Unsupervised and Domain Independent Ontology Learning. Combining Heterogeneous Sources of Evidence

David Manzano-Macho*, Asunción Gómez-Pérez*, Daniel Borrajo†

*Facultad de Informática, Universidad Politecnica de Madrid
Campus de Montegancedo, sn, 28660 Boadilla del Monte, Spain
{dmanzano,asun}@fi.upm.es
†Departamento de Informática, Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911 Leganés, Spain
dborrajo@ia.uc3m.es

Abstract

Acquiring knowledge from the Web to build domain ontologies has become a common practice in the Ontological Engineering field. The vast amount of freely available information allows collecting enough information about any domain. However, the Web usually suffers a lack of structure, untrustworthiness and ambiguity of the content. These drawbacks hamper the application of unsupervised methods of building ontologies demanded by the increasingly popular applications of the Semantic Web. We believe that the combination of several processing mechanisms and complementary information sources may potentially solve the problem. The analysis of different sources of evidence allows determining with greater reliability the validity of the detected knowledge. In this paper, we present GALEON *(General Architecture for Learning Ontologies)* that combines sources and processing resources to provide complementary and redundant evidence for making better estimations about the relevance of the extracted knowledge and their relationships. Our goal in this paper is to show how combining several information sources and extraction mechanisms is possible to build a taxonomy of concepts with a higher accuracy than if only one of them is applied. The experimental results show how this combination notably increases the precision of the obtained results with minimum user intervention.

1. Introduction

Building minimally-supervised and domain-independent methods for the construction of ontologies is still an open issue. Nowadays, ontologies are used in a wide range of fields such as information extraction, natural language processing, knowledge engineering, etc. The great volume of currently available information demands for technology that automates as much as possible the acquisition process, allowing a rapid creation of new ontologies from the available sources. The Ontology Learning (OL) community has proposed several methods of acquiring ontological knowledge from of a great variety of textual resources and through the application of several techniques (Gómez-Pérez and Manzano-Macho, 2005). There are several techniques that can potentially be used to achieve a domain-independent and unsupervised OL method. Among other possibilities, statistical and syntactic analysis (Downey et al., 2004), pattern-based extraction (Ruiz-Casado et al., 2007), clustering (Cimiano et al., 2005) and other Machine Learning techniques are commonly used.

Not all available techniques are equally applicable to all types of domains and sources. They need to be carefully selected to achieve the desired results. Thus, the use of patterns has shown its utility in many applications for extracting knowledge from text. However, these patterns usually appear scattered though the corpus, so they present a low frequency of appearance and the information extracted from them may be biased. Besides, the context where the pattern can be applied is limited to a sentence, while relations between relevant terms to the domain can really appear in broader contexts such as a document or across the whole corpus. The statistical analysis compiles relevant terms through a corpus, and also detects frequent collocations among them. Collocations focus on detecting attached terms, and often suffer similar problems to the pattern analysis. Clustering methods group related terms and can potentially organise them into a hierarchical structure. The main difficulty in using them is to define the algorithm for joining clusters. However, the knowledge that clustering techniques acquire is often too generic to be used in very specific domains. In many cases, the precision of all the text-based methods depends on the quality of the texts used. Therefore, we can decrease efficacy when we apply them to generic texts extracted from the Web, whose quality is unknown. Web documents usually follow a layout to show their content and link information provided among several of them, which is rarely taken into account in the OL processes. So, another set of techniques use HTML features, which has been useful to improve the accuracy for Information Extraction tasks (Craven et al., 2000).

Individually, all the approaches mentioned provide modest levels of accuracy in their results. But all suffer similar shortcomings when trying to become domain independent and unsupervised, as we have explained before. The combination of several sources of information and extraction techniques may partially overcome this problem, because redundant information (as coming from several of these techniques analysing the same input) represents a measure of the information relevance and trustiness for a certain domain (Kurshmerick et al., 2005). An statement can be compared and contrasted across several documents using multiple evidences provided by different methods. In addition, a method can provide valuable information that the other methods are not able to detect. So, our goal in this paper is to show how through the combination of statistical, syntactic, semantic and visual layout analysis of HTML docu-

ments for building a taxonomy of concepts we obtain better results than when using each of them individually. The experimental results have been obtained by GALEON combining the aforementioned types of analysis in the domains of (*Universities* and *Economics*).

2. Conceptual Design of the General Architecture for Learning Ontologies

The *General Architecture for Learning Ontologies* (GALEON) is an open, extensible and domain-independent architecture that automates the process of building or extending domain ontologies. GALEON allows to learn the new ontology or parts of it, using the given sources with minimum user intervention. As shown in Figure 1, the architecture is composed of six main phases: *processing, acquisition, action, consolidation, evaluation* and *knowledge augmentation*.

Once the user has provided the set of sources that sufficiently describes the domain of the ontology, these are morphosyntactically analysed while at the same time the document structure is being processed. Afterwards GALEON extracts and selects the core terminology, called in this context *candidate elements* (CE), out of the processed sources. For example, CEs found on the experiments with respect to the domains of *Economics* and *Universities* were *economic school of thought* or *computer science department*. Every CE has attached its morphosyntactical, textual, statistical, and visual information produced as a result of the processing phase. These features are augmented and refined with information through all phases creating the corresponding *hypotheses*. A hypothesis is a probabilistic statement that describes the relevance of a CE and its relationships with other elements. It reflects what the system believes to be true at each stage and also allows to know what assumptions have been used to generate the element. Some hypotheses can reinforce others by providing additional sources of evidence about their correctness and suitability. Usually, their redundancy represent a good sign about its relevance and validity for the domain. After each phase, the proposed hypotheses are evaluated to compute their relevance. Thus, the relevance of each hypothesis is justified in terms of the sources combined and the degree of confidence in the methods responsible of its generation, that provide its evidential support. Sometimes, the information provided by a hypothesis may contradict others. For instance, a hypothesis might propose to define an element as a subclass of another element, and another hypothesis might propose the second element to be the subclass of the first element. In these cases, the hypotheses are filtered, selecting only the hypothesis with the highest relevance factor. At the end of the process, the remaining hypotheses relate CEs to actions that include those elements into the ontology. They can be included as *"X is a concept", "X is a subclass-of Y", "X is an attribute-of Y", "X is an instance-of Y"*, or *"X is related through Y with Z"*. The evaluation in the final layer allows to estimate the level of precision and recall achieved through the whole process.

Each layer is composed of a set of independent operators (*OPs*) that take as input the set of hypotheses produced in the previous layer, and generate as output another set of hypotheses for the next layer. While the operators at the acquisition layer relate CEs based on their syntactic, semantic and visual relationships, the action operators decide, based on the hypotheses generated at the acquisition layer, about the type of knowledge the CEs can be and detect hierarchical relationships between them. All operators are associated with a priority value that defines the degree of credibility of the assumptions they generate. To carry out the experiments, we have considered that all the operators have the same priority. One of the characteristics of the architecture is its extensibility, given that incorporating a new acquisition technique simply implies adding a new operator that adds as hypothesis the information extracted by the new technique.

GALEON also accepts as an optional input an existing domain ontology that will be enriched as a result of the learning process. The provided knowledge acts as background knowledge of the domain and is used by the operators to improve the accuracy of the learning process. The following section describes the layers of the architecture and how through them CEs are organised into hierarchies of concepts.

3. Transforming data into knowledge: Layers of the architecture

In this section we sketch the conceptual design of the proposed architecture that allows to extract the core terminology about the domain and to detect the relationships that appear among terms in order to build a hierarchy of concepts.

3.1. Processing Layer

GALEON performs the following types of processing over the selected sources: statistical, morphosyntactical,[1] semantic (accessing to WordNet) analysis, and HTML parser to analyse the visual layout of the documents.

3.1.1. Creation of the candidate elements repository

The textual content of every document is analysed and split into chunks by means of a shallow parser. The creation of the CEs relies on the analysis of the detected chunks. For every chunk, a new CE is created. The CEs are identified using the stem of their corresponding chunks and their syntactical category, at this moment being just noun or verb (other categories are not considered as valid CEs for the purposes of the work herein presented). Examples of extracted CEs are *computer science department* or *Public Economy*. Besides, a headword is also associated to every element, that refers to the central word that summarises the main meaning of the element. In the previous example, *department* and *economy* are the corresponding headwords. While the CEs are created, GALEON collects the statistical distribution of each term through the corpus. Through this process, the frequency of appearance, number of documents and sentences where the CE has appeared in are collected for every CE. These values will be used in order to have an idea about how relevant an element may be from the statistical point of view.

[1] In our experiments, we have used the OpenNLP package available at http://opennlp.sourceforge.net

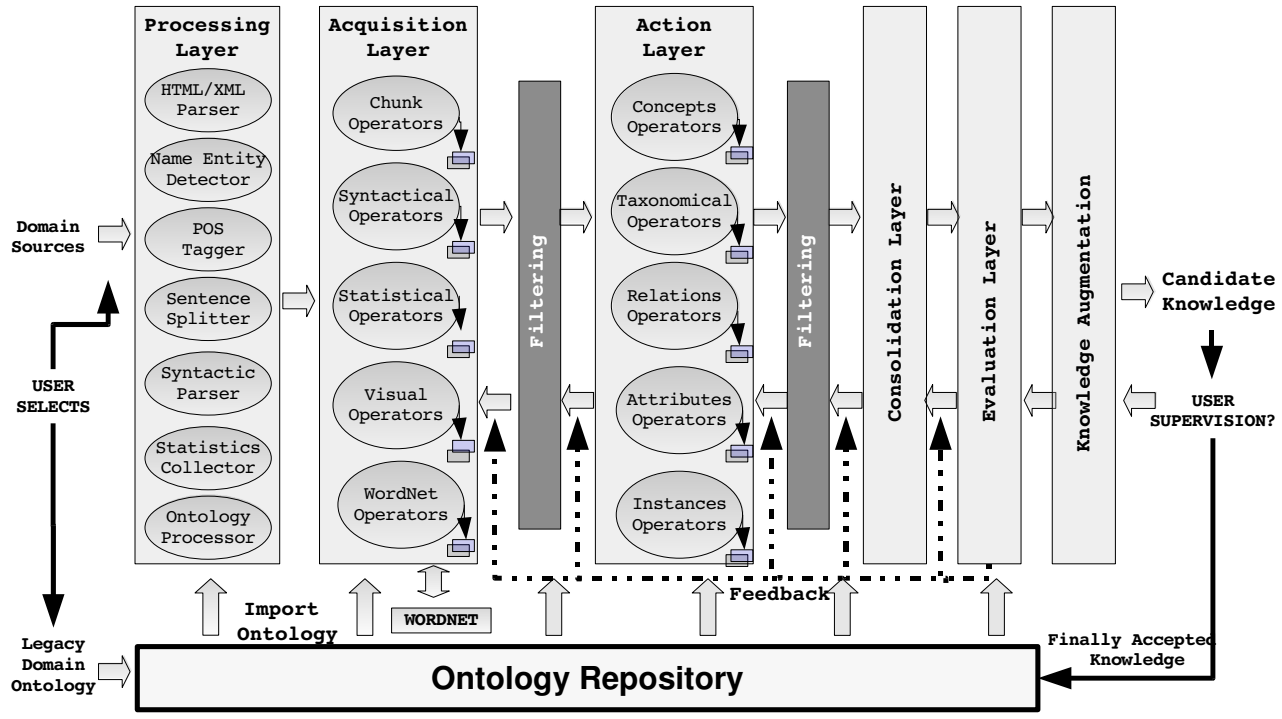


Figure 1: Design of GALEON: a General Architecture for Learning Ontologies.

Once the analysis of the sources has been completed, CEs are semantically analysed using WordNet to create clusters of related terms. The most probable WordNet synset is associated to each element. This process is performed using several similarity measures based on WordNet inspired in (Resnik, 1995). For each CE, the set of WordNet synsets is gathered. If the element does not appear in WordNet, the headword is used instead. In case of multiple meanings, for each of them the procedure computes the number of CEs that are hyponyms and hypernyms of the current meaning, selecting the synset with a highest number of matches. The method computes the distance over which the synsets appear within WordNet. Thus, the shorter the distance between the synsets, the greater would be their relationship to the domain. And in case of multiple candidates, GALEON computes the similarity using the similarity measures implemented in the WordNet::Similarity package.[2] Using the synset information, GALEON connects synonym elements, while similar elements are clustered together. The most relevant term to the domain, from the statistical point of view, will be assigned as the cluster label. CEs are also clustered through the application of the distributional hypothesis (Harris, 1968). This process reduces noticeably the number of CEs to be considered through the following layers.

3.1.2. Collecting facts about the candidate elements

While the elements are generated, different *facts* about them are collected based on the aforementioned analysis perspectives performed at the early beginning of the processing layer. In this context, facts are properties that describe a CE. The main facts collected for every CE and the corresponding predicates that give access to such information throughout the architecture are:

- Statistical measures. Using the statistical values associated to every element, GALEON calculates the idf (Jones, 1973) and $tf \cdot idf$ (Salton and Buckley, 1987).

- Frequent co-ocurrencies among CEs. This analysis compiles frequent collocations among CEs in a sentence and within a document (*in-the-same-doc? (e_1 e_2)*, *in-the-same-sent? (e_1 e_2)* and *appears-together? (e_1 e_2)*).

- Name entity recognition. Using the OpenNLP NE module some CEs are considered as potential entities (*is-an-entity? (e)*), such as personal name, organisation, location, etc.

- Textual characteristics analysis. It aims to measure how often a CE appears in uppercase through the corpus compared to its total frequency of appearance. This measure allows to estimate the probability that a CE can be considered as a type of entity on the basis of how it has been written throughout the whole corpus (*appears-in-uppercase? (e)*).

- Visual properties analysis. Relevant information usually appears within the document title (*in-title? (e)*), keywords or meta (*in-content? (e)*) and headers (*in-header? (e)*). Also, highlighted information usually appears in bold or with a different type-case (*is-highlighted-element? (e)*). All these features are cap-

[2] http://wn-similarity.sourceforge.net

tured as a source of evidence for relevant domain elements. Besides, following the notion of Semantic Textual Units (STUs) proposed in (Buyukkokten et al., 2001) (text that appears between a couple of HTML tags is somewhat related, which extents the notion of context that usually is a sentence or a window of words), the HTML structure is also used to detect frequent collocations (*in-the-same-unit? (e_1 e_2)*).

3.1.3. Filtering elements

In order to generate the core terminology that describes the domain, it is necessary to remove those CEs that are less related to the domain. GALEON implements two filtering methods. The first one analyses the frequency of appearance of every CE. Thus, CEs that have appeared less than or equal to two times in the whole corpus are removed because it is difficult to decide whether they are relevant or not. The second approach compares the idf measure that a CE has in the corpus and in the Web (we called this value idf_{web}). The CEs have to be single words because the repository does not include multi-words. To perform this process, GALEON includes a repository [3] where a set of words have associated their idf in the Web. The formulation is as follows:

$$\forall CE_i \mid if\ idf(CE_i) >> idf_{web}(CE_i) \rightarrow remove(EC_i)$$

The assumption here is that a relevant element to the domain has to appear in more documents into the domain sources than in others non directly related to the domain. The corpus has to be vast enough to assume that the knowledge for understanding the domain is written down somewhere in the selected documents. In contrast, elements that appear more frequently out of the domain may reflect common words of the language. In our experiments, the CE *university* has associated an idf value of 2.72 and a idf_{web} of 3.74 what may mean that it is potentially relevant to the *universities* domain.

3.2. Acquisition Layer

The acquisition *Ops* look for relations that may appear between CEs at several levels covering all the processing perspectives previously mentioned: *chunk*, *statistical*, *syntactical*, *visual*, and *semantic*. Thus, they generate hypotheses that relate the elements with their visual places, syntactical and semantic relations to other elements, etc. Besides, *OPs* have to analyse the statistical values attached to each CE specifying its relevance to the domain. The reason to distinguish between the aforementioned facts and the hypotheses is that facts describe observable data about the elements and the hypotheses are assumptions about them and their relationships to others.

3.2.1. Chunk level

These *OPs* analyse each CEs to decide whether an element has any relation to others, based on its syntactic and/or textual structure. This set of *OPs* is organised into three groups. The first group aims at knowing whether an element *specifies* or *generalises* the meaning of other elements within the CE repository. If an element specifies/generalises another, the *OPs* create the corresponding hypothesis. The *OPs* analyse the internal structure of every CE looking for their possible variations (at morphological, lexical, syntactic and semantic levels), following a similar approach to the one presented in (Jaquemin, 2001). Another complementary approach to perform the process relies on using the headword heuristics based on which a CE can be a hyponym of its headword. Thus, in the case of a CE that is no entity, does not always appear in capital letters, is an NP and is a multi-word element these *OPs* create a hypothesis stating that the CE is an hyponym of its headword. These *OPs* find, for example, that *computer science department* or *Public Economy* are candidate subtypes of *department* and *Economy* respectively. Analysis of the modifiers attached to some elements may denote the same kind of relationship, as it is the case with the elements modified by any adjective. The last group of *OPs* detects elements that are *composed* of other elements. This is the case, for example, of elements like *department web page*, that it is composed of *department* and *web page*. Finally, GALEON is also able to detect, at this level, that elements such as *Department of Computer Science* or *Dept. of Computer Science* are synonyms of *Computer Science Department*. The predicates to access to this information are: *is-hypernym-of? (e_1 e_2)*, *is-hyponym-of? (e_1 e_2)*, *is-a-composed-element? (e)* and *appears-attached-to? (e_1e_2)*.

3.2.2. Statistical level

Once most of the non-domain related CEs have been removed as a result of the previous filtering process, all remaining elements are supposed to contain the most relevant terminology to the domain. The purpose of the statistical *OPs* is to distinguish among elements that are general knowledge to the domain to those that are more specific. The assumption is that using a domain corpus, the general terminology will appear frequently, being the core vocabulary to the domain. However, the most specific terminology will be found scattered throughout the corpus. For example, in the context of *education*, the words *professor, university, college* will have a high frequency of appearance, but most of the names of the authors of the web pages belonging to the selected corpus will have a significantly lower frequency. Thus, the idea is to sort the elements according to their relevance. The hypothesis generated will allow knowing whether an element is more general than another. Thus, in the subsequent phases of GALEON more general terminology will be placed in the learned taxonomies at a higher level. This process is performed using the available statistical measures (attached to every element). Thus, the elements are grouped into two clusters; one cluster contains the *general knowledge* to the domain (higher level concepts), and the other cluster contains the more *specific knowledge* (candidates to be leaf concepts in a taxonomy). The predicates that give access to this information are: *is-general-knowledge? (e)*, *is-specific-knowledge? (e)*, and *is-more-general-than? (e_1e_2)*.

3.2.3. Syntactical level

Using a set of syntactical patterns, this type of *OPs* aims to find syntactical relations among the CEs that appear within

[3] http://elib.cs.berkeley.edu/docfreq/index.html

a sentence. Each *OP* implements a type of regular expression, that is based on syntactic features (syntactical annotation). Some interesting relations that are captured are *hypernym*, *hyponym*, *partOf*, or which verb *links* two elements in a sentence. Besides, how the elements appear within a sentence can reflect some type of relationship among them. The experiments herein presented only use *OPs* that implement the set of Hearst's patterns (Hearst, 1992). As with the rest of *OPs*, the set of patterns can be easily extended with new patterns defined in terms of the syntactical structure by adding new *OPs*. The corresponding predicates are in this case: *is-hypernym-of? ($e_1 e_2$)* and *is-hyponym-of? ($e_1 e_2$)*.

3.2.4. Visual level

The visual *OPs* rely on the assumption that the places where CEs appear within an HTML document may denote some kind of relationship among them. For example, texts placed in a list or in a table denote interconnected information. Using the HTML structure, the STUs extracted during the processing layer are arranged into a hierarchy of units. The notion of a hierarchical order among the different units within a document can show an implicit structuring of the information within the document, which may suggest certain types of hierarchical relationships among different units. The main relationships detected using the visual level which may indicate the existence of a hierarchical relationship between the CEs involved are the following. The list also includes the predicates that allow to access the generated hypotheses.

- *Belong to*. Appears between CEs that appear with the title and meta content of the Web document. In these tags, the author of the document usually describes the content of the document, and thus they may have a close meaning. (predicate *x-belongs-to-y? (e_1 e_2)*).
- *Hierarchical context*. This relationship appears between CEs that appear one of them within the title and each of the headers that are part of the document, or between CEs that appear within a header and others that are within another header with higher level of indentation. (*in-a-hierarchical-link? (e_1 e_2)*).
- *In the Same Hierarchical Context*. This relates CEs that appear within two headers of the document with the same level of indentation. This also occurs between CEs contained within two consecutive list items of an HTML list. (*in-the-same-hierarchy? (e_1 e_2)*).
- *Thematic Context*. This relates CEs that appear within list items, table captions and textual paragraphs and the text that appears in the header above them. This means that the content below the *headers* has a connection to the CEs that appear within the *headers*.
- *Document Link*. Relates CEs that are linked by means of hyper-link included within the document. The relation can appear within the same document or spread over several documents of the corpus. In the first case, the target CEs of the relation are those included within the header section where the link points to. For the second case, the target CEs are considered those that appear within the head section of the document. (*does-x-link-to-y? (e_1 e_2)*).

The analysis of the visual layout offers more types of relationship than the aforementioned. However, these are the most relevant to detect hierarchical relations between CEs. The figure (2) shows a typical definition of an acquisition *OP*. In that example, the *OP* has a name that identifies the *OP* and an internal variable *op* that defines the *OP*. That *OP* includes a *selection-method* which allows to consider only those CEs with enough statistical relevance. Thus, as considered STUs may contain a long text (i.e. long paragraph within a header), they may include several CEs. Only the selected ones will be analysed by the *OP*. In this example, the *OP* selects from each STU only those CEs that have appeared in several documents of the corpus. The *OP* implements a set of rules (*clauses*) which contain a set of *conditions* and the corresponding *actions*. The *conditions* combine facts about the elements ($?e1$ and $?e2$). If a pair of elements presents those characteristics, the *OP* will execute the corresponding actions. Each *action* generates a hypothesis about the existence of a relationship within the HTML schema. The new hypothesis is associated with the documents from which it has been generated, together with which facts have been combined to create it. This information is used to evaluate how relevant the hypothesis is within the provided sources. The conditions require that one of the elements should be in the title ($?e1$) while the other in a header ($?e2$) of the same document. In addition, this particular *OP* dismisses those elements which are name entities. Finally, the element ($?e1$) has also to appear in bold or in type-case through the corpus (the more documents where the fact appears in, the higher its relevance will be).

```
(html-operator hierarchical-rels-html
:vars (?e1 ?e2)
:selection-method max-idf-based-relevance
:clauses
((:conditions
   (AND
    (in-header? ?e2)
    (in-title? ?e1)
     (not (is-an-entity? ?e1))
     (not (is-an-entity? ?e2))
    (is-highlighted-element? ?e1))
  :actions
   (html-hyp ?e1 :hierarchicalContOf ?e2
     :where docs :evidences facts)) ))
```

Figure 2: Detecting visual relationships among CEs.

3.2.5. Semantic operators

Once the CEs have been annotated using WordNet synsets, the semantic *OPs* look for semantic relations among them using the WordNet data. The closer the synsets of a pair of elements appear within the WordNet structure, the more probable the existence of that relationship among those elements is. This type of hypotheses acts as another source of evidence to the hypotheses generated by the syntactical

OPs. The generated hypotheses based on WordNet consider how far the elements appear into the WordNet structure. The predicates to access the generated hypotheses are the same ones as those presented for the syntactical *OPs*.

3.3. Evaluation of the acquired hypotheses

The extracted facts and the generated hypotheses have to be evaluated in order to know how relevant they are to the domain. The measurement of the relevance is based on the previously mentioned assumption that the redundancy of information can represent a measure of its relevance to and trustiness for the domain. Thus, the more documents where the fact or the hypothesis appears in, the higher its relevance will be. Therefore, the relevance of a fact is measured using the following *Relevance Formula (rf)*:

$$rf(f_i) = \frac{log(d_i)}{log(D)} \; / \; rf(f_i) \; \epsilon \; [0,1] \qquad (1)$$

which reflects these ideas. $rf(a_i)$ measures the relevance of the assumption f_i to the current problem, where d_i is the number of documents from which f_i has been generated, and D is the size of the input document collection. The bigger the evidence is, the bigger its relevance is. The values are normalised between $[0, 1]$. The generated hypotheses are evaluated following the same schema, but in this case, the measure has to consider that a hypothesis may have been generated by several *OPs*. So, the formula 2 includes these requirements.

$$r(h_k^{acq}) = \frac{\sum_{i=1}^{N} op_i^{acq} \frac{\sum_{j=1}^{M} rf(fact_j)}{M}}{N} \qquad (2)$$

where op_i^{acq} is the priority of the *OP* which generates the hypothesis; N is the number of *OPs* that generate h_k^{acq}; M is the number of facts applied in the process; and $r(fact_j)$ is the relevance function of each fact. The values are also normalised to fall between zero and one. Once the hypotheses have been evaluated, they pass through a *Filtering phase*. There are two scenarios where the filtering is applicable. The first one occurs when a hypothesis contradicts others. To reduce the degree of inconsistency, the *Filtering phase* selects the hypotheses with greater relevance from those that have multiple options. And, in the second case, the generated hypotheses or facts have been generated only from one document within the collection. These hypotheses are not considered enough relevant.

3.4. Action layer

Combining the facts and the hypotheses generated through the acquisition layer, the action operators generate hypotheses about whether an element is a concept or not and the hierarchical relationships that may appear among them.

3.4.1. Domain concepts detection

These *OPs* detect which CEs can be considered as relevant knowledge of the domain, generating candidates to be new concepts of the target ontology. These *OPs* combine several *acquisition hypotheses* to perform the process. Thus, domain concepts have to be general knowledge from the statistical point of view and highlighted in some way from the visual perspective. Besides, they only consider CEs that are NPs with some textual characteristics, such as those that appear sometimes in uppercase within the text. For example, those CEs that are NPs and statistically relevant, appear in the *title, meta content or keywords* of several documents, and are not any kind of *name entity* are primarily considered as *candidate concepts* to the domain. Besides, the *OPs* do not consider at this moment domain-specific knowledge, such as name entities or elements that appear totally in uppercase, that are supposed to be *instances*. Finally, the *Op* creates a hypothesis that proposes a new domain concept for every CE that follows all the previous conditions. In our experiments in the *Universities* domain, these *OPs* detect candidates to be concepts such as *department, student,* or *publication*. In the case of the *Economics* domain, it detects concepts such as *economics, microeconomics,* or *public economy*. The proposed candidate concepts are accessible by means of the *is-candidate-concept? (*$?e_1$*)* predicate.

3.4.2. Detection of hierarchical relationships

CEs are also proposed as candidate concepts by means of the hierarchical discovery process. *OPs* responsible for the detection of taxonomic relationships combine different evidence generated through the acquisition layer and indicate the possible existence of a hierarchical relationship between a couple of CEs. These evidences are:

- From the visual layout, the hierarchical relationships are evidenced through the hypotheses that state the existence of a *hierarchical context* or *in the same hierarchy* relation between a couple of CEs.
- The chunk, syntactic and semantic analyses, provide evidence about a candidate hierarchical relationship by means of the *hyponym* and *hypernym* hypotheses.
- From a statistical point of view, a CE that is more general than another one can reinforce the validity of such relationship. However, the statistical analysis does not provide alone plausible evidences to detect such relationships.

The figure (3) shows an example of one of the *OPs* that combines the hypotheses and facts provided by the visual layout analysis. In the example, the *OP* uses the hypothesis that both elements are connected by a *hierarchical link* as evidence to identify the existence of a hierarchical relationship between $?e_1$ and $?e_2$. The CEs have also to be relevant both from the visual (to appear in the head section of some documents and to be highlighted elements) and the statistical (to be general knowledge) points of view. In contrast, negative evidence for detecting the relationship is that the CE has been considered as a type of entity (the *OP* looks for relationships between candidate concepts) or both CEs appear frequently collocated within the same STU. In case of a positive match, the *OP* creates the corresponding action hypotheses stating that both CEs are potentially candidate concepts and are hierarchically related.
To define other types of *OPs*, their conditions have to use the evidences that correspond to their corresponding level of analysis (e.g. syntactical based operators will use the hyponym and hypernym evidences, among others).

```
(action-operator taxonomy-html-structure-1
 :vars (?e1 ?e2)
 :type :html-based
 :clauses (
 (:conditions
  (AND
   (is-a-hierarchical-link? ?e1 ?e2)
    (OR (in-meta? ?e1) (in-title? ?e1))
   (in-header? ?e2)
   (is-highlighted-element? ?e1)(is-highlighted-element? ?e2)
   (is-general-knowledge? ?e1)(is-general-knowledge? ?e2)
   (is-more-general-than? ?e1 ?e2)
    (not (in-the-same-unit? ?e1 ?e2))
    (not (is-an-entity? ?e1)) (not (is-an-entity? ?e2))
   (is-a-noun? ?e1) (is-a-noun? ?e2)
    (not (appears-together? ?e1 ?e2)) )
  :actions (
   action-hyp op ?e1 :concept :evidences sources)
   action-hyp op ?e2 :concept :evidences sources)
   action-hyp op ?e2 :is-a ?e1 :evidences sources)) ) ) )
```

Figure 3: Example of an operator to detect hierarchical relationships using the visual layout analysis.

Similar to the acquisition phase, once the action phase finishes and the generated hypotheses have been evaluated using equation (1), they pass through a *Filtering layer*. A hypothesis may establish a subclass relationship between a couple of elements and others pointing in the opposite direction. In these cases, the filtering process selects the hypothesis with greater relevance.

3.5. Consolidation layer

It aims at keeping the correct level of abstraction about what a set of hypotheses is stating, performing two tasks. On the one hand, gathering instances and their relationships and generalising them to their respective concepts. On the other hand, the concepts can be subclasses of several concepts. In these cases, the highest level of generality among all the possibilities is selected, removing the others.

3.6. Evaluation layer

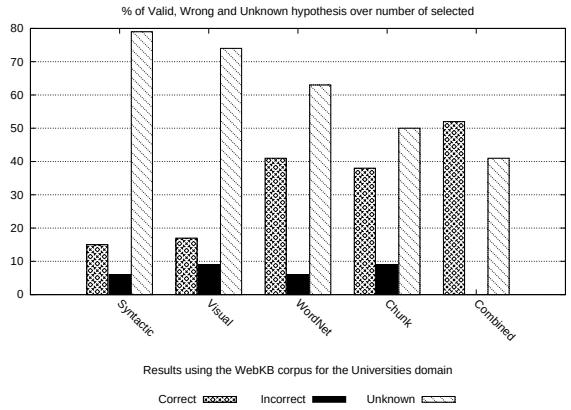
It analyses the final hypotheses computing their relevance to the domain. The evaluation is performed by the procedure of comparing the results with a gold-standard. So, after assessing the relevance of all the generated hypotheses, the most significant ones are selected. The comparison with the standard can produce three types of results. The hypothesis is *Valid* meaning that it proposes an action consistent with the reference ontology. *Incorrect* hypotheses propose something inconsistent with the reference ontology. And, when GALEON cannot decide the correctness of the hypothesis, the hypothesis is labelled as *Unknown*. The number of valid hypotheses selected allows for the estimation of the precision of the process. In addition, by comparing this value with the total number of valid generated hypotheses, we can measure the recall.

3.7. Knowledge augmentation

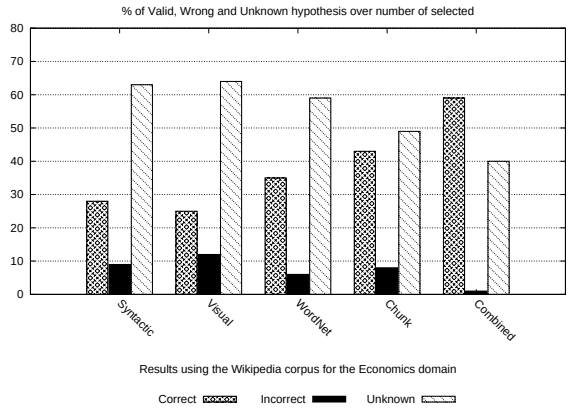
The final knowledge is released into the Ontology repository to make it available. This layer is monitored by the user, who has to make decisions over cases that present ambiguous actions.

4. Experimental results

GALEON has been tested in two different domains: *Universities* and *Economics*. For each domain, a corpus of web documents has been selected. The first one is composed of 9,000 web documents, and the second one is composed of nearly 2,000 web documents (extracted from Wikipedia and dmoz directory). Additionally, in order to evaluate the results, we use a reference ontology for each domain.[4] The experiments aim to show how much of the two reference ontologies the system is able to rebuild. Once the hypotheses have been generated by each configuration, they are automatically classified by GALEON into three groups (*valid, unknown* and *incorrect*) comparing them to the selected reference domain ontology. Finally, we measure precision as the number of *valid* hypotheses divided by the number of selected hypotheses. The recall compares the valid hypothesis globally generated to those valid hypothesis finally selected. When all processing methods are used jointly, they



(a) Results for the *Universities* domain.



(b) Results for the *Economics* domain.

Figure 4: Results obtained using GALEON.

[4]The ontology about universities is located at http://www.cs.umd.edu/projects/plus/DAML/onts/univ1.0.daml. In the Economics domain, the ontology has been made based on the combination of the Agrovoc Thesaurus (http://www.fao.org/agrovoc/) and dmoz taxonomy (http://www.dmoz.org/)

perform with greater precision than any of them in isolation because each one constitutes a source of supplementary evidence to the other. As an example, the use of *syntactic patterns* alone achieves a 15% precision in *Universities* (4a) and 27% of precision in the *Economics* (4b). But, a hypothesis generated by the syntactic analysis can be corroborated by other sources of evidence such as WordNet. This increases the relevance and trustiness of that assumption. Thus, the precision rises in both cases to 51% for the *Universities* domain and to 62% for the *Economics*. Similarly, each source provides new assumptions about elements not found by the individual analysis of each source. In this sense, the accuracy achieved by the combination of all methods together comes from the union of the assumptions properly generated by each of them individually. As a result of the combination, the recall reaches 66% and 71% respectively. In relation to the error rate, it decreases to almost zero in both domains. This is also due to the combination of complementary assumptions that makes the wrongly generated hypotheses receive a lesser reinforcement. Thus, they are ruled out through the filtering stages.

Other state of the art methods also combine several information sources to enrich a domain ontology. Cimiano et al. (Cimiano et al., 2005) combine lexical patterns, WordNet and clustering techniques to enrich an ontology. This combination reaches a lower precision rate than ours, mainly because our method combines more complementary sources of evidences that reinforce more accurately the proposed hypotheses. Ontolearn (R. Navigli, 2004) reaches better levels of both precision and recall, but the process is semi-supervised by an expert who has to decide on the validity of the extracted vocabulary.

5. Conclusions

In this article, a new method for learning ontologies is presented. The method, implemented in our architecture GALEON, combines heterogeneous sources of information and knowledge as well as various processing techniques associated with each of them to improve the detection of potential useful knowledge. First, it extracts the core vocabulary (CEs) to the domain using a parsing process. How a CE appears within the text, how often, in which HTML places and what type of relationships it has with other CEs may denote the relevance of the element to the domain as well as the type of ontological knowledge it is. The underlying idea of our method is that the combination of all these additional sources of evidence improves the accuracy of the OL process decreasing the error. Thus, the CEs are analysed at five different levels at this moment: chunk, statistical, syntactical, visual and semantical. From each of the CEs, GALEON generates hypotheses stating their type, statistical relevance, syntactical and visual relations with other candidates. The experimental results obtained processing a set of HTML documents belonging to two domains, *Universities* and *Economics*, have shown the potential benefit of its use to learn or enrich ontologies following an unsupervised learning approach.

The next step in our work is to introduce a feedback into the process. Given that each *Op* already has a priority value, we plan to automatically adjust their values for producing better results. The feedback from the evaluation will change how the *Ops* combine the available evidence selecting those that may produce better results. The feedback, combined to the automatic evaluation, will allow to train the system to be used in different domains with higher precision.

Acknowledgements

This work has been partially supported by the Spanish MEC projects TIN-2004-02660 and TIN2005-08945-C06-05, a grant from the Spanish MEC, and regional CAM-UC3M project CCG06-UC3M/TIC-0831.

6. References

O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. 2001. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *Proceedings of 10th International World-Wide Web Conference*.

P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab, 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, chapter Learning Taxonomic Relations from Heterogeneous Sources of Evidence, pages 59–73. IOS Press, July.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Sean Slattery. 2000. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, (118):69–113.

O. Downey, D. Etzioni, S. Soderland, and D. Weld. 2004. Learning text patterns for web information extraction and assessment. In *Proceedings of the AAAI Workshop on Adaptive Text Extraction and Mining*.

A. Gómez-Pérez and D. Manzano-Macho. 2005. An overview of methods and tools for ontology learning from texts. *Knowledge Engineering Review*, 19:187–212.

Z. Harris. 1968. *Mathematical Structures of Language*. John Wiley and Sons.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Int. Conf. on Computational Linguistics*, Nantes, France.

C. Jaquemin. 2001. *Spotting and discovering terms through Natural Language Processing*. MIT Press.

K. Sparck Jones. 1973. Index term weighting. *Information Storage and Retrieval*, 9:619–633.

N Kurshmerick, F. Ciravegna, A. Doan, C. Knoblock, and S. Staab, editors. 2005. *Proceedings of the 2005 Seminar on Machine Learning for the Semantic Web*.

R. Navigli, 2004. *Web Mining: Applications and Techniques*, chapter Ontology learning from Domain Web Corpus, pages 69–98. IGI Publishing, PA, USA.

P. Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on AI*, pages 448–453.

M. Ruiz-Casado, E. Alfonseca, and P. Castells. 2007. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data and Knowledge Engineering*, 61:484–499.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.