A Proposal for the Automatic Generation of Instances from Unstructured Text

Roxana Danger[1], I. Sanz[2], Rafael Berlanga-Llavori[2], and José Ruiz-Shulcloper[3]

[1] University of Oriente, Santiago de Cuba, Cuba
roxana@csd.uo.edu.cu
[2] Universitat Jaume I, Castellón, Spain
berlanga@uji.es
[3]Institute of Cybernetics, Mathematics and Physics, La Habana, Cuba
recpat@icmf.inf.cu

Abstract. An ontology is a conceptual representation of a domain resulted from a consensus within a community. One of its main applications is the integration of heterogeneous information sources available in the Web, by means of the semantic annotation of web documents. This is the cornerstone of the emerging *Semantic Web*. However, nowadays most of the information in the Web consists of text documents with little or no structure at all, which makes impracticable their manual annotation. This paper addresses the problem of mapping text fragments into a given ontology in order to generate ontology instances that semantically describe this kind of resources. As a result, applying this mapping we can automatically populate a Semantic Web consisting of text documents that concern with a specific ontology. We have evaluated our approach over a real-application ontology and a text collection both in the Archeology domain. Results show the effectiveness of the method as well as its usefulness.

1 Introduction

The Semantic Web (SW) project is intended to enrich the current Web by creating knowledge objects that allow both users and programs to better exploit the Web resources [1]. The cornerstone of the Semantic Web is the definition of an ontology that conceptualizes the knowledge within the resources of a domain [2]. Thus, the contents of the Web will be described in the future by means of a large collection of semantically tagged resources that must be reliable and meaningful.

Nowadays most of the information in the Web consists of text documents with little or no structure at all, which makes impracticable their manual annotation [3]. As a consequence, automatic or even semi-automatic resource tagging will become a necessary and urgent task to populate the contents of the SW.

For this purpose, we can benefit from the research works done in both the automatic document classification and the Information Extraction areas. In the former area, several methods have been proposed to automatically classify documents according to a given taxonomy of concepts (e.g. [4]). However, these methods require a big amount of training data, and they do not generate *instances* of the ontology. In the latter area, much work has been focused on recognizing predefined entities (e.g. dates, locations, names, etc.), as well as on extracting relevant relations between them by using natural language processing. However, current Information Extraction (IE) systems are restricted to extract flat and very simple relations, mainly to feed a rela-

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 462–469, 2004.

tional database. In an ontology, relationships are more complex as they can involve nested concepts and they can have associated inference rules.

This paper addresses the problem of mapping text fragments into a given ontology in order to generate the ontology instances that semantically describe them. Our proposal is based on generating a specific mapping between text fragments and the subgraphs of the ontology that better fit them. As in current IE systems, we also recognize predefined entities that can be directly associated to ontology concepts. These entities are extracted by applying a set of pre-defined regular expressions. However, in contrast to these systems, our approach does not use any syntactic nor semantic analysis to extract the relations appearing in the text. As a consequence, the proposed method performs very efficiently, and as the results show, it achieves a good effectiveness.

2 Generation of Ontology Instances

In this work we have adopted the formal definition of ontology from [6]. This definition distinguishes between the conceptual schema of the ontology, which consists of a set of concepts and their relationships, and its associated resource descriptions, which are called *instances*. In our approach, we represent both parts as oriented graphs, over which the ontology inference rules are applied to extract and validate complex instances.

2.1 Preliminary Concepts

The next paragraphs present the definitions of the concepts used when generating partial and complete instances from the words and entities appearing within a text fragment.

Definition 1. An *ontology* is a labeled oriented graph $G = (N, A)$ where the set of nodes N contains both the concept names and the instance identifiers, and A is a set of labeled arcs between them representing their relationships. Specifically, the labels *is_a* and *instance_of* account for the taxonomy and classification relationships respectively. Additionally, over this graph we introduce the following restrictions according to [6]:

- The set of the ontology concepts C is the subset of nodes in N that are not pointed by any arc labeled as "*instance_of*".
- The taxonomy of concepts C consists of the subgraph that only contains the *is_a* arcs. It defines a partial order over the elements of C, denoted with $\leq_{C.}$
- The relation signature is a function σ: $R \rightarrow C$ x C, which contains the subsets of arc labels that involve only concepts.
- The function dom: $R \rightarrow C$ with $dom(r) = \prod_1(\sigma(r))$ gives the domain of a relation $r \in R$, whereas $range$: $R \rightarrow C$ with $range(r) = \prod_2(\sigma(r))$ gives its range.
- The function dom_C: $C \rightarrow 2^{|\gamma|}$ gives the domain of definition of a concept.

As an example, Figure 1 shows an ontology to describe artifacts made by artists and where they are exhibited. Double line arrows represent the taxonomy relation $\leq_C$, single line arrows denote the different relationships between concepts, dotted line arrows represent the partial order $\leq_R$, and finally shaded nodes represent instances.

In order to express the proximity of two concepts *c* and *c*' we define the following relevance function:

$$Rel(c,c') = 1\text{- } \log_{|C|} \ \min \ \{d_T(c,c'),\ d_R(c,c')\}$$

where $d_T(c, c')$ and $d_R(c, c')$ are the distances between the two concepts *c* and *c*' through paths across the taxonomy and the relations in *R*, respectively.

For the sake of simplicity, we will represent these graphs only with the set of arcs, which are triples of the form (n_1, *label*, n_2), omitting the set of nodes.

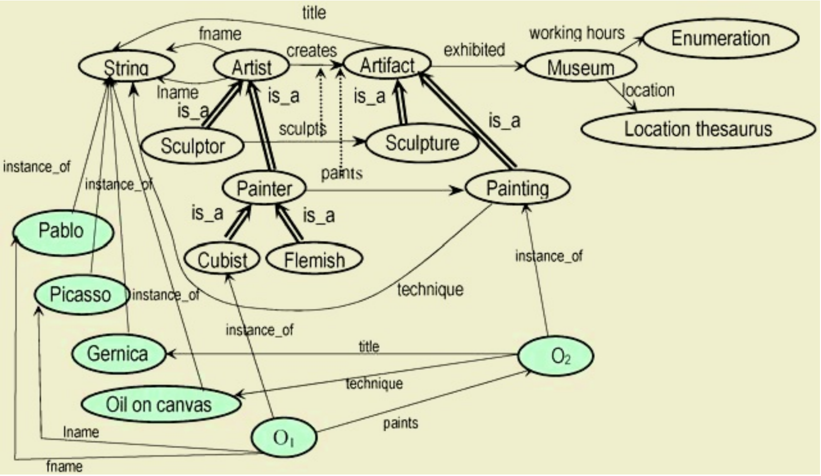

Fig. 1. Example of an ontology graph.

Definition 2. We denote with $I\big|_o^c$ an *instance related to the object o of the class c or simply an instance*, as the subgraph of the ontology that relates the node *o* with any other, and there exists an arc (*o*, *instance_of*, *c*) ∈ *A*.

$$I\big|_o^c = \{\ (o,r,o')/(o,\ instance\_of,\ c) \in A, \exists r \in R, \sigma(r) = (c^*, c'),$$
$$o \in dom_c(c^*), o' \in dom_c(c'), c^* \leq_C c, \neg\exists r', c^* \leq_C dom(r') \leq_C c\}$$

Notice that the arc (*o*, *instance_of*, *c*) is not included in the instance subgraph.

An instance is empty if the object *o* only participates in the arc (*o*, *instance_of*, *c*), which is represented as the set {(*o*, *, *)}

According to the ontology of Figure 1, the following sets are examples of instances:

$$I\big|_{o_1}^{cubist} = \{(o_1, fname, \text{"}Pablo\text{"}), (o_1, lname, \text{"}Picasso\text{"}), (o_1, paints, o_2), (o_1, paints, o_3)\}$$

$$I\big|_{o_2}^{painting} = \{(o_2, techniques, \text{"}Oil\ on\ canvas\text{"})\}$$

$$I\big|_{o_3}^{painting} = \{(o_3, techniques, \text{"}Oil\ on\ canvas\text{"}), (o_3, exibited, o_4)\}$$

In the next paragraphs we introduce several definitions to operate over instance subgraphs by applying the ontology inference rules over the taxonomy and the concept relations. These operations are intended to merge the partial information extracted from the text (concepts and relationships) in order to form maximal and coherent graphs covering as much as possible the text terms.

Definition 3. We call a *specialization of the instance* $I\big|_{o}^{c}$ to the class *c'*, *c*$\leq_{c}$*c'*, denoted with $I\big|_{o\to o'}^{c\to c'}$, the following instance:

$$I\big|_{o'}^{c'} = \{\, (o', r, x) \,/\, (o, r, x) \in I\big|_{o}^{c}, \neg\exists r' \in R, r \leq_R r', dom(r) \leq_C c', range(r) = c_x,$$
$$x \in dom_C(c_x), o' \in dom_C(c')\}$$

Basically, this definition says that an instance can be specialized by simply renaming the object with a name from the target class in all the instance triples that do not represent an overridden property.

For example, the specialization of the instance

$$I\big|_{o}^{artist} = \{\, (o, fname, "Pablo"), (o, lname, "Picasso"), (o, creates, o_2), (o, creates, o_3)\}$$

to the class *painter* is:

$$I\big|_{o'}^{painter} = I\big|_{o\to o'}^{artist\to painter} = \{(o', fname, "Pablo"), (o', lname, "Picasso")\}$$

Definition 4. We call an *abstraction of the instance* $I\big|_{o}^{c}$ to the class *c'*, *c'*$\leq_{c}$*c*, denoted with $I\big|_{o\uparrow o'}^{c\uparrow c'}$, the instance

$$I\big|_{o'}^{c'} = \{(o', r, x) \,/\, (o, r, x) \in I\big|_{o}^{c}, \exists r, c_x, dom(r) \leq_C c', range(r) = c_x, x \in dom(c_x)\}\,.$$

Similarly to the previous definition, we can obtain an abstract instance by selecting all the triples whose relation name can be abstracted to a relation of the target class *c'* and renaming accordingly the instance object.

For example, the abstraction of the instance $I\big|_{o_1}^{cubist}$ to the class painter is the following subgraph:

$$I\big|_{o'}^{painter} = I\big|_{o_1\uparrow o'}^{cubist\uparrow painter} = \{(o', fname, "Pablo"), (o', lname, "Picasso"), (o', paints, o_2),$$
$$(o', paints, o_3)\}$$

Definition 5. We call *union of the instances* $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$ and denote it by $I_1\big|_{o}^{c} \hat{\cup} I_2\big|_{o'}^{c'}$ the set:

$$I_1\big|_{o}^{c} \hat{\cup} I_2\big|_{o'}^{c'} = \begin{cases} I_1\big|_{o}^{c} \cup I_2\big|_{o'}^{c'} & \mathit{if} \quad I_1\big|_{o}^{c} \neq \{(o,*,*)\} \wedge I_2\big|_{o'}^{c} \neq (o',*,*) \vee (\neg c \leq_{c} \mathrm{c'} \wedge \neg \mathrm{c'} \leq \mathrm{c}) \\ I_1\big|_{o\to o'}^{c\to c'} & \mathit{if} \quad I_2\big|_{o'}^{c'} = \{(o',*,*)\} \wedge \mathrm{c} \leq_{c} \mathrm{c'} \\ I_1\big|_{o}^{c} & \mathit{if} \quad I_2\big|_{o'}^{c'} = \{(o',*,*)\} \wedge \mathrm{c'} \leq_{c} \mathrm{c} \\ I_2\big|_{o'}^{c'} & \mathit{if} \quad I_1\big|_{o}^{c} = \{(o,*,*)\} \wedge \mathrm{c} \leq_{c} \mathrm{c'} \\ I_2\big|_{o'\to o}^{c'\to c} & \mathit{if} \quad I_1\big|_{o}^{c} = \{(o',*,*)\} \wedge \mathrm{c'} \leq_{c} \mathrm{c} \end{cases}$$

Definition 6. We call *difference of the instances* $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$, denoted with $I_1\big|_{o}^{c} \mathbin{\hat{-}} I_2\big|_{o'}^{c'}$, the following set: $I_1\big|_{o}^{c} \mathbin{\hat{-}} I_2\big|_{o'}^{c'} = \begin{cases} I_1\big|_{o}^{c} - I_2\big|_{o'\uparrow o}^{c'\uparrow c} & if \quad c \leq_c c' \\ I_1\big|_{o}^{c} - I_2\big|_{o'\to o}^{c'\to c} & if \quad c' \leq_c c \end{cases}$

Definition 7. We call *symmetric difference of the instances* $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$, denoted with $I_1\big|_{o}^{c} \mathbin{\hat{\div}} I_2\big|_{o'}^{c'}$ the result of $I_1\big|_{o}^{c} \mathbin{\hat{-}} I_2\big|_{o'}^{c'} \mathbin{\hat{\cup}} I_2\big|_{o'}^{c'} \mathbin{\hat{-}} I_1\big|_{o}^{c}$.

Definition 8. We say that two instances $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$ related to the objects *o* and *o*' of classes *c* and *c*', respectively, $c \leq_C c'$, are *complementary* if they satisfy at least one of the following two conditions:

1. $(\neg\exists(o', r, x), (o', r, x') \in I_1\big|_{o\to o'}^{c\to c'} \div I_2\big|_{o'}^{c'}, x \neq x'$, *r* is biyective) and $(\neg\exists(o, r, x), (o', r', x') \in I_1\big|_{o}^{c} \mathbin{\hat{\cup}} I_2\big|_{o'}^{c'}, r \leq_R r', r \neq r', x \neq x')$ or
2. at least one of the instances $I_1\big|_{o}^{c}$ or $I_2\big|_{o'}^{c'}$ is empty.

Two instances are complementary if there not exist contradictions between the values of their similar relations.

For example, the following instances are complementary:

$$I\big|_{o}^{artist} = \{(o, fname, "Pablo"), (o, lname, "Picasso")\}$$

$$I\big|_{o'}^{painter} = \{(o', paints, o_2), (o', paints, o_3)\}$$

Definition 9. We say that two complementary instances $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$ are *unifiable* in $I\big|_{o_u}^{c'} = I_1\big|_{o\to o_u}^{c\to c'} \mathbin{\hat{\cup}} I_2\big|_{o'\to o_u}^{c'\to c'}$, $c \leq_C c'$.

For example, $I\big|_{o}^{artist}$ and $I\big|_{o'}^{painter}$ are unifiable, producing the following instance:

$$I\big|_{o_u}^{painter} = I\big|_{o\to o_u}^{artist\to painter} \mathbin{\hat{\cup}} I_2\big|_{o'\to o_u}^{painter\to painter} = \{ (o_u, fname, "Pablo"), (o_u, lname, "Picasso"), (o_u, paints, o_2), (o_u, paints, o_3)\}$$

Definition 10. We say that two instances, $I_1\big|_{o}^{c}$ and $I_2\big|_{o'}^{c'}$ are aggregable in $I\big|_{o}^{c^*} = I_1\big|_{o\to o}^{c\to c^*} \cup \{(o, r, o'')\}$, if $\exists r \in R$, $\sigma(r) = (c^*, c'')$, $c \leq_C c^*$, $c'' \leq_C c'$, and *o*'' is the name of the instance $I\big|_{o''}^{c''} = I_2\big|_{o'\to o''}^{c'\to c''}$.

For example, $I\big|_{o}^{artist}$ and $I\big|_{o_2}^{painting}$ are aggregable in

$$I\big|_{o}^{painter} = I\big|_{o\to o}^{artist\to painter} \cup \{(o, paints, o'')\} = \{(o, fname, "Pablo"), (o, lname, "Picasso"), (o, paints, o'')\}$$

where $I\big|_{o''}^{painting} = \{(o'', techniques, "Oil\ on\ canvas")\}$

2.2 Extracting Instances from Texts

Let T be a text fragment formed by the sequence of terms $(w_1,\ldots,w_n)$, which can be either words or extracted entities. It is worth mentioning that stop-words are removed from this sequence, and the different word forms are reduced to their lemmas. We will denote with $I\big|^{c}_{o,[i,j]}$ the instance described by the subsequence of terms between w_i and w_j ($j \geq i$).

In this context, two instances of the text T, $I_1\big|^{c}_{o,[i_1,j_1]}$ and $I_2\big|^{c'}_{o',[i_2,j_2]}$, with $c \leq_C c'$, are unifiable if they are complementary according to Definition 8, and there not exists any instance related to them. In this case, the unification of both instances is the following instance:

$$I\big|^{c'}_{o_u,[\min\{i_1,i_2\},\max\{j_1,j_2\}]},\ \text{where}\ I\big|^{c'}_{o_u} = I\big|^{c}_{o} \hat{\cup} I\big|^{c'}_{o'}.$$

For example, let us consider the following text, where the extracted entities have been denoted with pairs *entity-type*:*entity*,

```
fname:Pablo lname:Picasso was a very famous painter, specifically
he was an eminent cubist. In his paintings title:Woman and
title:Guernica he used oil on canvas.
```

From this text, we can identify the following partial instances:

$$I\big|^{painter}_{o_1,[1,7]} = \{(o_1 fname, "Pablo"), (o_1, lname, "Picasso")\}$$

$$I\big|^{cubist}_{o_2,[13,13]} = \{(o_2,*,*)\} \qquad I\big|^{painting}_{o_3,[16,17]} = \{(o_3, title, "Woman")\}$$

$$I\big|^{atifact}_{o_4,[18,18]} = \{(o_4, title, "Guernica")\} \qquad I\big|^{painting}_{o_5,[21,23]} = \{(o_5, techniques, "Oil\ on\ canvas")\}$$

According to the previous definitions, the instances $I\big|^{painter}_{o_1,[1,7]}$ and $I\big|^{cubist}_{o_2,[13,13]}$ are complementary and they are unifiable in

$$I\big|^{cubist}_{o_2,[1,7]} = \{(o_2, fname, "Pablo"), (o_2, lname, "Picasso")\}$$

Moreover, the instance $I\big|^{cubist}_{o_2,[1,13]}$ can aggregate $I\big|^{painting}_{o_3,[16,17]}$ and $I\big|^{atifact}_{o_4,[18,18]}$, being the result the following one

$$I\big|^{cubist}_{o_2,[1,23]} = \{(o_2, fname, "Pablo"), (o_2, lname, "Picasso"),$$
$$(o_2, paints, o_3), (o_2, paints, o_4)\}$$

Taking into account the definitions above, the system generates the set of instances associated to each text fragment as follows. Firstly, it constructs all the possible partial instances with the concept, entity and relation names that appear in the text fragment. Then it tries to unify partial instances, substituting them by the unified ones. Finally, instances are aggregated to form complex instances. The following algorithm resumes all the process.

```
1. For each concept c in the ontology appearing in the text
    a. Create an empty instance of c
    b. Let $R_{concepts} = \{c' / Rel(c', c) \geq 3\}$ be the relevant concepts related to c
    c. For each value/entity, v, appearing in the text so that it participates in a relation r of a
       concept c', in $R_{concepts}$, create an instance $I|_{o'}^{c'} = \{(o', r, v)\}$
2. Apply Definitions 9, 10 and the considerations remarked in Section 2.2, to merge complementary instances and to group aggregable instances.
```

Alg. 1. Algorithm for generating ontology instances.

3 Experiments

In order to validate the proposed approach, we have tested it over an ontology for the Archeological Research[1]. This ontology contains 164 concepts, 92 relations, and 390 predefined literal values. The number of entity types that can be recognized is around 20 (e.g. dates, locations, numbers, measures, etc.). We have selected 16 excavation reports from [8] as the test document collection.

In this work we evaluate the quality of the extracted instances by computing the percentage of the correct ones and the percentage of missing ones. For this purpose, we have manually annotated 15 complete instances from these text collection. Table 1 shows the global results[1] obtained by applying different ways of fragmenting the document contents, namely: sentences, paragraphs and sections.

Table 1. Precision and missing percentages for different levels of fragmentation.

	Number	Number of generated instances	Precision (%)	Missing (%)
Sentences	7262	5841	96.1	15.3
Paragraph	4441	3923	94.6	16.8
Section	254	4157	84.2	15.6

From these results, we can conclude that when we use sentences and paragraph to fragment documents, the proposed method performs very effectively, extracting a good number of instances. When we use the logical sections of documents, the precision notably decreases, as the ambiguity of relations increases. The obtained missing percentages can be attributed to the linguistic references to objects that currently are no detected by our method.

4 Conclusions

This work presents a new approach to the automatic generation of ontology instances from a collection of unstructured documents. This approach does not use any grammar nor extraction rules to obtain the ontology instances. Instead, our system tries to form correct partial instances by taking words and entities appearing in texts and automatically relating them. This makes our approach very efficient. Our experiments on a real-application show that we can obtain a good extraction precision, but it de-

[1] http://tempus.dlsi.uji.es/TKBG/Arqueologia.rdfs, http://tempus.dlsi.uji.es/TKBG/Instancias.rdf

pends on the fragmentation level used to find the mapping between fragments and the ontology. As future work, we plan to include further techniques of current IE systems for detecting concept references, which can help us to combine partial instances even if they stem from different documents.

Acknowledgements

This work has been partially funded by the research project of the Spanish Program of Research CICYT TIC2002-04586-C03.

References

1. Berners-Lee, T., Hendler, J., Lassila, O. "The Semantic Web". Scientific American, 2001.
2. Gruber, T.R. “Towards Principles for the Design of Ontologies used for Knowledge Sharing”, International Journal of Human-Computer Studies Vol. 43, pp. 907-928, 1995.
3. Forno, F., Farinetti, L., Mehan, S. “Can Data Mining Techniques Ease The Semantic Tagging Burden?”, SWDB 2003, pp. 277-292, 2003.
4. Doan, A. et al. “Learning to match ontologies on the Semantic Web”. VLDB Journal 12(4), pp. 303-319, 2003.
5. Appelt, D. “Introduction to Information Extraction”, AI Communications 12, 1999.
6. Maedche, A., Neumann, G. and Staab, S. “Bootstrapping an Ontology based Information Extraction System”. Studies in Fuzziness and Soft Computing, Springer, 2001.
7. Danger, R., Ruíz-Shulcloper, J., Berlanga, R. "Text Mining using the Hierarchical Structure of Documents" Current Topics in Artificial Intelligence (CAEPIA 2003), Lecture Notes in Computer Science (In Press), 2004.
8. Dirección General del Patrimonio Artístico. http://www.cult.gva.es/dgpa/