# Mining for lexons: applying unsupervised learning methods to create ontology bases

Marie-Laure Reinberger[1], Peter Spyns[2], Walter Daelemans[1], and Robert Meersman[2]

[1] CNTS - University of Antwerp,
Universiteitsplein 1, B-2610 Wilrijk - Belgium,
{reinberg,daelem}@uia.ua.ac.be
[2] STAR Lab - Vrije Universiteit Brussel,
Pleinlaan 2 Gebouw G-10, B-1050 Brussel - Belgium
{Peter.Spyns,Robert.Meersman}@vub.ac.be

**Abstract.** Ontologies in current computer science parlance are computer based resources that represent agreed domain semantics. This paper first introduces ontologies in general and subsequently, in particular, shortly outlines the DOGMA ontology engineering approach that separates "atomic" conceptual relations from "predicative" domain rules. In the main part of the paper, we describe and experimentally evaluate work in progress on a potential method to automatically derive the atomic conceptual relations mentioned above from a corpus of English medical texts. Preliminary outcomes are presented based on the clustering of nouns and compound nouns according to co-occurrence frequencies in the subject-verb-object syntactic context.
*Keywords*: knowledge representation, machine learning, text mining, ontology, semantic web, clustering, selectional restriction, co-composition.

## 1 Introduction and General background

### 1.1 The Semantic Web

Internet technology has made IT users aware of both new opportunities as well as actual needs for large scale interoperation of distributed, heterogeneous, and autonomous information systems. Additionally the vastness of the amount of information already on-line, or to be interfaced with the WWW, makes it unfeasible to depend merely on human users to correctly and comprehensively identify, access, filter and process the information relevant for the purpose of applications over a given domain. Be they called software agents, web services, or otherwise, this is increasingly becoming the task of computer programs equipped with *domain knowledge* . Presently however there is an absence of usable formal, standardised and shared domain knowledge of what the information stored inside these systems and exchanged through their interfaces actually means. Nevertheless this is a prerequisite for agents and services (or even for human users) wishing to access the information but who, obviously, were never involved when these systems were created. The pervasive and explosive proliferation of computerised information systems (databases, intranets, communication systems, or other) quite simply makes this into the key problem of the application layer of the current internet and

its semantic web successor. The equally obvious key to the solution of this problem therefore lies in a better understanding, control and management of the semantics of information in a general sense.

### 1.2 Ontologies

The semantic principles and technology underlying such solutions are emerging in the form of *ontologies*, i.e. practically usable computer-based repositories of formal and agreed semantics about application domains [35]. Ultimately these ontologies will coalesce, more or less tightly, into a vast knowledge resource about the entire ”information universe” that is the web. In the present parlance [17], a computer-implemented ontology roughly is constituted of :

1. a computer-based lexicon, thesaurus, glossary, or other type of controlled and structured vocabulary of linguistic terms; the terms in those vocabularies are assumed to refer in well-defined ways to concepts.
2. an extension with explicit ”knowledge” about a given domain, under the form of relationships between concepts, often including a taxonomy of those concepts.
3. an extension with a set of general rules and constraints supporting reasoning about the concepts. [3]

Roughly speaking, in the realm of information systems a (first order) formal semantics of an application may be defined by a *formal interpretation*, viz. mapping, of some given computer representation of that application (in a suitable computer language) in terms of a given world domain, or rather some suitably elementary and agreed conceptualisation of it. This common classic formalism, also the most amenable to ontologies, is called declarative or Tarski semantics [34] and may be found in various places in the database and AI literature, in Reiter’s seminal paper [31] linking the two fields through first order logic, or in the textbook by Genesereth & Nilsson [14].

Essentially this approach replaces ”the world” (the domain) by a *conceptualisation*, a mathematical object that typically consists of very elementary constructs such as a set of objects and a set of (mathematical) relations. Conceptualisations theoretically are language-, context-, and usage independent formalisations of this world, or domain of discourse. A formal ontology on the other hand is a formal rendering of such a conceptualisation through e.g. an ontology language [35]. For a proper understanding, the actual notion of ontology should therefore be seen as separate from this conceptualisation of the ”world” [17]. Note that this distinction is not always made in parts of the recent literature on ontologies.

### 1.3 Mining for DOGMA terms and lexons

The *DOGMA* (Developing Ontology-Guided Mediation for Agents) ontology engineering approach of VUB STAR Lab is based on the three rather evident observations that:

1. agreements become easier if the items involved are simpler

[3] This latter item falls outside the scope of this paper.

2. most relevant human knowledge is massively available in natural language in text documents and other "lexical" sources such as databases
3. conceptualisations -and hence ontologies- should be as independent as possible of intended application and design(er) context, and of the language used to reach the agreement

A DOGMA inspired ontology [4] is based on the principle of a double articulation: an ontology is decomposed into an ontology base, which holds (multiple) intuitive conceptualisation(s) of a domain, and a layer of ontological commitments, where each commitment holds a set of domain rules.

The *ontology base* consists of sets of intuitively "plausible" domain fact types, represented and organised as sets of context-specific binary conceptual relations, called *lexons*. They are formally described as $< \gamma, \lambda : term_1, role, co-role, term_2 >$, where $\gamma$ is a context identifier, used to group lexons that are intuitively "related" in an intended conceptualisation of a domain for a specific natural language $\lambda$. Informally we say that a lexon is a fact that may hold for some application, expressing in that case that within the context $\gamma$ the $term_1$ may plausibly have $term_2$ occur in an associating $role$ (with $co-role$ as its inverse) with it. Lexons are independent of specific applications and should cover relatively broad domains. For each context $\gamma$ and term $term$ for a natural language $\lambda$, the triplet $(\gamma, \lambda, term)$ is assumed to refer to a unique concept $concept$. Formulated alternatively, a $concept$ is lexicalised for a natural language $\lambda$ by a specific $term$ depending on a specific context $\gamma$ of use. More details on the DOGMA engineering approach (e.g., the commitment layer) can be found in [19, 33].

The main research hypothesis is that lexons, representing the "basic facts" expressed in natural language about a domain can be extracted from textual sources. Other potential sources are database schemas or semi-structured data (e.g. XML files). (Semi-)automatic mining for lexons to populate the ontology base would allow to short-circuit the knowledge acquisition bottle-neck (human knowledge elicitation and acquisition implies a high cost in terms of time and resources). We have opted for extraction techniques based on unsupervised learning methods since these do not require specific external domain knowledge such as thesauri and/or tagged corpora [5]. As a consequence, the portability of these techniques to new domains is much better [27]:p.61].

A first step in order to mine for DOGMA lexons is the discovery and grouping of relevant terms. A domain expert will then distill concepts from the set of terms and determine which relationships hold between the various newly discovered concepts. Note that the terms and lexons operate on the language level, while concepts and conceptual relationships are considered to be, at least in principle, language independent. By doing so, the domain expert - together with the help of an ontology modeller - shapes the conceptualisation of a domain as it is encoded in the textual sources (taking synonymy into account). The second step will most probably be repeated several times before an adequate and shared (formal) domain model is agreed upon.

[4] An overview of other ontology representation techniques can be found in [32].
[5] Except the training corpus for the general purpose shallow parser - see below.

### 1.4 Selectional Restrictions and Co-composition

A lot of information about the meaning of words can be inferred from the contexts in which they occur [20]. For example, information about the functionality and properties of the concepts associated with a word can be inferred from the way nouns and verbs are combined. Of course, a fine-grained representation of the meaning of a word cannot be reached without the use of large amounts of syntactically analysed data about their use. The use of powerful and robust language processing tools such as shallow parsers allows us to parse large text collections (available in massive quantities) and thereby provide potentially relevant information for extracting semantic knowledge.

The linguistic assumptions underlying this approach are (i) the principle of selectional restrictions (syntactic structures provide relevant information about semantic content), and (ii) the notion of co-composition [28] (if two elements are composed into an expression, each of them imposes semantic constraints on the other). The fact that heads of phrases with a subject relation to the same verb share a semantic feature would be an application of the principle of *selectional restrictions*. The fact that the heads of phrases in a subject or object relation with a verb constrain that verb and vice versa would be an illustration of *co-composition*. In other words, each word in a noun-verb relation participates in building the meaning of the other word in this context [11, 12]. If we consider the expression "write a book" for example, it appears that the verb "to write" triggers the informative feature of "book", more than on its physical feature. We make use of both principles in our use of clustering to extract semantic knowledge from syntactically analysed corpora.

## 2 Objectives

Our purpose is to build a repository of lexical semantic information from text, ensuring evolvability and adaptability. This repository can be considered as a complex semantic network. An important point is that we assume that the method of extraction and the organisation of this semantic information should depend not only on the available material, but also on the intended use of the knowledge structure. There are different ways of organising it, depending on its future use and on the specificity of the domain. In this paper, we deal with the medical domain, but one of our future objectives is to test our methods and tools on different specific domains.

In the remainder of this paper, we will shortly introduce in the next section (Material and Methods) the shallow parser to extract subject-verb-object structures and the English medical corpora used (section 3.1), the clustering methods applied to the task (section 3.2), and an evaluation of their accuracy (section 3.3) using WordNet [24] as a gold standard. Section 4 describes the various experiments in detail. We have tested similarity based clustering algorithms, applying some variations to the set of data and to the algorithm in order to compare and improve the quality of the clusters: soft (section 4.1) and hard clustering (section 4.2) are compared (section 4.3 and 4.4) and merged (section 4.5). Particular attention is paid to compound nouns (section 4.6). The results are briefly discussed and related to other on-going work in this area (section 5). Some ideas about future work are also presented before concluding this paper (section 6).

## 3 Material and Methods

### 3.1 Shallow parsing

In a specific domain, an important quantity of semantic information is carried by the nouns. At the same time, the noun-verb relations provide relevant information about the nouns, due to the semantic restrictions they impose. In order to extract this information automatically from our corpus, we used the memory-based shallow parser which is being developed at CNTS Antwerp and ILK Tilburg [4, 5, 9] [6]. This shallow parser takes plain text as input, performs tokenisation, POS tagging, phrase boundary detection, and finally finds grammatical relations such as subject-verb and object-verb relations, which are particularly useful for us. The software was developed to be efficient and robust enough to allow shallow parsing of large amounts of text from various domains.

The choice of the specific medical domain has been made since large amounts of data are freely available. In particular, we decided to use Medline, the abstracts of which can be retrieved using the internal search engine. We have focused on a medical subject that was specific but common enough to build a moderately big corpus. Hence, this corpus is composed of the Medline abstracts retrieved under the queries "hepatitis A" and "hepatitis B". It contains about 4 million words. The shallow parser was used to provide a linguistic analysis of each sentence of this corpus, allowing us to retrieve semantic information of various kinds.

### 3.2 Clustering

Different methods can be used for the extraction of semantic information from parsed text. Pattern matching [2] has proved to be a efficient way to extract semantic relations, but this method involves the predefined choice of the semantic relations that will be extracted. We rely on a large amount of data to get results using clustering algorithms on syntactic contexts in order to also extract previously unexpected relations.

Clustering requires a minimal amount of "manual semantic pre-processing" by the user. Clustering on nouns can be performed by using different syntactic contexts, for example noun+modifier relations [7] or dependency triples [20]. As we have mentioned above, the shallow parser detects the subject-verb-object structures. This gives us the possibility to focus on the noun-verb relations with the noun appearing as the head of the subject or the object phrase, but also on the relation noun-verb-noun, where the verb features a link between the two head nouns. From now on, we will refer to the nouns appearing as the head of the subject or object phrase as "nouns".

The first step of the similarity-based clustering algorithm we are using consists of processing the parsed text to retrieve the co-occurring noun-verb-noun relations, and remembering whether the noun appeared in a subject or in an object position. This step is performed with the use of a stop list that skips all pairs containing the verbs *to be* or *to have*. We want to point out that we are not implying by doing so that those two verbs do not provide relevant information. They simply are too frequent and have such a broad range of meanings that we cannot, with this method and at this stage of

[6] See `http://ilk.kub.nl` for a demo version.

the experiments, take them into account. The words are then lemmatised, before we select from the resulting list the most frequent relations. Those relations are organised in classes before the processing of a clustering algorithm. We will describe in the next section the evaluation method that we have used.

### 3.3 Evaluation

Evaluation of extracted clusters is problematic, as we do not have any reference or model for the clusters that we want to build. At the same time, we wanted an automatic evaluation method. We chose to use WordNet, which is freely available. As WordNet is not devoted to a particular domain, it can be used for the different corpora we are experimenting with. WordNet has been used by [20] for the evaluation of an automatically constructed thesaurus. Wordnet was transformed into the same format as the thesaurus, and a comparison was carried out between the entries of the thesaurus and the entries of the transformed WordNet, allowing a global evaluation of the constructed thesaurus.

We want to validate the relations between words that are established through our clustering process on medical text, but as WordNet does not contain all information related to the medical domain, it will provide us with only a sample of the correct associations. The semantic information provided by WordNet is only used in the evaluation process. We do not intent to correct or enlarge the clusters with this information, as we wish to stay as much as possible within the paradigm of purely unsupervised learning.

From the list of all nouns appearing in the clusters, we have kept the sublist that belongs to WordNet (WN words). Then, we have used this list to build the list containing all the pairs of nouns connected in WordNet through a relation of synonymy, hypernymy, hyponymy, meronymy or holonymy. Here are some examples of the relations found in WordNet:

hepatitis - disease (hypernymic relation)
blood - cells (meronymic relation)
aim - purpose (synonym)

This list of pairs (WN pairs) allows us to compute a recall value R, with:
R = # WN pairs in the clusters / # WN pairs

Computing a precision value was more difficult as Wordnet is not complete or even representative for the medical domain. Our clusters depend on subject-verb and object-verb relations, and consequently some of them will stand for functional relations. One cluster for example will contain the list of elements that can be “gathered” or “collected”, namely “blood”, “sample” and “specimen”. Another cluster will link “infection” and “disease” as object of the verbs “to cause” and “to induce”. “Syringe” and “equipment” appear in the same cluster, gathered by the verbs “to share” and “to reuse”. Those relations do not appear in WordNet. Therefore, the precision values we give must be considered as a “lower bound precision” or “minimum precision” mP. It is computed by dividing the number of correct WordNet pairs found in the clusters by the total number of pairs of words (formed with WordNet words) in the clusters:
mP=# WN pairs in the clusters / # pairs

In order to balance this minimum precision, we have made an attempt to build a more exhaustive set of relations, in order to extrapolate a more realistic precision value.

We have worked on a sample of words W and clusters C, associated to a set of WordNet pairs WnP. They correspond to a minimum precision mP=(# WnP in C) / (# pairs in C). We have derived manually all possible pairs, including functional relations like the ones mentioned above. We obtain an augmented set of pairs AugP, that allows us to find the new set of correct pairs in the set of clusters C and a new precision:
newP = (# AugP in C) / (# pairs in C)

We have used the ratios obtained with this sample to compute a new value that we will call extrapolated precision (eP) for the various clustering experiments. To do this, we assume that WnP/AugP is a constant value, and that:
(# WnP in C)/WnP = (# AugP in C)/AugP

This extrapolated precision will allow us to propose an estimation for the real precision. In the next sections, we will give a description of the different steps of our experiment and of the evaluation of the different results.

## 4 Description of the experiments

The first step of the experiment was to measure if and to what extent the information provided by the shallow parser is relevant for the extraction of semantic relations. Even if the syntactic analysis supplies useful information, it requires some processing time as well, and this cost is only motivated if it improves the semantic analysis. In order to evaluate this, we carried out a comparative study on the results of the clustering applied to raw text and parsed text. We have compared the results using three different clustering algorithms: a soft (or disjunctive) similarity-based clustering algorithm, a hard bottom-up hierarchical similarity-based clustering algorithm, and a non-hierarchical clustering algorithm (AutoClass [8]). We have applied the hard clustering algorithms to two different corpora: our Medline corpus and the Wall Street Journal (WSJ) corpus (1M words). We refer here to “soft clustering” as clustering algorithms that allow an element to appear in more than one cluster, contrary to “hard clustering” where an element can appear in one and only one cluster. We will give a description of the clustering algorithms we have been using before commenting on the results.

### 4.1 Soft clustering

The bottom-up soft clustering we have performed gives us the possibility to take into account the ambiguity of the words [10] by allowing a word to belong to different clusters. The soft similarity-based clustering algorithm applied on parsed text starts with the processing the parsed text to retrieve the co-occurring noun-verb pairs, and remembering whether the noun appeared in a subject or in an object position. We then select from the list we get the most frequent co-occurrences: the 100 most frequent noun-verb relations with the nouns appearing in the subject group, and the 100 most frequent relations where the noun is part of the object group. What we obtain is a list of verbs, each verb associated with a list of nouns that co-occur with it, either as subjects only or as objects only. Here is an extract of the list (“_o” (resp “_s”) indicates that the list of nouns appears as object (resp. subject)):

acquire_o: hepatitis infection virus disease
compensate_o: liver cirrhosis disease
decompensate_o: liver cirrhosis disease
decrease_s: rates prevalence serum incidence proportion number percentage
estimate_o: prevalence incidence risk number
transmit_o: hepatitis infection disease

It appears, for example, that the same set of nouns occur as object of the verbs "compensate" and "decompensate", or that "acquire" and "transmit" present a very similar set of nouns occurring as object. Some cases are more interesting, for example the fact that the set of nouns appearing as the subject of "decrease" present strong similarities with the set of nouns appearing as object of "estimate".

The next step consists of clustering these classes of nouns according to their similarity. The similarity measure takes into account the number of common elements and the number of elements that differ between two classes. Each class is compared to all other classes of nouns. For each pair of classes C1-C2, the program counts the number of nouns common to both classes (sim), the number of nouns only present in C1 (dif1) and the number of nouns only present in C2 (dif2). If sim, dif1 and dif2 respect some predefined values the matching is considered to be possible. After the initial class has been compared to all other classes, all the possible matchings are compared and the one producing the largest new class is kept (in case of ties, the first one is kept). Each time a new cluster is created, the 2 classes involved are removed from the processed list. The whole process is iterated as long as at least one new matching occurs, resulting in the creation of a new cluster.

### 4.2 Hard clustering

The hard clustering experiments have been performed on the most frequent vectors associating nouns to their co-occurring verbs. On raw text, we have extracted 2-grams, 3-grams and 4-grams and built vectors representing co-occurring words. The input of the bottom-up similarity-based algorithm is a list of nouns (or words), each of them associated to its list of co-occurring verbs (or words). Contrarily to the soft clustering algorithm, the nouns (words) are clustered according to the similarity between the classes of verbs. The similarity measure, as for the soft clustering algorithm, takes into account the number of common elements and the number of elements that differ between two classes of verbs. The classes are compared two by two and the process is iterated as long as a cluster can be modified. When no change is possible, the similarity measure is lowered and the process is iterated again until we obtain one cluster containing all the nouns. The resulting tree is cut according to the percentage of nouns that have been clustered. A cluster is valid when at least two nouns are in it.

The second hard clustering algorithm we have used is a non-hierarchical hard clustering algorithm called AutoClass. AutoClass is fed with ordered vectors of attribute values and finds the most probable set of class descriptions by performing successive reallocations, given the data and prior expectations. We have selected AutoClass among

other existing clustering algorithms for this comparative experiment as it showed good results on our data. For raw text, the best results have been observed on 3-grams.[7]

### 4.3 Comparison

The comparative study parsed text/raw text showed very different results for the soft and for the hard clustering. Concerning the soft clustering, we have observed a better recall on parsed text, but a better precision on raw text. This part of the experiment is described in detail in [29].

For the hard clustering, Table 1 shows better results on parsed text for the Medline corpus. But the comparative study we have carried out with two hard clustering algorithms (similarity-based and AutoClass) and two corpora (Medline and WSJ) shows less clear results on the WSJ corpus, and with AutoClass (Table 2). Further experiments are necessary in order to check to what extent the size and the specificity of the corpus influence the results.

| | % of words clustered | R | mP | eP | Nb of pairs |
|---|---|---|---|---|---|
| SP | 90% | 15% | 13% | 33% | 250 |
| 3-grams | 90% | 10% | 13% | 23% | 250 |

**Table 1.** *Recall (R), minimum precision (mP) and extrapolated precision (eP) values for the hard clustering similarity-based algorithm on parsed text and on plain text (n-grams). The experiment has been carried out for about 150 words, 90% of them are clustered and the set of clusters contains 250 pair relations.*

### 4.4 Performing hard and soft clustering

Next we discuss the performance of the similarity-based hard and soft clustering algorithms applied only to parsed text. The soft clustering has been performed on two different sets of data. The first set consisted in the 200 vectors associating a noun to its co-occurring verbs and corresponding to the most frequent co-occurrences (poor verbal information). In the second set, each of the 200 nouns was associated to the list of verbs frequently co-occurring, and consisted therefore of 200 couples noun-list of verbs (rich verbal information). The purpose was to vary the amount of verbal information, in order to decide whether considering more verbal information would improve the results or increase the noise.

This soft clustering algorithm tends to produce too many clusters, especially too many large clusters (although we get good small clusters as well, see Figure 1), and it is difficult to sort them. Restricting the similarity conditions reduces the number of "bad" clusters as well as the number of "good" clusters. The evaluation shows that

[7] The results of this study have been presented at CLIN-02 [30].

| | Sim. based | | | AutoClass | | |
|---|---|---|---|---|---|---|
| | R | mP | eP | R | mP | eP |
| WSJ corpus | | | | | | |
| n-grams | 7% | 10% | 16% | 8% | 7% | 10% |
| SP | 11% | 12% | 19% | 6% | 10% | 15% |
| MEDLINE corpus | | | | | | |
| n-grams | 10% | 13% | 23% | 30% | 2% | 4% |
| SP | 15% | 13% | 33% | 11% | 8% | 12% |

**Table 2.** *Comparison of 2 hard clustering algorithm: the hierarchical similarity based algorithm vs. the non-hierarchical AutoClass algorithm, on 2 corpora*

*Cluster 1:* aim objective purpose study
*Cluster 2:* immunization vaccine vaccination

**Fig. 1.** *Examples of soft clusters*

good information is found in the small clusters, and that we obtain the best results with rich verbal information (with poor verbal information, we have observed, for the same number of words clustered, a lower recall and a lower precision) and by dismissing the biggest clusters. The results, using rich verbal information, are displayed in Table 3.

The experiment on hard clustering has been carried out with poor verbal information. When we compare with the soft clustering results, we notice an important decrease of the number of clusters, and a reduction of the average size of the clusters (see Figure 2). Inconveniences are that we miss every case of polysemy, that we cannot get an exhaustive set of the possible relations between the nouns, and that the recall is very low. Nevertheless, a positive aspect lies in the fact that here, in accordance with the co-composition hypothesis, the nouns are clustered according to the semantic information contained in the sets of verbs, whereas for the soft clustering, only the initial classes of nouns are built using verbal information.

*Cluster 1:* month year
*Cluster 2:* children infant
*Cluster 3:* concentration number incidence use prevalence level rate
*Cluster 4:* course therapy transplantation treatment immunization

**Fig. 2.** *Examples of hard clusters*

The modification of the similarity measure produced only minor changes for this experiment in the results, and the ratio between recall and precision was steady. We give a summary of the results in Table 4. Both methods (soft and hard) present a balance

between advantages and shortcomings and produce good clusters and we would like to keep the best clusters resulting from each method. That lead us to the idea of merging, or combining the two sets of results.

| | Nb of cl. | % wds in cl. | Size cl. | R | mP | eP |
|---|---|---|---|---|---|---|
| E1.1 | 120 | 94% | 8.87 | 75% | 4% | 10% |
| E2.1 | 155 | 91% | 5.39 | 74% | 6% | 15% |
| E1.2 | 28 | 64% | 10.71 | 57% | 7% | 18% |
| E2.2 | 32 | 66% | 9.81 | 65% | 8% | 19% |

**Table 3.** *Number of clusters, % of those words clustered, average size of the clusters, recall, min. and ext. precision values for the different soft clustering experiments (rich verbal information), for about 150 words. E1.1 is the initial experiment. E2.1 has been carried out easing the similarity measure but discarding big clusters. E1.2 and E2.2 are based resp. on E1.1 and E2.1, the small clusters (2 elements) being discarded.*

| | Nb of cl. | % wds in cl. | Size cl. | R | mP | eP |
|---|---|---|---|---|---|---|
| Hd cl. | 45 | 90% | 4 | 15% | 13% | 33% |

**Table 4.** *Best recall, min. and ext.precision values for the hard clustering experiment (about 160 words clustered).*

### 4.5 Merging soft and hard clustering

To summarise the results described above, we can say that the soft clustering provides too many clusters and too many large clusters, and the hard clustering does not build enough clusters, hence not enough relations. But both sets of results present as well strong similarities, and numerous clusters are formed whatever algorithm used.

In consequence, our next attempt has been to try to combine the results of both algorithms (soft and hard). More precisely, we assume that the relations between nouns that are produced with both methods are more reliable, and we have filtered the soft clustering results, using the hard clustering results. By doing this, we keep the possibility for a noun to belong to more than one cluster, which represents the situation that a noun can share relations of different kinds, and also represents the polysemic properties of some nouns. We have used the similarity measure described for the previous clustering algorithms, and we have compared each hard cluster with every soft cluster, considering the number of common elements and the number of differing elements to decide if the soft cluster would be kept or not.

We give below some concrete examples of this operation of merging. We indicate successively in the examples below the hard cluster in which the word appears, followed by extracts of some soft clusters in which it appears, and finally the clusters obtained by combining both results.

*Merging clusters, Example 1: "Disease"*
1. Hard clustering:
- disease transmission
2. Soft clustering: 8 clusters, including:
- drug disease treatment clinic
- prevalence infection correlation disease...
- ...
3. Merging: 2 clusters
- hepatitis infection disease case syndrome
- disease liver cirrhosis carcinoma vaccine HCC HBV virus history method model

In the hard clustering results, the noun "disease" appears in a two-element cluster, the second element being "transmission". This is very poor if we consider the importance of this word and the various relations it shares. Alternatively, "disease" appears in 8 different soft clusters, some of them containing about 20 words and including non relevant relations. After combination, "disease" is associated to 2 clusters. These 2 clusters belong to the 8 soft clusters set containing "disease". They have been kept because they hold relations that appear as well in a hard cluster. One of them contains general words (case, syndrome, infection) and the other more specific information related to "disease"(cirrhosis, carcinoma, virus...)

*Merging clusters, Example 2: Chemotherapy*
1. Hard clustering:
- therapy transplantation immunization treatment
2. Soft clustering:
- hepatitis blood factor HBV doses chemotherapy treatment vaccine vaccines vaccination injection drug immunization
- liver chemotherapy treatment transplantation
3. Merging:
- liver transplantation chemotherapy treatment

The noun "chemotherapy" does not appear in a hard cluster, and appears in 2 soft clusters, including a big cluster (13 words). But as a hard cluster links "transplantation" and "treatment", the merging operation keeps the soft cluster that associates "chemotherapy" to "liver", "treatment" and "transplantation".

The operation of merging has also revealed reliable two-element clusters composed of strongly related words, such as "hepatitis infection" or "hepatitis virus" for example. The best results for the merging have been obtained by sorting the soft clusters obtained with rich verbal information with the hard clusters. Comparative results are displayed in Table 5.

| | R | mP | eP | Nb of pairs |
|---|---|---|---|---|
| Random | 4% | 2% | 10% | (250) |
| Hard cl. | 15% | 13% | 33% | 250 |
| Soft cl. | 74% | 6% | 15% | 8000 |
| Merging | 62% | 12% | 31% | 1100 |

**Table 5.** *Recall, min. and ext. precision values for the different clustering experiments, considering 150-200 words (summary)*

The merging of soft and hard clustering has improved the results of the soft and hard clustering experiments. But we have only considered in this first step the clustering of head nouns, without taking into account the numerous compound nouns that are used in the medical domain.

In the last section of this paper, we will describe the first set of clustering experiments we have carried out on compound nouns.

### 4.6 Turning to compound nouns...

| | Tot. nb of words | Nb of WN words | R | mP | eP |
|---|---|---|---|---|---|
| Sbj | 252 | 45 | 24% | 12% | 65% |
| Obj | 241 | 71 | 22% | 11% | 49% |

**Table 6.** *Recall and precision values for the clustering experiments on compound nouns*

Compound nouns are an important source of semantic information, especially in the medical domain. As we process in an unsupervised way, we do not know which association of nouns is a compound noun, but the syntactic analysis allows us to detect associations “noun noun” and “adjective noun”frequently occurring in the subject and object phrases. Not all of them fit with the formal definition of a “compound nouns”. However, we have performed the clustering on all the frequently occurring associations, and we will refer from now on to those expressions as “compounds”. We have chosen to perform a hard clustering on the compounds for two reasons. As we have shown it above, the hard clustering takes into account the notion of co-composition (by clustering nouns through the semantic verbal information), which could allow us in a next step to build semantic relations between classes of nouns and verbs. The second reason is based on the fact that a noun can appear in different compounds, each of them standing for a different semantic feature of this noun. In consequence, this noun can appear in different clusters. At the same time, we have modified the similarity measure in order to take into account more sparse data.

The first results show an improvement when we compare them to the previous clustering on nouns, especially in the quality of "large" clusters (more than 10 words). For comparison, we have clustered the compounds appearing in the subject phrase, and the compounds appearing in the object phrase separately. It seems that the clustering of compounds belonging to the object phrase is more efficient than the clustering of compounds appearing in the subject phrase. That appears to be the case especially for the extrapolated precision value. We can advance two hypotheses for this difference in the results. On the one hand, the high proportion of passive sentences limits the number of subject-verb structures found by the shallow parser. On the other hand, a higher proportion of the compounds occurring in the object phrases are domain-specific. At this point of our study, we have only evaluated the results using WordNet. The examples below show clusters containing sparse data that were not taken into account before. Those clusters could be validated by WordNet:

- face mask, mask, glove, protective eyewear
- woodchuck, mouse, animal, chimpanzee

We observe (Table 6) a low recall in the results if we compare it to the recall values of the soft clustering experiments described above, but still this recall is better than the one obtained in the hard clustering experiment. We must signal here that the proportion of words and consequently the proportion of relations we can evaluate is inferior to the proportion evaluated in the previous experiments, due to the high percentage of compounds that do not belong to WordNet. Actually, a third of the compounds appearing in the objects clusters has been evaluated, but only a fifth of the nouns appearing in the subjects (which means that the evaluation concerning the subjects cannot be considered as reliable). The evaluation with WordNet allows us to compare the results with the previous clustering experiments, but we are planning to perform now a more reliable evaluation using UMLS (Unified Medical Language System [18]). We hope to get a better and more reliable evaluation by making use of this specific ontology, as we can spot the presence of many clusters that WordNet could not evaluate but the content of which looks "interesting". Here are some of them, whose content has been validated using UMLS:

- immunoadsorbent, immunoassay, immunospot, immunosorbent, immunosorbent assay
- passive haemagglutination, transcriptase activity, transcriptase inhibitor, transcription, transcriptase, transcriptase polymerase chain reaction, transcription-polymerase chain reaction, transcription polymerase chain reaction
- blood sample, information, sample, sera, blood specimen, serum sample, specimen

## 5 Discussion and Related Work

Unsupervised clustering allows us to build semantic classes. The main difficulty lies in the labelling of the relations for the construction of a semantic network. The ongoing

work consists in part in improving the performance of the shallow parser by increasing its lexicon and training it on passive sentences taken from our corpus, and in part in refining the clustering and using UMLS for the evaluation. At the same time, we work on using the verbal information to connect clusters of nouns-subject and clusters of noun-objects, and we turn as well to pattern matching in order to label semantic relations.

Related work in the medical area happens in the context of the MuchMore project [27]. However, the UMLS is used as an external knowledge repository to discover additional terms on basis of attested relations between terms appearing in a text. Relations themselves are not the focus of the research. Earlier work on creating medical ontologies from French text corpora has been reported on by [25]. Instead of using shallow parsing techniques, "full parse" trees are decomposed into elementary dependency trees. The aim is to group bags of terms or words according to semantic axes. Another attempt involving clustering on specific domains, including the medical domain, is described in [3]. Term extraction is performed on a POS-tagged corpus and followed by a clustering operation that gathers terms according to their common components, in order to build a terminology. An expert provides some help in the process, and performs the evaluation.

Some work of the same kind has been done for other specific domains, for example the terrorist attacks domain in French, as described in [10]. Nouns are gathered in classes and clustered according to their semantic similarity. Here as well, an expert participates in the process, sorting the information after each step of clustering, in order to obtain classes of nouns and frames of sub-categorization for verbs. Unsupervised clustering has been performed as well on general domains. In [20], a thesaurus is built by performing clustering according to a similarity measure after having retrieved triples from a parsed corpus. Here, a big corpus (64M words) was used, and only very frequently occurring terms were considered. A domain independent recent work is presented in [21]. But here again, external knowledge (a semantic dictionary that relates terms to concepts and an external taxonomy) is used. This method allows to calculate whether a relation should involve a particular concept or rather one of its ancestor nodes. A very recent overview of ontology learning methods and techniques is provided by [15].

Unsupervised clustering is difficult to perform. Often, external help is required (expert, existing taxonomy...). However, using more data seems to increase the quality of the clusters ([20]). Clustering does not provide you with the relations between terms, hence the fact that it is more often used for terminology and thesaurus building than for ontology building.

Performing an automatic evaluation is another problem, and evaluation frequently implies a manual operation by an expert [3, 10], or by the researchers themselves [16]. An automatic evaluation is nevertheless performed in [20], by comparison with existing thesauri like WordNet and Roget.

## 6 Future Work and Conclusion

Note that the terms and lexons operate on the language level, while concepts and conceptual relationships are considered to be, at least in principle, language independent. The next step of this research could consist on one side in adding more medical information, especially a kind of information that would enrich the corpus and the semantic relations, and could consist for example of more basic medical facts than the ones encountered in Medline abstracts. However, care has to be taken to maintain the general purpose and flexibility of the method, by avoiding to rely too heavily on external knowledge sources (taxonomy, semantic lexicon, semantic corpus annotations, ...). Therefore, we are planning similar experiments for other specific domains, as the comparison with the Wall Street Journal corpus seems to show that different data can have an effect on the semantic extraction operations. Although it is too early for solid conclusions, we feel that the method presented in this paper merits further investigations, especially regarding the discovery of semantic relations. Only then, genuine lexons, as defined according to the DOGMA approach, could be automatically mined from text corpora. This would constitute a major breakthrough in the field of ontological engineering.

**Acknowledgments** Most part of this research was carried out in the context of the OntoBasis project (GBOU 2001 #10069), sponsored by the IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders).

## References

1. Agirre E. and Martinez D., Learning class-to-class selectional preferences. In *Proceedings CoNLL-01*, 2001.
2. Berland M. and Charniak E., Finding parts in very large corpora. In *Proc. of ACL-99*, 1999.
3. Bourigault D. and Jacquemin C., Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings of EACL-99*, 1999.
4. Buchholz S., *Memory-Based Grammatical Relation Finding*. 1999.
5. Buchholz S., Veenstra J., and Daelemans W., Cascaded grammatical relation assignment. PrintPartners Ipskamp, 2002.
6. Caraballo S., Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of ACL-99*, 1999.
7. Caraballo S. and Charniak E., Determining the specificity of nouns from text. In *Proceedings of SIGDAT-99*, 1999.
8. Cheeseman P. and Stutz J., Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. 1996.
9. Daelemans W., Buchholz S., and Veenstra J., Memory-based shallow parsing. In *Proceedings of CoNLL-99*, 1999.
10. Faure D. and Nédellec C., Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system Asium. In *Proc. of EKAW-99*, 1999.
11. Gamallo P., Agustini A., and Lopes G., Selection restrictions acquisition from corpora. In *Proceedings EPIA-01*. Springer-Verlag, 2001.
12. Gamallo P., Alexandre Agustini A., and Lopes G., Using co-composition for acquiring syntactic and semantic subcategorisation. In *Proceedings of the Workshop SIGLEX-02 (ACL-02)*, 2002.

13. Gamallo P., Gasperin C., Agustini A., and Lopes G., Syntactic-based methods for measuring word similarity. In *Proceedings TSD-01*. Springer, 2001.
14. Genesereth M. and Nilsson N., *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann, 1987.
15. Gómez-Pérez A. and Manzano-Macho D., (eds.)., A survey of ontology learning methods and techniques. OntoWeb Deliverable #D1.5, Univ. Politécnica de Madrid, 2003.
16. Grishman R. and John Sterling J., Generalizing automatically generated selectional patterns. In *Proceedings of COLING-94*, 1994.
17. Guarino N. and Giaretta P., Ontologies and knowledge bases: Towards a terminological clarification. In Mars N., editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25 – 32, Amsterdam, 1995. IOS Press.
18. Humphreys B. and Lindberg D., The unified medical language system project: a distributed experiment in improving access to biomedical information. In Lun K.C., (ed.), *Proc. of the 7th World Congress on Medical Informatics (MEDINFO92)*, pp. 1496–1500, 1992.
19. Jarrar M. and Meersman R., Formal ontology engineering in the dogma approach. In Meersman R., Tari Z., and al., (eds.), *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE; Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings*, LNCS 2519, pp. 1238 – 1254. Springer, 2002.
20. Lin D., Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL-98*, 1998.
21. Maedche A. and Staab S., Discovering conceptual relations from text. Technical Report 399, Institute AIFB, Karlsruhe University, 2000.
22. Maedche A. and Staab S., Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16, 2001.
23. McCarthy D., Carroll J., and Preiss J., Disambiguating noun and verb senses using automatically acquired selectional preferences. SENSEVAL-2, 2001.
24. Miller G., Wordnet: a lexical database for english. *Comm. of the ACM*, 38(11):39–41, 1995.
25. Nazarenko A., Zweigenbaum P., Bouaud J., and Habert B., Corpus-based identification and refinement of semantic classes. In R. Masys, editor, *Proceeding of the AMIA Annual Fall Symposium - JAMIA Supplement*, pages 585–589. AMIA, 1997.
26. Pantel P. and Lin D., Discovering word senses from text. In *Proceedings of ACM SIGKDD-02*, 2002.
27. Peeters S. and Kaufner S., State of the art in crosslingual information access for medical information. Technical report, CSLI, 2001.
28. Pustejovsky J., *The Generative Lexicon*. MIT Press, 1995.
29. Reinberger M.-L., and Daelemans W., Is shallow parsing useful for the unsupervised learning of semantic clusters? In *Proceedings CICLing03*. Springer-Verlag, 2003.
30. Reinberger M.-L., Decadt B., and Daelemans W., On the relevance of performing shallow parsing before clustering. Computational Linguistics in the Netherlands 2002 (CLIN02), Groningen, The Netherlands, 2002.
31. Reiter R., *Readings in AI and Databases*, chapter Towards a Logical Reconstruction of Relational Database Theory. Morgan Kaufman, 1988.
32. Ribière M. and Charlton P., Ontology overview motorola labs. Technical report, Networking and Applications Lab - Centre de Recherche de Motorola Paris, 2000.
33. Spyns P., Meersman R., and Jarrar M., Data modelling versus ontology engineering. *SIGMOD Record Special Issue*, 31 (4), 2002.
34. Tarski A., *Problems in the philosophy of language*, chapter The semantic concept of truth. Holt, Rinehart & Winston, New York, 1969.
35. Ushold M. and Gruninger M., Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93 – 155, 1996.