Evaluating Automatically a Text Miner for Ontologies: A Catch-22 Situation?

Peter Spyns

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

Abstract. Evaluation of ontologies is increasingly becoming important as the number of available ontologies is steadily growing. Ontology evaluation is a labour intensive and laborious job. Hence, the importance to come up with automated methods. Before automated methods achieve reliability and widespread adoption, these methods themselves have to be assessed first by human experts. We summarise experiences acquired when trying to assess an automated ontology evaluation method. Previously we have implemented and evaluated a light-weight automatic ontology evaluation method that can be easily applied by knowledge engineers to rapidly determine whether or not the most important notions and relationships are represented in a set of ontology triplets. Domain experts have contributed to the assessment effort. Various assessment experiments have been carried out. In this paper, we focus particularly on the practical lessons learnt, in particular the limitations that result from real life constraints, rather than on the precise method to automatically evaluate results of an ontology miner. A typology of potential evaluation biases is applied to demonstrate the substantial impact conditions in which an evaluation happens can have on the reliability of the outcomes of an evaluation exercise. As a result, the notion of "meta-evaluation of ontologies" is introduced and its importance illustrated. The main conclusion is that still more domain experts have to be involved, which is exactly what we try to avoid by applying an automated evaluation procedure. A catch-22 situation?

1 Introduction and Background

The development of the Semantic Web (of which ontologies constitute a basic building block) has become a very important research topic for the information based society. However, the process of conceptualising an application domain and its formalisation require substantial human resources and efforts. Therefore, techniques applied in human language technology (HLT) and information extraction (IE) are used to create or grow ontologies with a quality as high as possible in a period of time as limited as possible. Work is still in progress - recent

R. Meersman and Z. Tari (Eds.): OTM 2008, Part II, LNCS 5332, pp. 1404–1422, 2008.

overviews of the state of the art (in particular for machine learning techniques) can be found in [5,6,26].

Even in the ideal case that (semi-) automated ontology learning methods have become mature, there still remains the problem of assessing and evaluating the results. Various proposals for evaluation methods[1] have recently been put forward [2,6,14,34]. All these approaches basically share the same problem, i.e. how to evaluate the outcomes of automated ontology learning methods in a way that goes beyond the context of a specific evaluation setting (task, domain, ...). Rare are the experts willing to devote their precious time to validate output generated by a machine or establish in agreement with colleague stakeholders and experts a gold standard. In addition, current evaluation methods require specialised skills and infrastructure almost solely available in an academic environment.

In an answer to these issues, we have tried to define a light-weight assessment procedure that is easy to understand and apply by "standard knowledge workers" (basically a domain expert, a computer scientist, an engineer, ...) outside academia [28]. The evaluation method should be generally applicable (any kind of text miner, any kind of text collection) and able to provide a rough but good enough and reliable indication whether or not results of a text miner on a particular corpus are worthwhile. Ontologies can be evaluated from many angles [7,12]. Our method wants to measure to which extent an ontology includes the important domain notions. Hence, in this paper quality of an ontology refers to the degree with which the lexical material delivered by the ontology miner covers the important notions conveyed in a text corpus. Furthermore, this is only one dimension of judging the quality of an ontology. Other dimensions are equally important and should also be taken into account - see the related work in section 6. Typical of our approach will be that only the "raw" corpus (lemmatised [2] but otherwise unmodified) constitutes the reference point, and not an annotated corpus or another reference ontology. However, the automatic evaluation procedure itself still needs validation, and therefore we do need human experts and/or a gold standard built by human experts.

As this is an ambitious endeavour, we have to realise it in several stages. The first step has been to define and try out some lexicometric scores for triplets generated automatically by a text miner [28,29,33]. A next step is to validate the evaluation procedure using these scores [31,32]. Trying out the method in various situations and synthesising the outcomes is a subsequent logical step. Finally, the experiences from the validation experiments have to be summarised as provide valuable insights to determine the set-up of new experiments.

The remainder of this paper is organised as follows. The next two sections present the material (section 2) and methods (section 3). An overview of the various experiments and their setting is presented in section 3.1. Subsequently, we

[1] The EON2006 workshop has been devoted to ontology evaluation - see http://km.aifb.uni-karlsruhe.de/ws/eon2006

[2] Lemmatise means to reduce words to their base form. E.g., working, works, worked → work. Incidentally note that in this paper, the terms 'word', 'term', and 'lemma' are used interchangeably.

explain how the machine gold standard is established (section 3.2) and validated (section 3.3). In section 4 (Results), we discuss how the automated evaluation procedure rates the results of an ontology miner on the one hand (section 4.1) as well as how the domain experts rate the automated procedure (section 4.2) on the other. In addition, not only the results of the ontology miner (section 5.1) and the evaluation experiments (section 5.2) are discussed but also their organisation (section 5.3). Related work is outlined in section 6. Indications for future research are given in section 7, and some final remarks (section 8) conclude this paper.

2 Material

The *memory-based shallow parser for English*, being developed at CNTS Antwerp and ILK Tilburg [4] [3], has been used. It is an unsupervised parser that has been trained on a large general purpose language model. No additional training sessions (= supervised) on specific corpora are needed. Hence, the distinction between learning and test corpus has become irrelevant for our purposes. Semantic relations that match predefined syntactic patterns have been extracted from the shallow parser output. Additional statistics and clustering techniques using normalised frequencies and probabilities of occurrence are calculated to separate noise (i.e. false combinations generated) from genuine results. The unsupervised memory-based shallow parser with the additional statistical modules constitute the ontology miner. More details can be found in [22,23].

The privacy and VAT corpora (two separate documents) consist of 72,1K resp. 49,5K words. They constitute two *directives* (English version), namely the 95/46/EC of 18/12/2000 (privacy) and the 77/388/EC of 27/01/2001 (VAT), which EU member states have to adopt and transform into local legislation. The VAT directive has served as input for the ontology modelling and terminology construction activities in the EU FP5 IST FF Poirot project[4] (IST-2001-38248). These two documents are the sole official legal reference texts for the domain. The texts have been lemmatised. The size of both texts however is rather small, when compared to other machine learning experiments. As a consequence, the quality of the ontology miner might be compromised. A possible workaround is to include unofficial documents that provide comments or points of view on the official directives. However, this might distort the outcomes as well as these don't represent an official EU position.

We were lucky to be able to use a list of *900 VAT terms selected manually* by domain experts on basis of the EU VAT Directive. According to the VAT experts the notions represented by these terms should be included in a VAT ontology.

The *Wall Street Journal (WSJ) corpus* (a collection - 1290K words - of English newspaper articles) serves as a corpus representing the general language that is to be contrasted with the specific technical vocabulary of the two Directives. The WSJ is not really a neutral corpus (the articles are about economic topics).

[3] See http://ilk.kub.nl for a demo version.
[4] http://www.ffpoirot.org/

It is easily available, also included in the WordSmith tool - see below - and a standard in corpus linguistics.

An off-the-shelf available *lexicographic program* (Oxford WordSmith Tools v4[5]) has been used to create the frequency lists and to easily filter out non words[6]. Further manipulation of exported WordSmith files and the calculation of the statistics are done by means of small scripts implemented in *Tawk* v.5 [35], a commercial version of (G)awk, in combination with some manipulations of the data in *MS Excel*.

3 Methods

3.1 Overview

As an ontology is supposed to represent the most relevant concepts and relationships of a domain of discourse or application domain, all the terms lexicalising these concepts and relationships should be retrieved from a corpus of texts about the application domain concerned when building an ontology for the domain. The key question is thus how to determine in an automated way to which extent the important terms of a text corpus have been retrieved. In addition, an algorithm is needed to distinguish relevant combinations (i.e. two concepts in a valid relationship - a triplet) from irrelevant ones. These key issues evidently hold for any ontology miner as well as for any method producing a machine gold standard [8](section 3.2).

The machine gold standard has to be validated by humans (section 3.3) in order to assign some authority to the automated method. This step is in essence similar to a manual evaluation of an individual ontology, but the purpose is different. In addition, we are well aware that current term extractors are more sophisticated than the methods we use. For our experiments, we have intentionally sacrificed scientific state of the art for the simplicity of an off the shelve product.

We also discuss in detail how the validation by human experts has been organised. The work of Friedman and Hripcsak [8] has been our main source of inspiration for a meta-evaluation.

Note that the human-based evaluation step only serves to validate the automated procedure. Figure 1 displays the process flow of the evaluation experiments. Part A of the figure shows the CNTS ontology miner that produces lexical triplets that are validated by human experts (part C), as happens usually. Parts B and D represent "the automated expert" (a term extractor combined with

[5] URL:http://www.lexically.net/wordsmith/

[6] Another interesting tool is the on-line available term identification tool described in [19]. Other heuristics, next to the classical term frequency and document frequency statistics, are taken into account, such as domain relevance, domain consensus, lexical cohesion, and stylistic relevance [7].

[8] In the ideal case an ontology miner's result completely coincides with the machine gold standard.

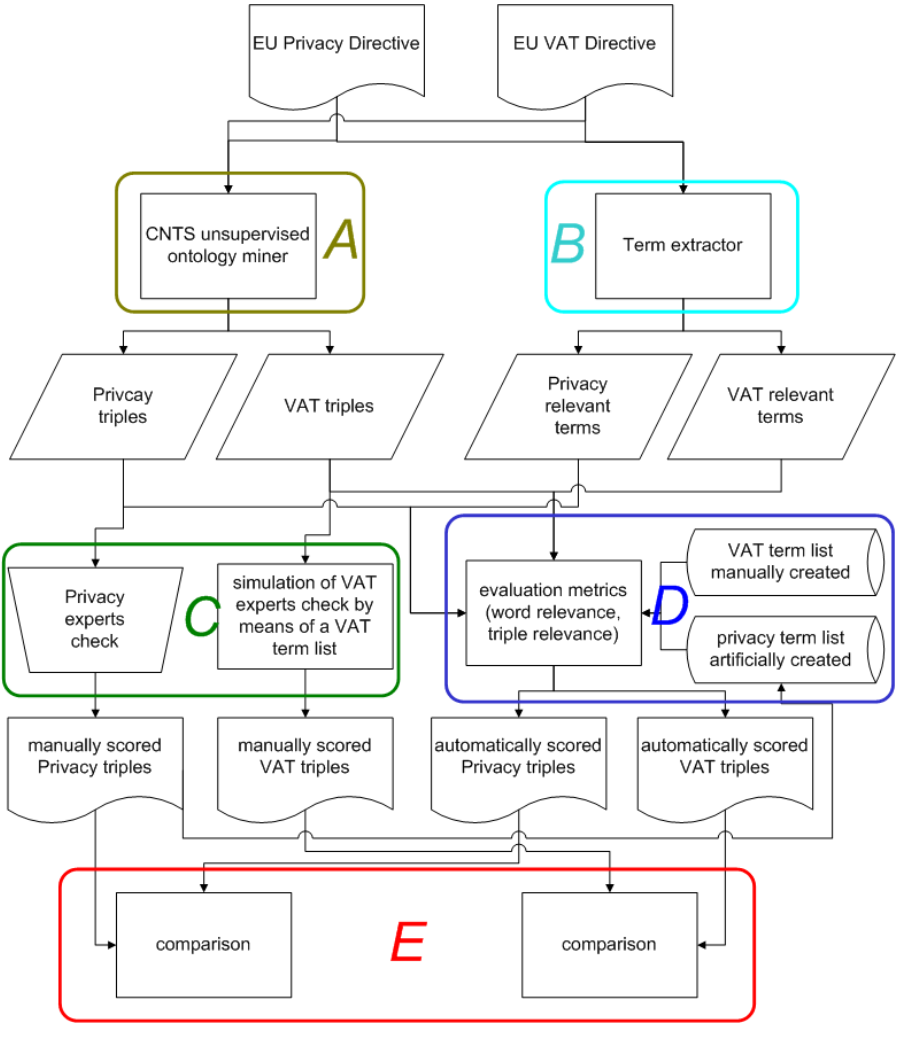

Fig. 1. Overall process flow of the experiments

term and triplet scoring heuristics) as a cheap and fast alternative for the human experts. Part E of the figure stands for the comparison of triplets validated by the human experts (human gold standard) and the ones automatically validated (machine gold standard). The greater the overlap between these two sets, the more closely the automated expert resembles the human experts. In the ideal case, the automated expert can consistently (no inter and intra rater differences) create a gold standard for any text and give a rough but fast impression of the quality of the material mined.

3.2 Establishing the Machine Golden Standard

Finding the relevant words. Basically, we try to answer the following fundamental questions by calculating an associated lexicometric score. Guarino [11] has proposed similar metrics but without a concrete implementation.

- is the vocabulary of the triplets retrieved representing the domain ? *coverage*
- is the vocabulary of the triplets retrieved not too general but reflecting the specialised terms of the domain ? *accuracy*
- has all the relevant domain vocabulary been captured by the triplets retrieved ? *recall*
- is the vocabulary of the triplets retrieved relevant for the domain ? *precision*

We have combined various insights from quantitative linguistics, in particular foundational insights by Zipf and Luhn, a statistical formula to compare two proportions, with the traditional IE evaluation metrics (recall and precision). The central notion linking everything together is "frequency class" (FC), i.e. the set of (different) lemmatised words that appear n times in a document d. E.g., for the Privacy Directive, there are 416 words that appear only once (hence FC 1 contains 416 elements), and there is one word that appears 1163 times (FC 1163 is a singleton). According to Zipf's law [39], the latter one ('the') is void of meaning, while the former ones (e.g., 'assurance') are very meaningful, but may be of only marginal interest to the domain. Subsequently Luhn [16] introduced the notion of "resolving power of significant words" by defining intuitively a frequency class upper and lower bound. in his view, the most significant words are found in the middle of the area of the frequency classes between these boundaries.

We propose to approximate the resolving power of significant words by simply calculating whether a FC is relevant or not. Only if a FC is composed by 60% or more of relevant words, the FC is considered to be relevant. A word is said to be relevant or not based on the outcome of a statistical formula that compares two relative proportions. Technically speaking, we compute the z-values of the relative difference between the frequency of a word in a technical text (the Privacy resp. VAT Directives) vs. a more general text (WSJ), which enables us to determine the words that are statistically typical of the technical text. These are the relevant words. Calculations have been done with a 99% confidence level. The gold standard for words is now defined in a very easy, fast and cheap way.

The assumption is that the ontology miner should be able to retain the words that belong to the relevant frequency classes, and hence simulate "the resolving power of words". The notion of relevant words is distributed over all members of a FC if 60% and more of its population is statistically relevant (see above). Subsequently, we have defined the following lexicometrics:

- The *coverage* of a text by the vocabulary of triplets automatically mined is measured by counting for each frequency class (FC) the number of words, constituting the triplets, that are identical with words from the text for that FC. This number is compared to the overall word count for the same FC . The mean value of these proportions constitutes the overall coverage percentage.

- The *accuracy* of triplets automatically mined to lexically represent the important notions of a text is measured by averaging the coverage percentage for the relevant frequency classes. An FC is considered to be relevant if it contains more than 60% of typical vocabulary, i.e. words considered as characteristic of a text on basis of statistical calculations (= machine gold standard). Characteristic words of a domain specific corpus are determined by comparison with a general language corpus (by calculating the relative difference of relative frequencies).
- The *recall* is defined as the vocabulary common to the triplets mined and the machine gold standard compared to the machine gold standard.
- The *precision* is defined as the vocabulary common to the triplets mined and the machine gold standard compared to the vocabulary of the triplets mined.

Finding the relevant triplets. After having determined how well (or bad) the overall triplet vocabulary (= all the words making up the triplets generated by the ontology miner) covers the terms representing important notions of the domain (as established by the machine gold standard for words), entire triplets are examined.

Again, the machine gold standard is used as reference. A triplet is considered relevant if it is composed by at least two terms statistically relevant (i.e. belong to the machine gold standard). We did not use a stopword list, as this list might change with the nature of the corpus, and as a preposition can be potentially relevant since they are included in the triplets automatically generated. The lexicometrics should cope with these issues.

A triplet score indicates how many characters of the three triplet parts (expressed as an averaged percentage) are matched by words of the machine gold standard. E.g., the triplet $< rule, establish, by\_national\_competent\_body >$ receives a score of 89 as only 'competent' is not included in the machine gold standard with a 95% confidence level (89 = ((4/4)*100 + (11/11)*100 + (17/25)* 100)/3) [9].

3.3 Validating the Machine Golden Standard

The ontology miner itself has not been modified during the experiments. In addition, the developers of the memory-based shallow parsers have not been involved in the experiments, and the developer (computational linguist) of the additional statistical measures only became knowledgeable of the test corpora and results when performing the batch runs of the ontology miner. She has not been involved in the evaluations. Nor had the experts performing the evaluation experiment anything to do with the ontology miner. The computer scientist responsible for the automated evaluation procedure had no knowledge of the internals of the ontology miner and was not involved in the actual assessment by the domain experts. He

[9] A slight imprecision occurs due to the underscores that are not always accounted for.

merely distributed and collected the files (input to computational linguist, mining results to domain experts, assessments from experts) and implemented and ran the automated evaluation procedure. This strict separation of roles guarantees that the various persons involved do not influence each other. Also the set up of the experiments is not biased in one way or the other. Here we fully respect the criteria of Friedman and Hripcsak to minimise the bias [8, p.335].

Assessing the relevant terms. We have determined a baseline against which the results of our method can be compared. In earlier work, we showed that our method performs better than this random baseline (see [29,32]).

Unfortunately, the VAT experts were not available to evaluate the VAT terms and triplets automatically generated. The vocabulary of the 900 VAT terms manually selected constitutes a substitute for humans directly assessing the triplet vocabulary automatically generated. It has not been communicated how the VAT experts have reached agreement on the terms, which constitutes a negative aspect [8, p.336].

Even if not ideal from a scientific point of view, this corresponds to real life situations where on the one hand lists of terms generally accepted by a community are put forward as standard reference, and on the other hand, experts check machine generated outcomes. The former situation can be problematic for consistency and completeness, while the latter corresponds to what is called "'leading the witness"' [10] [8, p.336]. From a methodological point of view, one can argue that the list of terms collected by experts does not necessarily adequately reflect the important terms in the text(s) submitted to the ontology miner. On the other hand, in many cases such term lists are compiled by several representative experts on behalf of standardisation committees and are (publicly) available. Thus, even if not ideal, it is as close as one can get to some objective and qualitative reference if experts are otherwise not available.

Note that a similar reference term list for the privacy domain was not available. As the privacy experts, at the time of the experiments, still had to construct a term list, such a list has been constructed artificially for the sake of the experiments. The terms contained in the triplets produced by the ontology miner that have been positively assessed by the experts make up the privacy machine gold standard.The privacy experts did not assess the machine gold standard for the privacy directive. Instead, the human gold standard was constituted by the vocabulary of the privacy triplets judged relevant by the privacy experts. Inevitably, such an approach runs the danger of missing terms. Not only can the ontology miner fail to erroneously produce a triplet for a relevant term, also human experts can (falsely) reject or unfortunately miss to approve a triplet containing such a term.

Assessing the relevant triplets. The basic questions to assess the quality of the automatic triplet scoring procedure are:

[10] Without a golden reference, evaluators show a tendency to agree with the system output - unless there is a glaring error.

- Have all the relevant triplets been positively scored ? *recall*, also called *sensitivity*
- Are the triplets positively scored indeed relevant for the domain ? *precision*
- Are the triplets that have been negatively scored not relevant for the domain? *specificity*

These metrics using the machine gold standard are applied to estimate the precision score of the miner. Note that we do not determine whether the miner has retrieved all the relevant triplets (recall score for the miner) as there will be no gold standard available (this is the point of setting up an automatic evaluation procedure instead of having experts produce a reference). We use the lexicometric scores to indirectly answer this question.

Two experts in privacy protection matters have been asked to independently validate the list of privacy triplets as produced by the ontology miner. One has been a privacy data commissioner and still is a lawyer while the other is a knowledge engineer specialised in privacy and trust. Ontology engineering involves experts of various background and affiliations to come to a commonly agreed upon conceptualisation. Hence, we consider them as appropriate for the experiments. Unfortunately, we didn't receive any information on the VAT experts involved. Friedman qualifies this as a source of potential bias [8, p.336]. It would have been better if more than two human (privacy) experts would have been involved [8, p.335], but unfortunately many experts are quite reluctant to perform this kind of validation as it is quite tedious and boring. That is also why we have not been able to perform a similar human assessment on the VAT triplets. As an approximation, we have re-applied the automated triple scoring procedure with the VAT term list, instead of the machine gold standard, to the VAT triplets.

The experts only knew they had to assess a set of triplets and were unaware of its origin and related purposes. For them, the goal was to assist in the semi-automated construction of a privacy ontology. The experts have assessed all the privacy triplets output by the ontology miner. They were unaware of the scores of the automatic validation procedure as well as each other's scores (so there was no mutual influence). The experts have marked the list of triplets with '+' or '-' indicating whether or not the triplet is valid, i.e. useful in the context of the creation of a privacy ontology. Their assessments have been merged subsequently. Only those triplets positively scored by both experts have been retained as the human triplet reference. More or less one year after the first rounds of experiments, the privacy experts agreed to perform a second round of scoring - again to all the triplets generated by the ontology miner. This round of scoring was meant to calculate the intra rater agreement. The long interval between the experiments served to avoid a learning effect with the experts. Otherwise, they might remember their previous assessment (or desired outcome).

In addition, a suggestion put forward by the privacy experts has been tested. When discussing the results of the first round of experiments, they had suggested to cut away manually irrelevant parts of the Privacy Directive before inputting it to the ontology miner. Even though the amount of text to be processed decreases (which might compromise the statistical calculations), the hypothesis was that

important terms would be detected more easily. As - except for a rough manual cutting away of irrelevant deemed sections - nothing else in the set up of the experiment has been changed, a comparison with the results obtained earlier became possible (regression test)[11].

4 Results

The CNTS ontology miner has been applied to the VAT and privacy texts. After some format transformation, the miner outputs 315 "subject-verb-object" triplets, such as <*person, pay, tax*>, and 500 "noun phrase-preposition-noun phrase" triplets such as <*accordance, with, article*> resulting in a total of 815 VAT triplets. Concerning the privacy corpus, 1115 *privacy* triplets have been generated by the ontology miner: 276 "subject-verb-object" triplets <*person, have, right*>, 554 "noun phrase-preposition-noun phrase" triplets (<*protection, of, individual*>)and 285 "subject-verb-prepositional object" triplets <*situation, result from, finding*>.

4.1 The Ontology Miner Rated by the Automated Evaluation Procedure

For the VAT corpus, only a list of terms was available. For the Privacy corpus, the "human" gold standard is approximated (consisting of the vocabulary of the triplets positively assessed by the human experts).

Table 1. Lexicometric scores

metrics	VAT	Privacy
coverage	49,26%	79,88%
accuracy	55,97%	95,04%
recall	36,06%	89,91%
precision	55,44%	27,43%

Table 2. VAT machine gold standard vs. human gold standard: $\kappa = 0,3757$

Word reference	"Expert" +	"Expert" -	
Statistics +	299	153	452
Statistics -	379	1375	1754
	788	1528	2206

Word relevance. Table 1 shows the lexicometric scores. There is a clear difference between the two sets of scores (VAT vs. privacy). We explain them by the origin of the human gold standard. For the VAT test, a list of expressions, not necessarily including the same words as used in the VAT directive, has been used. Hence, it is not a surprise that less terms match. For the privacy test, it basically concerns the same words, which might account for the good recall score. Precision is quite low. Probably because only the vocabulary of the positively scored triplets is considered as reference, which might be a too drastic limitation (both the miner and the experts might discard or miss out valid words). Hence, we only calculated the agreement (expressed by the κ-value) between the machine and human gold standard - see Table 2. A rather modest agreement was found.

[11] Due to space restrictions, this aspect is not presented here.

Triplet relevance. In a previous experiment [29], we have investigated 22 different scenarios to distinguish relevant triplets from superfluous ones.

In the situation of ontology engineering we estimate that a high specificity is more interesting than a high sensitivity (less false positives at the detriment of less true positives): a relevant triplet might be missed in order to have less rubbish triplets. The rationale is that it is probably more efficient to reduce the extent of the material ontology engineers have to check and reject compared to their effort needed to detect missing material. Fully automated ontology learning is still not achievable to completely dismiss human experts, so important misses will most probably not remain unnoticed. In the experiments described in [31], we have kept the 95% word relevance confidence level and set the treshold for the triplet scores at 70% (V1 and P1), 70% (V2 and P2) and 90% (V3 and P3). 3091 VAT triplets (V) and 1116 privacy triplets (P) have been generated by the ontology miner. They have been automatically validated in the way described above. Figure 2 [12] displays the outcomes.

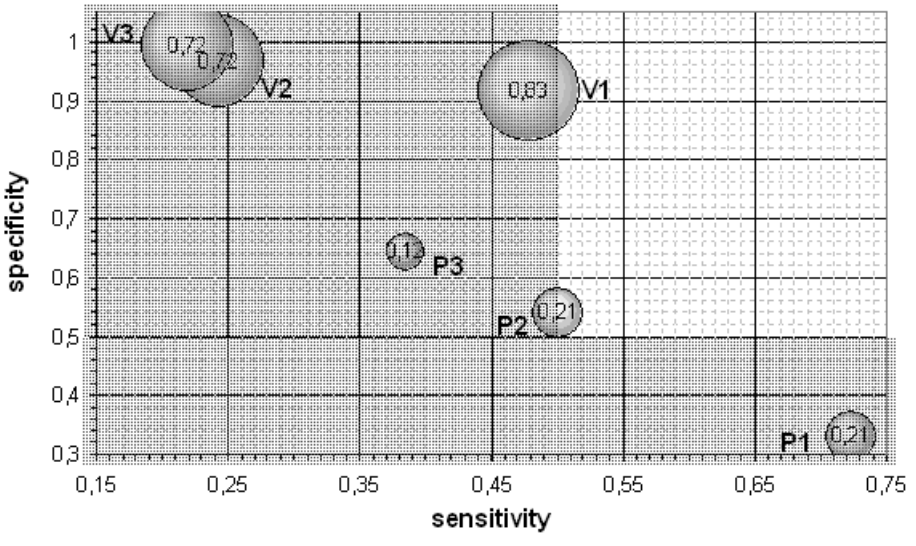

Fig. 2. Sensitivity, specificity and precision scores for the VAT and privacy corpora

As one can see on Figure 2, the 70% threshold produces the best results for the VAT corpus (V1) even if the sensitivity is slightly below 0,5 (shaded zone), and the worse for the privacy corpus (P1: low precision and specificity scores), while the 90% threshold results in almost moderate scores for the privacy corpus (P3) but in an unsatisfactory one in the VAT case (V3: low sensitivity). The 70% threshold gives the most acceptable results for the privacy corpus (P2) but a low sensitivity score in the VAT case (V2).

[12] The size of the bubble represents the precision score.

Table 3. Inter rater agreement (first round) on privacy triplets: $\kappa = -0,0733$

triplets mined	expert 1 -	expert 1 +	
expert 2 -	463	292	755
expert 2 +	248	112	360
	711	404	1115

Table 4. Inter rater agreement (second round) on privacy triplets: $\kappa = 0,1169$

triplets mined	expert 1 -	expert 1 +	
expert 2 -	793	117	910
expert 2 +	216	65	281
	1009	182	1191

Table 5. Intra rater expert 1 agreement on privacy triplets: $\kappa = 0,292$

triplets mined	round 1 -	round 1 +	
round 2 -	672	279	951
round 2 +	39	125	164
	711	404	1115

Table 6. Intra rater expert 2 agreement on privacy triplets: $\kappa = 0,4714$

triplets mined	round 1 -	round 1 +	
round 2 -	679	165	844
round 2 +	76	195	271
	755	360	1115

4.2 The Automated Evaluation Procedure Rated by Domain Experts

The 1116 privacy triplets have been rated by two human experts. The inter rater agreement expressed by the κ value is -0,0733 (first round) and 0,1169 (second round). This almost equals contradiction - see Tables 3 and 4, which means that they agree in a way even less than expected by chance. One of the experts clearly behaved in a rather inconsistent way (intra rater agreement of κ = 0,2936 vs. 0,4714) over the two test rounds - see Tables 5 and 6. The very low inter rater agreement becomes less surprising. These findings support the statement by Friedman and Hripcsak that two experts are not enough [8, p.335] to establish a gold standard. Only the privacy triplets commonly agreed upon by both experts (in a positive (112) and negative (463) sense) have been retained as the human triplet reference. For the VAT corpus, the triplet reference or gold standard has been constructed artificially (see section 3.3).

This probably explains why a modest agreement between the automated scoring procedure and the artificially simulated experts is found (κ value = 0,407). Contrarily, the privacy experts (during the first round of experiments) apparently behaved almost completely in contradiction with the automated procedure.

Table 7. Automated scoring procedure vs. VAT simulated "experts" (threshold 70%) with $\kappa = 0,407$

triplets mined	"Expert" +	"Expert" -	
automaton +	684	136	2271
automaton -	748	1523	820
	1432	1659	3091

Table 8. Automated scoring procedure vs. Privacy experts (threshold 70%, round 1) with $\kappa = 0,026$

triplets mined	Expert +	Expert -	
automaton +	56	213	269
automaton -	56	250	306
	112	463	575

5 Discussion

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). Currently this has not been explored yet, although the required data are available. As soon as the scores for a particular (and commonly agreed upon) textual source have been scientifically validated, the source and the scores together can be re-used as an evaluation standard in bench-marking tests involving other ontology miners, or even to some extent any RDF-based ontology producing tool. A logical next step would be that ontologies, automatically created by an ontology miner, are documented with performance scores on their textual source material as well as with scores for that particular miner on an evaluation reference (commonly agreed corpus and outcomes) - as e.g. customarily happens in the speech recognition industry.

5.1 Rating the Ontology Miner

Unfortunately, we cannot add a lot to our findings in the previous section for the VAT corpus as the VAT experts did not participate in assessing the triplets generated by the miner. The privacy experts did provide some comments. As a result, the following improvements could be implemented.

- The background ("neutral") corpus (here the WSJ) is key in establishing the machine gold standard. However, legal documents have many terms which are relevant to the legal domain in general, but not relevant to the particular legal domain under consideration. In future experiments using legal documents it is recommended to use a background corpus of terms taken from a set of European legal documents. For example, the term "Member State" is highly relevant to European Legislation in general, but has no specific relevance to the privacy domain. This is an example of a term that was judged highly relevant by the miner, but totally irrelevant by the experts.
- An issue not addressed is that of abstraction. Human experts extracting terms from a corpus are able to amalgamate synonyms and instances of higher level concepts where the use of lower level terms is of no use to the application domain. For instance, the privacy directive gives a list of data types which it is prohibited to collect without the data subject's consent. To a human expert, these classes of data are clearly what is known as "sensitive data". The inclusion of a synonym dictionary would go some way towards term abstraction although it can only take account of equivalence and not subclass relationships between terms. Currently, the tests and calculations depend too much on string matching.

The automated evaluation procedure assessed the ontology miner as rather "modestly" producing material reliably suitable for ontology engineering. Not only the miner misses more or less half of the interesting material but additionally

the quality of the material generated is not consistent (see Figure 2). It currently seems infeasible to define an appropriate threshold setting suited for both cases. Even if these results might be not so unexpected for an unsupervised miner (for supervised miners still perform better), ontology engineers in the field most probably are less impressed or helped by such material.

5.2 Rating the Automated Evaluation Procedure

The experiments do not allow to draw valid conclusions concerning the "goodness of fit" of the automated scoring procedures. The main reason is the insufficient (or even completely missing) availability of experts in general to establish a valid human gold standard. Although two privacy experts have participated, they did not rate in a consistent way casting doubt on the validity of the privacy human gold standard. A more detailed analysis should reveal whether or not this is due to one or the other expert. The mere fact that the machine gold standard does not represent state of the art techniques and methods is irrelevant in this matter.

5.3 "Rating" the Evaluation Set-Up

Following the criteria of Friedman and Hripcsak [8], we have clearly described the method applied to evaluate the results. Inter and intra rater agreement scores have been calculated, showing that one of experts did not score in a consistent way. Also the lexicometric and triplet overlap scores have been described in detail as they are used to establish a gold standard. This also allows to easily discover the limits of our experiments, which also complies with criteria set by Friedman and Hripcsak [8, pp.336-337]. In particular, the fact that the experiments are not completely symmetric. Also, it would have been interesting to have experts build an ontology completely by hand and use this as a human gold standard instead of validating machine generated output.

By involving two different domains in the evaluation experiment, we tested to which extent outcomes can be generalised over several application domains. Currently, due to the practical circumstances of the evaluation, one should not generalise the findings, either in a positive or negative way. Basically, conclusions can only be indecisive as for the VAT corpus, the expert involvement was to a large extent lacking, while for the Privacy corpus, not enough experts have been involved. One can wonder how many conclusions concerning evaluations of ontological material or ontologies reported in the literature will survive an analysis of the evaluation set-up as scrutinous as the presented here. E.g., one rarely finds inter and intra rater agreement numbers. Often developers of ontology learning applications also perform the evaluation - sometimes even as a sole evaluator.

6 Related Work

Previous reports on our work contain additional details on the unsupervised miner [22], its application to a bio-medical corpus and a qualitative evalua-

tion [23]. The method and previous quantitative experiments have been presented in [28,29,30,31,32]. Various researchers are working on different ways to evaluate an ontology from various perspectives. Good overviews of the recent state of the art that also contain a comparison of the characteristics of the various methods are [2,6,7,9,14,21].

A somewhat related topic is that of ontology selection and ranking: ontologies are evaluated as part of a selection process to choose the most appropriate ontology for a purpose or task (e.g. [1,24]). Some researchers have evaluated methods and metrics to select the most appropriate terms (e.g, [10,19]) from texts for building an ontology. However, these latter do not evaluate entire triplets. Others are active in ontology based information extraction (OBIE) and present metrics to evaluate OBIE performance - e.g., [18]. Next to that, one could consider additionally work that measures the similarity between two ontologies [17].

Only a few other approaches address the quantitative and automated evaluation of an ontology by referring to its source corpus. *Brewster* and colleagues have 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 [3]. Unfortunately, till now no concrete results or test cases have been presented.

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 [36]. 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, which enables us to filter out words that are only seemingly typical of the technical text. In [19], 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* [10] 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).

There is the - no longer continued - work of *Sabou* [25] who has examined how to learn ontologies for web services from their descriptions. Although the practical aspects of her work on the ontology learning aspects are quite tailored towards the application domain, the evaluation method resembles ours. She has "established a one-to-one correspondence between phrases in the corpus and derived concepts", so that our lexicometric scores are comparable to her ontology ratios. In more or less the same vein, *Gulla* et colleagues [13] use a

keyphrase extraction techniques to semi-automatically build an ontology. They involve domain experts to evaluate the ontology in a more or less task independent way. Queries are run against a separate manually built ontology and the semi-automatically constructed one.

Concerning the meta-evaluation of ontologies, *Gangemi* [9] provides in his impressive overview (and ontology) of ontology evaluation metrics and measures some elements and insights that come close to some of our findings. We have rather focused on a meta-analysis of how content material for an ontology can be assessed by means of a gold standard approach and which pitfalls are to be avoided to obtain methodologically sound outcomes.

7 Future Work

The same experiments can be repeated using another text miner - e.g., [15] - based on other algorithms or heuristics. Also other scoring measures (e.g. the weirdness measure [10]) to determine whether or not terms are relevant can be tried in the future. In the same line of thinking, we could look for other user friendly term extraction tools. We hope to test our method on other domains, pending the availability of sufficient appropriate domain experts. In addition, a regression test can be performed with the set up as described in this paper. All these experiments will also provide new insights to extend the framework for the meta-evaluation of ontologies.

8 Conclusion

The current experiments give an indecisive answer to the question whether the automatic evaluation procedure is up to providing a reliable indication on the quality of triplets produced by a ontology miner. The main reason is the imperfect manner in which the experiments had to be set up, constrained as they were by practical limitations in working conditions. The main lesson to be drawn is that ontology evaluation, especially when it concerns aspects that go beyond mere consistency checking, counting of ontology nodes or other "mechanical" or structural checks [38], is a fragile exercise. In order to produce scientifically valid results, an important number of conditions has to met with - as illustrate our experiences. Evaluating how ontology evaluation should happen is a rather novel research topic, which is not a surprise as the topic of ontology evaluation itself still offers many further avenues of research to be explored. The growing number of recent publications in this area illustrates that the topic is becoming a valuable research area. And the criterion whether or not an ontology adequately covers a domain cannot be addressed only in an impressionistic way by having people rate ontologies- cf. [20]. If well designed (e.g., [37]), computer assisted evaluation of ontologies is possible. And maybe introducing gaming aspects [27] could alleviate the psychological burden?

The lightweight automated evaluation procedure reported on in this paper aims at reducing the need to call upon experts, who are, in general, reluctant

to participate in evaluation procedures. However, the experiments and results show clearly that an active involvement of several appropriate experts of various backgrounds is still crucially needed at this stage. How to break this catch-22 (or deadlock) situation?

Acknowledgments

Parts of this research have been supported by the the OntoBasis project (GBOU 2001 #10069) of the IWT Vlaanderen (Institution for the Promotion of Innovation by Science and Technology in Flanders) and by the EU FP6 IP PRIME (IST 2002-507591) project. We are particularly indebted to dr. Marie-Laure Reinberger (at the time at the Universiteit Antwerpen - CNTS), who has produced the VAT and pricavy triplets as well as a lemmatised version of the WSJ, to dr. Giles Hogben (at the time at the EU Joint Research Centre IPSC, Italy) and to drs. John Borking (Borking Consultancy, The Netherlands). Both acted as the privacy domain experts. In addition, we gratefully acknowledge Prof. dr. Patrick Wille (VAT@ NV, Belgium and partner of the EU FP5 IST FF Poirot consortium) for putting at our disposal the hand-crafted list of 900 VAT terms.

References

1. Alani, H., Brewster, C., Shadbolt, N.: Ranking ontologies with aktiverank. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 1–15. Springer, Heidelberg (2006)
2. Brank, J., Grobelnik, M., Mladenić, D.: Ontology evaluation. SEKT Deliverable #D1.6.1, Jozef Stefan Institute, Prague (2005)
3. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Shadbolt, N., O'Hara, K. (eds.) Advanced Knowledge Technologies: selected papers, pp. 164–168. AKT (2004)
4. Buchholz, S., Veenstra, J., Daelemans, W.: Cascaded grammatical relation assignment. In: Proceedings of EMNLP/VLC 1999, PrintPartners Ipskamp (1999)
5. Buitelaar, P., Cimiano, P., Loos, B. (eds.): Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Association for Computational Linguistics (2006)
6. Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press, Amsterdam (2005)
7. Burton-Jones, A., Storey, V., Sugumaran, V.: A semiotic metrics suite for assessing the quality of ontologies. Data and Knowledge Engineering 55(1), 84–102 (2005)
8. Friedman, C., Hripcsak, G.: Evaluating natural language processors in the clinical domain. Methods of Information in Medicine 37(1-2), 334–344 (1998)
9. Gangemi, A., Catenacci, C., Ciaramita, M., Gil, R., Lehmann, J.: Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. Technical report (2005), `http://www.loa-cnr.it/Publications.html`
10. Gillam, L., Tariq, M.: Ontology via terminology? In: Ibekwe-San Juan, F., Lainé Cruzel, S. (eds.) Proceedings of the Workshop on Terminology, Ontology and Knowledge Representation (2004), `http://www.univ-lyon3.fr/partagedessavoirs/termino2004/programgb.htm`

11. Guarino, N., Persidis, A.: Evaluation framework for content standards. OntoWeb Deliverable #D3.5, Padova (2003)
12. Guarino, N., Welty, C.: Evaluating ontological decisions with OntoClean. Communications of the ACM 45(2), 61–65 (2002)
13. Gulla, J., Borch, H., Ingvaldsen, J.: Ontology learning for search applications. In: Meersman, R., Tari, Z., et al. (eds.) OTM 2007, Part I. LNCS, vol. 4803, pp. 1050–1062. Springer, Heidelberg (2007)
14. Hartmann, J., Spyns, P., Maynard, D., Cuel, R., de Figueroa, S., Sure, Y.: Methods for ontology evaluation. KnowledgeWeb Deliverable #D1.2.3, 3 (2005)
15. Judge, J., Sogrin, M., Troussov, A.: Galaxy: IBM ontological network miner. In: CSSW. LNI, vol. 113, pp. 157–160 (2007)
16. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2(2), 159–195 (1958)
17. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 251–263. Springer, Heidelberg (2002)
18. Maynard, D., Peters, W., Li, Y.: Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D. (eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Paris, European Language Resources Association (2008)
19. Navigli, R., Velardi, P.: Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics 30(2), 151–179 (2004)
20. Noy, N., Guha, R., Musen, M.: User ratings of ontologies: who will rate the raters? In: AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors (2005)
21. Obrst, L., Ashpole, B., Ceusters, W., Mani, I., Ray, S., Smith, B.: Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. In: The Evaluation of Ontologies: toward Improved Semantic Interoperability, pp. 139–158. Springer, Heidelberg (2007)
22. Reinberger, M.-L., Spyns, P.: Unsupervised text mining for the learning of DOGMA-inspired ontologies. In: Buitelaar, Ph., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Applications and Evaluation, pp. 29–43. IOS Press, Amsterdam (2005)
23. Reinberger, M.-L., Spyns, P., Pretorius, A.J., Daelemans, W.: Automatic initiation of an ontology. In: Meersman, R., Tari, Z. (eds.) OTM 2004. LNCS, vol. 3290, pp. 600–617. Springer, Heidelberg (2004)
24. Sabou, M., Lopez, V., Motta, E., Uren, V.: Ontology selection: Ontology evaluation on the real semantic web. In: Proceedings of the 4th International EON Workshop, Evaluation of Ontologies for the Web (2006), `http://eprints.aktors.org/487/`
25. Sabou, M., Wroe, C., Goble, C., Mishne, G.: Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In: Proceedings of the 14th International World Wide Web Conference (2005)
26. Shamsfard, M., Barforoush, A.: The state of the art in ontology learning: a framework for comparison. Knowledge Engineering Review 18(4), 293–316 (2003)
27. Siorpaes, K., Hepp, M.: Games with a purpose for the semantic web. IEEE Intelligent Systems 23(3), 50–60 (2008)
28. Spyns, P., Reinberger, M.-L.: Lexically evaluating ontology triples automatically generated from text. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 563–577. Springer, Heidelberg (2005)

29. Spyns, P.: Evalexon: assessing triples mined from texts. Technical Report 09, STAR Lab, Brussel (2005)
30. Spyns, P.: Object role modelling for ontology engineering in the DOGMA framework. In: Meersman, R., Tari, Z., Herrero, P., et al. (eds.) OTM-WS 2005. LNCS, vol. 3762, pp. 710–719. Springer, Heidelberg (2005)
31. Spyns, P.: Validating evalexon: validating a tool for evaluating automatically lexical triples mined from texts. Technical Report x6, STAR Lab, Brussel (2005)
32. Spyns, P., Hogben, G.: Validating an automated evaluation procedure for ontology triples in the privacy domain. In: Moens, M.-F., Spyns, P. (eds.) Proceedings of the 18th Annual Conference on Legal Knowledge and Information Systems (JURIX 2005), pp. 127–136. IOS Press, Amsterdam (2005)
33. Spyns, P., Pretorius, A.J., Reinberger, M.-L.: Evaluating DOGMA-lexons generated automatically from a text. In: Cimiano, P., Ciravegna, F., Motta, E., Uren, V. (eds.) EKAW 2004. LNCS (LNAI), vol. 3257, pp. 38–44. Springer, Heidelberg (2004)
34. Stvilia, B.: A model for ontology quality evaluation. First Monday 12(12) (2007)
35. Thompson Automation Software, Jefferson OR, US. Tawk Compiler, v.5 edition
36. Velardi, P., Missikoff, M., Basili, R.: Identification of relevant terms to support the construction of domain. In: Maybury, M., Bernsen, N., Krauwer, S. (eds.) Proc. of the ACL-EACL Workshop on Human Language Technologies (2001)
37. Vossen, P., Agirre, E., Calzolari, N., et al.: Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D. (eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Paris, European Language Resources Association (2008)
38. Vrandecic, D., Sure, Y.: How to design better ontology metrics. In: May, W., Kifer, M. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 311–325. Springer, Heidelberg (2007)
39. Zipf, G.K.: Human Behaviour and the Principle of Least-Effort. Addison-Wesley, Cambridge (1949)