An Approach Based in LSA for Evaluation of Ontological Relations on Domain Corpora

Mireya Tovar[1](✉), David Pinto[1], Azucena Montes[2], and Gabriel González[3]

[1] Faculty of Computer Science,
Benemérita Universidad Autónoma de Puebla, Puebla, Mexico
{mtovar,dpinto}@cs.buap.mx
[2] TecNM, Instituto Tecnológico de Tlalpan, Mexico City, Mexico
amr@cenidet.edu.mx
[3] Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET),
Cuernavaca, Mexico
gabriel@cenidet.edu.mx

Abstract. In this paper we present an approach for the automatic evaluation of relations in ontologies of restricted domain. We use the evidence found in a corpus associated to the same domain of the ontology for determining the validity of the ontological relations. Our approach employs Latent Semantic Analysis, a technique based on the principle that the words in a same context tend to have semantic relationships. The approach uses two variants for evaluating the semantic relations and concepts of the target ontologies. The performance obtained was about 70% for class-inclusion relations and 78% for non-taxonomic relations.

Keywords: Ontology evaluation · Latent semantic analysis · Natural language processing

1 Introduction

The continuous increase in the number of documents produced on the Web makes it more complex and costly to analyze, categorize and retrieve documents without considering the semantics of each document. One way to represent knowledge of documents is through ontologies.

An ontology, from the computer science perspective, is "an explicit specification of a conceptualization" [1].

Ontologies can be divided into four main categories, according to their generalization levels: generic ontologies, representation ontologies, domain ontologies, and application ontologies. Domain ontologies, or ontologies of restricted domain, specify the knowledge for a particular type of domain, for example: medical, tourism, finance, artificial intelligence, etc. An ontology typically includes

This work is partially supported by the Sectoral Research Fund for Education with the CONACyT project 257357, by PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, and VIEP-BUAP project 00478.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2017, LNCS 10267, pp. 225–233, 2017.
DOI: 10.1007/978-3-319-59226-8_22

the following components: classes, instances, attributes, relations, constraints, rules, events and axioms.

The ontologies are resources that allow to capture the explicit knowledge in the data, through concepts and relationships. In this paper we are interested in the process of discovering and evaluating ontological relations, thus, we focus our attention on the following two types: taxonomic relations and/or non-taxonomic relations. The first type of relations are normally referred as relations of the type "is-a" (hypernym/hyponymy or subsumption) or class-inclusion.

In order to evaluate concepts and semantic relations of three domain ontologies using Latent Semantic Analysis, in this research work we present two variants, the first one based on the cosine similarity, and second one based on clustering by committee.

The experiments carried out and the obtained results are discussed through the remaining of this paper, which is organized as follows: in Sect. 2 we present the related work, in Sect. 3 we present the concept of latent semantic analysis, whereas in Sect. 4 we describe the concept of clustering by committee, both employed in this research work. The method proposed is presented in Sect. 5. The experimental results are shown and discussed in Sect. 6. Finally, in Sect. 7 the conclusions of the work are given.

2 Related Work

Different approaches employing LSA for task related with ontologies can be found in literature. For example, in [2] it is presented an automatic method for ontology construction using latent semantic, clustering and Wordnet over a collection of documents.

In [3] they show methods for improving both, the recall and the precision of automatic methods for extraction of hyponymy (IS-A) relations from raw text. By applying latent semantic analysis (LSA) to filter extracted hyponymy relations, they reduce the error rate of their initial pattern-based hyponymy extraction by 30%, achieving precision of 58%. By applying a graph-based model of noun-noun similarity learned automatically from coordination patterns to previously extracted correct hyponymy relations, they achieve roughly a five-fold increase in the number of correct hyponymy relations extracted.

In [4], the authors describe an approach that extracts hypernym and meronym relations between proper nouns in sentences of a given text. Their approach is based on the analysis of the paths between noun pairs in the dependency parse trees of the sentences.

In [5] techniques of machine learning and statistical natural language processing are used to attempt to construct a domain concept taxonomy. They employ different evaluation measures such as: Precision, Recall, F-measure, and others. Their work focused on the integration of knowledge acquisition with machine learning techniques for the ontology creation.

We purpose it is evaluate semantic relationships with evidence in the domain corpus through of latent semantic analysis method. For the evaluation, we use the mesure of accuracy.

3 LSA

Latent Semantic Analysis (LSA) is a computational model used in natural language processing, considered in its beginnings as a method for representing knowledge [6]. LSA is considered an unsupervised dimensionality reduction tool, such as principle component analysis (PCA) [7]. The rationale behind this model indicates that words in the same semantic field tend to appear together or in similar contexts [8,9].

LSA has its origin in an information retrieval technique called Latent Semantic Indexing (LSI) whose purpose is to reduce the size of an array of document terms using a linear algebra technique called Singular Value Decomposition (SVD). The difference with LSA is that it uses a word-context matrix. The context can be a word, a sentence, a paragraph, a document, a test, etc.

Venegas [6] considers that LSA is characterized for being a mathematical-statistical technique that allows the creation of multidimensional vectors for the semantic analysis of the relationships that exist among the different contexts.

The purpose of dimensionality reduction in LSA is to eliminate noise present in the relationships between terms and contexts, since it is usually possible to express the same concept with different terms.

LSA does not consider the linguistic structure of contexts, but the frequency and co-occurrence of terms. However, it has been possible in some cases to identify semantic relationships such as synonymy using LSA [8].

This technique is based on the principle that the words in a same context tend to have semantic relationships, and consequently, indexing of documents with similar contexts should be included by the words that appear in similar contexts even if the document does not contain that words.

4 Clustering By Committee

The Clustering By Committee algorithm (CBC) allows automatic discovery of concepts from text [10,11]. Initially it discovers a set of strict groups called committees that are scattered in the space of similarity. The feature vector that represents a group is the centroid of the committee members, and the clustering method proceed to assign elements to their most similar groups.

The CBC algorithm consists of three phases:

1. To find the most similar elements. In order to calculate the most similar words of a word w, first the characteristics of the word w are ranked according to their mutual information with w.
2. To discover the committees. Each committee that is discovered in this phase defines one of the final groups for the output of the algorithm.
3. To assign elements to the groups. Each element is assigned to the group containing the most similar committee.

CBC has also been used to find the meanings of a word w [12] (algorithm in its flexible version), and for clustering texts (algorithm in its strong version) [13].

Other authors, such as Chatterjee and Mohan [14], have successfully used this algorithm in its flexible version for the discovery of word meanings, including *Random Indexing* to reduce the dimensionality of the context matrix.

5 The Proposed Approach

The proposed approach uses the method of latent semantic analysis, with the purpose of identifying the semantic relationships between the concepts existing in the ontology and looking for evidence in the domain corpus for further evaluation.

LSA points out that words in the same semantic field tend to appear together or in similar contexts, therefore, we considered that the concepts that are semantically related can be in the same sentence or in different sentences sharing information in common.

Based on this assumption, we present the following algorithm that takes into account two variants: (a) Cosine similarity and (b) Grouping by committees (CBC) that assign a weight w to each evaluated relation of the domain ontology. The algorithm performs the following steps:

1. *Pre-processing the domain corpus and domain ontologies.* The domain corpus is divided into sentences and the empty words (such as prepositions, articles, etc.) are removed. The Porter stemming algorithm is applied to the words contained in these sentences [15]. The concepts are also extracted from the ontology[1]. The same process is applied to each one of the concepts of the ontology in order to maintain consistency in the terminology representation (empty words elimination and the Porter stemming algorithm).
2. Application of the LSA algorithm to reduce the dimensionality of the context matrix. In this case, we use the S-Space[2] package and the LSA algorithm[3]. The algorithm receives as parameters the sentences of the corpus of Domain and K dimensions (we use 300 dimensions). The output of the LSA algorithm are semantic vectors of dimension K for each word identified by LSA in the corpus.
3. Construction of concepts. The words obtained by the LSA method are clustered by using the cosine similarity to form the concepts of the ontology.
4. Dimension reduction of vocabulary (vectors) in the LSA matrix. Only the concepts obtained in the previous step are kept in the next step, the rest of the words of the original matrix are removed.
5. Application of variants. At this point two variants are used: cosine similarity for each relation and CBC algorithm to cluster concepts.
   – Similarity cosine

[1] We used Jena for extracting concepts and semantic relations (http://jena.apache.org/).

[2] https://github.com/fozziethebeat/S-Space.

[3] http://code.google.com/p/airhead-research/wiki/LatentSemanticAnalysis.

(a) Calculation of cosine similarity. The concepts obtained in the previous step are used to determine the degree of similarity between each pair of concepts that form the class-inclusion and non-taxonomic relations.

(b) Calculation of threshold u and weight w assigned to the relation. The threshold u is calculated as the sum of the similarities between the total of relationships divided by 2. If the value of the degree of similarity of the relation is greater than the threshold u, the relation takes the weight of $w = 1$, otherwise $w = 0$.

- CBC Algorithm

(a) Application of the CBC algorithm in its flexible version. The concepts formed by similarity, in the previous step, are the input to the CBC algorithm. The output of the algorithm are the clustered concepts.

(b) Identification of the concepts that form the relationship in the clusters generated by CBC. If the pair of concepts that form the relation (class-inclusion and non-taxonomic) are in the cluster, the relation takes the weight $w = 1$ otherwise it receives the weight $w = 0$.

6. Ontology evaluation. We used the metric of accuracy for evaluating the concepts and semantic relations obtained with our approach for each input domain ontology.

The next section, we present the obtained results with this approach.

6 Experimental Results

Below, we present the dataset and the results obtained with the aforementioned approach.

6.1 Dataset

The domains used in the experiments are Artificial Intelligence (AI)[4], standard e-learning SCORM [16] and OIL taxonomy.

In Table 1 we present the number of concepts (C), class-inclusion relations (S) and non-taxonomic relations (R) of the ontology evaluated. The characteristics of its reference corpus are also given in the same Table: number of documents (D), number of tokens (T), vocabulary dimensionality (V), and the number of sentences (O)

6.2 Results

The number of vectors or words retrieved by the LSA algorithm from the domain corpus are shown in Table 2. After concepts discovering by employing the cosine similarity, the approach reduces the matrix to the total of concepts of the domain ontology. Por example, from 1,659 words obtained by LSA for the ontology IA, the matrix is reduced to 276 concepts included in the ontology (see Table 1).

[4] The ontology together with its reference corpus can be downloaded from http://azouaq.athabascau.ca/goldstandards.htm.

Table 1. Datasets

Domain	Ontology			Corpora			
	C	S	R	D	O	T	V
AI	276	205	61	8	475	11,370	1,510
SCORM	1,461	1,038	759	36	1,621	34,497	1,325
OIL	48	37	–	577	546,118	10,290,107	168,554

Table 2. Vocabulary obtained by the LSA algorithm for each domain

Domain	LSA_V
AI	1,659
SCORM	1,473
OIL	168,762

The LSA method with cosine similarity obtained favorable results for the three domains ontologies evaluated, finding more than 70% of the class-inclusion relations (see Table 3). The CBC method obtained the best results in the OIL ontology with 54% of accuracy.

Table 3. Experimental results of the LSA approach to class-inclusion relation in each domain ontology

Ontology	Total	Variant	Enc	Accuracy
AI	205	LSA-cosine	179	87.32%
		LSA-CBC	32	15.61%
SCORM	1038	LSA-cosine	908	87.48%
		LSA-CBC	194	18.69%
OIL	37	LSA-cosine	26	70.27%
		LSA-CBC	20	54.05%

In Table 4 we show the total of concepts that integrate a class-inclusion relation (CO) in the domain ontology and the total of these obtained by the LSA approach (Enc) for this type of relation.

The accuracy of the concepts found by the LSA method is greater than 79% with the cosine similarity variant (see Table 4). However, the CBC variant does not report satisfactory results for the first two ontologies. In the case of the OIL ontology it obtained a better behavior by achieving 62% accuracy, but without exceeding the result of the cosine variant (79%). The CBC method does not cluster all the concepts, so it was expected that most of the relations would not be found.

Table 4. Experimental results of concepts that maintain only a class-inclusion relation using the LSA approach for each domain ontology

Ontology	Variant	CO	Enc	Accuracy
AI	LSA-cosine	233	219	93.99%
	LSA-CBC	233	47	20.17%
SCORM	LSA-cosine	1154	1069	92.63%
	LSA-CBC	1154	285	24.70%
OIL	LSA-cosine	43	34	79.07%
	LSA-CBC	43	27	62.79%

In the case of non-taxonomic relations, the results obtained by the approach are presented in Table 5. Again, the cosine variant obtains better results (78% accuracy) than the CBC variant for this type of relation. As the CBC variant failed to cluster all concepts (see Table 6), the approach does not achieve a satisfactory accuracy in such relations. A first approximation of this approach is presented in [17] reporting only the concepts found by LSA.

Table 5. Experimental results of the LSA approach to non-taxonomic relations in each domain ontology.

Ontology	Total	Variant	Enc	Accuracy
AI	61	LSA-cosine	51	83.61%
		LSA-CBC	16	26.23%
SCORM	759	LSA-cosine	594	78.26%
		LSA-CBC	113	14.89%

In the case of concepts, the cosine variant obtains 85% accuracy in comparison with that obtained with the CBC variant (see Table 6).

Table 6. Experimental results of concepts that keep a non-taxonomic relation using the LSA approach for each domain ontology

Ontology	Variant	CO	Enc	Accuracy
AI	LSA-cosine	69	61	88.41%
	LSA-CBC	69	21	30.43%
SCORM	LSA-cosine	570	485	85.09%
	LSA-CBC	570	123	21.58%

7 Conclusions

The LSA method has been widely used in the state of the art to represent semantic at the context level, and with the proposed approach it was possible to obtain more than 70% of the semantic relations of each domain ontologies.

The results of the LSA approach, considering only the cosine similarity variant, obtained satisfactory results. But when the CBC variant was employed, it was not possible to find in the clustered concepts all the ontology relations (approximately only 10% of the total concepts).

The CBC method is very costly at runtime and did not produce satisfactory results. We consider that this is because we do not have enough information from each domain ontology, that this variant can process.

The LSA based approach requires a robust corpus (in terms of domain and size), including a large vocabulary that this allows more terms to be clustered. However, the accuracy offered is acceptable for one of the variants presented.

As future work we consider to increase the number of documents processed by the approach, as well as, the reviewing of other alternatives of concept clustering for the evaluation of domain ontologies.

References

1. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Technical report KSL-93-04, Knowledge Systems Laboratory, USA (1993)
2. Novelli, A.D.P., de Oliveira, J.M.P., Maria, J.: Simple method for ontology automatic extraction from documents. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **3**(12), 44–51 (2012). http://ijacsa.thesai.org/
3. Cederberg, S., Widdows, D.: Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (CONLL 2003), Stroudsburg, vol. 4, pp. 111–118. Association for Computational Linguistics (2003)
4. Sheena, N., Jasmine, S.M., Joseph, S.: Automatic extraction of hypernym and meronym relations in English sentences using dependency parser. Procedia Comput. Sci. **93**, 539–546 (2016). Proceedings of the 6th International Conference on Advances in Computing and Communications
5. Sankat, M., Thakur, R., Jaloree, S.: Design of ontology learning model based on text classification for domain concept taxonomy. IJSRSET **2**, 138–142 (2016). Themed Section: Engineering and Technology
6. Venegas, V.R.: Análisis Semántico Latente: una panorámica de su desarrollo. Revista signos **36**, 121–138 (2003)
7. Sidorov, G.: Non-linear Construction of n-grams in Computational Linguistics: Syntactic, Filtered, and Generalized n-grams. Instituto Politécnico Nacional, Mexico (2013)
8. Landauer, T.K., Dutnais, S.T.: A solution to platoś problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**, 211–240 (1997)

9. Vázquez Pérez, S.: Resolucin de la ambigedad semntica mediante mtodos basados en conocimiento y su aportacin a tareas de PLN. Ph.D. thesis, Universidad de Alicante (2009)
10. Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Stroudsburg, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
11. Pantel, P.A.: Clustering by committee. Ph.D. thesis, University of Alberta (2003)
12. Pantel, P., Lin, D.: Discovering word senses from text. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 613–619. ACM, New York (2002)
13. Pantel, P., Lin, D.: Document clustering with committees. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), pp. 199–206. ACM, New York (2002)
14. Chatterjee, N., Mohan, S.: Discovering word senses from text using random indexing. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 299–310. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78135-6_25
15. Porter, M.F.: Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., San Francisco (1997)
16. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)
17. Tovar, M., Pinto, D., Montes, A., González, G., Ayala, D.V., Beltrán, B.: Validación de conceptos ontoógicos usando métodos de agrupamiento. Res. Comput. Sci. **73**, 9–16 (2014)