From Laws and Decrees to a Legal Dictionary

Ismahane Kourtin[1,2](✉), Samir Mbarki[2], and Abdelaaziz Mouloudi[2]

[1] ELLIADD Laboratory, Bourgogne-Franche-Comté University,
Besançon, France
[2] MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Abstract. The mass of information in the legal field, which is constantly increasing, has generated a capital need to organize and structure the content of the available documents, and thus transform them into an intelligent guide, capable of providing complete and immediate answers to queries in natural language. Therefore, the Question Answering System (QAS), which is an application of the Automatic Language Processing domain (NLP), responds perfectly to this need by offering different mechanisms to provide adequate and precise answers to questions expressed in natural language. The general context of our work is the construction of an ontology-based legal question-answering system, allowing users to ask questions about desired information using natural language without having to browse through documents.

In this article, we will mainly focus on the construction of a legal dictionary from textual laws and decrees, for the natural language automatic processing platform NooJ. The legal dictionary that we propose to build from laws and decrees, will bring together the terminological material that will serve as a linguistic resource for the automatic processing of users' questions in natural language, and in particular during the information extraction step which is necessary for the formulation of SPARQL queries equivalent to users' questions.

Keywords: Legal field · Information retrieval · Automatic Language Processing (NLP) · NooJ · Legal dictionary

1 Introduction

Question-answering systems (QASs) offer different mechanisms to provide adequate and precise answers to questions expressed in natural language. Indeed, this type of system allows user to ask a question in natural language and receive a precise answer to his request instead of a set of documents deemed relevant, as in the case of search engines.

The first process in QASs is to extract the information from users' questions that are expressed in natural language. One of the crucial steps in the extracting information from texts is the recognition of named entities. The term named entity appeared during the MUC6 conference (Message Understanding Conference) [1]. These are the entities that have a determined designator (e.g. "EDF", "Jules Verne"). They include proper names or expressions such as the species names (e.g. "Bengal tiger"), diseases, or

M. Bigey et al. (Eds.): NooJ 2021, CCIS 1520, pp. 187–197, 2021.
https://doi.org/10.1007/978-3-030-92861-2_16

chemicals. This definition has also been extended to temporal expressions such as dates and times, or to numeric values (e.g. 2.3 g/l).

By legal entities, we mean named entities specific to the legal field such as acts and facts. Detecting such entities requires the availability of resources describing the domain vocabulary and / or training corpus allowing the learning of the common characteristics to these entities.

Our goal in this article is to build a legal dictionary that will be used for the automatic analysis of the users' questions expressed in natural language in order to extract the information that is needed to formulate SPARQL queries equivalent to users' questions.

The rest of this document is organized as follows: First, Sect. 2 presents related work on extracting terms from texts. Subsequently, Sect. 3 presents the legal field and its complexity. Then, Sect. 4 describes the methodology used for the construction of the legal dictionary. Finally, we end this article with the results of the experimentation of the legal entities recognition by applying our legal dictionary in Sect. 5, and conclude in Sect. 6.

2 Extracting Terms from Texts

A term is an expression with a unique meaning for a particular domain [2]. In the legal field, the words "tax service" become a term in relation to the field, it has a unique meaning in this field.

Term extraction consists of identifying potential terms in a specific text or a set of texts (corpus) as well as the relevant information related to the use of these terms or to the concepts to which they refer (definition, context, etc.).

Extracting terms is an important step in building a dictionary from a corpus. Terms are words or expressions having a precise meaning in a given context, and represent the linguistic supports of the concepts. The problem of building up resources is at the heart of terminological activity. If the notion of "term", which appeals to that of concept and is often based on a particular act of reference, does not seem to lend itself to computer processing, a certain number of tools aiming to extract the terms of a corpus have seen the day [3].

The definition of the term given above exerts strong constraints on the form and the functioning of the terminological units. These constraints constitute the operational principles of terminology extraction software that have been developed in recent years. The objective of these software is to automatically provide a more or less structured lexicon of the domain.

We can distinguish three types of approaches for the automatic term extraction: (i) linguistic approaches that use lists of named entities and manually written recognition patterns [4, 5], (ii) statistical approaches based on learning techniques from annotated texts [6, 7] and (iii) hybrid approaches which integrate the first two methods [8, 9]. Table 1 gives a brief description of each approach for the automatic term extraction.

Table 1. Approaches for the automatic term extraction

Approach	Description
Linguistic methods	These methods generally call upon linguistic knowledge which can be syntactic, lexical or morphological Linguistic methods consider that the construction of the terminological units obeys more or less stable syntax rules, they are mainly phrases formed of nouns and adjectives. Based on this knowledge, these systems perform the extraction of candidate terms using syntactic schemes [10] We can also use grammars and a lexicon acquired during analysis or through collaboration with specialists to generate all the potential terms of a domain [11] These tools therefore require a preprocessing of the corpus by a syntactic analyzer. The quality of the results depends closely on the quality of these analyzers. They have the disadvantage of depending directly on the language of the texts processed and require linguistic resources (dictionaries, stop-word list, etc.). In addition, they are only effective on small corpora
Statistical methods	The statistical approach offers undeniable advantages, since it makes it possible to tackle large data sets that it would be completely impossible to process manually [11] The first works in this field, using statistical data, date from the 80s, they were carried out by Ludovic Lebart and André Salem on the repeated segments [12]. These works exploit similarity measures There are several statistical methods applied to term extraction, most of which are based on mutual information or the Dice coefficient [13]. The principle is that the recurring association of two words cannot be due to chance. Therefore, it is necessarily significant [14]
Hybrid methods	Hybrid models, as their name suggests, are at the crossroads between linguistic and statistical approaches. The existing studies adopt a varying order of treatment. Indeed, some authors prefer to start processing corpora with linguistic analysis, and then filter the results using statistical techniques, while others do the opposite

3 The Legal Field

The legal field is a complex field by its terms which can be:

- Terms with only a legal meaning;
- Terms with at least one legal and non-legal meaning;
- Terms designated by their synonyms in different texts;
- Terms appearing in different morphological forms;
- Non-synonymous terms with the same legal meaning.

In addition, there are different lexical forms that legal terms can take. Table 2 gives some examples of legal terms with their lexical form.

Table 2. Examples of legal terms with their lexical form

Legal term	Lexical form
acquisition	Noun
action nominative	Noun Adjective
adressé par lettre recommandée avec accusé de réception	Verb Preposition Noun Adjective Preposition Noun Preposition Noun
disposition d'ordre législatif	Noun Preposition Noun Adjective
domicile fiscal	Noun Adjective
droit social	Noun Adjective
ensemble immobilier	Noun Adjective
établissement des sociétés non résidentes	Noun Preposition Determiner Noun Adverb Adjective
groupement d'intérêt économique	Noun Preposition Noun Adjective
impôt sur les sociétés	Noun Preposition Determiner Noun
libre disposition	Adjective Noun
membre de la société	Noun Preposition Determiner Noun
nommément désigné	Adverb Verb (PP)
occupé en majeure partie	Verb (PP) Preposition Adjective Noun
opération à caractère lucratif	Noun Preposition Noun Adjective
organisme légalement assimilé	Noun Adverb Adjective
remise contre récépissé	Noun Preposition Noun
service des impôts	Noun Preposition Determiner Noun
société à objet immobilier	Noun Preposition Noun Adjective
société de fait	Noun Preposition Noun
société en participation	Noun Preposition Noun
société immobilière transparente	Noun Adjective Adjective
société nouvellement créée	Noun Adverb Verb (PP)

These examples of legal terms show the diversity and the infinity of the lexical forms of the legal terms. We find terms in the form of "Noun", "Noun-Adjective", "Noun-Preposition-Noun", etc. This lexical diversity makes it impossible to automatically extract the legal terms based on lexical grammars.

No resource on the legal terms has been developed for the legal field. Therefore, we decided to build a NooJ legal dictionary describing the legal terms and their categorization, which will be used for the automatic analysis of the users' questions that are expressed in natural language, using the natural language automatic processing platform NooJ [15]. The latter makes it possible to build, test and manage formal descriptions in a wide coverage of natural languages, in the form of electronic dictionaries and grammars.

4 The Legal Dictionary

The description of natural languages is formalized in the form of electronic dictionaries and grammars represented by organized sets of graphs. NOOJ dictionaries are used to represent, describe and recognize simple and compound words. Dictionaries are.nod files compiled from editable.dic source files.

Our goal is to build an electronic dictionary of legal terms for NOOJ. A term can be simple if it contains one word, or compound if it contains more than one. A compound word is built from simple words. Silberztein M. [16] defines a compound noun as a consecutive sequence of at least two simple forms and blocks of separators. A simple form is a consecutive nonempty sequence of characters of the alphabet appearing between two separators. A single word is a simple form that constitutes a dictionary entry.

The legal dictionary that we propose to build from laws and decrees, will bring together the terminological material necessary for the automatic processing of legal texts, and in particular during the stage of transforming users' questions, in natural language, to SPARQL queries in our question-answering system. We have adopted a methodological framework in 6 steps for the construction of the legal dictionary (see Fig. 1).

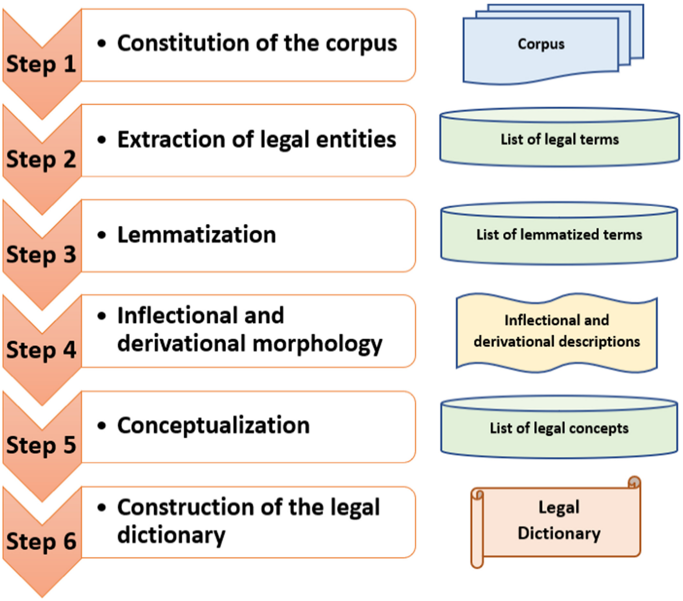

Step 1
Constitution of the corpus
Corpus
Step 2
Extraction of legal entities
List of legal terms
Step 3
Lemmatization
List of lemmatized terms
Step 4
Inflectional and derivational morphology
Inflectional and derivational descriptions
Step 5
Conceptualization
List of legal concepts
Step 6
Construction of the legal dictionary
Legal Dictionary

Fig. 1. Construction stages of the legal dictionary

4.1 The Constitution of the Legal Corpus

In this step we have built up a legal corpus from laws and decrees. We focused our study initially on the general tax code of Morocco. The general tax code has 3 books (see Fig. 2).

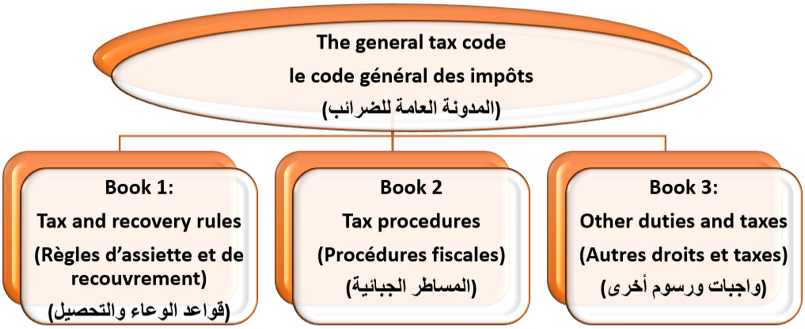
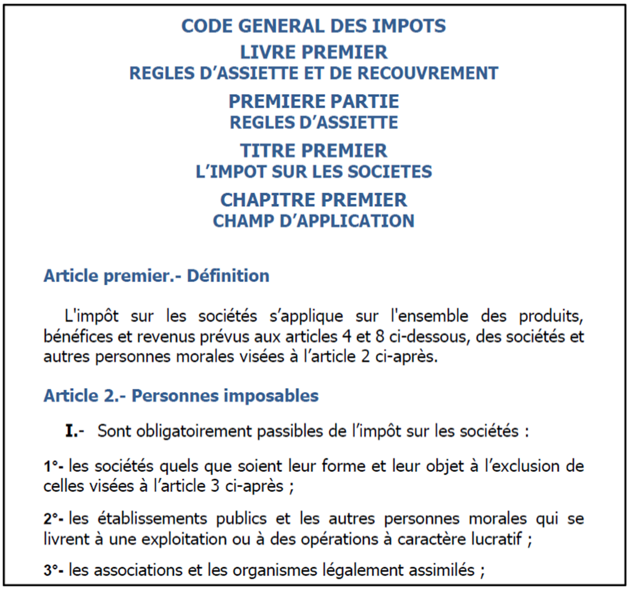


Fig. 2. The general tax code

The first book deals with the tax and recovery rules, and has 9 titles and 209 articles. Book 2 deals with the tax procedures and has 3 titles and 39 articles. Book 3 deals with other duties and taxes and has 5 titles and 40 articles.

We started with the first title of the first book of the general tax code, on "corporation tax" (see Fig. 3).



CODE GENERAL DES IMPOTS

LIVRE PREMIER
REGLES D'ASSIETTE ET DE RECOUVREMENT

PREMIERE PARTIE
REGLES D'ASSIETTE

TITRE PREMIER
L'IMPOT SUR LES SOCIETES

CHAPITRE PREMIER
CHAMP D'APPLICATION

Article premier.- Définition

L'impôt sur les sociétés s'applique sur l'ensemble des produits, bénéfices et revenus prévus aux articles 4 et 8 ci-dessous, des sociétés et autres personnes morales visées à l'article 2 ci-après.

Article 2.- Personnes imposables

I.- Sont obligatoirement passibles de l'impôt sur les sociétés :

1°- les sociétés quels que soient leur forme et leur objet à l'exclusion de celles visées à l'article 3 ci-après ;

2°- les établissements publics et les autres personnes morales qui se livrent à une exploitation ou à des opérations à caractère lucratif ;

3°- les associations et les organismes légalement assimilés ;

Fig. 3. The first title of the first book of the general tax code

4.2 Extracting the Legal Entities

In this step we have manually analyzed the corpus and extracted the legal entities. We identified 679 legal entities.

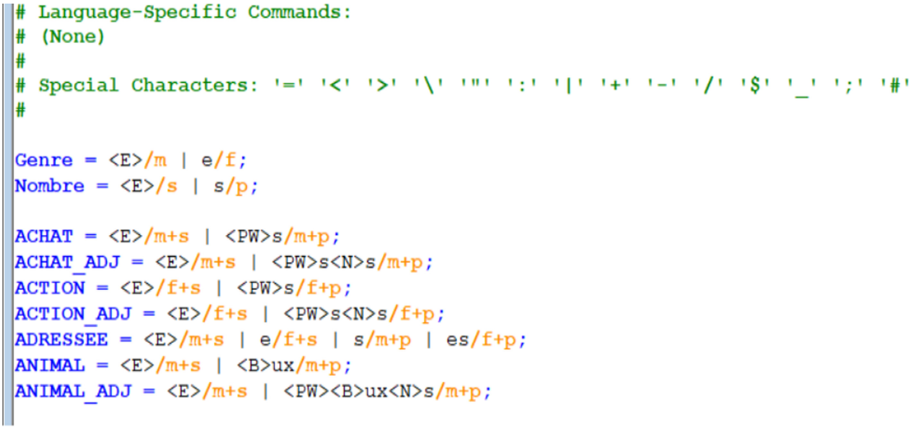
4.3 Lemmatization of Legal Entities

Then, we proceeded to the lemmatization of the extracted legal entities by passing words bearing inflection marks (plural, conjugated form of a verb…) to their reference forms (lemma or canonical form).

For example, the legal entity “Personnes imposables” (Taxable persons) becomes “Personne imposable” (Taxable person).

4.4 Inflectional and Derivational Morphology

In this step, we established the inflected and derived forms of the legal entities using NooJ grammars. An extract is given in Fig. 4.



```
# Language-Specific Commands:
# (None)
#
# Special Characters: '=' '<' '>' '\' '"' ':' '|' '+' '-' '/' '$' '_' ';' '#'
#

Genre = <E>/m | e/f;
Nombre = <E>/s | s/p;

ACHAT = <E>/m+s | <PW>s/m+p;
ACHAT_ADJ = <E>/m+s | <PW>s<N>s/m+p;
ACTION = <E>/f+s | <PW>s/f+p;
ACTION_ADJ = <E>/f+s | <PW>s<N>s/f+p;
ADRESSEE = <E>/m+s | e/f+s | s/m+p | es/f+p;
ANIMAL = <E>/m+s | <B>ux/m+p;
ANIMAL_ADJ = <E>/m+s | <PW><B>ux<N>s/m+p;
```

Fig. 4. An extract of the inflectional grammar

For example, the inflectional model “ACHAT” is defined by “ACHAT = <E>/m +s | <PW> s/m+p;” and means that the legal term that uses this inflectional model has two forms:

- The term as it is: masculine singular
- The term with an “s” at the end of the first word: masculine plural.

4.5 Conceptualization

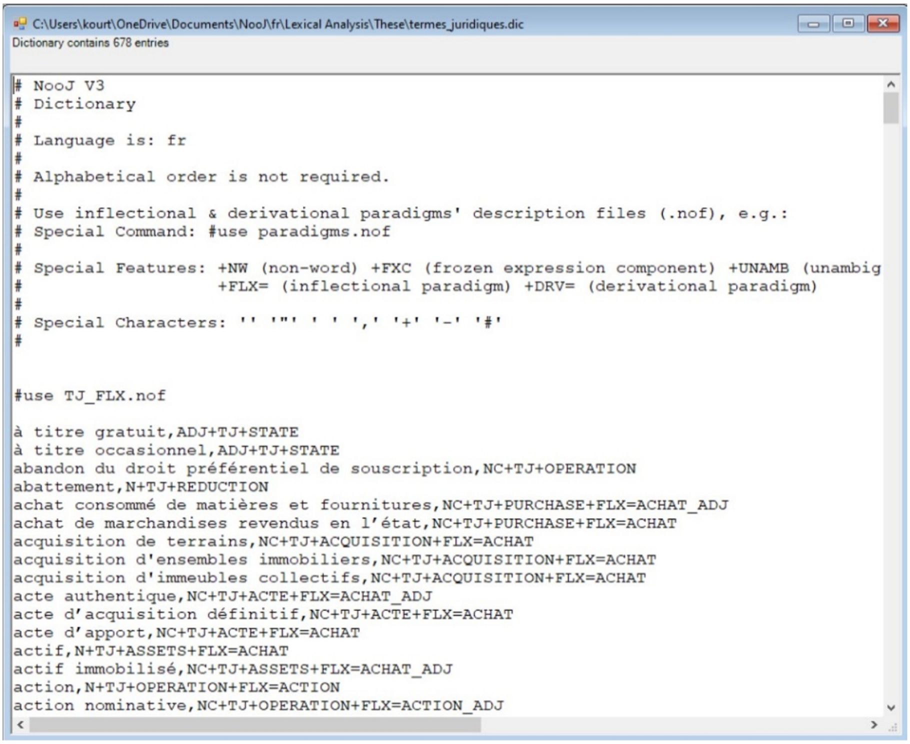
After having established the list of the legal entities, we proceeded to group these entities into semantic classes by establishing a list of concepts. We have established 42 concepts. Table 3 gives some examples of legal concepts with their description and some examples.

Table 3. Examples of legal concepts

Code	Meaning	Examples
ORGANIZATION	Organization	établissement d'animation touristique (tourist entertainment establishment)
COMPANY	Company	jeune entreprise innovante (young innovative company)
ADMINISTRATION	Administration	service des impôts (Tax service)
AGENCY	Agency	agence de développement social (social development agency)
ASSOCIATION	Association	association sportive (sports Association)
BANK	Bank	Banque Européenne d'Investissements (B.E.I.) (European Investment Bank)
PERSON	Person	adhérent, associé, bénéficiaire (member, partner, beneficiary)
VEHICLE	Vehicle	ambulance (ambulance)
ANIMAL	Animal	animal vivant (living animal)
STATE	State	capacité d'hébergement (accommodation capacity)
PURCHASE	Purchase	achat de marchandises revendus en l'état (purchase of goods resold as is)
ACQUISITION	Acquisition	acquisition de terrains (land acquisition)
ACTE	Act	autorisation, avance, agrément, exonération (authorization, advance, approval, exemption)
ACTIVITY	Activity	assistance technique (technical assistance)
HELP	Help	aide au logement (housing assistance)

4.6 The Construction of the Legal Dictionary

Finally, we proceeded to the structuring of the legal terms by building an electronic dictionary of legal terms. The electronic computer dictionary was developed with NooJ [17–19] and has 679 entries. An extract is given in Fig. 5.



C:\Users\kourt\OneDrive\Documents\NooJ\fr\Lexical Analysis\These\termes_juridiques.dic

Dictionary contains 678 entries

```
# NooJ V3
# Dictionary
#
# Language is: fr
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' description files (.nof), e.g.:
# Special Command: #use paradigms.nof
#
# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB (unambig
#                   +FLX= (inflectional paradigm) +DRV= (derivational paradigm)
#
# Special Characters: '' '"' ' ' ',' '+' '-' '#'
#


#use TJ_FLX.nof

à titre gratuit,ADJ+TJ+STATE
à titre occasionnel,ADJ+TJ+STATE
abandon du droit préférentiel de souscription,NC+TJ+OPERATION
abattement,N+TJ+REDUCTION
achat consommé de matières et fournitures,NC+TJ+PURCHASE+FLX=ACHAT_ADJ
achat de marchandises revendus en l'état,NC+TJ+PURCHASE+FLX=ACHAT
acquisition de terrains,NC+TJ+ACQUISITION+FLX=ACHAT
acquisition d'ensembles immobiliers,NC+TJ+ACQUISITION+FLX=ACHAT
acquisition d'immeubles collectifs,NC+TJ+ACQUISITION+FLX=ACHAT
acte authentique,NC+TJ+ACTE+FLX=ACHAT_ADJ
acte d'acquisition définitif,NC+TJ+ACTE+FLX=ACHAT
acte d'apport,NC+TJ+ACTE+FLX=ACHAT
actif,N+TJ+ASSETS+FLX=ACHAT
actif immobilisé,NC+TJ+ASSETS+FLX=ACHAT_ADJ
action,N+TJ+OPERATION+FLX=ACTION
action nominative,NC+TJ+OPERATION+FLX=ACTION_ADJ
```

Fig. 5. An extract of the legal dictionary

For example, for the dictionary entry “acte d’acquisition définitif”:
acte d’acquisition définitif, NC+TJ+ACTE+FLX = ACHAT.

- acte d’acquisition définitif: the legal entity
- +NC+TJ: the categories are compound noun and legal term
- +ACTE: the semantic class “ACTE”
- ACHAT: the inflectional model “ACHAT”

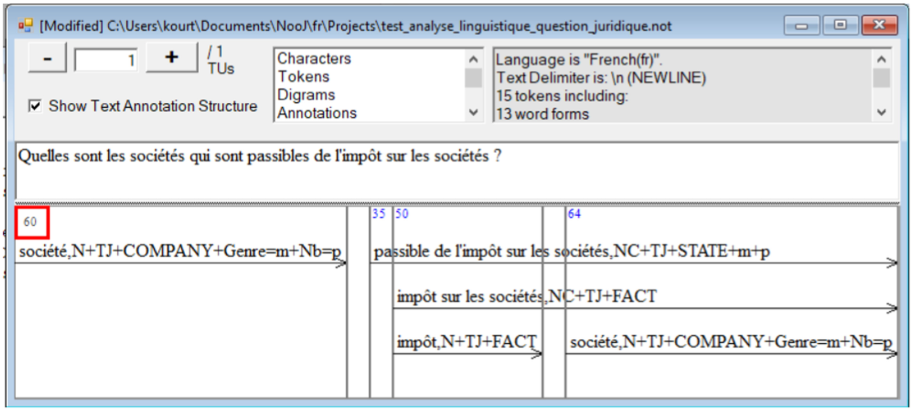
The inflectional model “ACHAT” is defined by “ACHAT = <E>/m+s | <PW>s/m +p;” which means that the legal term has two inflected forms:

- acte d’acquisition définitif: masculine singular
- actes d’acquisition définitif: masculine plural.

5 Experimentation

The NooJ legal dictionary, which we have developed, is able to annotate and recognize legal entities in natural language text. However, with the legal dictionary one is able to automatically analyze and recognize legal terms in natural language questions, using the natural language automatic processing platform NooJ.

Figure 6 shows the result obtained from the annotation, with the NooJ legal dictionary that we built, of the question in French "Quelles sont les sociétés qui sont passibles de l'impôt sur les sociétés?" (Which companies are liable to corporation tax?). The result of the annotation shows that the term "société" (company) was identified by: noun and legal term masculin plural, of semantic class "COMPANY"; and that the term "passibles de l'impôt sur les sociétés" (liable to corporation tax) was identified by: noun and legal term masculin plural, of semantic class "STATE".


[Modified] C:\Users\kourt\Documents\NooJ\fr\Projects\test_analyse_linguistique_question_juridique.not
-
1
+
/ 1
TUs
☑ Show Text Annotation Structure
Characters
Tokens
Digrams
Annotations
60
35
50
64

Language is "French(fr)".
Text Delimiter is: \n (NEWLINE)
15 tokens including:
13 word forms

Quelles sont les sociétés qui sont passibles de l'impôt sur les sociétés ?

société,N+TJ+COMPANY+Genre=m+Nb=p

passible de l'impôt sur les sociétés,NC+TJ+STATE+m+p

impôt sur les sociétés,NC+TJ+FACT

impôt,N+TJ+FACT

société,N+TJ+COMPANY+Genre=m+Nb=p

Fig. 6. The result of the annotation with the NooJ legal dictionary

6 Conclusion

In this work we have developed an electronic NooJ dictionary that allows annotating and recognizing legal terms in natural language texts. We have adopted a methodological framework in 6 steps for the construction of the legal dictionary: (1) we have constituted a legal corpus of laws and decrees focusing on the first title of the first book of the general tax code, on "corporation tax"; (2) we manually analyzed the corpus and extracted the legal entities by identifying 679 legal entities; (3) we lemmatized the extracted legal entities by passing words bearing inflection marks (plural, conjugated form of a verb…) to their reference forms; (4) we have built grammars describing the inflectional and derivational morphology of the legal entities; (5) we have grouped the legal entities into semantic classes by establishing 42 concepts; (6) we have structured legal entities by building a NooJ electronic legal dictionary capable of annotating and identifying legal terms in natural language texts.

As perspectives, we will integrate the legal dictionary into our question-answering system, by using it in the automatic processing of the users' questions in natural language, which the objective is to extract the information necessary for the formulation of SPARQL queries equivalent to users' questions.

References

1. Grishman, R., Sundheim, B.: Message understanding conference 6 - a brief history. In: Proceedings of COLING, Copenhagen, Denmark, (AUG 1996), pp. 466–471 (1996). (Cited pages 17 & 19)
2. Azé, J., Heitz, T.: Cours sur la Fouille de textes et Apprentissage (2004). http://www.lri.fr/~aze/enseignements.php
3. Piwowarski, B.: Techniques d'apprentissage pour le traitement, d'informations structurées: Application à la recherche d'information, Doctoral thesis, University of Paris 6 (2003)
4. Poibeau, T.: Le repérage des entités nommées, un enjeu pour les systèmes de veille. In: Terminologies Nouvelles (actes du colloque Terminologie et Intelligence Artificielle, TIA'99, Nantes), no. 19, pp. 43–51 (1999). (Cited page 17)
5. Elkateb-Gara, F.: Extraction d'entités nommées pour la recherche d'informations précises. Dans 4e Congrès ISKO France, Grenoble (2003). (Cited page 17)
6. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, features induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (2003). (Cited page 17)
7. Raymond, C., Wei, W.: Named entity recognition using hybrid machine learning approach. In: IEEE ICCI, pp. 578–583 (2006). (Cited page 17)
8. Kosseim, L., Poibeau, T.: Extraction de noms propres à partir de textes variés: problématique et enjeux. In: TALN 2001, pp. 365–371 (2001). (Cited page 1)
9. Fourour, N.: Nemesis: un système de reconnaissance incrémentielle des entités nommées pour le français. In: TALN 2002, pp. 255–264 (2002). (Cited page 17)
10. Malaisé, V.: Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels, Doctoral thesis, University of Paris 7 – Denis Diderot France (2005)
11. Drouin, P.: Acquisition automatique des termes: l'utilisation des pivots lexicaux spécialisés, Doctoral thesis, University of Montreal (2002)
12. Lebart, L., Salem, A.: Analyse statistique des données textuelles. Dunod, Bordas, Paris (1988)
13. Velardi, P., Missikof, M., Fabriani, P.: Using text processing techniques to automatically enrich a domain ontology. In: Proceeding of ACM-FOIS (2001)
14. L'Homme, M.-C.: Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. In: L'impact des nouvelles technologies sur la gestion terminologique, Toronto (2001)
15. Silberztein, M.: NooJ manual (2006)
16. Silberztein, M.: Le dictionnaire électronique des mots composés. In: Langue Française, No. 87, septembre 1990
17. Aoughlis, F.: A computer science electronic dictionary for NOOJ. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) NLDB 2007. LNCS, vol. 4592, pp. 341–351. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73351-5_30
18. Aoughlis, F.: Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales. Thèse, Université Mouloud Mammeri, Tizi Ouzou (2010)
19. Hildebert, J.: Dictionnaire des technologies de l'informatique. vol. 2, Français/Anglais, La maison du dictionnaire (Paris), Hippocrene Books Inc., New York (1998)