Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations

Guiraude Lame

Centre de Recherche en Informatique, Ecole des mines de Paris, 35 rue Saint-Honoré,
77305 Fontainebleau, France
lame@cri.ensmp.fr
http://www.cri.ensmp.fr

Abstract. A method to identify ontology components is presented in this article. The method relies on Natural Language Processing (NLP) techniques to extract concepts and relations among these concepts. This method is applied in the legal field to build an ontology dedicated to information retrieval. Legal texts on which the method is performed are carefully chosen as describing and conceptualizing the legal domain. We suggest that this method can help legal ontology designers and may be used while building ontologies dedicated to other tasks than information retrieval

1 Introduction

If the semantic Web is more than a vision but the future of the Web and if the semantic Web is to rely on ontologies, these ontologies cannot be entirely built by hand. Many methods of ontology design have been suggested (see [10], [19], [7], [12], [11]). Most of them include these different steps:

- a preliminary step to determine the reasons why an ontology is needed ;
- the precise definition of the domain of the ontology ;
- the specification of the task to which the ontology is dedicated ;
- the identification of the domain concepts and relations among them ;
- the collection of the concepts and relations in an ontology formalized in an appropriate language to become machine readable ;
- the integration of the ontology in a system.

We focus on the step consisting in identifying concepts and relations among them. We claim that this step can be improved if ontology designers use Natural Language Processing (NLP) techniques.

Ontologies are composed of concepts and relations among them, structuring an overview of entities [16]. We assume that concepts are embodied in terms and that semantic links among concepts are embedded in syntactical relations among these terms.

Legal concepts are known as being open textured concepts meaning that their definition may vary depending on many factors (context, source etc). Many ontologies of law may be defined, their components depending mainly upon the task for which these ontologies are built for [2].

V.R. Benjamins et al. (Eds.): Law and the Semantic Web, LNAI 3369, pp. 169–184, 2005.

In this article, we present a general method for identifying legal concepts and semantic relations among them using NLP techniques. All these elements are the structuring blocks of an ontology. This method is inspired by the one defined in [1]. A similar approach is taken in [8]. In our context, the ontology is an ontology of French law and is dedicated to information retrieval. Our method relies on the principle that legal concepts are often[1] defined and conceptualized by the legislator himself. We propose to use the legal norms that are the Codes in French law to infer legal concepts and semantic relations among them. We claim that such an ontology is useful in information retrieval contexts such as interactive query expansion systems or didactical access to legal texts bases.

2 The Codes: A Previously Existing Conceptualization of the Law

We assume that the legislator, while making the law, conceptualizes the legal field. The legislator himself performs another conceptualization task when he decides to rationalize the legal field by compiling norms into Codes.

In French law, two different types of Codes may be distinguished. The first ones are those initially created. These Codes are known as the Codes Napoléon: the Civil Code or the Penal Code for example. The second ones are those created more recently, resulting on thematic compilations of previously existing norms. In French law, this process is called codification [5]; many codes have been created since the beginning of the 1990's. Independently of their types, all Codes can be viewed as conceptualizations of legal fields. First of all, their structure is logically defined: one division for one theme, from the more generic to the more specific (See Table 1). Second, the concepts are one by one defined. These definitions may be more or less explicit. For example, the definition of a *record of birth for persons born abroad* (C. civ., art. 98) is explicit. We know under which conditions a birth record may be established; we also know what the elements composing such a birth record are: "*A record taking the place of a record of birth shall be drawn up for any person born abroad who acquires or recovers the French nationality unless the record drawn up at his birth was already entered on a register kept by a French authority. That record shall state the name, first names and sex of the party concerned and indicate the place and date of his birth, his parentage, his residence at the date of his acquiring French nationality*".[2] The definition of the concept of *divorce* is less explicit (C. civ., art. 227), referring to breach of marriage: "*A marriage is dissolved 1° by the death of one of the spouses; 2° by lawfully pronounced divorce*".[3]

If the task to which the ontology is dedicated relies on inferences, i.e. on reasoning, one would need to define a *record birth for persons born abroad* with its components,

[1] In the case of a legal system based on texts.

[2] « Un acte tenant lieu d'acte de naissance est dressé pour toute personne née à l'étranger qui acquiert ou recouvre la nationalité française à moins que l'acte dressé à sa naissance n'ait déjà été porté sur un registre conservé par une autorité française. Cet acte énonce les nom, prénoms et sexe de l'intéressé et indique le lieu et la date de sa naissance, sa filiation, sa résidence à la date de l'acquisition de la nationalité française » (C. civ, art. 98).

[3] « Le mariage se dissout : 1° par la mort de l'un des époux ; 2° par le divorce légalement prononcé » (C. civ., art. 227).

and the conditions of its drawing. Then, the concept of *abroad* must be defined, relying on a precise definition of the countries [2]. The concept of *time* may also have to be defined, to establish the value of the concept of *residence at the date of his acquiring French nationality*. In our context of information retrieval, we claim that the only elements we need are the concepts of *birth record for people born abroad* linked to all its components (*name*, *first names*, *sex*, *place of the birth*, *date of the birth*, *parentage*, *residence*) related to the more general concept of *birth record*. With the same logic, we claim that in our ontology, we only need to define *divorce* as *breach of marriage*.

Table 1. Some Sections and Subsections of the Civil Code

Civil Code
BOOK I OF PERSONS
TITLE ONE OF CIVIL RIGHTS
TITLE ONE bis OF FRENCH NATIONALITY
Chapter I - General Provisions
Chapter II - Of French Nationality by Birth
Section I - Of French Persons by Parentage
Section II - Of French Persons by Birth in France
Section III - Common Provisions
Chapter III - Of the Acquisition of French Nationality
Section I - Of the Modes of Acquiring French Nationality
…
TITLE TWO OF RECORDS OF CIVIL STATUS
Chapter I - General Provisions
Chapter II - Of Records of Birth
Section I - Of Declarations of Birth
Section II - Of Changes of First Names and Name
Section III - Of Record of Acknowledgement of an Illegitimate Child
Chapter III - Of Records of Marriage
BOOK II OF PROPERTY AND OF THE VARIOUS MODIFICATIONS OF OWNERSHIP
TITLE ONE OF THE VARIOUS KINDS OF PROPERTY
Chapter I - Of Immovable
Chapter II - Of Movables
Chapter III - Of Property in its Relations with Those Who own it
TITLE TWO OF OWNERSHIP
Chapter I - Of the Right of Accession to what is Produced by a Thing
Chapter II - Of the Right of Accession to What Unites or Incorporates Itself with a Thing
TITLE THREE OF USUFRUCT, OF USE AND OF HABITATION
Chapter I - Of Usufruct
…
BOOK III OF THE VARIOUS WAYS OWNERSHIP IS ACQUIRED
…

3 Legal Terms and Legal Concepts

3.1 Definitions

Concepts are labeled with terms. For example, *breach of contract* or *liability* are terms that label legal concepts.

Law, tending to regulate human activities, conceptualizes the world. As a consequence, the legal domain deals with various domains such as medicine or science. Consequently, many terms, general or specific to given domains, may be assimilated to legal terms since they label objects or artifacts apprehended by law. We assume that as law regulates things, conceptualizing them, these things turn out to become legal things and legal concepts.

We define legal terms as terms labeling specific legal concepts such as *contract* or *liability* but also labeling general or specific concepts such as *passenger*, *doctor*, or *weapon*: all world objects or artifacts apprehended by law. Legal terms are defined as terms labeling world objects apprehended by law and artifacts created by law.

3.2 Seeking Legal Terms

To identify legal terms labeling concepts, the future components of our ontology, we have performed Natural Language Processing (hereafter NLP) techniques on the French Codes. The experiment took place on the 57 Codes available on the governmental web site for French law: Légifrance[4]. All these Codes compose our corpus of experiments. We have used a syntactical analyzer of texts called Syntex [3]. This tool performs syntactical analysis on texts, identifying nouns, verbs, adjectives and adverbs and syntactical dependencies among them (subject of verb, object of verb etc…). On these bases, applying a set of syntactical rules, the tool is able to identify complex terms such as noun phrases, verb phrases, adjective phrases etc…

Table 2. Terms extracted from the Codes

Terms
budgetary
eventually
Hauts-de-Seine
decision
elaborated
designed for disabled persons
breach of contract
notified of the decision
to acquire French nationality
to state

[4] http://www.legifrance.gouv.fr

Used on our corpus of experiment, the tool has extracted more than 500 000 terms. This list gathers terms from all syntactical categories: verbs, adverbs, nouns, noun phrases etc. Table 2 gives an example of these outputs, translated in English.

Our experiment then consisted on trying to identify among this list of more than 500 000 terms those that could be qualified as legal terms (complying with the definition given above) and those that could not.

3.2.1 Statistical Indices to Seek Legal Terms

The first step of the method removes some classes of terms from the initial list. First of all, we have decided to only consider terms belonging to just one syntactical category: the nouns and noun phrases. This choice relies on the idea that most concepts are embedded in nouns. Legal concepts that are labeled as adjectives or adverbs are then not included in our ontology.

Secondly, terms with non-alphabetical characters are removed from the initial list. Most of such terms in our list are internal or external references to texts such as *article 1382*, or values of various rates. As our ontology is dedicated to information retrieval and not to reasoning, we assume that the useful term in our ontology is, for example, the term *taxation rate* and not *taxation rate of 19.6%*.

Applying these two principles on the initial list, we obtain a list of about 300 000 terms.

The second step of our method to identify legal terms uses statistical methods classically used to weigh index terms. The idea was to weigh the terms of our list and, on the basis of these weights, determine which are legal and which are not. Various statistical indices have been used to weigh our 300 000 terms.

1. Term frequency (tf):

Term frequency (tf) corresponds to the number of times a given term occurs in all the Codes. Term frequency characteristics in our corpus of experiments are as listed in Table 3.

Table 3. Term frequency characteristics

Term frequency	
Minimum	1
1st quartile	1
Median	1
Mean	16.6
3rd quartile	2
maximum	106 386

Among 300 000, 188 158 terms (63%) appear only once. Manually analyzing some of these terms, we have concluded that they could not all be assumed as non-legal terms. Figure 3 lists important legal terms that have a frequency rate of 1.

The term presenting the maximum frequency rate is *article*. This result is not a surprise, knowing that our corpus is composed of Codes, each code being divided in various numbers of articles ; every article starts with its own reference, for example *article 1382*.

Table 4. Important legal terms that have a frequency rate of 1

Terms
chargeable activities
agricultural activities
drug forwarding
potential vendee
risk completion

That way, high frequency rates (more than 50 000) can be used to identify empty terms that can not be assumed as legal terms such as *chapter*, *code*, or *provision*. Unfortunately, manual analysis allowed us to state that terms presenting high frequency rates include legal terms such as *decree* or *law*.

This manual analysis led us to the conclusion that fixing thresholds under or above which terms may be valuably assumed as legal terms is not possible. Such a result would require complex heuristics that probably could not be applied in other contexts and experiments.

We have concluded that the frequency rate is useless in trying to distinguish legal terms from non-legal ones.

2. Inverse document frequency (idf):
Idf ([17], [15]) establishes term distribution among a corpus, relying on the principle that term importance is inversely proportional to the number of documents from the corpus in which the given term occurs (See Equation 1). Documents are defined as articles of Codes. Our corpus gathers a total of 59 275 documents. Inverse document frequencies for terms in our corpus are as listed in Table 5.

$$idf_i = \log \frac{N}{n_i} \tag{1}$$

Where N = total number of documents in the corpus,
And n_i = number of documents of the corpus in which term i occurs

Table 5. Inverse document frequency characteristics

idf	
Minimum	0.6932
1st quartile	4.2767
Median	4.7049
Mean	5.0690
3rd quartile	5.0690
maximum	7.8136

Firm is the term presenting the lowest idf, *application to the Préfecture* is the one presenting the highest idf. Traditionally, terms presenting a low rate of idf are not considered interesting because occurring in most of the documents of the corpus. Inversely, terms presenting a high level of idf are supposed to be interesting.

We have manually analyzed terms and their idf weights. It appears that legal terms may have high (superior to 7.7) as well as low idf weights (inferior to 2.5). Table 6 lists some terms with high and low idf rates that are legal terms.

Table 6. Legal terms with high and low idf rates

Idf < 2.5	Idf > 7.7
moral aid	*application to the Préfecture*
judicial guarantee	*notification through bill sticking*
educational obligation	*bond subscription*
Minister	*period for candidature registration*
Firm	*quantity of voting paper*

As for frequency rates, fixing thresholds on idf weights would require long and complex heuristics that probably could not be applied in others contexts. We then conclude that idf cannot be used to distinguish legal terms from non-legal terms.

3. Tf.idf
Combining tf with idf, the idea is to distinguish terms that, although appearing in a few numbers of documents of the corpus, present at the same time a high frequency rate in the corpus [17].

The same conclusion is drawn with tf.idf ; manual analysis allowed us to conclude that legal terms may present various rates of tf.idf, high or low. Fixing thresholds of tf.idf under or above which terms could be assumed legal terms would require long and complex heuristics.

4. Entropy
Entropy is used to measure disorder. We have computed the entropy of the distribution of terms in the corpus. A term largely distributed in a corpus, say occurring in a large number of documents of the corpus, will present a high level of entropy, meaning that this term adds little information to the general distribution of terms in the documents of the corpus. Distribution *r* of term *i* on document *x* is as described in Equation (2). Entropy is as described in Equation (3). Entropy rates for terms in our corpus are as listed in Table 7.

$$r(i)_x = \frac{tf_i(x)}{TF_i} \tag{2}$$

Where $tf_i(x)$ is the frequency of term i in document x,
And TF_i is the total frequency of term i in the corpus

$$S = -\sum_{i,x} r(i)_x \log r(i)_x \tag{3}$$

Table 7. Entropy chracteristics

Entropy	
Minimum	0.0000
1^{st} quartile	0.0000
Median	0.0000
Mean	0.6103
3^{rd} quartile	0.3466
maximum	60.2408

Part is the term presenting the highest rate of entropy. As for all indices, a manual analysis of entropy weights of terms did not allow us to identify thresholds on which relying to distinguish legal terms from non-legal terms.

3.2.2 Irrelevance of Statistical Indices in Seeking Legal Terms

The experiments described above suggest that statistical indices, classically used to identify index terms cannot be used to distinguish legal terms from non-legal terms and, more generally, domain terms from non-domain terms. This conclusion is indeed confirmed by a second experiment here detailed.

In the first step of this second experiment, we identified, among our list of 300 000 terms, a sub-list of terms that are surely legal terms. To obtain a sub-list of terms known to be legal terms, we have used French legal dictionaries available on the Internet. Browsing these dictionaries, we collected 1 490 terms defined in these dictionaries. Using a pattern matching procedure, we extracted from the initial list of 300 000 terms a sub-list of 111 202 (hereafter "legal terms") assumed to be legal terms as they exactly match or include a term known as being a legal term. The rest of the list (185 478) is called "other terms".

In the second step of the experiment, "legal terms" have been compared with the "other terms" on the basis of statistical indices above presented. For each of these indices (tf, idf, tf.idf, entropy), "legal terms" and "other terms" appear to exhibit exactly the same behavior in our corpus. The linear correlation coefficient between "legal terms" and "other terms" is 0.9998 for frequency, 0.9979 for idf, 0.9997 for tf.idf and 0.9999 for entropy.

As an example, Figure 1 shows the number of "legal terms" (in red, or grey) and "other terms" (in black), depending on tf.idf values. To build this graph, we considered terms presenting a tf.idf between 2 and 60, which represent more 97% of our 300 000 terms. Numbers of terms have been computed for each value of tf.idf from 2 to 60 with a step of 2.

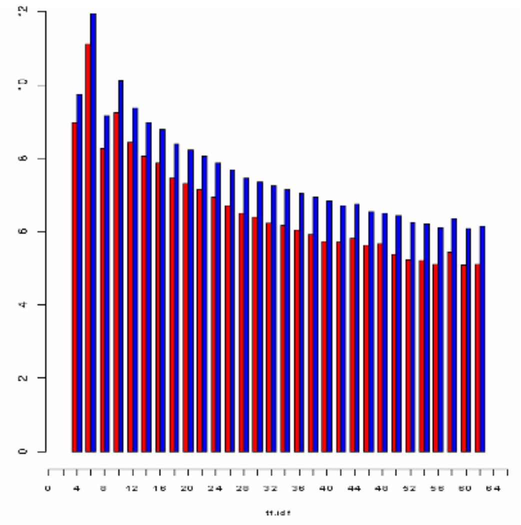

Fig. 1. Number of legal terms and non legal terms depending on tf.idf values

The general conclusion drawn from these two experiments is that statistical indices usually used to identify index terms are useless in domain term identification. This general conclusion can be derived into several statements:

1. Statistical indices such as those used in our experiments can at least be useful in identifying what we call "empty terms". As described in our experiment, high level of tf, tf.idf or entropy allowed us to detect terms inherent to our corpus such as *article*, *chapter* or *title*. We have chosen to use frequency to elaborate a list of 22 empty terms that we manually validated (see Table 8). All the terms of the initial list of 300 000 exactly matching or including one of these 22 empty terms have been removed; we then obtain a list of 118 000 terms.

Table 8. Empty terms

"Empty" term	Frequency
Title	111
chapter	107
Book	91
general provisions	87
common provisions	80

2. Domain terms such as legal terms cannot be assimilated to index terms. Index terms are usually considered as descriptors for document contents. Statistical indices used to detect them tend to single out terms that are discriminating within a given corpus of documents. Domain terms are different from discriminating terms;

a given domain term may occur in most of the documents of the corpus. For example, *contract* in the French Codes is a domain term but cannot be a good index term as it occurs in most of the documents of the corpus.

3. In such a method to identify domain terms, the choice of an appropriate corpus is fundamental. The main result of our experiment is that legal terms and other terms have the same behavior in our corpus. Another statement may be inferred from this: assuming that "other terms" are in fact "legal terms". This statement could be enforced by the fact that we have worked on a carefully chosen corpus: the Codes. This corpus has the particularity to be specific to the legal domain and to have been rationally elaborated (no repetition for example).
4. We have finally decided, on the basis of these statements, to consider our list of 118 000 terms as legal terms. Figure 9 lists a few examples of these 118 000 terms. Meanwhile, a sub-list of this 118 000 terms list has been elaborated and used in the process of detecting relations among terms (see Section 4 below). This list is called "fundamental legal terms". This list has been elaborated using discourse structures [15]. The principle of using discourse structure is to exhibit terms used by the author in specific parts of the text: titles, summary etc… Terms from our list of 118 000 occurring in the titles of Codes structures are considered as "fundamental legal terms". Our list of fundamental legal terms gathers a total of 16 681 terms.

Table 9. Examples of legal terms

Terms
chargeable activities
agricultural activities
updating scientific data
drug forwarding
potential vendee
risk completion

Legal terms have been identified, being assumed that they label legal concepts. These concepts will be one of the components of our ontology.

In the second step of our method, we identify the relations that exist among these legal terms that label legal concepts.

4 Relations Among Terms and Concepts

Semantic relations exist among concepts such as the one linking *divorce* with *marriage* or *damages* with *obligation*. These semantic relations are expressed in texts through syntactic forms such as "*a marriage is dissolved by lawfully pronounced divorce*" or "*the damages result from the non-performance of an obligation*". We then look for syntactic relations among terms to identify semantic relations.

4.1 Texts Analysis Methods

The text analysis we perform on Codes blends syntactical analysis with statistical analysis. We use different methods: syntactical analysis combined with statistical methods, simple syntactical analysis, and purely statistical analysis.

4.1.1 Syntactical Analysis

We used a tool called Syntex [2] to identify terms in our documents (see Section 3.2).

Based on syntactical analysis, Syntex also establishes syntactical dependencies among terms, determining for example that a given noun phrase is subject or object of a given verb. For example, in Article 98 of the Civil Code given above (see Section 2), the tool outputs that *French nationality* is object of *acquire* and *recover*. Contexts are then defined, merging terms with syntactical roles. In our example, the contexts of *French nationality* are [to acquire, OBJECT] and [to recover, OBJECT]. With these results, comparisons of terms with the syntactic contexts they share can be performed, allowing validating semantic relations among terms. For example, *child* and *minor* share contexts such as [guardianship, OBJECT] or [to endanger, SUBJECT].

4.1.2 Analysis of the Coordination Relations

In this methods, documents are parsed and terms that are separated with the conjunctive phrase *and* or *or* [13] are identified. This method relies on the previously established list of legal terms ("fundamental legal terms"). Given this list, a program parses the documents, identifying these terms and checking whether two of them are separated by *and* or *or*. To narrow the results of such a method, it has been applied on the titles of sections and subsections of all the Codes, not on all the texts of the Codes. In the example given above (see Section 2) of the sections and subsections of the Civil Code, such a program outputs that *first names* and *name* may be related, as are *property* and *various modifications of ownership* and *use* and *habitation*. These outputs have to be manually checked to validate which relations are semantically relevant, and which are not.

4.1.3 Statistical Analysis

A statistical method has been performed on the Codes using the previously defined list of legal terms to identify relations among them. The method relies on the idea that two semantically related terms often occur in similar contexts. In this method, contexts are words surrounding a given term, independently of their syntactic roles. Context words may be defined as a given number of words occurring before and after a given term. In our case, context words are defined as all words surrounding a given term occurrence in an article of a Code. In the example presented above (see Section 2), if *French nationality* is the given term, its context words will be *record*, *place*, *birth* etc. Previously defined terms are called target words [9] and the words surrounding these terms are called context words. Each context word is weighted with a mutual information measure which quantifies the dependency existing in texts among the context word and a given target word [9] (see Equation 4).

$$MI_{(cw)} = \log\left(\frac{f_{cw}}{f_c f_w} + 1\right) \tag{4}$$

Where MI = mutual information,
c = context words,
w = target words,
f_{cw} = joint frequency for c and w,
f_c, f_w = individual frequencies of c and w.

A vector linking each context word to its weight is associated to each target word. These vectors are compared two by two with the cosine measure [9] (see Equation 5).

$$Sim_{a,b} = \frac{\sum_{ab} p_a p_b}{\sqrt{\sum_a p_a{}^2 \sum_b p_b{}^2}} \tag{5}$$

Where $Sim_{a,b}$ = cosine similarity measure for terms a and b,
p_a = weight of context words for term a,
p_b = weight of context words for term b,
ab = number of context words shared by term a and term b.

Consequently, each tuple of target words is associated to a similarity score. A threshold has to be defined, above which tuples are considered valid. A manual validation may also be performed on these results.

4.1.4 Pattern Matching

This method relies on a previously defined list of terms. It consists in linking terms with the ones that include them. As an example, with this method, *contract* will be related *to breach of contract*, *contract of deposit* etc. A program parses the list of legal terms and identifies, with a pattern matching function, those that need to be linked together. This method is coarse but, applied to a list of well-identified legal terms, can give good results, especially in our context of an ontology dedicated to information retrieval.

4.2 Results

All the methods presented above have been applied on the 57 Codes available on the governmental site publishing French law on the Internet[5]: the Penal Code, the Civil Code, the Intellectual Property Code etc… Each Code being divided in articles, the 57 Codes represent more than 59 000 articles, gathering a total of more than 6 millions words. Fundamental legal terms, such as defined above (Section 3.2.2), have been used as a previously established list of terms.

On the basis of the list of legal terms, applying the methods above presented, we have identified relations among terms. Most of the methods used to identify relations among terms need manual validation or experimental threshold determination.

The analysis of the coordination relations needs human validation of the outputs of the program parsing the Codes. Applied on the titles of sections and subsections of the Codes, we obtain a list of more than 5 000 sequences of text. Validating these results took us 15 hours to identify 2 580 relevant relations established among 3 762 different terms.

[5] http://www.legifrance.gouv.fr

The statistical analysis based on the outputs of Syntex requires thresholds determination. As stated above, terms are compared on the basis of the syntactic contexts they share. Comparison is quantified with various indices[6]: number of shared contexts, terms and contexts' productivity (number of contexts and terms they respectively occur with) etc. Each of these indices needs a threshold above which it is assumed that results are good. Determining these thresholds requires empirical approximations and tests, comparing the relevance of results for each value of the indices. These experiments have been done, fully described in [4] and in [14].

The statistical analysis that compares terms on the basis of the words they co-occur with also needs a threshold. Contexts are compared with the cosine similarity measure. We have fixed a threshold of 0.8, meaning that two terms are supposed to be related when they share more than 80% of their contexts.

Gathering all the results of these methods, we obtain a list of 103 994 terms, each being related at least once to another term. Among them, 17 688 are related to more than one term. Typical results of all these methods are as described in Table 10.

Table 10. Related terms

Term 1	Term 2	Method
teaching	*Research*	Coordination relation
offence	*Crime*	Syntactical analysis
offence	*Infringement*	Syntactical analysis
minor	*Child*	Syntactical analysis
usufructuary	*exercise of undivided rights*	Statistical analysis
birth	*record of birth*	Pattern matching
contract	*breach of contract*	Pattern matching

5 Toward a Legal Ontology

Legal terms, assumed to label legal concepts, and relations, assumed to match semantic relations among these terms, have been identified. Terms and relations among them put together constitute a graph that we call "ontological resource". This graph can be seen as a description of the legal domain, but an ontology is more than that. An ontology is constituted of concepts and semantic relations among them. In an ontology, concepts are defined by the semantics of the relations established between each concept and others.

The next step of our method is then to infer semantic relations from relations more or less automatically identified. To reach that goal, we have first identified a list of semantic relations labeling ontological relations.

First of all, there is the relation of subsumption is_a. We distinguish two relations of subsumption, a legal one and a general one. The legal one is established between a concept and a legal qualification of its concept, and the general one is established

[6] All these indices are described in [4] and in [14].

between a concept and a general sort of this concept. For example, *universal legacies*, *legacies by universal title* or *specific legacies* are legal sorts of *legacies* defined in French law while *legacy of movables* is a general sort of *legacy*. This means that a given *legacy of movables* may be a *universal legacy*, a *legacy by universal title*, or a *specific legacy*. Depending on this legal qualification, different sets of legal rules may be applied to the given *legacy of movable*. We believe that this distinction made between two kinds of relations of subsumption is specific to the legal domain. The main reason being that the legal is_a relation infers legal qualification and, thus, application of specific sets of legal rules. The second type of relations is the one linking a concept and its components. As an example, the relation between *price of a sale* and *sale*. The third type of relations is the one linking a concept to a related one. For example, the relation existing between *legacy* and *gift*. The last type of relation is the one allowing identifying another sense of the one assumed for the initial concept. For example, if the concept *exchange* is defined as follows: *international exchange* is a legal *exchange*, *multilateral exchange* is a legal *exchange* and *parties of the exchange* is a component of *exchange*; it is clear that *exchange of glances* doesn't have the same meaning. *Exchange of glances* will then be related to exchange with the relation "is another sense of".

All these relations are listed in Table 11.

Table 11. Relations

Relations
Is_a_legal_sort_of
Is_a_general_sort_of
Is_a_component_of
Is_related_to
Is_another_sense_of

We assume that attributing semantic relations to legal terms labeling legal concepts amounts to conceptualization operation, in the sense that these concepts are then defined. This enables us to infer an ontology from our "ontological resource", derived from texts analysis.

Our ontology is integrated in a legal information system that offers interactive request expansion and didactical access to legal documents. This system is available on the Internet: http://ontologie.w3sites.net.

6 Conclusion

In this article, we present a general method relying on text-based NLP techniques to identify components (concepts and relations among them) of an ontology dedicated to information retrieval (IR). Text analysis is performed on particular legal documents: the Codes. These documents have been chosen for their characteristics: the Codes are logically structured and each legal concept is defined. We assume that a conceptualization of the legal field is expressed in these Codes.

This method mainly relies on automatic techniques and tools such as syntactical analyzers of texts or statistics. These automatic techniques do not substitute ontology designers but assist them in the process of ontology design consisting in identifying concepts and relations. NLP techniques are of course relevant for building ontologies dedicated to IR. Meanwhile, we claim that part of these methods may be used while building ontologies dedicated to other tasks such as educational systems [6], decision making systems, or ontologies providing interoperability between systems [18].

References

1. N. Aussenac-Gilles, B. Biébow and S. Szulman. Revisiting ontology design : A method based on corpus analysis. In: Proceedings of Knowledge Engineering and Knowledge Management. Methods, Models and Tools, Juan-les-Pins, France (2000) 172-188
2. T. Bench-Capon. Task Neutral Ontologies, Common Sense Ontologies and Legal Information Systems. In: Second International Workshop on Legal Ontologies, JURIX 2001, Amsterdam, Neederlands (2001)
3. D. Bourigault. Analyse distributionnelle étendue. In: Proceedings of Traitement Automatique des Langues, Nancy, France (2002)
4. D. Bourigault and G. Lame. Analyse distributionnelle et structuration de terminologie, application à la construction d'une ontologie documentaire du droit. Traitement automatique des langues, vol.43, n°1, Ed. Hermès, Paris, France (2002) 129-150
5. G. Braibant. La problématique de la codification. Savoir Innover en Droit. Concepts, Outils, Systèmes. Ed. La documentation française (1999) 55-65
6. J. Breukers and A. Muntjewerff. Ontological modelling for designing educational systems. In: Proceedings of Workshop on Ontologies for Intelligent Educational Systems, Le Mans, France (1999)
7. M. Fernandez, A. Gomez-Perez and N. Juristo. Methontology: From ontological art towards ontological engineering. In: Proceedings of AAAI Spring Symposium Series on Ontological Engineering, Stanford, USA (1997) 33-40
8. A. Gangemi, D. Pisanelli and G. Steve. An Overview of the ONIONS Project : Applying Ontologies to the Integration of Medical Terminologies. Data and Knowledge Engineering, vol. 31 (1999) 183-220
9. S. Gauch, J. Wank, S. Rachakonda. A Corpus Analysis Approach for Automatic Query Expansion and Its Extension to Multiple Databases. ACM Transactions on Information Systems, vol. 17, n°3 (1999) 250-269
10. M. Gruningen and M. Fox. Methodology for the design and evaluation of ontologies. In: Proceedings of IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada (1995)
11. C. Holsapple and K. Joshi. A collaborative Approach to Ontology design. Communications of the ACM, vol. 45, n°2 (2002) 42-47
12. D. Jones, T. Bench-Capon and P. Visser. Methodologies for ontology developement. In: Proceedings of IT-KNOWS Conference, XV IFIP World Computer Congress, Budapest, Hungaria (1998)
13. G. Lame. Constructing an IR-oriented legal ontology. In: Second International Workshop on Legal Ontologies, JURIX 2001, Amsterdam, Neederlands (2001)
14. G. Lame. Construction d'ontologie à partir de textes. Une ontologie de droit dédiée à la recherche d'information sur le Web. PhD dissertation, Ecole des mines de Paris, Paris, France, http://www.cri.ensmp.fr/ (2002)

15. M.-F. Moens, Automatic indexing and abstracting of document texts, Kluwer (2000)
16. L. Mommers. A Knowledge-based ontology of the legal domain. In: Second International Workshop on Legal Ontologies, JURIX 2001, Amsterdam, Neederlands (2001)
17. K. Spark-Jones, Index term weighting. Information storage and retrieval (1973) 619-633
18. H. Stuckenschmidt, E. Stubkjaer and C. Schleider. Modeling Land Transactions : Legal Ontologies in Context. In: Second International Workshop on Legal Ontologies, JURIX 2001, Amsterdam, Neederlands (2001)
19. M. Uschold. Building Ontologies: Towards a unified methodology. In: Proceedings of Expert System, Conference of the British Computer Society Specialist Group on Expert Systems, Cambridge, England (1996)