Book Title
Book Editors
IOS Press, 2003

Validating an Automated Evaluation Procedure for Ontology Triples in the Privacy Domain

Peter Spyns [a,1] and Giles Hogben [b]
[a] *Vrije Universiteit Brussel - STAR Lab, Belgium*
[b] *European Commission Joint Research Centre - IPSC, Ispra, Italy*

Abstract. In this paper we validate a simple method to objectively assess the results of extracting material (c.q. triples) from text corpora to build ontologies. The EU Privacy Directive has been used as corpus. Two domain experts have manually validated the results. Several experimental settings have been tried. As the evaluation scores are rather modest (sensitivity or recall: 0.5, specificity: 0.539 and precision: 0.21), we see them as a baseline reference for future experiments. Nevertheless, the human experts appreciate the automated evaluation procedure as sufficiently effective and time-saving for usage in real-life ontology modelling situations.

Keywords. Ontology Evaluation, Privacy Ontology Mining

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 (future) knowledge based society. However, the process of conceptualising an application domain and its formalisation require substantial human efforts. Therefore, techniques applied in human language technology (HLT) and information extraction (IE) are used to create or grow ontologies. Work is still in progress - recent overviews of the state of the art (in particular for machine learning techniques) can be found in [1,2]. Even in the ideal case that (semi-)automated HLT and IE methods have become mature, there still remains the problem of assessing and evaluating the results. Recent proposals for evaluation methods are [1,3].

All these approaches share basically the same problem, i.e. how to determine a gold standard valid for various application situations. Rare are the experts who are willing to devote their time to validate output generated by a machine or establish in agreement a gold standard for a particular technical domain. Therefore, we want to define a lightweight assessment procedure that is easy to understand and apply by "'standard knowledge workers"' (basically a domain expert, a computer scientist, an engineer, ...) outside the specialised academic environment.

[1]correspondence to: Peter Spyns, Vrije Universiteit Brussel - STAR Lab, Pleinlaan 2, Gebouw G-10, B-1050 Brussel, Belgium. Tel.: +32-2-629 3753; Fax: +32-2-629 3819; Email: Peter.Spyns@vub.ac.be.

The method should be generally applicable (any kind of text miner, any kind of text) 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. Typical of our approach will be that only the corpus (lemmatised [1] but otherwise unmodified) constitutes the reference point, and not an annotated corpus or another gold standard ontology. The reason being that these artifacts require quite some human effort to be built. However, the evaluation procedure itself should first be validated, and therefore we have to rely on human experts. This paper focusses on this latter aspect.

As a test corpus, we used the European Data Protection Directive (95/46/EC) in the application domain of the PRIME project [2]. The use-case for the final ontology is the creation and processing of machine-readable data-handling practice statements (privacy policies) by data processors within the European Union jurisdiction. The semantics of the ontology will support the evaluation of practice statements when deciding whether to release data, either by users or automated evaluators mandated by users (user-agents). It will also be used to support administators in creating machine-readable policies, which are legally compliant.

The remainder of this paper is organised as follows. The next two sections present the material (section 2) and methods (section 3): the lexicometric scores are shortly explained in section 3.1, after which the procedure to evaluate triples is discussed (section 3.2) as well as its validation method (section 3.3). The outcomes of the lexicometric measures (section 4.1) as well as the triple scoring procedure (section 4.2) are described and discussed subsequently in section 5. 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], has been used. Semantic relations that match predefined syntactic patterns have been extracted from the shallow parser output. Additional statistics using normalised frequencies and probabilities of occurrence are calculated to separate noise (i.e. false combinations generated by chance during clustering) from genuine results. More details on the linguistic processing can be found in [5].

The *Privacy corpus* (a single long document) consists of 72,1K words. It constitutes the EU directive on Privacy (95/46/EC of 18 dec 2000 - English Version) that has to be adopted and transformed into local legislation by every Member State. This was chosen because the document provides the legal context in which the application domain's use cases will operate. The ontology is used to model privacy enhanced access control policies (access rules controlling data processing events) within Europe. E.g., it will be able to provide an automated judgement as to the legality of a data processing event within the EU.

The CNTS text miner has been applied to this corpus. After some format transformation, it outputs "subject-verb-object" triples, such as *<third_country, ensure,*

[1] 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.

[2] See http://www.prime-project.eu.org

level_of_protection>, "noun phrase-preposition-noun phrase" triples such as <*Treaty, on, European_Union*> and subject-verb-prepositional object triples, such as <*controller, establish, in_Member_State*>, in total 1116 triples. In addition, the *Wall Street Journal (WSJ) corpus* (a collection - 1290K words - of English newspaper articles) serves as a "neutral" corpus that is to be contrasted with the specific technical vocabulary of the Privacy Directive.

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

3. Methods

An ontology is supposed to represent the most relevant concepts and relationships of a domain of discourse. The terms lexicalising these concepts and relationships are to be retrieved from a corpus of texts about the domain. The key question is how to determine in an automated way which are the important terms (section 3.1). In addition, an algorithm is needed to distinguish relevant combinations (i.e. two concepts in a valid relationship - a triple) from irrelevant ones (section 3.2). Both steps have to be validated. In this paper, we'll primarily discuss the validation of triples (section 3.3), the validation of the relevant terms having been presented elsewhere [7].

3.1. Determining the Relevant Words

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 (FC 1), and there is one word that appears 1163 times (FC 1163). According to Zipf's law [8], 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 [9] introduced the notion of "resolving power of significant words", by defining intuitively a frequency class upper and lower bound. The most significant words are found in the middle of the area of the FCs between these boundaries.

We propose to calculate whether the FC is relevant or not. Only if a FC is composed by 60%[3] 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. Calculations have been done for 95% and 99% confidence level. The relevant words are used as the gold standard for the precision and recall metrics, which has otherwise to be created by human experts. The coverage (involving all FCs) and accuracy (involving only relevant FCs) metrics indicate to which degree the vocabulary of a triple mining tool coincides with the set of relevant words. Due to space restrictions, we have to refer the reader to [7] for more details on the metrics.

[3] Currently, this threshold has been choosen arbitrarily.

3.2. Determining the Relevant Triples

Again, the relevant terms are used as reference (one set computed with a 95% confidence level and one with a 99% level). Several basic scenarios have been set up in which a triple (= Subject-Predicate-Object) is considered relevant (see also Table 3 and Table 4):

- Two (A1, B1) or three (C1,D1) of the three triple parts contains a term statistically relevant.
- The degree with which a triple lexically overlaps with terms of the reference set depasses a certain level: 65% (X1, Y1), 70% (X2, Y2), and 90% (X3, Y3).
- The degree of lexical overlap is combined with the number of triple parts containing a relevant term.

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 (unlike e.g. in IE applications) since they are included in the triples automatically generated.

3.3. Validating the (Relevant) Triples

Two experts have been asked to independently validate the list of triples as produced by the text miner. One has been a privacy data commissioner and still is a lawyer while the other an knowledge engineer specialised in the field of privacy and trust. They have marked the list of triples with '+' or '-' indicating whether or not the triple is valid, i.e. usefull in the context of the creation of a privacy ontology. It allows us to evaluate the "goodness of fit" of the automated procedure compared to a human reference. Of course, it would have been better if more than two human experts would have been involved, but many experts are reluctant to perform this kind of validation as it is quite tedious and boring.

4. Results

In total 1116 triples have been generated by the CNTS unsupervised text miner: 278 have a verb, 557 a preposition and 286 a verb-preposition combination on the predicate position. 49 FCs are considered as relevant. It is important to remember that the scores mentioned in section 4.1 measure the intersection of corpus vocabulary considered relevant with vocabulary derived from a series of triples generated automatically. The scores given in section 4.2 concern the goodness of fit of an automatic evaluation procedure compared to (two) human experts who have evaluated the same set of triples.

4.1. Word Relevance

When only taking into consideration the 49 relevant FCs of the privacy corpus, the recall is 95,58%, the precision is 96,29% and the accuracy is 95,04%. The κ value comparing the miner results with the statistical results (again only for the relevant FCs) is 0,69%, which signifies a good agreement. These good results are probably obtained because the FCs from 1 till 10 are not considered relevant. However, they contain the bulk of the vocabulary. From another experiment (using the EU VAT Directive as corpus) in which

an expert manually had created a list of relevant terms, there is a strong indication that the lower FCs cannot be ruled out [7]. An option would be to relax the 60% treshold. When taking all the privacy FCs into consideration (which is a too simplistic relaxation), the recall is 89,91%, the precision 27,43% and the coverage 79,88%. The κ value now is 0,14% (very close to contradiction).

4.2. Triple Relevance

The 1116 triples have been rated by two human experts ('+' for appropriate vs. '-' for non appropriate). Their ratings resulted in a κ value of -0,0742 (= almost contradiction - see Table 1), which means that they agree in a way even less than expected by chance. For subsequent tests, only the triples commonly agreed upon (in a positive (112) and negative (463) sense) have been retained as the reference.

Table 1. Inter experts agreement: $\kappa = -0,0742$.

	Expert 2 +	Expert 2 -	
Expert 1 +	112	292	404
Expert 1 -	249	463	712
	361	756	1116

Some examples of triples are shown in Table 2. E1 and E2 correspond to how expert 1 resp. expert 2 rated a triple. X% and Y% indicate the absolute lexical overlap percentage (see below). In addition, various different evaluation scenarios have been investigated - see Table 3 for the exact settings.

Table 2. Some triples and their status in eight evaluation scenarios - see Table 4.

triple	E1	E2	X%	Y%	A1	B1	C1	D1
breach, of, right	+	+	66	66	+	+	+	-
functioning, of, internal_market	-	+	53	33	+	+	+	+
recording, of, personal_data	+	+	66	66	+	+	+	+
rise, to, damage	+	+	33	33	-	-	-	-
chairman_term, be, two_year	+	-	0	0	-	-	-	-
exercise, be, public_administration	+	-	44	44	+	-	+	-
own_initiative, make, recommendation_on_matter	+	+	9	0	+	-	-	-
course_activity, fall, outside_scope	+	+	48	35	+	-	+	-
staff_supervisory_authority, even, after_employment	-	-	28	28	-	-	-	-

In a first stage, confidence levels (i.e. the confidence level, with which the separate lemmas are considered to be relevant) of 95% and 99% have been combined with the fact whether two ('-++', '+-+' and '++-') or three ('+++') constituting elements of a triple contain relevant vocabulary (see Table 3). E.g., in the scenario A1 (two triple parts have to contain a lemma relevant with a confidence of 95%) 1012 triples are accepted as appropriate ('+'), while 104 are rejected ('-').

Subsequently, the qualitative criterium (does the triple contain a relevant word) has been replaced by a score indicating how many characters of the three triple parts (ex-

Table 3. Some of the evaluation scenarios and the corresponding results.

expertimental settings	label	-	+	+++	++-	+-+	-++
3/3 (expert 1)	E1	712	404				
3/3 (expert 2)	E2	755	361				
95% conf, 2/3	A1	104	1012	637	59	224	92
95% conf, 65% score	X1	346	770				
95% conf, 2/3, 65% score	A2	104	770	594	34	77	65
95% conf, 3/3	B1	479	637	637			
95% conf, 70 %score	X2	582	534				
95% conf, 3/3, 70% score	B3	479	534	534			
99% conf, 2/3	C1	148	968	497	73	267	131
99% conf, 65% score	Y1	464	652				
99% conf, 2/3, 65% score	A2	148	652	445	42	79	86
99% conf, 3/3	D1	619	497	497			
99% conf, 70% score	Y2	720	396				
99% conf, 3/3, 70% score	D3	619	396	396			

pressed as an averaged percentage) are matched by words statistically relevant. E.g., the triple $< rule, establish, by\_national\_competent\_body >$ receives a score of 89 as only 'competent' is not a relevant word with a 95% confi dence level (89 = ((4/4)*100 + (11/11)*100 + (17/25)*100)/3) [4]. Finally, 22 combinations have been tried - see Table 4. E.g., if there is no perfect match for the entire triple (that would equal a 100% score), one could still establish a treshold (e.g., 90%) that on the one hand relaxes the '+++' criterium and on the other hand still rejects matches that are too partial (scenarios B4 and D4). Note in that respect that 66% corresponds to a perfect match of two of the three triple elements.

Table 4. Sensitivity (Se), precision (P), and specificity (Sp) values.

setting	label	Se	P	Sp
95% cf, 2/3	A1	0,955	0,211	0,133
95% cf, 3/3	B1	0,580	0,205	0,455
99% cf, 2/3	C1	0,910	0,212	0,179
99% cf, 3/3	D1	0,464	0,211	0,578
95% cf, 65%	X1	0,723	0,208	0,332
95% cf, 70%	X2	0,50	0,208	0,539
95% cf, 90%	X3	0,383	0,117	0,645
99% cf, 65%	Y1	0,616	0,205	0,421
99% cf, 70%	Y2	0,410	0,212	0,630
99% cf, 90%	Y3	0,321	0,226	0,734

setting	label	Se	P	Sp
95% cf, 2/3, 65%	A2	0,723	0,168	0,136
95% cf, 2/3, 70%	A3	0,508	0,157	0,341
95% cf, 2/3, 90%	A4	0,392	0,226	0,673
99% cf, 2/3, 65%	C2	0,723	0,176	0,179
99% cf, 2/3, 70%	C3	0,410	0,236	0,678
99% cf, 2/3, 90%	C4	0,321	0,263	0,781
95% cf, 3/3, 65%	B2	0,553	0,163	0,311
95% cf, 3/3, 70%	B3	0,508	0,178	0,431
95% cf, 3/3, 90%	B4	0,392	0,164	0,431
99% cf, 3/3, 65%	D2	0,428	0,151	0,416
99% cf, 3/3, 70%	D3	0,410	0,175	0,531
99% cf, 3/3, 90%	D4	0,321	0,142	0,531

We have also calculated the sensitivity or recall (indicates the power of the evaluation procedure to accept true positive facts), specifi city (indicates the power of the evaluation

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

procedure to reject true negative facts) and precision (indicates to which extent the results obtained are correct) for the triples rated in the same way by both experts - see Table 4.

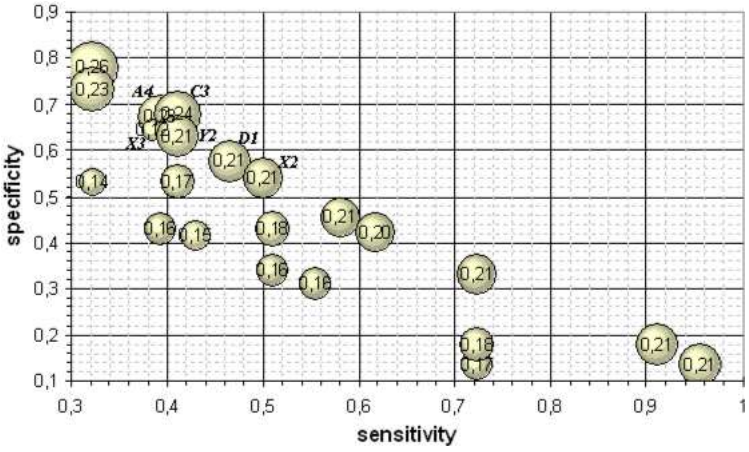

Figure 1. Sensitivity versus specificity and precision.

Figure 1 graphically represents the data of Table 4. The size of the bubbles (each representing a different experimental setting - e.g. X3, A4, C3, Y2, D1 and X2) indicate the precision (value inside the bubble) for a combination of sensitivity (X-axis value) and specifi city (Y-axis value) scores.

5. Discussion

It is to be expected that taking care of synonyms will improve the scores, but it is unclear to what extent. On the other hand, synonym lists or domain specifi c vocabularies with explanatory glosses will not always be available. In those cases, a corpus will be the only source of information (next to human experts). Both the human experts jointly examined their initial evaluation. They stated that many disagreements were due to a difference in understanding of the basis criteria. This could be improved by a clearer statement of evaluation criteria in combination with a test run on a small number of terms where evaluators explain each choice and are given the opportunity to comment on any differences (although to each other to avoid bias). E.g., one evaluator erroneously removed all object terms containing a preposition because he decided they were not strictly nouns, whereas the other did not. This difference in criteria would quickly be detected and removed by the test run and subsequent check. Unlike for classical IE evaluation settings, inter-rater disagreement is less problematic as it reflects a different position regarding the ontology depending on the (type of) stakeholder involved. For the moment, hierarchical and other relationships are not yet assessed. We have to bare in mind that the entire automatic triple evaluation procedure is "blind", i.e. mainly based on term frequencies and string overlap instead of domain knowledge.

5.1. Word Relevance

In a previous experiment [7], we've tested our metrics on the EU VAT Directive. When comparing the machine reference with a human reference (a list of terms produced by human experts), a κ value of 0,758 has been reached, indicating a good agreement. Therefore, we assumed that, in case of the Privacy Directive, a similar result, i.e. the metrics approximate human experts, would be obtained (as we don't dispose yet of a list of privacy terms made by human experts).

5.2. Triple Relevance

Minimally the sensitivity and specificity values should be higher than 0,5. This happens only with scenario X2 (95% confidence and 70% score levels). However, the precision only reaches 0,208. We see that the scoring algorithm mainly enhances the sensitivity. The confidence level has a positive impact on the specificity and a negative one on the sensitivity. 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 triple might be missed in order to have less rubish triples. 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. Therefore, scenario C3 reaching a precision of 0,236 (with of sensitivity and specificity of 0,41 and 0,678 resp.) can be an alternative. We consider X2 and C3 as scenarios that provide *baseline* results as the precision scores are too low.

The experts stated that too many irrelevant results are produced - the text miner not being able to skip over sections that are only of marginal interest for the privacy topic. To remedy this, they have suggested the following improvements:

- The background ("neutral") corpus (here the WSJ) is key in defining what constitutes relevance. 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. This would eliminate a large number of classification discrepancies between experts themselves, and between the experts and the unsupervised miner. 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 higly relevant by the miner, but totally irrelevant by the experts.
- European legislative texts have a uniform structure and therefore lend themselves to a pre-processing stage where irrelevant material can be quickly and repeatably removed. For example, in creating a privacy ontology for user-agents, sections about member states' obligations to inform the European Commission when legislation has been implemented are clearly not relevant. It is suggested that the text be pre-processed according to a well-defined set of criteria (e.g., because our ontology is applied to the modelling of normative rules, we were able to remove the "whereas sentences" in the preprocessing step, using them only for later grounding of term semantics) before being processed by the text miner. This will significantly reduce the number of irrelevant triples.

- An issue not addressed by the above process 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 synomym dictionary would go some way towards term abstraction although it can only take account of equivalence and not subclass relationships between terms.

6. Related Work

Reports on previous experiments contain additional details on the unsupervised miner [5]. The method and previous quantitative experiments have been presented in [7]. To the best of our knowledge, so far only one other approach addresses the quantitative and automated evaluation of an ontology by referring to its source corpus [10]. Others have evaluated methods and metrics to select the most appropriate terms (e.g, [11]) for building an ontology. However, they don't evaluate entire triples. Various researchers are working on methods to evaluate results of ontology learning and population [1,3].

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 [10]. Unfortunately, no concrete results or test case are presented. Next to that, there is the work of *Sabou* who, in her latest work, tries 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 well ours. She has "established a one-to-one correspondence between phrases in the corpus and derived concepts [12], so that our lexicometric score are comparable to her ontology ratios.

7. Future Work

The core topic concerns research on the goodness of fit of the evaluation procedure as an approximation of the human expert behaviour. This necessitates the involvement of minimally two human experts to account for inter-rater (dis)agreement. Some axes for future work emanate from the discussion section:

- Study alternative term selection methods available from the domains of lexico- and terminography, quantitative linguistics or library information science - e.g., the TF/IDF and domain relevance and consensus metrics. [5] - as well as alternative heuristics to asses entire triples.
- Study the "move" towards the conceptual level, which necessitates the integration of semantic distance measures such as the WordNet similarity functions [13]. Brewster et al. add two levels of WordNet hypernyms [10, p.166] for that purpose. That implies that (novel) compound terms should be assigned a semantic interpretation as is done e.g., by Navigli and Velardi [11].

[5] Up till now, it was pointless to use these as the corpus consists of one (resaonably large) document.

8. Conclusion

We have presented a validation effort for an easily to apply automatic evaluation procedure for triples as material for a privacy ontology to be created. Compared to a human reference, the automatic evaluation procedure is able in more or less half of the cases to successfully accept an appropriate triple and reject an irrelevant triple. The precision score however is too low. Nevertheless, the automatic evaluation procedure is considered practically useful by the human experts to speed up the ontology creation process. The current outcomes can be considered as a baseline reference for further experiments.

Acknowledgements

This research has been financed by the Flemish OntoBasis project (IWT GBOU 2001 #10069) and the EU FP6 IP PRIME (IST 2002-507591). It represents the view of the authors only. We are particularly indebted to dr. Marie-Laure Reinberger (Universiteit Antwerpen - CNTS) who has produced the pricavy triples and to drs. John Borking (Borking Consultancy, The Netherlands), who was the second domain expert.

References

[1] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press, Amsterdam, 2005.

[2] M. Shamsfard and A. Barforoush. The state of the art in ontology learning: a framework for comparison. *Knowledge Engineering Review*, 18(4):293 – 316, 2003.

[3] J. Hartmann, P. Spyns, D. Maynard, R. Cuel, M. Carmen Suarez de Figueroa, and Y. Sure. Methods for ontology evaluation. KnowledgeWeb Deliverable #D1.2.3, 2005.

[4] Sabine Buchholz, Jorn Veenstra, and Walter Daelemans. Cascaded grammatical relation assignment. In *Proceedings of EMNLP/VLC-99*. PrintPartners Ipskamp, 1999.

[5] Marie-Laure Reinberger and Peter Spyns. *Ontology Learning from Text: Methods, Applications and Evaluation*, chapter Unsupervised Text Mining for the Learning of DOGMA-inspired Ontologies. IOS Press, Amsterdam, 2005.

[6] Thompson Automation Software, Jefferson OR, US. *Tawk Compiler*, v.5 edition.

[7] P. Spyns and M.-L. Reinberger. Lexically evaluating ontology triples automatically generated from text. In A. Gómez-Pérez and J. Euzenat, editors, *Proceedings of the second European Semantic Web Conference*, volume 3532 of *LNCS*, pages 563 – 577. Springer Verlag, 2005.

[8] George K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, 1949.

[9] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159 – 195, 1958.

[10] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data driven ontology evaluation. In N. Shadbolt and K. O'Hara, editors, *Advanced Knowledge Technologies: selected papers*, pages 164 – 168. AKT, 2004.

[11] Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2):151 – 179, 2004.

[12] Marta Sabou, Chris Wroe, Carole Goble, and Gilad Mishne. Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In *Proceedings of the 14th International World Wide Web Conference*, 2005.

[13] 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.