VRIJE UNIVERSITEIT BRUSSEL

FACULTEIT WETENSCHAPPEN

VAKGROEP INFORMATICA EN TOEGEPASTE INFORMATICA
SEMANTICS TECHNOLOGY AND APPLICATIONS RESEARCH LAB

STAR Lab Technical Report

Lexically evaluating ontology triples generated automatically from texts

Peter Spyns & Marie-Laure Reinberger°

affiliation:	°: Universiteit Antwerpen - CNTS, Universiteitsplein 1, B-2610 Wilrijk, Belgium, tel.: +32-3- 820.2766; fax: +32-3-820.2762
keywords	ontology evaluation
number	STAR-2005-04
date	24/03/2005
corresponding author	Peter Spyns
status	final
reference	Gómez-Pérez A. & Euzenat J., (eds.), Proceedings of the 2nd European Semantic Web Conference, LNCS 3532, Springer, pp. 563-577

Lexically Evaluating Ontology Triples Generated Automatically from Texts

Peter Spyns[1] and Marie-Laure Reinberger[2]

[1] Vrije Universiteit Brussel - STAR Lab,
Pleinlaan 2 Gebouw G-10, B-1050 Brussel - Belgium
tel.: +32-2-629.1237; fax: +32-2-629.3819
Peter.Spyns@vub.ac.be
[2] University of Antwerp - CNTS,
Universiteitsplein 1, B-2610 Wilrijk - Belgium
tel.: +32-3- 820.2766; fax: +32-3-820.2762
marielaure.reinberger@ua.ac.be

Abstract. Our purpose is to present a method to lexically evaluate the results of extracting in an unsupervised way material from text corpora to build ontologies. We have worked on a legal corpus (EU VAT directive) consisting of 43K words. The unsupervised text miner has produced a set of triples. These are to be used as preprocessed material for the construction of ontologies from scratch. A quantitative scoring method (coverage, accuracy, recall and precision metrics resulting in a 38.68%, 52.1%, 9.84% and 75.81% scores respectively) has been defined and applied.

1 Introduction and Background

A recent evolution in the areas of artificial intelligence, database semantics and information systems is the advent of the Semantic Web [1]. It evokes "futuristic" visions of intelligent and autonomous software agents including mobile devices, health-care monitoring, ubiquitous and wearable computing. E.g., a heartbeat monitoring device integrated in a person's shirt could trigger, in case of observed rhythm deviations, a web agent that schedules an appointment with his/her doctor via the mobile network.

An essential condition to the actual realisation and unlimited use of these smart devices and programs is the possibility for interoperability, which is currently still lacking to a large extent. Indeed, intelligent agents have to be able to exchange "meaningful" messages[1] while continuing to function autonomously (interoperability with local autonomy as opposed to integration with central control). Exchange of meaningful messages is only possible when the intelligent devices or agents share a common conceptual system representing their "world"[2], as is the case for human communication. Meaning

[1] We make abstraction here of the feasibility of physically connecting these devices and services or agents to a (global) network.

[2] See [28] for more details on the semantics of the Semantic Web.

A. Gómez-Pérez and J. Euzenat (Eds.): ESWC 2005, LNCS 3532, pp. 563–577, 2005.

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 [10].

The development of ontology-driven applications is currently slowed down due to the knowledge acquisition bottleneck. Indeed, the process of conceptualising an application domain and its formalisation need substantial human resources and efforts. 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, semi-structured data (XML, web pages), ontologies[3] 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 extracting the information [14].

In addition, there is hardly any method available to thoroughly evaluate the results of (unsupervised) text mining for ontologies. We have looked to the domain of information science to suggest a quantitative method - see [21, 25] - that will be refined in this paper. Previoulsy, criteria for ontology evaluation have been put forward by Gruber [9–p.2] and taken over by Ushold and Grüninger [27]: clarity, coherence, extendibility, minimal encoding bias and minimal ontological commitment. Gómez-Pérez [8–p.179] has proposed consistency, completeness and conciseness. Neither set of criteria are well suited to be applied in our case as the triples produced by the unsupervised miner are merely "terminological combinations" (i.e., no explicit meaning for the terms and roles is provided, not to mention any formal definition of the intended semantics). Recent proposals for evaluation methods have been discussed during the ECAI2004 workshop on ontology learning and population [4]. The majority of them proposes to evaluate an ontology mediating improvement measures of an existing application or by a comparison with another ontology acting as a gold standard. Typical of our approach will be that only the corpus (lemmatised but otherwise unmodified) constitutes the reference point, and not an annotated corpus or some other ontology. We aim at defining an evaluation method that is extremely easily applicable by laymen.

We have been mainly inspired by the criteria proposed by Guarino [11–p.7] and the classical information extraction measures [29]. In the current ontology engineering field, it is problematic to objectively evaluate ontologies in an automated way as in the overwhelming majority of cases (suitable) gold standards are lacking [12]. Below (section 3), we give our definition of these criteria that allow computation, which are closer to the traditional information extraction definitions of recall and precision.

The remainder of this paper is organised as follows. The next two sections present the material (section 2) and methods (section 3). The evaluation results are described in section 4 and discussed subsequently (section 5). Related work (section 6) is presented. Indications for future research are given in section 7, and some final remarks (section 8) conclude this paper.

[3] This is called ontology aligning and merging

2 Material

2.1 Unsupervised Text Mining

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. As a consequence, these techniques are expected to be more easily portable to new domains. In order to extract this information automatically from our corpus, we used the memory-based shallow parser for English, which is being developed at CNTS Antwerp and ILK Tilburg [3][4]. This shallow parser takes plain text as input, performs tokenisation, part of speech (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. We extract from the shallow parser output semantic relations that match predefined syntactic patterns. Additional statistics using normalised frequencies and probabilities of occurrence are calculated to separate noise (i.e. false combinations generated by chance) from genuine results. More details on the linguistic processing can be found in [20, 21, 22].

2.2 Corpus

The VAT corpus (a single long document) consists of 49,5K words. It constitutes the sixth EU directive on VAT (77/388/EEC of 27 January 2001 - English Version) that has to be adopted and transformed into local legislation by every Member State[5]. We applied the memory based shallow parser to this corpus. After some format transformation, the text miner outputs 315 triples subject-verb-object, such as <*person pay tax*>, and 500 triples noun phrase-preposition-noun phrase such as <*accordance with article*> resulting in a total of 815 triples. In addition, the Wall Street Journal corpus (a collection - 1290K words[6] - of English newspaper articles) serves a "neutral" corpus representing the general language use - see below.

To compute the necessary frequencies and statistics about the corpora, specific Perl scripts have been used. Further manipulation of the numbers is done by means of other small scripts implemented in Tawk v.5 [32] in combination with some standard DOS or Linux commands (mainly "sort").

2.3 DOGMA Ontology Engineering Framework

Before presenting the actual experiments, we shortly discuss the framework for which the results of the experiments are meant to be used, i.e. the *VUB STAR Lab DOGMA* (Developing Ontology-Guided Mediation for Agents) ontology engineering approach[7].

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

[5] This directive serves as input for the ontology modelling and terminology construction activities in the EU FP5 IST FF Poirot project (IST-2001-38248).

[6] The Linux wc -c command has been used to count the words of the VAT and WSJ corpora.

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

The results of the unsupervised mining phase are represented as *lexons*. These are binary fact types indicating the entities as well as the roles assumed in a semantic relationship [24]. Formally, a lexon is described as $<(\gamma, \lambda)$*: term*$_1$ *role co-role term*$_2>$. For the sake of brevity, the context (γ) and language (λ) identifiers will be omitted. 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 [17]. With some simplications, one can state that a lexon can be considered as a combination of two RDF-triples.

2.4 Combining all the Above

As the triples resulting from the unsupervised mining consist of three elements[8] (two terms consisting of one or several words and one role represented by the verb or the preposition[9]) extracted from the VAT corpus, it is possible to investigate to what extent the vocabulary of triples (to be converted afterwards to DOGMA lexons) adequately represents the notions of a particular application domain. Note that this technique in principle could be applied not only to DOGMA lexons but also to RDFS and OWL Lite ontologies.

3 Methods

3.1 Introduction

The starting point in this paper is that triples, representing the basic binary facts expressed in natural language about a domain, can be extracted from the available textual sources using the unsupervised text miner described above. The basic research question is whether or not suitable metrics can be defined to quantitatively evaluate the goodness of fit between the vocabulary of the triples extracted and the intended domain model "embodied" in the textual sources.

We have combined the criteria of Guarino [11] with the more classical information extraction measures [29]. We stress that text mining does not deal with an actual conceptualisation, but rather with its representation or lexicalisation in a text, meaning that we cannot access directly the conceptualisation (meaning level) but have to stay on the linguistic level [26]. However, as many ontology engineers seem to overlook this distinction, the evaluation method proposed here can be applied to existing ontologies as well.

The four measures are:

- *coverage*: are the triples retrieved representing the domain ?
- *accuracy*: are the triples retrieved not too general but reflecting the specialised terms of the domain ?

[8] In fact, the words composing an element have been lemmatised, i.e. reduced to their base form. E.g., working, works, worked → work. In this paper, the terms 'word', 'term', and 'lemma' are used interchangeably.

[9] Co-roles and context are not provided by the CNTS unsupervised miner.

- *recall*: have all the relevant triples been retrieved
- *precision*: are the triples retrieved relevant for the domain ?

In the following sections, we shall elaborate on a computable definition of these criteria and on the ideas that form the basis of the metrics. The exact formulas will be explained as well. The core of the method relies on decomposing the triples into their constituting words and performing calculations on the individual words.

3.2 Coverage

A simplistic metric to determine the coverage would be to calculate the intersection between the vocabulary of the triples and the entire corpus. As many words do not represent domain concepts (e.g. adverbs, determiners, particles, ..., which are by definition not retained by the unsupervised text miner) the triples generated automatically most probably will not attain a high domain coverage rate. In order to differentiate more important words from less important ones, the frequency of a word can be taken into account. Naively, one would expect that important domain words are mentioned more often than others. Therefore, the words are grouped into frequency classes, i.e. the absolute number of times a word appears in a corpus. E.g., in the VAT corpus, the word 'the' appears 3573 times while it is the only element in the frequency class *3573*. Conversely, 'by-product' and 'chargeability' each occur only once, but there are 1521 different words in the frequency class *1*. For each frequency class the ratio of the vocabulary intersection and the frequency class is calculated, and subsequently averaged over the number of classes.

coverage(triples,text) =

$$\frac{\sum_{i=1}^{n} \frac{\#(words\_triples\_freq\_class_i \bigcap words\_text\_freq\_class_i)}{\#words\_text\_freq\_class_i} * 100}{n}$$

The coverage of a text by the vocabulary of triples automatically mined will be measured by counting for each frequency class the number of words, constituting the triples, that are identical with words from that frequency class and comparing this number to the overall word count for the same class . The mean value of these proportions constitutes the overall coverage percentage.

3.3 Precision and Recall

It is difficult to compute the precision, i.e. determining if the triples retrieved are correct, whereby correct is to be interpreted as making sense for the application domain. These decisions require the involvement of human evaluators, and/or an established gold standard. An earlier experiment on evaluating the precision of unsupervised text mining for ontologies is reported in [20] using UMLS [13] as gold standard.

In the approach proposed here, we use a metric from quantitative linguistics [6] to automatically build a gold standard. The standard consists of a set of words that characterise an application domain text resulting from a quantitative comparison with another text. Regarding technical texts, one can easily assume that the specialised vocabulary

constitutes the bulk of the characteristic vocabulary, especially if the other corpus with which to compare is the Wall Street Journal (= collection of general newspaper articles), as is the case here.

The following statistical formulas (used to calculate the difference between two proportions) determine which words are typical of one text compared to another:

$$\tilde{f} = (\frac{f_{word\_text}}{N}) * 100$$

with f being the absolute frequency of a word in a text and N being the total number of words of that text.

$$z = \frac{\tilde{f}_1 - \tilde{f}_2}{\sqrt{(\frac{\tilde{f}_1 * (100 - \tilde{f}_1)}{N_1}) + (\frac{\tilde{f}_2 * (100 - \tilde{f}_2)}{N_2})}}$$

with z expressing a significance value for the deviation between the relative frequencies $\tilde{f}_1$ and $\tilde{f}_2$. Depending on one's preference for the threshold, values of z (expressed in units of σ) below 1,96 ($p < 5\%$) or 2,57 ($p < 1\%$) are statistically not significant.

recall(triples,text) =

$$(\frac{\#(words\_of\_triples\_mined \bigcap statistically\_relevant\_words)}{\#statistically\_relevant\_words}) * 100)$$

The ratio of the vocabulary common to the retrieved triples and statistically significant (threshold = 1,96) characteristic words and these characteristic words determines the recall value.

precision(triples,text) =

$$(\frac{\#(words\_of\_triples\_mined \bigcap statistically\_relevant\_words)}{\#words\_of\_triples\_mined}) * 100)$$

The ratio of the vocabulary common to the triples mined and statistically significant (threshold = 1,96) characteristic words and the vocabulary of the triples mined determines the precision value.

As is done for the coverage, one could also compute the average over the frequency classes of their recall and precision values.

3.4 Accuracy

The purpose of calculating the accuracy is to refine the coverage measure that is based only on word frequency, by combining it with the precision measure. The source of inspiration is Zipf's law [31]. It states that the product of the frequency and the rank order is approximately constant [29–p.2]. Or said in a simpler way, in each text there is a small set of words that occur very often and a large set of words that rarely occur. Zipf has discovered experimentally that the more frequently a word is used, the less meaning it carries. Hence his observation that the higher frequency classes (i.e. containing the

few words that appear very often) contain mostly ”empty” words (also called function or stop words).

A corollary from Zipf’s law is that domain or topic specific vocabulary is to be looked for in the middle to lower frequency classes. Consequently, triples mined from a corpus should preferably contain terms from these ”relevant” frequency classes. Luhn [15] has defined intuitively a frequency class upper and lower bound between which the most significant words are found in the middle of the area of the frequency classes between these boundaries. He called this the ”resolving power of significant words”.

The metrics from quantitative corpus linguistics mentioned above are re-used to objectively determine the range of relevant frequency classes. The frequency classes that contain a high number of typical words will be considered as ”relevant” frequency classes. Currently, we assume that a frequency class should contain at least 60% of characteristic words in order to be a relevant class. Additionally, one could apply the statistical significance threshold (not done for these experiments). Notions represented by words of the relevant frequency classes should be maximally included in an ontology for that particular application domain.

accuracy(triples,text) =

$$\frac{\sum_{i=1}^{n} \frac{\#(words\_triples\_rel\_freq\_class_i \bigcap words\_text\_rel\_freq\_class_i)}{\#words\_text\_rel\_freq\_class_i} * 100}{n}$$

The accuracy of automatically mined triples to lexically represent the important notions of a text will be measured by averaging the coverage percentage for the relevant frequency classes. A frequency class is considered to be relevant if it contains more than 60% of typical vocabulary.

3.5 Experiments

In the experiment, various scripts have been used to calculate the absolute and relative frequencies as well as the coverage, recall, precision and accuracy measures mentioned above. The input texts have not been filtered or modified except for the lemmatisation.

88,44% of the lemmas (=types) falls into the first 110 FCs, which represents 10,98% of the total word occurrences (=tokens). There are 66 FCs more above 110. The ten highest are 752, 790, 1011, 1110, 1157, 1260, 1378, 2011, 2401 and 3573, all consisting of one word (see 1).

We have also determined a baseline against which the results of our method will be compared. The core of the baseline algorithm is a random number generator (built-in TAWK function [32]) that is used to pick out a word from the lemmatised corpus vocabulary (3210 unique base forms). An equal amount of lemmas is randomly selected as there are lemmas in the triples list.

4 Results

It should be clear from the on-set that high scores will not be attained. Only terms in a verb-object and a subject-object grammatical relation are selected by the shallow

parser, combined by clustering and submitted subsequently to several selection thresholds , which already constitutes a substantial reduction of the number of lemmas that constitute the triples.

4.1 Coverage

A coverage rate of 39.68% is obtained (the naive coverage rate being 8.62%). Figures 1 and 3 show that, especially for the first six frequency classes (FC) (i.e. lemmas appearing once up to six times) the coverage rate is below 10%. In figure 1, for reasons of graphical visibility, a ceiling for the number of lemmas (Y-axis) has been established on 170. FC 1 contains 1521, FC 2 442 and FC 3 223 lemmas respectively. The high dispersion of the coverage for FCs starting from class 40 is to be explained by the low number of lemmas in these classes (rarely higher than 5). From class 82 onwards, a FC consists of a single lemma (FC 93, 108, 120, 128, 131 and 169 being the exceptions containing two and 100 three lemmas).

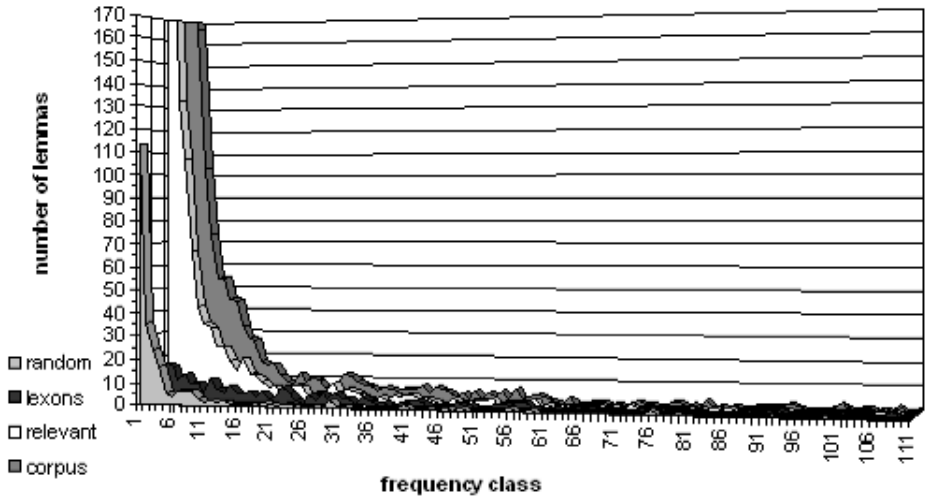

Fig. 1. absolute coverage of frequency classes

The unsupervised miner seemingly misses a lot of low frequency terms that are considered as typical of the VAT corpus. Even a naive random selection mechanism scores ”better” for the FC 1 till (and including) 4.

4.2 Recall and Precision

The precision ratio is of 58.78% while the recall is 9.84%. In absolute numbers, it means that 211 lemmas have been selected as representing domain knowledge by both the unsupervised miner and the statistical comparison formula. Figure 2 shows the distribution of the recall ratios per frequency class. The averaged recall is 48,74% and the averaged precision is 58,27%.

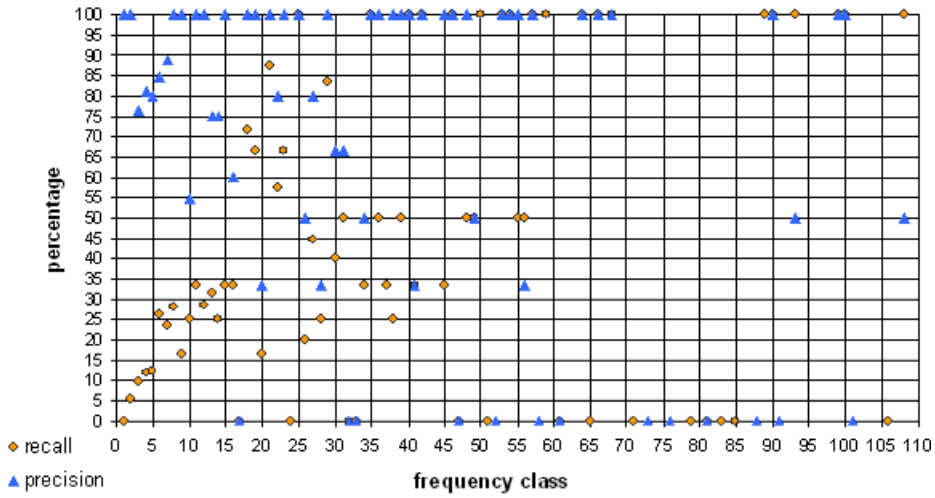

Fig. 2. recall and precision per frequency class

4.3 Accuracy

There are 34 typical frequency classes (i.e. containing at least 60% of words that are judged to be statistically typical). The classes are 1 - 5, 7, 13 -15, 18, 21, 22, 24 - 27, 29, 30, 34, 36 - 38, 45, 48, 51, 55, 57, 64 - 66, 68, 71, 79, 83, 85, 90, 99, 106, 111, 119, 121, 145, 169, 173, 181, 199, 219, 276, 277, 385, 597, 727, 1011, and 1378. It has to be noted that, from class 60 onwards the FCs only contain a single word and it is judged as typical (except for class 72: two words and both typical). The average coverage ratio (= the accuracy) for these 34 typical frequency classes is 52,1%.

5 Discussion

5.1 The Material

The two corpora exhibit the expected behaviour as expressed by Zipf's law. The corresponding FCs of the two corpora contribute more or less to the same extent to the overall vocabulary. Therefore, it is rather disappointing that the text miner only attains low coverage and recall scores on the one hand, and it is surprising on the other that the lowest FCs are to be considered as relevant - see figure 3.

Table 1 illustrates the effect of applying the accuracy calculation. The first two data rows show Zipf's law in practice (the top frequency classes contain empty words), while the other two data rows display the ten topmost typical frequency classes. Calculating the accuracy measure apparently does not adequately filter out the empty words or non relevant words.

A closer examination of the entire corpus revealed an important part of non-words (numbers of all kinds, section indications, percentages, typos, ...) in both the VAT (655 items or 20,24% of the lemmas) and WSJ corpora (6236 items or 14,71%). This is

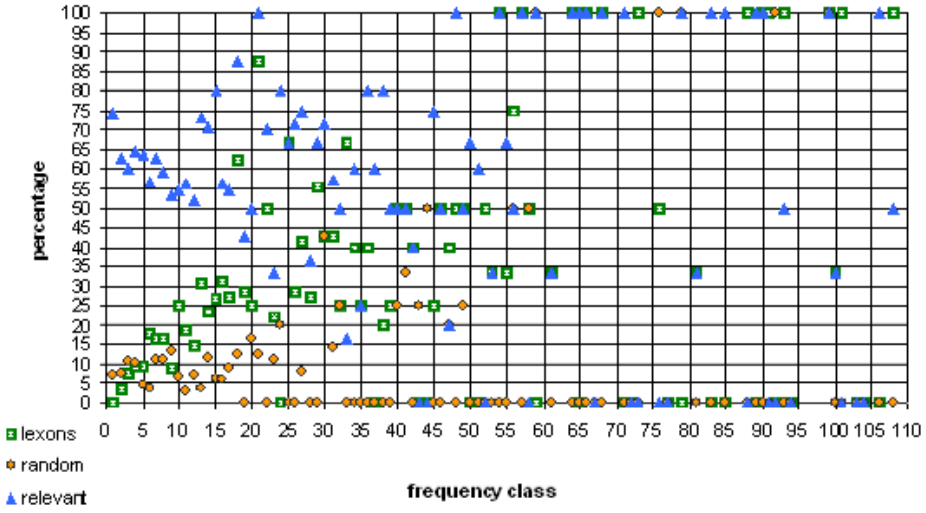

Fig. 3. relative coverage of frequency classes: by mining triples or lexons, by randomly picking words versus selecting statistically typical words

Table 1. Ten topmost frequency classes and their members (before and after accuracy calculation)

”raw” data										
FC	3573	2401	2011	1378	1260	1157	1110	1011	790	752
word	the	of	,	)	to	in	be	(	and	.
”accurate” data										
FC	1378	1011	727	597	385	277	276	219	199	181
word	)	(	or	shall	-	;	:	/	add	refer

particularly annoying for the VAT corpus as almost all (551) of these non-words are considered to be characteristic (on a total of 1965 characteristic words). As, naturally, the text miner does not retain these non-words, the coverage and recall scores are biased in a negative way. These non-words also bias the accuracy score as they influence the status of a FC (being typical or not). This explains to a large extent why even very low FCs are considered as relevant FCs, which contradicts Zipf’s and Luhn’s findings. Therefore, we plan to redo the experiment, but with an adequate definition of what a ”good” formally word consists of. A professional concordancy program (e.g., WordSmith) will be used to this aim in a next iteration of the experiments.

Luckily, the precision score is not affected by this problem - see figure 2. A score of a bit less than 60% is not spectacularly good, but neither particularly bad. If we look at it from a positive angle, it means that a knowledge engineer disposes, with a sufficient degree of trust, of two thirds of the important domain words. It would be interesting to investigate which kind of notions the words represent. We believe that these words are representative for the ”middle out” ontology engineering approach,

and therefore it is most likely that human domain experts are able to rapidly fill in the more general domain notions that are missing. More research, involving application domains of various nature, is needed to investigate how to mine the very specialised, and therefore, less often used notions. However, it is our intuition that the reduction of the cognitive load for a knowledge engineer (studying some 815 lexons instead of an entire text) is already substantial.

5.2 The Unsupervised Text Miner

The text miner clearly behaves in a non-random way: the distribution of the lemmas randomly picked follows the overall corpus distribution - see figure 1. Because of the high number of non-words in this experiment, it is almost certain that randomly picking words will produce a lot of garbage.

What is evident from this evaluation experiment is that the CNTS unsupervised text miner currently misses too many important notions, but that the results produced are of an acceptable quality. It is unclear to the authors how they would have reached this objective conclusion in a fast way without the support of the evaluation method reported on.

Some mistakes made by the shallow parser have a strong influence on the quality of the semantic extraction process. This happens if words unknown to the parser are unproperly tagged, or if syntactic relations are missed or wrongly identified. The structure of the corpus also plays a role in that respect. The VAT corpus contains a lot of enumerations that are difficult to analyse for the parser due to the distance between the main verb and some complements. The fact that the shallow parser has not been trained on legal material plays a role as well. It is the nature of unsupervised mining that no tuning to a specific corpus is done. Therefore, the overall results are worse than with supervised mining. There is a trade-off to be made between resource investment and quality of the results.

Concerning the extraction of triples, the size of the corpus matters a lot as one common technique used to judge the appropriateness of a term relies on its frequency in the corpus. The extraction process of the text miner discards some relevant terms because they appear only once in the VAT corpus. A new experiment (without the non-words) including human validation should determine if the statistical thresholds of the unsupervised miner are to be relaxed.

5.3 The Evaluation Method

It is quite evident that the coverage measure is a too ”naive” measure to be useful, except as an intermediary step to calculate the accuracy. Table 1 shows why. Recall and precision are considered traditionally as complementary (and are often combined in the F-measure). Accuracy could be an alternative to recall as it tries to somehow combine Luhn’s theoretical findings on the resolving power of significant words with using a gold standard. More practical work should be done in order to validate this assumption.

Note that the evaluation method proposed does not give any indication on the correctness of a triples as a whole. It means that, if the words ”fish” and ”exception”

are typical of the application domain, the method cannot rate the triple <fish with exception> as invalid. We did not yet examine these aspects.

As already mentioned, the method stays on the word level. It is to be expected that grouping synonyms might improve the score, but it is unclear to what extent. Eventually some way of abstraction (especially for the RDF predicate or lexon role) will have to be done. Also these aspects require further investigations.

The important point of applying these metrics, how imperfect they currently might be, is that the scores can be used to monitor changes (preferably improvements) in the behaviour of the text miner (regression tests). As soon as the scores for a particular (and commonly agreed) textual source have been scientifically validated, the source and the scores together can become an evaluation standard in bench-marking tests involving other text miners, or even to some extent any RDF-based ontology producing tool. A logical next step would be that ontologies, automatically created by a text miner, are documented with performance scores on their textual source material as well as with scores for that particular text miner on the evaluation standard (commonly agreed text and outcomes).

6 Related Work

Previous reports on our work contain additional details on the unsupervised miner [22], its application to a bio-medical corpus [21], and a qualitative evaluation [25]. To the best of our knowledge, so far only one other approach has been presented that addresses the quantitative and automated evaluation of an ontology by referring to its source corpus.

Brewster and colleagues have recently presented a probabilistic measure to evaluate the best fit between a corpus and a set of ontologies as a maximised conditional probability of finding the corpus given an ontology. The specific probabilistic formula to compute the conditional probability of a concept label given a term occurrence takes synonyms into account mediating a form of query expansion [2–p.166]. It seems that some training needs to be done on basis of the annotated corpus, which is something we explicitly want to avoid with our approach. Unfortunately, no concrete results or test case are presented.

In addition, *Velardi* and colleagues have proposed to use the combination of ”domain relevance” and ”domain consensus” metrics to prune non domain terms from a set of candidate terms [30]. They use a set of texts typical of the domain next to other ones. *Domain relevance* is in fact the proportion of the relative frequency of a term in the domain text compared to the maximum relative frequency of that term over several non domain texts. *Domain consensus* is defined as the entropy of the distribution of a term in all the texts of the corpus. In our approach, we have computed the difference between two proportions, more specifically the z-values of the relative difference between the frequency of a word in a technical text vs. a general text (WSJ), which enables us to filter out words that are only seemingly typical of the technical text. In [18], the authors also present a method to semantically interprete novel complex terms with the help of WordNet and to organise them in a hierarchy. An evaluation of these latter aspects is also provided. Remark that both of

the proposed methods clearly (and correctly) differentiate a term or word from a concept.

Another statistical approach is elaborated by *Gillam and Tariq* [7] as part of a method to extract technical complex terms. They as well try to compare a specific text with a general text and characterise words by their weirdness (z-score for the ratio of the two relative frequencies of a word). More research is to be done to determine the exact difference with our approach.

Finally, although the main focus of the reports does not cover exactly the work presented here, the methods for ontology evaluation presented in [5, 12] can provide complementary information and inspiration. In particular, we could extend the notion of "relevant" as used above to an entire triple and define additional metrics, as has been done by Sabou [23] for extracted significant pairs. In the same vein, one could consider additionally the work of Maedche and Staab [16] who include the Levenshtein edit distance in their approach to measure the similarity between two ontologies. However, it is our conviction that the Levenshtein measure is too crude and naive to be of any use for our purposes.

In short, our approach is a first step to evaluate quantitatively and objectively triples generated by an unsupervised text miner. It does not aim directly at selecting relevant compounds terms and providing their semantic compositional interpretation. Although it would be interesting to see whether for our VAT corpus sensible interpretations could be generated using the structural semantic interconnections algorithm of Velardi and colleagues [18]. Also, their domain relevance measure, when applied to two texts, is equivalent to the corpus linguistic statistical formulas presented in section 3.3. could be an alternative metric to be taken into account for our evaluation purposes.

7 Future Work

There still remain some major points for improvement. An important issue is to extend the evaluation techniques presented above to multiple documents (i.e. by using the inverse document frequency (TF/IDF) metric, chi-square or the domain relevance and consensus metrics instead of simple word frequency for a single document). Although the unsupervised text miner detects compounds, the evaluation component currently takes only simple words into account. A compound detection module should thus be added.

Concerning the text miner itself, alternative statistical measures could be considered or thresholds could be relaxed to capture more low frequencies words. Additional syntactical patterns (e.g., subject - verb - prepositional object) should be retained. Ideally, the choice for a specific pattern should be done automatically in function of the structure and content of the corpus.

A next step would be to compare manually created ontologies with their source texts, which necessitates the integration of semantic distance measures such as the WordNet similarity functions [19] to operate on the meaning level instead of the linguistic level. Brewster et al. add two levels of WordNet hypernyms [2–p.166] for that purpose. That implies that (novel) compound terms should be assigned a semantic interpretation as is done by Navigli and Velardi [18].

8 Conclusion

We have presented the results of a simple quantitative evaluation method for the outcomes of an unsupervised mining algorithm applied to a financial corpus. Coverage, accuracy, recall and precision measures have been defined and calculated accordingly resulting in a 38.68%, 52.1%, 9.84% and 75.81% score respectively. These results (although biased because of the presence of many non-words in the corpus) have permitted us to identify a weak spot in the performance of the text miner, which will be improved in the future. New experiments to further validate the method are scheduled. An outline of a future research agenda has been given.

Acknowledgments. This research has been carried out during the OntoBasis project (IWT GBOU 2001 #10069), sponsored by the IWT Vlaanderen (Institution for the Promotion of Innovation by Science and Technology in Flanders). In additon, some parts have served as a contribution to the joint research activity program [12] of the EU FP6 IST NeO KnowledgeWeb (IST-2004-507482).

References

1. T. Berners-Lee, *Weaving the Web*, Harper, 1999.
2. Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data Driven Ontology Evaluation. In, N. Shadbolt and K. O'Hara (eds.), *Advanced Knowledge Technologies: selected papers 2004*, pp. 164 – 164, 2004 (reprint from LREC2004).
3. Sabine Buchholz, Jorn Veenstra, and Walter Daelemans, Cascaded grammatical relation assignment, in *Proceedings of EMNLP/VLC-99*. PrintPartners Ipskamp, 1999.
4. Paul Buitelaar, Siegfried Handschuh, and Bernardo Magnini (eds.). *Proc. of the ECAI04 Workshop on Ontology Learning and Population*, 2004.
5. Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini (eds.). *Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press, Amsterdam, 2005 (forthcoming).
6. Josse De Kock. *Elementos para una estilística computacional - tomo I*. Editorial Coloquio, Madrid, 1984.
7. Lee Gillam and Mariam Tariq. Ontology via Terminology? In F. Ibekwe-San Juan and S. LainCruzel (eds.), *Proceedings of the Workshop on Terminology, Ontology and Knowledge Representation*, 2004. http://www.univ-lyon3.fr/partagedessavoirs/termino2004/programgb.htm
8. Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. Springer Verlag, 2003.
9. T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **6** (2):199–221, 1993.
10. N. Guarino and P. Giaretta, 'Ontologies and knowledge bases: Towards a terminological clarification', in *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, ed., N. Mars, pp. 25 – 32, IOS Press, Amsterdam, 1995.
11. Nicola Guarino. Towards a Formal Evaluation of ontological Quality. *IEEE Intelligent System*, **19** (4):78–80, 2004.
12. Jens Hartmann, Peter Spyns, Diane Maynard, Roberta Cuel, Mari Carmen Suarez de Figueroa and York Sure. Methods for Ontology Evaluation, KnowledgeWeb Deliverable #D1.2.3, 2005.

13. B. Humphreys and D. Lindberg. The unified medical language system project: : a distributed experiment in improving access to biomedical information. In K.C. Lun, (ed.), *Proc. of the 7th World Congress on Medical Informatics (MEDINFO92)*, pp. 1496–1500, 1992.
14. H. Karanikas and B. Theodoulidis, 'Knowledge discovery in text and text mining software', Technical report, UMIST - CRIM, Manchester, 2002.
15. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2** (2):159 – 195, 1958.
16. Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In, *Proceedings Of the European Conference on Knowledge Acquisition and Management (EKAW02)*, pp. 251 – 263, LNAI 2473, Springer Verlag, 2002
17. Robert Meersman. Ontologies and databases: More than a fleeting resemblance. In A. d'Atri and M. Missikoff (eds.), *OES/SEO 2001 Rome Workshop*. Luiss Publications, 2001.
18. Roberto Navigli and Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, **30** (2):151–179, 2004.
19. T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *The Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, 2004.
20. 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.), *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA and ODBASE*, LNCS 2888, pp. 803 – 819, 2003. Springer.
21. Marie-Laure Reinberger, Peter Spyns, A. Johannes Pretorius, and Walter Daelemans. Automatic initiation of an ontology. In Robert Meersman, Zahir Tari et al. (eds.), *On the Move to Meaningful Internet Systems 2004: CooPIS, DOA and ODBASE (part I)*, LNCS 3290 , pp. 600 – 617 , 2004. Springer Verlag.
22. Marie-Laure Reinberger and Peter Spyns. Unsupervised Text Mining for the Learning of DOGMA-inspired Ontologies. In P. Buitelaar, Ph. Cimiano, and B. Magnini, (eds.), *Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press Amsterdam, 2005.
23. Marta Sabou. Extracting Ontologies from Software Documentation: a Semi-automatic Method and its Evaluation. In P. Buitelaar, Ph. Cimiano, and B. Magnini, (eds.), *Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press Amsterdam, 2005.
24. Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *SIGMOD Record Special Issue*, **31** (4):12–17, 2002.
25. Peter Spyns, A. Johannes Pretorius and Marie-Laure Reinberger. Evaluating DOGMA-lexons generated automatically from a text corpus. In Cimiano P., Ciravegna F., Motta E. and Uren V. (eds.), *Proceedings of the EKAW2004 Workshop on Human Language Technology and Knowledge Management*, pp. 38 – 44, 2004.
26. Peter Spyns and Jan De Bo. Ontologies: a revamped cross-disciplinary buzzword or a truly promising interdisciplinary research topic? *Linguistica Antverpiensia, new series* (3), 2004 (forthcoming).
27. M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge Sharing and Review*, **11** (2), June 1996.
28. M. Ushold, 'Where are the semantics in the semantic web?', *AI Magazine*, **24** (3):25 – 36, 2003.
29. C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
30. Paola Velardi, Michele Missikoff, and Roberto Basili. Identification of relevant terms to support the construction of Domain Ontologies. In Maybury M., Bernsen N., and Krauwer S. (eds.)*Proc. of the ACL-EACL Workshop on Human Language Technologies*, 2001.
31. George K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA, 1949.
32. Tawk Compiler v.5. Thompson Automation Software, Jefferson OR, US.