Extracting Ontology Hierarchies From Text

Jone Correia, Rosario Girardi, Carla Faria
Federal University of Maranhão
São Luís, Brazil
jonecorreia@ufma.br, rgirardi@deinf.ufma.br, carlafaria@ifma.edu.br

Abstract — Ontologies are an approach for knowledge representation capable of expressing a set of entities and their relationships, constraints, axioms and vocabulary of a given domain. Manual construction of ontologies by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and/or semi-automatic approaches are needed. Ontology Learning looks for automatically or semi-automatically identifying ontology elements like classes, taxonomic and non-taxonomic relationships, properties and axioms from textual resources. This article proposes a process for the automatic extraction of ontology taxonomic relationships from English texts using natural language processing techniques. Some experiments using a legal corpus were conducted in order to evaluate it. Initial results are promising.

Keywords - Ontology, Ontology Learning, Natural Language Processing, Taxonomic Relationships, Ontology Hierarchies

I. INTRODUCTION

Ontologies are used by modern knowledge-based systems allowing the representation and sharing of knowledge about an application domain [9]. They provide a formal means of knowledge representation capable of expressing a set of entities, their relationships, constraints and rules (conditional statements) of a given domain [10][15].

Ontology Learning looks for identifying ontology elements like classes, taxonomic and non-taxonomic relationships, properties and axioms from textual resources. Manual construction of ontologies by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and/or semi-automatic approaches are needed.

This paper proposes a process for automatic learning of ontologies from text based on Natural Language Processing (NLP) [1][4] techniques. Particularly, the process looks for identifying taxonomic relationships. An experiment conducted to evaluate this the process using a legal corpus is also described.

The article is organized as follows. Section II introduces the ontology definition used in this work. Section III presents an overview of the proposed process. Section IV describes an experiment conducted to evaluate it. Section V summarizes related work and, finally, section VI concludes the article discussing results and future work.

II. AN ONTOLOGY DEFINITION

Ontologies are formal specifications of concepts in a domain of interest. Their classes, relationships, constraints and axioms define a common vocabulary to share knowledge [10].

Formally, an ontology can be defined as the tuple:

\[ O = (C, H, I, R, P, A) \] (1)

where,

\[ C = \{c \mid c \in C\} \]

\[ H = \{\text{kind_of}(c_1, c_2) \mid c_1 \in C, c_2 \in C\} \]

\[ I = \{\text{is_a}(c_1) \mid c_1 \in C\} \cup \{\text{prop}(c, \text{value}) \mid c \in C\} \]

\[ R = \{\text{rel}(c_1, c_2, \ldots, c_n) \mid \forall i, c_i \in C\} \]

\[ P = \{\text{condition} \mid c \in C\} \]

\[ A = \{\text{axiom} \mid c \in C\} \]

This work is supported by the Brazilian Government agencies: CNPq, CAPES and FAPEMA.
indicates that if two daughters have the same mother then, the daughters are sisters.

III. A PROCESS FOR ACQUIRING TAXONOMIC RELATIONSHIPS

Figure 2 illustrates the proposed process for the automatic acquisition of ontology taxonomic relations, that is, the $H$ set of the ontology definition in section II. It consists of four steps: “Tagging”, “Extraction of Candidate Classes”, “Identification of Hyponyms and Synonyms” and “Identification and Representation of Taxonomic Relationships”.

The input of the process is a corpus consisting of a set of textual documents in a particular application domain. The step “Tagging” is intended to identify the tokens, sentences, grammatical classes and lemmas with the application of NLP techniques. Then, the “Extraction of Candidate Classes” phase is responsible for separating the tokens which are likely to be classes of the ontology hierarchy. The step “Identification of Hyponyms and Synonyms” identifies the synonyms and hyponyms of the candidate classes obtained in the previous step. The last stage of the process is the “Identification and Representation of Taxonomic Relations” phase which aims at identifying the final taxonomic relations using heuristic patterns and their representation in an ontology specification language. In the next sub-sections the transformations made at each stage of the process are presented in detail.

The “Tokenization” activity identifies the terms (tokens) in the document. The activity “Division in Sentences” organizes the identified tokens by grouping them into sentences. The activity “Lemmatization” performs the reduction of each token to its basic form. This form generalizes the inflected forms of a token making possible the grouping of tokens. The activity “Lexical Analysis” aims at identifying the grammatical classes for each token selected in the tokenization activity.

The result of this step is the corpus tagged with tokens, sentences, lemmas and grammatical categories. For example, Figure 4 shows the processing results of applying the “Tagging” phase to the text fragment “Mens have mothers. Mothers are great people!”.

The first frame shows the terms that were marked as tokens: “Men”, “Have”, “Mothers”, “Mothers”, “Are”, “Great” and “People”. Dots and spaces are also tokens, however, for illustrative purposes, only the terms were shown. The second frame shows the result of dividing the text into sentences: “Men Have mothers” and “Mothers are great people”. Next “Lemmatization” is performed on each term. For example, the term “mothers” has “mother” as lemma. The last frame corresponds to Lexical Analysis and shows each token and its corresponding part of speech. For example, the term “mothers” is tagged NNS [14] indicating that this is a plural noun.

A. Tagging

The step “Tagging” has a corpus as input and aims at transforming it into a model that can be processed computationally. This step consists of the following activities: “Tokenization”, “Division in Sentences”, “Lemmatization” and “Lexical Analysis” (Figure 3).

B. Extraction of Candidate Classes

This phase aims at selecting candidate classes among concrete and abstract nouns.

The hypothesis assumed in this work is that only nouns may be considered classes. Thus, conceptually speaking, we can find two kinds of nouns: concrete nouns and abstract nouns. Concrete nouns are used to represent people names (anthroponyms) or place names (toponyms), which essentially characterizes class instances. Therefore, in the “Extraction of Candidate Classes” anthroponyms and toponyms are ignored and only concrete and abstract nouns are selected.

The products of this step are candidate classes, sentences, lemmas and grammatical classes.
C. Identification of Hyponyms and Synonyms

The "Identification of Hyponyms and Synonyms" looks for the identification of synonyms and hyponyms in WordNet [21], a lexical database that contains natural language terms, their definitions and their semantic relationships as synonymy, hyponymy and hypernym. The classes selected in the previous step are located in Wordnet and whenever a synonym or hyponym is found this occurrence in the text is stored. That is, each class is associated with an array of synonyms and hyponyms from the text. The product of this step is the same product from the previous stage plus the relations of synonymy and hyponymy, of each select term, found in Wordnet.

D. Identification and Representation of Taxonomic Relationships

The "Identification and Representation of Taxonomic Relationships" aims at discovering taxonomic relationships through the application of heuristic patterns and at representing them in an ontology specification language. Each sentence is analyzed in order to match the patterns. To test the existence of hyponymy relations we apply regular expressions on sentencing patterns with heuristic examples as the Hearst patterns [11]. Regular expressions are a formal method to specify a text pattern [12]. For example, applying the regular expression "\([A-Za-z]+,+.+\)[A-Za-z]+ and other [A-Za-z]+" to "My daughter loves dolls, masks and other toys", the matched pattern should be "My daughter loves dolls, masks and other toys". The tokens "Dolls" and "Masks" are related to "Toys" in a relationship of hyponymy, thus being "Toy" a superior class of "Doll" and "Mask" in the hierarchy of an ontology. This matching represents the pattern (iv) in the table of Hearst heuristics patterns of Figure 5, where NP0 represents a token hierarchically superior to the other NP.

\[(i) \text{NP0 such as } \text{NP1}, \text{NP2} \ldots, \text{(and | or) NP1} \\
(ii) \text{such NP0 as } \{\text{NP}, \text{NP}\} \ast \text{(and | or) NP} \\
(iii) \text{NP, (NP) } \ast \text{ (and | or) other NP0} \\
(iv) \text{NP, (NP) } \ast \text{ (and | or) other NP0} \\
(v) \text{NP, (NP) } \ast \text{ (and | or) other NP0} \\
(vi) \text{NP, (NP) } \ast \text{ (and | or) other NP0} \]

Figure 5. Heuristic Patterns of Hearst (Hearst, 1992)

The products of this phase are the taxonomic relationships, the set H of the definition of ontology in Section II, represented in an ontology specification language.

IV. EVALUATION

A case study in the area of Family Law allows for a preliminary assessment of the effectiveness of the proposed process. For that, a prototype tool called T-NLPDumper has been developed for automating the process using Java [13].

A corpus from the Family Law Doctrine [7] has been used in the case study. One hundred sentences randomly selected from the corpus has been manually analyzed by a domain expert, who did the manual identification of taxonomic relationships. These results (Figure 6) were manually compared with the results found on the automatic extraction of taxonomic relationships with the T-NLPDumper tool.

The results were also compared with an adaptation of the precision measure from the Information Retrieval area [5], considering the number of correctly extracted taxonomic relations.

Precision is the ratio between the number of taxonomic relationships extracted correctly (NREC) and number of taxonomic relationships extracted (NRE).

\[
P = \frac{\text{NREC}}{\text{NRE}}
\]

The precision value obtained in this experience was 81.57%.

Figure 6. Part of the ontology hierarchy obtained as a result of the experience.

V. RELATED WORK

Table 1 shows the main approaches for automatic and semi-automatic identification of taxonomic relationships.

The main techniques used (second column of Table 1) are based on lexical syntactic patterns [11][19], statistics with Markov Logic Networks (MLR) [6] and Machine Learning techniques (ML) [3][2][16]. More recent works have in common the application of techniques of Natural Language Processing (NLP) in association with other techniques.

The approaches described in [6][16][19] and the one proposed in this article use WordNet as a lexical base for the extraction of hyponyms and GATE [8] to perform the Natural Language Processing tasks. Weka [20] was used as a framework for the clustering tasks in [16], while the tools TreeTagger [18] and Lopar [17] are used in [3] to find grammatical categories and as a parsing tool, respectively.

The effectiveness is shown in terms of the precision obtained for each approach in their experiments. However, these numbers cannot be considered for comparative purposes, because the evaluation of each approach was performed with different corpora and ontologies.
The approaches in [2], [3], [11] and the one proposed in this work provide solutions for a fully automatic acquisition of taxonomic relations, while the approaches in [16] and [6] propose just semi-automatic solutions.

Main advantages of the technique proposed here are its domain independency and high precision value (more than 80%).

VI. CONCLUSION

The automatic process for ontology learning proposed in this paper is based on natural language processing and consists of four steps: "Tagging", "Extraction of Candidate Tokens", "Identification of Hyponyms and Synonyms" and "Identification and Representation of Taxonomic Relations".

The process has been evaluated through a case study conducted in the domain of Family Law, showing good results, and demonstrating that the use of techniques of natural language processing represents a promising approach for learning taxonomic relationships of ontologies.

Currently, new heuristic patterns are being developed in order to improve the identification of taxonomic relationships. The development of a tool for implementing all the proposed process is also under construction.

Finally, to improve the results of the proposed tool it is necessary to compare them with concept extractions done manually by domain experts in other domains.