Semantic mark-up of Italian legal texts through NLP-based techniques

Roberto Bartolini°, Alessandro Lenci§, Simonetta Montemagni°, Vito Pirrelli°, Claudia Soria°

°Istituto di Linguistica Computazionale – CNR
Via Moruzzi 1 - 56100 Pisa, Italy
§Department of Linguistics
University of Pisa
Via S. Maria, 36 – 56100 Pisa, Italy
firstname.lastname@ilc.cnr.it

Abstract

In this paper we illustrate an approach to information extraction from legal texts using SALEM. SALEM is an NLP architecture for semantic annotation and indexing of Italian legislative texts, developed by ILC in close collaboration with ITTIG-CNR, Florence. Results of SALEM performance on a test sample of about 500 Italian law paragraphs are provided.

1. Introduction

The huge amount of documents available in the legal domain calls for computational tools supporting search and filtering of information. As a matter of fact, the overwhelming majority of processes for legal content creation constantly produce volumes of written material that is poorly structured, not easily predefined and variably associated with information. Laws are a clear example of this state of affairs. To our knowledge, very few tools are available for the automatic management of Italian law texts. Some of them, such as the tool described in (Bolioli *et al.*, 2002), are used for the automatic recognition of structural elements of the law text, and allow for intra- and inter-textual browsing of documents. Current legal knowledge management tools, however, are usually limited to formal and structural analyses of texts, while the need is felt for automatic or semi-automatic systems that carry out a *semantic* analysis of texts, thus providing a representation of their content. Notable exceptions are the DIAsDEM system (Graubitz *et al*., 2001) and, albeit not restricted to specific domain texts, the approach of De Busser *et al*. (2002). Under many respects our approach is similar to the one adopted by Sayas and Quaresma (2003), who exploit NLP techniques to yield a syntactic annotation of law texts and populate a legal ontology.

SALEM (Semantic Annotation for LEgal Management), the NLP system illustrated in the following pages and currently used as an advanced module of the NIR[1] legal editor (Biagioli *et al.*, 2003), has the potential of automatically tagging the semantic structure of Italian law paragraphs through an integration of NLP and information extraction-inspired technology.

2. Methodology and motivations

We model legal text semantic mark-up as a by-product of information extraction, intended as indicated by MUC as "the extraction of information from a text in the form of text strings and processed text strings which are placed into slots labelled to indicate the kind of information that can fill them"[2]. In particular, we automatically extract information about the *legislative provision* contained in a law paragraph and the *legal entities* (i.e. actors, actions and properties) referred to therein. Our task is therefore twofold: a) assign each law paragraph to a given provision type; b) automatically tag the parts of the paragraph with domain-specific semantic roles identifying the entities referred to in the legislative provision.

Automatic identification of the provision type expressed by a law paragraph is important for effective management of law texts. Law databases could be queried through fine-grained "semantic" searches according to the type of legal event reported by a law paragraph. Furthermore, automatic extraction of text portions of law that are subject to modifications could enable (semi)automatic updating of law texts, or make it possible for the history of a law to be traced throughout all its modifications; the original referenced text could be imported and modified, etc. Finally, automatic assignment of the relevant paragraph parts to semantic slots is bound to have an impact on effective legal content management and search, allowing for fine-grained semantic indexing and query of legal texts, and paving the way to real-time analysis of legal corpora in terms of logical components or actors at the level of individual provisions. In the near future, it will be possible to search an on-line legislative corpus for all types of obligation concerning a specific subject, or to highlight all possible legislative provisions a given action or actor happens to be affected by.

3. The legal text

As textual units, (Italian) laws are typically organized into hierarchically structured sections, the smallest one being the so-called *law paragraph*. Law paragraphs are usually numbered sections of an article, as in Example 1 below:

Article 6
1. The Commission shall be assisted by the committee set up by Article 5 of Directive 98/34/EC.
2. The representative of the Commission shall submit to the committee a draft of the measures to be taken.

Example 1

[1] NIR ("Norme in Rete", *Laws on the web*) is a national project sponsored by the Ministry of Justice for the free access by citizens to Italian jurisdiction. The DTDs defined by the project are the current standard format for the encoding of Italian legislative texts.

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc.

From the point of view of the content of a law, a law paragraph is associated with a particular legislative provision, which could somehow be seen as the illocutionary point of a law section. For instance, a paragraph may express an obligation for some actor to perform or not to perform a certain action, as in Example 1 above. Similarly, a paragraph may express a permission or an obligation as in Examples 2 and 3.

DirectiveA Member State may provide that a legal body the head office of which is not in the Community may participate in the formation of an SCE provided that legal body is formed under the law of a Member State, has its registered office in that Member State and has a real and continuous link with a Member State's economy.

Example 2: A Permission

Licence applications shall be accompanied by proof of payment of the fee for the period of the licence's validity.

Example 3: An Obligation

Law paragraphs may also have an inter-textual content, i.e. they can contain some sort of amendments to existing laws. In this case they are said to be *Modifications*. For instance, a paragraph may contain an insertion with respect to another law, or a replacement, or a repeal, as the following examples illustrate.

The following point shall be inserted after point 2g (Council Directive 96/61/EC) in Annex XX to the Agreement: "2h. 399 D 0391: Commission Decision 1999/391/EC of 31 May 1999 concerning the questionnaire relating to Council Directive 96/61/EC concerning integrated pollution prevention and control (IPPC) (implementation of Council Directive 91/692/EEC) (OJ L 148, 15.6.1999, p. 39)."

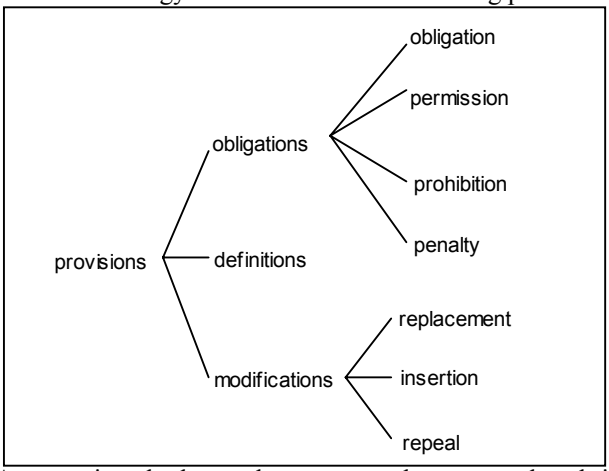
Example 4: an Insertion

The text of point 2eg (Commission Decision 95/365/EC) in Annex XX to the Agreement shall be replaced by the following: "399 D 0568: Commission Decision 1999/568/EC of 27 July 1999 establishing the ecological criteria for the award of the Community eco-label to light bulbs (OJ L 216, 14.8.1999, p. 18)."

Example 5: A Replacement

4. A frame-based legal ontology

SALEM includes a small ontology of legislative provisions types. The ontology distinguishes three major categories of provisions: *obligations*, *definitions* and *modifications*. A main distinction can be made between obligations, addressing human actors, and modifications, which is rather aimed at modifying the textual content of pre-existing laws. Obligations in turn divide into the following classes: *obligation*, *prohibition*, *permission*, and *penalty*. In their turn, modifications are further subdivided into *replacement*, *insertion* and *repeal*. The ontology was developed by a pool of experts in the legal domain at ITTIG-CNR (Florence). The taxonomical structure of the SALEM ontology is illustrated in the following picture:



As mentioned above, law paragraphs are analysed in SALEM not only according to the particular type of legislative provision they express, but also with respect to the main legal entities involved by the law. Consistently, the classes in the SALEM ontology are formally defined as *frames* with a fixed number of (possibly optional) slots corresponding to the semantic roles played by the legal entities specified by a given provision type. For instance, in Example 1 above, which expresses an obligation, the relevant roles in the first sentence of paragraph 2 are the *addressee* of the obligation (i.e. *The representative of the Commission*), the *action* (what the addressee is obliged to do, in this case *submit to the committee a draft of the measures to be taken*) and, optionally, a *third_party* (the action recipient, here *the committee*). In a similar way, a modification such as an insertion can have up to four relevant roles: (1) the reference text being modified, or *rule* (in Ex. 4 above, the text *(Council Directive 96/61/EC) in Annex XX to the Agreement),* (2) the *position* where the new text is going to be inserted (here, *after point 2g*); (3) the new text or *novella* (here, the captioned text); (4) the verbatim text to be replaced by the *novella* (*novellato*, not occurring in the example above).
The slot types required for the description of the 8 bottom classes in the SALEM taxonomy are illustrated in Table 1:

Provision class	**Slots**
Obligation	Addressee Action Third-Party
Permission	Addressee Action Third-Party
Prohibition	Action Third-Party
Penalty	Addressee Action Object Rule
Definition	Definiendum Definiens
Repeal	Rule Position *Novellato*

Provision class	Slots
Replacement	Rule Position *Novella* *Novellato*
Insertion	Rule Position *Novella*

Table 1: Frame-based description of the different provision types

5. SALEM architecture

General overview

SALEM is a suite of NLP tools for the analysis of Italian texts (see Bartolini *et al.*, 2002), specialized to cope with the specific stylistic conventions of the legal parlance, with the aim to automatically classify law paragraphs and identify legal entities.

A first prototype of SALEM has just been brought to completion and its performance evaluated. The NLP technology put to use is relatively simple, but powerful, also thanks to the comparative predictability of law texts.

SALEM takes in input single law paragraphs in raw text and outputs a semantic tagging of the text, where its classification together with the semantic roles corresponding to the different frame slots are rendered as XML tags. An output example (translated into English for the reader's convenience) follows:

<obl:addressee> The Member State </obl:addressee> shall <obl:action> pay the advance within 30 calendar days of submission of the application for advance payment </obl:action>.

Example 6: SALEM output example

where it can be seen that the input paragraph has been classified as an obl(igation) and portions of the text have been assigned to the addressee and action slots.

In SALEM, the approach we adopted for the semantic mark-up of legal texts follows a two-stage strategy. During the first step, a general purpose parsing system, which has been specialized to handle Italian legal texts, provides an organized structure corresponding to an initial syntactic analysis of each law paragraph: at this stage, the input text is tokenized, lemmatized, POS-tagged and shallow parsed into non-recursive constituents called "chunks". During the second step, chunks are fed into the semantic annotation component, a specialized version of the ILC finite-state compiler of grammars for functional analysis, with the result of deriving and making explicit the information implicitly stored in provisions.

Reasons for using chunks

With the benefit of the hindsight, we can say that a chunked text is an optimal starting point for semantic annotation of legal texts for different reasons. Besides its robustness and flexibility in the face of parse failures (which remain local), chunking combines low level textual features (e.g. indication of punctuation) with a first level of syntactic grouping which is instrumental for the identification of deeper levels of linguistic analysis (e.g. dependency-based).

As a matter of facts, semantic mark-up of legal text along the lines described above requires simultaneous consideration of both low level textual features like punctuation, and functional syntactic roles such as subject, object, and indirect object. In the analysis of modifications, a crucial role is played by punctuation marks, in particular by quotes and colons, which can effectively be used to identify the text of the amendment (*novella*) and the amending text (*novellato*). On the other hand, the mark-up of both modifications and obligations requires knowledge of the syntactic structure underlying the provision text. To give the reader but one example, the addressee of an obligation typically corresponds to the syntactic subject of the sentence, while the action (s)he is obliged to carry out is usually expressed as an infinitival clause, as in the example below:

Il comitato misto e' tenuto a raccomandare modifiche degli allegati secondo le modalita' previste dal presente accordo
[The Joint Committee shall be responsible for recommending amendments to the Annexes as foreseen in this Agreement].

Example 7

Note, however, that this holds only when the verbal head of the infinitival clause is used in the active voice. By contrast, the syntactic subject can express the third-party if the action verb is used in the passive voice and is governed by specific lexical heads.

Summing up, there are three main advantages in taking chunked syntactic structures as the starting point of semantic mark-up of legal texts. First, at this stage information about punctuation is still available, whereas this information type is typically lost at further analysis levels. Second, chunked representations can profitably be used as the starting point for partial functional analyses, aimed at reconstructing the range of functional relations within the provision text, that are instrumental in the annotation of legal entities. Last but not least, chunking does not "balk" at domain-specific constructions that do not follow general grammar rules; rather it actually carries on parsing, while leaving behind an ill-formed chunk unspecified for its category.

Semantic mark-up component

The chunked representation of each law paragraph is fed into the semantic annotation component proper, which is responsible for populating the legal ontology through a two-step semantic markup:

1. each paragraph is assigned to an ontology class (corresponding to the legislative provision expressed in the text);
2. slots of the class identified at step (1) are turned into an extraction template and instantiated through specific parts of the law paragraph.

The current version of the semantic annotation prototype uses a specialized grammar including (i) a core group of syntactic rules for the identification of basic syntactic dependencies (e.g. subject and object), and (ii) a battery of

semantic mark-up rules covering the classes of law provisions of Table 1 above.
Rules in the grammar are written according to the following template:

chunk-based regular expression
WITH {battery of tests} => {actions}.

Patterns of structural conditions, expressed through regular expressions over sequences of chunks, often combine with lexical conditions defined through a battery of tests, which include possible specification of a given dependency role. The action type ranges from identification of basic dependency relations to XML annotation of entire law paragraphs.

6. Evaluation results

SALEM preliminary results are very encouraging. The system has been tested on a sample of 473 law paragraphs, covering 7 ontology classes of Table 1. The test corpus was built and hand-annotated by law experts at ITTIG-CNR. The aim of the evaluation was to assess the system's performance on two tasks: paragraph classification and semantic role mark-up.
Table 2 below summarizes the results achieved for the paragraph classification task wrt the following 7 bottom classes of provisions:

provision class	tot	SALEM			
		answers	ok	recall	prec
Prohibitions	15	15	14	93,33%	93,33%
Permissions	15	18	15	100,00%	83,33%
Obligations	19	19	18	94,74%	94,74%
Penalties	122	117	109	89,34%	93,16%
Repeals	70	69	69	98,57%	100,00%
Insertions	121	119	119	98,35%	100,00%
Replacements	111	111	111	100,00%	100,00%
Total	473	468	455	96,19%	97,22%

Table 2: SALEM classification results

where precision is defined as the ratio of correctly classified provisions over all SALEM answers, and recall refers to the ratio of correctly classified provisions over all provisions in the test corpus. Note that here a classification is valued as correct if the automatically assigned class and the manually assigned one are identical. The classification performance is even better if it is related to the corresponding first level ontology classes (i.e. obligations and modifications). In fact, in some cases, mostly penalties and permissions, multiple answers are given due to the fact that obligations bottom classes share a great deal of lexical and morphosyntactic properties; yet, these answers are to be considered correct if classification is evaluated wrt first level classes. On the other hand, when unambiguous linguistic patterns are used, the system easily reaches 100% Precision and Recall, as with the class of Modifications.
Table 3 below illustrates the performance of the system as a semantic annotator. We distinguish three possible cases: a) the system correctly identifies all relevant semantic roles instantiated in the provision text ("Success"); b) the system identifies only a subset of the relevant semantic roles ("Partial Success"); c) the system utterly fails.

Provision Class	Success	Partial success	Failure
Prohibitions	73,33%	26,67%	-
Permissions	66,67%	20,00%	13,33%
Obligations	88,89%	11,11%	-
Penalties	47,93%	45,45,%	6,61%
Repeals	95,71%	2,86%	1,43%
Insertions	97,48%	1,68%	0,84%
Replacements	96,40%	3,60%	-
Total	82,09%	15,35%	2,56%

Table 3: SALEM performance in the semantic mark-up

7. Future Work

Although they are quite stable as a language genre, laws can also be stylistically variable depending on the personal inclinations of the author, the particular domain they apply to, not to mention variations determined by historical changes. The system needs to be tested on a larger sample of laws, witnessing a wider variety of personal and temporal parlances. Whether the addition of semantic information derived from a domain-specific lexicon can improve SALEM performance will be the object of further investigation in the near future. We also intend to test the performance of SALEM augmented with hybrid architectures making use of both probabilistic and categorical constraints on dependency parsing.

References

Bartolini R., Lenci A., Montemagni S, Pirrelli V. (2002) "The Lexicon-Grammar Balance in Robust Parsing of Italian", in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Gran Canaria.

Biagioli C., Francesconi E., Spinosa P., Taddei M. (2003) **"**The NIR Project. Standards and tools for legislative drafting and legal document Web publication", in *Proceedings of the International Conference of Artificial Intelligence and Law*, Edinburgh, June 24, 2003.

Bolioli, A., Dini, L., Mercatali, P. and F. Romano (2002) "For the automated mark-up of Italian legislative texts in XML", in *Proceedings of JURIX 2002*, London, 16-17 December 2002.

De Busser, R., Angheluta, R. & Moens, M.-F. (2002) "Semantic Case Role Detection for Information Extraction", in *Proc. of COLING 2002 - Proceedings of the Main Conference*. New Brunswick, pp. 1198-1202.

Graubitz H., Winkler, K., Spiliopoulou, M. (2001) "Semantic Tagging of Domain-Specific Text Documents with DIAsDEM", in *Proc. of the 1st International Workshop on Databases, Documents, and Information Fusion (DBFusion 2001)*, Gommern, Germany, May 2001, pp. 61-72.

Saias, J., Quaresma, P. (2003) "Using NLP techniques to create legal ontologies in a logic programming based web information retrieval system", in *Proc. of ICAIL 2003 Workshop on Legal Ontologies & Web Based Legal Information Management*. Edinburgh, June 24-28.