Ontology Enrichment Through Automatic Semantic Annotation of On-Line Glossaries

Roberto Navigli and Paola Velardi

Dipartimento di Informatica
Università di Roma "La Sapienza"
00198 Roma Italy
{navigli, velardi}@di.uniroma1.it

Abstract. The contribution of this paper is to provide a methodology for automatic ontology enrichment and for document annotation with the concepts and properties of a domain core ontology. Natural language definitions of available glossaries in a given domain are parsed and converted into formal (OWL) definitions, compliant with the core ontology property specifications.

To evaluate the methodology, we annotated and formalized a relevant fragment of the AAT glossary of art and architecture, using a subset of 10 properties defined in the CRM CIDOC cultural heritage core ontology, a recent W3C standard.

Keywords: Ontology Learning, Core Ontology, Glossaries.

1 Introduction

Large-scale, automatic semantic annotation of web documents based on well established domain ontologies would allow various Semantic Web applications to emerge and gain acceptance. Wide coverage ontologies are indeed available for general-purpose domains (e.g. WordNet, CYC, SUMO[1]), however semantic annotation in unconstrained areas seems still out of reach for state of art systems. Domain-specific ontologies are preferable since they would limit the domain and make the applications feasible.

Recently, certain web communities began to believe in the benefits deriving from the application of Semantic Web techniques. Accordingly, they produced remarkable efforts to conceptualize their competence domain through the definition of a *core ontology*. Relevant examples are in the area of enterprise modeling (Fox et al. 1997; Uschold et al. 1998) and cultural heritage (Doerr, 2003).

Core ontologies are indeed a necessary starting point to model in a principled way the basic concepts, relations and axioms of a given domain. But in order for an ontology to be really usable in applications, it is necessary to enrich the core structure with the thousands of concepts and instances that "make" the domain.

[1] WordNet: http://wordnet.princeton.edu, CYC: http://www.opencyc.org, SUMO: http://www.ontologyportal.org

S. Staab and V. Svatek (Eds.): EKAW 2006, LNAI 4248, pp. 126 – 140, 2006.

In this paper we present a methodology to automatically annotate a glossary $\mathcal{G}$ with the semantic relations of an existing *core ontology* $\mathcal{O}$. The annotation of documents and glossary definitions is performed using regular expressions, a widely adopted text mining approach. However, while in the literature regular expressions seek mostly for patterns at the lexical and part of speech level, we defined expressions enriched with syntactic and semantic constraints. A word sense disambiguation algorithm, SSI (Velardi and Navigli, 2005), is used to automatically replace the high level semantic constraints specified in the core ontology with fine–grained sense restrictions, using the sense inventory of a general purpose lexicalized ontology, WordNet.

From each gloss *G* of a term *t* in the glossary $\mathcal{G}$, we extract one or more semantic relation *instances* $R(C_t,C_w)$, where R is a relation in $\mathcal{O}$, C_t and C_w are respectively the *domain* and *range* of R. The concept C_t corresponds to its lexical realization *t*, while C_w is the concept associated to a word *w* in G, captured by a regular expression.

The annotation process allows to automatically enrich $\mathcal{O}$ with an existing glossary in the same domain of $\mathcal{O}$, since each pair of term and gloss (t,G) in the glossary $\mathcal{G}$ is transformed into a formal definition, compliant with $\mathcal{O}$. Furthermore, the very same method can be used to automatically annotate free text with the concepts and relations of the enriched ontology $\mathcal{O}'$. We experimented our methodology in the cultural heritage domain, since for this domain several well-established resources are available, like the CIDOC-CRM core ontology, the AAT art and architecture thesaurus, and others.

The paper is organized as follows: in Section 2 we briefly present the CIDOC and the other resources used in this work. In Section 3 we describe in detail the ontology enrichment algorithm. Finally, in Section 4 we provide a performance evaluation on a subset of CIDOC properties and a sub-tree of the AAT thesaurus. Related literature is examined in Section 5.

2 Semantic and Lexical Resources in the Cultural Heritage Domain

In this section we briefly describe the resources that have been used in this work.

2.1 The CIDOC CRM

The core ontology $\mathcal{O}$ is the *CIDOC CRM* (Doerr, 2003), a formal core ontology whose purpose is to facilitate the integration and exchange of cultural heritage information between heterogeneous sources. It is currently being elaborated to become an ISO standard. In the current version (4.0) the CIDOC includes 84 taxonomically structured concepts (called *entities*) and a flat set of 141 semantic relations, called *properties*. Entities are defined in terms of their subclass and super-class relations in the CIDOC hierarchy, and a *scope note* is used to provide an informal description of the entity. Properties are defined in terms of *domain* (the class for which a property is formally defined) and *range* (the class that comprises all potential values of a property), e.g.:

P46 is composed of (forms part of)
Domain: E19 Physical Object
Range: E42 Object Identifier

The CIDOC is an "informal" resource. To make it usable by a computer program, we replaced specifications written in natural language with formal ones. For each property $\mathcal{R}$, we created a tuple $R(C_d, C_r)$ where C_d and C_r are the domain and range entities specified in the CIDOC reference manual.

2.2 The AAT Thesaurus

The domain glossary $\mathcal{G}$ is the Art and Architecture Thesaurus (AAT) a controlled vocabulary for use by indexers, catalogers, and other professionals concerned with information management in the fields of art and architecture. In its current version[2] it includes more than 133,000 terms, descriptions, bibliographic citations, and other information relating to fine art, architecture, decorative arts, archival materials, and material culture. An example is the following:

maestà
Note: Refers to a work of a specific iconographic type, depicting the Virgin Mary and Christ Child enthroned in the center with saints and angels in adoration to each side. The type developed in Italy in the 13th century and was based on earlier Greek types. Works of this type are typically two-dimensional, including painted panels (often altarpieces), manuscript illuminations, and low-relief carvings.
Hierarchical Position:
Objects Facet
.... Visual and Verbal Communication
..... Visual Works
..... <visual works>
..... <visual works by subject type>
..... maestà

We manually mapped the top CIDOC entities to AAT concepts, as shown in Table 1.

Table 1. Mapping between AAT and CIDOC

AAT topmost	CIDOC entities
Top concept of AAT	CRM Entity (E1), Persistent Item (E77)
Styles and Periods	Period (E4)
Events	Event (E5)
Activities Facet	Activity (E7)
Processes/Techniques	Beginning of Existence (E63)
Objects Facet	Physical Stuff (E18), Physical Object (E19)
Artifacts	Physical Man-Made Stuff (E24)
Materials Facet	Material (E57)
Agents Facet	Actor (E39)
Time	Time-Span (E52)
Place	Place (E53)

[2] http://www.getty.edu/research/conducting_research/ vocabularies/aat/

2.3 Additional Resources

To apply semantic constraints on the words of a definition (as clarified in the next Section), we need additional resources. WordNet (Miller, 1995) is used to verify that certain words in a gloss-string *f* satisfy the range constraints $R(C_d,C_r)$ in the CIDOC. In order to do so, we manually linked the WordNet topmost concepts to the CIDOC entities. For example, the concept E19 (Physical Object) is mapped to the WordNet synset "*object, physical object*". Furthermore, we created a gazetteer $\mathcal{I}$ of named entities extracting names from the Dmoz[3], a large human-edited directory of the web, the Union List of Artist Names (ULAN) and the Getty Thesaurus of Geographic Names (GTG) provided by the Getty institute, along with the AAT.

3 Enriching the CIDOC CRM with the AAT Thesaurus

In this Section we describe in detail the method for automatic semantic annotation and ontology enrichment in the cultural heritage domain. Let $\mathcal{G}$ be a glossary, *t* a term in $\mathcal{G}$ and *G* the corresponding natural language definition (gloss). The main steps of the algorithm are the following:

1. Part-of-speech analysis. Each input gloss is processed with a part-of-speech tagger, TreeTagger[4]. As a result, for each gloss $G = w_1\ w_2\ \ldots\ w_n$, a string of part-of-speech tags $p_1\ p_2\ \ldots\ p_n$ is produced, where $p_i \in \mathcal{P}$ is the part-of-speech tag chosen by TreeTagger for word w_i, and $\mathcal{P}$= { *N*, *A*, *V*, *J*, *R*, *C*, *P, S, W* } is a simplified set of syntactic categories (respectively, nouns, articles, verbs, adjectives, adverbs, conjunctions, prepositions, symbols, wh-words).

2. Named Entity recognition. We augmented TreeTagger with the ability to capture named entities of locations, organizations, persons, numbers and time expressions. In order to do so, we use regular expressions (Friedl, 1997) in a rather standard way, therefore we omit details. When a named entity string $w_i\ w_{i+1}\ \ldots\ w_{i+j}$ is recognized, it is transformed into a single term and a specific part of speech denoting the kind of entity is assigned to it (*L* for cities (e.g. Venice), countries and continents, *T* for time and historical periods (e.g. Middle Ages), *O* for organizations and persons (e.g. Leonardo Da Vinci), B for numbers).

3. Annotation of sentence segments with CIDOC properties. We developed an algorithm for the annotation of gloss segments with properties grounded on the CIDOC-CRM relation model. Given a gloss *G* and a property[5] *R*, we define a *relation checker* c_R taking in input *G* and producing in output a set F_R of fragments of *G* annotated with the property *R*: <*R*>*f*</*R*>. The selection of a fragment *f* to be included in the set F_R is based on different kinds of constraints:

[3] http://dmoz.org/about.html
[4] TreeTagger is available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.
[5] In what follows, we adopt the CIDOC terminology for relations and concepts, i.e. properties and entities.

- a **part-of-speech constraint** *p*(*f*, *pos-string*) matches the part-of-speech (*pos*) string associated with the fragment *f* against a regular expression (*pos-string*), specifying the required syntactic structure.
- a **lexical constraint** *l*(*f*, *k*, *lexical-constraint*) matches the lemma of the word in *k-th* position of *f* against a regular expression (*lexical-constraint*), constraining the lexical conformation of words occurring within the fragment *f*.
- **semantic constraints on domain and range** s_D(*f*, *semantic-domain*) and *s*(*f*, *k*, *semantic-range*) are valid, respectively, if the term *t* and the word in the *k-th* position of *f* match the semantic constraints on domain and range imposed by the CIDOC, i.e. if there exists at least one sense of *t* C_t and one sense of *w* C_w such that: $R_{kind\text{-}of}^{*}(C_d, C_t)$ and $R_{kind\text{-}of}^{*}(C_r, C_w)$[6].

More formally, the annotation process is defined as follows: a *relation checker* c_R for a property *R* is a logical expression composed with constraint predicates and logical connectives, using the following production rules:

$c_R \rightarrow s_D(f, semantic\text{-}domain) \wedge c_R'$
$c_R' \rightarrow \neg c_R' \mid (c_R' \vee c_R') \mid (c_R' \wedge c_R')$
$c_R' \rightarrow p(f, pos\text{-}string) \mid l(f, k, lexical\text{-}constraint) \mid s(f, k, semantic\text{-}range)$

where *f* is a variable representing a sentence fragment. Notice that a relation checker must always specify a semantic constraint s_D on the *domain* of the relation *R* being checked on fragment *f*. Optionally, it must also satisfy a semantic constraint *s* on the k-th element of *f*, the range of *R*.

For example, the following excerpt of the checker for the *is-composed-of* relation (*P46*):

$(1)\ c_{is\text{-}composed\text{-}of}(f) = s_D(f, physical\ object\#1) \wedge p(f, \text{“}(V)_1(P)_2R?A?[CRJVN]^*(N)_3\text{”})$
$\wedge\ l(f, 1, \text{“}\hat{}(consisting|composed|comprised|constructed)\$\text{”})$
$\wedge\ l(f, 2, \text{“of”}) \wedge s(f, 3, physical\_object\#1)$

reads as follows: “the fragment *f* is valid if it consists of a verb in the set { *consisting*, *composed*, *comprised*, *constructed* }, followed by a preposition “of”, a possibly empty number of adverbs, adjectives, verbs and nouns, and terminated by a noun interpretable as a *physical object* in the WordNet concept inventory”. The first predicate, s_D, requires that also the term *t* whose gloss contains *f* (i.e., its domain) be interpretable as a *physical object*.

Notice that some letter in the regular expression specified for the part-of-speech constraint is enclosed in parentheses. This allows it to identify the relative positions of words to be matched against lexical and semantic constraints, as shown graphically in Figure 1.

Checker (1) recognizes, among others, the following fragments (the words whose part-of-speech tags are enclosed in parentheses are indicated in bold):
(consisting)$_1$ (of)$_2$ semi-precious **(stones)$_3$** (matching part-of-speech string: **(V)$_1$(P)$_2$ J(N)$_3$**)
(composed)$_1$ (of)$_2$ (knots)$_3$ (matching part-of-speech string: **(V) $_1$(P)$_2$(N)$_3$**)

[6] $R_{kind\text{-}of}$* denotes zero, one, or more applications of $R_{kind\text{-}of}$.

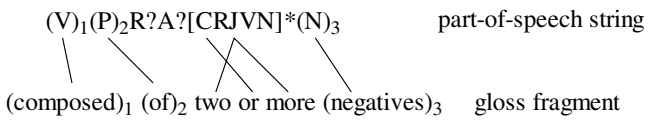

Fig. 1. Correspondence between parenthesized part-of-speech tags and words in a gloss fragment

As a second example, an excerpt of the checker for the *consists-of* (*P45*) relation is the following:

$$(2)\ c_{consists\text{-}of}(f) = s_D(f, \textit{physical object\#1}) \wedge p(f, \text{“(V)}_1\text{(P)}_2\text{A?[JN,VC]*(N)}_3\text{”}) \wedge l(f, 1, \text{“\^{}(make|do|produce|decorated)\$”}) \wedge l(f, 2, \text{“\^{}(of|by|with)\$”}) \wedge \neg s(f, 3, \textit{color\#1}) \wedge \neg s(f, 3, \textit{activity\#1}) \wedge (s(f, 3, \textit{material\#1}) \wedge s(f, 3, \textit{solid\#1}) \wedge s(f, 3, \textit{liquid\#1}))$$

recognizing, among others, the following phrases:

- **(made)$_1$ (with)$_2$** the red earth pigment **(sinopia)$_3$** (matching part-of-speech string: **(V)$_1$(P)$_2$**AJNN(**N)$_3$**)
- **(decorated)$_1$ (with)$_2$** red, black, and white **(paint)$_3$** (matching part-of-speech string: **(V)$_1$(P)$_2$**JJCJ(**N)$_3$**)

Notice that in both checkers (1) and (2) semantic constraints are specified in terms of WordNet sense numbers (*material#1*, *solid#1* and *liquid#1*), and can also be negative (¬*color#1* and ¬*activity#1*). The motivation is that CIDOC constraints are coarse-grained due to the small number of available core concepts: for example, the property *P45 consists of* simply requires that the range belongs to the class *Material* (*E57*). Using WordNet for semantic constraints has two advantages: first, it is possible to write more fine-grained (and hence more reliable) constraints, second, regular expressions can be re-used, at least in part, for other domains and ontologies. In fact, several CIDOC properties are rather general-purpose.

4. Formalisation of glosses. The annotations generated in the previous step are the basis for extracting *property instances* to enrich the CIDOC CRM with a conceptualization of the AAT terms. In general, for each gloss *G* defining a concept C_t, and for each fragment $f \in F_R$ of *G* annotated with the property *R*: <*R*>*f*</*R*>, it is possible to extract one or more property instances in the form of a triple $R(C_t, C_w)$, where C_w is the *concept* associated with a term or multi-word expression *w* occurring in *f* (i.e. its language realization) and C_t is the *concept* associated to the defined term *t* in AAT. For example, from the definition of *tatting* (a kind of lace) the algorithm automatically annotates the phrase *composed of knots*, suggesting that this phrase specifies the *range* of the *is-composed-of* property for the term *tatting*:

$$R_{is\text{-}composed\text{-}of}(C_{tatting}, C_{knot})$$

In this property instance, $C_{tatting}$ is the *domain* of the property (a term in the AAT glossary) and C_{knot} is the *range* (a specific term in the definition *G* of *tatting*).

Selecting the concept associated to the domain is rather straightforward: glossary terms are in general not ambiguous, and, if they are, we simply use a numbering policy to identify the appropriate concept. In the example at hand, $C_{tatting}$=*tatting#1* (the first and only sense in AAT). Therefore, if C_t matches the domain restrictions in the regular expression for R, then the domain of the relation is considered to be C_t. Selecting the range of a relation is instead more complicated. The first problem is to select the correct words in a fragment f. Only certain words of an annotated gloss fragment can be exploited to extract the range of a property instance. For example, in the phrase "depiction of fruit, flowers, and other objects" (from the definition of *still life*), only *fruit*, *flowers*, *objects* represent the range of the property instances of kind *depicts* (*P62*).

When writing relation checkers, as described in the previous paragraph of this Section, we can add *markers of ontological relevance* by specifying a predicate $r(f, k)$ for each relevant position k in a fragment f. The purpose of these markers is precisely to identify words in f whose corresponding concepts are in the range of a property. For instance, the checker (1) $c_{is\text{-}composed\text{-}of}$ from the previous paragraph is augmented with the conjunction: $\wedge$ ***r(f*, 3)**. We added the predicate $r(f, 3)$ because the third parenthesis in the part-of-speech string refers to an ontologically relevant element (i.e. the candidate *range* of the *is-composed-of* property).

The second problem is that words that are candidate ranges can be ambiguous, and they often are, especially if they do not belong to the domain glossary G. Considering the previous example of the property *depicts*, the word *fruit* is not a term of the AAT glossary, and it has 3 senses in WordNet (the fruit of a plant, the consequence of some action, an amount of product). The property *depicts*, as defined in the CIDOC, simply requires that the range be of type *Entity* (E1). Therefore, all the three senses of *fruit* in WordNet satisfy this constraint. Whenever the range constraints in a relation checker do not allow a full disambiguation, we apply the SSI algorithm (Navigli and Velardi, 2005), a semantic disambiguation algorithm based on structural pattern recognition, available on-line[7]. The algorithm is applied to the words belonging to the segment fragment f and is based on the detection of relevant semantic interconnection patterns between the appropriate senses. These patterns are extracted from a lexical knowledge base that merges WordNet with other resources, like word collocations, on-line dictionaries, etc.

For example, in the fragment "depictions of fruit, flowers, and other objects" the following properties are created for the concept *still_ life#1*:

$$R_{depicts}(still\_\ life\#1, fruit\#1)$$
$$R_{depicts}(still\_\ life\#1, flower\#2)$$
$$R_{depicts}(still\_\ life\#1, object\#1)$$

Some of the semantic patterns supporting this sense selection are shown in Figure 2.

A further possibility is that the range of a relation R is a concept *instance*. We create concept instances if the word w extracted from the fragment f is a named entity. For example, the definition of *Venetian lace* is annotated as "Refers to needle lace

[7] SSI is an on-line knowledge-based WSD algorithm accessible from http://lcl.di.uniroma1. it/ssi. The on-line version also outputs the detected semantic connections (as those in Figure 2).

created **<current-or-former-location>** in Venice**</current-or-former-location>** […]". As a result, the following triple is produced:

$$R_{has\text{-}current\text{-}or\text{-}former\text{-}location}(Venetian\_lace\#1, Venice{:}city\#1)$$

where *Venetian_ lace#1* is the concept label generated for the term *Venetian lace* in the AAT and *Venice* is an instance of the concept *city#1* (*city, metropolis, urban center*) in WordNet.

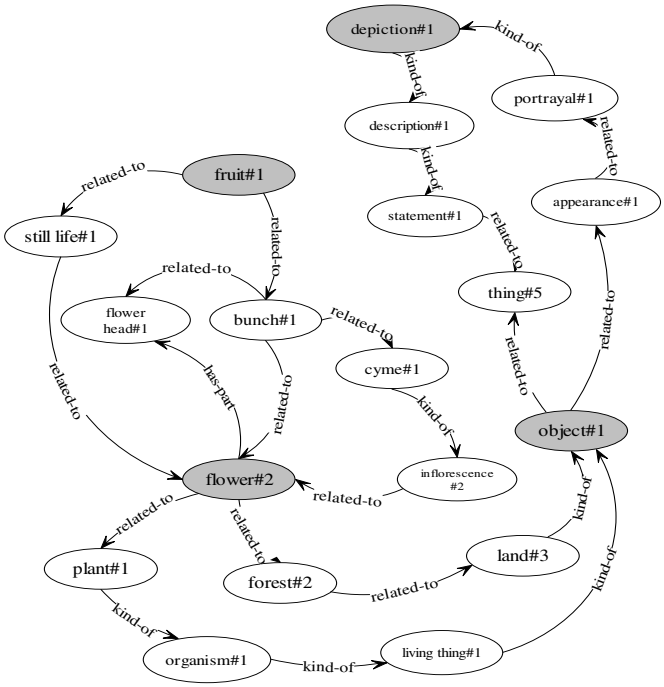

Fig. 2. Semantic Interconnections selected by the SSI algorithm when given the word list: "depiction, fruit, flower, object"

4 Evaluation

Since the CIDOC-CRM model formalizes a large number of fine-grained properties (precisely, 141), we selected a subset of properties for our experiments (reported in Table 2). We wrote a relation checker for each property in the Table. By applying the checkers in cascade to a gloss G, a set of annotations is produced. The following is an example of an annotated gloss for the term "vedute":

Refers to detailed, largely factual topographical views, especially **<has-time-span>**18th-century**</has-time-span>** Italian paintings, drawings, or prints of cities. The first vedute probably were **<carried-out-by>**painted by northern European artists**</carried-out-by>** who worked **<has former-or-current-location>**in Italy**</has former-or-current-location><has-**

time-span>in the 16th century**</has-time-span**>. The term refers more generally to any painting, drawing or print **<depicts>**representing a landscape or town view**</depicts>** that is largely topographical in conception.

Figure 3 shows a more comprehensive graph representation of the outcome for the concepts *vedute*#1 and *maestà*#1 (see the gloss in Section 2.2).

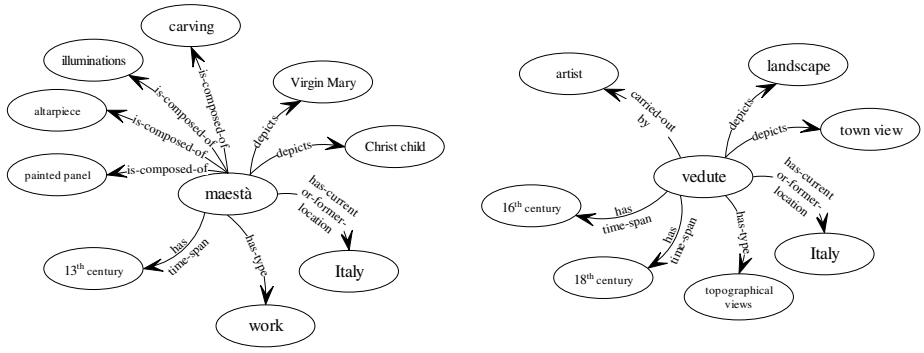

Fig. 3. Extracted conceptualisation (in graphical form) of the terms maestà#1 and vedute#1 (sense numbers are omitted for clarity)

To evaluate the methodology described in Section 3 we considered 814 glosses from the *Visual Works* sub-tree of the AAT thesaurus[8], containing a total of 27,925 words. The authors wrote the relation checkers by tuning them on a subset of 122 glosses, and tested their generality on the remaining 692. The test set was manually tagged with the subset of the CIDOC-CRM properties shown in Table 2 by two annotators with adjudication (requiring a careful comparison of the two sets of annotations).

We performed two experiments: in the first, we evaluated the *gloss annotation task*, in the second the *property instance extraction task*, i.e. the ability to identify the appropriate domain and range of a property instance. In the case of the gloss annotation task, for evaluating each piece of information we adopted the measures of "*labeled" precision* and *recall*. These measures are commonly used to evaluate parse trees obtained by a parser (Charniak, 1997) and allow the rewarding of good partial results. Given a property *R*, labeled *precision* is the number of *words* annotated correctly with *R* over the number of words annotated automatically with *R*, while labeled *recall* is the number of words annotated correctly with *R* over the total number of words manually annotated with *R*.

Table 3 shows the results obtained by applying the checkers to tag the test set (containing a total number of 1,328 distinct annotations and 5,965 annotated words). Note that here we are evaluating the ability of the system to assign the correct tag to *every word* in a gloss fragment *f,* according to the appropriate relation checker. We choose to evaluate the tag assigned to single words rather than to a whole phrase,

[8] The resulting OWL ontology is available at http://lcl.di.uniroma1.it/tav

because each misalignment would count as a mistake even if the most part of a phrase was tagged correctly by the automatic annotator.

The second experiment consisted in the evaluation of the property instances extracted. Starting from 1,328 manually annotated fragments of 692 glosses, the checkers extracted an overall number of 1,101 property instances. We randomly selected a subset of 160 glosses for evaluation, from which we manually extracted 344 property instances.

Table 2. A subset of the relations from the CIDOC-CRM model

Code	Name	Domain	Range	Example
P26	moved to	Move	Place	P26(installation of public sculpture, public place)
P27	moved from	Move	Place	P27(removal of cornice pictures, wall)
P53	has former or current location	Physical Stuff	Place	P53(fancy pictures, London)
P55	has current location	Physical Object	Place	P55(macrame, Genoa)
P46	is composed of (is part of)	Physical Stuff	Physical Stuff	P46(lace, knot)
P62	depicts	Physical Man-Made Stuff	Entity	P62(still life, fruit)
P4	has time span	Temporal Entity	Time Span	P4(pattern drawings, Renaissance)
P14	carried out by (performed)	Activity	Actor	P14(blotted line drawings, Andy Warhol)
P92	brought into existence by	Persistent Item	Beginning of Existence	P92(aquatints, aquatint process)
P45	consists of (incorporated in)	Physical Stuff	Material	P45(sculpture, stone)

Table 3. Precision and Recall of the gloss annotation task

Property	Precision		Recall	
P26 – moved to	84.95%	(79/93)	64.23%	(79/123)
P27 – moved from	81.25%	(39/48)	78.00%	(39/50)
P53 – has former or current location	78.09%	(916/1173)	67.80%	(916/1351)
P55 – has current location	100.00%	(8/8)	100.00%	(8/8)
P46 – composed of	87.49%	(944/1079)	70.76%	(944/1334)
P62 – depicts	94.15%	(370/393)	65.26%	(370/567)
P4 – has time span	91.93%	(547/595)	76.40%	(547/716)
P14 – carried out by	91.71%	(343/374)	71.91%	(343/477)
P92 – brought into existence	89.54%	(471/526)	62.72%	(471/751)
P45 – consists of	74.67%	(398/533)	57.60%	(398/691)
Average performance	**85.34%**	**(4115/4822)**	**67.81%**	**(4115/6068)**

Two aspects of the property instance extraction task had to be assessed:

- the extraction of the appropriate *range words* in a gloss, for a given property instance
- the precision and recall in the extraction of the appropriate *concepts* for both *domain* and *range* of the property instance.

An overall number of 233 property instances were automatically collected by the checkers, out of which 203 were correct with respect to the first assessment (87.12% precision (203/233), 59.01% recall (203/344)).

In the second evaluation, for each property instance $R(C_t, C_w)$ we assessed the semantic correctness of both the concepts C_t and C_w. The appropriateness of the concept C_t chosen for the domain must be evaluated, since, even if a term t satisfies the semantic constraints of the domain for a property R, it still can be the case that a fragment f in G does not refer to t, like in the following example:

pastels (visual works) -- *Works of art*, typically on a paper or vellum support, to which designs are applied using crayons **made of ground pigment** held together with a binder, typically oil or water and gum.

In this example, *ground pigment* refers to *crayons* (not to *pastels*).

The evaluation of the semantic correctness of the domain and range of the property instances extracted led to the final figures of 81.11% (189/233) precision and 54.94% (189/344) recall, due to 9 errors in the choice of C_t as a domain for an instance $R(C_t, C_w)$ and 5 errors in the semantic disambiguation of range words w not appearing in AAT, but encoded in WordNet (as described in the last part of Section 3). A final experiment was performed to evaluate the generality of the approach presented in this paper.

As already remarked, the same procedure used for annotating the glosses of a thesaurus can be used to annotate web documents. Our objective in this third experiment was to:

- Evaluate the ability of the system to annotate fragments of web documents with CIDOC relations
- Evaluate the domain dependency of the relation checkers, by letting the system annotate documents not in the cultural heritage domain.

We then selected 5 documents at random from an historical archive and an artist's biographies archive[9] including about 6,000 words in total, about 5,000 of which in the historical domain. We then ran the automatic annotation procedure on these documents and we evaluated the result, using the same criteria as in Table 3.

Table 4 presents the results of the experiment. Only 5 out of 10 properties had at least one instance in the analysed documents. It is remarkable that, especially for the less domain-dependent properties, the precision and recall of the algorithm is still high, thus showing the generality of the method. Notice that the historical documents

[9] http://historicaltextarchive.com and http://www.artnet.com/library

influenced the result much more than the artist biographies, because of their dimension.

In Table 4 the recall of P14 (*carried out by*) is omitted. This is motivated by the fact that this property, in a generic domain, corresponds to the *agent* relation ("an active animate entity that voluntarily initiates an action"[10]), while in the cultural heritage domain it has a more narrow interpretation (an example of this relation in the CIDOC handbook is: "the painting of the Sistine Chapel (E7) was *carried out by* Michelangelo Buonarroti (E21) *in the role of* master craftsman (E55)"). However, the domain and range restrictions for P14 correspond to an agent relation, therefore, in a generic domain, one should annotate as "carried out by" almost any verb phrase with the subject (including pronouns and anaphoric references) in the class Human.

Table 4. Precision and Recall of a web document annotation task

Property	**Precision**		**Recall**	
P53 – has former or current location	79.84%	(198/248)	77.95%	(198/254)
P46 – composed of	83.58%	(112/134)	96.55%	(112/116)
P4 – has time span	78.32%	(112/143)	50.68%	(112/221)
P14 – carried out by	60.61%	(40/66)	-	-
P45 – consists of	85.71%	(6/7)	37.50%	(6/16)
Average performance	**78.26%**	**(468/598)**	**77.10%**	**(468/607)**

5 Related Work and Conclusions

This paper presented a method, based on the use of regular expressions, to automatically annotate the glosses of a thesaurus, the AAT, with the properties (conceptual relations) of a core ontology, the CIDOC-CRM. The annotated glosses are converted into OWL concept descriptions and used to enrich the CIDOC.

Several methods for ontology population and semantic annotation described in literature (e.g. (Thelen and Riloff, 2002; Califf and Mooney, 2004; Cimiano et al. 2005; Valarakos et al. 2004)) use regular expressions to identify named entities, i.e. concept *instances.* Other methods extract hypernym relations using syntactic and lexical patterns (Snow et al. 2005; Morin and Jaquemin 2004) or supervised clustering techniques (Kashyap et al. 2003). Evaluation of hypernymy learning methods is usually performed by a restricted team of experts, on a limited set of terms, with hardly comparable results, usually well over 40% error rate (Caraballo, 1999; Maedche et al, 2002). When the evaluation is an attempt to replicate the structure of an already existing taxonomy, the error rate is over 50-60% (Widdows, 2003).

As far as the adopted ontology learning technique is concerned, in our work we automatically learn *formal concepts* (not simply instances or taxonomies, as in the literature) compliant with the semantics of a well-established core ontology, the CIDOC. In AAT the hypernym relation is already available, since AAT is a thesaurus, not a glossary. However we developed regular expressions also for hypernym

[10] http://www.jfsowa.com/ontology/thematic.htm

extraction from definitions[11] (Velardi et al. 2006). When applying these patterns to the AAT (for sake of space this is not discussed in this paper) we found that in 34% of the cases the automatically extracted hypernym is the same as in AAT, and in 26% of the cases, either the extracted hypernym is more general than the one defined in AAT, or the contrary, wrt the AAT hierarchy. This result quite favorably compares with available results in the literature.

Semantic annotation with relations other than hypernymy are surveyed in (Reeve and Han, 2005), and again, regular expressions are a commonly used technique. Reeve and Han's survey presents a table to compare system's performance, but in absence of well-established data sets of annotated documents, a fair comparison among the various techniques is not possible. Similarly, comparing the performance of our system with those surveyed in (Reeve and Han, 2005) is not particularly meaningful.

The method presented in this paper is unsupervised, in the sense that it does not need manual annotation of a significant fragment of text. However, it relies on a set of manually written regular expressions, based on lexical, part-of-speech, and semantic constraints. The structure of regular expressions is rather more complex than in similar works using regular expressions, especially for the use of automatically verified semantic constraints. The issue is however how much these expressions generalize to other domains:

- A first problem is the availability of lexical and semantic resources used by the algorithm. The most critical requirement of the method is the availability of sound *core ontologies*, which hopefully will be produced by other web communities stimulated by the recent success of CIDOC CRM. On the other side, *in absence of an agreed conceptual reference model, no large scale annotation is possible at all.* As for the other resources used by our algorithm, glossaries, thesaura and gazetteers are widely available in "mature" domains. If not, we developed a methodology, described in (Navigli and Velardi, 2005b), to automatically create a glossary in novel domains (e.g. enterprise interoperability), extracting definition sentences from domain-relevant documents and authoritative web sites.
- The second problem is about the generality of regular expressions. Clearly, the relation checkers that we defined are tuned on the CIDOC properties, however many of these properties are rather general (especially locative and temporal relations) and could easily apply to other domains, as demonstrated by the experiment on automatic annotation of historical archives in Table 4. Furthermore, the method used to verify semantic constraints is fully general, since it is based on WordNet and a general-purpose, untrained semantic disambiguation algorithm, SSI.

Finally, the authors believe with some degree of convincement that automatic pattern-learning methods often require non-trivial human effort just like manual methods[12]

[11] In the referenced paper we apply hypernymy-seeking patterns to automatically learn a taxonomy from an (automatically extracted) glossary of terms in the field of enterprise interoperability. The results have been evaluated in the large by the members of the INTEROP EC network of excellence (http://www.interop-noe.org).

[12] A similar concern is expressed also in the concluding remarks of the already mentioned survey by Reeve and Han: "all SAPs [semantic annotation platforms] require some type of lexicon and resource. Rule-based systems require rules, pattern discovery systems require an intial set of seeds, machine learning system require a training corpus.".

(because of the need of annotated data, careful parameter setting, etc.), and furthermore they are unable to combine in a non-trivial way different types of features (e.g. lexical, syntactic, semantic). A practical example is the full list of automatically learned hypernymy-seeker patterns provided in (Morin and Jacquemin, 2004). The complexity of these patterns is certainly lower than the regular expression structures used in this work, and many of them are rather intuitive, they could have easily written by hand.

However, we believe that our method can be automated to some limited degree (for example, semantic constraints can be learned automatically), a research line we are currently exploring.

References

S. A. Caraballo “Automatic construction of a hypernym-labeled noun hierarchy from text” In 37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference, pages 120-126,1999

M. E. Califf and R.J. Mooney, “Bottom-up relational learning of pattern matching rules for information extraction” Machine Learning research, 4 (2)177-210, 2004

E. Charniak, “Statistical Techniques for Natural Language Parsing”, AI Magazine 18(4), 33-44, 1997

P. Cimiano, G. Ladwig and S. Staab, “Gimme the context: context-driven automatic semantic annotation with C-PANKOW” In: Proceedings of the 14th International WWW Conference, Chiba, Japan, May, 2005. ACM Press.

M. Doerr, “The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata”. AI Magazine, Volume 24, Number 3, Fall 2003.

M. S. Fox, M. Barbuceanu, M. Gruninger, and J. Lin, "An Organisation Ontology for Enterprise Modeling", In Simulating Organizations: Computational Models of Institutions and Groups, M. Prietula, K. Carley & L. Gasser (Eds), Menlo Park CA: AAAI/MIT Press, pp. 131-152. 1997

J.E. F. Friedl “Mastering Regular Expressions” O’Reilly eds., ISBN: 1-56592-257-3, First edition January 1997.

V. Kashyap, C. Ramakrishnan, T. Rindflesch. "Toward (Semi)-Automatic Generation of Biomedical Ontologies", Proceedings of American Medical Informatics Association, 2003

A. Maedche V. Pekar and S. Staab , “Ontology learning part One: On Discovering Taxonomic Relations from the Web” in In Web Intelligence. Springer, Chapter 1, 2002.

G. A. Miller, ``WordNet: a lexical database for English." In: Communications of the ACM 38 (11), November 1995, pp. 39 - 41.

E. Morin and C. Jacquemin “Automatic acquisition and expansion of hypernym links” Computer and the Humanities, 38: 363-396, 2004

R. Navigli and P. Velardi, “Structural Semantic Interconnections: a knowledge-based approach to word sense disambiguation”, Special Issue-Syntactic and Structural Pattern Recognition, IEEE TPAMI, Volume: 27, Issue: 7, 2005.

R. Navigli, P. Velardi. Automatic Acquisition of a Thesaurus of Interoperability Terms, Proc. of 16th IFAC World Congress, Praha, Czech Republic, July 4-8th, 2005b.

Reeve, L., & Han, H. (2005). Survey of Semantic Annotation Platforms. Proceedings of the 20th Annual ACM Symposium on Applied Computing, Web Technologies

R. Snow, D. Jurafsky, A. Y. Ng, "Learning syntactic patters for automatic hypernym discovery", NIPS 17, 2005.

M. Thelen, E. Riloff, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts", Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002.

M. Uschold, M. King, S. Moralee and Y. Zorgios, “The Enterprise Ontology”, The Knowledge Engineering Review , Vol. 13, Special Issue on Putting Ontologies to Use (eds. Uschold. M. and Tate. A.), 1998.

Valarakos, G. Paliouras, V. Karkaletsis, G. Vouros, “Enhancing Ontological Knowledge through Ontology Population and Enrichment” in Proceedings of the 14th EKAW conf., LNAI, Vol. 3257, pp. 144-156, Springer Verlag, 2004.

Paola Velardi, Alessandro Cucchiarelli and Michaël Pétit “Supporting Scientific Collaboration in a network of Ecellence through a semantically indexed knowledge map” I-ESA 2006, Bordeaux, France, March 2006

D. Widdows “Unsupervised methods for developing taxonomies by combining syntactic and statistical information” HLT-NAACL 2003, Edmonton, May-June 2003