Automatic initiation of an ontology

Marie-Laure Reinberger[1], Peter Spyns[2], A. Johannes Pretorius[2], and Walter Daelemans[1]

[1] University of Antwerp - CNTS,
Universiteitsplein 1, B-2610 Wilrijk - Belgium,
{Firstname.Lastname}@ua.ac.be
[2] Vrije Universiteit Brussel - STAR Lab,
Pleinlaan 2 Gebouw G-10, B-1050 Brussel - Belgium
{Firstname.Lastname}@vub.ac.be

Abstract. We report on an a set of experiments carried out in the context of the Flemish OntoBasis project. Our purpose is to extract semantic relations from text corpora in an unsupervised way and use the output as preprocessed material for the construction of ontologies from scratch. The experiments are evaluated in a quantitative and "impressionistic" manner.

We have worked on two corpora: a 13M words corpus composed of Medline abstracts related to proteins (SwissProt), and a small legal corpus (EU VAT directive) consisting of 43K words. Using a shallow parser, we select functional relations from the syntactic structure subject-verb-direct-object. Those functional relations correspond to what is a called a "lexon". The selection is done using prepositional structures and statistical measures in order to select the most relevant lexons. Therefore, the paper stresses the filtering carried out in order to discard automatically all irrelevant structures .

Domain experts have evaluated the precision of the outcomes on the SwissProt corpus. The global precision has been rated 55%, with a precision of 42% for the functional relations or lexons, and a precision of 76% for the prepositional relations. For the VAT corpus, a knowledge engineer has judged that the outcomes are useful to support and can speed up his modelling task. In addition, a quantitative scoring method (coverage and accuracy measures resulting in a 52.38% and 47.12% score respectively) has been applied.

Keywords: machine learning, text mining, ontology creation, quantitative evaluation, clustering, selectional restriction, co-composition.

1 Introduction

A recent evolution in the areas of artificial intelligence, database semantics and information systems is the advent of the Semantic Web [5]. It evokes "futuristic" visions of intelligent and autonomous software agents including mobile devices, health-care, ubiquitous and wearable computing. An essential condition to the actual realisation and unlimited use of these smart devices and programs is the possibility for interconnection and interoperability, which is currently still lacking to a large extent. Exchange of meaningful messages is only possible when the intelligent devices or agents share

a common conceptual system representing their "world" [3], as is the case for human communication. Meaning ambiguity should be, by preference, eliminated. Nowadays, a formal representation of such (partial) intensional definition of a conceptualisation of an application domain is called an ontology [25].

The development of ontology-driven applications is currently slowed down due to the knowledge acquisition bottleneck. Therefore, techniques applied in computational linguistics and information extraction (in particular machine learning) are used to create or grow ontologies in a period as limited as possible with a quality as high as possible. Sources can be of different kinds including databases and their schemas - e.g. [52], semi-structured data (XML, web pages), ontologies [4] and texts. Activities in the latter area are grouped under the label of Knowledge Discovery in Text (KDT), while the term "Text Mining" is reserved for the actual process of information extraction [28].

This paper wants to report on a joint research effort on the learning of ontologies from texts by VUB STAR Lab and UA CNTS during the Flemish IWT OntoBasis project [5]. The experiments concern the extraction and clustering of natural language terms into semantic sets standing for domain concepts as well as the detection of conceptual relationships. For this aim, the results of shallow parsing techniques are combined with unsupervised learning methods [45, 44].

The remainder of this paper is organised as follows. The next section (2) gives an overview of research in the same vein (section 2.1). Methods and techniques including others than the ones applied for this paper are mentioned (section 2.2). In section 3, a short overview of the DOGMA ontology engineering framework is given as it is the intention that the experiments described in this paper lead to a less time consuming process to create DOGMA-inspired ontologies. The objectives are presented in section 4.1, while the methods and material (section 4.2) as well as the evaluation techniques (sections 5.2 and 5.3) are explained. The results are described in sections 6.1 and 6.2. Related work (section 7) is presented. Indications for future research are given in section 8, and some final remarks conclude (section 9) this paper.

2 Background

2.1 Overview of the field

Several centres worldwide are actively researching on KDT for ontology development (building and/or updating). An overview of 18 methods and 18 tools for text mining with the aim of creating ontologies can be found in [22]. A complementary overview is provided in [33] [6]. It is worth to mention that in France important work (mostly applied to the French language) is being done by members of the TIA ("Terminologie et Intelligence Artificielle") working group of the French Association for Artificial

[3] See [51] for more details on the semantics of the Semantic Web.

[4] This is called ontology aligning and merging - e.g. [41]

[5] see http://wise.vub.ac.be/ontobasis

[6] We refer the interested reader to these overviews rather than repeating all the names of people and tools here.

Intelligence (AFIA) [7]. TIA regroups several well known institutes and researchers included in the overviews mentioned above and organises at a regular basis "Ontologies and Texts" (OLT) workshops linked to major AI-conferences (e.g., EKAW2000 [1], ECAI2002 [2]). Other important workshops on ontology learning were linked to ECAI2000 [49] and IJCAI2001 [34].

In addition to tools and researchers listed in the two overviews, there are the EU IST projects Parmenides [8] and MuchMore [9]. These projects have produced interesting state-of-the-art deliverables on KDT [27] - in particular section 3 - and related NLP technology [40]. The NLP groups of the University of Sheffield and UMIST (Manchester) are also active in this area [8, 28]. A related tool is SOOKAT, which is designed for knowledge acquisition from texts and terminology management [39]. A specific corpus-based method for extracting semantic relationships between words is explained in [19]. Mining for semantic relationships is also - albeit in a rather exploratory way - addressed in the Parmenides project [46].

2.2 Overview of methods

In essence, one can distinguish the following steps in the process of learning ontologies from texts (that are in some way or another common to the majority of methods reported):

1. collect, select and preprocess an appropriate corpus
2. discover sets of equivalent words and expressions
3. validate the sets (establish concepts) with the help of a domain expert
4. discover sets of semantic relations and extend the sets of equivalent words and expressions
5. validate the relations and extended concept definitions with the help of a domain expert
6. create a formal representation

Not only the terms, concepts and relationships are important, but equally the circumscription (gloss) and formalisation (axioms) of the meaning of a concept or relationship. On the question how to carry out these steps, a multitude of answers can be given. Many methods require a human intervention before the actual process can start (labelling seed terms - supervised learning, compilation/adaptation of a semantic dictionary or grammar rules for the domain ,...). Unsupervised methods don't need this preliminary step - however, the quality of their results is still worse. The corpus can preclude the use of some techniques: e.g., machine learning methods require a corpus to be sufficiently large - hence, some authors use the Internet as additional source [14]. Some methods require the corpus to be preprocessed (e.g., adding POS tags, identifying sentence ends, ...) or are language dependent (e.g., compound detection). Again, various ways of executing these tasks are possible (e.g., POS taggers can be based on handcrafted rules, machine-induced rules or probabilities). In short, many linguistic engineering tools can

[7] http://www.biomath.jussieu.fr/TIA
[8] http://www.crim.co.umist.ac.uk/parmenides/
[9] http://muchmore.dfki.de/demos.htm

be put to use. To our knowledge no comparative study has been published yet on the efficiency and effectiveness of the various techniques applied to ontology learning.

Selecting and grouping terms can be done by means of tools based on distributional analysis, statistics, machine learning techniques, neural networks, and others. To discover semantic relationships between concepts, one can rely on valency knowledge, already established semantic networks or ontologies, co-occurrence patterns, machine readable dictionaries, association patterns or combinations of all these. In [28] a concise overview is offered of commercially available tools that are useful for these purposes. Due to space restrictions, we will not discuss in this paper how the results can be transformed in a formal model (e.g., see [3] for an overview of ontology representation languages).

3 DOGMA

Before presenting the actual text mining experiments, we want to shortly discuss the framework for which the results of the experiments are meant to be used, i.e. the *DOGMA* (Developing Ontology-Guided Mediation for Agents) ontology engineering approach [10]. Within the DOGMA approach, preference is given to texts as objective repositories of domain knowledge instead of referring to domain experts as exclusive knowledge sources [11]. Apparently, this preference is rather recent [1] and probably more popular in language engineering circles (see e.g. [11]).

Notice that also restrictions on a semantic relationship, e.g. indicating its mandatory aspect or its cardinality, should be mined from the corpus. These constraints serve to define more precisely the concepts and relations in the ontology. This is a step that should be added before the formal model is created, and that currently is hardly mentioned in the KDT literature. But one will easily agree that, e.g. when modelling a law text, there can be a huge difference between "must" and "may". This issue will not be further addressed in the present paper.

The results of the unsupervised mining phase are represented as *lexons*. These are binary fact types indicating which are the entities and the roles they assume in a semantic relationship [48].

Formally, a lexon is described as $< (\gamma, \lambda) : \ term_1 \ role \ co{-}role \ term_2 >$. For the sake of brevity, abstraction will be made of the context (γ) and language (λ) identifiers. For the full details, we refer to [6]. Informally we say that a lexon expresses that the $term_1$ (or head term) may plausibly have $term_2$ (or tail term) occur in an associating $role$ (with $co - role$ as its inverse) with it. The basic insights of DOGMA originate from database theory and model semantics [35].

In the near future, a strict distinction in the implementation of the DOGMA ontology server will be made between concept labels and natural language words or terms [6]. In many cases, "term" is interpreted in the ontology literature as "logical term" (or concept) of the ontology first order vocabulary and, at the same time, as a natural language term. Without going too much in detail here, we separate the conceptual

[10] see http://www.starlab.vub.ac.be/research/dogma

[11] This does not imply that texts will be the sole source of knowledge.

level from the linguistic level (by using WordNet-like synsets - see also [20]), which has its impact on the KDT process, namely in step (3) mentioned in section 2.2. One of the rather rare KDT methods that also takes this distinction into account is described in [37]. It is easy to understand that the first step to initiate an ontology is situated on the linguistic level: lexons constitute a necessary but intermediary step in the process of creating a (language-independent) conceptualisation and its corresponding implemented artefact, i.e. an ontology [25].

4 Unsupervised Text Mining

In the following sections, we will report on experiments with unsupervised machine learning techniques based on results of shallow parsing.

4.1 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. 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 this knowledge, depending on its future use and on the specificity of the domain.

Currently, the focus is on the discovery of concepts and their conceptual relationships, although the ultimate aim is to discover semantic constraints as well. We have opted for extraction techniques based on unsupervised learning methods [45] since these do not require specific external domain knowledge such as thesauri and/or tagged corpora [12]. As a consequence, the portability of these techniques to new domains is expected to be much better [40, p.61].

4.2 Material and methods

The ***linguistic assumptions*** underlying this approach are

1. the principle of selectional restrictions (syntactic structures provide relevant information about semantic content), and
2. the notion of co-composition [43] (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 [17, 18]. 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

[12] Except the training corpus for the general purpose shallow parser.

its physical feature. We make use of both principles in our use of clustering to extract semantic knowledge from syntactically analysed corpora.

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 [9, 10, 13] [13]. 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.

Different methods can be used for the ***extraction of semantic information*** from parsed text. Pattern matching [4] has proved to be a efficient way to extract semantic relations, but one drawback is that it involves the predefined choice of the semantic relations that will be extracted. On the other hand, clustering only requires a minimal amount of "manual semantic pre-processing"by the user. We rely on a large amount of data to get results using pattern matching and clustering algorithms on syntactic contexts in order to also extract previously unexpected relations. Clustering on terms can be performed by using different syntactic contexts, for example noun+modifier relations [12] or dependency triples [30]. As mentioned above, the shallow parser detects the subject-verb-object structures, which gives us the possibility to focus in a first step on the term-verb relations with the term appearing as the head of the object phrase. This type of structure features a functional relation between the verb and the term appearing in object position, and allows us to use a clustering method to build classes of terms sharing a functional relation. Next, we attempt to enhance those clusters and link them together, using information provided by prepositional structures.

The SwissProt corpus (see below) provides us with a huge number of those syntactic structures associating a verb to two nominal strings (NS), namely the subject nominal string (SNS) and the object nominal string (ONS). A nominal string is the string composed of nouns and adjectives appearing in a NP, the last element being the head noun of the NP.

However, we have to deal with the fact that the parser also produces some mistakes (f-score for objects is 80 to 90%), and that not all verb-object structures are statistically relevant. Therefore, we need to find a way to select the most reliable dependencies, before applying to them automatic techniques for the extraction of ontological relations. This step can be achieved with the help of pattern matching techniques and statistical measures.

Therefore, the stress is put in this experiment on the operation of filtering we are carrying out through pattern matching and statistical measures in order to discard automatically the irrelevant lexons. In a first step, we apply a pattern on the corpus in order to retrieve all the syntactic structures: NS-Preposition-NS. This structure has been chosen for its high frequency and because it generates few mistakes from the parser.

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

In a second step, the most relevant prepositional structures NS1-P-NS2 are selected, using a statistical measure. We want this measure to be high when the prepositional structure is coherent, or when NS1-P-NS2 appears more often than NS1-P and P-NS2. Therefore, it takes into account the probability of appearance of the whole prepositional structure (#NS1-P-NS2), as well as the probability of appearance of the two terms composing the whole structure (#NS1-P and #P-NS2):

$$\frac{\frac{\#NS1-P-NS2}{\min(\#NS1,\#NS2)}}{\frac{\#NS1-P}{\#NS1}+\frac{\#P-NS2}{\#NS2}}$$

The final step consist in the selection of the lexons. We consider the N prepositional structures with the highest rate, and we elect the relevant lexons or SNS-Verb-ONS structures by checking if the SNS and the ONS both appear among the N prepositional structures selected by the statistical measure.

We have worked with the 13M words *SwissProt* ***corpus*** composed of Medline abstracts related to genes and proteins. In a specific domain, an important quantity of semantic information is carried by the noun phrases (NP). At the same time, the NP-verb relations provide relevant information about the NPs, due to the semantic restrictions they impose. Therefore, we applied to this corpus the memory based shallow parser mentioned above. This shallow parser gives us the possibility to exploit the subject-verb-object dependencies. The selectional restrictions associated with this structure imply that the NPs co-occurring, as the head of the object, with a common set of verbs, share semantic information. This semantic information can be labeled as "functional", due to the semantic role of the verb, and therefore refers to the notion of "lexon" we have described in section 3. The smaller *VAT* corpus consists of 43K words. It constitutes the EU directive on VAT that has to be adopted and transformed into local legislation by every Member State. The VAT corpus has been chosen to validate the results of the unsupervised mining process on the SwissProt corpus.

5 Evaluation Criteria

5.1 Preliminary remarks

The main research hypothesis in this paper is that lexons, representing the basic binary facts expressed in natural language about a domain, can be extracted from the available textual sources. Thus, a first step is the discovery and grouping of relevant terms. Using the lexons, a domain expert will, in a second step, distill concepts and determine which relationships hold between the various newly discovered concepts. Unambiguous definitions have to be provided. 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 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 domain model is commonly agreed upon (third step). Formalising the model is a subsequent step. The following sections discuss how the mining results will be evaluated.

5.2 The SwissProt corpus

The results have been evaluated by experts of the biological domain. They were asked to consider a set of 261 relations corresponding to a subset of nominal strings appearing frequently in the corpus and including lexons as well as more general relations (spatial, part of...) issued from the prepositional relations. They had to rate each relation, regarding its relevance to the gene/protein domain as:

- false/irrelevant
- general information/weak relevance
- specific information/strong relevance

The subset of nominal strings considered for the evaluation contained every relation involving at least one of those keywords: DNA, cDNA, RNA, mRNA, protein, gene, ATP, polymerase, nucleotide, acid.

5.3 The EU VAT Directive corpus

In this section, a more empirical evaluation method will be provided. Criteria for ontology evaluation have been put forward by Gruber [24, p.2] and taken over by Ushold and Grüninger [50]: clarity, coherence, extendibility, encoding bias and minimal ontological commitment. Gómez-Pérez [21, p.179] has proposed consistency, completeness and conciseness. Neither set of criteria are well suited to be applied in our case as the lexons produced by the unsupervised miner are merely "terminological combinations" (i.e. no explicit definition of the meaning of the terms and roles are provided not to mention any formal definition of the intended semantics). We have been mainly inspired by the criteria proposed by Guarino (coverage, precision and accuracy) [26, p.7], although there are problems to "compute" them in the current practice (unlike in information extraction evaluation exercises) as there are no "gold standards" available.

Qualitative method Therefore, a human knowledge engineer has been asked to evaluate the practicality and usefullness of the results. A manually built lexon base is available, but this is a single person's work, which means that the "shared" and "commonly agreed" aspects - typical of an ontology - are lacking. Or stated in another way, a person - even an expert - maybe be wrong and therefore not the sole reference for a valid evaluation. Nevertheless, some questions have been formulated independently of the knowledge engineer/evaluator who is supposed to rely on his past experience. The evaluator/knowledge engineer was given a list of questions regarding the lexon bass as produced by the unsupervised miner.

- Do you think that w.r.t. the domain being modelled the lexon based produced is :
  - "covering" (are all the lexons there)
  - precise (are the lexons making sense for the domain)
  - accurate (are the lexons not too general but reflecting the important terms of the domain)

- concise (are the lexons not redundant [14])

– Would you have produced (more or less) the same lexons (inter-modeller agreement) ?
– Do you think that, using these lexons, ontology modelling happens faster (practicality)?
– Is it possible to create additional lexons from the original set to improve the coverage and accuracy while remaining precise ?

Note that this kind of evaluation implicitly requires an ontological commitment from the evaluator, i.e. he/she gives an intuitive understanding to the terms and roles of the lexons.

Quantitative method In addition, for the coverage and accuracy criteria we have tried to define a quantative measure and semi-automated evaluation procedure that will be explained subsequently. We don't define a computable precision measure here (see [45] for an earlier attempt). The underlying idea is inspired by Zipf's law [54]. It states that the frequency of the occurrence of a term is inversely proportional with its frequency class. Zipf has discovered experimentally that the more frequently a word is used, the less meaning it carries. E.g., the word "the" appears 3573 times and there is only 1 element in the frequency class 3573. "by-product" and "chargeability" occur only once, but there are 1155 words in the frequency class 1. Important for our purpose is the observation that the higher frequency classes contain mostly "empty"(also called function words). A corollary is that domain or topic specific vocabulary is to be looked for in the middle to lower frequency classes (see also [31, 32]).

As the DOGMA lexons resulting from the unsupervised mining consist of three words[15] (two terms and one role[16]) extracted from the corpus, it is possible to investigate to what extent the produced lexons cover the corpus vocabulary, and more importantly how accurate they are. Note that the same technique can be applied to RDFS ontologies.

Coverage will be measured by comparing for each frequency class the number of terms from the lexons with the number of terms from the corpus. Accuracy will be estimated on basis of the coverage percentage for particular frequency classes. However, some caveats should be made from the on-set. It should be clear that a coverage of 100% is an illusion. Only terms in a V-O and S-O grammatical relation are selected and submitted subsequently to several selection thresholds (see section 4.2). Regarding the accuracy, determining exactly which frequency classes contain the terms most characteristic for the domain is still a rather impressionistic and intuitive entreprise. It should be kept in mind that no stopword list has been defined because lexons have been produced with a preposition assuming the role function.

[14] This could be a tricky criterion as the terms and roles can have synonyms.
[15] In fact, the words have been lemmatised, i.e. reduced to their base forms. E.g., working, works, worked $\rightarrow$ work.
[16] Co-roles are not provided.

6 Results

Evaluation typically has to do with avoiding all kinds of biases (e.g., the evaluator and developer is the same person, there is only one evaluator, evaluation is only done on machine produced output, etc. [16]). The results on the SwissProt and VAT corpus have been given to domain experts and a knowledge engineer for a qualitative evaluation. In addition, the quantitative measures (as defined in the previous section) have been applied on the VAT results. Below, the outcomes of the evaluation rounds are presented.

6.1 The SwissProt corpus

Among the 261 relations that have been evaluated, we count 165 lexons and 96 other relations. What we obtain is a global precision of 55%, of which 47% have been evaluated as specific information, and 8% as general information. If we consider the lexons, we have a precision of 42%, with 35% of specific relations and 7% of general relations. Finally, considering the other relations, the precision is 76%, with 67% of specific relations and 9% of general relations.

Here are some examples of relations evaluated as specific information:

- DNA_damage induce transcription
- amino_acid_sequence reveal significant_homology
- fusion_protein with glutathione_S-transferase
- oligonucleotide_probe from N-terminal_amino_acid_sequence

And some examples of relations evaluated as general information:

- DNA contain human_chromosome
- amino_acid_sequence provide support
- uracil into DNA
- asparagine for aspartic_acid

6.2 The EU VAT Directive corpus

Qualitative method When applied to the VAT corpus, the unsupervised mining exercise outlined above resulted in the extraction of 817 subject-verb-object structures. These were analysed by a knowledge engineer using the LexoVis lexon visualisation tool [42]. This analysis was rather informal in the sense that the knowledge engineer was largely guided by his intuition, knowledge and experience with the manual extraction of lexons from the VAT legislature domain.

A first important aspect to consider is whether the domain (VAT legislature) is adequately described (or *covered*) by the set of extracted triples. In this regard, it soon became apparent that there is a significant amount of noise in the mining results; the triples need to be significantly cleaned up in order to get rid of inadequate (and often humorous) structures such as <*fishing, with, exception*>. The percentage of inadequate triples seems to fall in excess of 53%. According to this percentage, approximately 384 of the resulting 817 triples may be deemed usable. If this is compared to the number of lexons resulting from a manual extraction exercise on the same corpus of knowledge

resources (approximately 900) there is doubt as to whether the domain is adequately covered by the results. As mentioned above, there is a significant portion of the unsupervised mining exercise results which are deemed inadequate. Firstly, this can be contributed to the fact that many resulting triples are not *precise* (intuitively, they do not make sense in the context of the VAT domain as the fishing example above illustrates). Furthermore, many of resulting triples were not considered *accurate* in the sense of describing important terms of the domain. In this respect, only the term VAT only occurs in three subject-verb-object structures, <*VAT, in, member*>, <*VAT, on, intra-Community_acquisition*> and <*VAT, to, hiring*> which are not considered appropriate to accurately describe the concept of VAT in the domain under consideration. In the same respect, there is only one mention of the term *Fraud*. In essence, the triples analysed form numerous disconnected graphs instead of one coherent and richly connected semantic network. The view is held that significant additions, in terms of roles, will need to be made in order to ensure that all applicable interrelationships in the domain are described. In this same vein, it is the case that no co-roles are defined. Clearly it would be a great advantage if this were the case. In Figure 1, the interpretation of the visual representation from left to right suggests that for any triplet $< t_i, r_{i-j}, t_j >$, the ontology engineer simply identifies t_j on the left arc, t_i on the right arc (or for a particular term in the object position, identify the same term in all subject positions). Consequently, r_{j-i} should then be presented. In this way, - triplets may be combined to form lexons in which co-roles are also defined. However, as is evident from the symmetry of the visual representation in Figure 1, this is seldom the case [42].

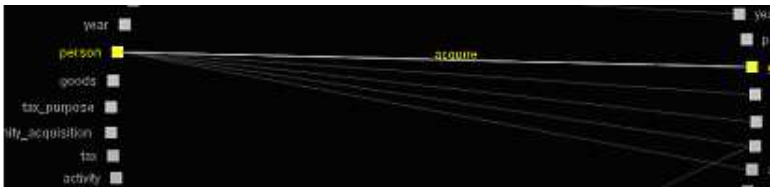

Fig. 1. irreversibility

For instance, in the triplet <*person, acquire, goods*> exists, but there is no triplet of the form <*goods, acquire*$^{-1}$*, person*>, where *acquire*$^{-1}$ signifies the inverse or co-role of *acquire* (see Figure1). This has the implication that in order to finalize the lexon base with which to describe the VAT domain, the knowledge engineer has to consider all machine extracted triplets in order to define co-roles, which could be quite an arduous task. However, a triplet such as <*person, acquire, goods*> does intuitively suggest a lexon of the form <*person, acquire, acquired_by, goods*> which should lessen the cognitive overhead required from the knowledge engineer. Furthermore, it is often the case that in the set of triples resulting from unsupervised mining of the VAT corpus, instances are identified rather than instance types. For example, the body of lexons includes the triplet <*Republic, of, Austria*>. Although this is clearly not satisfactory, such a triplet does suggest to the ontology engineer the inclusion of a lexon such as

<country, isA, isA, republic>. It is striking that many roles take the form of prepositions. This includes triplets such as, *<application, of, exemption>*, *<adjustment, in, purchaser>*, *<agricultural_product, for, derogation>*, *<agricultural_product, of, agricultural_service>*, *<electronic_mean, to, data>*. Even though this might be conceptually correct, there exist many richer roles in a domain such at VAT legislature. One example might be *<agricultural_product, yields, agricultural_service>*, for instance.

Finally, the notion of *redundancy* is harder to evaluate, since terms and roles may have synonyms. However, the intuitive impression of the results was that redundancy was not a critical problem. In conclusion, the subject-verb-object structures resulting from unsupervised mining of the VAT corpus is not considered sufficient to represent the VAT domain. Even though the number of resulting triples approach the number manually extracted form the same texts, the impreciseness, inaccuracy and inconciseness results in many not being usable. However, the above analysis does have interesting methodological implications. Indeed, it suggests a *subtractive* approach to ontology engineering. That is, as opposed to an *additive* approach where the ontology engineer starts with an empty set of lexons to which he or she adds lexons to describe an universe of discourse.

Instead, the lexons resulting from a machine extraction exercise presents the ontology engineer with an initial corpus of lexons. These lexons are analysed, noise in the form of meaningless lexons removed or annotated, and new lexons added. In this regard, it is contended that through the analysis of such an initial body of lexons other lexons may be suggested to the ontology engineer and subsequently added to the resulting ontology base. Such an approach could significantly reduce the time investment needed from the knowledge engineer, since he or she does not have to start from scratch. It is further held that if unsupervised mining approaches such as those outlined in this paper can guarantee consistent results (that is, the same algorithm applied to the same corpus at different time instances results in similar results), then the knowledge engineer would be able to come up with an initial set of lexons by a process of elimination. Based on this set of lexons, the ontology engineer can then proceed to ensure that the domain is adequately described by considering this set.

Quantitative method In order to produce illustrative graphics the highest frequency classes have been omitted (e.g., starting from class 300: member (336), which (343), article (369), taxable (399), person (410), tax (450), good (504), by (542), will (597), a (617), for (626), or (727), and (790), be (1110), in (1156), to (1260), of (2401), and the (3573)). At the other end, the classes 1 to 4 are also not displayed: class 1 containing 1165 lemmas, class 2 356, class 3 200 and class 4 has 132 members. Also some non-word tokens have been removed (e.g., 57.01.10, 6304, 7901nickel, 2(1, 8(1)(c, 2(2)). However, some of these non-word tokens have survived (which might influence the outcomes, especially in the lowest frequency classes).

The content of the frequency classes (FC) shows that they be can rated "content-wise" as follows:

- $FC < 3$: many non-words and/or too loosely related to the domain
- $3 < FC < 20$: domain related technical language
- $20 < FC < 50$: general language used in a technical sense

– $50 < FC < 300$: mixture of general language and domain technical language
– $300 < FC < 500$: general language and highly used domain terms
– $FC < 500$: function words and highly used general language terms

We determine the area with "resolving power of significant words" [32, p.16] to be the range of frequency classes 3 till 40. The range encompasses 596 terms that one would expect to be covered by the lexons. Figures 2 and 3 show that the coverage improves with the increasing rank of the frequency class. On average, the coverage ratio is 52.38%. The accuracy (i.e. the coverage percentage for the selected interval) ratio for the 3-40 interval is 47.31%.

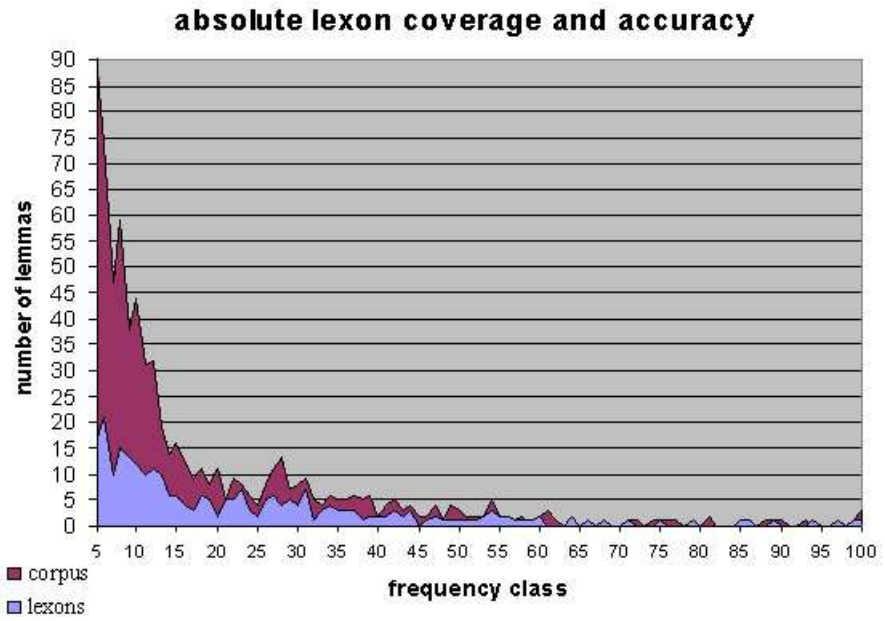

Fig. 2. absolute coverage and accuracy of frequency classes by lexon terms

7 Discussion & 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. At the same time, we turn to pattern matching in order to label semantic relations. 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 ([30]). Clustering does not provide you with the relations

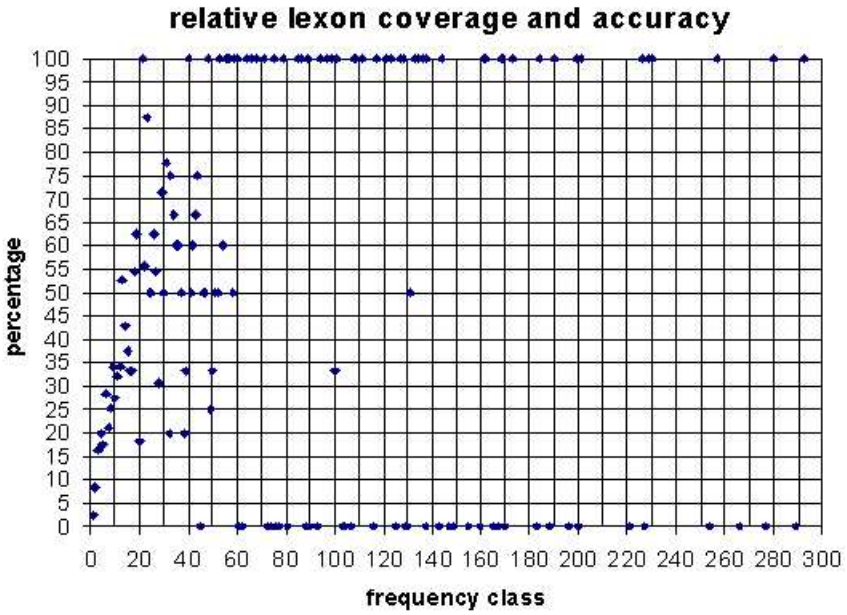

Fig. 3. relative coverage of frequency classes by lexon terms

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 [7, 15], or by the researchers themselves [23]. An automatic evaluation is nevertheless performed in [30], by comparison with existing thesauri like WordNet and Roget. Our attempt takes the corpus itself as reference and reduces the need for human intervention. Humans are still needed to clean the corpus (e.g. to choose the stopwords and to remove the non-words), but do not intervene in the evalution process itself, except for setting the frequency class interval. Regression tests can be done. Currently, we estimate that the accuracy should be improved. Taking synonyms into account might help. On the other hand, more research should be done to determine the proportion of domain technical terms vs. general language terms in the ”relevant” frequency class interval. If we look at it from a positive angle, we could argue that already half of the work of the domain specialist and/or terminographer to select the important domain terms is done. We were specifically (but happily) surprised by the fact that the different evaluation techniques performed in an independent way lead to similar conclusions.

8 Future Work

Some topics for future work can be easily sketched. From the work flow point of view, the lexons resulting from the unsupervised mining should be entered into an ontology

modelling workbench that includes appropriate visualisation tools [42] and hooks to thesauri, controlled vocabularies and dictionaries, e.g. (Euro)WordNet [53, 36]), on the one hand and (formal) upper ontologies, e.g. SUMO [38] or CyC [29] on the other. This workbench embodies the DOGMA ontology engineering methodology (see [47] for a limited illustration).

With respect to the quantitative evaluation of the outcomes of the unsupervised mining, insights from information science technology should be taken into account to answer some questions. E.g. does the length of a document influence the determination of the most meaningful frequency class interval ? Is it possible to establish a statistical formula that represents the distribution of meaningful words over documents ?

Once this interval can be reliably identified, one could apply the unsupervised learning algorithm only to sentences containing words belonging to frequency classes of the interval. This could be easily done after having made a concordance (keyword in context) for the corpus.

Part of the mistakes is due to the difficulty of parsing negative and passive forms. In the future, we are planning to increase the global number of structures, by considering also the verbal structures introducing a complement with a preposition. Also, spatial and part_of relationships should become more precise.

9 Conclusion

We have presented the results of an experiment on initiating an ontology by means of unsupervised learning. In addition, we have performed both a qualitative and quantitative evaluation of the outcomes of the mining algorithm applied to a protein and a financial corpus. The results can be judged as moderately satisfying. We feel that unsupervised semantic information extraction helps to engage the building process of a domain specific ontology. Thanks to the relatedness of a DOGMA lexon and an RDF triple, the methods proposed above can also be applied to ontologies represented in RDF(S).

Acknowledgments The major 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). Hannes Pretorius is supported by the EU IST FP5 FF Poirot project (IST-2001-38248).

References

1. Nathalie Aussenac-Gilles, Brigitte Bi´ebow, and Sylvie Szulman, editors. *EKAW'00 Workshop on Ontologies and Texts*, volume http://CEUR-WS.org/Vol-51/. CEUR, 2000.
2. Nathalie Aussenac-Gilles and Alexander Maedche, editors. *ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, volume http://www.inria.fr/acacia/OLT2002, 2002.
3. Sean Bechhofer (ed.). Ontology language standardisation efforts. OntoWeb Deliverable #D4, UMIST - IMG, Manchester, 2002.
4. Matthew Berland and Eugene Charniak. Finfing parts in very large corpora. In *Proceedings ACL-99*, 1999.

5. T. Berners-Lee. *Weaving the Web*. Harper, 1999.
6. Jan De Bo and Peter Spyns. Creating a ”dogmatic” multilingual ontology to support a semantic portal. In Z. Tari et al. R. Meersman, editor, *On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops*, volume 2889 of *LNCS*, pages 253 – 266. Springer Verlag, 2003.
7. Didier Bourigault and Christian Jacquemin. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings EACL-99*, 1999.
8. Christopher Brewster, Fabio Ciravegna, and Yorick Wilks. User centred ontology learning for knowledge management. In Birger Andersson, Maria Bergholtz, and Paul Johannesson, editors, *Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems (NLDB 2002) - Revised Papers*, volume 2553 of *LNCS*, pages 203 – 207. Springer Verlag, 2002.
9. Sabine Buchholz. *Memory-Based Grammatical Relation Finding*. 1999.
10. Sabine Buchholz, Jorn Veenstra, and Walter Daelemans. Cascaded grammatical relation assignment. PrintPartners Ipskamp, 2002.
11. Paul Buitelaar, Daniel Olejnik, and Michael Sintek. A Prot´eg´e plug-in for ontology extraction from text based on linguistic analysis. In Frank Van Harmelen, Sheila McIlraith, and Dimitris Plexousakis, editors, *Proceedings of the Internal Semantic Web Conference 2004*, LNCS. Springer Verlag, 2004.
12. Sharon A. Caraballo and Eugene Charniak. Determining the specificity of nouns from text. In *Proceedings SIGDAT-99*, 1999.
13. Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. Memory-based shallow parsing. In *Proceedings of CoNLL-99*, 1999.
14. A. Dingli, F. Ciravegna, David Guthrie, and Yorick Wilks. Mining web sites using adaptive information extraction. In *Proceedings of the 10th Conference of the EACL*, 2003.
15. David Faure and Claire N´edellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proceedings EKAW-99*, 1999.
16. C. Friedman and G. Hripcsak. Evalutating natural language processors in the clinical domain. *Methods of Information in Medicine*, 37:334 – 344, 1998.
17. Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Selection restrictions acquisition from corpora. In *Proceedings EPIA-01*. Springer-Verlag, 2001.
18. Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Using co-composition for acquiring syntactic and semantic subcategorisation. In *Proceedings of the Workshop SIGLEX-02 (ACL-02)*, 2002.
19. Pablo Gamallo, Marco Gonzalez, Alexandre Agustini, Gabriel Lopes, and Vera de Lima. Mapping syntactic dependencies onto semantic relations. In Nathalie Aussenac-Gilles and Alexander Maedche, editors, *ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, volume http://www.inria.fr/acacia/OLT2002, 2002.
20. Aldo Gangemi, Roberto Navigli, and Paola Velardi. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In Robert Meersman, Zahir Tari, and Douglas Schmidt et al. (eds.), editors, *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, number 2888 in LNCS, pages 820 – 838, Berlin Heidelberg, 2003. Springer Verlag.
21. Asunci´on G´omez-P´erez, Mariano Fern´andez-L´opez, and Oscar Corcho. *Ontological Engineering*. Advanced Information and Knowledge Processing. Springer Verlag, 2003.
22. Asunci´on G´omez-P´erez and David Manzano-Macho (eds.). A survey of ontology learning methods and techniques. OntoWeb Deliverable #D1.5, Universidad Polit´ecnica de Madrid, 2003.

23. Ralph Grishman and John Sterling. Generalizing automatically generated selectional patterns. In *Proceedings of COLING-94*, 1994.
24. T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221, 1993.
25. N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In N. Mars, editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25 – 32, Amsterdam, 1995. IOS Press.
26. Nicola Guarino and Andreas Persidis. Evaluation framework for content standards. Technical Report OntoWeb Deliverable #3.5, Padova, 2003.
27. Haralampas Karanikas, Myra Spiliopolou, and Babis Theodoulidis. Parmenides system architecture and technical specification. Parmenides Deliverable #D22, UMIST, Manchester, 2003.
28. Haralampos Karanikas and Babis Theodoulidis. Knowledge discovery in text and text mining software. Technical report, UMIST - CRIM, Manchester, 2002.
29. D.B. Lenat and R. V. Guha. *Building Large Knowledge Based Systems*. Addison Wesley, Reading, Massachusetts, 1990.
30. Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL-98*, 1998.
31. Robert Losee. Term dependence: A basis for luhn and zipf models. *Journal of the American Society for Information Science and Technology*, 52(12):1019 – 1025, 2001.
32. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159 – 195, 1958.
33. Alexander Maedche. *Ontology Learning for the Semantic Web*, volume 665 of *The Kluwer International Series in Engineering and Computer Science*. Kluwer International, 2003.
34. Alexander Maedche, Steffen Staab, Claire N´edellec, and Ed Hovy, editors. *IJCAI'01 Workshop on Ontology Learning*, volume http://CEUR-WS.org/Vol-38/. CEUR, 2001.
35. Robert Meersman. Ontologies and databases: More than a fleeting resemblance. In A. d'Atri and M. Missikoff, editors, *OES/SEO 2001 Rome Workshop*. Luiss Publications, 2001.
36. G. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39 – 41, 1995.
37. Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22 – 31, 2002.
38. I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, 2001.
39. Päivikki Parpola. Managing terminology using statistical analyses, ontologies and a graphical ka tool. In Nathalie Aussenac-Gilles, Brigitte Bi´ebow, and Sylvie Szulman, editors, *EKAW'00 Workshop on Ontologies and Texts*, volume http://CEUR-WS.org/Vol-51/. CEUR, 2000.
40. Stanley Peeters and Stefan Kaufner. State of the art in crosslingual information access for medical information. Technical report, CSLI, 2001.
41. H. Pinto, A. G´omez-P´erez, and J.P. Martins. Some issues on ontology integration. In R. Benjamins and A. G´omez-P´erez, editors, *Proceedings of the IJCAI'99 Workshop on Ontology and Problem-solving methods: lesson learned and future trends*, pages 7.1–7.11. CEUR, 1999.
42. A.Johannes Pretorius. Lexon visualization: visualizing binary fact types in ontology bases. In *Proceedings of the 8th international conference on information visualisation (IV04)*, London, 2004. IEEE Press. In press.
43. James Pustejovsky. *The Generative Lexicon*. MIT Press, 1995.

44. Marie-Laure Reinberger and Peter Spyns. Discovering knowledge in texts for the learning of dogma-inspired ontologies. In Paul Buitelaar, Siegfried Handschuh, and Bernardo Magnini, editors, *Proceedings of the ECAI04 Workshop on Ontologies, Learning and Population*, 2004.
45. Marie-Laure Reinberger, Peter Spyns, Walter Daelemans, and Robert Meersman. Mining for lexons: Applying unsupervised learning methods to create ontology bases. In Robert Meersman, Zahir Tari, and Douglas Schmidt et al. (eds.), editors, *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, number 2888 in LNCS, pages 803 – 819, Berlin Heidelberg, 2003. Springer Verlag.
46. Fabio Rinaldi, Karel Kaljurand, James Dowdall, and Michael Hess. Breaking the deadlock. In Robert Meersman, Zahir Tari, and Douglas Schmidt et al. (eds.), editors, *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, number 2888 in LNCS, pages 876 – 888, Berlin Heidelberg, 2003. Springer Verlag.
47. Peter Spyns, Sven Van Acker, Marleen Wynants, Mustafa Jarrar, and Andriy Lisovoy. Using a novel orm-based ontology modelling method to build an experimental innovation router. In Enrico Motta and Nigel Shadbollt, editors, *Proceedings of EKAW 2004*, LNAI. Springer Verlag, 2004. in press.
48. Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *SIGMOD Record Special Issue*, 31 (4): 12 - 17, 2002.
49. Steffen Staab, Alexander Maedche, Claire N´edellec, and Peter Wiemer-Hastings, editors. *Proceedings of the Workshop on Ontology Learning*, volume http://CEUR-WS.org/Vol-31/. CEUR, 2000.
50. M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge Sharing and Review*, 11(2), June 1996.
51. Michael Ushold. Where are the semantics in the semantic web? *AI Magazine*, 24(3):25 – 36, 2003.
52. R. Volz, S. Handschuh, S. Staab, L. Stojanovic, and N. Stojanovic. Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1:187 – 206, 2004.
53. P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.
54. George K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA, 1949.