ICT Tools for the Discovery of Semantic Relations in Legal Documents

Marco Bianchi[1], Mauro Draoli[1], Giorgio Gambosi[2,3], Maria Teresa Pazienza[4], Noemi Scarpato[4] and Armando Stellato[4]

[1] Italian National Center for ICT in the Public Administrations (CNIPA)
{marco.bianchi, draoli}@cnipa.it
[2] Dept. of Mathematics , Univ. of Rome "Tor Vergata"
gambosi@mat.uniroma2.it
[3] NESTOR - Laboratory, Univ. of Rome "Tor Vergata"
[4] Dept. of Computer Science, Systems and Production, Univ. of Rome "Tor Vergata"
{pazienza, scarpato, stellato}@info.uniroma2.it

Abstract. This paper reports the experience of the development and the evaluation of a set of pre-competitive tools to support legal professionals in exploring a complex corpus of norms and documents in the legal domain. The research addresses two complementary goals: using ICT to support the simplification of the corpus of norms and using ICT to facilitate search and retrieval of information in large archives in the legal domain. The contribution of this work is in the development of tools beyond state of the art for the e-discovery of relationships between sections of norms or other legal documents. To reach the best results in terms of effectiveness, the tools combine statistical and the semantic approaches. The effectiveness of the statistical tool has been measured in terms of precision, through an assessment procedure that involved some experts of the legal domain.

Keywords: E-Government, Information Retrieval, Semantic Annotation, Collaborative Tagging.

1 Introduction

The field of law involves a large number of professionals and represents a multi-billion business. The workers of the legal domain have to afford a challenging amount of text, information and knowledge represented in thousands of documents. Especially in continental Civil Law countries, like Italy, the system of the laws is particularly complex. The Italian system of laws, by example, is composed by more than one hundred thousand of different norms. Thousands of civil servants, judges, public prosecutors and lawyers, highly specialised, need to manage this huge corpus of laws, archives of court verdicts and rulings or just written opinions.

A recent user survey conducted in the framework of the European funded *Judicial Management by Digital Libraries Semantics* (JUMAS) project [1] offers

a quite clear view of the need for effective retrieval of textual and multimedia documentation in the criminal courts. More than 40% of the users evaluated the effectiveness of current tools for retrieving documentation or searching of relevant information as "poor" or critical.

Information technology is expected to be crucial in simplifying the task of the workers of the legal domain. The research is pursuing two complementary goals: using ICT to support the simplification of the corpus of norms and using ICT to facilitate search and retrieval of information in large heterogeneous archives in the legal domain.

In this paper we present the experience of the development and the evaluation of a set of pre-competitive tools to support legal professionals for e-discovery in a complex corpus of norms. Main functionalities regard the automatic detection of correlations between paragraphs of different norms and the discovery of semantic "paths" that cross related norms.

The research has been coordinated by the ICT Laboratory of the CNIPA in the context of a partnership framework involving administrations, universities and research centres. The need of a system to support legal experts in designing, modifying, integrating, simplifying laws has been faced putting together a team of computer engineers, researchers, administrative people coming from Public Administration and Governmental Research. From the ICT point of view, the main possible theoretical approaches to afford the problem are the following:

- the *information retrieval (IR) approach*, mainly based on the possibility of using statistics in order to compare strings and portions of texts;
- the *semantic approach*, mainly based on the possibility of defining a set of cross related concepts and to annotate the corpus of laws with such concepts.

The first relevant contribution from ICT experts has been the choice of adopting the statistical approach, since it promised to have a first prototype, referred to as *NavigaNorme*, in a shorter time. This revealed to be very effective, since it quickly allowed to get the feedback of the users. Moreover, such agile technical approach facilitates the evolution of the tool according to opinion of the users.

The second contribution from ICT experts has been to start-up an activity for the development of a tool, *STIA*, that allows users to add semantic annotations in the normative documents. These annotations explicit semantic relations between fragments of normative texts and can be used both to build a valuable knowledge base and to improve the effectiveness of the NavigaNorme.

The participation of CNIPA to the JUMAS project is further enriching the research, bringing the contribution of the industry and other consolidated requirements from international legal communities.

The working team comprehends ICT professionals, researchers, graduate students; a very important role has been played by the potential users of the system. The team worked together with the users in order to better understand the way they think, organise the information, develop their work, exploit traditional and already available ICT tools.

The users themselves have been instructed in order to systematically evaluate the effectiveness of the tool, through an assessment procedure. To the best of our

knowledge this assessment activity also represents the first attempt in the Italian Public Administration context to evaluate the performance of an Information retrieval system in the legal domain. Furthermore, the outcome of the assessment procedure also set up the core of an evaluation benchmark to be adopted for comparing different retrieval software solutions.

This paper is organised as follows: Section 2 describes the NavigaNorme platform, Section 3 introduces the STIA annotation tools, Section 4 reports on assessment results. Concluding remarks are reported in Section 5.

2 NavigaNorme: IR techniques in the Legal Domain

NavigaNorme [2, 3] is a platform allowing specialists of the legal domain to identify correlations between paragraphs of normative texts[1]. More precisely, starting from a paragraph selected by the user (input paragraph), NavigaNorme returns a list of related paragraphs sorted by score: the greater the score value the stronger should be the correlation between the input and the related paragraph. Potential users pointed out that the capability of choosing the level of granularity (paragraphs, section or the whole norm) is a relevant point of strength of the tool.

Given an input paragraph, NavigaNorme assigns scores to related paragraphs on the basis of the following strategies:

- *text similarity strategy* - the score is evaluated on the basis of the text similarity between the input paragraph and the others paragraphs of the corpus;
- *in-references strategy* - the score is evaluated on the basis of the presence, in the input paragraph, of references to other paragraphs of the corpus;
- *out-references strategy* - the score is evaluated on the basis of the presence, in the related paragraphs, of references to the input paragraph.

These strategies can be easily set by users on the basis of their information needs. In fact users can enable, disable, or tune a strategy just by setting the associated weight. Strategies can also be used in combination.

2.1 Architecture

NavigaNorme is developed in Java and is released as a 3-tiers Web application. The presentation layer is implemented using dynamic web pages (JSP), whereas the persistence layer is mainly encapsulated by the Terrier API.

Terrier [4] is an open-source search engine readily deployable on large-scale collections of documents. Furthermore, Terrier implements state-of-the-art indexing and retrieval functionalities and provides a platform for the rapid development of large-scale retrieval applications.

[1] Note that a norm is hierarchically structured in sections, paragraph, and optionally in sub-paragraph.

From the Information retrieval point of view, all paragraphs are indexed as different document-unit. Furthermore, since NavigaNorme has to retrieve *correlations* between paragraphs, the full text of the input paragraph is submitted to the Terrier as input query on the system in order to retrieve a list of related paragraphs.

The network of references is stored in a graph data structure managed by means of the Java Universal Network/Graph Framework (JUNG) library [5].

3 STIA: Semantic Annotation in the Legal Domain

STIA is a tool allowing domain experts (lawyers, administrative staff, researchers, etc.) to inspect - through an ordinary Web browser - laws, sections and paragraphs from two different electronic sources (Web sites, digital repositories, etc.), to compare their content and to annotate relations of pertinence between them.

Due to its functional and architectural requirements (i.e. to offer web browsing capabilities, to provide data management and acquisition capabilities, to store data according to available Knowledge Representation standards) STIA has been designed as an extension of Semantic Turkey [6–8], a Semantic Web tool for Knowledge Management and Acquisition. Semantic Turkey (ST from now on) is an extension for the Web browser Mozilla Firefox [9] providing Ontology Development capabilities and facilitating the population of ontologies with new data by acquiring it from the Web. Through ST, users can literally select textual information from Web Pages, drag it & drop over ontology definitions to semi-automatically generate ontological data. ST offers a versatile extension mechanism combining OSGi standard [10] and Mozilla extension support thus allowing for the creation of completely new application residing on ST and on the hosting web browser.

STIA extends ST offering an easy interface for annotating qualified relationships between text fragments of different normative documents. Successively, the resulting conceptual annotations feed a complex process to retrieve specific relationships between texts (groups of sentences inside a section) of two norms for new application task.

Annotations become thus first class citizens in the domain ontology of this tool.

STIA adopts a specific ontology for handling concepts from jurisprudence and those needed for the annotation (e.g. laws, constraint relationship between different part of law, and some relevant relationship between part of laws), and provides dedicated forms for managing them.

Furthermore it is possible for users, by using the ontology editor features, to add resources representing new similarity relationships and to delete existing ones, and these changes will be dynamically accounted into STIA.

3.1 Architecture

STIA is deployed as an XPInstall (cross-platform installer) package [11] which, once installed inside Firefox, is handled by Semantic Turkey extension discovery

system, which extracts OSGi bundles and installs them in the main application. From an architectural point of view STIA is composed by two main logical components: the user interface (UI) and dedicated services.

The UI provides the presentation layer exploiting the Mozilla overlaying mechanism.

STIA services provide functionalities to store annotated relationship, to retrieve relationship related on a law, to remove previously stored annotation and manage information about structure of law to populate interface.

3.2 User Interaction

Fig. 3.2 shows the STIA UI: in the upper section of the interface the user can select two laws to browse and inspect. A list of sections is automatically filled with those from the selected law and the same is done for a paragraph list when a section is selected. Further, the user can select the kind of relationship by selecting one of the proposed relation types in the menubox labelled as "Tipologia di Relazione" (i.e. "Relation type").

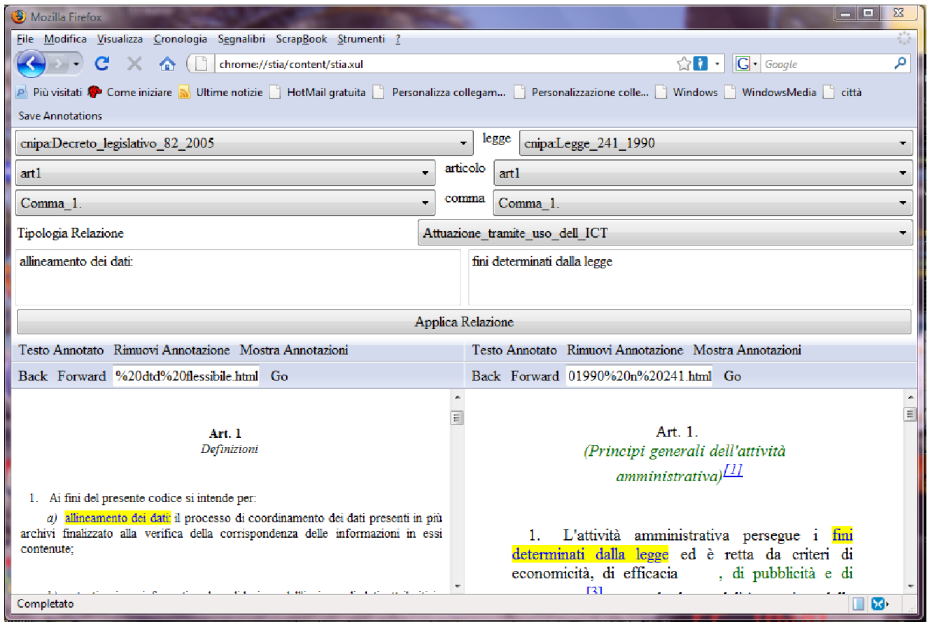

Fig. 1. STIA main interface

Finally, there are two text boxes containing the text that qualifies the relation: they are automatically filled when the user selects a section of text and click the button labelled as "Testo annotato" (i.e. "Annotated text").

With respect to the bottom part of the UI, STIA shows into different panels the two laws selected by the user. STIA also allows to navigate backward or forward in the visited pages as in any traditional browser.

As shown in Fig. 3.2, STIA makes it possible to highlight the semantic annotation that were previously taken in a document. This is a very useful feature that simplifies the work of annotators.

4 Assessment results

4.1 Assessing NavigaNorme effectiveness

NavigaNorme has been tested to estimate how helpful for the lawyer is the set of relations computed by the system, that is how precise is the answer set computed by it. Such an evaluation is usually based on a test reference collection and on an evaluation measure. The benchmark consists of:

1. a collection of normative texts: a set of 120 norms dealing with the ICT sub-domain have been selected, for a total of 8368 paragraphs;
2. a statistically significant number of paragraphs to submit as input to NavigaNorme, in order to test its effectiveness. In fact, since such a system effectiveness is known to vary widely across paragraphs, the greater the number of paragraphs used in the experiment the more confident we can be in our conclusions [12]. Since the evaluation process is also a time consuming activity, we have selected 20 paragraphs to evaluate our system effectiveness. A a team of specialists of the Legal domain, consisting of three master degree students (junior assessors), each one having a degree in law, and the head of the Office for Legislative Studies of CNIPA (senior assessor), have been involved in the assessment procedure. For each test paragraph, assessors have evaluated the first 40 results of the answer set computed by NavigaNorme.

Regarding the evaluation measure, we considered the precision (P) metric [13]. P is defined as the ratio between the number of correlations retrieved by the system that are judged valid by the majority of the assessors (relevant correlations), and the total number of correlations retrieved by NavigaNorme (retrieved correlations). P can be computed considering different levels of depth. More precisely, P@K denotes the precision computed considering the K most highly scored correlations retrieved.

The subjective evaluation of the system reveals that, simply applying the *text similarity strategy*, *NavigaNorme* has a precision of more than 55% when computed on the first 40 relations detected for each topic, and of more than 90% when derived for the first 5 relations. This performance has been judged as "very satisfying" by all the experts involved in the research. Furthermore, it has been demonstrated that NavigaNorme can improve its performance by properly combining all developed strategies (the average P@5 improves of +20%, and the average P@10 of +17%).

4.2 Assessing STIA usability

With respect to the STIA annotation tool, it is worth to note that graphical user interface has been incrementally designed and implemented on the basis of the feedback collected by the domain experts from CNIPA. The goal of this incremental development process was to improve the usability of the interface and to make the task of annotation easier and faster. At the end of the implementation phase, all the experts involved in the research considered the UI "very satisfying" since it well fits the way they think, organise the information and develop their works.

5 Conclusions

In this paper we have reported the experiences conducted during the development and the evaluation of a set of pre-competitive tools to support legal professionals in exploring a complex corpus of norms and documents in the legal domain. The set of tools is composed by NavigaNorme, an innovative software system allowing the automatic discovery of correlations in large legislative frameworks and by STIA, an novel tool for the inspection, the comparison and the annotation of normative texts.

NavigaNorme combines classical information retrieval techniques with some ad-hoc strategies able to improve the precision of the statistical retrieval exploiting implicit information extracted from the logical structure of legal and normative texts. The subjective evaluation of the system, conducted with four domain experts, reveals that NavigaNorme has a precision of more than 55% when computed on the first 40 relations detected for each topic, and of more than 90% when derived for the first 5 relations. This performance has been judged as "very satisfying" by all the experts involved in the research. It has also been verified that combining all developed strategies NavigaNorme significantly improves its performance. The NavigaNorme performances are particularly encouraging, especially in consideration of the fact that we are considering relations among specific paragraphs of the norms, and not between whole norms. This has been considered by the experts as a "killer" feature, since it allows to immediately find the specific sections of interest in a law.

STIA simplifies and speeds up the usage of large legal document collections mainly enabling the annotations explicit semantic relations between fragments of normative texts and the setting-up, consequently, a valuable knowledge base. All the experts involved in the evaluation of the STIA usability considered its user interface "very satisfying" since it well fits the way they think, organise the information and develop their works.

The implementation of both the systems have been reasonably quick (whilst the subjective evaluation required a significant human effort) and the participation of domain experts to the project greatly contributed to design the features of the tool with the purpose of meeting the real expectations of final users.

Finally, let us remind that NavigaNorme is a sort of modular framework that allows the experimentation and the evaluation of new search strategies

with as less effort as possible. Currently, we are implementing both an ad hoc strategy based on the knowledge of relevant dictionaries of terms, and a more sophisticated strategy based on the semantic annotation of the relations between paragraphs. Because of this semantic annotations produced by STIA can be used also to improve the effectiveness of NavigaNorme.

Acknowledgements

This paper has been supported by CNIPA and by the European Union in the framework of the JUMAS project (EC DG INFSO ICT Programme Project grant n. FP7-214306).

References

1. Judicial Management by Digital Libraries Semantics (JUMAS) Project homepage. *http : //www.jumasproject.eu/*
2. Bianchi, M., Draoli, M., Gambosi, G., Stilo, G.: A support system for the analysis and the management of complex ruling documents. In: 2nd Workshop on Legal Informatics and Legal Information Technology (LIT 2009). Poznan, Poland (2009).
3. Bianchi, M., Draoli, M., Gambosi, G., Petrucci, A., Stilo, G.: ICT tools for the simplification of legislative frameworks. In: First International Conference on eGovernment & eGovernance (ICEGOV 2009), Ankara, Turkey (2009).
4. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier - A High Performance and Scalable Information Retrieval Platform. In: ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006), Seattle, Washington, USA (2006)
5. The JUNG Web site. *http : //jung.sourceforge.net*
6. Griesi, D., Pazienza, M.T., Stellato, A.: Gobbleing over the Web with Semantic Turkey. In : Semantic Web Applications and Perspectives, 3rd Italian Semantic Web Workshop (SWAP2006) (2006).
7. Griesi, D., Pazienza, M.T., Stellato, A.: Semantic Turkey - a Semantic Bookmarking tool (System Description). In : 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria (June 3-7, 2007)
8. Pazienza, M.T., Scarpato, N., Stellato, A., Turbati, A.: Din din! The (Semantic) Turkey is served! In : SWAP 2008, Roma (2008).
9. Firefox Project homepage. *http : //www.mozilla.com/en − US/firefox/*
10. OSGi: OSGi Bundle Repository Specification. In: OSGi RFC0112. (Accessed 2005). *http : //www2.osgi.org/Download/File?url = /download/rfc − 0112_BundleRepository.pdf*
11. XPInstall Project homepage. *http : //www.mozilla.org/projects/xpinstall/*
12. Voorhees E.M.: The Philosophy of Information Retrieval Evaluation. In: CLEF '01 - Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, Springer-Verlag, 355–370 (2002)
13. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)