Evaluating DOGMA-lexons generated automatically from a text corpus

Peter Spyns[1], A. Johannes Pretorius[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,Hannes.Pretorius}@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 was to devise a method to evaluate the results of extracting semantic relations from text corpora in an unsupervised way. We have worked on a legal corpus (EU VAT directive) consisting of 43K words. Using a shallow parser, a set of ”lexons” has been produced. These are to be used as preprocessed material for the construction of ontologies from scratch. A knowledge engineer has judged that the outcome is useful to support and speed up the ontology modelling task. In addition, a quantitative scoring method (coverage and accuracy measures resulting in a 52.38% and 47.12% score respectively) has been applied.
Keywords: text mining, ontology creation, (quantitative) evaluation

1 Introduction and Background

The development of ontology-driven applications is currently slowed down due to the knowledge acquisition bottleneck. Therefore, techniques applied in computational linguistics and information extraction (in particular unsupervised machine learning [13, 14]) can be used to create or grow ontologies in a period as limited as possible with a quality as high as possible. 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, next to a more ”classical” qualitative evaluation.

Criteria for ontology evaluation have been put forward by Gruber [6, p.2] and taken over by Ushold and Grüninger [17]: clarity, coherence, extendibility, minimal encoding bias and minimal ontological commitment. Gómez-Pérez [4, p.179] has proposed consistency, completeness and conciseness. Neither set of criteria are well suited to be applied in our case as the lexons produced by the unsupervised miner are merely ”terminological combinations” (i.e., no explicit meaning for the terms and roles is provided, not to mention any formal definition of the intended semantics). We have been mainly inspired by the criteria proposed by Guarino (coverage, precision and accuracy) [7, p.7], although there are problems to ”compute” them in the current practice (unlike in information extraction evaluation exercises) as there are no (suitable) ”gold standards” available.

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

2 Methods and Material

2.1 DOGMA

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 [3]. 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 [16]. 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 [10].

2.2 Unsupervised Text Mining

We have opted for extraction techniques based on unsupervised learning methods [13] 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, which is being developed at CNTS Antwerp and ILK Tilburg [2] [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.

2.3 Corpus

In a specific domain, an important quantity of semantic information is carried by the noun phrases (NP). At the same time, the NP-verb relations provide relevant information about the NPs, due to the semantic restrictions the verbs impose. The VAT corpus (a single long document) consists of 43K 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

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

[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.

based shallow parser to this corpus. After an additional selection step and some format transformation, the parser produces triples, such as <*person pay tax*> (see [13, 15] for more details on the linguistic processing).

2.4 Evaluation Criteria

The main research hypothesis in this paper is that lexons, representing the basic binary facts expressed in natural language about a domain, can be extracted from the available textual sources using the shallow parser described above. In this section, two empirical evaluation methods are described.

Qualitative method. A human knowledge engineer who has manually built a lexon base for this domain using the same corpus, has been asked to evaluate the usefulness of the results. Some questions have been formulated independently of the evaluator. The evaluator/knowledge engineer, relying on his past experience, was given three criteria to rate the lexons produced by the unsupervised miner.

- coverage (have all the lexons to be discovered actually been discovered)
- precision (are the lexons making sense for the domain ? [6])
- accuracy (are the lexons not too general but reflecting the important terms of the domain ?)

Quantitative method. In addition, for the coverage and accuracy criteria we have tried to define a quantitative measure and semi-automated evaluation procedure. We don't define a computable precision measure here (see [13] for an earlier attempt).

The underlying idea is inspired by Zipf's law [19]. It states that the frequency of the occurrence of a term is inversely proportional to its frequency class. Zipf has discovered experimentally that the more frequently a word is used, the less meaning it carries. E.g., in the 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 1155 words in the frequency class 1. Important for our purpose is the observation that the higher frequency classes contain mostly "empty" words (also called function words). A corollary is that domain or topic specific vocabulary is to be looked for in the middle to lower frequency classes. Consequently, lexons mined from a corpus should preferably contain terms from these frequency classes.

As the DOGMA lexons resulting from the unsupervised mining consist of three words[7] (two terms and one role[8]) extracted from the corpus, it is possible to investigate to what extent the produced lexons cover the corpus vocabulary, and more importantly how accurate they are. Note that the same technique can be applied to RDFS ontologies.

Coverage will be measured by counting for each frequency class the number of lexon terms that are identical with terms from the corpus and comparing this number

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

[7] In fact, the words have been lemmatised. E.g., working, works, worked → work.

[8] Co-roles are not provided by unsupervised miner.

to the overall frequency class term count. *Accuracy* will be estimated on basis of the coverage percentage for particular frequency classes.

However, *some caveats* should be made from the on-set. It should be clear that a coverage of 100% is an illusion. Only terms in a Verb-Object and Subject-Object grammatical relation are selected and submitted subsequently to several selection thresholds (see section 2.2). Regarding the accuracy, determining exactly which frequency classes contain the terms most characteristic for a domain is still an exercise based largely on intuition and impression. It should also be kept in mind that no stopword list has been defined because lexons have been produced with a preposition assuming the role function.

3 Evaluation Results

3.1 Qualitative method

When applied to the VAT corpus, the unsupervised mining exercise outlined above resulted in the extraction of 817 triples or lexons. These were analysed by a knowledge engineer using the LexoVis lexon visualisation tool [12]. This analysis was rather informal in the sense that the knowledge engineer was largely guided by intuition, knowledge and experience gained with the manual extraction of lexons from the VAT legislature domain.

A first important aspect to consider is whether the domain (VAT legislature) is adequately described (or *covered*) by the set of extracted triples. It soon became apparent that there is a significant amount of noise in the mining results. The triples need to be significantly cleaned up in order to get rid of inadequate (and often humorous) structures such as <*fishing, with, exception*>. The percentage of inadequate lexons seems to fall in excess of 53%. According to this percentage, approximately 384 of the resulting 817 lexons may be deemed usable. If this is compared to the number of lexons resulting from a manual extraction exercise on the same corpus of knowledge resources (approximately 900), there is doubt as to whether the domain is adequately covered by the results.

As mentioned above, there is a significant portion of the unsupervised mining exercise results which are deemed inadequate. Firstly, this can be contributed to the fact that many resulting lexons are not *precise* (intuitively, they do not make sense in the context of the VAT domain as the fishing example above illustrates). Furthermore, many of resulting triples were not considered *accurate* in the sense of describing important terms of the domain. In this respect, only the term VAT occurs in three lexons, <*VAT, in, member*>, <*VAT, on, intra-Community_acquisition*> and <*VAT, to, hiring*> which are not considered appropriate to accurately describe the concept of VAT in the domain under consideration. In the same respect, there is only one mention of the term *Fraud*. The notion of *redundancy* is harder to evaluate, since terms and roles may have synonyms. However, the intuitive impression of the results is that redundancy is not a critical problem.

3.2 Quantitative method

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

- $FC < 3$: many non-words and/or too loosely related to the domain
- $3 < FC < 20$: domain related technical language
- $20 < FC < 50$: general language used in a technical sense
- $50 < FC < 300$: mixture of general language and domain technical language
- $300 < FC < 500$: general language and highly used domain terms
- $FC < 500$: function words and highly used general language terms

We determine the area with ”resolving power of significant words” [9, p.16] to be the range of frequency classes 3 till 40. The range encompasses 596 terms that one would expect to be covered by the lexons. Figure 1 show that the coverage improves with the increasing rank of the frequency class. On average, the coverage ratio is 52.38%. The accuracy (i.e. the coverage percentage for the selected interval) ratio for the 3-40 interval is 47.31%.

4 Discussion & Related Work

Performing an automatic evaluation is hard task, and evaluation frequently implies a manual operation by an expert [1, 3], or by the researchers themselves [5]. An automatic evaluation is nevertheless performed in [8, 11] by comparison with existing thesauri like WordNet and Roget. Our attempt takes the corpus itself as reference and reduces the need for human intervention. Humans are still needed to clean the corpus (e.g. to choose the stopwords and to remove the non-words), but do not intervene in the evaluation process itself, except for setting the frequency class interval. Regression tests can also be done. Currently, we estimate that the accuracy should be improved. Taking synonyms into account might help. On the other hand, more research should be done to determine the proportion of domain technical terms vs. general language terms in the ”relevant” frequency class interval. Measures as ”domain relevance” and ”domain consensus” [18] might be useful, but were not applied here as the corpus consisted of one document (for one domain). If we look at it from a positive angle, we could argue that already half of the work of the domain specialist and/or terminographer of selecting the important domain terms/lexons is done. We were specifically (but happily) surprised by the fact that the different evaluation techniques performed in an independent way lead to similar conclusions.

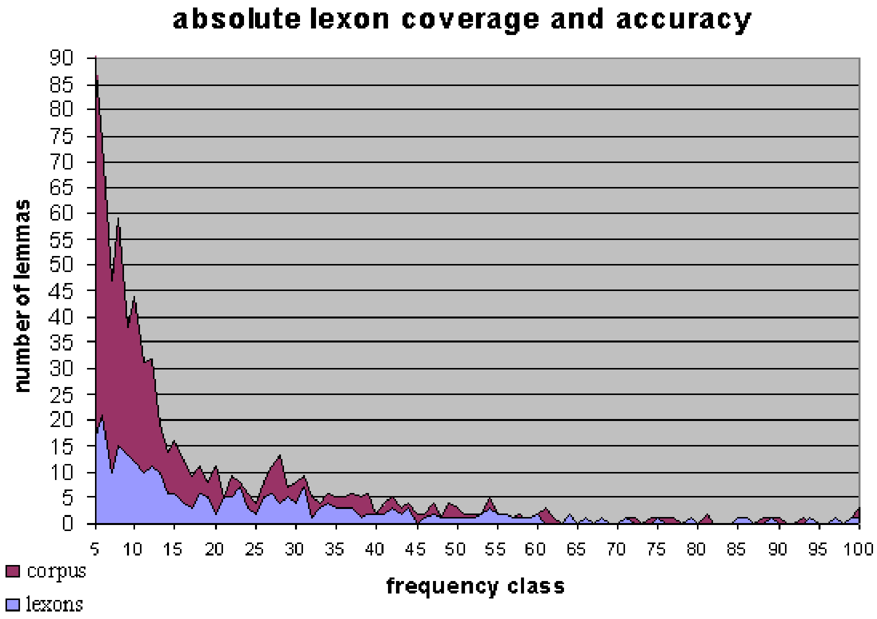

Fig. 1. absolute coverage and accuracy of frequency classes by lexon terms

5 Future Work

With respect to the quantitative evaluation of the outcomes of the unsupervised mining, insights from information science technology should be taken into account to answer some remaining questions. E.g. does the length of a document influence the determination of the most meaningful frequency class interval ? Is it possible to establish a statistical formula that represents the distribution of meaningful words over documents ? Once this interval can be reliably identified, one could apply the unsupervised learning algorithm only to sentences containing words belonging to frequency classes of the interval. This could be easily done after having made a concordance (keyword in context) for the corpus. We would like to carry out this experiment on a corpus of another domain, thereby also applying the domain relevance and domain consensus measures [18].

6 Conclusion

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

Acknowledgments The major part of this research was carried out in the context of the OntoBasis project (GBOU 2001 #10069), sponsored by the IWT Vlaanderen. Hannes Pretorius is supported by the EU IST FP5 FF Poirot project (IST-2001-38248).

References

1. Didier Bourigault and Christian Jacquemin. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings EACL-99*, 1999.
2. Sabine Buchholz, Jorn Veenstra, and Walter Daelemans, Cascaded grammatical relation assignment, in *Proceedings of EMNLP/VLC-99*. PrintPartners Ipskamp, (1999).
3. David Faure and Claire Nédellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proc. EKAW-99*, 1999.
4. Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. Springer Verlag, 2003.
5. Ralph Grishman and John Sterling. Generalizing automatically generated selectional patterns. In *Proc. of COLING-94*, 1994.
6. T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221, 1993.
7. Nicola Guarino and Andreas Persidis. Evaluation framework for content standards. Technical Report OntoWeb Deliverable #3.5, Padova, 2003.
8. Dekang Lin. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL-98*, 1998.
9. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159 – 195, 1958.
10. Robert Meersman. Ontologies and databases: More than a fleeting resemblance. In A. d'Atri and M. Missikoff, editors, *OES/SEO 2001 Rome Workshop*. Luiss Publications, 2001.
11. Roberto Navigli and Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151–179, 2004.
12. A.J. Pretorius. Lexon visualization: visualizing binary fact types in ontology bases. In *Proceedings of the 8th international conference on information visualisation (IV04)*, London, 2004. IEEE Press. in press.
13. 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, pages 803 – 819, 2003. Springer.
14. Marie-Laure Reinberger and Peter Spyns. Discovering knowledge in texts for the learning of dogma-inspired ontologies. In Paul Buitelaar, Siegfried Handschuh, and Bernardo Magnini, editors, *Proc. of the ECAI04 Workshop on Ontologies, Learning and Population*, 2004.
15. 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*, LNCS (in press), 2004. Springer Verlag.
16. Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *SIGMOD Record Special Issue*, 31 (4): 12 - 17, 2002.
17. M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge Sharing and Review*, 11(2), June 1996.
18. 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
19. George K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA, 1949.