Extracting domain ontologies from Italian texts: an approach with Finite State Transducers

{Edoardo Salza} edoardo.salza@gmail.com - LiCoTT Università del Piemonte Orientale

Abstract

Ontologies play an important role in different fields of technology, so the need of automatic extraction system is increasingly felt. This is primarily due to the fact that most information is scattered between data in the form of raw text written in natural languages, therefore NLP techniques are fundamental to the goal of analyzing sentences for the purpose of extracting information.

Here is presented such an approach using a combination of finite state automata and linguistic resources used to extract in a semi-automatic way a basic structure of a domain ontology.

Introduction

I have developed a NLP-based ontology extraction method suitable for implementing information analysis systems in specific domains, suitable to carry out tasks like automatic document classification.

The system extracts the taxonomy (the backbone of an ontology) and the basic relations between the represented concepts starting from plain text. This approach uses finite-state transducers both to recognize the terminology and relations and to map them into an ontology and needs the training corpus (used to elicit the rules) to be small enough to allow a trained human reader to analyze it in a reasonable time.

In specific domains the availability of large training corpora is often low, so a rule-based method is recommended. Moreover, such a system is easily adaptable to a variety of specific domains without the need of retraining, while changing the related language resource.

A deterministic rule-based method is also well suitable to recognize idiosyncratic linguistic phenomena. The drawbacks of this approach are the need of comprehensive and standardized language resources, longer developing time due to the need to manually analyzing the corpus, time needed to write and implement extraction and mapping rules and to test and debug the system.

The method uses UNITEX software to implement RTN (Recursive Transition Network) to recognize language patterns and to map them into ontology elements. This software is based on INTEX, a similar development environment developed by Max Silberztein [2].

References

[1] Gross M. (1993). Local grammars and their representation by finite automata.

[2] Silberztein M. (1993). Dictionnaires électroniques et analyse automatique de textes : le système INTEX.

[3] Saias J.; Quaresma P. (2003). A methodology to create ontology based information retrieval system.

[4] Voghera M. (2004). Polirematiche. In Grossmann M., Rainer F., La formazione delle parole in italiano.

Method

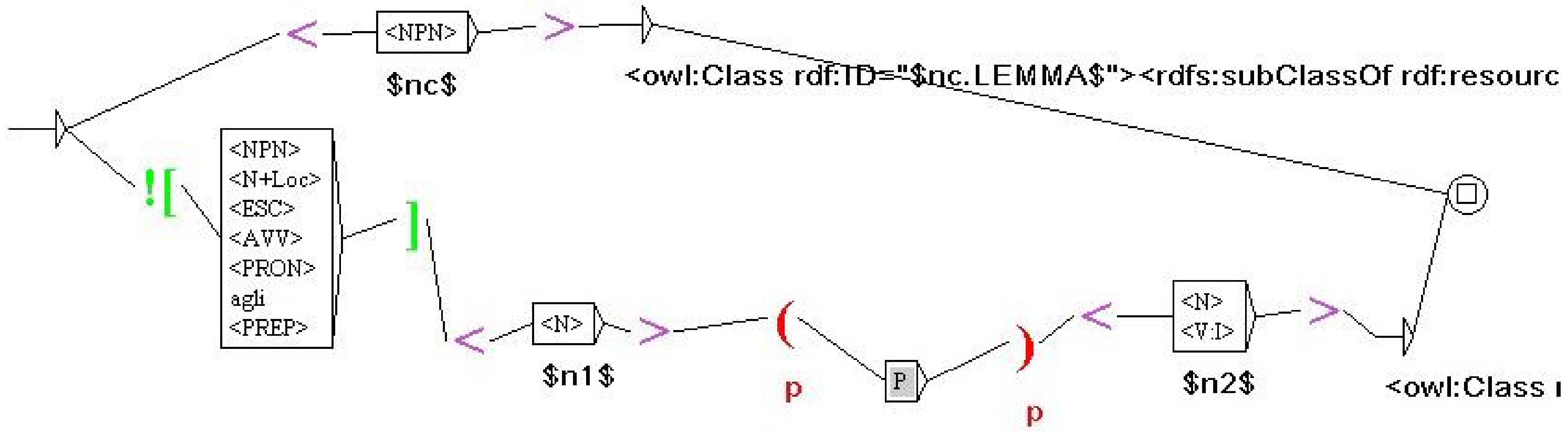
The Finite State Automata are used here for recognizing patterns and to map them into the ontology: patterns are represented by FSA using a context free grammar formalism involving lexical, syntactical and semantical elements called *local grammar*, first introduced by the linguist Maurice Gross [1].

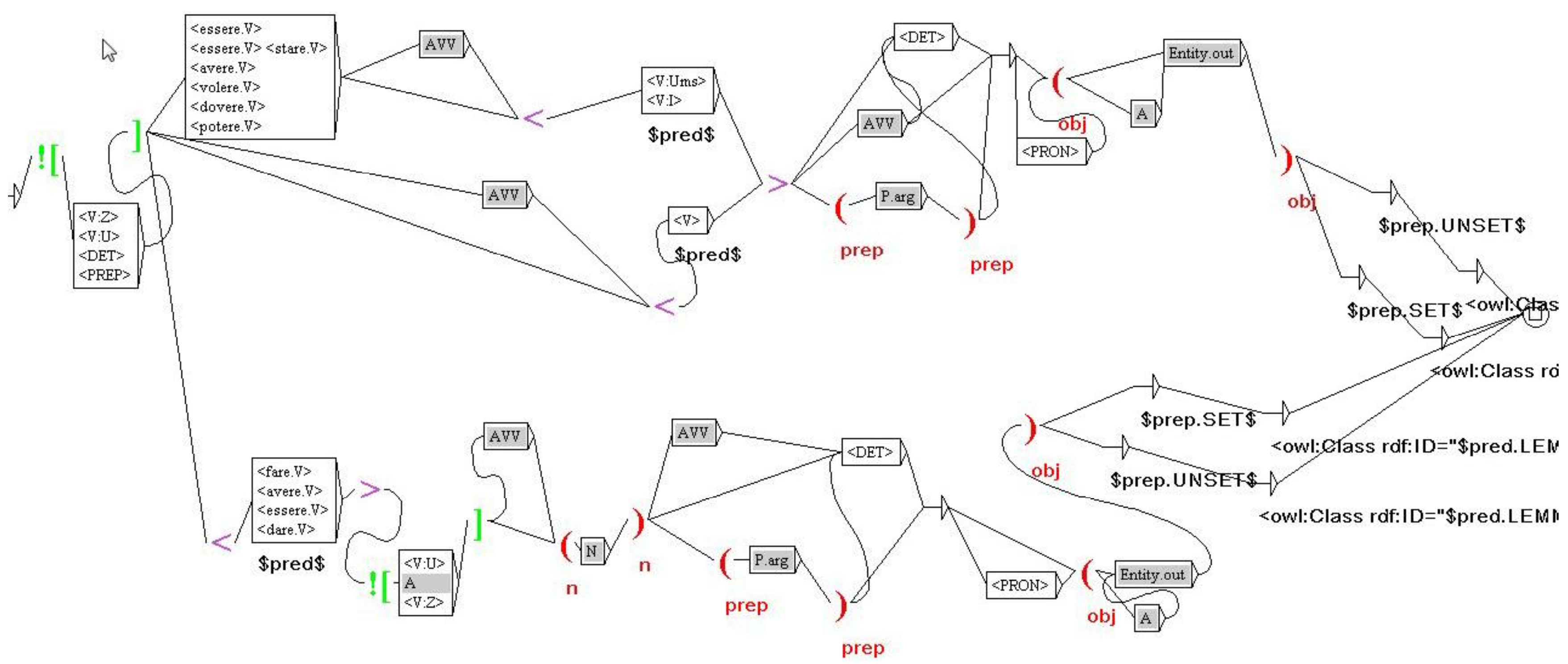
The first step is to compile a domain dictionary of compound nouns (DELAC), that form the bulk of the domain lexicon. The dictionary should be readable from the automata in order to check the searched constructions, so the chosen format is the DELA (*Dictionnaires Eléctronique LADL*, a standard proposed by the linguist Maurice Gross for his RELEX international network. This dictionary is a first step in order to disambiguate between compound nouns that form a semantic or syntactic unit (and need to be mapped as a singular entry) and those who are not. In the latter case the head of the compound should be mapped into a class while the whole entry will be mapped as a subclass of the respective head. The ontology extraction method works in the following way: first we should set up manually a TBox to host the ontology elements as the following [3]: $\exists(a, e) \wedge Entity(e) \wedge Action(a) \wedge subject(a,e) \wedge object(a,e)$. The system can then expand automatically the TBox with new subclasses of *Entity* and it recognizes both simple and compound nouns. The searched elements match a set of most frequent patterns found in italian [4] like: *N ADJ* (e.g. *adempimenti burocratici*) or *N Prep N* (e.g. *piano di studi*). The system also recognizes an expansion of this set consisting in more complex pattern such as *N ADJ Prep N* (e.g. *numeri telefonici di recapito*) or *N ADJ Prep N Prep N* (e.g. *stage urgente della studentessa di LIRAP*). This is done using an appropriate FSA. The mapping of the extracted terminology is done as the following:

- Entity $\sqsubseteq$ *piano di studi*: if the compound noun is in the DELAC dictionary as a monolithic fixed form
- Entity $\sqsubseteq$ *progetto* $\sqsubseteq$ *progetto VOICE*: the noun is not in the dictionary, thus is not considered a monolithic unit

Finally, the ontology is populated with the predicate-argument pairs to extract the information about the action performed by the verb and their participants. Each recognized verb is mapped into a subclass of the upper class *Action*. For each sentence an instance of the respective verb subclass is generated: each instance will have a specific value for the properties *Subject* and *Object* linking the *Action* subclasses to the taxonomy extracted above. This allows in applications like automatic document classification to have an immediate feedback of the main elements involved (i.e. the main topics).

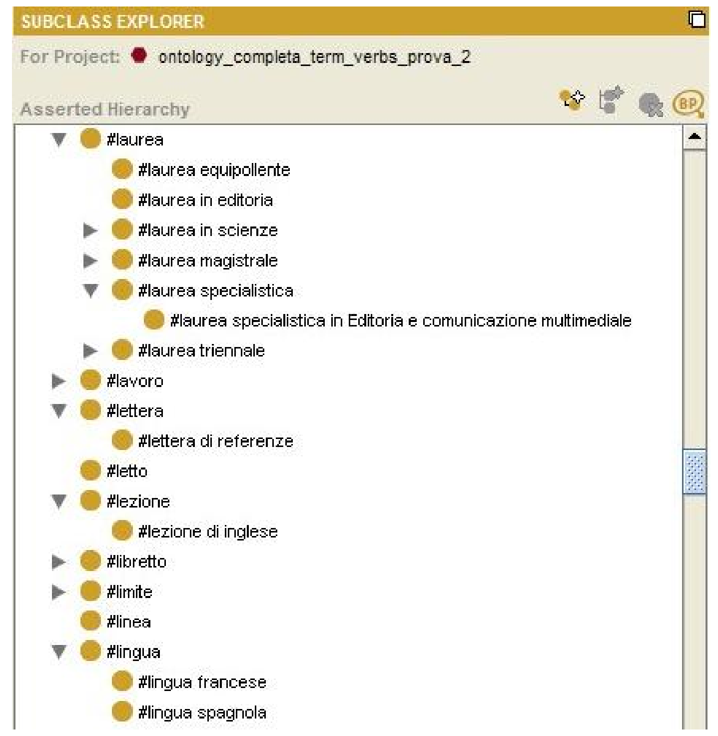
The system recognizes the following clause patterns (or verb phrases): standard SVO active voice (e.g. *scrivere una lettera di referenze*), passive voice (e.g. *la data della prova è stata concordata*), relative and infinitive clauses (e.g. *esami che ho sostenuto*) and topicalized clauses with pronominal resumption (e.g. *La bozza della lettera gliel'ho spedita*). An example automaton used for extracting a pattern (*N Prep N*) and an automaton used for recognize predicate-argument pairs in an active clause are showed below. Note that the automata put out directly the ontology code in OWL format:

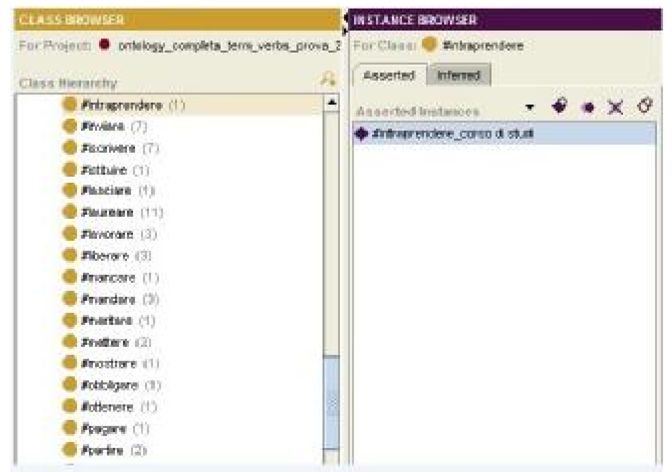




Extracted Ontology

Here an example of the taxonomy and of the predicate-argument structures extracted from a small corpus of emails collected in an academic context viewed with software Protégé.





Conclusions

This method, while extracting the verbal structure of a sentence, is suitable to many different applications. It could be used not only for automatic document classification (e.g. emails) but could be also useful also to populate language resources (e.g. FrameNET), where it could be used to search lexical units according to specified patterns or to accomplish tasks as semantic role labelling.

A more semantically specified TBox could then be used to extract a more complex taxonomy involving different categorization levels. An approach could be to write automata for each verb (or verb class) of the domain of interest using the respective subcategorization information. It should be then possible to implement FSA to search complex dependencies. Also, considering information about reflexivity, inchoativity, change of state, etc. embedded in text could allow to extract more complex properties in order to produce more expressive ontologies.