Semi-automatic Parsing for Web Knowledge Extraction through Semantic Annotation

Maria Pia di Buono
University of Salerno, DSPSC
Via Giovanni Paolo II, 182 – 84084 Fisciano (SA), IT
mdibuono@unisa.it

Abstract

Parsing Web information, namely parsing content to find relevant documents on the basis of a user's query, represents a crucial step to guarantee fast and accurate Information Retrieval (IR).
Generally, an automated approach to such task is considered faster and cheaper than manual systems. Nevertheless, results do not seem have a high level of accuracy, indeed, as also Hjorland (2007) states, using stochastic algorithms entails low precision, low recall and generic results.
Usually IR systems are based on invert text index, namely an index data structure storing a mapping from content to its locations in a database file, or in a document or a set of documents.
In this paper we propose a system, by means of which we will develop a search engine able to process online documents, starting from a natural language query, and to return information to users.
The proposed approach, based on the Lexicon-Grammar (LG) framework and its language formalization methodologies, aims at integrating a semantic annotation process for both query analysis and document retrieval.

Keywords: Knowledge Extraction, Semantic Annotation, Parsing

1. Introduction

Parsing Web information, namely parsing content to find relevant documents on the basis of a user's query, represents a crucial step to guarantee fast and accurate Information Retrieval (IR).
Generally, an automated approach to such task is considered faster and cheaper than manual systems. Nevertheless, results do not seem have a high level of accuracy, indeed, as also Hjorland (2007) states, using stochastic algorithms entails:

- Low precision due to the indexing of common Atomic Linguistic Units (ALUs) or sentences.
- Low recall caused by the presence of synonyms.
- Generic results arising from the use of too broad or too narrow terms.

Usually IR systems are based on invert text index, namely an index data structure storing a mapping from content to its locations in a database file, or in a document or a set of documents.
Most traditional IR systems process each document separately to retrieve terms in free-text query, which means that they do not compare results provided from different sources.
Such lack of integration in results causes overlapping and decreasing in the positive predictive value, due to the fact that shared content are indexed several times. Various approaches have been proposed to overcome this boundary, increasing recall and precision in results.
For example, Lempel from Yahoo! Labs deals with query evaluation strategies which are based on Term-at-a-Time (TAAT) and Document-at-a-Time Evaluation (DAAT) processing.
Furthermore, different researches employ concept-based and semantic approach in order to process both documents and queries through semantic entities and concepts.
Baziz et al. (2005) and Boubekeur et al. (2010) describe their methods for assigning document ALUs to the correct ontological entries.
Boubekeur and Azzoug (2013) propose an approach for semantic indexing based on concepts identified from a linguistic resource. In their work, authors use WordNet and WordNetDomains lexical databases with the aim to identify concepts and they also apply a concept-based indexing evaluation.
In this paper we propose a system, by means of which we will develop a search engine able to process online documents, starting from a natural language query, and to return information to users.
The proposed approach applies the Lexicon-Grammar (LG) framework and its language formalization methodologies, developed by Maurice Gross during the '60s.

2. Methodology

As presented in di Buono (2014, 2015), we have developed the Archaeological Italian Electronic Dictionary (AIED). We also develop other Italian LRs, namely Finite State Automata/Finite State Transducers (FSA/FSTs) (Silberztein, 2013, 2015) and LG tables, for the Archaeological Domain, starting from the NooJ [1] module for Italian. Such module and its relate resources have been created and maintained by the research group of University of Salerno[2] , under the LG framework.
Our approach aims at integrating a semantic annotation

[1] www.nooj-association.org.
[2] http://labgross.unisa.it/.

process for both query analysis and document retrieval. Indeed, semantic annotation represents a key step in our procedure, in order to annotate texts matching correctly a natural language formalism and a data model formalism.
The system workflow is based on representation models applied to all resources, which represent objects of linguistic processing, namely Knowledge Bases (KBs), Web pages and full texts. Therefore, we develop an architecture, which takes advantage from the semantic information stored in Linguistic Resources (LRs) and is based on the integration of NooJ. Such system architecture integrates NooJ into a Web application in order to (re)use the representation models.

Our system recognizes RDF predicate in sentence structures and electronic dictionaries entries (simple words and ALUs) which are the subject and the object of the RDF triple.
In the following image, we will show a sample of FSA/FST which may be used to analyze users' queries. Such automaton allows us to recognize entities involved in RDF relationships, namely Person and date. In such RDF triple, the subject, *Person*, and the object, *date*, are trigged by a predicate, namely a Verb Phrase (VP). This VP is represented by a class verb which may co-occur together with the given entities in sentence contexts. Therefore, the VP may hold verbs such as *live* or *born* followed by a preposition. It is worth noticing that in our

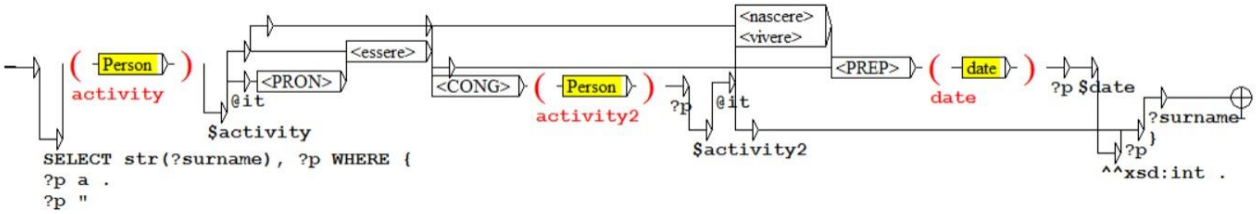

Figure 1: FSA for annotating users' queries.

3. Semantic Annotation

Please In developing our LRs, we model data semantically using two kinds of conceptual schemata: an upper level ontology, namely a cross-domain ontology, and a specific-domain ontology.

- DBpedia (upper level) ontology which is composed of:
  Classes: 734
  Properties: 2975
- CIDOC Conceptual Reference Model CRM (domain) ontology which is composed of:
  Classes: 90
  Properties: 148.

In our system, we present a semantic annotation process which works simultaneously on two sides: it analyses (I) the user's query (natural language analysis), and (II) documents stored in KBs (data representation).
Thus, we propose an architecture, which takes advantage from semantic information stored both in electronic dictionaries and FSA/FSTs. Furthermore, this architecture may also map linguistic tags (i.e. POS) and structures (i.e. sentences, ALUs) to domain concepts employing metadata from conceptual schemata.

3.1 Representation Model

The first task concerns the processing of user's queries in order to annotate them, domain-independent semantic data modelling (DBpedia cross-domain ontology) and inferring Boolean relationship among elements in a free-text query and relative meta-data. Starting from the entries retrieved and from their specific tags, stored in electronic dictionaries and in FSA/FSTs.

sample we insert two nodes containing the same entity (Person), which stands for two different variables, namely *activity* and *activity2*. Such variables refer to a specific CIDOC CRM class tag, which is used to identify a specific attribute, namely profession, for elements belonging to the generic class *Person*.
Values, produced by variables (activity, activity2 and date), are employed to generate a SPARQL query, able to retrieve surname of such persons which perform a specific activity/job/profession in a determinate interval.
In other words, the previous automaton may process a query as the following one:

(1) *Tutti gli archeologi che sono stati anche scrittori nati nel '900* (All the archaeologists who have been also writers and were born in 19th century)[3].

The following sample shows the result of FSA applied to the previous query

```
SELECT str(?surname), ?p WHERE {
      ?p a .
      ?p "scrittore"@it .
      ?p "archeologo"@it .
      ?p "1900"^^xsd:int .
      ?p ?surname
}
```

[Example of pseudo-code query in SPARQL which may be used into an Endpoint]

Thus, the output of FSA may be used in order to generate

[3] Such example is adapted from the one proposed on the page of Italian DBpedia.
http://it.dbpedia.org/esempi/.

a query which may be run against any SPARQL endpoint or repository in which documents are formalized using RDF.
The second subtask, namely data representation, involves appropriate operations on the RDF-based data layer, mapping OWL concepts to object-oriented classes with methods for interrelations and domain-specific rules used to generate and consolidate processes.
Such process of data representation aims at analysing information stored in RDF documents, which means that we may retrieve information from any repository directly. Actually, we use RDF data representation in order to process documents and create a match between users' requests and concepts stored in KBs.
We develop NooJ FSA in order to process information stored in DBpedia KB, matching values of semantic attributes with the ones retrieved from users' queries analysis.

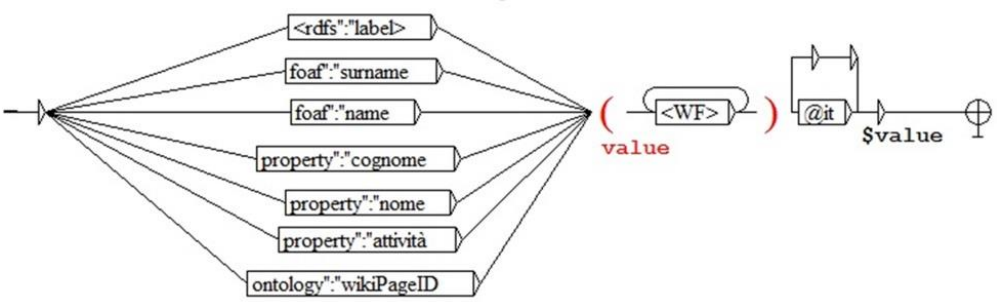

Figure 2: Sample of FSA for analyzing DBpedia documents.

In the previous FSA (Figure 2) we use the nodes on the left in order to recognize labels used inside RDF documents, which are stored, for example, in DBpedia KB. It means that first we process tags which describe elements semantically and subsequently we analyze which values are assumed for such descriptions (Table1). Actually, in the following node, we insert a generic WF class, in order to recognize each word form which is present inside documents. These word forms represent values stored for each specific semantic descriptive tag, i.e. for foaf:surname Levi value.
On the other hand, the final node @it indicates language tag in resource description schemata (i.e., Italian).

4. Future Work and Conclusion

In our experiment, we test DBpedia database as knowledge source of structured data in RDF/XML and we test our system outputs using its SPARQL (Protocol and RDF Query Language) Endpoint.
Thus, if we run the given query against the Italian BDpedia Endpoint, we obtain a list of results which match with user's information need (Table 2).
After being tested and debugged, the LRs described so far are actually under final development and completion and they will be proposed as part of the NooJ Italian module.
We will integrate such LRs into an web environment[4], considering that our final aim is to propose the development of a SPARQL endpoint based on NooJ.
All the annotations produced by the application of our method and resources can be reused to enrich lexical databases or ontologies referred to the CH domain. Noticeably, the size and quality of the enrichment is strictly dependent on the largeness and on the content of the corpus on which the NooJ resources are applied. Therefore, in order to obtain widespread CH databases, it is preferable to use corpora able to cover the larger group of CH domain possible.
Our future research work aims at integrating different RDF formats in the parser and writer registries, i.e. Turtle, JSON-LD, RDF/JSON and so on.
Future work also aims at integrating manually constructed rules with supplementary rules, in order to improve not-probable word removal. In addition, we are planning to develop grammars useful to recognize discontinuous expressions inside NPs and VPs, and to implement an anaphora-resolution task.

[4] Endpoint for Semantic Knowledge (ESK) is the beta-version of a Web environment based on the proposed approach.
http://dsc.unisa.it/mariapiadb/esk/project.html.

Path	Output
<rdfs:label> Peter Levi @it</rdfs:label>	Peter Levi @it
<ontology:wikiPageID>2168662</ontology:wikiPageID>	2168662
<foaf:name> Peter Chad Tigar @it </foaf:name>	Peter Chad Tigar @it
<ontology:wikiPageLength>11646 </ontology:wikiPageLength>	11646
<ontology:birthYear>1931</ontology:birthYear>	1931
<ontology:deathYear>2000</ontology:deathYear>	2000
<foaf:surname> Levi @it </foaf:surname>	Levi @it
<property:nome> Peter Chad Tigar @it </property:nome>	Peter Chad Tigar @it
<property:cognomen> Levi @it </property:cognomen>	Levi @it
<property:sesso> M @it </property:sesso>	M @it
<property:attività> Scrittore @it</property:attività> <property:attività> … </property:attività>	Scrittore @it … Poeta @it … Archeologo @it

Table 1: Results from the analysis of DBpedia documents.

Name/Surname Value	Resource
Peter Levi	http://it.dbpedia.org/resource/Peter_Levi
Paolo Matthiae	http://it.dbpedia.org/resource/Paolo_Mattiae
Thorkild Hansen	http://it.dbpedia.org/resource/Thorkild_Hansen
Glenn Cooper	http://it.dbpedia.org/resource/Glenn_Cooper
Alfred Duggan	http://it.dbpedia.org/resource/Alfred_Duggan
Max Mallowan	http://it.dbpedia.org/resource/Max_Mallowan
Almerico Meomartini	http://it.dbpedia.org/resource/Almerico_Meomartini
Michael Coe	http://it.dbpedia.org/resource/Michael_D._Coe
Thanos Kondylis	http://it.dbpedia.org/resource/Thanos_Kondylis
Vincenzo Zecca	http://it.dbpedia.org/resource/Vincenzo_Zecca
En Bellis	http://it.dbpedia.org/resource/En_Bellis
Sebastiano Consoli	http://it.dbpedia.org/resource/Sebastiano_Consoli

Table 2: Results from the analysis of DBpedia documents.

5. References

Baziz, M., Boughanem, M., Aussenac-Gilles, N. (2005). A Conceptual Indexing Approach based on Document content Representation. In *CoLIS5: Fifth International Conference on Conceptions of Libraries and Information Science*, Glasgow, UK, 4 Berlin Heidelberg : Springer-Verlag, pp. 171--186.

Boubekeur, F. and Azzoug, W. (2013). Concept-based indexing in text information retrieval, *International Journal of Computer Science & Information Technology (IJCSIT)* Vol 5, No 1, February 2013

Boubekeur F., Boughanem M., Tamine L., Daoud M., (2010). Using WordNet for Concept-based document indexing in information retrieval. In *Fourth International Conference on Semantic Processing (SEMAPRO)*, Florence, Italy, October 2010.

di Buono, M.P. (2015). Information Extraction for Ontology Population Tasks. An Application to the Italian Archaeological Domain. *International Journal of Computer Science: Theories and Applications*, Vol 3, No 2 (2015). ORB Academic Publisher, pp. 40-50 (2015).

di Buono, M.P., Monteleone, M., Elia, A. (2014). Terminology and Knowledge Representation Italian Linguistic Resources for the Archaeological Domain. In *Proceedings of 25th International Conference on Computational Linguistics (COLING 2014)* - Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014) (2014).

Hjorland, B. (2007). *Semantics and Knowledge Organization.* Annual Review of Information Science and Technology 41, pp. 367--405.

Lempel R. (2010), http://webcourse.cs.technion.ac.il/236621/Winter2010-2011/ho/WCFiles/lec4-evaluation.pdf.

Silberztein, M. (2015). *La Formalisation des Langues. L'Approche de NooJ*. ISTE Edition.

Silberztein, M. (2013). NooJ Computational Devices. In Donabédian A., Khurshudian V. and Silberztein M. (eds.): *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2012 International Conference.* Cambridge Scholars Publishing. Newcastle.