Reference Extraction and Resolution for Legal Texts

Mercedes Martínez-González[1], Pablo de la Fuente[1], and Dámaso-Javier Vicente[2]

[1] Grupo de Investigación en Recuperación de Información y Bibliotecas Digitales (GRINBD), Universidad de Valladolid, Spain
{mercedes, pfuente}@infor.uva.es
[2] Dpto. de Derecho Mercantil, Derecho del Trabajo e Internacional Privado, Universidad de Valladolid, Spain
damaso@der.uva.es

Abstract. An application to the legal domain of information extraction is presented. Its goal is to automate the extraction of references from legal documents, their resolution, and the storage of their information in order to facilitate an automatic treatment of these information items by services offered in digital libraries. References are extracted matching the texts in the collection against sets of patterns, using grammars.

1 Introduction

Legal documents[1] are highly related by references within them. These references are precious information: a rule can not be correctly interpreted without reading some of the referenced items, or they can be used to answer queries as *What documents reference this one?, What is the most referenced document in this category?*.

References in legal texts always refer to intellectual entities (laws, decrees, conventions, ...). One of these entities can be addressed in several different manners, all of them correct and without ambiguity. A document can be referenced, e.g., by its year and number or by its title. The *internal structure* of legal texts is used by legal texts authors to precise in references "where" inside a document can be found the rules governing a given subject-matter [5].

We present a information extraction problem. The information extracted are references, which follow regular patterns [9]. References are used to extend services in digital libraries: querying, hypertext services.

2 The Reference Extraction and Resolution Software

The extraction of references and their resolution consists in an analysis of document content. The document is checked from start to end, searching strings

[1] The term "document" here designates the intellectual entity users or authors have in mind. It may or not correspond unidirectionally to any of the physical objects stored in a digital system databases.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 218–221, 2005.
© Springer-Verlag Berlin Heidelberg 2005

corresponding to references, either to the analyzed document (internal references) or to other documents (external references). The extraction is a sequence of three subprocesses: the **detection** inside document content of the references (strings), the **resolution** of those references to some item (document, document fragment) between the collection of available documents, and, finally, the **storage** of the information associated to the reference in some database.

These tasks are the responsibility of software tools shown in figure 1. The *document analyzer* gets a document to extract references. It collaborates with a *pattern recognizer*, which extracts patterns from pieces of text it receives from the document analyzer. This component has information about the vocabulary used to name legal items, and knows the grammar associated to each type of reference (references are classified according to a taxonomy). The result of the analysis is a set of data: for each reference, the string found, and some data containing information about it (source document, fragment that contains the reference, target document, referred fragment).

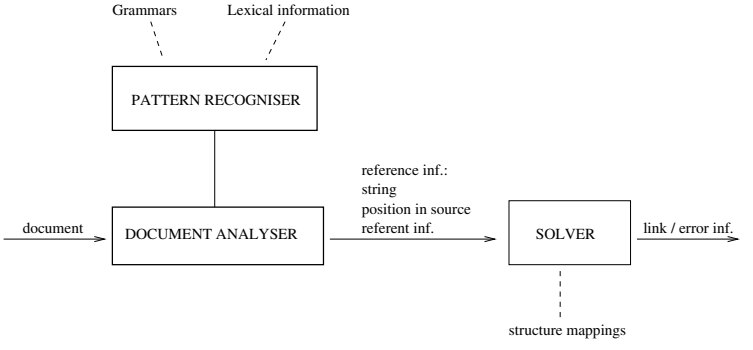

Fig. 1. Software for a reference extraction prototype

These data are processed by the *solver*, which tries to match each reference to some legal item. If it is not able to do it, the solver generates an error message which is associated to the reference, so as the other services in the digital library know that the reference has not been correctly solved. Otherwise, it generates a link that is stored in a link database.

3 Methodology

The extraction of references, their resolution, and the validation of both processes is organized in the following steps:

1. A human examination of 50 documents is made to extract references (strings). In parallel we studied some manuals about legal texts creation [5], that provide some rules about citing legal texts.
2. *Patterns* and vocabularies are characterized and a set of grammar rules is obtained.

Table 1. Results of the experiments

Collection	References detected	Relevant references	Recall
Collection 1	312	423	74%
Collection 2	268	496	54%

3. The extraction is *implemented* (using the grammars obtained in the previous step) and tested on two collections of legal texts. A set of data about references is obtained. The first collection ('*Collection 1*') is the set of 50 documents used in the previous steps to obtain the patterns. The second collection ('*Collection 2*') is a set of 14 documents provided by a legal text publisher –*LexNova S.A.*–, in which references are already tagged. This phase includes the three extraction steps described in section 2.
   As the aim was also to evaluate the process, we extracted other data useful for this goal:
   (a) >From the *Collection 1* we select a subset of six documents, on which we carefully revise the set of references extracted manually (this manual work is the reason to limit the number of documents used for the evaluation). That is, we have two types of output data: the reference data set, which is the normal output of the extraction process, and the set of references found manually, generated for validation purposes.
   (b) For the 14 documents provided by the legal publisher we mine the marks that this publisher had included in documents to point to references. This is indeed another extraction, in which we analyze every document to find the marks (and associated references).
4. The last step is the *validation*, which also differs for the two collections:
   (a) For the references extracted from *Collection 1*, the strings set obtained during the extraction is compared to the set of references obtained by human examination.
   (b) With the references of the publisher's collection what we do is to compare the data generated by the application with the data mined from the legal publisher's marks.

Table 1 shows the results for both collections. It shows the number of references detected by the automatic extractor, the number of references present in the collection and the percentage of success (number of relevant references detected over the number of relevant references present in the collection), name *recall*. The results are better with the collection used to obtain the grammars (*Collection 1*), which is logical. Some variations in the patterns of references that appear in the legal text publisher's collection are not present in the first collection. For example, a reference as '*art. 20 y la DA sexta*' (*article 20 and the Sixth Additional Disposition*) contains abbreviations ('*art.*', '*DA*') which are specific to this collection, and in consequence, they are not discovered by the recognizer.

4 Related Work

Some experiences in the area of automatic hypertext creation using information extraction techniques are applied to legal documents [3,1,2]. Pattern recognition is done with grammars in some of them; grammars are appropriate because patterns in legal documents have a high level of regularity. Outside the legal domain, references (mainly scientific references) have also been the target of several efforts [7,6,4,8].

5 Conclusions and Future Work

An application of pattern matching for reference extraction was presented. The work includes the extraction of references, but also their resolution, and validation, whose results are encouraging. There is a wide range of open possibilities for future work. Currently, we are working on extending the tests and evaluation to more types of legal documents, including European Union legal texts.

References

1. Agosti, M., Crestani, F., and Melucci, M. On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing and Management 33*, 2 (1997), 133–44.
2. Bolioli, A., Dini, L., Mercatali, P., and Romano, F. For the automated mark-up of italian legislative texts in XML. In *Legal Knowledge and Information Systems. JURIX 2002: 15th Annual Conference.* (London, UK, Dec. 2002), T. Bench Capon, D. A., and W. R., Eds., IOS Press, Amsterdam, Netherlands, pp. 21–30.
3. Choquette, M., Poulin, D., and Bratley, P. Compiling legal hypertexts. In *Database and Expert Systems Applications, 6th International Conference, DEXA'95* (Sept. 1995), N. Revell and A. M. Tjoa, Eds., vol. 978 of *Lecture Notes in Computer Science*, Springer, pp. 449–58.
4. Ding, Y., Chowdhury, G., and Foo, S. Template mining for the extraction of citation from digital documents. In *Proceedings of the Second Asian Digital Library Conference* (Nov. 1999), pp. 47–62.
5. Grupo de Estudios de Técnica Legislativa. *Curso de técnica legislativa GRETEL.* Serie de Técnica Legislativa I. Centro de Estudios Constitucionales, Madrid, 1989.
6. Lawrence, S., Giles, C. L., and Bollacker, K. Digital libraries and autonomous citation indexing. *IEEE Computer 32*, 6 (1999), 67–71.
7. Lawson, M., Kemp, N., Lynch, M., and Chowdhury, G. Automatic extraction of citations from the text of english language patents: An example of template mining. *Journal of Information Science 22*, 6 (1996), 423–36.
8. Moens, M.-F., Angheluta, R., and Dumortier, J. Generic technologies fonr single- and multi-document summarization. *Information processing & Management 41* (2005), 569–86.
9. Wilson, E. Links and structures in hypertext databases for law. In *European Conference on Hypertext, ECHT'90* (Paris (France), 1990), A. Rizk, N. A. Streitz, and J. André, Eds., The Cambridge Series on Electronic Publishing, Cambridge University Press, pp. 194–211.