ICITST-2008
3rd International Conference for Internet Technology and Secured Transactions
June 23–28, 2008, Dublin, Ireland (www.icitst.org)

Constructing ontologies for narrow domains: Methodology for term extraction and relationship discovery

Perez Fernando* **, Pinto David**, Rosso Paolo**, Cardiff John*

* Institute of Technology Tallaght, Dublin, Ireland
fernandoperez@itnet.ie, John.Cardiff@ittdublin.ie
** Natural Language Engineering Lab
Universidad Politécnica de Valencia, Spain
{dpinto, prosso}@dsic.upv.es

Abstract. Ontologies are commonly used in diverse natural language processing tasks. The fact that most of the ontologies are manually constructed arises the challenge of developing novel techniques for the automatic construction of linguistic resources of this kind. In this paper, as a first step of ontology construction, we present a methodology for term extraction and detection of relationship between term.

Keywords. Ontology, n-gramas, Mutual Information.

1 INTRODUCTION

The understanding of texts is something that computers cannot perform in an easy way. For this reason machines are provided with methods to interact between them and also interchange or reuse information.

In recent years, web applications have increased the necessity of interaction between computers and share information automatically. For this reason ontologies are proposed to fulfill the necessity of represent information in a semantic level.

An ontology can be defined as "an explicit specification of a conceptualization" [5]. In other words, an ontology is a model that represents concepts and relations of a specific domain in a semantic level and its objective is to provide information of implicit relations between concepts.

Designing and constructing ontologies are time consuming tasks that will need the knowledge of an expert in the domain in order to understand the conceptualization of the information needed for a given domain.

Ontology can help new technologies to use conceptual representation of knowledge as Semantic Web. Semantic Web uses the information given by an ontology to interpret and process the knowledge in order to provide better interaction with human beings. Some approaches as in [16] have been taken advantage of this representation of knowledge.

The paper is organized as follows. In the next section previous related approaches on ontology construction are presented. Section 3 introduces the proposed methodology for extracting terms and discovering relationships. Finally, in Section 4 conclusions are drawn and future work is discussed.

2 ONTOLOGY CONSTRUCTION

Ontologies are complex to build due to knowledge is not usually expressed in an explicit way. For Semantic Web, fast construction of specific domain ontologies is crucial for the success and proliferation of comprehensive and transportable machine understanding. In this Section some approaches that extract information to build ontologies are described.

2.1 Manual ontology construction

Up to now many approaches have appeared to help people in the manual construction of ontologies such as Ontolingua[13] and Protégé[15], but the demanding necessity for specific domain ontologies force ontology engineers to look for new methods for automatic or at least semi-automatic ontology construction. Some

advantages of the manual ontology construction are that we can detect co-existing concepts and avoid this situation. Unfortunately, manual construction is a very time-consuming task and the update of the ontology becomes another problem. These are some reasons why some approaches in automatic ontology construction are proposed.

In order to solve the problems of manual construction of ontologies, new techniques for automatic construction of this kind of resources are proposed. There is a clear necessity to detect relationships between terms, as a first step of ontology design, followed by the need to detect the kind of relation, e.g. synonymy, antonymy, hyperonymy or hyponymy.

2.2 Automatic ontology construction

In this section we describe recent approaches that confront the problem of ontology construction in a different perspective.

There are some techniques for automatic ontology construction based on linguistic patterns that take advantage of lexico-syntactic patterns to capture relations [7]. These techniques have shown very high precision but very low recall and they are used in order to extract explicit knowledge from the text.

On the other hand, there are other approaches that apply clustering methods for instance, in [9] the k-means algorithm [8] and html structure are exploited. Location of words is another important feature for context representation. Others use the Formal Concept Analysis (FCA) [4] which is based on the distributional hypothesis which says "words that occur in the same context tend to have similar meanings" [6].

The FCA method is investigated in many research works such as [2] and [1] which take also advantage of syntactic regularities using syntactic parsers and parse trees. Another work [3] joins two methods like Latent Semantic Indexing and k-means clustering technique to implement a system for semi-automatic topic ontology construction.

In [14] the authors propose a pattern-based method for automatically acquiring hyponyms from the web. Another work is [11], in which the authors identify pairs of words in opposite relationship (antonymy relations) from plain text; the method is based on relations between words and features are extracted from contexts of word pairs.

As we can see, to detect relations between terms is a crucial factor in ontology construction. In this paper, we present a novel methodology for detecting relations among n-grams (sequence of words); the use of n-grams is one component in our methodology which ultimately is more wide ranging in the field of natural language processing. Our technique is based on two steps: the extraction of important terms and the detection of relations between them.

3 PROPOSED METHODOLOGY

In our approach we look for implicit relations between terms without considering what kind of relationship will be found. Our proposal is based on the assumption that the observation in the co-occurrence between terms could give us semantic information and it could help to detect implicit relations.

3.1 N-grams

An N-gram is a sequence of words with a predefined order that can be used for statistical purposes in the area of natural language processing and it can be defined in terms of its length. Figure 1 shows how n-grams are constructed.

Figure 1. Example of n-grams

Sentence	The weather is cloudy
Unigrams	The, weather, is, cloudy
Bigrams	The weather, weather is, is cloudy
Trigrams	The weather is, weather is cloudy

3.2 Data set

Google n-grams: This corpus (Web 1T 5-gram Corpus Version 1.1) was provided by Google Inc. The data set contains English n-grams (sequence of n words) and their observed frequency counts. The length of the n-grams ranges from unigrams (single words) to five-grams.

Reuters Volume 1: This is a collection of newswires from Reuters for one year from 1996-08-20 to 1997-08-19.

3.3 Mutual Information

Mutual Information (MI) is a measure which is used to see the relationship and dependency between two terms (n-grams in our case) [12]. In other words, the amount of information one random variable contains about another.

Mutual Information is defined as: given x and y a pair of terms (n-grams in this case), Equation (1) , where $fr(x,y)$ is the number of sentences which contain both x and y; on the other hand $fr(y)$ and $fr(x)$ are the number of sentences which contain y and x respectively and N is the number of all the sentences in the corpus (it is a normalization factor).

$$MI(x,y) = \log_2(N \frac{fr(x,y)}{fr(x) \cdot fr(y)}) \qquad (1)$$

3.4 Methodology proposed

First a list of sentences of a specific domain (in our case "weather") was selected from Reuters corpus, n-grams were computed from raw text and a threshold was established to apply a filter and discriminate irrelevant terms.

The resulting set was used to calculate the intersection with n-grams provided by Google (Web1T5-gramCorpus Version1.1); this step is important because possible relations can be identified.

The next step was to identify the relations between terms or to calculate a factor to measure the degree of relation between them. Finally, a threshold for MI will be estimated

Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST-2008), June 23–28, 2008, DIT, Dublin, Ireland. ISBN 0-9546628-2-2 www.icitst.org/2008 © e-Centre for Infonomics

Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST-2008), June 23–28, 2008, DIT, Dublin, Ireland. ISBN 0-9546628-2-2 www.icitst.org/2008 © e-Centre for Infonomics

Figure 2. Methodology for term extraction andrelationship discovery

Let N be a set made of n-grams
$N=\{N_{A1}, N_{A2}, N_{A3}, N_{A4}, N_{A5}\}$
(unigrams N_{A1}, bigrams N_{A2}, trigrams N_{A3}, forthgrams N_{A4}, fivegrams N_{A5}) extracted from Reuters corpus.
Let NG be a set of n-grams (unigrams, bigrams, trigrams, forthgrams, fivegrams) Google n-grams.
And N_{AT} the set of n-grams N which frequency is equal or greater than a threshold T in the restricted domain corpus.
Let define $N_I = N_{AT} \cap NG$ as the terms extracted from the target corpus.
Calculate the Mutual Information to estimate the threshold for determining important relations between terms.

Table 1. Preliminary results

N-grams	MI Value
ice_storm remained_without_electricity	8,63038
northern_coast pacific_storm	8,5888
snowstorms european_russia	8,5563
agricultural_statistics statistics_service	8,4070
and_hurricane the_cayman_islands	8,19648
tunnel emergencies	7,79084
devastated coastal_communities	7,36875
increasing rain_could_fall	7,12639
russian closures_of_airports	7,11159
russian and_heavy_snowstorms	7,01206

to determine the existence of a significative relationship between terms.

The steps of the proposed methodology are illustrated in Figure 2.

In Table 1 we show some preliminary results obtained employing the proposed methodology for extracting terms. The results identify that some important terms can be representative for the specific weather domain.

4 CONCLUSIONS AND FURTHER WORK

Experimental results show that detecting relations between n-grams in specific domains is an important step that helps the identification of relations. As further work, the aim is to analyze another factor of correlation to detect relationship between words, and to implement the Levenshtein [10] distance although it is a measure of similarity between two strings but it will be applied to measure distance between n-grams.

ACKNOWLEDGMENTS

This work was partially funded by the MCyT TIN2006-15265-C06-04 research project. The research of the first author was made possible by a grant of ITT.

REFERENCES

1. Cimiano P., Hotho A., Staab S., (2004), Learning Concept Hierarchies form Text Corpora.
2. Faure D., Nedellec C., (1998), A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition.
3. Fortuna B., Mladenic D., Grobelnik M., (2006), Semi-Automatic Construction of Topic Ontology.
4. Ganter, B., Reuter, K., (1991), Formal Concept Analysis- Mathematical Foundations. Springer Verlag.
5. Gruber T. R., (1993), A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220.
6. Harris, Z., (1968), Mathematical Structures of Language. Wiley.
7. Hearst M., (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France.
8. Jain A. K., Murty M. N., Flynn P. J., (1999), Data Clustering: A Review, ACM Comp, Surv.
9. Karoui L., Aufaure M., Bennacer N., (2006), Context-based Hierarchical Clustering for the Ontology Learning, Proceedings of the IEEE WIC ACM International Conference on Web Intelligence.
10. Levenshtein V., (1996), Binary codes capable of correcting deletions, insertions, and reversals. Doklady akademii nauk SSSR, 163(4):845_848, 1965, in Russian. English Translation in Soviet Physics Doklady, 10(8) p. 707_710.
11. Lucero C., Pinto D., Jiménez-Salazar H., (2004), An Automatic Method to Identify Antonymy Relations, Workshop on Lexical Resources and the Web for Word Sense Disambiguation, Proceedings of Workshops on Artificial Intelligence, 105-111.
12. Manning C., Schütze H., (1999), Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.
13. Ontolingua: http://www.ksl.stanford.edu/software/ontolingua/ (2008).
14. Ortega-Mendoza R., Villaseñor-Pineda L. and Montes-y-Gómez M., (2007), Using Lexical Patterns for Extracting Hyponyms from the Web, MICAI 2007: Advances in Artificial Intelligence, Springer Berlin , Heidelberg, Volume 4827.
15. Protégé: http://protege.stanford.edu/ (2008).
16. Sampson, D. G., Lytras, M. D., Wagner, G., Diaz, P., (2004), Ontologies and the Semantic Web for E-learning. Educational Technology and Society, 7 (4), 26-28.