Arabic Legal Information System for Case-Law Indexing and Retrieval

AL ACHKAR Mona[a] , RAMMAL Mahmoud[b]
Legal Informatics Center, Lebanese University, Beirut , Lebanon
[a] maj_aj@hotmail.com ; [b]rammal.mahmoud@gmail.com

Abstract

This paper aims to illustrate the building and the implementation of a system to enhance access, retrieval, and indexation of Arabic legal language. The core of the system relies on accurate representation of the content using Arabic legal patterns, designed as linguistic tool, to assist indexation and as an improving tool of access and of retrieve of legal information in "case law documents".

The construction of patterns is based on the use of semantic chaining at the syntactic level, as well as, at the contextual one, at the concepts' networking. The work is the initial phase of a project to benefit from what we believe is effectiveness of semantic chaining and human analysis in reconstructing documents' content and representing legal knowledge. At trials the system revealed that Patterning Arabic legal language, by mapping out the "analysis' structures" (Patterns), has greatly helped homogenize indexation's vocabulary, guide Analysts, reconstruct more accurately the content of the document and improve quality of access and retrieval system.

Keywords: *Processing natural language, Arabic legal language, linguistic tool, semantic chaining, analysis' structures, information system.*

1. Introduction

To transmit the message law uses an extraordinarily definite and precise language. It employs words in strict accordance with definitions and by employing technical legal terms that meaning has been brought out and fixed, by long experience and use [12]. In this context, legal professional can have easier access to online legal information;

Nevertheless, and despite the potential of Information and communications' technologies and the growth of the Internet as the most important feature of the information society, users do still need to know how to avoid any confusion when using terms and words, especially that substituting one word for another can easily result in serious errors and misunderstandings. They need to be precise in order to access and retrieve relevant information. But precision here may also mean knowing the indexation language. Actually, processors and users are expected to use words carefully and precisely. They are expected to catch the legally significant differences between terms and their use, such in the case of: "*Living in*" and "*residing in*" "*domiciling*" and "*being a citizen*".

Effectively, the main challenge is still to find tools that can efficiently help users access to legal relevant information in huge corpus. In this context, this paper describes the work done in creating a system to insure better access and retrieval of Arabic legal information of case law on legal database through legal language patterning and concepts networking.

A tedious work of research and retrieval, on the existing jurisprudential database, has been conducted to dress an inventory of the actual situation of jurisprudential databases, at the legal informatics center in the Lebanese University, so we can define the challenges the system has to deal with.

We choose case law texts issued by courts in Lebanon [5] (cassation and appeal) as groundwork to create the integrated legal information system, because case law language represents the characteristics of natural language used in a scientific domain.

2. Legal Language: special vocabulary

When processing legal language, we shall consider the fact that legal language comes from natural language and adds to it specific words and particular meanings corresponding to the legal nature of that discourse [11]. As a matter of fact, some recognizable words in the general language take on different or new meanings when used in the law. In the penal code, for example, "*Malice*" used in the defamation case, does not mean *hatred or meanness* but rather "*with reckless disregard for the truth*". Consequently, one can notice that, the ordinary understanding of words is loosely shaken in order to absorb the legal meaning.

Moreover, the difference between natural language and legal language is semantic not syntactic. It depends on the words as well as on their specific meanings. Law's vocabulary is a little bit complicated due to many factors among which we can notice that it's made of:

- New words that will not probably be encountered in other disciplines or fields of sciences and knowledge. Hence, it's important to learn it, in order to understand, not only the legislative and regulatory texts, but also, the case law as well. These are the "legal terms". *Impleader*, *executory interest, demurrer* and *mens rea* are few examples.
- Some words describing complex concepts. These are words that have been interpreted by judges or doctrine and represent or adhere to large bodies of law. The terms "*unfair competition*," "*reasonable care*," "*absolute necessity*" and "*act of God*", are some examples.
- Some words said "terms of art" because of their distinct or specialized meanings. They represent the standard formulas of expression that the meaning has been sanctified through long use .
- Some words with expanding, contracting or changing meaning, depending on the context in which it is used. In one context (vote, for example), a person may be considered "legally capable in Lebanese law" if he is 21 years old. In another context (getting a driver's license in Lebanese law) a person may be considered "legally capable" when he gets 18.
- Some technical words may seem interchangeable while they are particularly technical and shall be distinguished such as: the difference between *residence* and *domicile*, *privilege* and *right*. Nevertheless, legal professional got to make the difference between them, giving the critical legal consequences they may have.

3. Case Law Language Particularity

Legal documents and texts represent informational tools aiming to communicate a message. Nevertheless, case-law texts have their own challenges stemming from:

- The personal linguistic input of the judge.
- The cohabitation of the formal and informal expressions and terminologies.
- The structure of the document itself intended to reveal the path to the result built on a methodology of reasoning.
- The rich vocabulary compared to other legal texts.

But, and despite this, case law doesn't contain the needed concepts and words that facilitate their retrieval on databases or on-line. This is mainly due to the characteristics of legal language on one hand and to the implicit concepts inherent to the legal language, and intended to be conveyed by the sentences and the document structures, on the other. Adding to this, that legal terms are highly sensitive to changes in context, we can realize that Legal discourse in case law is a highly specialized use of language requiring a special set of habits.

Obviously, processing case law requires painstaking attention to detail and to the consequences of contextual changes. For this reason, automatic processing of legal language in legal documents in general, and in case law in particular, is of limited usefulness in describing content, since, a good processing requires both a thorough understanding of the subject and a familiarity with the legal system and its components, which is not possible for computers yet.

So, when looking for answer, the most helpful element revealed to be considering the particularities of legal language in one hand, and relying on human analysis of each document based on predefined linguistic tools and language patterns on the other hand. This means achieving normalization and homogenization of vocabulary as well as precising the analysis degree and enhancing the human performance.

4. Aims and Characteristics of the System

In this context the system's aim is twofold: first to build an efficient communication oriented towards jurist and non-jurist and second, to help enhance information retrieval by boosting both recall and precision at the same time. Hence, the followings shall be the characteristics of the retrieval functions:

- An Arabic friendly-user interface.
- Language oriented indexing and query processing techniques.
- Enhanced sentences processing based on linguistic tools and patterns.
- Acceptance of semi-natural language sentences as query.

5. Methodology

Considering our aims and the characteristics we are looking for in the system, we adopted a linguistic approach involving syntactical and lexical analysis. Thus, our methodology is built in a way to:

- Allow for accurate representation of the document content
- Ensure better management of the special nature of legal language

- Provide a reliable solution for natural language problems stemming from the fact that it's not a language designed to be understood by machines.

Aiming to achieve access to legal information, through documents and texts classifications, semantic conceptualization, and documents re-structuring, our techniques heavily depend on elaborating tools that answers the reliability of this methodology, such as lexicon, language patterns, and analysis structures.

The methodology consists of building bridges between AI techniques and IR system to achieve unambiguous representations on which matching and retrieval can take place. Consequently, our work relies heavily on ontology built by concepts and analysis 'structures, and defined in [8] as an "explicit specification of a conceptualization: the objects concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them" and that make the context of words and terms.

6. Patterning Legal language

Giving the fact that the importance of a legal term could only be understood in its context and that classification mode is inherent to the juridical system, patterns are created to represent the discourse level and integrate all the documents. Patterns are as numerous as needed. They consist of classificatory concepts hierarchically represented. Key high-level legal concepts are chosen according to the traditional academic divisions of law families and sectors. They determine text context. Moreover they allow for top-down flow of information, so that constructed context of the words or concepts could influence their semantic meaning.

Structure of these patterns allows meaning to become hard-wired to even more pertinent contexts, and make search technology ever more useful.

Most relevant documents are not necessarily those with the highest incidence of the search terms, but rather those in which the concepts represented by the search terms are the focus of the document, and of its representative's parts.

Most of the natural language processing steps performed were done within the framework of the words and concepts' disambiguation through the reconstruction of their relation to other words and through their semantic relations [13].

7. Defining the Context

In their quest to retrieve legal texts and information jurists search by legal concepts not random keywords. Concepts are distinguished from the words that refer to them. Different words could refer to one concept, and more than one concept could be referred to by one word. So, using concept in a structured knowledge database achieves better disambiguation of word and concepts as well. It helps jurists better express their information needs and get quickly to information they seek.

Defining concepts and patterning language represent the most interesting part of our work, because they provide the possibility of handling language and enhance accuracy of the indexation and of the retrieval results.

Concepts are manually extracted from analyzed documents and organized according to the order they belong to and to their relation to each other. To achieve a better definition of concepts, language is analyzed and structured into patterns. A complete list of all the possible meanings of a concept is established for the lexicon of concepts.

8. Analysis' Structures

"Analysis' structures", are built out of case law texts analysis done by legal analysts. They represent the legal issues before the Lebanese courts in a given domain and organized according to the academic divisions of the law: civil, penal, administrative, commercial and so on.

They are sets of concepts hierarchically organized from the general to the specific and from legal concepts to factual ones. Figure (1) shows a sample of an analysis structure in Maritime Law.

المادة	رقم الهيكلية	الهيكلية
ضمان	2	2
	3	ضمان بحري، فسخ عقد الضمان (نعم،لا)، مادة 302 قانون تجارة بحرية، - (افلاس المضمون/ افلاس الضامن) - (توقف المضمون عن الدفع/ توقف الضامن عن الدفع) - عدم دفع قسط ضمان مستحق، تقديم كفالة (نعم،لا)، - احتفاظ الضامن بالاقساط المدفوعة (نعم،لا)، - اصول الانذار والتبليغ - ضمان بضائع، عدم الابلاغ عن الشحنات،مادة 302 قانون تجارة بحرية، استحقاق اقساط الضمان (نعم،لا)، - فسخ اختياري/ مادة 300 قانون تجارة بحرية، عدم بدء المخاطرة (نعم،لا)،
	4	ضمان بحري، بطلان عقد الضمان (نعم،لا)، * (كتم معلومات/ تصريح كاذب/ اختلاف بين عقد الضمان واوراق النقل) من شأنه التقليل من فكرة الخطر (نعم،لا)، مادة 279 قانون تجارة بحرية، وجود علاقة سببية بين (كتم المعلومات/ التصريح الكاذب/ الاختلاف بين عقد الضمان واوراق الدعوى) والضرر (نعم،لا)، * زوال الخطر، مادة 318 قانون تجارة بحرية، انشاء العقد بعد (هلاك البضاعة/ وصول السفينة)، * ضمان على الانباء السارة او السيئة، (علم المضمون/ علم الضامن) قبل توقيع العقد (بهلاك السفينة/ بوصول السفينة)، * عدم ابلاغ الضامن الحوادث اللاحقة للعقد (نعم،لا)، مادة 298 قانون تجارة بحرية، * تراكم عقود الضمان، مادة 324 قانون تجارة بحرية، - (حسن نية المضمون/ مسؤولية المضمون) (نعم،لا)، - صحة العقود (نعم،لا)، * استحقاق قسط الضمان (نعم،لا)، * استحقاق تعويض الضمان (نعم،لا)،

Figure (1): Analysis structure used to reconstruct the secondary document

These sets are connected together by semantic relation.

They heavily rely on lexical chain technique that can be used to identify the central theme of a document or the points of law in a case law document. It is a sequence of related concepts in the text.

To construct lexical chain we have to identify relationships between words; this was made possible to us by the use of our legal databank lexicon.

By using the analysis' structures we were able to build more sophisticated indexing techniques than the simple keyword-based ones.

The Analysis' structures represent:

- Rigorous methods that take into account mastering language and its organization.
- An agent that helps reducing judicial vocabulary heterogeneity, especially in the discourse describing the factual concepts or events.

- Factor that allow indexers to restore implicit concepts.
- Tool of harmonizing processes of the reconstruction of the text among all the indexers.
- Helpful element to reduce semantic ambiguities and to homogenize classifying concepts and to define the analysis' degree.

When analyzing, indexers have to reinstate the concepts describing the point of law they extract from the document in the suitable set of concepts.

9. Representation of the Case Law Document

Actually, Jurisprudential documents like any given document will typically have a central theme, one or more focus points. These are the solutions or points of law discussed by the judge.

That is why, the matter of how we represent, what a document is about, is a key factor in the design of a documentary retrieval system and that's why we do consider our approach to be efficient and reliable, since we respect the particularities of the legal language in first place and the document's structure on the second.

In reconstructing jurisprudential documents our focus goes toward reasons adduced for a judgment which are divided into paragraphs. The case law is reconstructed in a secondary document according to the given solutions for the disputes. Legal issues are represented by hierarchical concepts arranged to convey the exact reasoning of the judge and to ensure complete representation of the entire information present in the judgment.

Legal concepts are complex and claim corresponding complexity in sentence structure. Sometimes, a great many qualifying phrases and dependent clauses are required in order to express a concept with the necessary precision.

Accordingly, the representation of the document is done with concepts and objects obtained from the text itself, as well as from the references it uses, such as codes and other case law, through linguistic techniques to help improving indexation of judge's discourse [6], as well as reconstructing the equivalent of the case law original document: the abstract, where linear text is transformed into structured one.

The secondary searchable document is composed of: Abstract, paragraphs and sentences.

- The abstract: the abstract is the secondary document established while processing the original document. It reflects the exact content's structure of the original document and can be described as the semantic representation of the case law. It has to be drawn up by specialized staff and consists of a set of paragraphs describing legal and factual questions discussed in the case law itself. It represents the whole document and hence the context in which the different questions are discussed and related in order to arrive at the given decision.
- The paragraphs are thematically coherent units. They are constructed by referring to Analysis' structures, they represent a sequence of hierarchical concepts semantically related to convey the meaning of the document by illustrating the argumentation and the solutions of legal issues discussed in the case law. Each paragraph represents one legal question. It starts by a key legal concept said general concept representing the legal family of the issue such as: " *commercial contracts*" which includes " *maritime*

transportation" *marine sales*" "*vessel chartering*" or the legal question that can include many related minor issues like "*general average*" that includes "*carrier liability*", " *conditions of contribution*", " *validity of the act*" . The following concepts are a mix of legal and factual concepts describing the issue and the solution and that are related semantically to convey the exact meaning of the solution given by the courts. Each concept is represented by a sentence and sometime, by one word, when the word alone describes a precise and well defined concept such as: expropriation. Figure (2) shows an abstract of case law composed of 3 paragraphs.

- The sentence: is a semantic chain intended to precisely define the meaning of single words that might be ambiguous. Sentences provide a context for the resolution of an ambiguous term and enable identification of the concept the term represents. Each given set of sentences determines the structure of a paragraph. The sentence is designed to convey precise legal concepts such as:" *Maritime carrier liability*" or a factual concept such as: "*falling apart*". Sentences are selected from a fixed list related to the convenient analysis' structure already mapped to include the legal issue or principle. More sentences are added to the analysis' structure in order to create a link between the apparent concepts and reveal the implicit ones.

رقم الحكم : 1974/983

تاريخ الحكم : 27/10/1974

المحكمة : محكمة الاستئناف المدنية-بيروت

مستخلص :

ضمان بحري، بطلان عقد الضمان (لا)، كتم معلومات من شأنه التقليل من فكرة الخطر (كلا)، مادة 297 قانون تجارة بحرية، جهل المضمون بالمعلومات، ابحار السفينة قبل توقيع العقد.

ضمان بحري، دعوى الخسارة البحرية، استحقاق تعويض الضمان (نعم)، اخطار مضمونة، ضرر اثناء النقل البحري، تبلّل البضاعة، سقوط امطار وتلوج، عقد خاضع لشروط وثيقة الضمان، وثيقة تضمن التبلل بمياه الشتاء والمياه الحلوة ومياه البحر، ضرر مضمون (نعم).

ضمان بحري، موجبات الضامن، تعويض الضمان، تسديد تعويض الضمان (نعم)، تأخر في تسديد تعويض الضمان، فائدة المبلغ لصالح المضمون، حق الضامن في الرجوع على الناقل، حكم على الناقل لصالح المضمون، حلول الضامن محل المضمون، حساب التعويض، ضرر فعلي، ربح فائت.

Figure (2): An abstract of a case law

10. The Retrieval System

It's well known, that there is an intimate link between language and the law, which makes the legal language properties of a major impact not only, on processing but also on retrieving methodologies adopted in legal databases. Thus, as far as the object of retrieving information is to find information by corresponding terms used for the indexation with those used by the user to formulate his query, processing natural language and retrieving information are not, but two sides of the same currency.

And if we admit that high precision information retrieval is an extremely hard problem, we also believe that we had already achieved a breakthrough. The most helpful element as mentioned was that we do completely consider the particularities of legal language in one hand, and that we rely on human analysis to reconstruct and rebuild each document based on predefined linguistic tools and language patterns on the other. This means we achieved normalizing and homogenizing of vocabulary as well as precising the analyze degree and enhancing the human performance.

The retrieval system allows the use of Boolean operators in addition to retrieval levels, made possible through many levels, the whole abstract, the paragraph, and the sentence. The most accurate answer for a search on a concept is given at the sentence level. Combination between two or more concepts, help to accurate retrieve of a specific legal issue or case, while search at the abstract level may yield less precise results. At the same time, Lexicon and thesaurus are used to deal with some linguistic difficulties such as polysemy, synonymy and morphology. Moreover, they are used in enlarging and refining the scope of query at the retrieval stage.

10. Search features

Yielding of relevant retrieving results while using the system largely depend on the complete and accurate representation of the case law content and structure, as well as on the NLP techniques used to provide matching between two different vocabularies: the one used in processing and the one used in searching and retrieving.

When searching in legal databases, users must translate their information needs into legal concepts. The system allows search on the secondary document built by indexers and based on the original document's structure. This methodology makes it more effective because it helps differentiate the role or function each concept plays in the text.

The document's structure is composed of an abstract, made of sentences, ordered into paragraphs to represent each legal issue along with the solution given by the court. It represents the search units where the information is well structured to help determination of terms and concepts 'semantic as well as its accurate location. Moreover it helps the user to rapidly decide the relevance of the case to its search criteria.

10.1 Searching at the sentence level

Representation of legal knowledge shows that each document contains many concepts, while each cluster of terms represents concepts. It shows also that some words are constituent parts of larger terms or may have a role relationship with each other. So to avoid ambiguity that single words may represent we designed the sentence as a search environment for the word following assumptions that:

a. The sentence is the most adequate frame to disambiguate a word and to give it the meaning intended in the search.

b. Word pairs that appear at the same sentence in the same order as in the query (forward pairs), are more likely to provide highly precision in retrieving information than the ones that appear in the inversed order (backward pairs).

And, in order to implement those assumptions we determined sentence as a search level for more effective matching between query terms and words in the documents.

We also used the thesaurus to expand query terms to their hyponyms and synonyms as well as to concepts.

10.2 Searching at the paragraph level

We designed the location model as an alternative system to the semantical net of relations, and based it on the following assumptions:

a. Presence of any given two concepts (represented by two different phrases) from the query in the same paragraph enhances the relevancy of the document.

b. Discourse information is very helpful in determining relevance of a document.

c. The more concepts are located at the same paragraph, the more relevant the document is.

10.3. Searching at the abstract level

Searching at the abstract level would be helpful, in case the research at the paragraph level by combining two or more concepts doesn't yield any result. And, despite the fact that the abstract (the input structured document) is well structured defined and composed of paragraphs while these latter are composed of concepts, searching by single words at this level is far from being accurate and effective. Nevertheless it is helpful to locate related legal issues in one case law.

11. Conclusion

On the grounds of building Arabic legal databases and information retrieval systems developed at the legal information center at the Lebanese university, it can be assumed that we have established an efficient approach in view of overcoming the Arabic legal barriers arising in information retrieval. The system paid particular attention as detailed to linguistics and conceptual aspects of Arabic legal language, which provides important insights for the building of an Arabic legal ontology.

Hence, we have combined techniques of natural legal language, documents 'inherent structures and legal concepts particularities. The system has two main components, a keyword lexicon that is a list of terms with semantic relations, and another lexicon of concepts used in constructing patterns. Elaborating "analysis' structures" as well as reconstructing the document is a must toward the integration of such system.

When we put the system to testing, we had the following results:

a - The patterns we mapped out revealed to be definitely reliable in the field of indexation and retrieval systems' improvement because:

- They emphasize on meaning processes and sense construction that result from the utterance sequences.
- They apply the semantic analysis of judicial argumentation to reconstruct the meaning of the case law. Actually, judicial argumentation illustration is a determining factor in reconstructing judge's reasoning which characterizes the jurisprudential documents.

- They improve the information retrieval and access by helping overcoming the ambiguities inherent in natural legal language, since they filter extraneous information and return only relevant ones.
- They introduce and provide the right basis to an Arabic legal ontology.

b – On the other hand, the document reconstruction allowed a better integration of all the needed keys to achieve a complete and accurate representation, such as, the division of paragraphs according to legal solutions and issues, the academic classifying concepts and indexation methodology. It also conveys the exact reasoning of the judge by permitting a better representation of all the parts of the case-law's document.

Legal language Patterns "analysis' structures", allow the Integration of a computer-assisted system for redaction of abstract and for indexing.

Actually, the development of the retrieval system based on using "analysis' structures" as legal language patterns has already proved to be efficient on the legal domain[9][10].

References

[1] Valente, A. (1995) "Legal knowledge engineering: a modeling Approach", IOS Press Amsterdam.

[2] Tom M. van Engers (206) "Legal Engineering: A structural approach to Improving Legal Quality", in Applications and innovations in intelligent systems, Vol. 1, 2006.

[3] Pietrosanti E, Graziadio (1999) "Advanced techniques for legal document processing and retrieval. Artificial Intelligence and law", vol 7, pages 341-361, Kluwer Academic Publishers, 1999.

[4] Marie-Francine Moens (2001) "Innovative techniques for legal text retrieval, Artificial Intelligence and law", vol 9, pages 29-57, Kluwer Academic Publishers, 2001.

[5] Rammal M, Mona Al Achkar, Philippe Nabhan (2001) "Computer Assisted research system on legal Lebanese document", in Workshop On Arabic Software, Lebanese American University, 2001

[6] Mona Al Achkar, Rammal M, Philippe Nabhan (2001) Pattering Arabic Legal Language, in Workshop On Arabic Software, Lebanese American University, 2001.

[7] Shapiro, Stewart (2006) "Vagueness in Context", Clarendon Press. Oxford.

[8] Gruber T, R, (1993) "A translation approach to portable ontologies". Knowledge Acquisition, vol.5 No. 2, pp:199-220, 1993

[9] Uyttendaele C, Moens M.-F. Dumortier J."Salomon (1996), "Abstracting of legal cases for effective access to court decisions", in Proceedings of JURIX 96 Ninth International Conference on LegalKnowledge Based Systems, Tilburg: University Press 1996.

[10] Grover C., B. Hachey, and C. Korycinski, (2003) "Summarising legal texts: Sentential tense and argumentative roles". In HLT-NAACL 2003 Workshop: Text Summarization Edmonton, Alberta, Canada, 2003,

[11] Wroblewski J, (1985), "Legal Language and Legal Interpretation" in Law and Philosophy,Vol. 4, No. 2, Legal Reasoning & Legal Interpretation, pp. 239-255, 1985

[12] Shartel, Burke (1947),"Our Legal System And How it Operates", Ann Arbor, Mich.: Overbeck Co.

[13] Navigli R, Lapata M, (2010) "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, 2010

[14] Karin C. Ryding, (2005) "A reference grammar of modern standard Arabic" Cambridge University press.

[15] Fahad A., Ibrahim A., Salah F, (2009) "Processing Large Arabic Text Corpora: Preliminary Analysis and Results", In 2nd International conference on Arabic language resources & tools, Cairo Egypt - 2009

[16] Hoeer, S., Bunzli, A, (2010) "Controlling the language of statutes and regulations for semantic processing" In: Proceedings of the LREC 2010 Workshop on Semantic Processing of Legal Texts - Valletta, Malta -2010