Reviewing the Problem of Learning Non-Taxonomic Relationships of Ontologies from Text

Ivo Serra[1], Rosario Girardi[1] and Paulo Novais[2]

[1] Federal University of Maranhão, Computer Science Departament, São Luís, Brazil
[2] University of Minho, Computer Science Departament, Braga, Portugal
ivocserra@gmail.com, rosariogirardi@gmail.com, pjon@di.uminho.pt

Abstract. Learning Non-Taxonomic Relationships is a sub-field of Ontology Learning that aims at automating the extraction of these relationships from text. This article discusses the problem of Learning Non-Taxonomic Relationships of ontologies and proposes a generic process for approaching it. Some techniques representing the state of the art of this field are discussed along with their advantages and limitations. Finally, a framework for Learning Non-Taxonomic Relationships being developed by the authors is briefly discussed. This framework intends to be a customizable solution to reach good effectiveness in the process of extraction of non-taxonomic relationships according to the characteristics of the corpus.

Keywords: Ontology, Ontology learning, Non-taxonomic relationships, Natural Language Processing.

1 Introduction

An ontology is a formalism for knowledge representation capable of expressing a set of entities, their relationships, constraints and rules (conditional statements) of a given domain [16]. They are used by modern knowledge-based systems for representing and sharing knowledge about an application domain. These knowledge representation structures allow the semantic processing of information and, through more precise interpretation of data, systems have greater effectiveness and usability [13].

Manual construction of ontologies by domain experts and knowledge engineers is a costly task, thus automatic and/or semi-automatic approaches for their development are needed. Ontology Learning (OL) [4] [5] aims at identifying from textual information sources, the constituent elements of an ontology, such as non-taxonomic relationships.

Some techniques have already been proposed for Learning Non-Taxonomic Relationships of Ontologies (LNTRO) [9] [18] [20] [22] [26]. All of them use Natural Language Processing (NLP) techniques [1] [7] to annotate the corpus with the information needed for subsequent processing. Information Extraction (IE) techniques [12] are used to extract from the annotated corpus possible relationships; and Machine Learning (ML) [19] or Statistic Techniques (ST) to make refinements of the relationships outputted from the previous phases.

This article discusses the problem of LNTRO, identifying its phases and what kind of techniques can be used to perform the activities of each phase. Some techniques of the state of the art on LNTRO are also described and the advantages and limitations of the solutions they adopt for each phase of LNTRO are discussed.

The paper is organized as follows. Section 2 introduces an ontology definition. Section 3 defines the problem of LNTRO, its phases and what techniques can be used to approach each one. Section 4 describes some representative techniques of the state of the art on LNTRO and which solutions they adopt for each of its phases described in section 3. Finally, section 5 presents the conclusions discussing general and open research topics on LNTRO.

2 A Formal Definition of an Ontology

An ontology is a formal and explicit specification of a shared conceptualization of a domain of interest [15]. *Conceptualization* refers to an abstract model of some phenomenon in the world. *Explicit*, means that the type of concepts used and the limitations of their use are explicitly defined. *Formal*, refers to the fact that the ontology should be machine readable. *Shared*, reflects the notion that an ontology captures consensual knowledge, that is, it's not private to some individual but

accepted by a group. Currently, ontologies are applied in areas such as communication of software agents [11], integration of information [2], composition of Web Services [23], knowledge management [17], description of contents to support information retrieval from textual sources [14] [16], in Semantic Web applications [3] and knowledge-based systems [6].

Formally, an ontology can be represented by a 6-tuple [13]:

$$O = (C, H, I, R, P, A). \quad (\mathbf{1})$$

where,

$C = C^C \cup C^I$ is the set of entities of the ontology. The C^C set consists of classes, i.e., concepts that represent entities (for example, "Person" $\in C^C$) describing a set of objects, class instances in the C^I set (for example "Erik Brown" $\in C^I$).

$H = \{kind\_of(c_1,c_2) \mid c_1 \in C^C, c_2 \in C^C\}$ 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(Lawyer, Person) ".

$R = \{rel_k (c_1,c_2,..., c_n) \mid \forall i, c_i \in C^C\}$ is the set of non-taxonomic ontology relationships like "represents(Lawyer, Client) ".

$P = \{prop^C (c_k, datatype) \mid c_k \in C^C\}$ is the set of properties of ontology entities. The relationship $prop^C$ defines the basic datatype of a class property. For instance, subject(Case, String) is an example of a $prop^C$ property.

$I = \{is\_a (c_1,c_2) \mid c_1 \in C^I, c_2 \in C^C\} \cup \{prop^I(c_k,value) \mid c_k \in C^I\} \cup \{rel_k(c_1,c_2,..., c_n) \mid \forall i, c_i \in C^I\}$ is the set of instance relationships related to the C^C (eg. "is_a (Anne Smith, Client)"), P (eg. "subject (Case12, "adoption")") and R (eg. "represents(Erik Brown, Anne Smith)") sets.

$A = \{condition_x \Rightarrow conclusion_y(c_1,c_2,..., c_n) \mid \forall j, cj \in C^C\}$ 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 $condition_x = \{(cond_1,cond_2,\ldots,cond_n) \mid \forall z, cond_z \in H \cup I \cup R\}$. For instance, "∀Defense_Argument, OldCase, NewCase, applied_to(Defense_Argument, OldCase), similar_to (OldCase, NewCase) ⇒ applied_to(Defense_Argument,

NewCase)" is a rule that indicates that if two legal cases are similar then, the defense argument used in one case could be applied to the other one.

As an example, consider a very simple ontology describing the domain of a law firm (Figure 1), which has lawyers responsible for cases of the clients they serve.

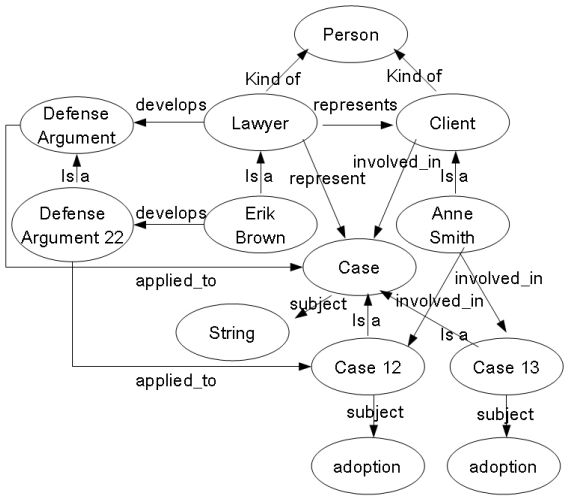

Fig. 1. Example of a simple ontology of a law firm.

According to the previous ontology definition, from the ontology in the Figure 1, the following sets can be identified.

C^C = {person, lawyer, client, case}.

C^I = {Erik Brown, Anne Smith, Case12, Case13, DefenseArgument22}.

H = {kind_of(Person, Lawyer), kind_of(Person, Client)}.

I = {is_a(Erik Brown, Lawyer), is_a(Anne Smith, Client), is_a(DefenseArgument22, DefenseArgument), is_a(Case12, Case), is_a(Case13, Case), subject(Case12, "adoption"), subject(Case13, "adoption")}.

R = {represents(Lawyer, Client), applied_to(DefenseArgument, Case), develops (Lawyer, Defense_Argument), involved_in(Client, Case)}.

P = {subject(Case, String)}.

A = {Defense_Argument, OldCase,NewCase, applied_to (Defense_Argument, OldCase), similar_to (OldCase, NewCase) ⇒ applied_to (Defense_Argument, NewCase)}.

3 The problem of LNTRO

LNTRO is an approach to automate or semi-automate the extraction of these relationships from textual information sources. Non-taxonomic relationships correspond to the R set of the ontology definition in section 2. For example, "solicit" is a non-taxonomic relationship between the classes "spouse" and "divorce" in the legal domain.

A generic process we are proposing for LNTRO [23] is illustrated in Figure 2 along with the process tasks, their sequence, task inputs and outputs and supporting techniques.

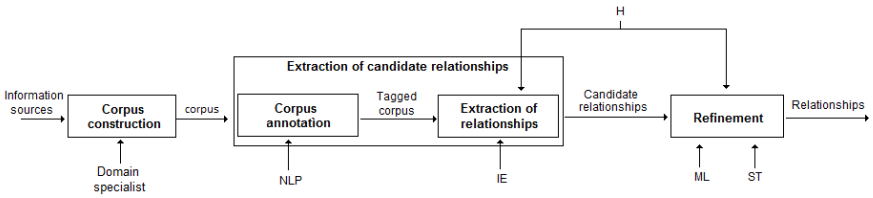

Fig. 2. A generic process for LNTRO.

The task of "Corpus construction" consists of selecting documents on the domain we expect to extract relationships from. This is usually a manually costly task and the outcome of any LNTRO technique depends on the quality of the used corpus.

The "Extraction of candidate relationships" task aims at identifying a set of possible relationships. It is composed of two sub-activities: "Corpus annotation" and "Extraction of relationships". The first subactivity consists of applying tags to the text using NLP techniques that are necessary for the next steps in LNTRO. The last searches in the annotated corpus for evidences suggesting the existence of relationships. For example, for Villaverde et al. [26], a relationship is identified by the

presence of two concepts of a given ontology in the same sentence with a verb between them. This sub-task can also receive the concepts of the ontology (the C set of the ontology definition in section 2) as input. In this case, there is a potential for achieving greater precision in the extraction of relationships.

Relationships outputted from the “Extraction of Relationships” task should not be recommended to the specialist, since there is usually a substantial amount of them that do not correspond to good suggestions. For this reason, Machine Learning (ML) or Statistic Techniques (ST) can be used in the "Refinement" phase. The ontology taxonomy (the set H of the ontology definition in section 2) can also be given as input. In this case, the corresponding LNTRO technique should be able to suggest to the specialist the best possible level in the hierarchy where to add the relationship. This functionality is explained with more detail in section 4.2. Table 1 summarizes the phases of LNTRO.

Table 1. Phases of the LNTRO generic process.

<table>
<tr><th colspan="2">Phases</th><th colspan="2">Description</th></tr>
<tr><td colspan="2">Corpus construction</td><td colspan="2">Selection of documents in quantity and quality required for LNTRO</td></tr>
<tr><td rowspan="2">Extraction of candidate relationships</td><td>Corpus annotation</td><td>Annotate the corpus using NLP techniques required for the continuity of LNTRO</td><td rowspan="2">Extraction of an initial set of relationships</td></tr>
<tr><td>Extraction of relationships</td><td>Application of the algorithm for extraction of relationships from the annotated corpus</td></tr>
<tr><td colspan="2">Refinement</td><td colspan="2">Application of ML or ST techniques to suggest the most probably relationships.</td></tr>
</table>

4 Techniques for LNTRO

In the following sections, some representative state of the art techniques for LNTRO are comparatively analyzed. The particular solutions adopted to approach the generic phases of LNTRO are highlighted and their positive aspects and limitations are discussed.

The technique based on the extraction of association rules [26], presented in section 4.1, applies NLP techniques to extract from text ontology concepts given as input and verb phrases located between the concepts to generate tuples of two concepts and a verb, that represent candidates relationships. These relationships are subjected to the algorithm for mining association rules [25] which suggest relationships in the form of association rules. These rules have the form X→Y, which means that the occurrence of X implies the occurrence of Y.

The technique based on mining generalized association rules [18], presented in section 4.2, differs from that of Villaverde et al. [26] mainly because of the use of the algorithm for mining generalized association rules [18] to refine the candidates and presents relationships in the form of association rules ($concept_1 \rightarrow concept_2$). This algorithm suggests the best level in the ontology taxonomy where to add the relationship. For example, in the case of a supermarket, the algorithm could suggest that "snacks are purchased together with drinks" rather than "chips are purchased with beer" and "peanuts are purchased with soda".

Section 4.3 presents a technique that retrieves non-taxonomic relationships based on statistics calculated upon the results of queries on a Web search engine [22]. From an initial keyword (eg. hypertension), representing an ontology concept, the technique suggests domain relationships (eg. *Hypertension* is caused by *hormonal problems*).

Section 4.4 presents Fader et al. [9] proposal which uses a logistic regression classifier [9] to rank non-taxonomic relationships according to the probability of being valid ones.

The last technique presented (section 4.5) [20] performs both learning and population of non-taxonomic relationships. It recommends non-taxonomic relationships from a corpus in English and infers from these new relationships and identify their instances.

4.1 LNTRO based on the Extraction of Association Rules

Villaverde et al. [24] proposed a technique for LNTRO with two phases: "Identification of occurrences of relationships" and "Mining associations". The first phase receives a corpus and a set of concepts of an ontology as inputs and outputs a set of tuples in the form $<c_1, v, c_2>$, where c_1 and c_2 are ontology concepts and v is a verb. Initially, to increase the recall of the search, each ontology concept is extended with its synonyms using Wordnet [10]. Then, in order to identify the verbs, the POS-tagging task is performed. For sentences that satisfy the following two conditions, a tuple $<c_1, v, c_2>$ is generated: (a) sentences that have exactly two concepts and a verb between them and (b) the two concepts are at a maximum distance of D terms. "D" is a parameter whose value is defined experimentally by the specialist and corresponds to the maximum number of terms that must exist between two concepts for them to be considered related. For example if D = 3 then, for the sentence "The court judged the custody in three days.", the tuple <court, judge, custody> is generated since there are two terms between the concepts.

Once a set of tuples outputted from the previous phase (candidate relationships) is obtained, the "Mining associations" task can be performed aiming at refining the results of the previous phase before suggesting relationships to the specialist. For this purpose, an algorithm for mining association rules [25] is used. This algorithm extracts rules of the form $X \rightarrow Y$, which means that the occurrence of X implies the occurrence of Y. A typical application is to extract from a database of sales transactions, rules representing the purchasing behavior of customers. For example <coffee, bread>→<butter> means that who purchase coffee and bread generally purchase butter. In the context of the present LNTRO technique, the extracted rules have the form $C \rightarrow v$, where 'C' denotes two concepts and 'v' the associating verb. Two thresholds are used to prune the rules: support and confidence. Support is the percentage of transactions containing all items that appear in the rule and is given by the formula: $Support(C \rightarrow v) = |\{t \in T | C \cup v \subseteq t\}| / |T|$, where "T" correspond to the set of transactions in the form $<c_1, v, c_2>$ from which the rules are extracted. Confidence corresponds to how one can trust the rule and is given by the formula: $confidence(C \rightarrow v) = support(C \rightarrow v) / support(C)$.

The product of this phase are non-taxonomic relationships represented by association rules in the form C→v, having values of support and confidence greater than the minimum defined experimentally by the specialist.

For example, in the sentence "Our data suggests that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation", taken from a corpus in the domain of medicine [21], Lipoxygenase (Li) and Reactive Oxygen Intermediates (ROI) are concepts and Activate (Ac) is a verb. In the first phase, the tuple <Li, ROI, Ac> is generated representing the fact that the two extraction conditions described previously were satisfied. In the second phase, if the rule <Li, ROI>→<Ac> has values of support and confidence greater than or equal to the minimum support and confidence, it is recommended to the specialist. Table 2 shows which solutions have been adopted for each one of the generic phases for LNTRO as defined in section 3.

A positive aspect of this proposal is that it labels with verbs the relationships between two concepts found in each sentence. In addition, ontology concepts are given as input to the technique, thus potentially leading to better results. Moreover, one restriction is the fact that no treatment is given to the possessive form "'s" that is one of the linguistic realizations of non-taxonomic relationships which can be present in the corpus with reasonable frequency. In addition, the authors refer to the verbs as single words when, in fact, they usually appear in the form of verb phrases. In Genia [21], the corpus used to illustrate and evaluate the technique, coincidentally most of the verb phrases are composed of a single term, which is a uncommon fact. Therefore, to be applied to corpora without this characteristic, the technique should be updated either to work with verb phrases or with the information of which verb, among those of the verb phrase, should be used as the label of the relationship.

Table 2. Solutions for LNTRO based on the Extraction of Association Rules.

Phase	Adopted solution
Corpus construction	Ad-hoc. A corpus already available in the medical field (Genia [21]) was used in its experiment.
Corpus annotation	POS-tagging
Extraction of relationships	Uses the algorithm already described to extract candidate relationships in the form of tuples ($<c_1, v, c_2>$)
Refinement	Uses a technique known as "Extraction of Association Rules" to suggest non-taxonomic relationships in the form of rules ($<c_1, c_2> \rightarrow <v>$)

4.2 LNTRO based on the Extraction of Generalized Association Rules

Maedech and Staab [18] propose a process to extract relationships from corpora in German composed of titles and text body. The technique consists of two phases: "Text processing" and "Mining associations". In the first phase, the objective is to extract pairs of concepts from the text that correspond to candidate relationships. For this purpose, the title and the sentence heuristics are used. The first one says that a pair of related concepts should be created for every concept in the text with every concept in the title. This heuristic is based on the intuition that the concepts that appear in the text body are related to the concepts that appear in the title. The second one sets up a tuple for each pair of concepts that are present in the same sentence.

In the second phase, the previous extracted relationships represented in the form of pairs of concepts are submitted to an algorithm for mining generalized association rules [18], a specialization of the algorithm for mining association rules [25], which goal is to extract non-taxonomic relationships in the form of association rules and suggest the best possible level in the hierarchy where to add the relationships. This algorithm [18] first extends each concept with its ancestors in the ontology taxonomy. Then, computes support and confidence for all possible association rules $X \rightarrow Y$ where Y does not contain an ancestor of X as this would be a trivially valid association. Finally, it prunes all those association rules $X \rightarrow Y$ that are subsumed by an "ancestral" rule $X_a \rightarrow Y_a$, the itemsets X_a, Y_a of which only contain ancestors or identical items of their corresponding itemset in $X \rightarrow Y$.

Table 3 shows some rules extracted from a corpus in the touristic domain [18]. The rule *area* → *hotel* is discarded because *area* → *accommodation* is an ancestral rule (its concepts are in the same or higher levels in the ontology taxonomy) and has support or confidence values greater or equal than the descendent rule. The same happens to the rules *room* → *television* and *room* → *furnishing*. The solutions adopted for each one of the generic phases of LNTRO are shown in Table 4.

Table 3. Relationships extracted with the generalized association rules algorithm [18].

Discovered relations	Confidence	Support
(area → accommodation)	0,38	0,04
~~(area → hotel)~~	~~0,1~~	~~0,03~~
(room → furnishing)	0,39	0,03
~~(room → television)~~	~~0,29~~	~~0,02~~
(accommodation → address)	0,34	0,05
(restaurant → accommodation)	0,33	0,02

Table 4. Solutions for LNTRO based on the Extraction of Generalized Association Rules technique.

Phase	Adopted solution
Corpus construction	Ad-hoc. A corpus already available in the touristic domain (Lonely Planet) was used in its experiment.
Corpus annotation	Uses chunking[1], stemming and NER[2].
Extraction of relationships	Uses sentence and title heuristics to extract candidate relationships as concept pairs ($CP = \{(a_{i,1}, a_{i,2}) \mid a_{i,j} \in C\}$).
Refinement	Uses a technique known as mining generalized association rules [18] to recommend relationships as association rules in the form ($c_1 \rightarrow c_2$).

A positive aspect of this proposal is the use of the algorithm for the extraction of generalized association rules that suggests the best possible level in the ontology taxonomy where the relationship should be added. Moreover since it uses NER it works with text with instances and not only with concepts like the solution of

[1] A NLP technique to tag noun and verb phrases.
[2] Named Entity Recognition

Villaverde et al. [26]. On the other hand, a limitation is the fact that the technique does not label the relationships but, only indicates what classes are related.

4.3 LNTRO based on Queries on Web Search Engines

This technique [22] is based on the premise that despite being diverse and unstructured, the redundancy of information in an environment as vast as the Web is a measure of its relevance and veracity. Initially, a keyword representative of the domain of what the specialist wants to learn the relationships have to be chosen. For example “hypertension” Then a Web search engine is used to retrieve pages related to it. Based on morphological and syntactic analysis, verbs that have a relationship with the keyword are extracted. Then, the degree of similarity between each verb and the domain keyword is measured. To do so, statistical measures are calculated about the term distribution on the web. The obtained values are used to rank the list of candidate verbs. This lets one choose the labels of non-taxonomic relationship that are closely related to the domain. The domain related verbs are used to discover non-taxonomic related concepts. To do so it queries the web with the patterns "Domain-keyword verb" or "Verb domain-keyword" that returns a corpus related to the specified query. The goal is to search 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 patterns. These concepts are candidate to be non-taxonomically related to the original keyword. This process is cyclic executed as shown in Figure 3. Table 5 shows which solutions have been adopted for each one of the generic phases for LNTRO as defined in section 3.

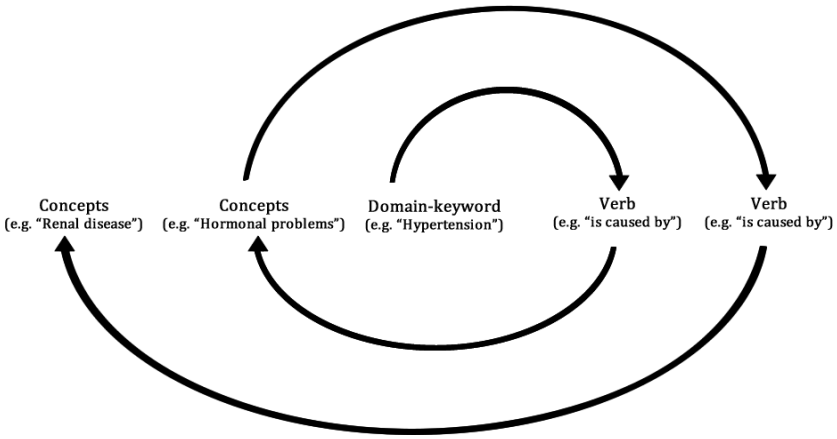

Fig. 3. Cyclic execution of LNTRO based on Queries on Web Search Engines.

Table 5. Solutions for LNTRO based on Queries on Web Search Engines.

Phase	Adopted solution
Corpus construction	Based on documents returned by a Web search engine.
Corpus annotation	Chunking.
Extraction of relationships	Extracts verb phrases as labels and noun phrases as concepts of non-taxonomic relationships.
Refinement	Statistical processing based on the result of queries in a web search engine.

A positive aspect in this proposal is that specialists do not have to deal with the construction or selection of corpora, a usually laborious task. They are automatically created with the help of a web search engine. In addition, the process is fully automatic. On the other hand, one limitation is that learning relationships is dependent of learning concepts.

4.4 LNTRO based on logistic regression

Fader, Soderland and Etzioni [9] propose a technique that is domain independent and extracts non-taxonomic relationships from corpora in English. It uses a syntactic and a lexical constraint.

The syntactic constraint requires that verb phrases match the following patterns: a verb (e.g. "invented"), a verb immediately followed by a preposition (e.g.

"located in"), or a verb followed by nouns, adjectives or adverbs ending with a preposition (e.g. "has atomic weight of"). The syntactic constraint reduces "uninformative" extractions, for example, for the sentence "Faust made a deal with the Devil" the tuple <Faust, made, devil> corresponds to a non informative extraction. The relationship extracted using the syntactic patterns would be <Faust, made a deal with, devil>, which is a valid relationship. However, it allows the extraction of relationships considered too "specific". As an example, let us consider the sentence "The Obama administration is offering only modest greenhouse gas reduction targets at the conference". The syntactic patterns will match the phrase "is offering only modest greenhouse gas reduction targets at". Thus, there are phrases that satisfy the syntax constraint, but are not relationships. To overcome this limitation the lexical constraint is used to separate sentences that represent real relationships from those very specific ones, such as the example sentence. The restriction is based on the intuition that a valid relational sentence must have many different arguments in a large corpus. The example sentence is specific to the pair of arguments "Obama administration" and "conference", so it is unlikely to represent a relationship. The lexical restriction is implemented by a repository of verb phrases that are considered sufficiently generic (have many different arguments). The repository is manually built and, whenever a verb phrase meets any of the syntactic patterns, it is checked against it. Verb phrases not present in the repository are not recommended as relationships. The technique has three phases as follows. The phases of "Extraction of relationships" and "Extraction of arguments" have high recall but low precision. Thus a refinement is required in order to reveal the most probable relationships among all extracted from the application of the syntactic and lexical constraints.

- Extraction of relationships: For each verb "v" in a sentence "s", find the longest sequence of words "r" such that (a) "r" is initiated by "v", (b) "r" satisfies the syntactic constraint and (c) "r" satisfies the lexical constraint.
- Extraction of arguments: For each verb phrase "r" identified in the previous step, find the noun phrase "x" closer to the left of "r" in the sentence such that "x" is not a relative pronoun, adverb "Who" or existential "there". Find the

noun phrase "y" closer to the right of "r" in "s". If a pair (x, y) has been found, return (x, r, y) as an extracted relationship.

- Refinement: In this phase a logistic regression classifier [19] is used to rank relationships according to the probability of representing valid relationships. Table 6 shows which solutions have been adopted for each one of the generic phases of LNTRO as defined in section 3.

Table 6. Solutions for LNTRO based on Logistic Regression.

Phase	Adopted solution
Corpus construction	Ad-hoc construction.
Corpus annotation	Chunking.
Extraction of relationships	Uses the algorithms described in the phases "Relations extraction" and "Extraction of arguments".
Refinement	Uses a logistic regression classifier to assign a probability to each relationship.

The technique extracts relationships with labels in the form of verb phrases that comply with the syntactic and lexical constraints. Moreover, the technique is capable of extracting relationships from very small corpus, such as a single sentence and from any area of knowledge thus, the technique is domain independent.

A limitation of this approach is that it uses a manually built repository of verb phrases, containing those that are considered sufficiently generic to be present in sentences relating various concepts. If a relationship potentially "valid" is represented by a verb phrase that is not in the repository, it will be discarded.

4.5 LNTRO based on the classification of relationships

Mohamed, Hruschka and Mitchell [20] propose a technique that performs both learning and population of non-taxonomic relationships. This technique recommends non-taxonomic relationships from a corpus in English and infers from these new relationships and identify their instances. This technique has as inputs the concepts of an ontology, a list of instances associated with each of them and a corpus and outputs: a set of non-taxonomic relationships represented by three elements, two ontology concepts and a label (eg RiverFlowsThroughCity (<River>, <City>)); for each relationship a set of instances (eg RiverFlowsThroughCity (<Nile>, <Cairo>))

and for each relationship extracted from the corpus a set of lexical patterns is generated for extracting new instances of this relationship. For example, "X in the heart of Y" with which the relationship between X and Y could be identified in the sentence "Thames in the heart of London". In the phase of "Preprocessing", for each pair of ontology concepts a set S is created consists of all sentences containing known instances of both concepts. In phase "Generation of relationships" a matrix of co-occurrence of contexts is generated for each pair of ontology concepts (Figure 4). In this matrix each cell corresponds to the number of instances of concept pairs with which both contexts co-occur. For example, for the sentences "Vioxx can cure Arthritis" and "Vioxx is a treatment for Arthritis" the contexts "can cure" and "is a treatment for" co-occur with a pair of instances "Vioxx" and "Arthritis." For example, consider that the preprocessing, for the pair of concepts <drug, disease> has obtained 122 contexts. Contexts such as "to tread", "for treatment of" and "medication", that indicate the same relationship (drug-to treat-disease) has high values of co-occurrence (Figure 4). The same happen for the contexts "can cause", "may cause" and "can lead to" that indicate the relationship "drug-can cause-disease".

Contexts / Contexts	may cause	can cause	can lead to	to treat	for treatment of	medication
may cause	0.176	0.074	0.030	0.015	0.011	0.000
can cause	0.051	0.150	0.039	0.018	0.013	0.010
can lead to	0.034	0.064	0.189	0.019	0.021	0.018
to treat	0.006	0.011	0.007	0.109	0.043	0.015
for treatment of	0.005	0.008	0.008	0.045	0.086	0.023
medication	0.000	0.011	0.009	0.030	0.036	0.111

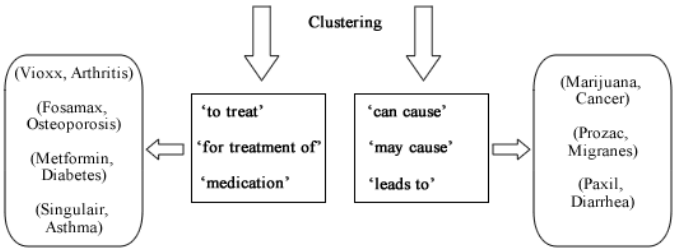

Fig. 4. Context by context sub-matrix (with six contexts) for the pair of concepts <Drug, Disease>.

Based on the values of co-occurrence taken from the matrix, the contexts are clustered. Each cluster is then used to propose a possible new relationship. The centroid of each cluster is used to suggest the name of the new relationship. For example, if the centroid of a cluster is "for treatment of" then the relationship name is "drug-for-treatment-of-disease". Then initial instances (seed instances) are generated for the new relationships. Instances of relationships (pairs of concepts) that correspond to the centroid or are close to it are the most representative of the relationship. For each seed instance of a relationship the formula 2 is calculated.

$$\sum_{C \in PatternCluster} Occ\ (c, s) / (1+sd(c)) \qquad (\mathbf{2})$$

"Pattern cluster" is the cluster of context patterns for the considered relationship. Occ(c,s) is the number of times that the instance of the relationship (s) co-occurs with the context pattern (c). Sd(c) is the standard deviation of the context (c) in relation to the centroid of its cluster. The instances are ranked by this measure and the first 50 are considered the seed instances of the proposed relationship.

Many relationships extracted from this phase do not correspond to valid ones. For this reason, in the phase "Classifying semantically valid relationships" some heuristics are used as criteria their classification. One of these concern about how specific a context pattern is in relation to a given relationship. For example, consider $<c_1, c_2>$ a pair of ontology concepts. If the same context connects instances of c_1 to a large number of instances other then c_2 then this context should not indicate a valid relationship. Table 7 shows which solutions have been adopted for each one of the generic phases of LNTRO as defined in section 3.

This technique is automatic and was developed in order to continuously extract relationships from the Web [20]. For this reason it is not suitable for the development of ontologies for specific situations such as the development of a ontology for a knowledge system in the legal domain.

Table 7. Solutions for LNTRO based on the classification of relationships.

Phase	Adopted solution
Corpus construction	Not approached
Corpus annotation	Tokenization, sentence splitter and NER.
Extraction of relationships	Identification of relationships in two ways: 1-Extraction of tuples (c_1, c_2) from the corpus. 2-Generation of new tuples from clustering context patterns as described in the technique.
Refinement	Uses a classifier based on some heuristics [20] to recommend only relationships considered valid.

5 Concluding Remarks

LNTRO techniques as well as any other in the area of ontology learning are subject to a great amount of noise because the source from which information is extracted is unstructured. Thus highly customizable solutions are needed for these techniques to be applied to the widest possible range of situations. As briefly discussed below LNTRO techniques proposed so far do not satisfy this requirement.

Villaverde et al. [26] implement the sentence extraction rule which says that two concepts are related if they are in the same sentence and there is a verb between them. However, relationships can be represented in other forms in the text and can be retrieved with rules not covered by this technique. An example is the possessive contract form rule, which says that if two concepts are joined by a possessive contract form ('s) they have a good chance to be related. An advantage of this rule is that despite being rarer, it has a higher degree of accuracy in comparison to the sentence rule.

Maedech and Staab [18] use the title rule, which creates a pair of concepts representing a relationship for each concept in the title with each concept in the text body. This strategy usually generates a great amount of candidate relationships, which is adequate if a statistical solution is used in the subsequent refinement phase, however this rule limits the texts from which relationships can be extracted to those that have titles and text body.

Sanchez and Moreno [22] extract relationships between any noun phrases within a sentence, once in their approach, ontology concepts are not known a priori. However informing the concepts reduces the search space for relationships and has the potential to lead to better results.

Reverb [9] extracts a relationship for each occurrence of a verb phrase between two noun phrases in a sentence. It has the advantage to give a label to the extracted relationship. However, unlike Villaverde et al. [26], concepts cannot be given as input to the technique, even if they are available. Furthermore, other extraction rules like the one implemented by Maedech [18], which has a greater recall, and the possessive contract, which has a better precision, are ignored.

Mohamed et al. [20] proposal is automatic and was developed in order to continuously extract relationships from the Web. For this reason it is not suitable for the development of ontologies for specific situations such as the development of an ontology for a knowledge system in the legal domain.

We are currently working on a framework for LNTRO that may help ontology developers to gain more efficiency in this task, because of its higher level of parametrization and more adequate solutions, namely the control over the execution of its extraction rules; the "apostrophe rule", for the phase of "Extraction of relationships" and the statistical solution "bag of labels" for the Refinement phase, explained in the next. Its phase of "Corpus annotation" can be customized based on the extraction rules that are selected to be executed in the next phase. Tokenization, sentence splitter and lemmatization are prerequisites to the sentence and apostrophe rules, while chunking is necessary for the sentence rule with verb phrase.

In the "Extraction of relationships" phase three extraction rules are provided. The sentence rule (SR) is based on the intuition that two consecutive concepts in the same sentence are probably non-taxonomically related. The sentence rule with verb phrase (SRVP) is based on the intuition that two consecutives concepts in the same sentence with a verb phrase in between are probably non-taxonomically related. This rule tends to present lower recall than SR because, for example, it does not extracts the tuple <marriage, spouse> from the sentence "The date and place of any previous marriages of either spouse as well as the date, place and circumstances under which they were terminated don't interfere in the present one.", which correspond to a valid

relationship. However, SRVP tends to have higher precision. The apostrophe rule (AR) is based on the intuition that two consecutive concepts with either strings "'s" or "'" in between have great probability of being non-taxonomically related. For example, for the sentence "While the court will generally honor the parties' agreements as set forth in the separation agreement," the extracted tuple would be <party, agreement>. The motivation for the definition of the AR is that it gives a differentiated treatment to extractions that have the highest probability of being non-taxonomic relationships.

In the "Refinement" phase the solution provided by the framework to refine the extractions made with SRVP is the statistical algorithm "bag of labels". Its general idea is to calculate the frequency of pairs of concepts (<c_1, c_2>), independent of their order, and store the corresponding verb phrases in its bag of labels. The specialist set the pruning parameter minimum frequency and chooses the most appropriate verbal label for each relationship. Finally the specialist decides which relationships should be actually added to the ontology. He/she can also re-execute the framework changing the solutions adopted for each phase and the execution parameters.

Initial evaluation still to be published showed that our framework presented better results in terms of recall and precision compared to the proposal of Villaverde et al. [26] in LNTRO from the corpus Genia [21].

Acknowledgments

This work is supported by CNPq, CAPES and FAPEMA, research funding agencies of the Brazilian government.

References

1. Allen, J.: Natural Language Understanding. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc. (1995)
2. Alexiev, V., Breu, M., de Bruijn, J., Fensel, D., Lara, R., Lausen, H.: Information Integration with Ontologies: Experiences from an Industrial Showcase, Wiley (2005)
3. Bontcheva, K., Cunningham, H.: The Semantic Web: A New Opportunity and Challenge for Human Language Technology, In Proceedings of the Workshop on Human Language

Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference, Sanibel Island (2003)
4. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: An Overview. DFKI, Language Technology Lab. AIFB, University of Karlsruhe. ITC-irst (2003)
5. Buitelaar, P., Cimiano, P., Magnini, P.: Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands (2006)
6. Cimiano, P., Volker, J., Studer, R.: Ontologies on Demand? – A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. In Information, Wissenschaft und Praxis 57 (6-7): 315-320 (2006)
7. Dale, R., Moisl, H., Somers, H. L.: Handbook of natural language processing. CRC (2000)
8. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In: Proceedings of the 5th International Semantic Web Conference. p. 228 – 241, Athens. Springer (2006)
9. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland (2011)
10. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge: MIT Press. 23-24 p. (1998)
11. Finin, T., Fritzson, R., McKay, D., McEntire, R.: KQML as an Agent Communication Language. In: Proceedings of the 3rd International Conference on Information and Knowledge management, p. 456-463 (1994)
12. Freitag, D.: Information extraction from HTML: Application of a general machine learning approach. In Proceedings of the 15th Conference on Artificial Intelligence. pp. 517-523, (1998)
13. Girardi, R.: Guiding Ontology Learning and Population by Knowledge System Goals. In: Proceedings of International Conference on Knowledge Engineering and Ontology Development, Ed. INSTIIC, Valence, pp. 480 – 484 (2010)
14. Girardi, R., Ibrahim, B.: Using English to retrieve software. Journal Of Systems Software: Special Issue on Software Reusability, New York, Elsevier. v. 30, n. 3, p. 249 - 270, (1995).
15. Gruber, T. R.: Toward Principles for the Design of Ontologies used for Knowledge Sharing, International Journal of Human-Computer Studies. Nº 43, pp. 907-928 (1995)
16. Guarino, N., Masolo, C., Vetere, C.: Ontoseek: Content-based Access to the web. IEEE Intelligent Systems, vol. 14 (3), pp. 70-80 (1999)
17. Jurisica, I., Mylopoulos, J., Yu, E.: Using ontologies for knowledge management: an information systems perspective. In: Knowledge: creation, organization and use – Proceedings of the 62nd annual meeting of the American Society for Information Science, Washington, DC, pp 482–496 (1999)
18. Maedche, A., Staab, S.: Mining non-taxonomic conceptual relations from text. In Knowledge Engineering and Knowledge Management. Methods, Models and Tools: 12th International Conference. Proceedings. 189-202. Berlin: Springer (2000)
19. Mitchell, T.: Machine Learning, McGraw Hill. (1997)
20. Mohamed, T. P., Junior, E. R. H., Mitchell, T. M.: Discovering Relations between Noun Categories. In Proceedings of the conference on empirical methods in natural language processing (EMNLP 2011), Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1447-1455 (2011)
21. Rinaldi, F. et al.: Mining relations in the GENIA corpus. Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics: 61 - 68. (2004)
22. Sanchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. Data and Knowledge Engineering, 64(3), p. 600-623 (2008)

23. Serra, I., Girardi, R., Novais. P.: The Problem of Learning Non-taxonomic Relationships of Ontologies from Text. In proceedings of the 9th Conference on Distributed Computing and Artificial Intelligence, Salamanca, Spain, pp. 485-492 (2012)
24. Sirin, E., Hendler, J., Parsia, B.: Semi-automatic composition of web services using semantic descriptions. In: Proceedings of the ICEIS Workshop on Web Services: Modeling, Architecture and Infrastructure (2002)
25. Srikant, R., Agrawal, R.: Mining generalized association rules. In Proc. of VLDB' 95, pages 407-419 (1995).
26. Villaverde, J., Persson, A., Godoy, D., Amandi, A.: Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Syst. Appl. 36(7): 10288-10294 (2009)