Discovery of relation axioms from the Web

Luis Del Vasto Terrientes, Antonio Moreno, David Sánchez

Universitat Rovira i Virgili. Departament d'Enginyeria Informàtica i Matemàtiques
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA)
Av Països Catalans, 26. 43007 Tarragona, Catalonia (Spain)
{luismiguel.delvasto, antonio.moreno, david.sanchez}@urv.cat

Abstract. Given the proven usefulness of ontologies in many areas, the representation of logical axioms associated to ontological concepts and relations has become an important task in order to create an expressive representation of domain knowledge. Manual inclusion of logical axioms into an ontology can be a harsh, time consuming task. As a result, very few ontologies include axioms in their formal definition. From the ontology learning point of view, axiom learning is one of the less tackled and unexplored problems. In this paper we introduce a preliminary methodology to learn axioms associated to ontological relationships in an automatic and unsupervised way using the Web as corpus.

Keywords: Ontologies, Axioms, Knowledge Discovery, Semantic Web

1 Introduction

An *ontology* represents the conceptual model underlying a certain domain, describing it in a declarative fashion and thus cleanly separating it from procedural aspects [2]. Ontologies play a key role in the Semantic Web [3] and other areas like agent communication, intelligent information integration and knowledge-based systems. Research in this area has increased considerably during the last years, bringing the development of modern ontological languages such as OWL 2, which provides a powerful formalism for knowledge representation and reasoning.

Ontologies are typically built by hand, requiring an arduous and time-consuming effort both from knowledge engineers and domain experts. As a result of the bottleneck introduced by knowledge acquisition approaches, manually composed ontologies are, in many cases, flat and lightweight, with very limited expressivity and mainly considering taxonomical knowledge and simple relations [4][5]. So, available ontologies very rarely exploit the expressiveness of languages like OWL 2 in defining logical constraints (axioms) associated to relations.

Ontology Learning methods, that extract automatically ontological elements from available corpora and build ontologies from them, mostly focus on taxonomic relationships [5] and, more rarely, on non-taxonomic knowledge [14]. Relation axioms involving logical connectives, role restrictions and other semantic features that can provide expressivity to ontologies remain largely unexplored [4].

This paper presents an automatic axiom learning algorithm which, starting from a set of non-taxonomic relations (denoted *xRy* or *R(x,y)*, in which *x* and *y* represent concepts or individuals and *R* the relation/property) defined in an input ontology, explores and verifies which axioms those relations fulfill. Specifically, the axioms studied in our approach are *symmetry*, *reflexivity*, *functionality*, *transitivity* and *inverse*. Our approach relies on linguistic techniques and statistical analyses to unsupervisedly and automatically extract and evaluate semantic evidences found in Web resources [1], which support the definition of ontological axioms.

The paper is structured as follows. Section 2 presents the related works dealing with automatic ontology axiomatization. Section 3 describes our approach in detail, explaining the different stages of the learning process and the techniques used. Section 4 explains the criteria and measures considered to evaluate the algorithm and provides some preliminary results. Finally, the last section presents conclusions and some lines of future research.

2 Related Work

Several natural language processes have been applied to acquire OWL axioms from lexical resources in tools such as LExO (that stands for Learning Expressive Ontologies), RELExO (Relational Exploration) and HASTI, among others.

The implementation of LExO basically relies on KAON2, an ontology management infrastructure for OWL, and the Minipar dependency parser [8]. Given a natural language definition of a class, LExO starts by analyzing the syntactic structure of the input sentence. The resulting dependency tree is then transformed into a set of OWL axioms (concept inclusion, transitivity, role inclusion, role assertions, concept assertions and individual equalities) by means of manually engineered transformation rules [4]. The axiomatization is a semi-automatic cyclic process [8], in which each iteration contains an ontology evaluation step to make sure that the axiom candidates are correct. The main problem is the semantic inconsistencies caused by the analysis of natural language expressions. To tackle this problem the relational exploration [15] method was used in the development of the RELExO application. This technique is an adaptation of attribute exploration from Formal Concept Analysis (FCA) to description logics, and is a good means to overcome the lack of completeness and precision in learned ontologies. It is natural to apply it to further specify the knowledge beyond the information extractable from the corpus, which makes relational exploration a perfect complement to automatic approaches for ontology generation based on lexical resources [7]. Another automated attempt was performed in [11], in which lexico-syntactic patterns are used to extract individuals and subclasses, and the OntoClean methodology is used to obtain evidence of strict disjointness between two classes. Pattern-extracted classes are analyzed and, if between two classes there aren't any common individuals or subclasses that taxonomically overlap, they are taken to be disjoint. In the OntoClean methodology, a pattern-based approach was used to analyze text taken from the Web specifically to verify that two classes have incompatible unity or identity criteria. Other patterns based in enumerations, defined in [9] for detecting disjointness, were not precise enough in comparison with average human disjointness detection.

HASTI is a system for automatic ontology construction, based on Persian (Farsi) text understanding. It uses an initial kernel, which has the essential metaknowledge (primitive concepts and operators) to build an ontology including axioms [6]. HASTI implements a rule base which contains linguistic rules, inference rules and heuristics to extract lexical elements from texts. Specifically, it contains Semantic Templates to extract taxonomic relations, non-taxonomic relations and axioms. It extracts axioms (disjointness, transitivity and functionality) from conditional and quantified sentences, using axiom templates written in Knowledge Interchange Format (KIF). Following the lexico-syntactic approach, AuCONTRAIRE is a Contradiction Detection (CD) system which relies on the TextRunner system, an open Information Extraction system that uses the Web as corpus to obtain relations between entities [10]. In order to detect functional relations, it checks if a domain and its relation map to a unique variable using probabilistic assessment based on aggregated textual evidence.

The lack of automatic and unsupervised approaches to axiomatize ontologies motivated us to aim our efforts into this area. Some of these works include natural language patterns in order to determine axioms. In LExO, RELExO and the disjointness detection system there are rules that transform a set of patterns, determined by a particular natural language syntax, into axioms. For example, the Disjunction Rule is defined as a pattern “NP_0 OR NP_1”, in which NP represents a noun phrase. If the text “Mazda or Toyota” is found, both brand concepts are set as disjoint. In our approach, the text patterns will not be only used to detect natural language structures from a text, but also to create Web search queries to extend the knowledge by using the biggest available repository. In this sense, on the contrary to these related works which are applied in a narrow context (i.e. a predefined document or even a sentence), our approach uses the Web as a massive learning corpus, enabling the discovery of relation axioms in a domain-independent fashion. On the other hand, HASTI uses structures of conditional sentences and compound sentences with a quantifier phrase to detect axioms, which may be a barrier to obtain complete information about the behaviour of the concepts and roles. Our proposed approach checks for evidences based only on the text pattern to analyse, avoiding the study of the sentence structure, taking as a matter of fact that logical axioms are not only stated on this type of sentences. AuCONTRAIRE uses text patterns and the Web as corpus, but its approach is based on detecting only contradictions by detecting functional relations. Unlike approaches such as LExO and RELExO, our approach does not require manual stages which, due to the Web size, would compromise its scalability.

3 Object Property Axioms Learning Methodology

Object properties define relations between pairs of ontological concepts. In contrast to hierarchical relations, an object property represents a non-taxonomic relation, which is typically expressed by a verb that relates a pair of concepts [12]. The definition of *object property axioms* allows the meaning of properties to be enriched and, therefore, the possibility to apply inference on them.

The pseudocode of the algorithm we propose is presented below, in which a domain ontology is received as input and a set of axioms associated to ontological relationships defined via object properties is obtained as output. First, the non-taxonomic relations from this ontology are read. For each relation, each potential axiom is analysed, generating a Web search query depending on a pattern create to analyse the axiom. The documents retrieved from the Web search query are analysed in order to find a match with the proposed pattern. Each match provides us with a term, which is a candidate to satisfy a given relationship (e.g. a candidate to be an inverse relation). Every candidate is evaluated, using Web-scale statistics, to assess if it provides enough evidence to guarantee the fulfilment of the axiom. All these steps are explained in detail in the following subsections. During the explanation of the method, we refer to algorithm lines as <*line*>.

```
1.  AxiomLearning (Domain Ontology DO)
2.  { AXS_OP /* Results variable */
3.    AXIOMS /* Constant holding the set of axioms of interest */
4.    THRESHOLD /* Constant holding the selection threshold */
5.  RELS ← get_nontaxonomic_relations(DO)
6.  /* for each relation, analyze each object property axiom */
7.  for all RELS_i ∈ RELS
8.  {   for all AXIOMS_j ∈ AXIOMS
9.        {  QUERY ← generate_query(RELS_i, AXIOMS_j)
10.          WEBPAGES ← retrieve_WebPages(QUERY)
11.          for all WEBPAGES_x ∈ WEBPAGES
12.          {  /* for each web page check for matchings */
13.              TEXT ← extract_matchings (QUERY, WEBPAGES_x)
14.              TEXT ← remove_stopwords(TEXT)
15.              KEY ← stem_word(TEXT)
16.              exists ← check_existence(AXS_OP, RELS_i, AXIOMS_j, KEY)
17.              if (!exists)
18.              {  /* calculate web scale statistics */
19.                  SCP = calculate_SCP(AXIOMS_j, RELS_i, TEXT)
20.                  /* add info. to result if SCP > THRESHOLD */
21.                  if (SCP > THRESHOLD)
22.                  AXS _OP ← AXS_OP + (RELS_i ∪ AXIOMS_j ∪ KEY ∪ TEXT ∪ SCP)
23.              }
24.          }
25.      }
26.  }
27. Return AXS_OP
28. }
```

3.1 Pattern construction

Lexico-syntactic patterns, as tools to discover and extract ontological entities, have been used in order to learn ontologies from unstructured text. Our approach bases the axiom learning in the exploitation of specially designed patterns aimed to retrieve (by means of querying a Web search engine) and extract (by means of a linguistic analysis) semantic evidences that support the fulfillment or not of each one of the studied Object Property Axioms. The patterns proposed below are domain-independent and can be adjusted to queries in many Web search engines. In base of these patterns, we have designed queries for the Yahoo! and Bing Web search engines <9>.

To describe the proposed patterns we use the nomenclature *Relation(Subject, Object*) in which *Relation* represents a relation and *Subject/Object* represent individuals. As an example, in a non-taxonomic relation defined as Borders(Spain, France), Spain is the *Subject*, France is the *Object* and Borders represents the *Relation*. In the following paragraphs the proposed patterns and their use to check some properties are described.

Relation (Subject, ?). This relation looks for possible *OBJECTS* for an individual *SUBJECT* and a *RELATION*.

Functional Object Property & Inverse Functional Object Property: Given a relation *xRy*, it is possible to detect if the subject *x* can be related to another object different from *y* by the relation *R* to check the *Functionality* of the Object Property *xRy* and the *Inverse Functionality* of the Object Property *xR'y (R´*being the inverse of *R).* In the query language we denote this pattern as ["x R" –"x R y"], in which ["*x R*"] is used to extract new objects while [–"x R y"] avoids the extraction of the initial object *y* for the relation.

Transitive Object Property: Given a relation *xRy*, to detect the *Transitivity* of the property for the subject *x*, we need first to detect if the object *y* can be related to another individual *z* by relation *R*. In this way we can detect a relation *yRz* in which the relation *R* is the same for both *xRy* and *yRz*. In query language we denote this pattern as ["y R" –"y R x"]. ["*y R*"] may help us to detect if *y* can be related to an object *z* by relation *R*, while [–"*y R x*"] avoids the extraction of the original subject *x* in *xRy* as object.

Relation (Subject, Object). We use this pattern to check if a particular relation R(x,y) holds.

Symmetric Object Property: Given a relation *xRy*, by switching the positions of the subject *x* and the object *y* we can check if the relation *R* can be applied from *y* to *x* as well, denoted in a web search query as ["*y R x*"].

Reflexive Object Property: Given a relation *xRy*, to check its reflexivity we used as object the reflexive pronoun 'itself', which is used to refer to things or animals. In query language we denote this pattern as "x R itself".

Transitive Object Property: As said in the previous pattern Relation (Subject, ?), we can detect *yRz* from an initial relation *xRy*. The second step uses this pattern in order to check that the subject in *xRy* can be related to the object in *yRz*. In query language we denote this pattern as ["x R z"]. In basic terms, we used this pattern to detect if its is possible to find evidence in the Web of a concrete relationship between two specific individuals.

? (Object, Subject): Given two individuals x and y, with this pattern we try to find verb phrases that represent relationships between them.

Inverse Object Property: Given a relation *xRy*, we can search for an inverse relationship by looking for a verb that relates *y* to *x*. In query language we denoted this pattern as ["y * x" OR "y * * x" OR "y * * * x"]. Note that the wildcard (*), which is used in the query in order to force the proximity (up to three words) between the subject and the object, has to correspond to a verb phrase.

3.2 Extraction

Applying the described patterns over a given relation and querying the resulting expression over a Web search engine, we are able to retrieve Web resources including one or several matchings <10>. The text is analysed in order to discover and extract the semantic evidences that would lead to the definition or not of the corresponding axiom. In order to minimize noise and ambiguity of textual processing, we imposed some syntactic boundaries. In general, we decided to get rid of interrogative sentences, cardinal candidates, candidates that are superclasses of the initial object, and verbs in future or gerund tense. Words between parentheses are deleted. If a text piece fulfills the conditions and match with the particular pattern, then the evidence is extracted <13> for further analysis.

3.3 Evaluating extractions and axiom definition

Applying the mentioned constraints to avoid natural language problems does not mean that the observations extracted for a certain relation are strong enough to grant the definition of an axiom. The next step consists on deciding which of the extractions are reliable and sufficiently related to the property. To perform this selection process we performed a Web-based statistical analysis relying on co-occurrence measures computed directly from Web search engines <19>. Co-occurrence measures are based on distributional hypothesis claiming that words that occur in the same context tend to have similar meanings.

Several scores have been proposed in the past to compute Web-scale statistics, adapting the notion of collocation and mutual information (computed as the probability of joint appearance of concepts in a corpus). Applied to the Web, probabilities (*p*) are estimated as the hit count (*hits*) returned by a Web search engine when querying the term. We used the Symmetric Conditional Probability (SCP) [13] computed from the Web (**1**) to evaluate the relatedness between two entities.

$$\mathrm{SCP}(x,y) = \frac{p(x,y)^2}{p(x)*p(y)} \approx \frac{\left(\frac{\mathrm{hits}("x"\ \mathrm{AND}\ "y")}{\mathrm{total\_web\_sites}}\right)^2}{\frac{\mathrm{hits}("x")}{\mathrm{total\_web\_sites}} * \frac{\mathrm{hits}("y")}{\mathrm{total\_web\_sites}}} = \frac{\mathrm{hits}("x"\ \mathrm{AND}\ "y")^2}{\mathrm{hits}("x")*\mathrm{hits}("y")} \quad \textbf{(1)}$$

To evaluate the relevance of the extracted terms, we need to measure the co-ocurrence of the relation and the extracted evidence. In order to construct queries that approximate term appearance probabilities, we consider a *discriminator phrase*, which represents the xRy relation. We call *instance* to the expression created with the extracted term plus the relation that links it to another individual, and *discriminator phrase without instance* to the expression that represents the relation without the extracted *instance*. So, when evaluating if the relation *xRy* “Spain borders France” is functional when a new object “Portugal” is retrieved, we can denote from (**1**):

p(x,y) - *Discriminator Phrase*: “Spain” AND “borders” AND “Portugal”

p(x) - *Instance*: “borders” AND “Portugal”

p(y) - *Discriminator Phrase without Instance*: “Spain” AND “borders”

The Web-based statistical evaluation process for the reflexive, symmetric and transitive (in its second step) properties implies the assessment of the relative relatedness (computed by means of the SCP) of the extracted evidences with respect to the initial relation. If the SCP of the extracted evidence is higher than a threshold, we may conclude that the relation is reflexive, symmetric or transitive (in its second step) <21, 22>. For the functional, inverse, inverse functional and transitive (in its first step) object properties, if the set of extractions contains an item with a co-occurrence higher than the threshold then we may conclude that the extractions are robust for the relation. Depending on these observations, an axiom may be fulfilled or not. In the functional and inverse functional case, if the extractions are robust then we may conclude that the relation is not functional or inverse functional <26, 27>. For the inverse axiom, we may conclude that an inverse relation has been found. In the first step of the transitivity test, we conclude that a good candidate to check for the transitivity in the second step has been found <21,22>.

4 Empirical evaluation

Due to the difficulty to carry out automatic evaluations of automatically acquired knowledge, we based them in a manual approach in which a human determines if a relation fulfills a certain axiom. A number of positive and negative relation examples of each of the studied axioms are presented, along with the results given by our methodology. In the tables shown in this section the following abbreviations are used:

- SCP O/P: SCP values of the original relation and those obtained for the best extraction from the constructed pattern.
- D: Decision of the algorithm concerning the evaluated axiom relation.
- HD: Decision of a human concerning the evaluated axiom relation.
- NC: Number of extractions for a given axiom and relation.
- BES: Best extraction selected (the one with the highest SCP value), when applicable.

The examples used to evaluate the properties have been selected from different domains, in order to show the genericity of the proposed methodology. In all cases the threshold used to select or discard an extraction has been empirically set to 1E-3. We used two Web search engines and the maximum number of Web pages retrieved for each pattern has been set to 500 per Web search engine.

Symmetric Object Property. The evaluation of the Symmetry of an Object Property has been tested with 10 relations, presented in Table 1 (5 of them are symmetric and 5 of them are not). The algorithm agrees in 9 out of 10 cases with the human evaluation. Note that, in this case, SCP O and SCP P coincide as the two individuals and the relationship are the same in the original relation *xRy* and in the relation *yRx*. In our test, 4 negative examples didn't obtain any result with the given pattern.

From the "Deep Blue defeated Gary Kasparov" example notice that, as a matter of fact, Gary Kasparov defeated Deep Blue but in a different match. This illustrates a limitation of the automatic process.

Table 1. Test of Symmetry for several Object Properties.

Original relation	Pattern	SCP O	SCP P	D	HD
Spain borders France	France borders Spain	0.22	0.22	Y	Y
Obesity is associated with cancer	Cancer is associated with obesity	0.09	0.09	Y	Y
Microsoft competes with Google	Google competes with Microsoft	0.29	0.29	Y	Y
Angelina Jolie is married to Brad Pitt	Brad Pitt is married to Angelina Jolie	0.34	0.34	Y	Y
Cigarettes are linked to cancers	Cancers are linked to cigarettes	0.001	0.001	Y	Y
Deep Blue defeated Gary Kasparov	Gary Kasparov defeated Deep Blue	0.004	0.004	**Y**	**N**
Heavy smokers get lung cancer	Lung cancer get heavy smokers	6.5E-4	-	N	N
Michael Jackson was born in Gary	Gary was born in Michael Jackson	0.006	-	N	N
Sociology is the study of society	Society is the study of Sociology	0.009	-	N	N
Madrid is located in Spain	Spain is located in Madrid	0.03	-	N	N

Reflexive Object Property. The evaluation of the reflexivity of an Object Property has been tested with 10 relations, presented in Table 2 (5 of them are reflexive and 5 of them aren't). A relation is considered reflexive if the SCP score of the pattern exceeds 1E-3. The algorithm agrees in 8/10 cases with the human evaluation.

In the evaluated examples, the algorithm found evidences for "people have itself" and "stress affects itself". The correlations of "*itself*" with their respective subjects are high, but that doesn't imply they are correct. For these two cases, the human expert evaluated the reflexivity as wrong.

Table 2. Test of Reflexivity for several Object Properties.

Original relation	Pattern relation	SCP O	SCP P	D	HD
Nature regulates CO2	Nature regulates itself	0.04	0.23	Y	Y
Market regulates capital production	Market regulates itself	2.88E-3	0.14	Y	Y
Fire consumes wood	Fire consumes itself	0.15	0.09	Y	Y
Science studies Nature	Science studies itself	0.07	0.03	Y	Y
Immune system attacks viruses	Immune system attacks itself	0.015	0.0048	Y	Y
Testicular cancer has good prognosis	Testicular cancer has itself	1.1E-4	3.6E-5	N	N
People have colon cancer	People have itself	0.002	0.16	**Y**	**N**
Pressure sensor measures water pressure	Pressure sensor measures itself	8.4E-3	1.9E-4	N	N
Hysterectomy reduces cancer	Hysterectomy reduces itself	0.012	7.2E-4	N	N
Stress affects cancer	Stress affects itself	0.05	0.046	**Y**	**N**

Functional Object Property. The evaluation of the Functionality of an Object Property has been tested with 10 relations, presented in Table 3 (5 of them are functional and 5 of them are not). The algorithm agrees in 8/10 cases with the human evaluation. The table shows the best extraction found with the relation(subject,?) pattern. The algorithm takes the relation to be functional if the best extraction has an SCP lower than the established 1E-3 threshold.

It can be observed that the functionality test presents some limitations. The original relation "Cadmium exposure is associated with prostate cancer" scored a SCP of 1.28E-4 and the BES "reduced pulmonary function" scored higher than the original relation with 1.3E-4, but the algorithm considered the relation as not functional. For this case the expert decided that the relation with the best extraction is in fact valid.

Regarding the relation “Michael Jackson was born in Gary”, the relation “was born in” may be related to either time or place giving unexpected results. The expert considered the relation as functional (giving an approach of “place” only) while the algorithm considered the contrary.

Table 3. Test of Functionality for several Object Properties.

Original relation	**NC**	**BES**	**SCP O**	**SCP P**	**D**	**HD**
Brasilia is the capital of Brazil	4	country	0.004	3.7E-4	Y	Y
Cadmium exposure is associated with prostate cancer	1	reduced pulmonary function	1.28E-4	1.3E-4	**Y**	**N**
Da Vinci Code was written by Dan Brown	1	Bart D. Ehrman	0.009	6.5E-4	Y	Y
Comet Halley is named for Edmund Halley	0	-	0.019	-	Y	Y
Microsoft competes with Google	42	Apple	0.29	0.23	N	N
Brazil borders Bolivia	8	Peru	0.14	0.20	N	N
Michael Jackson was born in Gary	3	August	0.005	0.006	**N**	**Y**
Raloxifene reduces breast cancer	32	osteoporotic fractures	9.4E-4	0.006	N	N
Spain borders France	3	Portugal	0.229	0.25	N	N
Chichen Itza is known as El Castillo	2	La Iglesia	0.008	8.25E-5	Y	Y

Inverse Object Property. The discovery of Inverse relations for an Object Property has been tested with 10 relations, presented in Table 4. The table shows the best extraction (i.e., the best candidate to inverse relation) found with the ?(object, subject) pattern. The algorithm takes the relation to have an inverse if the best extraction has an SCP (SCP P column) higher than the established 1E-3 threshold.

Table 4. Discovery of Inverse relations for several Object Properties.

Original relation	**NC**	**BES**	**SCP O**	**SCP P**	**D**	**HD**
Christopher Columbus discovered America	5	was discovered by	0.004	0.017	Y	Y
Market regulates capitalist production	1	is attached to	1.8E-3	5.8E-5	N	N
Global warming is caused by pollution	6	causes	0.11	0.06	Y	Y
Cigarette smoking is associated with lung cancer	1	is caused by	0.007	0.002	Y	Y
Sensors contain mechanical parts	1	required for	4.7E-3	7.1E-4	N	N
Viruses are associated with liver cancer	2	caused by	5.2E-3	0.003	Y	Y
Kasparov defeated Deep Blue	16	beaten by	0.003	0.001	Y	Y
Stress affects asthma	3	is linked to	0.02	0.025	Y	Y
Meningococcal disease is caused by Neisseria Meningitides	1	causes	0.03	0.002	Y	Y
Charles Darwin proposed evolution theory	8	proposed by	0.002	0.001	Y	Y

The examples show that it is possible to extract incomplete verb phrases. For example, we found evidences for “liver cancer caused by viruses”, whereas “is caused by” would be a more complete verb phrase for that relation.

Inverse Functional Object Property. The Inverse Functionality of an Object Property has been tested with 5 of the "inverse" relations shown in the previous table. Table 5 shows the best extraction found with the relation(subject,?) pattern. The algorithm takes the original relation to be inverse functional if the best extraction has an SCP (SCP P column) higher than the established 1E-3 threshold. The algorithm agrees with the human result in 4/5 cases. In the example (shown in the following table) that states "Evolution theory proposed by Charles Darwin", notice that a selected extraction is "Darwin". Both expressions refer to the same individual, but with the algorithm proposed this fact is not taken into account concluding in a wrong decision.

Table 5. Test of the Inverse Functionality of some Object Properties

Original relation	**NC**	**BES**	**SCP O**	**SCP P**	**D**	**HD**
Pollution causes global warming	76	acid rain	0.066	0.013	N	N
Asthma is linked to stress	67	allergies	0.026	0.108	N	N
Neisseria Meningitides causes meningococcal disease	2	inflammation	0.0022	3.08E-5	Y	Y
Evolution theory proposed by Charles Darwin	1	Darwin	0.001	0.001	**N**	**Y**
Lung cancer is caused by cigarette smoking	57	second hand smoke	0.003	0.0028	N	N

Transitive Object Property. To test the Transitivity of an Object Property we have followed two steps. Given a relation *xRy*, in the first step we look for individuals related to *y* via *R*, using the pattern *R(y,?)*. In the second step, we take the best extraction in each of the cases (*z*), and we try to evaluate whether *xRz* holds.

Table 6a shows the five relationships used in this test, along with the best extraction found in each case. An extraction is taken as correct by the algorithm if its SCP is above the 1E-3 threshold. From the first four relations given in Table 6a, we evaluate in Table 6b the best extraction (*z*) to see if *R(x,z)* holds. If it does, the algorithm takes the relation to be transitive.

Tables 6a and 6b. Two steps in the test of Transitivity for some Object Properties

Original relation	**NC**	**BES**	**SCP O**	**SCP P**	**D**	**HD**
Global warming is caused by pollution	54	Nature	0.11	0.05	Y	Y
Berlin is located in Germany	12	Europe	0.04	0.07	Y	Y
Pollution affects global warming	46	Weather	0.038	0.05	Y	Y
Microsoft competes with Apple	30	Google	0.29	0.18	Y	Y
Arsenic is associated with kidney cancer	5	Obesity	8.1E-5	4.3E-5	**N**	**Y**

Relation	**SCP O**	**SCP P**	**D**	**HD**
Global warming is caused by Nature	0.11	0.05	Y	Y
Berlin is located in Europe	0.04	0.02	Y	Y
Pollution affects weather	0.038	0.02	Y	Y
Microsoft competes with Google	0.29	0.31	Y	Y

Considering that the evaluation of the transitive property in our approach includes two steps, the possibility to find transitivity for a relation is difficult. However, if a transitivity property is found using this approach that is due to the fact that the evaluated relatedness is certainly strong, as in the relations "Berlin is located in Europe" or "Microsoft competes with Google".

5 Conclusion and Future Work

Axiomatizing ontologies by hand is a hard and time-consuming task which involves ontology engineers and domain experts. The design of automatic axiomatization solutions is fundamental to contribute to the success of the Semantic Web and many other fields in Computer Science.

Our approach is a step forward to fill this gap, by learning object properties axioms for non-taxonomic relations using the Web as corpus, shallow linguistic techniques and Web-scale statistics. The Web can be exploited to aid in the Ontology Learning process by using suitable queries in order to retrieve valuable information. The patterns used in this work are totally generic and can be easily applied to any Web search engine. Besides, the proposed method is domain-independent.

However, being completely automatic and unsupervised, the proposed algorithm presents several limitations, that can lead to the definition of lines of future work. Some of these shortcomings are the following:

- At the moment we study only the axioms of a particular relationship between two individuals, *xRy*, e.g. to test the symmetry, we check whether *yRx* holds. We don't cover the study of the mathematical properties of the *R* relation in general; for example, to guarantee the symmetry of *R*, we should check whether, for all *x*,*y* such that *xRy* holds, *yRx* also holds.
- Regarding the reflexive property, the search for "*xR itself*" may be too restrictive and hard to find explicitly. We can add to the query the words "may" and "can" in order to extend the reflexive property discovery in the form of *Subject may/can Relation itself*.
- The transitive property is checked only with a particular instance. Note that "Italy borders Switzerland" will be qualified as transitive depending on the country which is the best extraction for the "Switzerland borders" pattern (giving different results e.g. for France and Germany, as one borders Italy and the other doesn't). Thus, we should at least check whether all the neighbours of Switzerland are also neighbours of Italy.

It is also necessary to study more deeply the effect of many natural language phenomena such as synonyms and polisemy in order to provide better results in certain domains in which these ambiguities are common. This may lead to further research in composing a more concrete set of patterns or adding more constraints in their extraction and selection. The empirical tests also show that in some cases the semantic relatedness acquired from the extractions has a similar value than the one of the original relation presented in the ontology. Assuming that the relations in the ontology are correct, we can use the SCP of the relations as a criterion select or discard observation instead of using a predefined threshold.

6 Acknowledgements

Luis Miguel Del Vasto has been supported by a grant of Fundación Carolina. This work has been supported by the Universitat Rovira i Virgili (2009AIRE-04), the Spanish Ministry of Science and Innovation (DAMASK project, Data mining algorithms with semantic knowledge, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan).

References

1. Sánchez, D.: Domain Ontology Learning from the Web. VDM Verlag Dr. Mueller (2008)
2. Cimiano, P.: Ontology Learning and Population from Text, Springer Science, 9-10 (2006)
3. Euzenat, J.M.: Research Challenges and perspectives of the Semantic Web, Report of the EU-NSF strategic workshop, Sophia-Antipolis, France, pp. 86--88 (2001)
4. Volker J., Haase P., Hitzler P.: Learning Expressive Ontologies. In: Conference on Ontology Learning and Population, Germany, pp. 45--69 (2008)
5. Sánchez, D., Moreno, A.: Pattern-based automatic taxonomy learning from the Web, AI Communications, 21(1), 27--48 (2008)
6. Shamsfard M., Abdollahzadeh A.: Learning Ontologies from Natural Language Texts. In: International Journal of Human-Computer Studies, International Journal of Human-Computer Studies 60, 17--63 (2004)
7. Rudolph S., Volker J., Hitzler P.: Supporting Lexical Ontology Learning by Relational Exploration. In: 15th International Conference on Conceptual Structures ICCS, Germany, pp. 488--491 (2007)
8. Volker J., Hitzler P., Cimiano P.: Acquisition of OWL DL Axioms from Lexical Resources. In: 4th European conference on The Semantic Web, Germany, 670--685 (2007)
9. Hearts M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th Conference on Computational linguistics - Volume 2, France, pp. 539--545 (1992)
10. Ritter A., Downey D., Soderland S., Etzioni O.: It's a Contradiction-No it's not: A Case study using Functional Relations. In: Conference on Empirical Methods in Natural Language Processing, University of Washington, USA, pp. 11--20 (2008)
11. Volker J., Vrande D., Sure Y. and Hotho A.: Learning Disjointness. In: 4th European conference on The Semantic Web, Germany, pp. 175--189 (2007)
12. Schutz, A., Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In: 4th International Semantic Web Conference, pp. 593--606 (2005)
13. Dias G., Santos C., Cleuzious G.: Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining. In: Workshop on Information Extraction Beyond The Document, Sydney, pp. 36--47 (2006)
14. Sánchez, D., Moreno, M.: Learning non-taxonomic relationships from web documents for domain ontology construction, Data and Knowledge Engineering 63(3), 600--623 (2008)
15. Rudolph, T.: Exploring relational structures via FLE. In: Wolff, K.E., Pfeiffer, H.D., Delugach, H.S., (eds.) Conceptual Structures at Work, pp. 196--212 (2004)