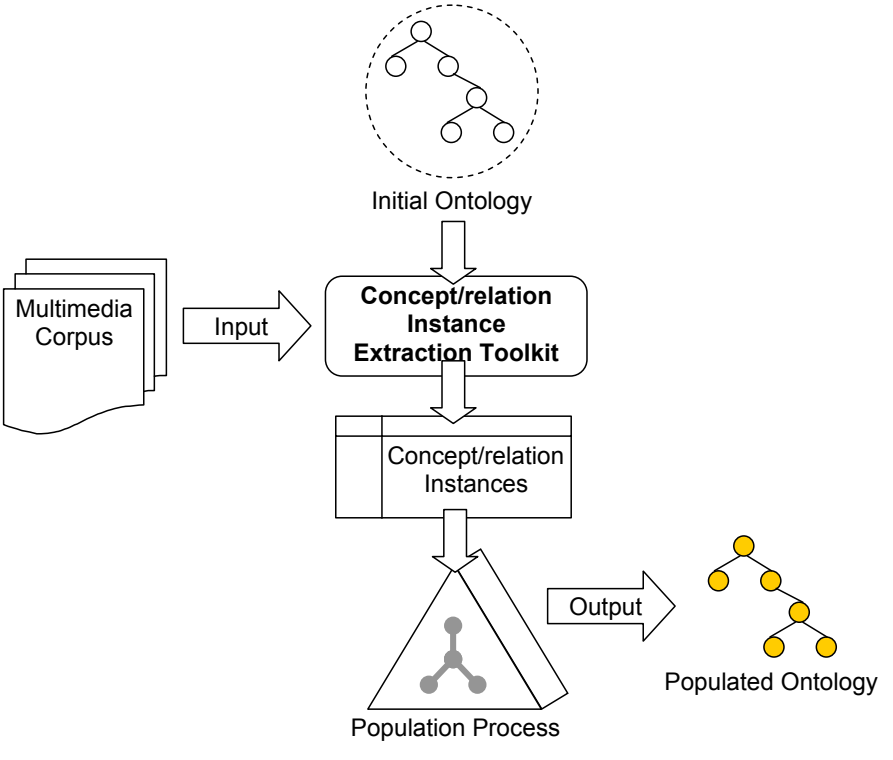

Fig. 3. The ontology population process

To our knowledge, only three approaches address this problem. The Artequakt system [4, 5, 6, 59] applies manually written heuristics, in order to merge instances that refer to the same real object or event. These heuristics are evaluated after a batch of instances has populated the ontology. The SOBA system [21], on the other hand, performs simple checks using special mapping rules, during instance creation (i.e. before the instances populate the ontology), in order to re-use instances that refer to the same real object or event instead of creating new ones. The approach followed by BOEMIE enhances that of Artequakt, through the use of machine learning instead of manually-developed heuristics.

3.1 The BOEMIE Approach to Ontology Population

BOEMIE [23] implements an ontology based information extraction system, that is able to extract objects from a variety of modalities, including texts, images, and videos. Due to its multimodal nature, the BOEMIE system clearly distinguishes entities from their realisations (through properties) in the various modalities. Exploiting the idea that you cannot find entities in corpora but rather their properties, BOEMIE adopts a different approach that separates the concepts into two types: "primitive" concepts that can be easily attributed to objects (i.e. have direct realisations) – mid-level concepts (MLCs) in BOEMIE terminology – and "composite" concepts (that represent real objects or events), usually build on top of primitive ones. These "composite" concepts do not have direct realisations as they cannot be mapped directly to an object and are named high-level concepts (HLCs) in BOEMIE. For example, consider a person that is referenced in a set of textual documents, images and videos. From the text modality BOEMIE can extract a person name, an age, a gender or a profession: this set of properties is considered instances of MLCs for the text modality. In addition, by exploiting linguistic information (such as verbs), relations may be extracted that relate these MLC instances with each other (i.e. suggesting that a specific age, gender and profession are related with a specific person name). Similarly, from an image anatomical parts (i.e. a person face) can be extracted, and possibly a person name from the caption or through OCR. Again, all these are instances of MLCs for this modality, possibly related to each other through spatial and proximity relations.

Despite the fact that instances of properties of a person have been extracted from the involved modalities, a person instance has not yet been identified. This is because "person" is a "composite" concept, an HLC. The identification of entities, and thus the instantiation of HLC instances, is performed as a second processing step: reasoning is employed, where through rules MLC instances (properties) extracted from the various modalities are fused and interpreted. During fusion and interpretation, relations between MLC instances will be examined in order to identify the number of involved entities (i.e. persons) and which properties belong to which person. The result of the interpretation process will be instances of HLC concepts, for all identified entities.

Since the vast majority of work in ontology learning does not discriminate between "primitive" and "composite" concepts, ontology population in these systems is performed as a single step, i.e. the instances that are assimilated into the ontology are identified directly by the instance extraction tool, thus requiring the incorporation of

considerable domain knowledge in the extraction tool. Instance extraction tools typically instantiate complex composite structures with groups of realisations (objects/terms) related to each other through ontology relations.

The population methodology proposed by the BOEMIE project distinguishes between two layers of complexity when populating an ontology with concept instances. Concepts are divided into "primitive", called mid-level concepts, and "composite" ones, called high-level concepts. In contrast to mid-level concepts that are populated by extraction tools as described above, the high-level concepts are populated by reasoning over the mid-level instances, since they are defined in terms of "primitive" concepts. The main differences between the BOEMIE approach and the state of the art are:

- The concept/relation instance extraction engine is not expected to extract instances of "composite" concepts. It is expected to extract only instances of "primitive" concepts. A clear advantage is the fact that the extraction engine becomes immune to changes in the organisation of the ontology, which is a desired property in environments where the ontology evolves over time. The extraction engine needs to adapt only when new "primitive" concepts or relations involving "primitive" concepts are modified.
- The ontology is used to create instances of "composite" concepts from populated "primitive" concept instances and populated relation instances, through non-standard reasoning[3]. The advantages of such an approach are two–fold: a) "composite" concept instances are always in sync with the current formal definition of the relevant concepts, and b) the formation of "composite" instances respects the constraints that may be imposed by the ontology, i.e. through rules, thus helping maintaining the consistency of the ontology.

To our knowledge, there is no method in the bibliography following this two-stage approach to ontology population.

3.2 Comparative Analysis of Ontology Population Tools

The vast majority of the systems found in the literature for ontology population, share the architecture depicted in Fig. 3: an extraction toolkit is used for object/term identification or named-entity recognition, in order to locate instances of concepts and in some cases also instances of relations between concepts, which are then assimilated into the ontology. Ontology population systems are closely related to ontology-based information extraction systems, since the latter provide mechanisms to associate pieces of the data with concepts of an ontology. Thus, every ontology-based information extraction system can be viewed as an ontology population system, as it can be extended to assimilate extracted instances into the ontology.

In the rest of this section we present a comparative analysis of the main approaches and practical systems that have been presented in the literature for ontology population.Table 1 presents a summary of the systems. The comparison is guided by our categorisation criteria described in subsection 2.7, relating also important features of

[3] BOEMIE employed abductive reasoning in order to create "composite" objects from "primitive" ones.

the BOEMIE project, such as portability to other thematic domains, preservation of the ontology consistency and entity disambiguation, as explained in subsection 3.1. Also, due to the focus of BOEMIE on multimedia corpora, we categorize the different systems according to the modality of the data they can handle. This parameter has proved particularly important, as the majority of the systems use textual corpora, and they rely heavily on linguistic processing, such as syntactic analysis, or exploitation of additional resources like thesauri and semantic hierarchies.

Elements extracted. Some systems are more complete in the sense that they populate an ontology with instances of both concepts and relations, such as Artequakt [4, 5, 6, 59], WEB→KB [26], SOBA [21], [85, 86], OPTIMA [58] and ISOLDE [113]. Others, such as Adaptiva [15], LEILA [106] and [7] concentrate only on relation instances. Finally, the KnowItAll system [34, 35] identifies only concept instances, while BOEMIE is able to extract both concept and relation instances in order to populate the ontology.

Table 1. Brief description of the different systems for ontology population

System	Description
Artequakt	Extracts knowledge from the web about artists, populates a knowledge base and uses it to generate personalized biographies. Once instances have been identified, the system uses a domain specific ontology and a generic one in order to extract binary relations between two instances. It uses heuristics to remove redundant instances from the ontology.
WEB→KB	Combines statistical and logical (FOIL rule learning) methods to learn concept instances and relation instances from web documents. The system employs document classification to identify and classify as instances whole pages from the web. Instances of relations are retrieved by examining hyperlink paths that connect web pages.
KnowItAll	Uses domain-independent lexico-syntactic patterns to extract possible instances. It selects the instances by evaluating their plausibility, using a version of the pointwise mutual information statistical measure.
Adaptiva	Employs a bootstrapping approach, extracting instances of relations from a corpus and asking an ontology expert to validate them. The outcome of validation is used by Amilcare [25], functioning as a pattern learner. Once the learning process is completed, the induced patterns are applied to unseen corpora and new examples are returned for further validation by the user.
SOBA	Automatically populates a knowledge base by information extracted from soccer match reports as found on the web. It employs standard rule-based information extraction to extract named entities related to soccer events. The extracted information is converted into semantic structures, as defined by the ontology, with the help of mapping rules.
[85, 86]	A pattern-based system to automatically enrich a core ontology with the definitions of a domain glossary. It uses manually developed lexico-syntactic patterns for extracting instances of concepts. These instances are processed in order to extract relation instances which associate extracted information with concept properties.

Table 1. (*Continued*)

LEILA	A system that learns to extract instances of binary relations from natural language corpora. The system employs statistical techniques to learn the extraction patterns for the relation.
[7]	Automatically learns extraction patterns for finding semantic relations in unrestricted text, based on statistical corpus processing.
OPTIMA	A (semi-)automated system for populating ontologies from unstructured or semi-structured texts. It extracts relational information with natural language processing techniques. It assigns instances to concepts by calculating a fitness value between a candidate instance and each concept in the ontology, using the hierarchical syntactic information of the ontology schema.
ISOLDE	Generates a domain ontology from a seed ontology by exploiting a general purpose NER system and lexico-syntactic patterns to extract concept candidates. Concept candidates are then filtered according to their statistical significance and the knowledge that can be derived from available Web resources.
BOEMIE	Combines an ontology-based information extraction (OBIE) engine based on machine learning, with an inference engine, in order to extract "primitive" concept instances from multiple modalities, which are then fused and interpreted (through abductive reasoning) to form instances of "composite" and more abstract concepts.

Initial requirements. In order to be self-sustained, an ontology population system should have as few initial requirements as possible, in terms of resources or background knowledge. Some systems do not perform object/term and synonym identification, but rather employ publicly available processing resources for this task. Artequakt is based on the information extraction toolkit GATE [27, 28] to perform named entity recognition, syntactic and semantic analysis. SOBA uses a standard rule-based information extraction system, an enhanced version of SProUT – [32], while [7] a part of speech tagger and a module for named entity recognition. Other systems, instead of employing a term/synonym extraction engine, require extraction patterns to be provided by the user. For example, KnowItAll uses domain-independent lexico-syntactic patterns, inspired by Hearst patterns [53]. On the other hand, the system presented in [85, 86] uses manual extraction patterns to populate the CIDOC CRM ontology with terms extracted from glosses of the Art and Architecture Thesaurus (AAT). OPTIMA uses user-defined named entity types, organized in a hierarchy, and user-defined binary relations. A name-entity recogniser based on these particular entity types is used for the extraction of instances. ISOLDE uses a general purpose named entity recogniser to find instances in a base ontology and then uses Hearst patterns to find class candidates. Systems like WEB→KB, Adaptiva and LEILA include an adaptable term/synonym extraction engine which can be taught with the help of concept/relation instance examples. BOEMIE adopts a similar term/synonym extraction approach. An adaptable term/synonym extraction engine is employed using examples of instances that are provided either through manually annotated corpora, or by the previous ontology population steps.

Learning approach. Machine learning seems to be the choice of the majority of systems, as all but three of the examined systems (Artequakt, SOBA, [85, 86]) employ some form of learning. The systems employing machine learning either use statistical methods to identify terms, or perform automated pattern extraction. For example, Adaptiva uses a tool for adaptive Information Extraction from text (IE), to learn patterns. KnowItAll uses an extended version of the pointwise mutual information [107] statistical measure, which selects the instances that will populate the knowledge base, by evaluating their plausibility. OPTIMA uses a trainable named entity recognizer, combining a boundary detector using CRFs [62] and a named-entity classifier using maximum entropy. ISOLDE employs a seed ontology and the general-purpose NER system SProUT [32] to extract instances for concepts in the seed ontology. Then lexico-syntactic patterns [53] are applied to identify possible new concepts, which are then filtered with the help of heuristics and knowledge obtained from online resources, such as Wikipedia[4], Wiktionary[5] and DWDS[6]. Finally, WEB→KB uses both a statistical and a symbolic approach (FOIL [88]) to learn classifiers that can detect instances and relations between instances. The three systems that do not use machine learning either employ an external, publicly available term/synonym extraction engine or require manually-constructed patterns as input, as they seem to rely mostly on linguistic information. The LEILA system also relies on linguistic knowledge, but employs additional filtering based on statistical approaches, such as adaptive k-Nearest-Neighbor-classifiers and Support Vector Machines. BOEMIE also uses machine learning. In particular, the term/synonym extraction engine makes use of both linguistic information (especially shallow syntactic analysis) and machine learning to identify concept instances and relations, while automated pattern extraction is used for relation extraction.

Degree of automation. This criterion examines the extent to which the domain expert needs to intervene during knowledge acquisition. With the exception of Adaptiva, all other systems examined here do not require interaction with the domain/ontology expert. This is an indication that the population process can be fully automated, which is also true for the approach adopted in BOEMIE. BOEMIE directly populates an ontology instead of producing an intermediate representation of instances. In addition, BOEMIE provides a graphical user-interface that enables the domain expert to examine and revise the populated instances, if such a need arises.

Consistency maintenance and redundancy elimination. These issues are only addressed by three systems (Artequakt, SOBA and BOEMIE). The Artequakt system uses manually-written heuristics, in order to merge populated instances that refer to the same real object or event. SOBA, on the other hand, performs simple checks during instance creation, i.e., before the instances populate the ontology, in order to reuse instances that refer to the same real object or event instead of creating new ones. The BOEMIE approach enhances the Artequakt proposal through the use of matching techniques instead of manually developed heuristics. More specifically, BOEMIE instance matching methods try to identify instances that refer to the same real entity or event and group them, rather than merging them into a single instance.

[4] http://en.wikipedia.org/
[5] http://en.wiktionary.org/
[6] http://www.dwds.de/

Domain portability. Some of the systems are domain-independent (KnowItAll, Adaptiva, LEILA, OPTIMA, ISOLDE, BOEMIE), as they do not use any domain-specific resources, while others are domain specific (SOBA, [85, 86] and [7]. There are also some systems that have limited portability, such as Artequakt and WEB→KB. The reason for this is either that they are applicable only to domains with specific characteristics, or that they require adaptation to the new domain, in ways not tested in their current work.

Corpora Modality. All the mentioned systems with the exception of BOEMIE are applied to text. No special effort has been made for other modalities, such as video, images or multimedia. BOEMIE explores this direction, by analysing multimedia corpora. The BOEMIE system supports the identification of objects from multiple modalities (such as text, image, video, audio and text from image/video OCR), which are then fused through reasoning (employing both deduction and abduction) to form instances of modality-independent concepts.

4 Ontology Enrichment

Ontology enrichment is the process of extending an ontology, through the addition of new concepts, relations and rules. It is performed every time that the existing domain knowledge is not sufficient to explain the information extracted from the corpus. Thus, the ontology enrichment activity is expected to extend the background knowledge, in order to better explain extracted information in the future. Since new concepts and relations can be added during enrichment, the structure of the ontology changes. Recalling our discussion of the "layer cake", the enrichment process involves all of the layers presented in section 2, unlike ontology population which is concerned only with the lower layers. The main approach adopted by the state-of-the-art methods starts with the identification of objects and their alternative realisations/synonyms. Each object, along with a possible set of alternative realisations, is a candidate concept to be added to an ontology. Advancing to the third layer of the "cake", each proposed cluster of objects and alternative realisations that possibly represent a concept must be evaluated in order to decide whether it constitutes a concept or not. In case the object represents a concept, the concept must be formulated by creating an intentional definition (section 2.3) and possibly augmented with evidence/instances that justify the addition of the new concept. At the next layer, relations (either taxonomic or non-taxonomic) must be identified between concepts, usually based on spatio-temporal information for modalities like image and video or linguistic information (either syntactic or semantic) for text. Finally, in order to support reasoning and derive facts not explicitly encoded but derivable from the ontology, rules and constraints must be acquired.

4.1 The BOEMIE Approach to Ontology Enrichment

Unlike ontology population which can be fully automated, ontology enrichment remains typically a semi-automated procedure. All systems presented in the literature require the manual intervention of a domain expert, in order to review and accept or reject the system's proposals (Fig. 4). The methodology proposed by the BOEMIE

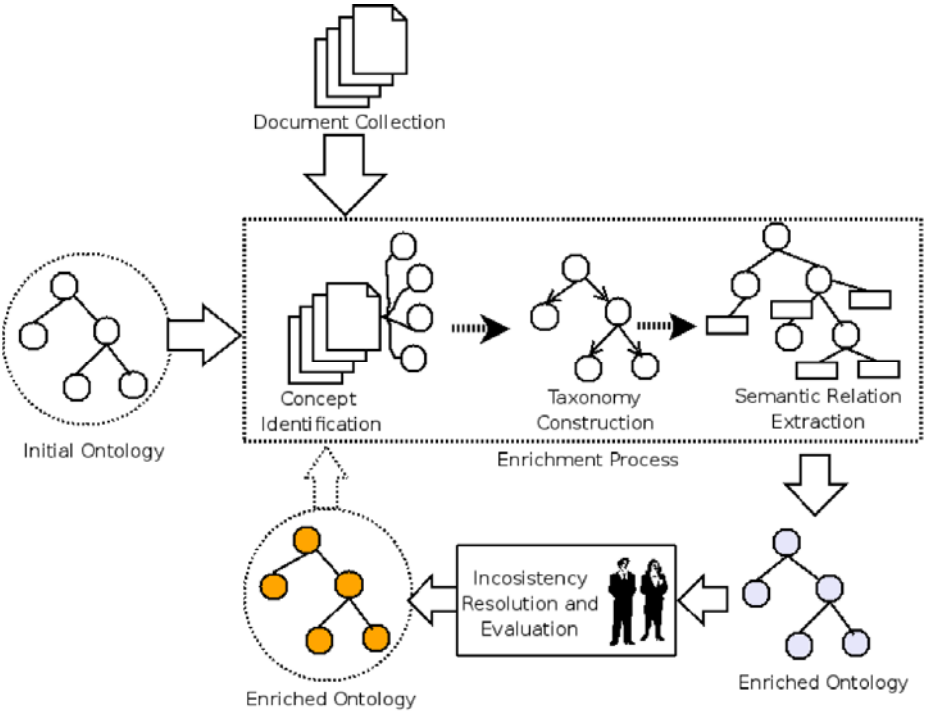

Fig. 4. The ontology enrichment process

project is not an exception. BOEMIE proposes a semi-automated approach which tries to minimise the role of the expert as much as possible.

As in ontology population, a two-stage approach is used. That is, the system distinguishes between high-level and mid-level concepts, as introduced in subsection 3.1. Ontology enrichment in BOEMIE is driven by the quality of the interpretation achieved for a multimedia resource: if a sufficient number of MLCs (properties) have been extracted from the involved modalities, and a large percent of these MLC instances have been successfully interpreted (through their relation to HLC instances), the background knowledge (ontology) is considered as sufficient to describe the multimedia resource. Ontology enrichment is triggered when the background knowledge is not sufficient to interpret adequately a resource: if a significant number of MLC instances are not part of the interpretation (i.e. not related to HLC instances), then the system tries to enrich the ontology through the addition of new HLC concepts. Similarly, if an inadequate number of MLC instances have been identified for one or more modalities, the system tries to enrich the ontology through the addition of new MLC concepts, by triggering the relevant modality-specific enrichment process for the involved modalities. Both enrichment processes rely on clustering techniques to perform proposal of possible new MLCs/HLCs, which are then enhanced with the use of external knowledge sources, through ontology matching techniques, before presented to a domain expert for final verification/approval. Once a concept has been approved

for inclusion into the ontology, the required fusion/interpretation rules used during reasoning are automatically created. Among the innovative aspects of BOEMIE, are the use of non-standard clustering, which tries to cluster ontological fragments, and the use of external knowledge sources aiming to provide the expert additional information during concept and relation definition. More information about this approach can be found in [23].

4.2 Comparative Analysis of Ontology Enrichment Tools

In this subsection we perform a comparative analysis of the most influential ontology enrichment systems. Table 2 presents the systems along with a brief description.

Elements learned. Some of the examined systems are more complete than others, in the sense that they cover several layers of the "cake" presented in section 2. Systems like ASIUM [39, 40], HASTI [97, 98, 100], TEXT-TO-ONTO [77], VIKEF[7] (Virtual Information and Knowledge Environment Framework) and KAON [79] perform learning of new concepts, relations and in some cases even rules. On the other hand, systems like SYNDIKATE [52], ABRAXAS [17, 55], ATRACT [82], and [104] concentrate on concept or relation learning. The BOEMIE ontology enrichment methodology incorporates methods to extract concepts, hierarchical and non-hierarchical relations and rules.

Initial requirements. Almost all systems rely on some form of linguistic analysis, exploiting syntactic relations to identify new concepts, relations or even rules. Besides linguistic knowledge, only a few systems require additional background knowledge, such as a domain ontology, domain specific lexicons or lexicon-syntactic patterns (SYNDIKATE, ABRAXAS, VIKEF, ATRACT). The BOEMIE approach follows a slightly different direction, as it has no initial requirements. Operating solely on the results of information extraction that have been enhanced through reasoning, BOEMIE learns concepts and relations through instance clustering. Furthermore, it tries to associate unknown objects with existing concepts/relations, through the use of external knowledge sources.

Learning approach. Machine learning seems to be the choice of most of the systems, especially in the form of clustering (e.g. ASIUM, HASTI, TEXT-TO-ONTO, KAON and BOEMIE) or lexico-syntactic pattern acquisition (ABRAXAS). BOEMIE also uses clustering on the results of multimedia interpretation through reasoning, rather than at the term/synonym level which is the common approach. As a result, clustering in BOEMIE effectively operates on ontological instances.

Degree of automation. In contrast to ontology population, the enrichment process cannot be fully automated, at least by the existing systems. Most systems interact with an ontology expert who has the final word on the modification of the ontology. Those systems that do not involve the expert either require significant background knowledge and/or support very limited knowledge acquisition (e.g. SYNDIKATE, ABRAXAS, VIKEF, ATRACT, [104]). SYNDIKATE requires an almost completeontology, which can be augmented with new concepts originating from unknown

[7] http://cordis.europa.eu/ist/kct/vikef_synopsis.htm, http://www.vikef.net/

Table 2. Brief description of ontology enrichement systems

System	Description
ASIUM	Learns terms, synonyms, concepts and hierarchical relations from unrestricted text corpora, based on syntactic analysis. It employs machine learning (hierarchical clustering) in order to learn concept hierarchies, with manual supervision by the domain expert.
HASTI	Learns terms, concepts, hierarchical and non-hierarchical relations and axioms in incremental and non-incremental modes. It starts from a small kernel ontology, using a hybrid approach, combining logical, linguistic, template-driven, and heuristic methods.
SYNDIKATE	A system for automatically acquiring knowledge from real-world texts and representing it into formal structures. Through reasoning, an unknown term is either added to an existing concept or creates a new one.
TEXT-TO-ONTO	Learning concepts and relations from unstructured, semi-structured, and structured data, using a multi-strategy method which combines association rules, formal concept analysis and clustering.
ABRAXAS	Performs concept and relation extraction, using automated lexico-syntactic pattern acquisition. This process spots all instances of concepts and relations already in the ontology and acquires extraction patterns using machine learning. These patterns are subsequently applied to the corpus, in order to detect new concepts and relations, the plausibility of which is accessed by a statistical measure.
KAON	Provides components for each subtask of the learning process. It contains an algorithmic library that supports clustering, classification and other techniques. It learns concepts, taxonomic relations and other general binary relations between concepts.
[104]	Learns instances of relations from unstructured corpora. It extracts triples that represent relations between entities/terms. The system employs various metrics for filtering the list of extracted triples in order to decide if a new relation has been discovered.
VIKEF	The system proposes a methodology for extracting information from product catalogues, aimed by an ontology to provide domain knowledge and guide the disambiguation process. The domain ontology can be enriched with parts from other ontologies, selected from a pool of ontologies.
ATRACT	Used for terminology recognition and clustering based on the C/NC-value method (a method for the automatic extraction of multi-word terms, which combines linguistic and statistical information) [43]. It specialises to the domain of molecular biology.
BOEMIE	BOEMIE employs an OBIE extraction engine along with a semantic interpretation engine orchestrated by a bootstrapping approach in order to enrich a seed ontology. The system continuously monitors the quality of interpretations achieved for multimedia resources and performs ontology enrichment when the background knowledge is found inadequate to interpret a set of resources, through a semi-supervised approach. Concept proposals expressed in natural language are automatically generated by exploiting both internal and external knowledge, which must be revised and approved by a domain expert.

terms. However, these concepts can be added mainly near the existing conceptual taxonomy, assuming that there is resemblance in the syntactic usage of the unknown term and concept lexicalisations already in the ontology. ATRACT serves mainly as a workbench for terminology recognition and clustering and is mainly targeting the domain of molecular biology. VIKEF also uses an initial ontology, which is created using a subset of the taxonomical glossary obtained from a product catalogue. This ontology forms the basis for the development of the final ontology about product catalogues. VIKEF applies pattern matching techniques to identify individual product descriptions. For each identified product, its natural language description is processed in order to identify relevant entities and relations between them. The learning process takes advantage of the results of the extraction to enrich the ontology. In addition, similar existing ontologies or parts of them are retrieved from a pool of available ontologies, and they are used to extend the domain ontology. ABRAXAS uses three external resources, namely a corpus of text, some lexico-syntactic textual patterns and an ontology. It considers ontology learning as a process that maintains these resources in some form of equilibrium, as a change in one resource triggers actions in the rest of the resources, in order to reach a consistent overall state. Specia and Motta [104] concentrate mainly on relation identification, thus supporting a very limited type of enrichment. BOEMIE belongs in the family of methods that interact with a domain expert, thus implementing a semi-automated approach to enrichment. However, BOEMIE aims to automate as many tasks as possible, employing also the use of diverse knowledge sources, in order to help the domain expert. It is worth noting that BOEMIE needs a domain expert and not an ontology expert, presenting in a natural-language format only part of the ontology. For example, when a cluster is identified as a candidate concept, a formal definition of the concept is automatically induced along with the required interpretation rules, augmented with its instances. In addition, external knowledge sources, such as other ontologies or Web directories sharing the same or similar thematic domain, are aligned to the concepts of the BOEMIE ontology and used to further enhance the suggested formal definition of a concept. Following the TEXT-TO-ONTO paradigm, BOEMIE provides a natural user interface to the domain expert, who is requested to revise, if needed, and approve the proposed definition. More details about the methodology proposed by BOEMIE can be found in [23].

Consistency maintenance and redundancy elimination. BOEMIE puts significant effort in maintaining the consistency of the ontology while at the same time keeping the ontology clean from redundant information. Consistency maintenance is an automated process performed with the help of reasoning, while redundancy elimination is performed mainly by the domain expert, who is responsible to evaluate whether the supportive information (i.e. clustered instances) for a new concept/relation is enough to justify its addition. Alternatively, this information can be associated with an existing concept/relation.

Domain portability. Most of the presented systems are domain independent, except SYNDIKATE and VIKEF that require significant background knowledge.

Corpora modality. As in the case of ontology population, most of the systems focus on text corpora. Only VIKEF uses both text and images extracted from product catalogues. BOEMIE goes a step further and tries to combine various modalities, such as text, images, video and audio.

5 Evaluation

Evaluation in the context of ontology learning measures the quality of a learned ontology with respect to some particular criteria, in order to determine the plausibility of the learned ontology for the purposes it was built for. Approaches for evaluating learned ontologies can be distinguished into four major categories:

- "Gold standard" evaluation: the learned ontology is compared to a predefined (and usually manually-constructed) "gold standard" ontology.
- Application-based evaluation: the learned ontology is used in an integrated system and is implicitly evaluated through the evaluation of the complete integrated system.
- Data-driven evaluation: the learned ontology is evaluated through comparison with a data source covering the same domain as the learned ontology.
- Human evaluation: the learned ontology is examined/evaluated by domain experts based on predefined criteria, requirements, standards, etc.

An ontology can be evaluated at different layers, such as:

- Lexical, vocabulary or data layer. The evaluation here focuses on which concepts and instances have been included in the ontology and the vocabulary used to identify them.
- Relational layer. The evaluation of this layer deals with the relations between the concepts of the ontology:
  - Hierarchy, taxonomy. An ontology almost always includes hierarchical inclusion relations between its concepts. Thus, the evaluation of these taxonomic relations is very important.
  - Semantic relations. This layer of the ontology concerns other relations besides inclusion and can be evaluated separately.
- Structure, architecture. At this layer we assess whether the design of the ontology has followed some predefined strategies and if it is possible to further develop the ontology easily.
- Philosophical layer. At this level we evaluate the ontology against highly general ontological notions, drawn from the field of philosophical ontology. Thus, we want to decide whether a property of a concept is essential for the specific concept, whether a concept is easily identified among others, etc.

The majority of the evaluation approaches fall into the first category, i.e. gold standard evaluation, and the last category, i.e. evaluation by humans. These categories can also be combined and thus, they are commonly viewed as different sides of the same coin. In what follows, we will discuss these two categories in more detail, while we will give some insights regarding the application-based and the data driven evaluation.

5.1 "Gold Standard" Evaluation

During the "gold standard" evaluation, a learned ontology is compared to a predefined ontology which is considered to be "correct" and which is usually developed by domain

experts. A typical strategy for evaluating against a "gold standard" ontology is as follows: As a first step, the "gold standard" ontology must be created, an action usually performed manually by the domain experts. Then, the "gold standard" ontology is deliberately damaged, usually some concepts, relations and rules are removed from the ontology. At the third step, the pruned ontology is enriched with ontology learning. What is measured is the degree to which learning managed to reconstruct the pruned knowledge.

The comparison can be performed at various levels of the ontology. At the lexical level various string similarity measures can be used, such as the Levenshtein edit distance [65], in order to measure the similarity of concept and relation names. The evaluation at this point is usually performed by measuring Term/Lexical Precision and Term/Lexical Recall [90]. At the relational level, precision and recall can also be used, in order to determine how many identified relations are correct and how many relations of the "golden standard" ontology were found. An interesting approach is presented in [78] based on the notion of *semantic cotopy*. The semantic cotopy of a concept in a given taxonomy is the set of its super and sub-concepts. The overlap of the semantic cotopies of two concepts can be used as a similarity measure between the two concepts. The *taxonomic similarity of concepts* [33, 89] compares the relative placement of concepts in the ontology, based on their distance (shortest path) to other concepts. This set of distances can be used to compare the learned ontology to the "golden standard". Similar ideas have been proposed in [80], where the measures of Augmented Precision and Recall have been used to measure the similarity between two ontologies, taking into account the distance of each concept from the root. Treating the hierarchical backbone as a partition of instances, the evaluation can also be performed using the OntoRand index [14]. This approach measures the similarity between concepts of different hierarchies based either on their common ancestors, their distances in the hierarchy, and the overlap of their sets of instances. Finally, the work in [116] introduces the measures of P-value and R-value, which measure the similarity between ontologies based on the cotopy sets of the concepts and the distance of the concepts, when treated as probability distributions over their instances.

Evaluation against a "gold standard" is an interesting approach but it also has some drawbacks. Besides the obvious problem of constructing manually the "gold" ontology, this approach is somewhat "subjective". The "gold" ontology models a domain in a specific way, chosen by the domain experts that crafted the ontology. Bad evaluation results of a learned ontology do not necessarily mean that the learned ontology is wrong. It is possible that the learned ontology conceptualises the domain with a slightly different model or even captures information not addressed by the domain experts and thus not contained in the "gold" ontology. Thus, the same learned ontology may exhibit different scores with a slightly modified "gold" ontology. Finally, the results of this method are affected by the quality of the matching between the learned and the gold ontology. Thus, a correct ontology matching [36, 81] between the two ontologies is of particular importance, in order to derive meaningful conclusions and penalize accordingly the learned ontology. A combination of matching methods with the measures of P-value and R-value and a relevant discussion can be found in [117, 118].

5.2 Application-Based Evaluation

An important reason for creating an ontology is, among others, to be used in a specific application. Thus, a reasonable approach in evaluating an ontology is to evaluate the performance of the system that uses this ontology, assuming of course that the quality of the ontology plays a role in the performance of the system. Possible measurable objectives in the performance of a system may include low query computation effort, efficient reasoning with the ontology, correctness and completeness of the provided answers. A disadvantage of this evaluation approach is that the results are affected by the dependency of the system on the used ontology. In other words, the evaluation figures depend on the way the ontology is used by the system and the aspects of the ontology that are exploited. As a result, various ontology aspects may not be evaluated.

Although many papers report good results and successful applications of learned ontologies in various tasks, the first experimental conclusions are given in [48]. In this work, the ontology supported a speech recognition task and its role was to determine how closely related the meaning of two concepts was. The task was to assign the correct senses to ambiguous lexical items. These senses were provided by the ontology concepts. The accuracy of the senses assigned to the lexical items was measured against a gold standard.

Similarly, the peculiarities of application-based ontology evaluation are also examined in [87], in the task of tagging the ontological relations that hold between ontologically marked-up entities. This mark-up is obtained from a concept tagging system and constitutes a form of sense disambiguation, whereby the specific senses correspond to items of the ontology's vocabulary. The authors measure the accuracy of the tagging task with respect to ground truth. In addition, they notice various shortcomings of the learned ontology, when comparing the results against those obtained with a gold ontology.

5.3 Data-Driven Evaluation

An ontology may also be evaluated on existing data sources. These are usually collections of text documents, Web pages or dictionaries. The most important requirement for these data sources is to be representative and to cover the domain of the ontology.

Data-driven evaluation has been applied at the lexical [110], and the relational [16] layer of the ontology. This kind of evaluation is particularly suitable for evaluating ontologies learned from textual sources, since we can use a corpus of documents as facts to check whether these facts can be logically derived from the ontology. The metrics of precision and recall are applicable, since they provide an indication of the information that the learning algorithm has captured from the document collection.

Evaluation can also be performed using a set of domain-specific terms or concepts extracted from a corpus, which is compared against the concepts in the ontology. The overlap of the two sets measures the fit between the ontology and the corpus [16]. In the special case that the learned ontology is the result of a document clustering algorithm, it can be evaluated against pre-categorized document collections, such as the Reuters corpus.

Data-driven evaluation requires representative and domain-specific data. Consequently, a question usually arises regarding the choice of the datasets that will be used for the evaluation and how to measure whether they are representative or not.

5.4 Human Evaluation

In human evaluation, the ontology is assessed by human experts, based on desired predefined criteria. The evaluation can be performed by ontology experts, usually the ones that have designed the ontology learning system, users testing the ontology in applications or both. Features evaluated by ontology experts usually include ontology consistency, completeness or conciseness of the model implemented by the ontology. Users on the other hand are interested in the applicability of the ontology to a target task.

The OntoMetric [72, 73] methodology is an example of a principled ontology evaluation by the users of the ontology. A tool is introduced which helps users determine the suitability of an ontology for a particular application, allowing them to compare the importance of the ontology objectives and carefully evaluate its characteristics based on multiple criteria.

A set of ten criteria that can be used for ontology evaluation, are presented in [11]. These criteria cover various ontology aspects like richness, i.e. number of features used, and lawfulness, i.e. frequency of errors, interpretability, clarity, comprehensiveness, accuracy, relevance, authority and history.

A different view to human evaluation focuses on the competence of the ontology [42]. Competence is measured by constructing queries in such a manner that helps the evaluator to check if the ontology meets predefined requirements. A set of generic criteria that are proposed in this work include: (a) efficient reasoning, (b) minimality, i.e. if the ontology contains only the necessary information, (c) functional completeness, i.e. if the ontology can represent the required information to support some task, (d) generality, i.e. if it can be shared among domains, and (e) perspicuity, i.e. if it is easily understood by the users.

From a philosophical point of view, the notion of rigidity, introduced in [46], can be used to check the taxonomical structure of the ontology. Rigidity is based on the more abstract notion of essence. A concept is essential for an instance, if and only if the instance is necessarily an instance of this concept among all universes and at all times. This method is supported by the OntoEdit tool. An important drawback of this approach, though, is that much manual tagging of the concepts participating in the ontology is required. AEON [112] is a tool that aims at enhancing this process by automatically tagging the ontology.

5.5 Comparing the Various Approaches

In the above subsections, various approaches for evaluating a learned ontology have been presented. Each of them has different advantages and disadvantages. First, in order to make data-driven evaluation applicable to a particular domain, a substantial set of data about this domain is required. However, it is not always easy to acquire such data, making the approach difficult to adopt. Similarly, application-based evaluation requires the whole application to be evaluated by humans, which is also a difficult task. In addition, evaluation must be performed by multiple users, in order for the evaluation results to have some statistical significance.

Human-based evaluation is the most complete approach, as all aspects of a learned ontology can be measured and evaluated. However, this evaluation approach is difficult to automate and must be supported by special tools, which help humans in the evaluation. The "gold standard" evaluation is a convenient approach for evaluating ontologies that provides a clear view of the performance of the ontology learning, by comparing the ontology to a predefined gold one in an automated way, using various metrics and measures from the field of information retrieval. To our view, all other approaches evaluate ontologies in an abstract way, which is not always operational and meaningful especially if the ontology is decoupled from the application that uses it. In addition, the fact that the "gold standard" ontology is developed manually provides the ontology engineers the opportunity to develop an ontology that will score well in human-defined criteria and is also suitable for the domain of application. Thus, measuring the closeness of a learned ontology to this "gold" ontology performs also an implicit evaluation according to criteria that are used in human evaluation.

6 Conclusions

In this chapter, we have attempted a detailed presentation of the state-of-the-art on ontology learning, focusing on ontology population and enrichment. A generic framework has been proposed, to facilitate the comparative presentation of the most influential approaches found in the literature.

The comparative presentation of both population and enrichment systems leads to a number of interesting conclusions. The first observation concerns the modality of corpora the systems use to learn ontologies. While a significant amount of work has been performed on text corpora, work on other modalities is practically non-existent. A second observation is that work on learning from text relies heavily on linguistic preprocessing, especially syntactic analysis and exploitation of additional resources like thesauri and semantic hierarchies, such as WordNet. This is due to the fact that many practical systems employ a pattern-based approach, especially for the discovery of relations between concepts. Finally, despite the wide use of machine learning, many systems still require significant manual intervention, usually by ontology experts who make the final decisions for modifying the ontology. Systems that perform ontology population seem to require less manual intervention, effectively automating a large portion of the population process.

In this context, BOEMIE addresses a number of problems identified in the state of the art. In particular, BOEMIE works on multimedia corpora instead of text. The distinction made between "primitive" and "composite" concepts helps in making the information extraction process independent of the ontology structure. Also, BOEMIE puts significant effort in handling redundancy and maintaining the consistency of the ontology. The BOEMIE approach supports interaction with a domain expert rather than an ontology expert, as it presents the discovered knowledge in a natural language format. Finally, as the approach is domain-independent, it is expected to have a wide range of applications in different domains.

References

[1] Agichtein, E., Gravano, L.: Snowball: Extracting Relations from Large Plain-Text Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries (ACM DL), pp. 85–94 (2000)

[2] Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching Very Large Ontologies Using the WWW. In: Workshop on Ontology Construction of the European Conference of A.I., ECAI 2000 (2000)

[3] Ahmad, K., Davies, A., Fulford, H., Rogers, M.: What is a term? The Semi-Automatic Extraction of Terms from Text. John Benjamins Publishing Company, Amsterdam (1994)

[4] Alani, H., Sanghee, K., Millard, E.D., Weal, J.M., Lewis, P.H., Hall, W., Shadbolt, N.: Automatic Extraction of Knowledge from Web Documents. In: Proceeding of (HLT 2003) (2003)

[5] Alani, H., Sanghee, K., Millard, E.D., Weal, J.M., Lewis, P.H., Hall, W., Shadbolt, N.: Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation. In: Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003), Florida, USA (2003)

[6] Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R.: Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems 18(1), 14–21 (2003)

[7] Alfonseca, E., Ruiz-Casado, M., Okumura, M., Castells, P.: Towards Large-scale Non-taxonomic Relation Extraction: Estimating the Precision of Rote Extractors. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 49–56 (July 2006)

[8] Aussenac-Gilles, N., Biebow, B., Szulman, S. (eds.): EKAW 2000 Workshop on Ontologies and Texts (2000), http://CEURWS.org/Vol-51/CEUR

[9] Aussenac-Gilles, N., Maedche, A. (eds.): ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontolology Learning (2002), http://www.inria.fr/acacia/OLT2002

[10] Baroni, M., Bisi, S.: Using cooccurrence statistics & the web to discover synonyms in a technical language. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, vol. 5, pp. 1725–1728 (2004)

[11] Burton Jones, A., Veda Storey, C., Sugumaran, V., Ahluwalia, P.: A Semiotic Suite for Assessing the Quality of Ontologies. Data and Knowledge Engineering (2004)

[12] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

[13] Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process. Advances in Neural Information Processing Systems 16 (2004)

[14] Brank, J., Mladenic, D., Grobelnik, M.: Gold standard based ontology evaluation using instance assignment. In: Proceedings of the EON Workshop (2006)

[15] Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 203–207. Springer, Heidelberg (2002)

[16] Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of the International Conference on Language Resources and Evaluation (2004)

[17] Brewster, C., Iria, J., Zhang, Z., Ciravegna, F., Guthrie, L., Wilks, Y.: Dynamic Iterative Ontology Learning. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria (2007)
[18] Buitelaar, P., Handschuh, S., Magnini, B. (eds.): Proceedings of the ECAI 2004 Workshop on Ontologies, Learning and Population (2004)
[19] Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005) ISBN: 1-58603-523-1
[20] Buitelaar, P., Cimiano, P., Loos, B.: Bringing the Gap between Text and Knowledge. In: Workshop on Ontology Learning and Population (2006)
[21] Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based Information Extraction with SOBA. In: Proceedings of the International Conference on Language Resources and Evaluation, pp. 2321–2324. ELRA (May 2006)
[22] Buitelaar, P., Cimiano, P., Paliouras, G., Spiliopoulou, M.: Proceedings of the ECAI 2008 Workshop on Ontology Learning and Population (OLP3) (2008)
[23] Castano, S., Peraldi, I.S.E., Ferrara, A., Karkaletsis, V., Kaya, A., Möller, R., Montanelli, S., Petasis, G., Wessel, M.: Multimedia Interpretation for Dynamic Ontology Evolution. Journal of Logic and Computation (September 2008)
[24] Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., New York (2006)
[25] Ciravegna, F., Dingli, A., Petrelli, D.: Document Annotation via Adaptive Information Extraction. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 11-15 (2002)
[26] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S.: Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence 118, 69–113 (2000)
[27] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Gate: an architecture for Development of Robust HLT Applications. In: Proceedings of ACL (2002)
[28] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Phil. USA (2002)
[29] Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge (2005)
[30] Damerau, F.J.: Evaluating domain-oriented multiword terms from texts. Information Processing and Management 29(4), 433–447 (1993)
[31] Downey, O., Etzioni, D., Soderland, S., Weld, D.: Learning Text Patterns for Web Information Extraction and Assessment. In: Proceedings of the AAAI Workshop on Adaptive Text Extraction and Mining (2004)
[32] Drozdzynski, W., Krieger, H.-U., Piskorski, J., Schäfer, U., Xu, F.: Shallow processing with unification and typed feature structures – foundations and applications. Künstliche Intelligenz 1, 17–23 (2004)
[33] Ehrig, M., Haase, P., Stohanovic, N., Hefke, M.: Similarity for Ontologies – a Comprehensive Framework. In: Proceedings of the European Conference in Inf. Sys. (2005)
[34] Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, D., Shaker, T., Yates, A.: Web-Scale Information Extraction in KnowItAll (Preliminary Results). In: Proceedings of the 13th International World Wide Web Conference (WWW 2004), New York, pp. 100–110 (2004)

[35] Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, D., Shaker, T., Yates, A.: Unsupervised named-entity extraction from the Web: An experimental Study. Artificial Intelligence 165, 91–134 (2005)
[36] Euzenat, J., Pavel, S.: Ontology Matching. Springer, Heidelberg (2007)
[37] Faatz, A., Steinmetz, R.: Ontology Enrichment with texts from the WWW. In: Semantic Web Mining 2nd Workshop at ECML/PKDD-2002. Helsinki, Finland (2002)
[38] Fellbaum, C.: WordNet: An On-Line Lexical Database and Some of its Applications. MIT Press, Cambridge
[39] Faure, D., Nedellec, C., 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)
[40] Faure, D., Poibeau, T.: First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: Proceedings of the ECAI 2000 Workshop on Ontology Learning (OL 2000) (2000)
[41] Fortuna, B., Mladevic, D., Grobelnik, M.: Visualization of Text Document Corpus. In: ACAI 2005 Summer School (2005)
[42] Fox, M.S., Barbuceanu, M., Gruninger, M., Lin, J.: An Organization Ontology for Enterprise Modelling. MIT Press, Cambridge (1998)
[43] Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: The c-value/nc-value method. International Journal on Digital Libraries 3(2), 115–130 (2000)
[44] Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing. Int. J. of Human and Computer Studies 43, 907–928 (1994)
[45] Gómez-Pérez, A., Manzano-Macho, D.: A survey of ontology learning methods and techniques. Onto-web IST Project, Deliverable 1.5, http://www.ontoweb.aifb.uni-karlsruhe.de/Members/ruben/Deliverable%201.5
[46] Guarino, N., Welty, C.: Evaluating ontological decisions with ontoclean. Communications of the ACM 45(2), 61–65 (2002)
[47] Gupta, K.M., Aha, D., Marsh, E., Maney, T.: An Architecture for engineering sublanguage WordNets. In: Proceedings of the First International Conference On Global WordNet, pp. 207–215. Central Institute of Indian Languages, Mysore (2002)
[48] Gurevych, I., Malaka, R., Porzel, R., Zorn, H.: Semantic coherence scoring using an ontology. In: Proceedings of the HLT/NAACL (2003)
[49] Haase, P., Stojanovic, L.: Consistent Evolution of OWL Ontologies. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 182–197. Springer, Heidelberg (2005)
[50] Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., Sure, Y.: A Framework for Handling Inconsistency in Changing Ontologies. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 353–367. Springer, Heidelberg (2005)
[51] Haase, P., Völker, J.: Ontology Learning and Reasoning — Dealing with Uncertainty and Inconsistency. In: da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2005 - 2007. LNCS (LNAI), vol. 5327, pp. 366–384. Springer, Heidelberg (2008)
[52] Hahn, U., Marko, K.G.: Ontology and Lexicon Evolution by Text Understanding. In: Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT 2002), Lyon, France (2002)

[53] Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France (1992)
[54] Harris, Z.: Mathematical Structures of Language. John Wiley & Sons, Chichester (1968); Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 268–275 (1990)
[55] Iria, J., Brewster, C., Ciravegna, F., Wilks, Y.: An Incremental Tri-Partite Approach To Ontology Learning. In: The 5th International Conference on Language Resources and Evaluation, May 24-25-26, pp. 24–25 (2006)
[56] Iwanska, L.M., Mata, N., Kruger, K.: Fully Automatic Acquisition of Taxonomic Knowledge from Large Corpora of Texts, pp. 335–345. MIT/AAAI Press (2000)
[57] Kietz, J.U., Maedche, A., Volz, R.: A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In: Proceedings of the ECAW 2000 Workshop Ontologies and Text, Juan-Les-Pins, France (2000)
[58] Kim, S.-S., Son, J.-W., Park, S.-B., Park, S.-Y., Lee, C., Wang, J.-H., Jang, M.-G., Park, H.-G.: OPTIMA: An Ontology Population System. In: 3rd Workshop on Ontology Learning and Population (July 2008)
[59] 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 (SAAKM 2002), the 15th European Conference on Artificial Intelligence (ECAI 2002), Lyon, France, pp. 1–6 (2002)
[60] Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., Cofino, T.: Glossary extraction and utilization in the information search and delivery system for IBM Technical SupportΣ. IBM System Journal 43(3) (2004)
[61] Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. Journal of Biomedical Informatics 37, 512–526 (2004)
[62] Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML) (2001)
[63] Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104, 211–240 (1997)
[64] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: The Fifth International Conference on Systems Documentation, ACM SIGDOC (1986)
[65] Levenshtein, I.V.: Binary codes capable of correcting deletions, insertions and reversals. Cybernetics and Control Theory 10(8), 707–710 (1966)
[66] Li, W., McCallum, A.: Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In: Proceedings of the 23rd Internationcal Conference on Machine Learning, pp. 577–584 (2006)
[67] Li, W., Blei, D., McCallum, A.: Nonparametric Bayes Pachinko Allocation. In: Uncertainty in Artificial Intelligence (2007)
[68] Lin, D., Pantel, P.: Induction of semantic classes from natural language text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 317–322 (2001)

[69] Lin, D., Pantel, P.: Dirt - Discovery of Inference Rules from Text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323–328 (2001)
[70] Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the International Conference on Computational Linguistics (COLING), pp. 577–583 (2002)
[71] Lisi, F.A.: Principles of Inductive Reasoning on the Semantic Web: A Framework for Learning in AL-Log. In: Fages, F., Soliman, S. (eds.) PPSWR 2005. LNCS, vol. 3703, pp. 118–132. Springer, Heidelberg (2005)
[72] Lozano-Tello, A., Gomez-Perez, A., Sosa, E.: Selection of Ontologies for the Semantic Web, pp. 413–416. Springer, Heidelberg (2003)
[73] Lozano-Tello, A., Gomez-Perez, A.: Ontometric: A method to choose the appropriate ontology. Journal of Database Management. Special Issue on Ontological Analysis, Evaluation, and Engineering of Business Systems Analysis Methods 15(2), 1–18 (2004)
[74] Maedche, A., Staab, S.: Semi-Automatic Engineering of Ontologies from Text. In: Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (2000)
[75] Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. In: Proceedings of ECAI 2000. IOS Press, Amsterdam (2000)
[76] Maedche, A., Staab, S., Nédellec, C., Hove, E. (eds.): IJCAI 2001 Workshop on Ontology Learning (2001), http://CEUR-WS.org/Vol-38/CEUR
[77] Maedche, A., Staab, S.: Ontology learning for the Semantic Web. IEEE Journal on Intelligent Systems 16(2), 72–79 (2001)
[78] Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
[79] Maedche, A., Staab, S.: Ontology Learning. In: Handbook on Ontologies (2004)
[80] Maynard, D., Peters, W., Li, Y.: Metrics for evaluation of ontology based information extraction. In: Proceedings of the EON 2006 Workshop (2006)
[81] Meilicke, C., Völker, J., Stuckenschmidt, H.: Learning Disjointness for Debugging Mappings between Lightweight Ontologies. In: Gangemi, A., Euzenat, J. (eds.) EKAW 2008. LNCS (LNAI), vol. 5268, pp. 93–108. Springer, Heidelberg (2008)
[82] Mima, H., Ananiadou, S., Nenadic, G.: The atract workbench: Automatic term recognition and clustering for terms. In: Matoušek, V., Mautner, P., Mouček, R., Tauser, K. (eds.) TSD 2001. LNCS (LNAI), vol. 2166, p. 126. Springer, Heidelberg (2001)
[83] Mimno, D., Li, W., McCallum, A.: Mixtures of Hierarchical Topics with Pachinko Allocation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 633–640 (2007)
[84] Morin, E.: Automatic Acquisition of Semantic Relations Between Terms from Technical Corpora. In: Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering - TKE 1999 (1999)
[85] Navigli, R., Velardi, P.: Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 1–9 (July 2006)
[86] Navigli, R., Velardi, P.: Ontology Enrichment Through Automatic Semantic Annotation of On-Line Glossaries. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 126–140. Springer, Heidelberg (2006)
[87] Porzel, R., Malaka, R.: A task-based approach for ontology evaluation. In: ECAI 2004 Workshop on Ontology Learning and Population (2004)

[88] Quinlan, J.R.: Learning logical definitions from relations. Machine Learning 5, 239–266 (1990)
[89] Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 17–30 (1989)
[90] Sabou, M., Wroe, C., Goble, C., Stuckenschmidt, H.: Learning domain ontologies for semantic web service descriptions. Journal of Web Semantics 3(4) (2005)
[91] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
[92] Sangun, P., Juyoung, K., Wooju, K.: A Framework for Ontology Based Rule Acquisition from Web Documents in Web Reasoning and Rule Systems (2007)
[93] Sclano, F., Velardi, P.: TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In: 9th Conf. on Terminology and Artificial Intelligence TIA 2007, Sophia Antinopolis (October 2007)
[94] Schütze, H.: Word space. Advances in Neural Information Processing Systems 5 (1993)
[95] Schutz, A., Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 593–606. Springer, Heidelberg (2005)
[96] Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
[97] Shamsfar, M., Barforoush, A.A.: An Introduction to HASTI: An Ontology Learning System. In: Proceedings of 6th Conference on Artificial Intelligence and Soft Computing (ASC 2002), Banff, Canada (June 2002)
[98] Shamsfard, M.: Designing the ontology learning Model, Prototyping in a Persian Text Understanding System, Ph.D. Dissertation, Computer Engineering Dept., AmirKabir University of Technology, Tehran, Iran (January 2003)
[99] Shamsfard, M., Barforoush, A.A.: The state of the art in ontology learning: a framework for comparison. Knowl. Eng. Rev. 18(4), 293–316 (2003), DOI: http://dx.doi.org/10.1017/S0269888903000687
[100] Shamsfar, M., Barforoush, A.A.: Learning Ontologies from Natural Language Texts. International Journal of Human-Computer Studies (60), 17–63 (2004)
[101] Schulte im Walde, S.: Clustering Verbs Semantically According to their Alternation Behaviour. In: Proceedings of the 18th International Conference on Computational Linguistics (COLINGS), pp. 747–753 (2000)
[102] Snow, R., Jurafsky, D., Ng, A.Y.: Semantic Taxonomy Induction from Heterogeneous Evidence. In: ACLY 2006 (2006)
[103] Snow, R., Jurafsky, D., Ng, A.Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery. In: Proceedings of Advances in Neural Information Processing Systems (2004)
[104] Specia, L., Motta, E.: A hybrid approach for extracting semantic relations from texts. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 57–64 (July 2006)
[105] Staab, S., Maedche, A., Nedellec, C., Wiemer- Hastings, P. (eds.): Proceedings of the Workshop on Ontology Learning (2000), http://CEUR-WS.org/Vol-31/CEUR
[106] Suchanek, F.M., Ifrim, G., Weikum, G.: LEILA: Learning to Extract Information by Linguistic Analysis. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge – OLP 2006, Sydney, Australia, pp. 18–25 (July 2006)

[107] Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
[108] Yang, H., Callan, J.: A Metric-based Framework for Automatic Taxonomy Induction. In: ACL 2009 (2009)
[109] Yarowsky, D.: Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In: COLING 1992, Nantes (1992)
[110] Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies. IOS Press, Amsterdam (2005)
[111] Velardi, P., Cucchiarelli, A., Petit, M.: A Taxomony learning Method and its Application to Characterize a Scientific Web Community. IEEE Transaction on Data and Knowledge Engineering (TDKE) 19(2), 180–191 (2007)
[112] Völker, J., Vrandečić, D., Sure, Y.: Automatic Evaluation of Ontologies (AEON). In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 716–731. Springer, Heidelberg (2005)
[113] Weber, N., Buitelaar, P.: Web-based Ontology Learning with ISOLDE. In: Proceedings of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference, USA (2006)
[114] Wei, W., Barnaghi, P.: Probabilistic Topic Models for Learning Terminological Ontologies. Transaction on Knowledge and Data Engineering 22(7), 1028–1040 (2010)
[115] Zavitsanos, E., Paliouras, G., Vouros, G.: Ontology Learning and Evaluation: A survey. Technical report, DEMO-(2006-3), NCSR Demokritos, Athens, Greece (2006)
[116] Zavitsanos, E., Paliouras, G., Vouros, G.: A Distributional Approach to Evaluating Ontology Learning Methods Using A Gold Standard. In: 3rd Ontology Learning and Population Workshop, ECAI 2008 (2008)
[117] Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S.: Learning Subsumption Hierarchies of Ontology Concepts from Texts. Web Intelligence and Agent Systems: An International Journal 8(1), 37–51 (2010)
[118] Zavitsanos, E., Paliouras, G., Vouros, G.A.: Gold Standard Evaluation of Ontology Learning Methods Through Ontology Transformation and Alignment. Transactions on Knowledge and Data Engineering (2010) (to appear)