Methodology for automatic evaluation of restricted domain ontologies *

Mireya Tovar[1,2], Azucena Montes[1,3], and David Pinto[2]

[1]Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Mexico
[2] Faculty Computer Science, Benemérita Universidad Autónoma de Puebla, Mexico,
[3]Engineering Institute, Universidad Nacional Autónoma de Mexico.
{mtovar,amontes}@cenidet.edu.mx
{dpinto,mtovar}@cs.buap.mx

Abstract. In this paper we present advances of the PhD research thesis entitled: "Automatic Evaluation of Restricted-Domain Ontologies". We discuss the methodology employed, so as the results obtained up to now. The evaluation has been carried out over two main components of the ontology: concepts and relationships. Thus, on the one hand we present methods for discovering and validating concepts. On the other hand, we show results when the degree of semantic similarity among concepts is computed over the relationships that already occur in the ontology to be evaluated.

Key words: Ontology evaluation, Natural Language Processing, lexical-syntactic patterns.

1 Introduction

In recent years, especially with the emergence of the concept of the Semantic Web, it is understood that there is a great interest in the management of ontological resources with the aim of help to comply the user information needs. The World Wide Web is one of the largest public repositories of information, but most of that information is designed for human consumption, thereby, it is almost innaccessible for machines because it is not properly organized, thus making difficult for software applications to use the WWW in an automatic way [1]. When semantic is used to organize or structure the information at the Web, this data receives the name of Semantic Web. Organization of knowledge in the Semantic Web is usually performed by means of ontologies. An ontology uses a predefined, reserved vocabulary of terms to define concepts and the relationships between them for one specific area of interest, or domain [1]. Gruber [2] defines an ontology as: "an explicit specification of a conceptualiation".

An ontology includes classes, instances, attributes, relationships, constraints, rules, events and axioms. In many cases, ontologies are structured as hierarchies

* This work is partially supported by CONACYT and PROMEP under grants: CONACYT 54371, PROMEP/103.5/12/4962 BUAP-792 and project CONACYT 106625.

Special Issue: Advances in Pattern Recognition
Research in Computing Science 61, 2013, pp. 63-72
Paper Received 28-02-2013 and Accepted 22-04-2013

of concepts modeled either by means of part-whole or class-inclusion semantic relationships. There exist, other types of semantic relationships that are not hierarchical such as synonyms, antonyms, etc.

Although there is sufficient research on methodologies, techniques, tools and software for building ontologies, an aspect that has not been considered in depth is the evaluation of ontologies. One of the reasons of this phenomenon is the difficulty for determining which items should be evaluated and what criteria should be considered to specify the quality of the ontology [3]. In this research work we aim to develop an automatic method for evaluating restricted-domain ontologies, employing Natural Language Processing (NLP) techniques. The methodology proposed assumes that one ontology has been automatically, semi-automatically or manually constructed. Therefore, we aim to "validate" the quality of the elements of the ontology, such as relationships and concepts. The evaluation is carried out twofold, 1) By using a reference corpus (document collection) of the same ontology domain, and, 2) by using self-evaluation, considering only the information stored in the ontology itself. Eventhough, two evaluations methods are considered in the PhD thesis, in this report, we only present results with respect to the first proposal.

The remaining of this paper is structured as follows. Section 2 describes with more detail the problem we are dealing with. The methodology proposed for solving this problem is presented in Section 3. A description of the contribution expected in this research work is presented in Section 4. The results obtained up to now for the evaluation of concepts and relationships is given in Section 5. Finally, in Section 6, the conclusions of the advances of this PhD thesis are given.

2 Research problem to solve

As previously stated, the problem we are intended to deal with is the automatic evaluation ontologies. We assume the concept of evaluation, in our context, to be associated with the process of determining the quality of such ontologies. According to literature, evaluation of ontologies can be performed in one or two of the following stages: a) During the construction of the ontology, and b) Once the ontology has been constructed.

If the ontology is small (for instance, with small number of concepts and relationships), the first approach is practical for verifying the quality of it. However, when the ontology contains a large number of components, the time it would take for a human expert for evaluating it could be very expensive in terms of time. Thus, the proposal of novels methods for automatic evaluation of ontologies (independently of the construction stage) would be of high benefit. The aim is to provide a framework which allows an extra tool to the ontological engineer for constructing better knowledge databases than when none evaluation tools is used.

There are various ontology evaluation approaches reportes in literature. Some of these approach descriptions follows: 1) Based on criteria ([4], [5], [6], [7], [8]);

2) Based on gold standard ([9], [10], [11]), corpus-based or data ([3], [12], [13], [14]); 3) Based in tasks ([15]) or application ([16]).

In particular, our proposal intend to use the criteria and corpus based approaches. One of the criteria we consider very important for the development of this research work is the correctness criterion. According to [5], **Correctness** specifies whether or not the information stored in the ontology is true, independently of the domain of interest.

In summary, we propose the automatic evaluation of domain ontologies, considering content level evaluation in the domain corpus, based in the correctness criterio, i.e., to verify whether or not the concepts and relationships of a domain ontology are true. For this purpose, we use boolean scores: 1 if the concept set or relationship set is correct in the domain corpus, or 0 otherwise. Additionally, we evaluate the semantic similarity degree of the ontology relationships using score values normalized between 0 and 1, such as Jaccard coefficient, Dice, etc.

3 Methodology

The methodological solution for this research works considers evaluate any type of domain ontology, assuming that exists a reference corpus. One of the initial conditions for the evaluation of a given ontology is that is should be well designed, structurally speaking, and that the reference corpus corresponds to the same domain of the ontology. Even if, this corpus can contain any domain-specific texts, i. e., scientific publications, project reports, books, medical notes, etc., it is expected that it sufficient diverse, i.e., to guarantee that there exist a reasonable amount of text [17] for executing statistical methods for searching evidence of the ontology components to be evaluated.

The proposed methodology for evaluation ontologies proposes three phases: a) Information filtering, b) Discovery of candidate terms and candidate relationships, and c) Evaluation of the ontology components, which are described in the following sub-sections.

3.1 Information filtering

Even if, there exist a reference corpus, it is normally made up of a huge amount of documents which need to be filtered in order to obtain the most specific features to be used in the process of evaluating the components of the ontology. The proposed method considers to find the most suitable information for validating each triple of the ontology. In this sense, the architecture considers the following three sub-modules:

Extraction of triples In this sub-module, we consider the theory of OWL[1], and we implement two algorithms in Jena[2] for extracting classes (or concepts) and relationships of restricted-domain ontologies.

[1] `http://www.w3.org/TR/owl-features/`
[2] `http://jena.apache.org/`

Query construction From the information extracted from the ontology, we construct queries that will be used for searching evidence in the reference corpus. This is perhaps one of the most critical steps of the methodology. Finding evidence of the triple quality in the reference corpus is actually the aim of this research work. Given a triple (S, R, O), with S the subject, R the relationship and O the object, the queries are constructed as: $S * O$ (documents containing S near of O), $*S*$ (documents containing S) and $*O*$ (documents containing O).

Information Retrieval System The third sub-module uses de queries constructed as input in an information retrieval system for finding information associated to the triples. The purpose of this phase is to filter out documents of the domain corpus containing terms of the ontology.

3.2 Discovery of candidate terms and relationships

The aim of this PhD thesis is to evaluate the quality of the ontology components. Thus, we might find or discover terms and relationships in the reference corpus in order to "validate" the terms and relationships already stored in the ontology. This process of discovering is explained as follows.

Discovery of candidate terms Up to now, we use a pattern-based approach for discovering candidate terms in the reference corpus, i.e. we find those terms that match one or more of the lexical-syntactic patterns already discovered in the ontology.

For discovering the candidate terms in the reference corpus, which may or not correspond to the concepts of the domain ontology, we implemented the following procedure:

1. To apply a Part-of-Speech (PoS) tagger (Tree Tagger[3]) to the concepts of the ontology to be evaluated [18].
2. To cluster similar concepts by using the morphological PoS tags.
3. To create morphological and/or lexical-syntactic patterns from the concept clusters.
4. To use the lexical-syntactic patters in order to extract candidate terms from the reference corpus.

The above procedure produces a list of candidate terms that will be further used for determining whether or not a concept should be present in the ontology, based on the evidence that exist in the reference corpus.

Discovery of candidate relationships We have also planned to identify lexical-syntactic patterns for identifying relationships of the ontology to be evaluated. For this purpose, we also need to find evidence in the reference corpus

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

about a given ontology relationship. We have considered to use a hybrid approach for the discoverying of the candidate relationships in the corpus, considering a pattern based approach together with a pure statistical approach. The former approach has not beed developed yet, however, the latter is considered in this paper. We use similarity measures for determining the degree of association between a pair of concepts. In other words, we are now only evaluating the quality of the relationship by measuring how related these two concepts are.

For this purpose, we identify the concepts associated to the relationship of the ontology and we construct frecuency vectors using the vocabulary of the domain corpus for each concept. We apply the similarity measure directly to the vectors and we determine the degree of term correlation associated to each relationship of the ontology. The hypothesis follows: "the most similar the concept vectors are, the better the quality of the relationship".

3.3 Evaluation measures

The third phase of the solution architecture is based on the evaluation of ontology triplets, using candidate terms and relationships generated in the automatic discovery phase.

Concept evaluation measures We have evaluated the performance of the presented approach by means of standard evaluation measures such as precision, recall and F-measure. The precision is the proportion of the candidate terms which truly are concepts in the ontology among all those which were identificated as candidate terms. The recall is the proportion of candidate terms which were identificated as concepts in the ontology, among all the real concepts of the ontology. The F-measure is the harmonic mean of precision and recall. These measures are defined as follows:

$$Precision = \frac{CO\_CexR}{reExCor} \tag{1}$$

$$Recall = \frac{CO\_CexR}{COnt} \tag{2}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Where:

CO_CexR	Number of multi-word terms extracted by the lexical-syntactic patterns that overlap between the ontology and the corpus.
$reExCor$	Total number of candidate terms extracted from the corpus using the lexical-syntactic patterns.
$COnt$	Total number of concepts of the ontology.

Evaluation of semantic relationships The similarity measures considered so far for evaluating the degree of relationship for a pair of concepts (x, y) are: cosine ($d(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t\boldsymbol{y}}{\|x\|\|y\|}$), Tanimoto or Jaccard ($d(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t\boldsymbol{y}}{\boldsymbol{x}^t\boldsymbol{x}+\boldsymbol{y}^t\boldsymbol{y}-\boldsymbol{x}^t\boldsymbol{y}}$), Sockal ($d(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t\boldsymbol{y}}{2\boldsymbol{x}^t\boldsymbol{x}+2\boldsymbol{y}^t\boldsymbol{y}-3\boldsymbol{x}^t\boldsymbol{y}}$), dice coefficient ($d(\boldsymbol{x}, \boldsymbol{y}) = 2\frac{\boldsymbol{x}^t\boldsymbol{y}}{\boldsymbol{x}^t\boldsymbol{x}+\boldsymbol{y}^t\boldsymbol{y}}$), and a variation of the previous measures (we have named it SimVar- $d(\boldsymbol{x}, \boldsymbol{y}) = 3\frac{\boldsymbol{x}^t\boldsymbol{y}}{\boldsymbol{x}^t\boldsymbol{x}+\boldsymbol{y}^t\boldsymbol{y}+\boldsymbol{x}^t\boldsymbol{y}}$) [19]. In order to be consistent in terms of the average results, we have reported random variables instead of similarity values.

Thus, the evaluation measures are summarized in terms of the following random variables: Mean ($\bar{\mu}$), Standard deviation (σ) and Coefficients of variation ($CV = \frac{\sigma}{\bar{\mu}}$).

Standard deviation shows the variation or dispersion of the data with respect to the mean. A low standard deviation indicates that the data tend to be close to the mean, a high standard deviation indicates that the data has high variation or dispersion. It is expressed in the same units as the data. To compare results between different data sets, regardless of the units, the coefficients of variation are used.

4 Main contribution

This proposal aims to evaluate the ontology independently of the construction phase, i. e. when the ontology is in development or when it is already constructed. Thus, providing the end user or the engineer for an ontological evaluation of the ontology in those cases when the knowledge database is too large. In this research proposal we have considered only restricted-domain ontologies, assuming that there exist a reference corpus for the evaluation process.

The main contribution of this proposal is the introduction of automatic methods for validating ontologies automatically, semi-automatically or even manually constructed. We provide mechanisms for discovering candidate terms and relationships from a reference corpus which may be further used for validating those terms and relationships that already exist in the ontology to be evaluated.

5 Results achieved and their validity

In this section we present the results obtained with the proposed approach.

5.1 Dataset

Table 1 presents two characteristics of the two ontologies already evaluated in this research paper. The number of concepts and "hierarchical" [4] relationships of the two ontologies.

[4] In this paper we have considered only those relationships extracted by means of the "subClassOf" axiom of OWL.

Table 1. Domain ontologies.

Domain	owl file	Total of concepts	Hierarchical relationships
Petroleum	petroleum.owl	48	37
Artificial Intelligence	ai.owl	276	205

Table 2 shows the number of documents processed from the reference corpus, together with the total of tokens analized. In that table, it can be seen a new corpus which results from filter the petroleum corpus by using the information filtering techniques described above.

In the case of the petroleum domain, the number of documents containing the terms of the concepts of the domain ontology are 575. For the case of Artificial Intelligence corpus, the filtering step resulted in a subcorpus exactly equal to the total of documents of the corpus, thus we have not added another row.

Table 2. Corpora to be evaluated.

Corpus	Documents	Tokens
Petroleum domain	577	9,730,495
Petroleum Filtered Subcorpus	575	9,727,092
Artificial Intelligence	8	10,805

5.2 Results for the evaluation of concepts

Tables 3 and 4 show the lexical-syntactic patterns obtained by the procedure presented in Section 3.2, and the results obtained by applying these patterns to the reference corpus for the Petroleum and Artificial Intelligence domain, respectively. The first column indicates the number of patterns identified in the ontology, column two indicates the frequency of the pattern, column three is the pattern identified in the ontology, finally, the last column shows the sum of term frequencies availables in the corpus. Only those terms that fulfill with the corresponding lexical-syntactic pattern. The last row indicates the number of candidate terms extracted from the reference corpus without repetition.

Consider the following tags for the petroleum ontology: IN is a preposition, RB is an adverb, NN is a Noun, JJ is an adjetive, VB is a verb in participle past, or verb gerund (VBP or VBG). For the Artificial Intelligence ontology, the tags follows: NN is a noun, proper noun, plural noun (NN, NP or NNS), JJ is an adjective or superlative adjective (JJ or JJS), CD is a cardinal number, IN is a preposition, VB is a verb in any form (VBG, VBZ, VB, VBN, VBP or VBD), FW is a foreign word and RB is an adverb.

Once the candidate terms were extracted from the reference corpus using the lexical-syntactic patterns, we match them with respect to the concepts of the ontology. We noted that from the 48 concepts defined in the Petroleum ontology,

Table 3. Lexical-syntactic patterns for the extraction of candidate terms in the Petroleum domain corpus.

n	Fr Ont	Lexical-syntactic patterns	Fr corpus
1	21	NN^+ $JJ?$	1,823,294
2	11	NN $VB(JJ\|(NN^+)?)$	125,308
3	11	JJ $(NN^+)?$	646,029
4	5	RB $(VB?\|NN?)$	223,301
5	4	VB $(NN\|JJ)$	71,358
6	1	IN NN^+	192,879
		Total of unrepeated terms:	378,465

Table 4. Lexical-syntactic patterns for the extraction of candidate terms in the Artificial Intelligence domain corpus.

n	Fr Ont	Lexical-syntactic patterns	Fr corpus
1	243	$(NN^+)((VB\ NN)?\|(VB^+)?)$	2693
2	84	$(JJ)^+(NN^+)?$	1000
3	35	$NN\ ((JJ\ NN)\|(IN)(JJ)(NN)\|(CD)\|(IN\ VB)\|(VB\ JJ\ NN)\|(IN\ NN^+))$	400
4	17	$(VB^+)(NN^+)?$	1582
5	10	$(JJ\ NN)((IN\ JJ\ NN?)\|(VB\ NN?)\|(JJ\ NN))$	135
6	6	$JJ\ ((NN\ IN\ JJ\ NN?)\|(VB\ NN?))$	43
7	3	$RB\ ((VB\ JJ\ NN)\|(JJ\ NN))$	27
8	2	$(VB\ IN)(NN\|(JJ\ NN))$	105
9	1	$IN\ JJ\ NN$	151
10	1	$FW\ NN\ IN\ JJ\ NN$	0
		Total of unrepeated terms:	3592

the system was able to find 42 concepts. For the case of the Artificial Intelligence domain, the system found 205 of 276 concepts. The results of precision, recall and F-measure for the Petroleum and Artificial Intelligence domains are shown in Table 5.

Table 5. Results of evaluation measures applied to concepts.

Ontology	Recall	Precision	F-measure
Petroleum	0.875000	0.00011098	0.00022192
Artificial Intelligence	0.742754	0.05707130	0.10599799

5.3 Results for the evaluation of semantic relationships

The similarity measures presented in Section 3.2 were applied to the hierarchical relationships of the ontologies to be evaluated. Table 6 shows the results

obtained. For the Petroleum ontology (with 37 hierarchical relationships), the mean similarity measure were very close to 1.0, indicating a high correlationship between concepts sharing a relationship in the ontology. The $\bar{\mu}$ value indicates that exist low variation for each relationship identified. The CV value shows a 5% of dispersion among the five similarity measures. For the case of correctness, we use an score of 0.90 for indicating that the relationship is true in the domain corpus. We observed that we have obtained more than 31 relationships of the Petroleum ontology. In the Artificial Intelligence ontology, the $\bar{\mu}$ is quite variable for the five similarity measures, though we always obtained a value greater than 40% of similarity. In the case of σ and CV, we observed that the best result was when the coseno similarity was used. By using correctness (score of 0.90), we obtained 87 of 205 hierarchical relationships of the ontology.

Table 6. Results of the evaluation measures for the hierarchical relationships.

	Petroleum				Artificial Intelligence			
Similarity	$\bar{\mu}$	σ	CV	$Correctness$	$\bar{\mu}$	σ	CV	$Correctness$
Cosine	0.9994	0.000444	0.0004	37/37	0.8185	0.162945	0.199	87/205
Jaccard	0.9754	0.028671	0.0293	36/37	0.5835	0.310579	0.532	54/205
Sokal	0.9535	0.051713	0.0542	31/37	0.4841	0.335557	0.693	43/205
SimVar	0.9915	0.010339	0.0104	37/37	0.7405	0.244661	0.330	77/205
Dice	0.9874	0.015193	0.0153	37/37	0.6846	0.272201	0.397	64/205

6 Conclusions

In this paper we present advances of the PhD thesis that tackles the problem of automatic evaluation of restricted-domain ontologies. Evaluating the quality of automatic, semi-automatic or manually constructed ontologies is of high importance and high challenging. Two different ontologies were evaluated showing encouraging results. In particular, a method for the automatic construction of lexical-syntactic patterns was presented with the aim of discovering candidate concepts and relationships which may be further used for validating concepts and relationships of the ontology to be evaluated. Still, a number of experiments to be carried out, however, we considering important to present the current results obtained in the framework of automatic evaluation of ontologies.

References

1. Hebeler, J., Fisher, M., Blace, R., Perez-Lopez, A., Dean, M.: Semantic Web Programming. Wiley (2011)
2. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL-93-04, Knowledge Systems Laboratory, USA (1993)

3. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of International Conference on Language Resources and Evaluation. (2004)
4. Staab, S., Studer, R., eds.: Ontology Evaluation. International Handbooks on Information Systems. Springer (2004) Chapter 13: Gómez- Pérez, Asunción.
5. Cantador, I., Ferández, M., Castells, P.: A collaborative recommendation framework for ontology evaluation and reuse. In: Actas de International Workshop on Recommender Systems, en la 17th European Conference on Artificial Intelligence (ECAI 2006), Riva del Garda, Italia. (2006) 67–71
6. Sleeman, D., Reul, Q.: CleanONTO: Evaluating Taxonomic Relationships in Ontologies. In Vrandecic, D., Mari, Gangemi, A., Sure, Y., eds.: Proceedings of 4th International EON Workshop on Evaluation of Ontologies for the Web, Edinburgh, Scotland (2006)
7. García-Ramos, S., Otero, A., Fernández-López, M.: Ontologytest: A tool to evaluate ontologies through tests defined by the user. In: Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. IWANN '09, Berlin, Heidelberg, Springer-Verlag (2009) 91–98
8. Völker, J., Vrandečić, D., Sure, Y., Hotho, A.: Aeon - an approach to the automatic evaluation of ontologies. Applied Ontology **3** (January 2008) 41–62
9. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Proceedings of European Knoeledge Ackquisition Workshop (EKAW). (2002)
10. Spyns, P., Reinberger, M.L.: Lexically evaluating ontology triples generated automatically from texts. In Gómez-Pérez, A., Euzenat, J., eds.: ESWC. Volume 3532 of Lecture Notes in Computer Science., Springer (2005) 563–577
11. Brank, J., Mladenić, D., Grobelnik, M.: Gold standard based ontology evaluation using instance assignment. In: Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006). (2006)
12. Netzer, Y., Gabay, D., Adler, M., Goldberg, Y., Elhadad, M.: Ontology Evaluation through Text Classification. Springer-Verlag, Berlin, Heidelberg (2009)
13. Murdock, J., Buckner, C., Allen, C.: Two methods for evaluating dynamic ontologies. In: Proceedings of the 2nd International Conference on Knowledge Engineering and Ontology Development (KEOD) Valencia, Spain. (2010)
14. Yao, L., Divoli, A., Mayzus, I., James, E.A., Rzhetsky, A.: Benchmarking ontologies: Bigger or better? PLoS Computational Biology **7**(1) (01 2011) 1–15
15. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Ontology evaluation and validation - an integrated formal model for the quality diagnostic task. Technical report, LOA , ISTC-CNR (2005)
16. Salem, S., AbdelRahman, S.: A multiple-domain ontology builder. In: Proceedings of the 23rd International Conference on Computational Linguistics. COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 967–975
17. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: Proceedings of the 3rd European Semantic Web Conference (ESWC2006), number 4011 in LNCS, Budva, Springer (2006)
18. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)
19. Choi, S.S., Cha, S.H., Tappert, C.: A Survey of Binary Similarity and Distance Measures. Journal on Systemics, Cybernetics and Informatics **8**(1) (2010) 43–48