Intelligent Information Management, **2011, 3, 119-124**
doi:10.4236/iim.2011.34014 Published Online July 2011 (http://www.SciRP.org/journal/iim)

A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text

Ivo Serra, Rosario Girardi
Federal University of Maranhão, Beren, Brazil
E-mail: *ivocserra@gmail.com, rosariogirardi@gmail.com*
Received February 15, 2011; *revised April* 11, 2011; *accepted April* 25, 2011

Abstract

Manual construction of ontologies by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and/or semiautomatic approaches are needed. Ontology learning looks for identifying ontology elements like non-taxonomic relationships from information sources. These relationships correspond to slots in a frame-based ontology. This article proposes an initial process for semi-automatic extraction of non-taxonomic relationships of ontologies from textual sources. It uses Natural Language Processing (NLP) techniques to identify good candidates of non-taxonomic relationships and a data mining technique to suggest their possible best level in the ontology hierarchy. Once the extraction of these relationships is essentially a retrieval task, the metrics of this field like recall, precision and f-measure are used to perform evaluation.

Keywords: Ontology, Ontology Learning, Non-Taxonomic Relationships, Natural Language Processing

1. Introduction

An ontology is a formal and explicit specification of a conceptualization of a domain of interest [9,15] that defines concepts and relationships between those concepts to represent knowledge in that domain. Ontologies have a great importance for modern knowledge systems since they provide a formalism for structuring knowledge bases and their reusing and sharing [13,16].

There are two fundamental aspects in ontology learning. The first is the availability of prior knowledge, which may be in the form of an ontology to be extended or to be transformed into the first version of an ontology. This version is then automatically extended by learning procedures or manually by the knowledge engineer. The other aspect is the format of data sources from which you want to extract knowledge. There are three different types of data sources: unstructured sources (documents in natural language like traditional Web pages), semi-structured sources (dictionaries and folksonomies) and structured sources (database schemas).

Some approaches for ontology learning from structured [4] and semistructured sources [8] were proposed and showed good results. However, even considering that these approaches provide support for the development of ontologies, most of the available knowledge, especially on the Web, is in the form of texts in natural language [1].

This paper proposes a process for automating the extraction of non-taxonomic relationships between concepts of ontologies from textual sources. These relationships correspond to slots in a frame-based ontology. For example, in the field of Family Law, we expect to extract relationships such as "represents" between the frames "Lawyer" and "Client" and "judge" between the frames "Court" and "Action". Issues related to this extraction are the definition of the label of the relationship, its hierarchical level in the ontology taxonomy and in which frame the slot should be added.

This paper is organized as follows. Section 2 presents the definition of ontology used in this work. Section 3 describes non-taxonomic relationships and its linguistic realizations. Sections 4 and 5 discuss an initial approach to extract these relationships from text and metrics to perform evaluations. Section 6 introduces main related work and finally, Section 7 concludes the work.

2. 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 [9]. Following, an ontology definition and a

simple example in the domain of family relationships are presented.

Formally, an ontology can be defined as the tuple (1):

$$O = (C, H, I, R, P, A) \quad (1)$$

where,

$C = C_C \cup C_I$ is the set of entities of the ontology. They are designated by one or more terms in natural language. The set C_C consists of classes, *i.e.*, concepts that represent entities that describe a set of objects (for example, "Mother" $\in C_C$), while the set C_I is constituted by instances, (for example, "Anne Smith" $\in C_I$).

$H = \{\text{kind\_of}\ (c_1,c_2) \mid c_1 \in C_C,\ c_2 \in C_C\}$ is the set of taxonomic relationships between concepts, which define a concept hierarchy and are denoted by "kind_of (c_1,c_2)", meaning that c_1 is a subclass of c_2, for instance, "kind_of (Mother, Person)".

$I = \{\text{is\_a}\ (c_1,c_2) \mid c_1 \in C_I \wedge c_2 \in C_C\} \cup \{\text{prop}_K\ (c_i, \text{value}) \mid c_i \in C_I\} \cup \{\text{rel}_K\ (c_1, c_2, ..., c_n) \mid \forall i, c_i \in C_I\}$ is the set of relationships between ontology elements and its instances. For example, "is_a (Anne, Mother)", birth (Anne Smith, 02/12/1980)" and "mother_of (Anne Smith, Clara Smith)" are relationships between classes, relationships, properties with its instances.

$R = \{\text{rel}_k\ (c_1,c_2, ..., c_n) \mid \forall i, c_i \in C_C\}$ is the set of ontology relationships that are neither "kind_of" nor "is_a". For example "mother_of (Mother, Daugther)".

$P = \{\text{prop}_K\ (c_i, \text{datatype}) \mid c_i \in C_C\}$ is the set of properties of ontology classes. For instance, "date_of_ birth (Mother, mm/dd/yyyy)".

$A = \{\text{condition}_x \Rightarrow \text{conclusion}_y\ (c_1, c_2, ..., c_n) \mid \forall j, cj \in C_C\}$ is a set of axioms, rules that allow checking the consistency of an ontology and infer new knowledge through some inference mechanism. The term condition_x is given by $\text{condition}_x = \{(\text{cond}_1, \text{cond}_2, ..., \text{cond}_n) \mid \forall z, \text{cond}_z \cup H \cup I \cup R\}$. For example, "$\forall$ Mother, Daughter_1, Daughter_2, mother_of (Mother, Daughter_1), mother_ of (Mother, Daughter_2) $\Rightarrow$ sister_of (Daughter1, Daughter2)" is a rule that indicates that if two daughters have the same mother then, they are sisters.

3. Non-Taxonomic Relationships

Non-taxonomic relationships can be classified in domain independent and domain dependent. The domain independent relationships are of two subtypes: ownership and aggregation. Aggregation is the "whole-part" relationship. For example, in the sentence "The car's wheel is out of order". We have a non-taxonomic relationship of aggregation between "car" and "wheel". The linguistic realization of the relationship of aggregation occurs in two forms: the possessive form of English (apostrophe) and the verb "have" in any conjugation. However, the converse is not true, *i.e.*, the occurrence of such linguistic accomplishments does not imply a relationship of aggregation as will be explained in the next case. Ownership relationships are held as in the example: "Father and mother will wait for the court's decision" in which there is a relationship of possession between "court" and "decision". The linguistic realization of this kind of relationship occurs in two forms: the possessive form of English (apostrophe) and the verb "have" in any conjugation. However, the converse is not true, *i.e.*, the occurrence of such linguistic accomplishments does not imply a relationship of possession. Domain dependent relationships are expressed by particular terms of an area of interest. For example, the sentence "The court will judge the custody in three days" shows the relationship "judge" between the terms "court" and "custody" and is characteristic of the legal field. **Table 1** summarizes the subtypes of non-taxonomic relationships and their dependence/independence of the domain.

4. The Proposed Process

The proposed process makes use of NLP [3,12] and data mining techniques [5] to extract, from textual sources in English, non-taxonomic binary relationships between two ontology classes. The technique retrieves the relationships indicated by verbs in a sentence, and suggests the possible best level in the ontology hierarchy where the relationship should be added. The process (**Figure 1**) is composed of three phases: extraction of candidate relationships, analysis of the hierarchical level and manual selection of relationships. These phases are detailed in the following sections.

4.1. Extraction of Candidate Relationships

The extraction of candidate relationships phase makes use of NLP techniques to extract from text an initial set of relationships (**Figure 2**). Initially, the text is split into sentences since relationships are identified only between terms in the same sentence. Then, a search is done in the sentences to find those that have at least two terms that can represent concepts from the class hierarchy of the ontology. For that, the class concepts of the ontology hierarchy are expanded with their synonyms and possibly with their hyponyms and heteronyms in a generalization/specialization level defined by the user. For example, beyond the term "wine" we can consider one level higher in the hierarchy of drink concepts and then include "alcoholic drink" in the search. These two parameters are intended to increase the recall of the search. Next, a lexical analysis is performed only on the sentences retrieved in the previous.

Table 1. Non-taxonomic relationships and its dependency/ independence with a domain.

Kind	Sub-kind	Linguistic realization	Extraction	Example
Domain independent	Aggregation	Possessive and the verb "have"	NLP techniques	"A typical car has four wheels"
	Ownership	Possessive and the verb "have"	NLP techniques	"Father and mother will wait for the court's decision."
Domain dependent	-	Domain verbs	Statistical methods	"The court will judge the custody in three days."

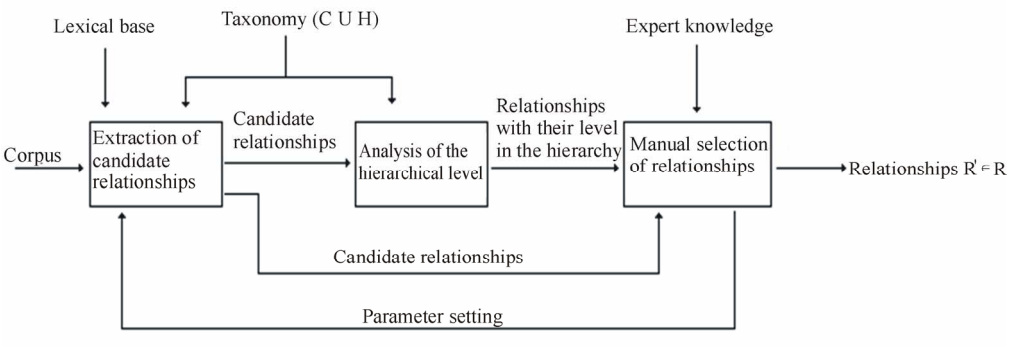

Figure 1. The proposed process for extracting non-taxonomic relationships.

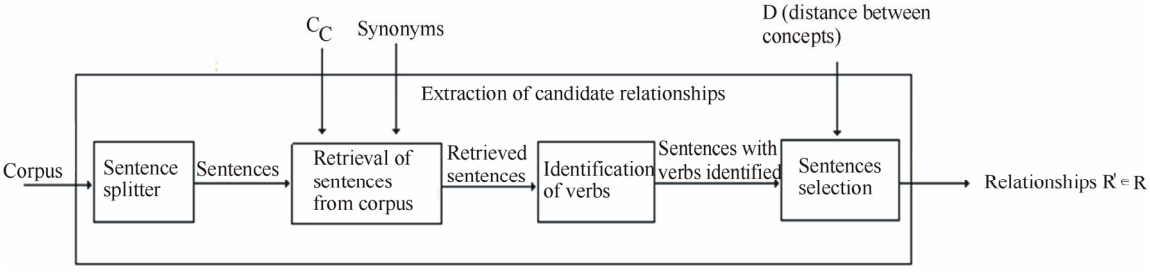

Figure 2. Phase of extraction of candidate relationships.

step. The goal is to find the verb forms as indicative of non-taxonomic relationships.

The last step consists on the generation of tuples composed of two concepts and a verb relating them from the sentences considered in the previous activity. Two situations are possible. First, there can be sentences having terms that represent ontology concepts that are at a maximum distance of D terms (being D a non-negative integer) and have a verbal form among them. For this situation, a tuple in the form <concept 1, verb form, concept 2> is generated. Second, there can be sentences that have the contract form "'", as in "Courts' decision". In this case, a tuple is generated with the format <concept 1, has, concept 2>. It is also generated an alert to the user that he/she needs to take a decision about the label of this relationship, since it may not be an aggregation, and so "has" may not be the best label. For example, in the sentence "Father and mother will wait for the court's decision", the best label for the relationship between "court" and "decision" might be "take".

4.2. Analysis of the Appropriate Hierarchical Level

To suggest the most appropriate level in the ontology hierarchy where to insert the relationship as a class slot, the algorithm for discovering generalized association rules proposed by Srikant and Agrawal [14] is used. One popular application of this algorithm is to find associations between products that are sold in a supermarket and describe them in a more appropriate hierarchical level. For example, a valid association could be "snacks are purchased together with drinks" rather than "chips are purchased with beer" and "peanuts are purchased with soda". The basic algorithm for extracting association rules uses a set of transactions $T = \{t_i \mid i = 1, ..., n\}$, each transaction t_i consists of a set of items, $t_i = \{a_{i,j} \mid j = 1, ..., m_i, a_{i,j} \in C\}$ and each item $a_{i,j}$ is an element of a set of concepts C. The algorithm computes association rules $X_k \Rightarrow Y_k$ (X_k, $Y_k \subset C$, $X_k \cap Y_k = \{\}$) which have values for the measures of support and confidence above a given threshold. The support of a rule $X_k \Rightarrow Y_k$ represents the percentage of

transactions that have $X_k \cup Y_k$ as a subset; the confidence is defined as the percentage of transactions that has Y_k as consequent when X_k is the precedent of the rule. Formally, support and confidence are given by the formulas: Support $(X_K \Rightarrow Y_K) = |\{t_i \mid X_K \cup Y_K \subseteq t_i\}| / n$ and Confidence $(X_K \Rightarrow Y_K) = |\{t_i \mid X_K \cup Y_K \subseteq t_i\}| / |\{t_i \mid X_K \subseteq t_i\}|$. To extract associations between concepts in the correct hierarchical level of a hierarchy, every transaction t_i is extended to include the ancestors of each item $a_{i,j}$, for example, $t_i' := t_i \cup \{a_{i,l} \mid (a_{i,j}, a_{i,l}) \in H\}$. Then, support and confidence are computed for all possible association rules $X_k \Rightarrow Y_k$, such that Y_k doesn't have an ancestor of X_k since this would be a trivial association. Finally, we exclude all association rules $X_k \Rightarrow Y_k$ that have lower values for support and confidence than an ancestor rule $\underline{X_k} \Rightarrow \underline{Y_k}$. Itemsets $\underline{X_k}$ and $\underline{Y_k}$ contain only ancestors or items found in itemsets of the rule $X_k \Rightarrow Y_k$.

4.3. Manual Selection of Relationships

No technique of NLP or Machine Learning (ML) replace better the expert decision in an environment of ambiguous nature as learning from natural language sources. Therefore, the goal of this phase is to make the best possible suggestions to the user and give him/her the control to take the final decision. Thus, the result of the technique should be evaluated by a specialist before the relationships can be definitely added to the ontology.

Issues such as scope of knowledge to be represented, level of generalization, the real need of adding a relationship, its direction and label must ultimately be evaluated, selected, and possibly adjusted by an expert.

5. Evaluation

Since the extraction of non-taxonomic relationships can be seen as an activity of information retrieval, measures to evaluate such systems can be used in this context. The usual measures of effectiveness of retrieval systems are precision and recall. Recall measures the system's ability to retrieve relevant information from all relevant information available, which in this context correspond to the extraction from the corpus as many tuples corresponding to real relationships. Precision measures the relevance of what was recovered, in other words, it is the system's ability to reject irrelevant information, that in the context of non-taxonomic relationships retrieval corresponds to ensure that among the retrieved tuples most of them represent real relationships. **Figure 3** defines recall and precision. To this end we define the following sets: "T^+" corresponds to all relevant tuples present in the corpus, *i.e.*, those that represent relationships. "T^-" corresponds to all irrelevant tuples present in the corpus, *i.e.*, those

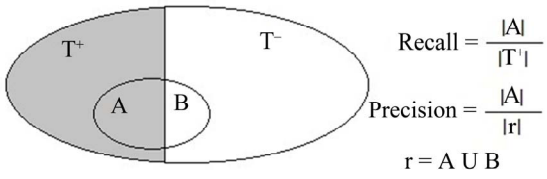

Figure 3. Measures of recall and precision.

that do not represent relationships and "*r*" is the retrieved set of tuples ($r = A \cup B$). "*A*" is the set of retrieved tuples that represent relationships and "B" corresponds to those that do not represent relationships.

Usually, mechanisms to improve recall reduce the precision and vice versa. Thus it is not desirable to provide good values only for recall (*R*) or precision (*P*). It is important to have a good combination of both. *A* measure frequently used to reflect this combination in a single value is the *F*-measure, a harmonic average of both that is given by (2).

$$F - measure = \left(2 * R * P\right) / \left(R + P\right) \quad (2)$$

6. Related Work

Table 2 shows a comparison of some approaches for the extraction of non-taxonomic relationships with the one proposed in this work. Most of them combine both NLP and ML techniques.

Villaverde *et al.* [5] proposed an approach based on the premise that non-taxonomic relationships are usually expressed by verbs that relate pairs of concepts (elements of the *C* set in the definition of section 2). First, sets of synonyms of concepts from an ontology are created, using Wordnet. The use of synonymous increases the recall of concepts extracted from a corpus. The corpus is then searched to identify pairs of concepts that occur in the same sentence with verbs that relate them. Thus, each occurrence of a relationship has the form of a tuple <concept 1, concept 2, verb> which make up the set of candidate relationships. It is then applied over this set a mining algorithm of association rules, which in this case will be of the form $\{<ci, cj> \Rightarrow v \mid c_i, c_j \in C$ and v is a verb$\}$. As a result in [5], rules are extracted that, according to the statistical evidence measures, support and confidence, represent good suggestions of non-taxonomic relationships. For example, given the concepts "parent" and "child" and the verb "has", the tuple <parent, child, has> represents the co-occurrence of these three terms in a sentence of at least one of the documents. If the rule <parent, child> ⇒ <has> has a support greater than the minimum, the strength of association between these two concepts, linked by this verb, is given by the confidence of the rule. The recommendation of an association rule is ultimately made based on the measure of his confidence,

Table 2. Processes for extracting non-taxonomic relationships from text.

	Employed techniques	Machine learning technique	Relationships are indicated by verbs	Suggests the hierarchical level
Villaverde et al [5]	NLP and data mining	Extraction of association rules	yes	no
Maedech and Staab [1]	NLP and data mining	Extraction of generalized association rules	no	yes
Sanchez and Moreno [2]	NLP and web search engines	-	yes	no
Serra and Girardi	NLP and data mining	Extraction of generalized association rules	yes	yes

which depends on its format. Thus, it would be necessary an evaluation about the consequences of the rule format in the final result. For example, the rule $<c_i, c_j> \Rightarrow v$ could be recommended, whereas $v \Rightarrow <c_i, c_j>$ could not be.

Maedech and Staab [1] propose a similar process to that of Villaverde *et al.* [5], with the difference that it uses an algorithm of generalized association rules [14] to suggest the possible most appropriate hierarchical level for the relationship. This approach works with texts in German. Sánchez and Moreno [2] propose an automatic and unsupervised technique for learning non-taxonomic relationships that is able to learn verbs from a domain, to extract related concepts and label them using the Web instead of a corpus as a source for the construction of an ontology. Despite being unstructured and diverse, according to the authors, the redundancy of information in an environment as vast as the Web is a measure of its relevance and veracity. The first stage is the extraction and selection of verbs that express typical relationships of the area. Based on morphological and syntactic analysis, verbs that have a relationship with the domain keyword are extracted. To avoid the natural language complexity, some constraints are used, for example, verb phrases containing modifiers in the form of adverbs are rejected. Then, it measures the degree of relationship between each verb and the domain. To do so statistical measures are made about the term distribution on the web. The values obtained are used to rank the list of candidate verbs. This let one choose the labels of non-taxonomic relationships that are closely related to the domain. The domain dependent verbs are used to discover concepts that are non-taxonomically related. To do so the system queries a web search engine with these patterns: "verb domain-keyword" or "domain-keyword verb" that returns a corpus related to the specified queries. The goal is to assess the content of documents to find concepts that precede ("High sodium diets are associated with hypertension") or succeed ("Hypertension is caused by hormonal problems") the constructed pattern which will be candidate to be non-taxonomic related to the original domain keyword.

7. Concluding Remarks

Most efforts on ontology development are required to identify and specify its non-taxonomic relationships and there is still a lack of effective techniques and tools to automate and even provide an appropriate help to these tasks.

This paper described a process to extract non-taxonomic relationships from English texts. The process is semi-automatic once it presents to the specialist a list of probable relationships that will be selected manually. The process uses NLP techniques to extract candidate relationships. It aims at extracting pairs of concepts in a sentence with the verb that probably link them. For this purpose, the specialist is asked to interactively adjust a parameter that indicates the maximum distance, in words, between two concepts in a sentence for them to be considered related by a verb located in between. The technique also includes a phase, which can be optionally performed, for the identification of the best level of the ontology hierarchy where a non-taxonomic relationship should be included. For this purpose a ML technique for the extraction of generalized association rules proposed by Srikant e Agrawal [14] is used.

A tool is been developed to automate and better evaluate [7] the proposed process. For the evaluation, a corpus of 500 documents in the Family Law doctrine will be searched for non-taxonomic relationships and the result will be compared against Family Law, a reference ontology in the same domain. The effectiveness of results will be measured using the traditional information retrieval measures (recall, precision and f-measure).

8. References

[1] A. Maedche and S. Staab, "Mining Non-taxonomic Conceptual Relations from Text," *Proceedings of Knowledge Engineering and Knowledge Manageme Methods, Models and Tools*: 12*th International Conference*, Berlin Springer, 2000, pp. 189-202.

[2] D. Sanchez and A. Moreno, "Learning Non-Taxonomic Relationships from Web Documents for Domain Ontology Construction," *Data and Knowledge Engineering*,

Vol. 64, No. 3, 2008, pp. 600-623. doi:10.1016/j.datak.2007.10.001

[3] J. Allen, "Natural Language Understanding," Redwood City, CA: The Benjamin/Cummings Publishing Company, 1995.

[4] J. Lehmann and P. Hitzler, "A refinement Operator Based Learning Algorithm for the ALC Description Logic," *Proceedings of International Conference on Inductive Logic Programming*, Springer-verlag, Corvallis Berlin, 2007, pp. 147-160.

[5] J. Villaverde, A. Persson, D. Godoy and A. Amandi, "Supporting the Discovery and Labeling of Nontaxonomic Relationships in Ontology Learning," Expert System Applications, Vol. 36, No. 7, 2009, pp. 10288-10294. doi:10.1016/j.eswa.2009.01.048

[6] K. Bontcheva and H. Cunningham, "The Semantic Web: A New Opportunity and Challenge for Human Language Technology," *Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, 2003.

[7] K. Dellschaft and S. Staab, "On How to Perform a Gold Standard Based Evaluation of Ontology Learning," *Proceedings of the 5th International Semantic Web Conference*, Athens Springer, 2006, pp. 228-241.

[8] L. Marinho and K. Buza, "Schmidt-Thieme, L. Folksonomy-based Collabulary Learning," *Proceedings of International Semantic Web Conference*, Karlsruhe Berlin: Springer-Verlag, 2008, pp. 261-276.

[9] N. Guarino, C. Masolo and C. Vetere, "Ontoseek: Content-Based Access to the Web," *IEEE Intelligent Systems*, Vol. 14, No. 3, 1999, pp. 70-80. doi:10.1109/5254.769887

[10] P. Buitelaar, P. Cimiano and P. Magnini, "Ontology Learning from Text: Methods, Evaluation and Applications," IOS Press, Amsterdam, 2006.

[11] P. Cimiano, J. Volker and R. Studer, "Ontologies on Demand?—A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text," *Wissenschaft and Praxis*, Vol. 57, No. 6-7, 2006, pp. 315-320.

[12] R. Dale, H. Moisl and H. L. Somers, "Cyclic Redundancy Check," VDM Publishing House Ltd, Saarbrücken, 2000.

[13] R. Girardi, "Guiding Ontology Learning and Population by Knowledge System Goals," *Proceedings of International Conference on Knowledge Engineering and Ontology Development*, Education INSTIIC, Valence, 2010, pp. 480-484.

[14] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," *Future Generation Computer Systerms*, Vol. 13, No. 2-3, 1997, pp. 161-180. doi:10.1006/ijhc.1995.1081

[15] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal of Human-Computer Studies*, Vol. 43, 1995, pp. 907-928

[16] V. Alexiev, M. Breu, J. De Bruijn, D. Fensel, R. Lara and H. Lausen, "Information Integration with Ontologies: Experiences from an Industrial Showcase," Wiley, Neu-Ulm, 2005.

