Semantic modelling of university regulations

Presented to the Faculty of New Information and Communication Technologies in Fulfillment of the Requirements for the Degree of Doctorate in

COMPUTER SYSTEMS & NETWORKS

Presented by

Nasria Bouhyaoui

Publicly defended on February 16, 2020

Jury members

Dr. Fouad Chebbara	University of Ouargla	President
Pr. Mohamed Benmohamed	University of Constantine	Examiner
Pr. Okba Kazar	University of Biskra	Examiner
Dr. Djamel Samai	University of Ouargla	Examiner
Dr. Mohamed Redouane Kafi	University of Ouargla	Examiner
Pr. Fatima Zohra Laallam	University of Ouargla	Supervisor

Academic Year: 2019/2020

ملخص

عدد كبير من الوثائق القانونية أصبح متوفرا إلكترونيا. تتعلق الوثائق القانونية بالعديد من المجالات مثل الرعاية الصحية والمالية والتعليم،إلخ. تلعب هذه النصوص القانونية دورا هاما، وتعتبر بمثابة مادة خام للمنظمات، سواء كانت عامة أو خاصة حيث يجب أن يكون كل طرف على علم بها وأن يمتثل لها. إلا أنه بسبب الصعوبات في المجال القانوني، يفضل الأشخاص الاعتماد على الخبراء بدلاً من اللجوء إلى البحث عن القوانين في مجموعة من الوثائق. في هذه الرسالة، نستخدم منهجًا قائمًا على القواعد يستند على طريقة الاستكشاف السياقية من أجل الشرح الدلالي للنصوص القانونية الجزائرية المكتوبة باللغة العربية. نحن مهتمون بتحديد المعلومات الدلالية وهي أنواع الأحكام (الالزام ، والجواز والحظر) وكذا العناصر (الدور و الفعل). بالإضافة إلى ذلك، نحن مهتمون بتمثيل النصوص القانونية من أجل تسهيل عملية الوصول إلى الوثائق القانونية واستردادها. قدمت التجارب الأولية التي أجريت على النصوص القانونية الجزائرية العربية نتائج واعدة في تحديد أنواع الأحكام القانونية.

كلمات مفتاحية: شرح دلالي، وثيقة قانونية، اللغة العربية، حكم، نهج قائم على القواعد، استكشاف سياقي.

Résumé

De nos jours, une quantité énorme de documents légaux est devenue disponible en version électronique, concernant de nombreux domaines tels que la santé, la finance et l'éducation. Les textes légaux jouent un rôle important dont ils sont considérés comme une matière première pour l'organisation, qu'elle soit publique ou privée, où chaque acteur doit connaitre et être conformer aux réglementations. En raison des difficultés du domaine légale, les acteurs préfèrent de s'appuyer sur l'expert plutôt que de recourir à la recherche de la réglementation dans une collection de documents. Dans cette thèse, nous utilisons une approche à base des règles. Ces règles sont basées sur la méthode d'exploration contextuelle pour l'annotation sémantique de textes légaux algériens écrits en langue arabe. Ici, nous nous intéressons à la spécification des informations sémantiques de types provision : obligation, permission et interdiction, et ainsi l'identification des arguments : rôle et action. Ce travail à pour l'objectif de faciliter le processus de la recherche et d'accès aux documents légaux. L'expérimentation menée sur les textes légaux algériens et arabes a présenté des résultats prometteurs pour la spécification des types de provisions.

Mot clés : Annotation sémantique, document légale, Langue Arabe, provision, approche à base de règle, exploration contextuelle.

Abstract

Nowadays, a huge amount of legal documents becomes electronically available. These legal documents concern many domains such as healthcare, finance and education. Legal texts play an important role, and they are considered as raw material for organization, be it public or private where each actor must be aware of, and comply with regulations. However, because of the difficulties of the legal domain, the actors prefer to rely on the expert rather than resorting to searching for the regulation in a collection of documents. In this thesis, we use a rule-based approach which based on the contextual exploration method for the semantic annotation of Algerian legal texts written in Arabic language. We are interested in the specification of the semantic information of the provision types: obligation, permission and prohibition, and the arguments: role and action. In addition we are interested in the representation of the legal texts in order to facilitate the process of accessing and retrieving legal documents. The preliminary experiment conducted over the Algerian Arabic legal texts presented promising results for the specification of provision types.

Keywords: semantic annotation; legal document; Arabic language; provision; rule-based approach; contextual exploration.

Dedication

To my beloved parents.

To my dear family and friends.

Acknowledgements

Table of Contents

List of Figures

List of Tables

Acronyms

AI: Artificial intelligence.

CE: Contextual Exploration.

GSG: General Secretariat of Government.

IE: Information Extraction.

IR: Information Retrieval.

ML: Machine Learning.

NER: Named Entity Recognition.

NLP: Natural Language Processing.

POS: Part-of-Speech

WSD: Word sense disambiguation.

XML: Extensible Markup Language.

RDF: Resource Description Framework.

RDFS: RDF Schema.

OWL: Ontology Web Language.

TP: True Positive.

FP: False Positive.

FP: False Negative.

General Introduction

General Introduction

Nowadays, a huge amount of legal documents becomes available electronically. These legal documents concern many domains such as healthcare, finance and education. Regulations affect the life of each individual by imposing a set of constraints in their actions. Hence, each individual and organization must comply with legislations. Failure to comply with the legislation can lead to penalties, lawsuits…etc.

Legal texts play an important role, and they are considered a raw material for organizations where each organization must be aware of laws and regulations. The mastery of these texts plays a crucial role in their management. With the growing importance of knowledge as the main engine of growth, the university is regarded as an essential element of human development in the world. The success of the University's mission is linked to the success of the tasks it undertakes. Where it must be organized, and their actors must know what to do at all times, to benefit from their time as best as possible.

The actors and the stakeholders of the Algerian administration are expected to know what their obligations are, and they must search for the relevant laws relating to their activities. Therefore, it is important to be able to access researched information in the context of the administration. However, several complexities arise when dealing with legal documents; because of the variety of legal sources and the dynamic character of law, access to the relative information becomes a laborious task. It is difficult to retrieve information existed among a large number of legal documents, and in particular, by citizens who are unfamiliar with the legislation. The legal text is also characterized by a complex nature (contradictory, ambiguous, etc.) and several challenges are encountered when it comes to dealing with, where it is difficult to be understood by a non-domain specialist. These complexities lead the actor to prefer to rely on the expert rather than resort to the search for legal rules in a collection of documents.

The government of Algeria makes the legal texts (laws, decrees, ordinances, etc.) available on the Web via the official portal. The system offers the possibility of accessing the legal text by filling out a set of defined fields; the latter is 1) Sector (Set of ministries: Education; Higher Education and Scientific Research; Health, Population, and Hospital

Reform; Foreign Affairs.). 2) Ministry. 3) Nature of the text (Law, Ordinance, Decree, etc.). 4) Journal number. 5) Date of publication. 6) The number of the text and its date of signature. 7) And finally, by a combination of keywords. A whole of this information requires familiarization of the laws while this is not the case with all actors and stakeholders.

These obstacles cause reluctance of the administrative actors and stakeholders to resort to the legal texts. This reluctance leads to several consequences including wasting time and helping to provide a climate conducive to the emergence of the phenomenon of lack of transparency. Therefore, there is a growing need for methods and tools that facilitate the mission of identification and extraction of information from legal texts. Legal texts contain a set of legal rules. The annotation of the semantics of the rules can help to improve the retrieval of the provisions.

Annotation is the process of augmenting a text or a portion of text (word, sentence, paragraph, etc.) with labels (Descles, 2006). The manual annotation of the legal text is a laborious, costly, and time-consuming task. Therefore, the use of the methods and the tools that facilitate this task becomes a necessary requisite (Kiyavitskaya et al., 2008). Legal text annotation is a challenging task in the field of Artificial Intelligence and Law (Lesmo et al., 2013).

Various researchers tend to process the information that exists in the legal texts; in which they claim that language processing is essential to extract the information contained in the legal document (Francesconi et al., 2010). In Algeria, legal texts are issued in two versions, Arabic and French; where, the first is the official language. Arabic is considered a Semitic language closely related to Hebrew and Aramaic. It ranks in fifth place of the most spoken language in the world. Arabic is the official Language in 26 countries, and it is the mother tongue of approximately 280 million speakers (Most Widely Spoken Languages, 2017). The Arabic alphabet consists of 28 letters, which can extend to ninety by added shapes, marks, and vowels (Khorsheed, 2002). The form of the letter differs depending on whether it occurs at the beginning, middle, end or alone. In contrary to Latin-based alphabets, the Arabic writing is from right to left (Amin, 1998). Also, it is characterized by a rich and complex morphology (Farghaly and Shaalan, 2009). The Modern Standard Arabic is different from classical Arabic, where the latter is the regular

version of Arabic. The Modern Standard Arabic is written without short vowels (Boudelaa, 2010), and this is one of the main features that distinguish it, where it is used in the official documents, books, newspapers, etc.

Some characteristics of legal language facilitate the task of processing natural language. However, some particularities may involve more complexities, such as ambiguity and the dynamic nature of the legislation. In the legal Arabic language, more complexities are confronted because of the particularities of Arabic language such as vocalization, the agglutinative nature and the free order of words.

Several studies have been proposed for the processing of legal concepts in the languages Italian, Dutch, English, etc., unlike Arabic, Hindi, Chinese, etc. in which there has been little research. The techniques relating to Morpho-Syntactic analysis did not yield satisfactory results in Arabic language where we observe that there is a lack of available tools. The indicators that indicate the existence of the provision and which can identify their type are written without diacritics, and some words are polysemy, meaning that the same word can express a different meaning. Moreover, the difficulty of identifying the role and action arguments; taking into consideration the insufficient results for Morpho-Syntactic analysis.

Previous studies such as SALEM (Bartolini et al.(2004); Biagioli et al.(2005); Soria et al.(2007); Spinosa et al.(2009)); Kiyavitskaya et al. (2008); Wyner and Peters (2011) and Zeni et al. (2015); etc. have used either the classification or the annotation for the specification of the type of provision. Where, in both, they have depended on the indicators, however, depending on the context; the indicator can take a different meaning. With the particularities addressed in Arabic legal texts, it is not possible to rely only on the words indicating the existence of the provisions in the sentence. Accordingly, it is not possible to depend on the tools and methods proposed by the above mentioned works.

The question asked in this research is, how to specify and represent the semantic information contained in the legal texts, to make them understandable and exploitable? In other words:

- What are the challenges of Arabic legal texts?

- What is the adequate method for the identification of the semantic information of Arabic legal texts?

- What is the main information in Arabic legal texts? How to specify the semantic information existed at the level of the rules?

- How to organize and represent this knowledge in order to facilitate the access and the retrieval of the semantic information?

Legal texts contain a set of legal rules (provisions or prescriptions), their types and arguments represent the semantics of rules (Francesconi and Passerini, 2007). The annotation of the semantics of the rules can help to improve the access and retrieval of the provisions. The annotation based on Information Extraction (IE) techniques: Rule-Based and Machine Learning approach. In this research, we adopt a rule-based approach based on the Contextual Exploration method (CE) proposed by Descles (1997). Our hypothesis is that the semantic annotation of legal texts is based on the identification of the indicators as well as on a set of clues to help disambiguate the terms. The contextual exploration approach responds to these needs. It is an approach based on a set of linguistic markers and formal rules. This research focuses on identifying the list of indicators for each type (obligation, permission and prohibition) and identifying the clues that may exist with each indicator, and creating a set of rules corresponding to the indicators. We are interested also in the representation of the legal texts in order to facilitate the access and the retrieval of the semantic information.

In this thesis, we are interested in the annotation of Algerian legal texts written in Arabic language. We have decided to address the problem of identifying legal rules from legal documents. In the legal texts, there are several categories, such as constructive rules, regulative rules, etc. we choose to focus on the provisions (prescriptions), specifically in the following types: obligation, permission and prohibition; these regulations are widely used and are of interest to the actors of administration and the stakeholders. Our objective is the annotation of the semantic information contained in the legal texts. We are more interested in the specification of the type of provision and their arguments, where the arguments on which we focus on in this thesis are the role and the action. We are interested also in the semantic representation of legal texts by building an Arabic legal ontology that

includes the main concepts for the aim of populating the extracted information in the created ontology.

The main contributions in this thesis are:

- Building a data set that concerns Algerian Arabic legal documents which concern the university.
- Specification of the semantic information contained at the level of Arabic legal provisions.
- Categorization of Arabic legal provisions.

The remains of the thesis are organized as follows.

- Chapter 1 presents the state of the art; it provides some main areas of Artificial Intelligence: Natural language processing, Text categorization and Ontology. Next, law text field is discussed by introducing the domain and describing the legal document structure and content. And then we discuss the related works on legal documents.

- Chapter 2 presents the characteristics of Arabic language and the challenges of the Arabic texts, and after that, we point out some particularities that distinguish the legal domain. Next the Contextual Exploration method is described.

- Chapter 3 provides the methodology of the research, we describe the kind of information contained in the provision and those that interest us, and then we define the markers that can help in categorizing the provisions and specifying the role and action. Next, we explain the method proposed for the specification of the semantic information and then the semantic annotation process is described. After that we present the Arabic legal ontology.

- Chapter 4 presents the results and the evaluations on the manually built corpus of Algerian Arabic legal texts and on the built ontology.

- Finally, the conclusion and future work of the thesis are pointed out.

Chapter 01

Artificial Intelligence and Law

1. Artificial Intelligence and Law

"Intelligent indexing, querying, searching, filtering, retrieving and annotating the ever increasing amount of legal text documents is a major challenge in the field of Artificial Intelligence and Law concerned with Natural Language Processing" (Lesmo et al., 2013). This chapter presents three parts, the technologies of AI, the domain of law, and the related works in AI and Law concerned with our research.

The first part describes the Natural Language Processing (NLP) and presents its different applications. And then we describe the application of relevance to our study which is text categorization. After that we study the difference between text categorization methods. And in the last point we describe the ontology.

The second part concerns the legal domain, first we depict the different definitions of law text and its fields, and next we present the main source of legislation in Algeria. Then we illustrate the structure of legal documents and their content. After that we present the different categories of legal rules. In last, we define the elements of the provision.

The last part presents synthesizes of some research issued in AI and Law domain.

1.1. Artificial Intelligence

1.1.1. Natural language processing

Natural Language Processing (NLP) is a subject of interest of many fields like computer science, artificial intelligence and linguistics, robotics, etc ((Liddy, 2001); (Chowdhury, 2003)). NLP has been investigated in many tasks including: question answering, text summarization, machine translation, speech recognition, opinion mining, and text categorization. The process of automatic understanding of texts is what we call natural language processing. So, NLP techniques are developed for the aim of understanding human language (Chowdhury, 2003); (Liddy, 2001)) by machines.

The author Liddy (2001) provided the following definition for NLP "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for

the purpose of achieving human-like language processing for a range of tasks or applications".

NLP starts as the intersection of Artificial Intelligence and Linguistics. "Linguistics is the science of language which includes Phonology that refers to sound, Morphology word formation, Syntax sentence structure, Semantics syntax and Pragmatics which refers to understanding" (Plisson et al., 2004).

1.1.1.1.   Natural Language Processing layers

NLP systems mission requires the use of several methods. NLP methods can consist of different representation levels including Phonology, Morphology, Lexical, Syntactic, Semantic, Discourse and Pragmatic (Liddy, 2001).

Generally, NLP methods fall in three levels: syntactic, semantic and pragmatics.

1) Syntactic level

The syntactic level focuses on the definition of sentence structure which is the constituent words of the sentence. In addition, it is interested in the detection of the sentence grammatical structure (Liddy, 2001). So, it concerns in the grammar and sentence structure. This level produces as a result a range of structural relationships between the sentence words. Syntactic layer has several modules including tokenization, Sentence Boundary Disambiguation, Lemmatization, Part-of- Speech tagging. In the following we describe some of the modules:

➢ Sentence Boundary Disambiguation

Sentence Boundary Disambiguation is the task of text segmentation into sentences. This deconstruction is an important phase in order to deal with more granular portion of text for the aim of achieving a particular treatment. In English and many other languages, the punctuation marks are used for recognizing boundaries of sentences. However, Finding Boundary of sentences is a challenging task in some languages (e.g. Arabic, Chinese). For example, the Arabic language does not follow a strict punctuation rules. Where, Arabic paragraphs can be written with

only one period at the end of the paragraph. Also, it is characterized by the use of coordination, subordination, etc. these features made the task of **Sentence Boundary Disambiguation** so complex.

➢ Lemmatization

Lemmatization aims for the construction of a normalized form (Plisson et al., 2004). Lemmatization is the process of converting words into their basic word form, which is called the lemma. In another words, lemmatization is the grouping together of different forms of the same word. Lemmatization process mapping all verb forms to infinite tense and converting nouns to a single form (e.g. The words help, helps, helped, helping are mapped to the verb help). Lemmatization is considered a hard task when it comes to process highly inflected languages (Toman et al., 2006).

➢ Part-of-Speech Tagging

Part-of-Speech Tagging (POS) tagging is considered one of the important task in NLP. POS tagging consists of defining a speech part of a particular sentence through the association of labels such as noun, pronoun, verb, adverb, adjective, etc. to each word or token in the given sentence. POS also called grammatical tagging is the task of automatic assigning a tag to a word in a text as corresponding to a particular part of speech . It is used for building NER and lemmatiser. Different techniques are used for POS tagging like lexical based methods, rule-based methods, Probabilistic methods and deep Learning Methods.

2) **Semantic level**

The Semantics level input is the output of the syntactic layer. Semantics level is applied when it comes to extract meaningful information. This level concerns the definition of the meaning of words and sentences (Chowdhury, 2003). The author Liddy (2001) said that:"Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the

sentence". Accordingly, the sentence meaning is derived relying on words meaning (Cambria et al., 2017). Semantic layer has several modules including Named Entity Recognition, Concept extraction, Word sense disambiguation and Anaphora resolution, etc. In the following we briefly describe some of modules that belong to this level:

> Named Entity Recognition

Named Entity Recognition (NER) also known as (*entity extraction and entity identification*) is a popular technique applied in order to extract relevant information or entities from unstructured text. NER is an important method used for identifying and classifying named entities in free text into pre-defined categories or classes such as names of persons, organizations, locations, times expressions, numerical quantities, etc.

> Word sense disambiguation

Word sense disambiguation (WSD) task consists of recognizing the meaning of an ambiguous word depending on its context (Stevenson and Wilks, 2003). Some words can convey multiple meanings; the correct word sense is detected according of the particular context. Identifying the correct sense is an important task in NLP. WSD is an open problem that needs to be solved in order to improve understanding of natural language (Cambria et al., 2017). Where, it is so difficult to define the right meaning of the word in different context.

3) Pragmatic level

The Pragmatic level requires in its task both syntactic and semantic layers. The pragmatic level is interested in the meaning of the text. That meaning extracted depending on the context of the text. The authors Cambria and White (2014) say that "Pragmatics deals with how meaning changes in the presence of a specific context and how the contexts affect the meaning of the sentences". Pragmatic layer has

several modules including Sarcasm Detection, Aspect extraction, Polarity detection, etc.
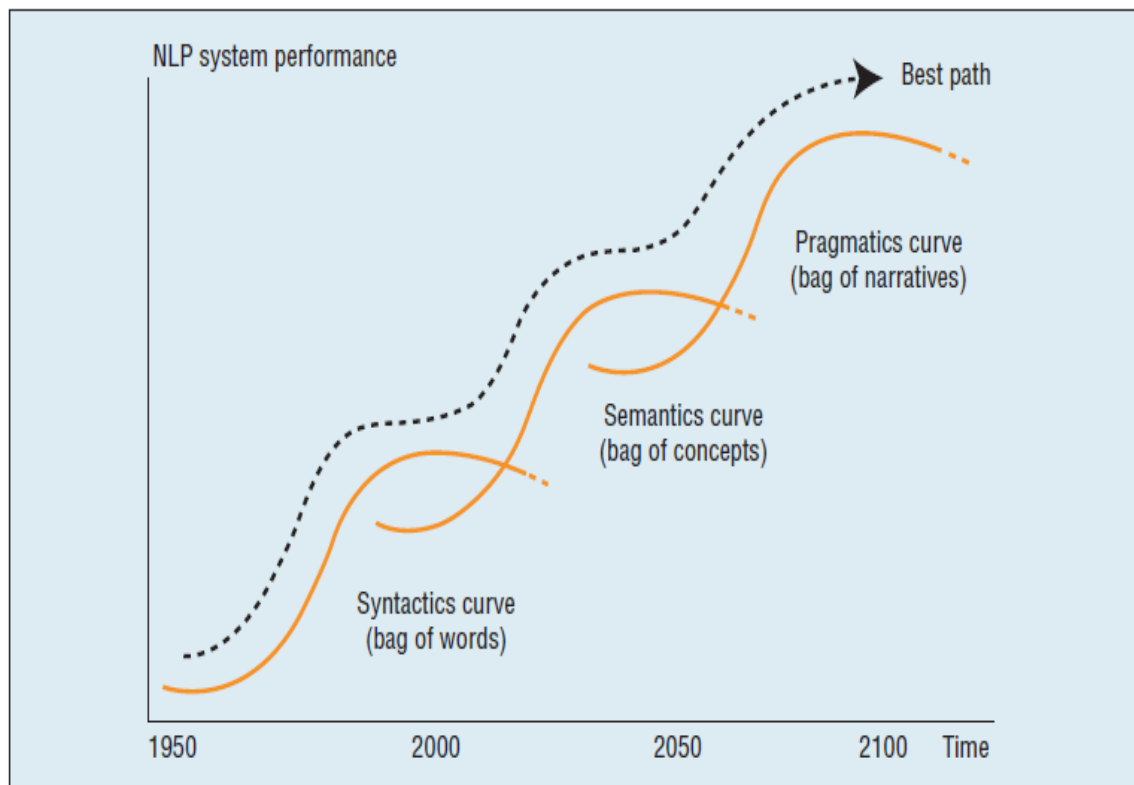
Fig.1.1 Jumping NLP curves (Cambria et al., 2017).

The process of understanding natural language is a complex task, where, many challenging tasks can be faced. The authors Cambria et al. (2017) illustrated the estimation of the evolution of NLP research through the three layers (syntactic, semantic and pragmatic) that lead the NLP research to arrive at the natural language understanding. The illustration is presented in Fig.1.1. The difficulty of NLP processing varies from language to language, following the nature or the characteristics of the language.

1.1.1.2.   Natural language processing application

1) Information Retrieval

The most known information retrieval systems are Google, Bing and Yahoo. Information Retrieval (IR) systems help users to find the information that they need. The information is

relevant to some query, which is a set of terms expressed in natural language that occur in the document. Where it does not return information but locate the document that may contain the information. So, IR is the process of finding the most relevant documents from a large datasets. Where it is interested in accessing process rather than locating relevant information within each text, which is the main purpose of Information Extraction.

2) Information Extraction

Information Extraction (IE) has been studied a lot in the past decade. Because of the necessity of extracting information automatically from text, IE methods aims for identifying important information in texts (Muslea, 1999). IE represents the process of identifying and extracting relevant information or main concepts from unstructured or semi-structured documents. This task permits to produce structured information (Jiang, 2012) depending on a set of text extraction rules (Cunningham et al., 2011). The popular example of IE is NER. IE relies on NLP methods to facilitate the location of information from natural language.

3) Question Answering

Unlike information retrieval systems that provide a set of document to the questions, the Question Answering provides the correct answer to the questions asked by users in natural language using a collection of documents or a database.

4) Text Summarization

Text Summarization is one of the challenging tasks in the NLP domain. The huge amount of documents on the web has leaded the researchers to make more interest on automatic text summarization field (Allahyari et al., 2017). Text summarization is the process of generation of a brief version from a long source text. The produced version should contain the most important information. According to Maybury (1995) "summary distills the most important information from a source (or sources) to produce an abridged version of the original information for a particular user(s) and task(s)". This technique is classified into two types which are the extractive and the abstractive. Text summarization technique can be used as feature selection method in the application of text categorization.

5) Machine Translation

Machine translation is defined in Oxford dictionary as "translation carried out by a computer". Machine translation concerns the task of automatic translation of text from one natural language into another. "Everybody has access to machine translation by visiting popular websites such as Google Translate and Systran's Babelfish" (Bikel and Zitouni, 2012).

6) Speech Recognition

Speech recognition "also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program" (Anusuya and Katti, 2009).

7) Text Categorization

Text categorization is one of the important applications in NLP. In the following section we will tackle this subject in detail.

1.1.2. Text categorization

Text Categorization (TC) has gained an important interest; it is applied in all applications that involve the organization of documents. Categorization is the classification of texts into predefined categories, where it depends on the key features that permit to distinguish the classes between each other. That features or characteristics are defined corresponding to the document or the class category. The organization of documents enables to facilitate the access. The annotation and the classification are the paradigms used for text categorization

1.1.2.1.   Text annotation

Metadata is the data that provide a description about other data. The association of this metadata is what we call annotation. Annotation permits to define what the subject of the text or portion of text is. Where it may informs about document information like author name, title, keywords, etc or about the knowledge provided at the level of the textual content of document.

Accessing information in a machine and human-readable state is so important. The traditional process followed for the annotation is expensive and time consuming (Lauser and Hotho, 2003) ; in addition, it is not possible to get a coherent annotation in different

documents with the intervention of different human annotators. So, it is so important to rely on methods that may facilitate this task. Therefore, several researchers have proposed methodologies for augmenting documents with metadata (Ciravegna et al., 2002).The researchers have relied on automatic or semi-automatic methods for document annotation.

Before embarking on describing details of annotation, we will provide some definitions.

"*A note by way of explanation or comment added to a text or diagram*". Oxford Dictionary of English

Annotation is the process of augmenting a text or a portion of text (word, sentence, paragraph, etc.) with labels (Descles, 2006). The automatic assignment of metadata is based on the information extraction techniques (Cunningham et al., 2011). That is considered the key element for tools of annotation (Ciravegna et al.,2002).

The annotation differs from the **semantic annotation.** Wide range of definitions is provided in the literature, we present from them the following:

Cunningham et al. (2011) have defined Semantic annotation as "the process of attaching metadata tags and/or ontology classes to text segments, as an enabler for knowledge access and retrieval tools."

Another definition is provided by Brank et al. (2018): "*Semantic annotation* is the task of augmenting an unstructured *textual document* with *semantic* information, such as concepts from an ontology."

As we can see, these definitions show that the label assigned to the text or portion of text is related to a knowledge base that represent the explicit semantic of a domain. Semantic annotation expresses the task of assigning meaning or semantic descriptions for the text.

What is the difference between the annotation and the semantic annotation?

The annotation is the process of affectation of metadata or key words to the corresponding text or other content. While the semantic annotation (also called semantic tagging or semantic enrichment) represents the process of assignment of labels or information that are

linked to a knowledge base that describe their semantics (the semantic of each concept (entities) and the concepts with each other).

1.1.2.2.   Text Classification

The massive data in the web that carry information involves new techniques like machine learning (ML) approach. Nowadays, Text Classification has become one of the key methods to organize and manage documents. Text classification (also referred as document classification or topic spotting) has made a considerable evolution in the AI technology and many other fields. The author Sebastiani (2002) defines text classification as "the activity of labeling natural language texts with thematic categories from a predefined set".

There are two main approaches in machine learning for text classification: supervised learning and unsupervised learning.

The **supervised classification** approach depends on the patterns that found in the training data in order to predict patterns in a new data where the classifier learns from the labeled examples in the training data. This approach is useful when the categories or the classes are well defined. So, it represents the process of automatic classifying a text or document into predefined categories or classes.   While in the **unsupervised classification** there is no indication that may define the probable categories hence, this means that there is not predefined categories to rely on in the classification task, here the clustering task are applied in order to detect the natural grouping of documents.

Text classification is carried out by employing ML paradigms. The ML provides methods and algorithms for classifying information automatically. There are different types of text classification techniques, the most popular algorithms include Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Neural Networks (NN), Classification Trees (CT), Logistic Regression (LR), Random Forest (RF), and Maximum Entropy (ME), Decision Trees (DT).

Document classification process includes mainly the following steps: document representation, feature selection, feature extraction, and performance evaluation.

In order to apply machine learning techniques the textual data must be transformed into vectors of numbers. **Document representation** is the task of representing the textual content in the appropriate form that is machine-understandable. There are several methods

for document representation such as **Bag-Of-Words**, **N-gram**, **Word2Vec**, etc. For decades, **Bag-Of-Words** (BoW) is the most commonly used document representations. Bag of Words approach has achieved good results in several tasks ((Wang and Manning, 2012); (Wu et al., 2010)). It is the simplest representation used for NLP. The text is represented as a collection of words that occur at least once. BoW considers each word as a feature where the words are independent of one another. The approach assumes that there is no semantic relationship that interlinks the words with each other. N-gram approach is an extension to BoW method. While Word2Vec method is a deep-learning-based distributed representation which is widely used recently.

BoW is considered the most commonly used document representations because it is the simplest representation. In BoW representation they assume that there is no semantic relationship that interlinks the words with each other. The challenge faced with BoW representation   is that they codify the words independently of their context. This means that the meaning of the words is not taken in consideration. However the context is so important because "*The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.*" (Firth, 1935)

Another challenge is with **Stop Words Removal**. The latter is an important step that is carried out in the preprocessing task. Stop words are the words that occurred frequently in the text such as Conjunctions, pronouns and prepositions, etc. (e.g. the, not, and). These words are considered as irrelevant information that doesn't add a significant value to the processed text. Therefore, stop words are usually excluded from text in the preprocessing task for the aim of improving processing time. However, eliminating stop words may lead to the modification of the context, and to the changing of the meaning of the given text.

1.1.2.3.   Methods of text categorization

IE techniques fall in two basic categories: linguistically oriented rule-based approaches and machine learning approaches (Cunningham et al., 2011).

1) Machine Learning- based approach

Machine Learning (ML) is a branch of Artificial Intelligence. ML is application that allows the development of algorithms which learn and improve their performance from experience. Nowadays, the ML method has gained much more interest, because it can be achieved more quickly. A great number of studies are carried out using ML methods.

"Machine learning, however, suffers from three big issues, namely: 1. Dependency: it requires (a lot of) training data and is domain-dependent. 2. Consistency: different training or tweaking leads to different results. 3. Transparency: the reasoning process is uninterpretable (blackbox algorithms)" (Cambria et al., 2017). In other words, concerning the dependency, in general, the automatic paradigms can be useful when a huge amount of data is used for the training and testing task (Cunningham et al., 2011). However, the task of manually annotating a corpus comprising thousands of documents by several specialists remains a tedious. In addition, it is so difficult to collect a considerable number of data if the corpus is not available. Concerning the consistency, the classification results depend on the training data. Concerning the transparency, the validity is one of the issues because the ML algorithms are considered as a "black box" hence it is too difficult to correct the errors that result from the classification process.

2) Rule-based approach

The classical approach used for identifying information is rule-based methods. Rule-based method is less popular, because it requires a long time and involves the intervention of domain expert. In the literature several studies have been carried out using rule-based methods.

Rule-based approaches aim for the definition of the corresponding information or knowledge using rules that depend on linguistic patterns collected by humans.

Constructing the adequate rules of extraction considered hard and tedious task, where, even if the approach has achieved high performance, the constructed rules remains domain dependent (Jiang, 2012). So, the disadvantage of rule-based methods is their dependency to a particular domain.

Generally, rule based approaches can be applied in these cases:

- "High precision system is needed" (Cunningham et al., 2011).
- There is not enough amount of data in order to use it for a machine learning training task (Cunningham et al., 2011).
- The dataset doesn't exist.
- The dataset needs a lot of preprocessing.

1.1.3. Ontology

The semantic representation of legal domain knowledge allows the resolution of problems related to its comprehension and management. The ontology has been widely used in different fields such as law, biology, economy, business, medicine, etc. ontology is considered as a polar in the Semantic Web Technologies. Where, it is a method used to capture knowledge within a particular domain (Cardellino et al., 2016), (Studer et al., 1992), to provide machine-readable of information and permits to implement the semantic into the human–machine communication. The ontology defined by Gruber (1993) as "a formal and explicit specification of a shared conceptualization". Ontology provides a formal and common conceptualization of the domain of knowledge. Also it enables the knowledge to be sharable and reusable. It defines, with different levels of formality, the meaning of the terms and the relations between them. Due to their explicit specification, the ontologies are widely used to model human knowledge (Buey et al., 2016). Where, the semantic web standards are used as mechanism to represent the ontologies.

1.2. Law text

1.2.1. Definition of Law

Wide range of definitions of law is provided in the literature. In the following we present some of them:

The definition provided by Oxford dictionary for the law is:

"The system of rules which a particular country or community recognizes as regulating the actions of its members and which it may enforce by the imposition of penalties."

The law is defined by تناغو (1984) as follow:

"القانون هو مجموعة القواعد العامة الجبرية، التي تصدر عن إرادة الدولة، وتنظم سلوك الأشخاص الخاضعين لهذه
الدولة أو الداخلين في تكوينها."

"The law is set of general and binding rules, issued by the will of the State, and governing the conduct of persons subject to that state or entrants in their composition."

The characteristic that distinguishes the law from other rules is that it is: "general" for all individuals. In addition, it is "mandatory" means that the individuals must do it. From this definition we realize also that the law is " العلاقات الاجتماعية أو الروابط تنظيم تستهدف ,اجتماعية قاعدة "; الأفراد بين it aims for organizing the relations between the different individuals.

The author (سليمان, 2015) defines the law as :

" مجموعة القواعد العامة التي تنظم سلوك الأفراد في المجتمع، والتي تكفل الدولة احترامها بالقوة عند الاقتضاء عن
طريق توقيع جزاءٍ على مَنْ يخالفها "

"The set of general rules governing the conduct of individuals in society, which the State guarantees to be respected by force when necessary by means of a punishment for those who violate it"

From these definitions we conclude that the legal texts are binding rules of conduct. These rules regulate and impose a set of constraints on the actions of individuals. And we realize that individuals are not excused because of their ignorance of law.

In general, the word "law" has two meanings, the term "law" is used for each rule, regardless of its source, and the term "law" is used also to indicate legislations which are " ما أمراً لتنظم التشريعية السلطة تضعها التي الملزمة العامة القواعد مجموعة "

1.2.2. Types of Law language

In legal domain there are different languages of law which are " التشريع لغة", "القضاء لغة", " لغة الاتفاقيات" and "الاكاديمية اللغة". "التشريع لغة" is the most prominent category, it concerns the legal documents that are issued by " الدستور", "التشريعات" and "اللوائح".

1.2.3. Legislative drafting

The legislative drafting represents the legal rules that are written rules (قواعد مكتوبة) and issued by the legislature (**سلطة مختصة**); these features permit to distinguish the legislation

from other rules. There are three main levels of legislative drafting which are "الدستور",
"التشريعات" and "اللوائح":

- "الدستور" represents "التشريع الأساسي" that is defined by "السلطة التأسيسية".
- "التشريعات" represents "التشريع العادي" which is issued by "السلطة التشريعية". Examples of "التشريع العادي" are (law and ordinance).
- "اللوائح" represents the subordinate legislation "التشريع الفرعي" which is issued by "السلطة التنفيذية". Examples of the subordinate legislations are (executive regulations " لوائح ", regulatory regulations "اللوائح التنظيمية" and policing regulations "اللوائح التنفيذية الضبط"

Each type of legislation has an impact and strength; these legislations follow the strategy of the pyramid (e.x Subordination in legislation hierarchy); "الدستور" ranked first followed by "التشريعات" and then "اللوائح" ranked the last. "الدستور" is considered the most strength law and it must not be violated by any other law.

1.2.4. Source of legal documents

The official portal of Algeria is named **General Secretariat of Government** (GSG) in Arabic language is "الأمانة العامة للحكومة", the GSG is the primary source of regulation on the Web. The legal documents in the journal are published in two versions, Arabic and French; where, the first is the official language. Researching of law texts is achieved by accessing to the GSG. The latter publishes all the Algerian legal texts (laws, decrees, ordinances..., etc.). And it contains legislation about all sectors (e.g. education, health, commerce, economy, security, etc.).

As pointed in the GSG homepage, the GSG offers the access in order to: "- Refer to the constitution, - Get access directly to the published issues of official journals, - Make a theme research of all texts published in the official journal, - Refer to other publications elaborated by the General Secretariat of the Government"[1]. The fig.1.2 presents the GSG homepage.

---

[1] https://www.joradp.dz/HAR/Index.htm 15/11/2018

Fig.1.2 The GSG homepage

The system offers the possibility of accessing to law text by filling a set of defined fields, it allows access by 1) Sector (Set of ministries: Education; Higher Education and Scientific Research; Health, Population and Hospital Reform; Foreign Affairs..., etc.). 2) Ministry. 3) Nature of the text (e.g. Law, Ordinance, Decree). 4) Journal number. 5) Date of publication. 6) Number of the text. 7) Date of signature. 8) And finally, by a combination of keywords. The fig.1.3 shows the different fields required in order to retrieving law texts, which concern the legal texts written in Arabic language.

The documents are offered in a Portable Document Format (PDF) and in a hardcopy. Where, the PDF format is published in protected format in order to provide a security.

Fig.1.3 the GSG field for legal document retrieving

1.2.5. Legal documents structure

Legal documents (الوثائق القانونية) are written in different forms, where for example the contract "العقود" documents have a form different than legislative documents. All the legislative documents have the same structure. The Fig.1.4 illustrates an example of a legislative text in Arabic language

أوامـر

أمر رقم 06-03 مؤرّخ في 19 جمادى الثانية عام 1427 الموافق 15 يوليو سنة 2006، يتضمّن القانون الأساسي العام للوظيفة العمومية.

‒ وبمقتضى القانون رقم 90-14 المؤرخ في 9 ذي القعدة عام 1410 الموافق 2 يونيو سنة 1990 والمتعلق بكيفيات ممارسة الحق النقابي، المعدل والمتمم،

‒ إن رئيس الجمهورية،

‒ وبمقتضى الأمر رقم 97-03 المؤرخ في 2 رمضان عام 1417 الموافق 11 يناير سنة 1997 الذي يحدد المدة القانونية للعمل،

‒ بناء على الدستور، لا سيّما المواد 51 و122 ‒ 26 و124 منه،

‒ وبمقتضى القانون رقم 99-07 المؤرخ في 19 ذي الحجة عام 1419 الموافق 5 أبريل سنة 1999 والمتعلق بالمجاهد والشهيد،

‒ وبمقتضى الأمر رقم 74-103 المؤرخ في أول ذي القعدة عام 1394 الموافق 15 نوفمبر سنة 1974 والمتضمن قانون الخدمة الوطنية، المعدل والمتمم،

‒ وبعد الاستماع إلى مجلس الوزراء،

‒ وبمقتضى الأمر رقم 76-111 المؤرخ في 17 ذي الحجة عام 1396 الموافق 9 ديسمبر سنة 1976 والمتضمن مهام الاحتياط وتنظيمه،

يصدر الأمر الآتي نصّه :

‒ وبمقتضى القانون رقم 78-12 المؤرخ في أول رمضان عام 1358 الموافق 5 غشت سنة 1978 والمتعلق بالقانون الأساسي العام للعامل، لا سيما المواد من 180 إلى 186 منه،

الباب الأول

أحكام عامة

المادة الأولى : يتضمّن هذا الأمر القانون الأساسي العام للوظيفة العمومية.

‒ وبمقتضى القانون رقم 83-11 المؤرخ في 21 رمضان عام 1403 الموافق 2 يوليو سنة 1983 والمتعلق بالتأمينات الاجتماعية، المعدل والمتمم،

يحدد هذا الأمر القواعد القانونية الأساسية المطبقة على الموظفين والضمانات الأساسية الممنوحة لهم في إطار تأدية مهامهم في خدمة الدولة.

‒ وبمقتضى القانون رقم 83-12 المؤرخ في 21 رمضان عام 1403 الموافق 2 يوليو سنة 1983 والمتعلق بالتقاعد، المعدل والمتمم،

الفصل الأول

مجال التطبيق

‒ وبمقتضى القانون رقم 83-13 المؤرخ في 21 رمضان عام 1403 الموافق 2 يوليو سنة 1983 والمتعلق بحوادث العمل والأمراض المهنية، المعدل والمتمم،

المادة 2 : يطبق هذا القانون الأساسي على الموظفين الذين يمارسون نشاطهم في المؤسسات والإدارات العمومية.

‒ وبمقتضى القانون رقم 83-14 المؤرخ في 21 رمضان عام 1403 الموافق 2 يوليو سنة 1983 والمتعلق بالتزامات المكلفين في مجال الضمان الاجتماعي، المعدل والمتمم،

يقصد بالمؤسسات و الإدارات العمومية، المؤسسات العمومية، والإدارات المركزية في الدولة والمصالح غير الممركزة التابعة لها والجماعات الإقليمية والمؤسسات العمومية ذات الطابع الإداري والمؤسسات العمومية ذات الطابع العلمي والثقافي والمهني والمؤسسات العمومية ذات الطابع العلمي والتكنولوجي وكل مؤسسة عمومية يمكن أن تخضع مستخدموها لأحكام هذا القانون الأساسي.

‒ وبمقتضى القانون رقم 88-07 المؤرخ في 7 جمادى الثانية عام 1408 الموافق 26 يناير سنة 1988 والمتعلق بالوقاية الصحية والأمن وطب العمل،

‒ وبمقتضى القانون رقم 90-02 المؤرخ في 10 رجب عام 1410 الموافق 6 فبراير سنة 1990 والمتعلق بالوقاية من النزاعات الجماعية في العمل وتسويتها وممارسة حق الإضراب، المعدل والمتمم،

لا يخضع لأحكام هذا الأمر القضاة والمستخدمون العسكريون والمدنيون للدفاع الوطني ومستخدمو البرلمان.

‒ وبمقتضى القانون رقم 90-11 المؤرخ في 26 رمضان عام 1410 الموافق 21 أبريل سنة 1990 والمتعلق بعلاقات العمل، المعدل والمتمم، لا سيما المادة 3 منه،

Fig.1.4 Excerpt of legal document content

Legal documents have some characteristics that distinguish them from other documents; they are characterized by what we call "accuracy", especially at the level of the structure. Algerian legal documents have a specific structure, and may include a set of legislative texts (law, decree, ordinance, etc.). These legal documents take the form illustrated in fig.1.5.
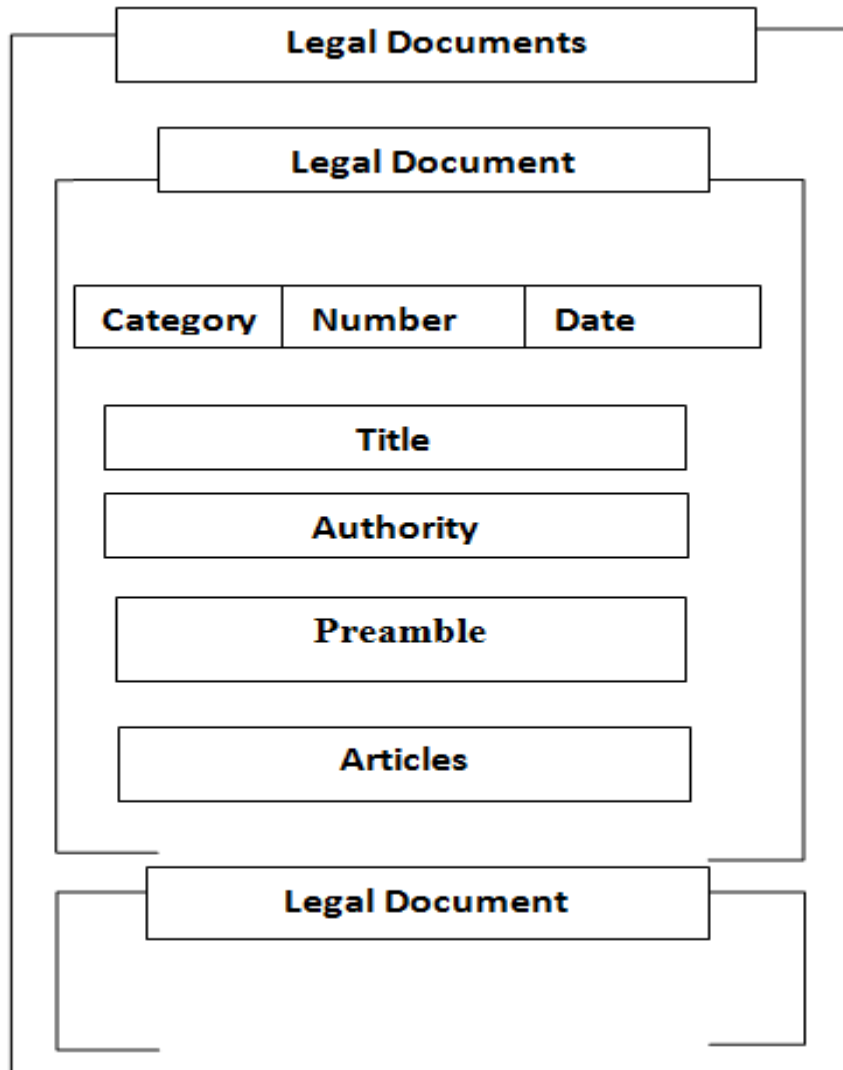
Fig.1.5 Legal document hierarchy.

This legislative text has a specific hierarchy, in general we can cluster their content in three parts, the first part defines the **metadata of the legislative document** (type of text, number of text, date, etc.); the second part concerns the **preamble** "الديباجة" and the third part is about **partitions of legal document** (e.g. the articles and chapters). In the following, we present a detailed description of the legislation structure:

- **Metadata of the legislative document:**

  – Category is the type of the legal document which could be: مرسوم, أمر,قرار, etc.

  – Number is the number of the document which is composed by two numbers separated by (-), the first one indicates the year of the construction and the second number specifies the number of document; for example (15-112).

  – Date corresponds to the construction day of the legal document, it is a succession of two dates :

    ➢ التاريخ-الهجري Hijric date.

    ➢ التاريخ-الميلادي Gregorian date.

  – Title represents the heading of the document, where it defines the document context.

  – Authority corresponds to the authority that constitutes or creates the document.

- Preamble

  – Preamble is the introduction of the legal document. It expresses for what the Authority was based on during the drafting phase of the legal document.

- **Partitions of legal document :**

      Algerian legal documents are a hierarchical, as we have mentioned above. The document may include a set of legislative texts (law, decree, ordinance, etc.); each one contains **chapters**, the latter can be divided into **sections**, which can have several **subsections** that contain set of articles. Each one of the parts: chapters, sections, and subsections have a title that describes their content. These parts are found in the case of the content of the legal text is big. In some legal documents, we can find a set of articles directly without chapters. Each article has a number. The article can be constituted of more than one legal rule. The parts are illustrated in Fig.1.6.
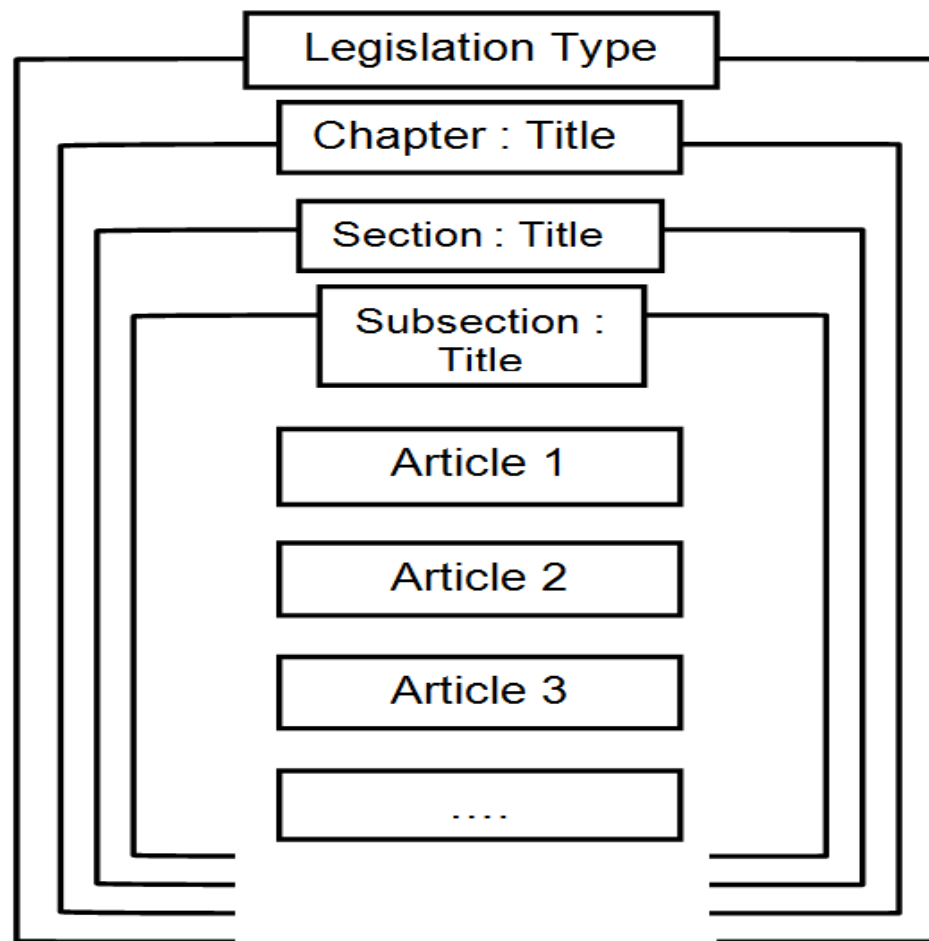
Fig.1.6 Arabic legal document structure.

1.2.6. Imperative and suppletive rules

As illustrated in Fig.1.7 the rules are divided into two main categories: imperative rules (**قواعد آمرة**)and suppletive rules(**قواعد مكملة**) .
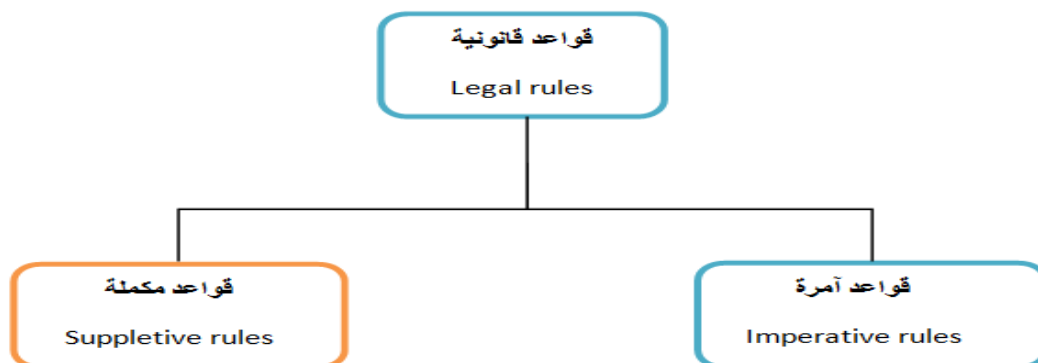
Fig.1.7 the two main categories of legal rules

The Arabic definition of the imperative rules is:

"هي القواعد التي لا يجوز للإفراد الاتفاق على خلاف ما جاء فيها من أحكام، وكل اتفاق على خلافها يكون باطلاً"
(تناغو، 1984)

"The imperative rules are the ones on which individuals shall not adopt the contrary"

The Arabic definition for the suppletive rules is:

"هي القواعد التي يجوز الاتفاق على خلاف ما جاء فيها من أحكام" (تناغو، 1984)

"The suppletive rules are the ones on which individuals may adopt the contrary".

All law rules are mandatory; however, they differ from one to the other, the imperative rule implies a penalty for the violation (e.g. "وعلى الموظف أن يسهر على حماية الوثائق الإدارية suppletive rule gives the parties the right to choose; where the sanction is "). And the [2] أمنها applied in case of violation, after the agreement of the parties on a particular issue (**تناغو،** يمكن للمستخدمين الذين يشغلون مناصب عمل معرضة بصفة دائمة للإرهاق و الأخطار أو تنطوي " e.g.) **1984** على ضغوط بدنية أو عصبية، للإستفادة في إطار ما يسمح به القانون، بتخفيض المدة الاسبوعية للعمل"[3].

1.2.7. Legal Rules categories

Legal texts play an important role, and are considered a raw material for organizations where each organization must comply with the governmental regulations. The mastery of these texts plays crucially in the effective management of organization, in particular the public ones.

In the legal system, there are various categories of legal rules. Legislative expressions are not used to express only the rules of type provisions, where besides of the provisions (that indicate what is obliged and what is prohibited), various types are existed like the definition, modification rules, the constitutive rules, etc.

In Arabic legal language, the authors صبره, (2008) have pointed out that the law is composed by: 1/ substantive articles "**المواد الموضوعية**" that are the provision rules, 2/ articles of definition "**الأحكام العامة**" which concerns: "مواد التعريف", 3/ Articles that relate to and 4/ articles of "**مواد الإصدار**" that are the articles that "تقرير مبادئ و سياسات واجبة الإتباع "

---

[2] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

[3] The internal regulations of Kasdi Merbah Ouargla University 2007.

relate to "**الأحكام الفنية**" :which contains: "" "أوامر تشريعية للمخاطبين به" " و مواد الالغاءات

.etc "التعديلات", "الأمر بالعمل بالقانون", "تفويض الصلاحيات التشريعية", "الأمر بنشر القانون و بدء العمل به"

 In general, the previous research in artificial intelligence and law has classified legal rules into the following categories: provision, constitutive and modification. (Biagioli and Gross (2008) have worked on two divisions that are the **rules** which composed by **constitutive** and **regulative provisions** and **rules on rules** which contains modificatory provisions.

In the following we are providing a description of some of legal rules types:

Prescriptions or provisions

The provisions are also called in the literature prescription rules. "The regulative provisions concern deontic concepts" (Biagioli and Grossi, 2008). "In legal domain, prescriptive rules constrain the activities by making actions obligatory, permitted or prohibition" (Hashmi, 2015). "An obligation is a situation or a course of action which legally binds some entity, and if it is not achieved results in violation" (Hashmi, 2015) in an other word, an obligation is a duty that the entities must do and comply with, and the non-compliance of the bound may lead to a penalty. In contrast, prohibition is the one that the entities must not do. Hashmi (2015) expresses a prohibition as "a situation or a course action that must not be performed, a violation is triggered otherwise". Permission is the one that the entities can choose to do or to adopt the contrary. The authors Hashmi (2015) say that: "something is permitted if neither an obligation nor a prohibition holds". Some researchers consider both of obligation and prohibition as obligations; where the author in (Franssen, 2007) describes obligations as "norm sentences that state that something must or must not be done".

Constitutive

"The constitutive provisions lay out the components of the relevant pieces of legislation by introducing new types of entities, defining new terms or procedures, creating new institutional bodies, and attributing powers" (Biagioli and Grossi, 2008). "Constitutive rules contain the conditions that have to be met in a particular context for a status to be instantiated. Regulative rules have turned out to be rules that link those conditions directly to normative attributes. The crucial difference is that regulative rules do not introduce new

status terms. Thus, the notion of a status rule was used to reveal that the distinction between constitutive and regulative rules is indeed a linguistic one." (Hindriks, 2009)

Here, rules of the type definition are included in the constitutive rules. In the definition rules we find the notion of the terms which are used in the legal document (Law, 2008).

Modificatory provisions

modificatory provisions are "rules on rules, which manage the dynamics of laws" (Biagioli and Grossi, 2008). Annotating modificatory provisions is an important task because that rules can have an effect on the whole legal system.

1.2.8. Provision Arguments

Provision arguments are the elements that constitute the legal rule. Where each provision should state (Gordon et al., 2009): "– who (the norm-subjects); – does what (the action-theme); – in what circumstances (the condition of application) and – the nature of their guidance (the mode)".
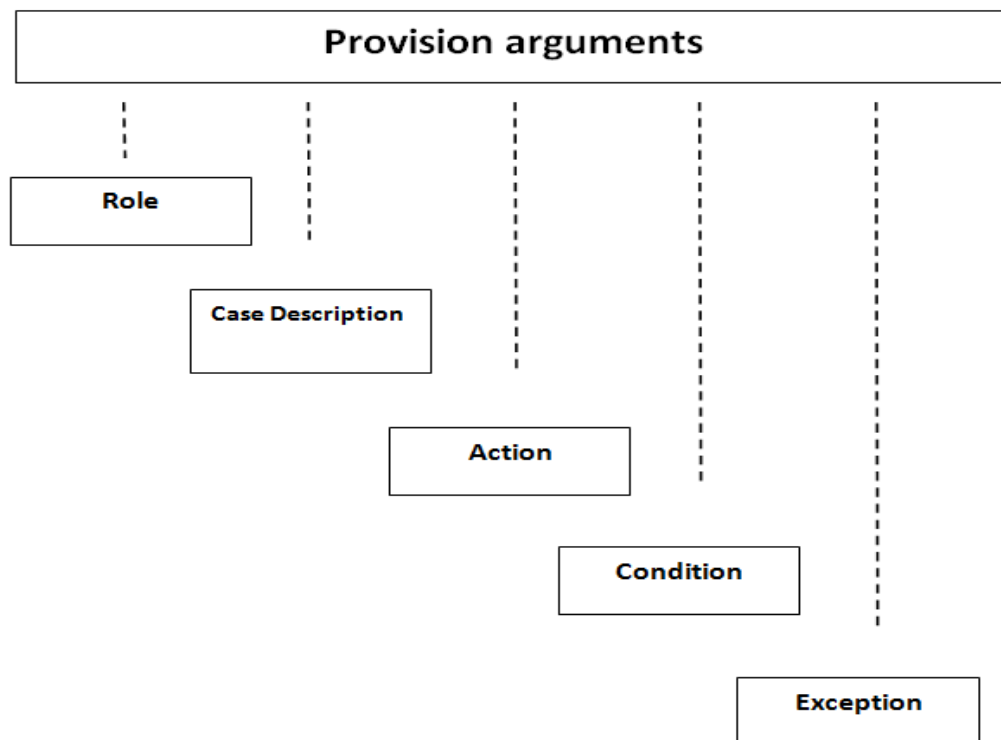
Fig.1.8 Arguments of provision rule.

Legislative drafting is characterized by the existence of the following essential elements (arguments): - role, - action, - case description, - condition, and – exception (دودين, 2009). These important elements are illustrated in Fig.1.8 in the following we provide a description for each element.

1.2.8.1 Role

The legal actor in the legislative sentence differs on the actor in ordinary sentence, where the actor in the legislative sentence is the person who is addressed to the judgment (صبره, 2008). The legal subject or the legal actor is " يفرض أو سلطة أو امتيازا أو حقا له يخول الذي الشخص

يحدد الفاعل القانوني نطاق القانون أولئك الذين ينطبق عليهم النص " Where. (السباعي) "عليه التزاما أو مسؤولية القانوني" : the legal subject defines law domain as on those to whom the legal provision is applicable on (صبره, 2008).

 "Prescriptions are associated with roles rather than individuals" (Boella et al., 2013). The role is the legal subject; it indicates the address (e.g., employer) to whom the rule is applied (Jiang et al., 2012). The role expresses the actor or the organization; the latter corresponds to a group of actors or organizations represented as one (Hoekstra et al., 2007).

1.2.8.2 Action

The action is the legal action in the provision. The legal action in the legislative drafting determines what is required from the legal actor in the legal provision (صبره, 2008). So, it determines the action assigned to the Role (Hoekstra et al., 2007). السباعي provides the following definition :

"يقصد بالفعل القانوني في الصياغة التشريعية بأنه ذلك الجزء من الجملة الذي يعبر عن حق أو امتياز أو سلطة أو التزام أو مسؤولية تخول أو تفرض على الفاعل القانوني"

1.2.8.3 Case description

Case description argument is an important element in the provision. Where the clear expression of that element enable to provide the exact meaning of the text (صبره, 2008)

 Case description defines the situation in which the rule is applied. Because generally " نادرا Legal" : "ما يسري الحكم القانوني على جميع الاوضاع و الظروف، و إنما يسري عادة على حالة معينة

provision applies to all situations and circumstances, but habituuelly applies to a specific situation" (صبره, 2008)

1.2.8.4 Condition

Condition specifies the particular circumstance required for the application of the provision. صبره, (2008) defines the condition as " الوضع الذي يجب تحققه قبل أن يصبح الحكم القانوني نافذ" : "The situation that must be accomplished before the legal provision becomes applicable"

1.2.8.5 Exception

Finally, exception indicates elements that are excluded from the application of the provision. Where صبره, (2008) defines the condition as " أو حذف عنصر ما من عدد أو استبعاد "وصف محدد و قد يكون المستثنى شخصا أو شيئا أو وضعا معينا": "Exclude or delete an item from a specific amount or specific description. An excluded person may be a person or a certain thing or situation."

The action and the role are mandatory in the permission and obligation, while, the remaining arguments: case description, condition, and exceptions may or may not exist in the provision rule. The arguments action and role are the main components of each legal sentence. The elements can be appeared explicitly in the text or can be existed implicitly (Biagioli and Grossi, 2008)

1.3. Overview of literature

1.3.1. Text categorization

Several studies have been conducted to address problems related to the manual annotation of legal texts (Lesmo et al., 2013). Previous studies have shown that language processing is an essential step in identifying the semantic information contained in the legal text (Francesconi and Passerini, 2007). Where, the authors Zeni et al. (2016) and Zeni et al. (2017) have pointed out that "analyzing and annotating legal documents in prescriptive natural language, still an open problem for research in the field".

The studies opt for the automatic or semi-automatic annotation based on machine learning approach or rule-based approach. In previous research, the treatment of the information has

been studied at different levels: title level, paragraph level, or sentence level. In our research, we have chosen to tackle the sentence level. For the identification of information, the previous studies have focused on classification and/or annotation. Some studies have adopted the technique of annotation through the use of the rule-based approach. In others, the classification technique is used for the categorization of the provisions, however, in the identification of the arguments, they depend on the technique of annotation. The difference between these two techniques is in the absence of the trace at the source with the classification technique. Furthermore, De Maat et al. (2010) have made a comparison between ML approaches and knowledge based approaches for the classification of legal texts that concern the legislation of Dutch, and he concluded that the pattern based approach is better than ML classifiers because the latter generalized poorly on new legislation (Neill et al. 2017). However, the authors in (De Maat et al. 2010) have argued that the results can be improved with the use of a bigger training set.

SALEM system (Semantic Annotation for Legal Management) (Bartolini et al. (2004); Biagioli et al. (2005); Soria et al. (2007); Spinosa et al. (2009)) has been developed for the semantic annotation of Italian Law texts. SALEM is based on linguistic characteristics of law paragraphs to classify provisions using Pattern matching technique, and rule-based approach for the annotation of the arguments of provision. The pattern matching technique is not sufficient in our case with the different indicators, as mentioned above, the context is essential to distinguish indicators. SALEM was interested in the classification of provisions while in our research we were interested in the identification of the provisions. Concerning the argument identification, SALEM used a particular version of the ILC finite–state compiler of grammars for dependency syntactic analysis carried out by Bartolini et al. (2002) which consists of a set of rules for the identification of fundamental syntactic dependencies and a set of rules for semantic annotation (Bartolini et al., 2004). However, we have relied on a set of rules that depend on the CE method. The pattern matching technique has limits in the case of the sentences are length. The SALEM output is a single law paragraph tagged in XML format. SALEM has some commonality with our work; however, it was interested in the classification of a set of legal rules, while in our research, we are interested in the annotation of the provision types: obligation, permission and prohibition.

kiyavitskaya et al. (2008) and Zeni et al. (2015) worked for the development of a tool called Gaius T, which is an extension of the method proposed by Kiyavitskaya et al. (2006) for the extraction of the provisions: right and obligation from the regulatory text. The system Gaius T aims at the semantic annotation of right, obligation, and constraint in the regulatory text that concerns two different languages English and Italian. In the development of the tool, the study based on the framework Cerno which was proposed by Kiyavitskaya et al. (2006), which they constructed the framework Cerno for the semi-automatic semantic annotation. The developed tool has been examined by both the U.S. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, and the Italian accessibility law. Zeni et al. (2016) and Zeni et al. (2017) have then developed the GaiusT 2.0 framework as an extension of the framework Gaius T. The authors have implemented the framework as a web-based system that extracts legal concepts semi-automatically. The works presented by authors are similar to ours in that they constructed a set of heuristic rules that help to identify rights or obligations.

Breaux et al. (2006) have also relied on a set of heuristic rules that help to identify rights or obligations.

The authors in (Bui et al., 2014) propose a system for the research of law text concerning Vietnamese Enterprise legal documents for citizens. The law text processed token from Vietnam Ministry of Justice database. Two phases followed for the annotation of the law text: the logical structure annotation and semantic annotation. In the logical structure annotation, the researchers have relied on the division of law article into the following parts: assumption, provision, and sanction. In this step, the authors have adopted a rule-based approach, relying on regular expressions and linguistic features using GATE framework (General Architecture for Text Engineering); GATE framework is an open source written in java for NLP tasks (Cunningham, 2002). In the semantic annotation phase, the authors based on the LKIF core ontology that was modified to cover the Vietnamese legal concepts, in which they classifying the selected terms in 5 concepts using SVM classifier and populating in. after that the ontology instances also are annotated in the document for the aim of effecting the retrieving in the document.

Wyner and Peters (2011) was interested in the annotation of semantic information contained in documents that concern the US regulations; they proposed a rule-based

approach for the annotation of conditional rules and deontic rules permission and obligation, and the annotation of the components: the antecedents, consequences, agents, themes, actions, and exceptions. In the proposed approach they adopted the environment Gate with the adaptation and development of modules to adapt to the requirements of English regulations. This approach requires pre-treatment phase that apply tokenization, NER, POS tagger; however, in Arabic language, these techniques have not returned satisfied results.

Boella et al. (2016) have been interested in classifying documents in different topics, however, we focused on the categorization of sentences.

Hashmi (2015) have used the structures of (If…then) for the automatic extraction of norms from regulatory texts. They were interested also in the extraction of the conditions.

Previous studies such as SALEM (Bartolini et al.(2004); Biagioli et al.(2005); Soria et al.(2007); Spinosa et al.(2009)); Kiyavitskaya et al. (2008); Wyner and Peters (2011) and Zeni et al. (2015); etc. have used either the classification or the annotation for the specification of the type of provision. Where, in both, they depended on the indicators, however depending on the context; the indicator can take a different meaning. With the particularities of Arabic legal texts, it is not possible to rely only on the words indicating the existence of the provisions in the sentence. Accordingly, it is not possible to depend on the tools and methods proposed by the above mentioned researchers.

The authors Neill et al. (2017) were also interested in the identification of relevant deontic modalities (obligations, prohibitions and permissions). They have used financial law domain as a use case. As opposite to the above mentioned works, the authors Neill et al. (2017) have token in consideration the contextual information in the input representation in order to classify the provision to a particular class. They have used a deep learning architecture in order to classify deontic modalities. Where, they have achieved an accuracy of 82.33 % on a held-out test set.

Few studies have focused on the processing of Arabic legal language, from which we can mention the work proposed by Rammal et al. (2015), which they proposed to use Local Grammar for the extraction of Key words from titles of Lebanese official journals. However, in our research, we focused on the provisions rather than document titles.

Berrazega et al. (2016) have used the CE methodology for the specification of categories of Arabic normative provisions, however; the work presented did not explain the method in which they were based on to distinguish the provision types. Where in our research we defined a set of classes according to the indicator occurred, to categorize to which type the sentence belongs. Besides of the identification of provision types, we are interested in the specification of the arguments role and action.

1.3.2. **Ontology**

Ontologies are one of the important fields, where it is used in several domains such natural language processing, artificial intelligence, knowledge representation…etc. The main purpose of ontology is providing a unified understanding. Ontology represents the knowledge of a specific domain in a formal and explicit way.

For the management of legal domain we represent its knowledge. Several researchers have used ontology to represent legal knowledge. Different legal ontologies have been proposed in the literature with different purposes, we mention from them the works proposed in (Hoekstra et al., 2007), (Casellas, 2011), (Boella et al., 2016) and (Athan et al., 2015), etc. The LKIF core ontology (Hoekstra et al., 2007) represents the general concepts in the legal domain where it is the most popular ontology in the legal domain.

In Arabic language few works have focused on the legal domain. The authors in (Dhouib and gargouri, 2015) proposed a legal ontology for Jurisprudence Decision written in Arabic language which covers the semantic content of jurisprudence decisions. The CrimAr ontology: A Criminal Arabic Ontology (Mezghanni and gargouri, 2017) authors have based on the criminal system in their domain application. In our research we have focused mainly on laws and regulations.

Zaidi et al. (2005) have proposed ontology for Arabic information retrieving that concerns Algerian law. However the authors have concentrated in the general hierarchy of the law as illustrated in fig.1.9. In our research we have focused on the structure and semantic content of the legal documents.
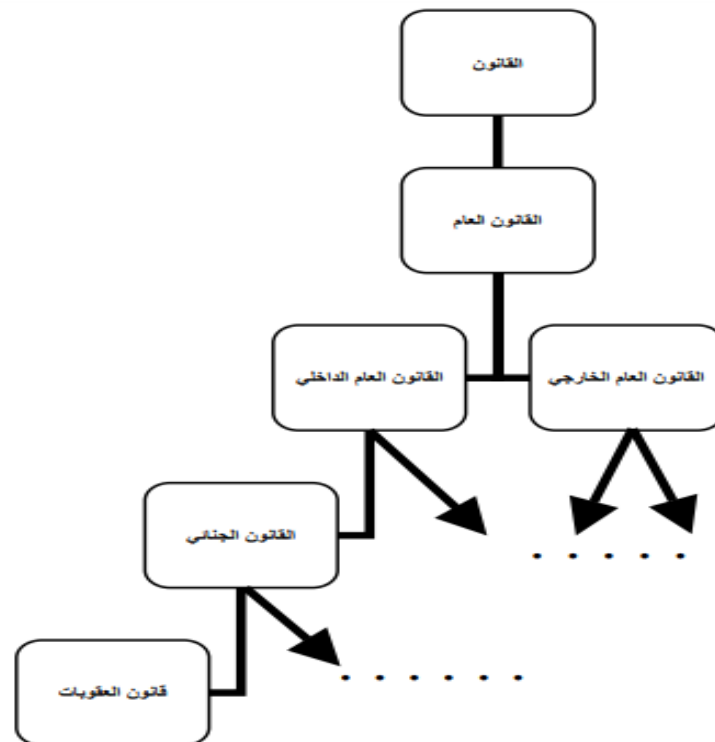
Fig.1. 9 The hierarchy of the concepts (Zaidi et al., 2005).

Accessing and retrieving legal documents is so important for most individuals, therefore we have to be interested in modelling Algerian legal texts. The Official portal of Algeria publishes all the legislation in two languages Arabic and French. The Algerian legal system offers the possibility to access law, but the method doesn't explore the semantic information provided at the rule level. And the set of the information which must be filled out in the system to find the corresponding document requires the familiarization of laws. However, this is not the case with all the users. Besides, because of the variety of legal sources and the dynamic character of law, the access to the relative information becomes a laboring task, it is so hard to retrieve information existed among a large number of legal documents.

We use NLP techniques in order to understand human language. The official language in Algeria is Arabic so we have to study the particularities of Arabic and the legal language in order to understand it. This point will be addressed in the next chapter.

Chapter 02

Arabic language and legal language nature challenges

2. Arabic language and legal language nature challenges

This chapter provides a description of Arabic language and its main characteristics, and then we present the nature of the legal language. Next we present some examples in order to illustrate the challenges of Arabic legal texts. After that we describe the method that we guess that it goes with the particularities of Arabic legal texts.

2.1. Arabic language
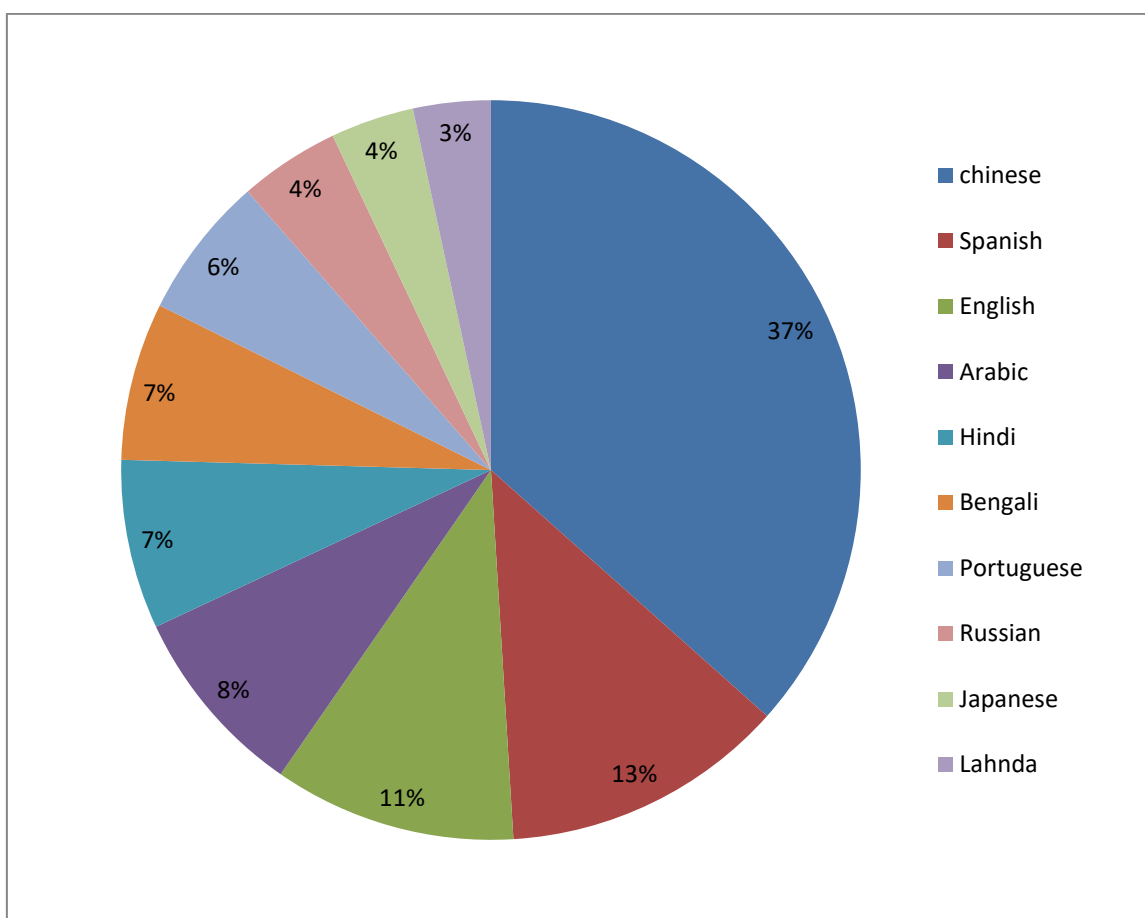
Fig.2. 1 The Top 10 spoken languages in the World (McCarthy, 2018)

In Algeria, legal texts are issued in two versions, Arabic and French; where, the first is the official language. Arabic is considered as Semitic language closely related to Hebrew and Aramaic. It ranks in fifth place of the most spoken languages in the world. Arabic is official Language in 26 countries and it is the

mother tongue of approximately 295 million speakers (McCarthy, 2018) as illustrated in Fig.2.1.

The Arabic alphabet consists of twenty eight alphabet letters; these letters are the consonants and the long vowels: (Aalif (ا), waw (و) and ya (ي)). The Fig.2.2 presents the Arabic alphabet. The letters can be extended to ninety by added shapes, marks, and vowels (Khorsheed, 2002).

alif	ا	za	ز	qaf	ق
ba	ب	sin	س	kaf	ك
ta	ت	shin	ش	lam	ل
tha	ث	sad	ص	mim	م
jim	ج	dad	ض	nun	ن
ha	ح	ta	ط	ha	ه
kha	خ	dha	ظ	waw	و
dal	د	ain	ع	ya	ي
dhal	ذ	ghain	غ		
ra	ر	fa	ف		

Fig.2. 2 Arabic alphabet[4]

The form of the letter differs depending on whether it occurs at the beginning, middle, end or independent as illustrated in Fig.2.3. In contrary to Latin-based alphabets, the Arabic writing is from right to left (Amin, 1998).

---

[4] https://www.metmuseum.org/learn/educators/curriculum-resources/art-of-the-islamic-world/unit-two/origins-and-characteristics-of-the-arabic-alphabet 25/02/2020
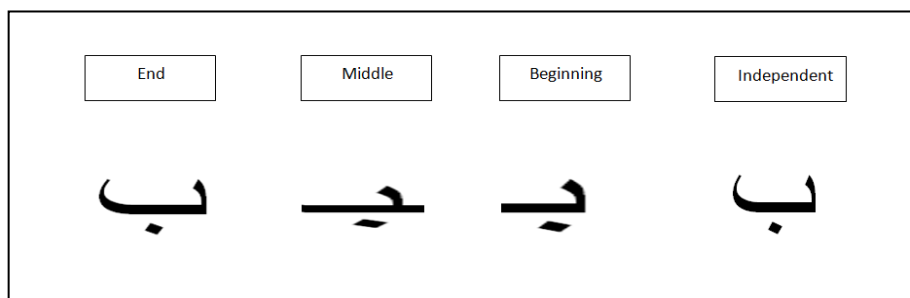
Fig.2. 3. Occurrence shape of the Arabic alphabet "ب".

Modern Standard Arabic is different from the classical Arabic, where the latter is the regular version of Arabic which is the language of the Holy Qur'an. The Modern Standard Arabic is written without short vowels (Boudelaa, 2010), and this is one of the main features that distinguish it, where it is used in the official documents, nooks, newspapers, etc.

Arabic is characterized by a rich and complex morphology (Farghaly and Shaalan, 2009). Therefore, Arabic language "presents significant challenges to many natural language processing (NLP) applications (Farghaly and Shaalan 2009)." (Zitouni, 2011).

2.2. Arabic language characteristics

Despite the efforts that are exerted in the field of automatic Arabic language processing, there are modest results. Several problems arise when dealing with the automatic processing of text. The difficulty of treatment lies in nature of the language treated (Harmanani et al., 2006). This particularity is illustrated in the Arabic language, in which the automatic processing of text is a difficult task due to their complex morphology (Al-Kharashi et Al-Sughaiyer, 2004; Alrahabi et al., 2006).

Here some features of Arabic language are mentioned:

1. The vocalization is highly essential in Arabic language (Al-Kharashi et Al-Sughaiyer, 2004), where the same word can express several meanings, according to the vocalization (Al-Kharashi et Al-Sughaiyer, 2004; Alrahabi et al., 2006; Bousmaha et al., 2013). Words are indicated by diacritics, or short vowels that are written above or below the letters, for example, the

non-vowel word "يسمح" may take a different meaning with the following diacritical marks:

- "yasmaho: يَسْمَحُ",
- "yosmaho: يُسْمَحُ",
- "yosmah: يُسْمَحْ",
- "yasmaha: يَسْمَحَ", etc.

Some words can take more than ten meanings with vocalization. However, recently vocalization is almost unused; in the Arabic writing, we find only some vowels used in some words. With experience and following the context, readers associate the corresponding diacritical marks during reading to get the correct meaning.

2. The agglutinative nature of Arabic language, for example, we take the word 'بإمكانهم': -which is resulted from the agglutination of the prefix "ب" which is a preposition, the stem "إمكان", and the suffix "هم" which is a possessive pronoun. The composition of the parts (prefix, stem, and suffix) can create a complex morphology. Fig.2.4 presents the agglutination in the word "بإمكانهم".
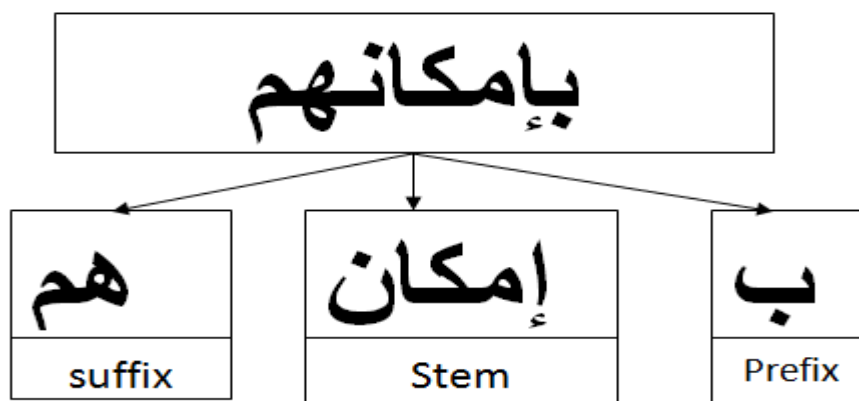
Fig.2. 4 Agglutination in the word "بإمكانهم".

3. Unlike other languages that depend on the capitalization feature such as the task of Named Entity Recognition, Arabic language does not use capitalization. Capitalization is essential in the applications such as classification, machine translation, information retrieval and clustering. In addition of the capitalization, Arabic language has not strict rules of

punctuation; and that makes the task of preprocessing Arabic texts so difficult; Because the punctuation play an important role in recognizing sentence boundaries.

4. The same sentence in Arabic can be expressed with different structures (El Kassas and Kahane, 2004); where the word in the sentence can take different orders, but the meaning remains the same. We mention for example:

- "يمكن الطالب أن يدرس".
- "الطالب يمكن أن يدرس".
- " يمكن أن يدرس الطالب".

5. Arabic language characterized also by the existence of the polysemy. Polysemy concerns the case in which a word or phrase can take several distinct meanings; word like "ذهب" can take the meaning of "Golden" or the meaning of "to go".

2.3. Legal language nature

Legal documents have some characteristics that distinguish them from other documents; they are characterized by what we call "accuracy", especially at the level of the structure. Algerian legal documents have a specific structure, and may include a set of legislative texts (law, decree, ordinance, etc.); each document can contains chapters, sections, subsections and articles. On each article, we can find more than one sentence.

The legal language is different from ordinary language; where the legal language uses an unambiguous syntax (Franssen, 2007); this is a contributory factor in the task of natural language processing, we mention as examples of the constraint imposed by the Arabic legal drafting, the use of the "simple imperfect tense". However, some particularities of legal language may imply more complexities such as:

- Ambiguity: legal language is difficult to be understood by non domain specialist (Otto and Antón, 2007);
- Contradictory: we can find contradiction between rules in the legal system;

41

- Dynamic character of law: the modification of provisions leads to make several versions and references. The detection of the cross references is challenging task, where several research have interested in handling the references or the cross-references existed between legal rules;

- Wordiness: the particularity of the use of long sentences in the legal text (Wyner and Peters, 2011); this feature leads to dealing with problems of parsing long sentences. Such as confronted for example by the technique of finite state automata, which gives weak performance when parsing long sentences. For example, the following legal rule contain 66 words;

"المادة 178: يمنع على كل موظف، مهما كانت وضعيته في السلم الإداري، أن يمتلك داخل التراب الوطني أو خارجه، مباشرة أو بواسطة شخص آخر، بأية صفة من الصفات، مصالح من طبيعتها أن تؤثر على استقلاليته أو تشكل عائقا للقيام بمهمته بصفة عادية في مؤسسة تخضع إلى رقابة الإدارة التي ينتمي إليها أو لها صلة مع هذه الإدارة، وذلك تحت طائلة تعرضه للعقوبات التأديبية المنصوص عليها في هذا القانون الأساسي[5]"

- The parts of the one sentence can be placed away from each other in comparison with another sentence in another field, for example, we can find the subject located far from the verb.

These characteristics make the operations of understanding and processing legal texts difficult.

Here we depict some cases where the indicator is not sufficient to define the category of the provision:

Depending on the occurrence of the verb "يمكن" in the following article which distinguishes the existence of the provision of type permission, we classify the following rule as expressing the permission;

---

[5] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

"المادة 178: تعتبر، على وجه الخصوص، أخطاء من الدرجة الأولى كل إخلال بالانضباط العام **يمكن** أن يمس بالسير الحسن للمصالح ."[6]

However, this rule does not express the permission. The rule expresses an explication of the expression "first-degree error". Where, we can find the same vocalization, but with a different meaning.

"المادة 146 : غير أنه يمكن للمستخدمين الذين يشغلون مناصب عمل معرضة بصفة دائمة للإرهاق و الأخطار أو تنطوي على ضغوط بدنية أو عصبية، للإستفادة في إطار ما **يسمح** به القانون، بتخفيض المدة الأسبوعية للعمل ."[7]

As we observe in the example shown above, the indicator "**يسمح**" that belongs to the permission category, in the illustrated example does not take the meaning of the permission.

If we rely on the occurrence of the indicator "**على**" at the beginning of the sentence: "...**على إثر فترة التربص،**" then we classify the rule as a type of obligation, however, the preposition "**على**" in the sentence, does not express an obligation. Where "**على**" in Arabic language could take several functions, we mention of them: preposition, gerund and adverb.

"المادة 49 : لا يمكن اعتبار الإنقاذ **حقا** مكتسبا للطالب، بل يعد من صلاحيات لجنة المداولات حصرا ."[8]

In the example shown above, the indicator "**حقا**" belongs to the permission category; however, it does not refer to the type of permission

"المادة 5 : تختلف الرتبة عن منصب الشغل. وهي الصفة التي **تخول** لصاحبها **الحق** في شغل الوظائف المخصصة لها ."[9]

In the example presented above, there are two indicators that belong to the permission category but do not express a permission type. The sentence expresses a definition.

---

[6] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

[7] The internal regulations of Kasdi Merbah Ouargla University 2007

[8] Order No. 711 of November 3rd, 2011. Setting the rules of organization and pedagogical management common to the university studies in order to obtain the degrees of License and Master.

[9] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

With these particularities, it is not possible to rely only on the words indicating the existence of the provisions in the sentence. Where, we conclude that the context is important for the specification of the word and the text cannot be represented as a bag of words, because of the significant relationships between the words (Bousmaha et al., 2013).

2.4. Contextual Exploration method

The Contextual Exploration CE method proposed by Descles (1997) is based on the linguistic markers occurring in the textual segment (title, sentence, paragraph, etc.) for the aim of the annotation. These markers are analyzed and accumulated by the help of linguist and a domain expert. The indicator in the textual segments could take several meaning according of the context of the text; hence the appearance of the indicator may not be enough for the association of the corresponding annotation. Therefore the CE method is based on the clues in addition of the indicators. So, the markers are indicators and a set of clues as shown in the Fig.2.5.

For the purpose of explaining that, we take the work presented by the authors Bertin et al. (2010), which they used the CE method for the annotation of definitional phrases written in French language. In the following example:" « *La culture de la* **paix est définie** *par les Nations Unies* **comme** *« un ensemble de valeurs, attitudes, comportements et modes de vie qui rejettent la violence et préviennent les conflits en s'attaquant à leurs racines par le dialogue et la négociation entre les individus, les groupes et les États* » »[10]. as we observe, the marker that correspond to the view point definition is the verb «définie»: « define » which refer to the indicator. However, the indicator in the sentence may carry another meaning such as determine. Therefore, for the disambiguation we depend on the other marker (clue) that exists in the sentence. The clue in this sentence is the preposition "comme" that follow the verb "définie" as shown by the authors Bertin et al. (2010). The authors argued that the verb "définir: to define " can take the meanining of definition if it is followed by a preposition like "*comme*": as, "*par*" : by, "*avec*": with, "*grâce à*": thanks to and "*au moyen de*": by means of. So the sentence correspends to the view point definition.
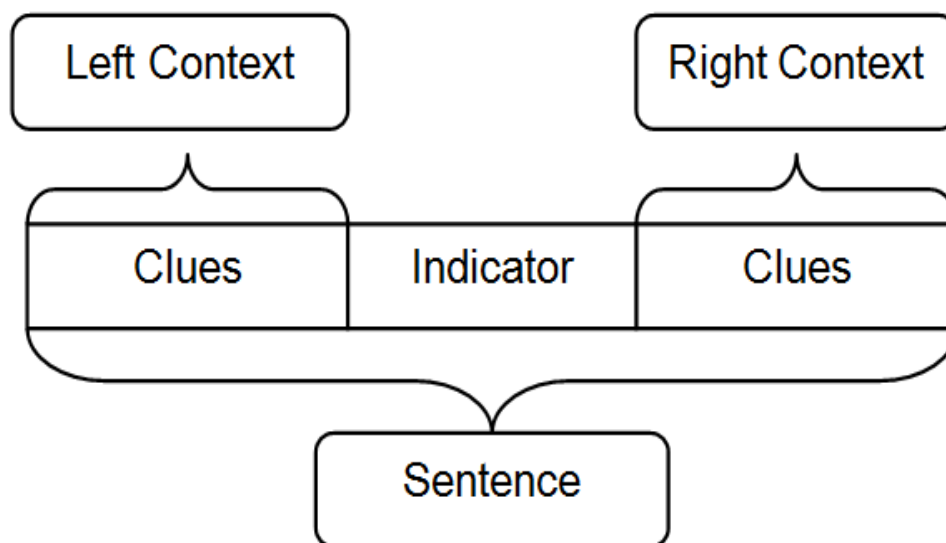
Fig.2.5 Contextual Exploration method markers and search spaces.

The CE method uses the clues to indicate the existence or the non-existence of searched information. Also, it uses the hierarchy between indicators and clues (Descles et al., 2009), unlike the rule-based techniques that rely on the occurrence of particular patterns in a text. Besides, the CE method can use one or more clues.

The CE approach is based on a set of linguistic markers and formal rules. its principle consists of searching for a set of clues at the right and/or the left context of the indicator or in the indicator itself, by the construction of a set of contextual rules, and the creation of a semantic map (linguistic ontology). Each CE rule relates a class of indicators with different classes of clues.

The CE method does not need a pretreatment phase, unlike techniques that involve the steps of tokenization, stemming, lemmatization, POS tagging, named entity recognition, etc. in addition, the main characteristic of the CE method is that it is independent of the domain of application (Le Pirol, 2011). Several studies are carried out by the use of the CE method in different languages such as French, Arabic, Korean and English.

 Examples of these works are: extraction of causal relationships (Jackiewicz, 1998), automatic summarization (Minel, 2002), identification of reported information (Alrahabi et al., 2006), automatic analyze and standardization of unstructured data

(Fadili and Jouis 2016), annotation of definitions (Teissedre, 2007), text segmentation (Mourad, 2002), events annotation (Elkhlifi and Faiz, 2009), Annotation of named entities (Bouhafs, 2005), Arabic quotations categorizes (Alrahabi, 2015), etc.

With the exploration of sentence contexts, we can significantly overcome some difficulties related to the Arabic legal language processing. The semantic information of the provision is the type of the provision and the arguments. Depending on the CE concept we can capture the semantic information contained in the text. In the next chapter, we present the methodology of our research.

Chapter 03

Methodology

3. Methodology

This chapter shows the methodology of the research. We introduce our proposition model for Algerian Arabic legal texts **specification** and **conceptualization**. The development of models that concern the legal rules helps individuals complying with legislation. The first section concerns specification of semantic information, the second section presents the classes of the mission of provision identification,   and then the semantic annotation process is described. Finally the last section is about the Arabic legal ontology.

It is obvious that it is not possible to reuse applications created to treat legal texts of other languages for the aim of processing Arabic Algerian legal texts. In our research we have to deal with different particularities of Algerian Arabic legal language that makes the task of processing more complicated than English language. Therefore, for texts written in Legal Arabic language we have to develop a dedicated system, in another word it requires specific tools designed for the Arabic legal language. In our research we chose a rule-based approach for the identification of information.

Why we chose a rule-based approach?

We decided to build a rule-based system. We took this decision because of the following reasons:

- Complex nature of Arabic legal texts.
- There is not an annotated legal texts corpus available for Algerian Arabic legal texts.
- The difficulty of accumulating the Algerian legal documents as the documents are published as secure or as scanned documents.
- There is not enough amount of data in order to use it for a machine learning training task (Cunningham et al., 2011).
- The legal language is different from ordinary language.
- Legal domain requires the accuracy, the rule-based approaches gives more precise results than machine learning approaches in general and in particular for Arabic language.
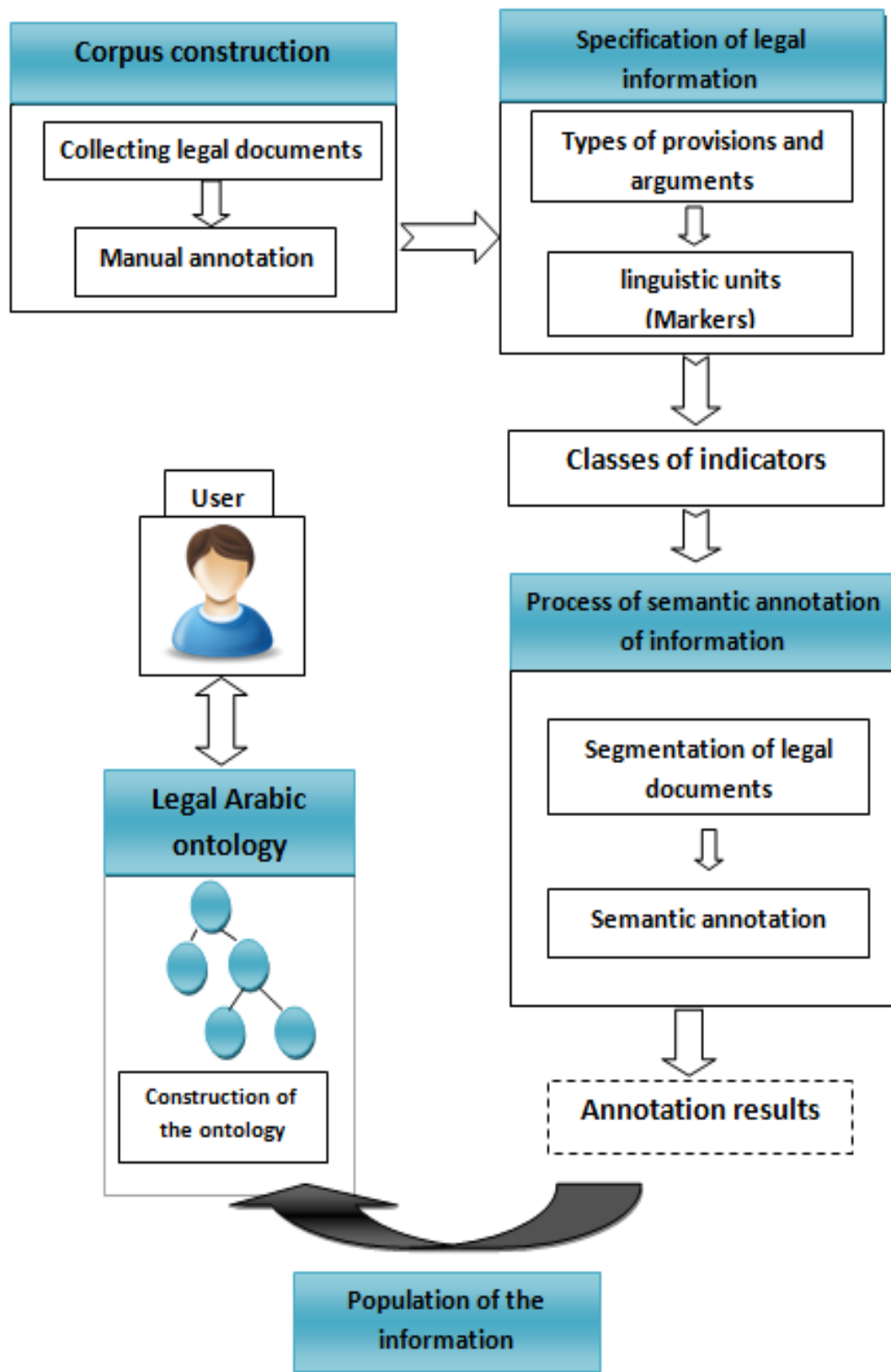- With the use of machine learning the traceability cannot be maintained.

Fig.3.1 Architecture of the model

- Concerning the transparency, the validity is one of the issues because the ML algorithms are considered as a "black box" hence it is so difficult to correct the errors that results from the classification process.

Concerning the general overview, first we construct the corpus by collecting a set of Arabic legal texts concerning the university, and annotating the collected corpus by a set of annotators. Next we pass to the task of specification of the semantic information, here, we starts by the definition of the types and arguments that we are interested on, and then we study and define the linguistic markers that concern each type and arguments. After that we define classes of the mission of provision identification. Next we move to the process of semantic annotation of information that includes two main steps which are the segmentation of legal documents and the semantic annotation. Finally, we build a legal Arabic ontology in order to populate the extracted information in it. The different stages of the proposed methodology are illustrated in Fig.3.1.

In the modelling phase we pass through two main methods which are the semantic annotation of information and the semantic representation of legal knowledge using the ontology. In the following we present the different phases in detail (the first phase which is the corpus construction will be studied in the next chapter).

3.1. Specification of semantic information

In the legal domain, there are several kinds of legal rules, where we choose to focus in this thesis only on the prescriptions or provisions, in particular, in the following types: obligation, prohibition, and permission. As illustrated in Fig.3.2, the following divisions: the obligation and prohibition are included in the imperative rule category, and the permission is integrated into the suppletive rule category. The obligation type is a duty that the entities must do and comply with, and the non-compliance of the bound may lead to a penalty. In contrast, the prohibition type is the one that the entities must not do. The permission type is the one that the entities can choose to do or to adopt the contrary.

Two criteria that can be relied on to distinguish the type of the legal rules, the verbal standard and pragmatic standard 2007 ,زعلاني)), on which we may depend on one or

both of them (2007 ,زعلاني 2007 ,جعفون;). The verbal standard depends on the words and expressions used in texts to define the type, except in some cases where the words and
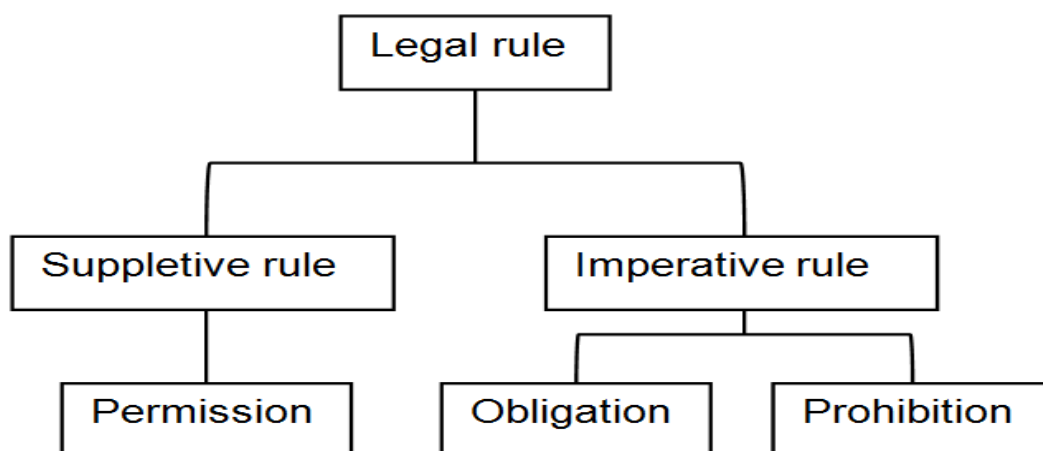
Fig.3.2 Subdivisions of imperative and suppletive rules.

expressions do not allow specifying the type of the rule; hence, we must depend on the pragmatic standard. The latter involves relying on the judges to determine the type of the legal rule.

Arabic legal language uses authority words to express an obligation (with the consideration that all legal rules are binding). We mention as an instance of authority words: "يجب", "يمكن", etc.

Legislative drafting is characterized by the existence of the following essential elements (arguments): - role, - action, - case description, - condition, and – exceptions (دودين, 2009). The role is the legal subject; it indicates the address (e.g., employer) to whom the rule is applied (Jiang et al., 2012). The role expresses the actor or the organization; the latter corresponds to a group of actors or organizations represented as one (Hoekstra et al., 2007). The action is the legal action in the provision, which determines the action assigned to the Role (Hoekstra et al., 2007). Case description defines the situation in which the rule is applied. Condition specifies the particular circumstance required for the application of the provision. Finally, exception indicates the exclusion elements of the application of the provision. The action and the role are mandatory in the permission and obligation, while, the remaining arguments: case description, condition, and

exceptions may or may not exist in the provision rule. In our thesis, we are interested in specifying the arguments role and action.

3.1.1. Specification of provision type

In the legal text, we find set of verbs, nouns, and a limited number of expressions and gerund expressing the meaning of obligation. Following this study, we assume that the specification of these words and expressions permit the identification of the provision types.
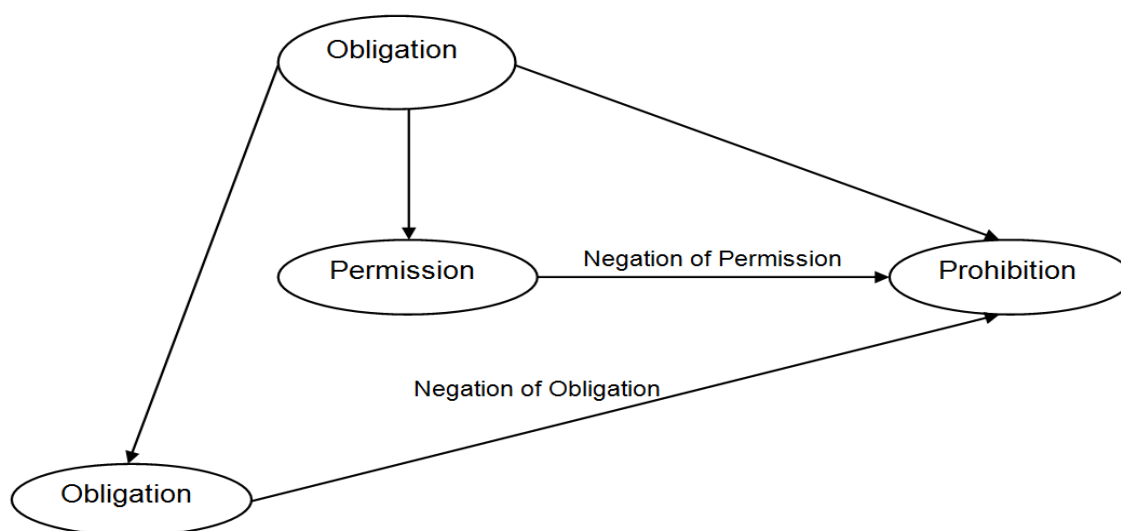
Fig.3.3 Types of obligation.

Each type has set of indicators that can lead to the distinction to which type belongs. As mentioned above all the provisions are binding rules (obligations), the Fig.3.3 illustrates the different cases of types: obligation, permission, and prohibition, where the negation of certain indicators expresses the prohibition type.

3.1.1.1. Provision markers

Here we will present the markers: indicators and clues.

a. The indicators

The indicators could be:

- Verbs (e.g. "يمنع": is translated to the verb "to prevent", "يجب": is translated to the modal verb "must", "يمكن": is translated to the modal verb "may"),
- Nouns (e.g. "واجب": duty),
- Gerund expresses the meaning of the obligation (e.g., "على": shall)
- Expressions (e.g. "ويقع باطلا كل شرط يخالف ذلك").

The detection of the indicators contributes to the identification of the type of provision; however, in some cases, it is not possible to depend only on the occurrence of these indicators in the sentence, because of the polysemy of the words. Therefore, for the identification of the right meaning of the terms, we rely on the clues exist in the sentence context, in addition to the indicator. English grammar contains modal verbs such as (must, should, would, can, may and might) however in Arabic language the modal verbs are used with the subordination of 1) "من ال" with a sentence, 2) verb with: particular particles and prepositions, or the infinitive verb "مصدر الفعل". English language uses the modal verb "must" to express an obligation, Arabic language uses, for example, the verb "يجب" with the addition of the particle "أن". However, in English, particles and prepositions cannot be used after modal verbs (e.g. must) while these cases are used in Arabic language; where the literal translation for ("يجب أن") to English language is: (must to).

Thus, the classification of the sentence to a given type depends on a structure that includes the indicator and a set of clues; these indicators may have different orders, it may occur in the sentence at the beginning, the end or another place. in addition, the clues can come directly after or before the indicator, or separated from each other (e.g. "يجب...أن").

Some of the indicators used can be agglutinative, and there are those that cannot be agglutinative; the indicator coordinated with a suffix in the last of the word which is a pronoun that associated to the role of the provision. For example, we take the indicator "يمكن"; it is written with the addition of "ه": "يمكنه"in the singular masculine, "ها": "يمكنها" in the feminine singular, and "هم": "يمكنهم" in the plural, etc. Besides, it can be coordinated at the beginning of the word with a prefix.

Here, we mention some indicators related to each type:

Permission type indicators:

يجوز، يمكن، يسمح، يرخص، يخول، يحق، حقا، "ما لم يوجد اتفاق أو عرف يقضي بغير ذلك"، "ما لم يوجد اتفاق أو نص قانوني يقضي بغير ذلك"،"ما لم يقضي الاتفاق بغير ذلك"...

Obligation type indicators:

يجب، وجب، واجب، على، يتعين، يلتزم، يلزم، ينبغي، ملزم، يجبر، إجباري، إلزامية، عليهم،، "ويقع باطلا كل شرط يخالف ذلك"، "ويقع باطلا كل اتفاق على خلاف ذلك"...

Prohibition type indicators:

يمنع، يمتنع، يحظر...

Besides the words of prohibition type, we can also find the negation structures of certain indicators of the obligation and permission types by adding negation particles (e.g. ,غير لا, لن), we state for example:

etc.لا يسمح, لا يمكن ,لا يجوز , لا يخول،...

The indicators are organized in a semantic map (linguistic ontology) where their categories are: obligation, permission, and prohibition.

   b.   The clues

The clues can play several functions, in our study, we are interested in the clues that are involved in defining the type of the provision, where each indicator is accompanied by particular clues, to give a specific meaning. Moreover, we are interested in some clues that are used to define roles and actions. Also, we rely on them to determine the sentence boundaries.

These clues could have different natures we mention from them:

Prepositions (e.g. "ل ":"Li" (for, to), "ب ":"Bi" (to), "على":"Aala" (on, for));

Particles: (e.g. "أن": "An" to);

Negation particles: (e.g. "لا" ,"غير" , "ليس" : not) ;

Words or expressions: (e.g. **"في هذه الحالة"**: in this case);

And location of the marker in the context (e.g. at the beginning, end).

3.1.2. *Specification of the argument role*

One of the arguments contained in the provision is the role, which is the legal subject to which the rule is applied. The role expresses the actor or the organization, and the latter is a group of actors represented as one (Hoekstra et al., 2007).

In order to specify the role, we relied on the analysis of the provisions that are studied to determine all roles; also, we add the roles that concern the domain of application; each one of them can be singular or plural. After that, we gave labels to all the actors or organizations that occur in the sentence. Since it can be the syntactic subject or the syntactic object, hence, we have tried to define the role for which the action is assigned in the provision. Therefore, we consider the actors and organizations labeled as expected roles, and we assign them the following label: (expRole), until the confirmation that the expRole is the role (or it is not), according to the method proposed below.

There is a kind in which the operation of the role selection, is accomplished after the specification of provision type. And in another kind, the determination of provision type, depends on the existence of role, hence, in this kind the role selected before the specification of the type of provision.

In both kinds, operation of the annotation passes through the following steps:

Check the existence of expRole, if this latter exist, we search for particular clues according to the indicator occurred in the sentence in the search spaces of expRole, which are (exp_role_Rc) and (exp_role_Lc), with this operation we prove whether the role exist or not.

In the following provision, the role is "الموظف", where, we based on the occurrence of the clue "على" in the right context of "الموظف" for the association of the annotation.
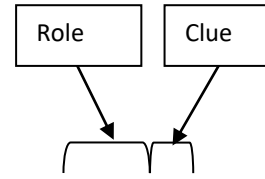
"المـادة 40 :يجب على الموظف، في إطار تأدية مهامه، احترام سلطة الدولة وفرض احترامها وفقا للقوانين

والتنظيمات المعمول بها. "[11]

3.1.3. Specification of the argument action

The action is one of the arguments of the provision, where it determines the act associated with the Role (Hoekstra et al., 2007). As we mentioned, the structure used in the Arabic provisions includes clues (e.g. "أن": Ann, "بـ": Bi) besides of the indicator of provision. For example, the subordination of the indicator with the particle "أن" which is informative; used to indicate what the action in the sentence is; and this corresponds to the argument action in the provision.

After specifying the role and the provision type, we specify the argument action. For the specification of the latter, we rely on some clues that may occur before the action, however in the case of the absence of these clues; it is difficult to address the action of the argument. Where in this case, we make the following assumptions:

- If the role exist then, in this case, we assume that the action is the part that follows the Role;
- If the role does not exist, then we assume that the action is the part that follows the indicator.

The identification of the action argument applied after the satisfaction of the condition of the existence of the type of provision.

For example, in the following provision:

---

[11]Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

```
                                                    ┌──────────┐
                                                    │   Clue   │
                                                    └──────────┘
                                                         │
                                                         ▼
                                                        ┌┐
"المادة 49: على الموظف أن يسهر على حماية الوثائق الإدارية وعلى أمنها. .."¹²
                                                    │
                                                ┌───────┐
                                                │ Action │
                                                └───────┘
```
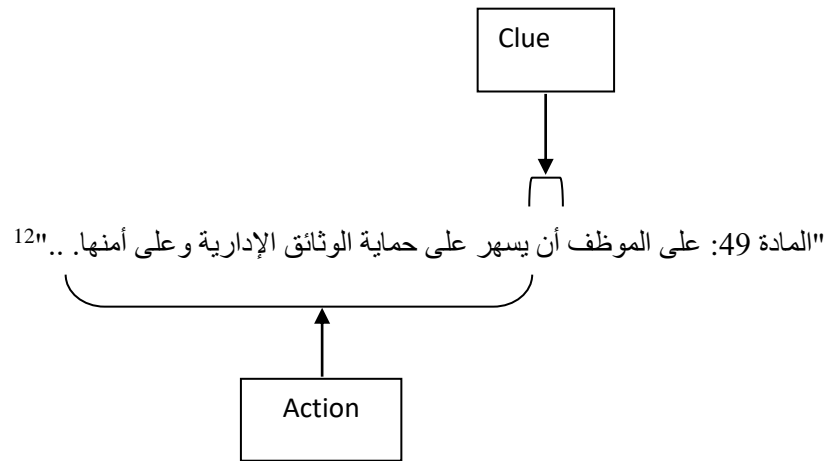
Here, we relied on the clue ("أن": "An") for the annotation of the argument action.

3.2. Classes of indicator

According to the analysis of different provisions, we decide to classify the mission of provision identification in different classes. As follows in the Fig.3.4, we define two categories, the provisions that contain an indicator that helps in their identification and those that do not contain, which are named respectively "without an indicator" and "with an indicator". We divide the latter into two classes, the first is "the indicator sufficient" and the second is "the indicator insufficient", in the latter we specify the classes "the indicator that needs clues" and "the indicator that needs a role" for the identification of their type. Each class that detected by an indicator follows a method for the specification to which type the indicator belongs. We collected a total of 57 indicators, 35 clues and 603 roles. And we created 18 rules for the identification of provision type.

Concerning the prohibition type, for the identification of anti-permission and anti-obligation (i.e. the negation of indicators belonging to the categories permission and obligation), we rely on the existence of negation clues. Hence, for all the classes mentioned above, we can check the presence of the negation clues for the definition of the prohibition type.

---

¹²Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.
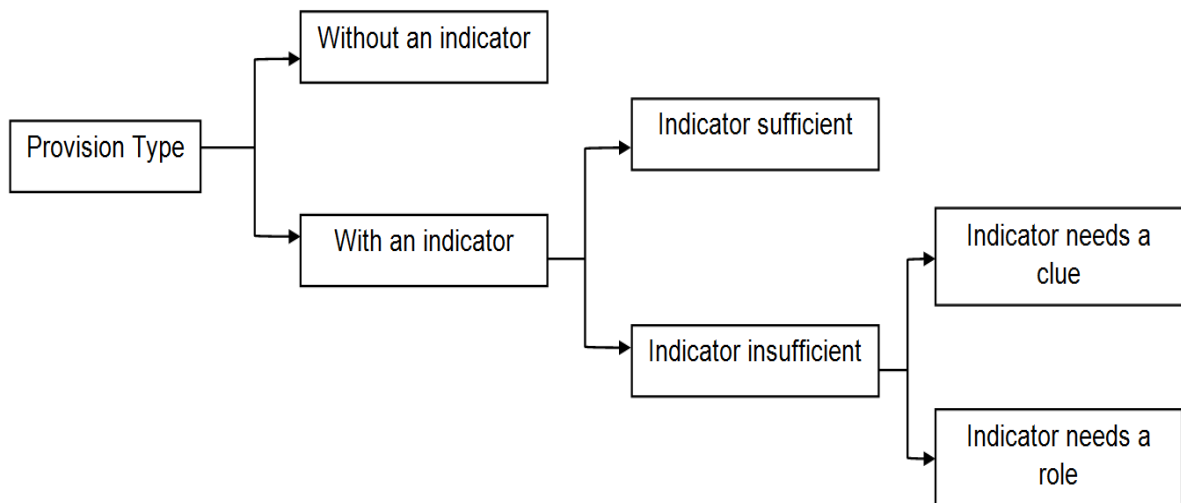
Fig.3. 4 The classes of provision type identification

3.2.1. With an indicator

3.2.1.1. Indicator sufficient

The indicator can be a term or an expression, where the detection of the expressions in the sentence context is sufficient to determine the type of provision.
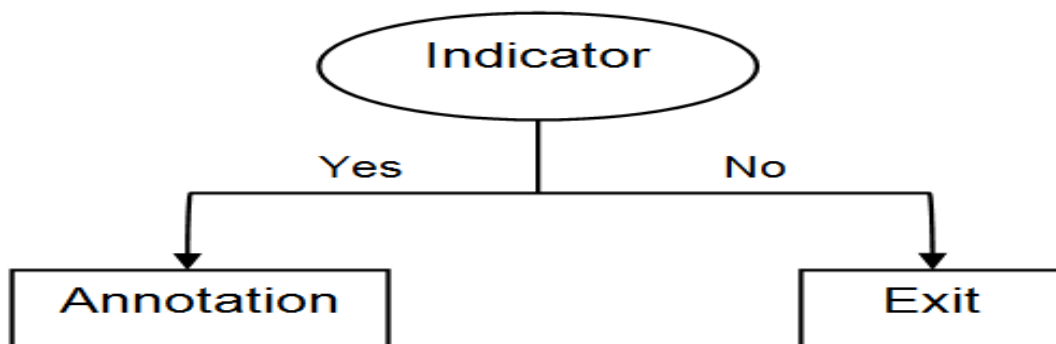
Fig.3.5 The method of indicator sufficient

Also, the occurrence of some terms is sufficient to indicate the right meaning of the term, because the word does not have a different meaning (e.g. "يمنع"), or some indicators although they may take more than one meaning, however, in the context of legal drafting we found that they have only one meaning (e.g. "يجب"). The method of indicator sufficient is given in Fig. 3.5.

3.2.1.2.   Indicator insufficient

This class includes two categories "indicator needs a clue" and "indicator needs a role":

a. The indicator needs a clue

In this class, the occurrence of an indicator and particular clues in the right and left context of the indicator allows the detection of provision type. The steps of this class are as follow: we start by the search for the occurrence of the indicator in the sentence, after that, the definition of contexts of the indicator which are the right and the left context.
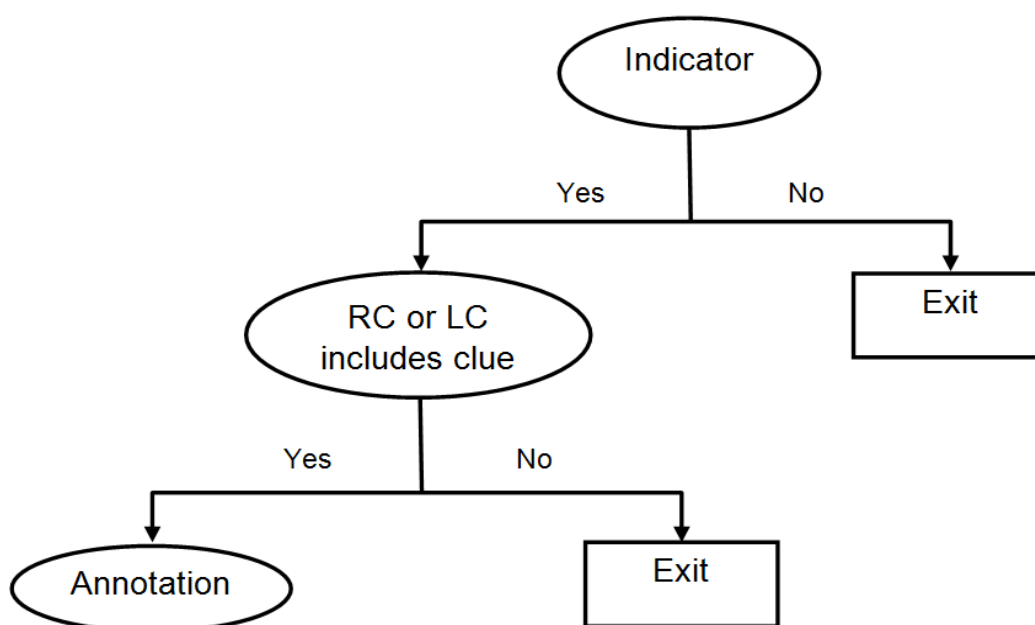
Fig.3.6 The method of the indicator needs a clue.

The next task is the search in the right context for the occurrence of particular clues according to the indicator detected, after that, the research is performed in the left context to find other particular clues. The indicator may require searching in both contexts, or only in one search space (the right or the left context). For example, we take the indicator "الحق" which requires the occurrence of a clue as "في" in the left context in order to categorize it. These steps enable to identify if the sentence is a provision and to associate the corresponding type for it. The steps are shown in Figs.3.6.

b. The indicator needs a role

This kind involves the existence of the role beside the indicator to take the meaning of the obligation (e.g. the indicator "على"), and for others, the presence of role helps in the categorization of the term.

The method of specifying this class is as follows; we start by searching for the possibility of existing of the indicator in the sentence, after that, defining the contexts of the indicator that are the right and the left context. Then, checking the presence of clue in the contexts. The next task is to check the presence of "role" in the determined contexts. If role exist, then we associate the type for the provision. The steps are shown in Figs.3.7.
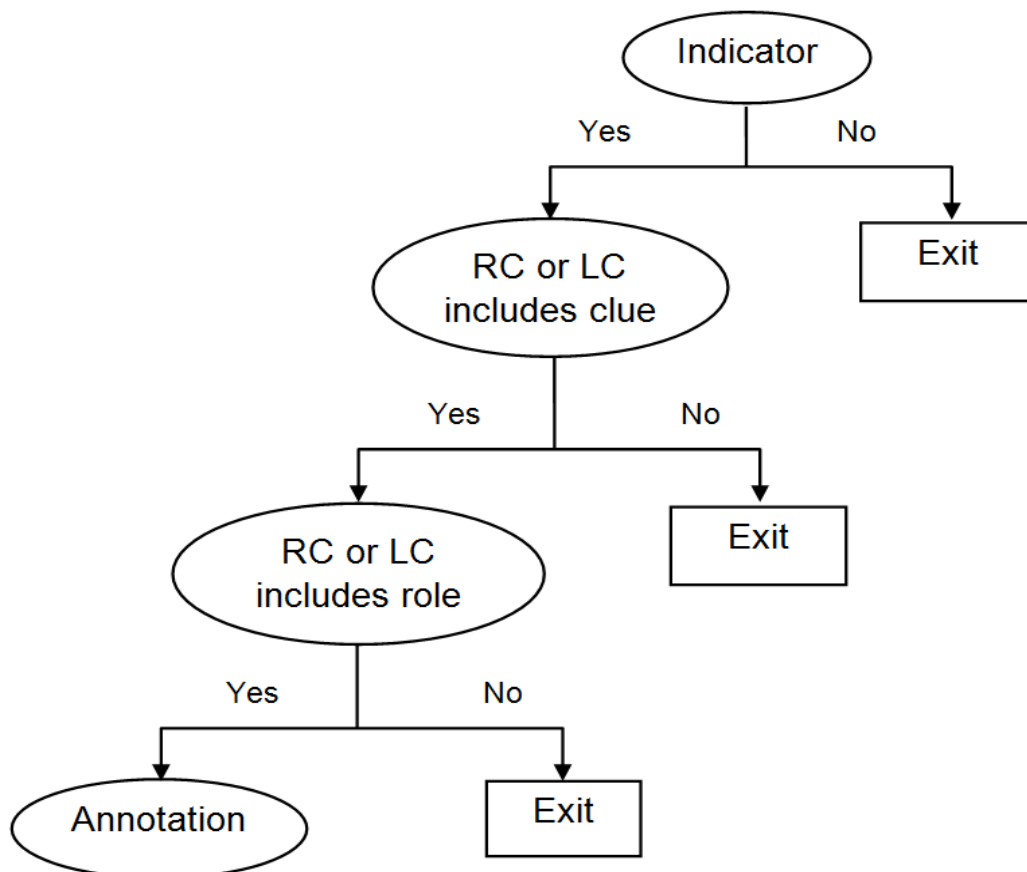

Fig.3.7 The method of the indicator needs a role.

3.2.2. Without an indicator

The obligation type is different from the permission and the prohibition types where the provision of the type obligation in some cases is expressed without any marker. Some legal rules their classification to a specific category carried out based on the markers; such as the use of modal verbs (must, shall, may, etc.). In contrast, certain obligations are expressed with the action to be obliged. We take for example the following provision :

"المادة 40 : سن الرشد تسعة عشرة (19) سنة كاملة"[13]

This case is one of the challenges presented in the artificial intelligence and law. This problem is not addressed in our study, where with these types of provisions we suggest the collection of legal rules that are not annotated by the system for manual annotation by a domain expert.

Two indicators or more in the sentence:

At the level of the provision, it is possible to find more than one indicator. We treated this case as follow:

1. We can find two successive indicators as illustrated in this example (يمنع على الموظف...), where "يمنع" and "على" are marked as indicators, however, "على" in this example play the role of the preposition. In this situation, we take into account only the first indicator.

2. Some provisions contain more than one indicator located away from each other. As shown in the first example below, the rule contains the indicators "يخول" preceded by "لا" and the indicator "الحق", where, the first indicator takes the priority in this case. The same decision is made for the second example, which includes the three indicators: "يمكن" proceeded by "لا", " تخول" and "الحق".

   Example 1 :

"المادة 18 : لا يخول تعيين غير الموظف في وظيفة عليا للدولة أو في منصب عال صفة الموظف أو الحق في التعيين بهذه الصفة"[14].

---

[13]Ordinance No. 75-58 of 20 Ramadhan 1395 corresponding to September 26, 1975 bearing the civil code, modified and completed

Example 2 :

"المادة 202: لا يمكن بأي حال أن تخول العطلة المرضية الطويلة الأمد كما يحددها التشريع المعمول به، مهما

كانت مدتها، الحق في أكثر من شهر واحد كعطلة سنوية. "[15]

3. If an indicator occurs with the expression indicator (e.g. " ويقع باطلا كل شرط يخالف ذلك

"), then, the expression indicator takes the priority.

4. Some terms, belong to categories that are not addressed in our research thesis, in
this case, depending on the term, it can take the priority and it can not. For example,
if the indicator is preceded by definition, the latter, takes the priority. Besides, we
take for example the term "**تعتبر**"; where, if "**تعتبر**" precedes the indicator "**حقا**" then
the latter takes the priority.

3.3. Process of semantic annotation

As shown in Fig.3.8, the process of automatic annotation of the provision types and
the arguments role and action follows two relevant steps: the first step is the definition
of the structure of the legal document by the segmentation of the text into articles, after
that, the segmentation of article into sentences. Before the segmentation phase, an
important step must be performed. It is the devocalization by removing the diacritics
from all the words.

The second step is the semantic annotation. In this phase, we apply the built rules on the
output of the first step. The database contains classes of indicators and clues and a set of
rules.

3.3.1. Segmentation of legal document

The identification of parts of Arabic legal documents is a challenging task, where several
obstacles are faced with this task. In English and many other languages, the punctuation
marks are used for recognizing boundaries of sentences. However, Finding Boundary of
sentences is a challenging task in Arabic languages. The Arabic language does not follow
a strict punctuation rules. Where, Arabic paragraphs can be written with only one period
at the end of the paragraph. Another obstacle is the absence of capitalizations (Douzidia

---

[14]Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

[15]Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

and Lapalme 2004, Farghaly and Shaalan 2009). Also, the difficulty of distinguishing conjunctions of coordination (Shaalan 2010, Biskri et al. 2017) and subordination , etc.
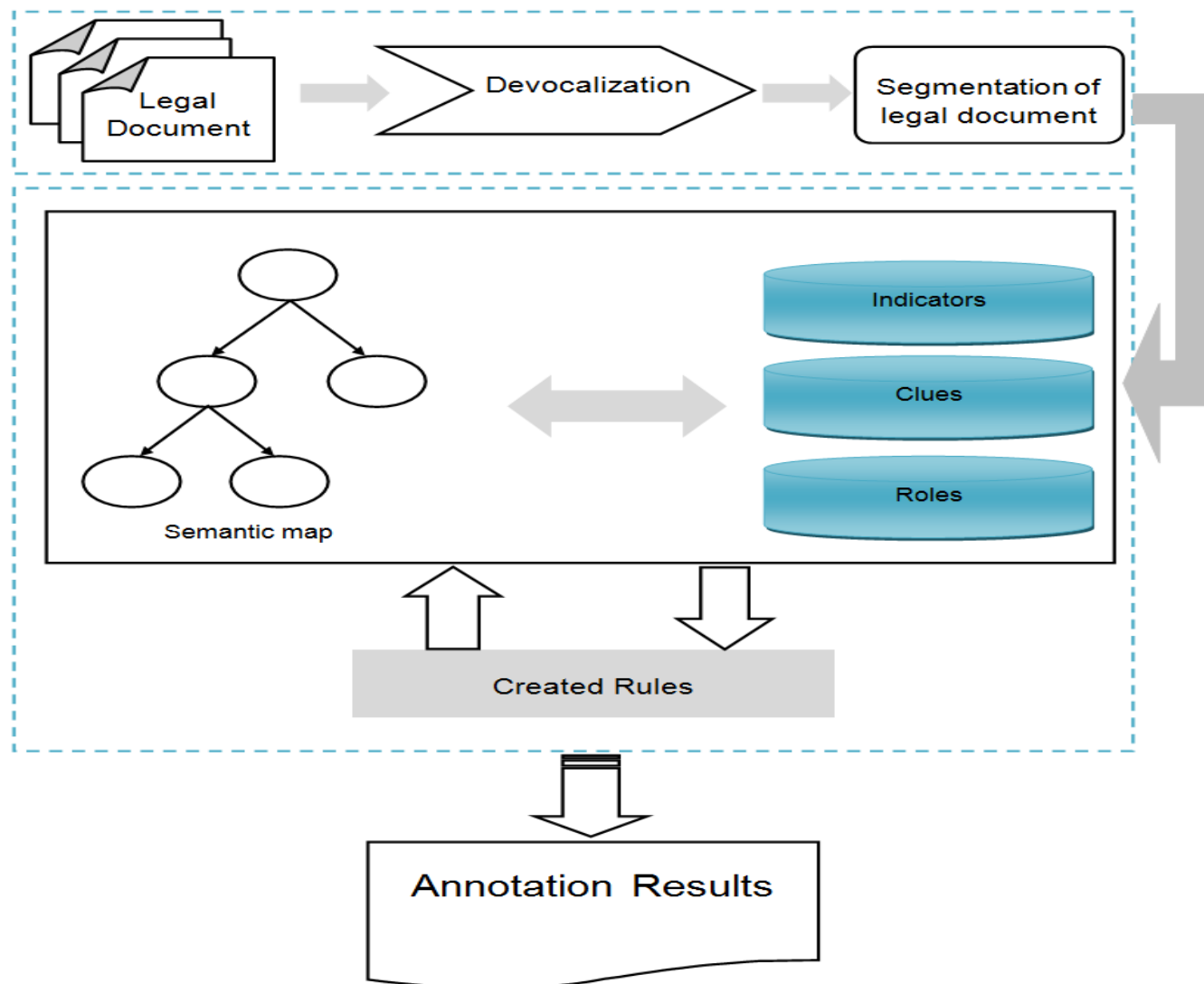
Fig.3. 8. The semantic annotation of Arabic provision.

First, we divided the legal text into articles where each article may have one or more provisions. In order to specify the parts, we depended on the linguistic terms (e.g. "المادة" "Article") and on the analysis of punctuation. The specified articles are then segmented into sentences. For the specification of the sentence boundaries, we relied on the indicators and set of characteristics such as the punctuation, coordinator, etc.

Taking into account the complexity of the segmentation phase, we added a manual segmentation or a correction to the results obtained from the segmentation phase. The result obtained is the input of the semantic annotation phase.

3.3.2. Semantic annotation

The semantic annotation phase aims to annotate the types of provisions and the arguments role and action.

For the semantic annotation, we construct a set of conditional rules in which the action of the association of semantic labels depends on the satisfaction of the entire rule conditions. According to the detected indicator, the call of the rule is made. Moreover, the selection of rules is achieved by the priority assigned to the rules.

"المادة 49: على الموظف أن يسهر على حماية الوثائق الإدارية وعلى أمنها. ..."[16]

We take, for example, the sentence shown above to illustrate the steps we have followed. The sentence includes the indicator "على": which expresses the meaning of the obligation (shall), we note that "على" in Arabic could be a preposition, a gerund and an adverb.

The rule is achieved by searching for clues in the right and the left context of the detected indicator, and searching for clues in the right and the left context of the (expRole). In the case of satisfaction of the whole rule conditions, then three annotations are made: the provision type, the role, and the action associated with the Role. The following steps are needed for the semantic annotation with the occurrence of the indicator "على":

As shown in Fig.3.9, we search in the search spaces of the indicator "على", if the right context is empty (empty is a clue) or includes the clue from the list (list_clue_beforeInd).

The second step is to check in the left context if role exist; where we give expRole label to all expected roles, after that, we check if the right context of expRole (exp_Role_Rc) is empty or if one of the clues of the list (list_clue_BeforeRole) exist then we assert that the expected role is the role in the provision. After that, we make two annotations: we declare that the sentence is a provision of the type Obligation, and we annotate the expRole with Role. In the later step, we search in the left context of role (exp_Role_Lc) for the occurrence of the clue "أن" if we find it then we attribute the annotation Action to all the

---

[16]Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

part that follows "أن" else if the latter does not exist, we attribute the annotation Action to all the part that follows Role.
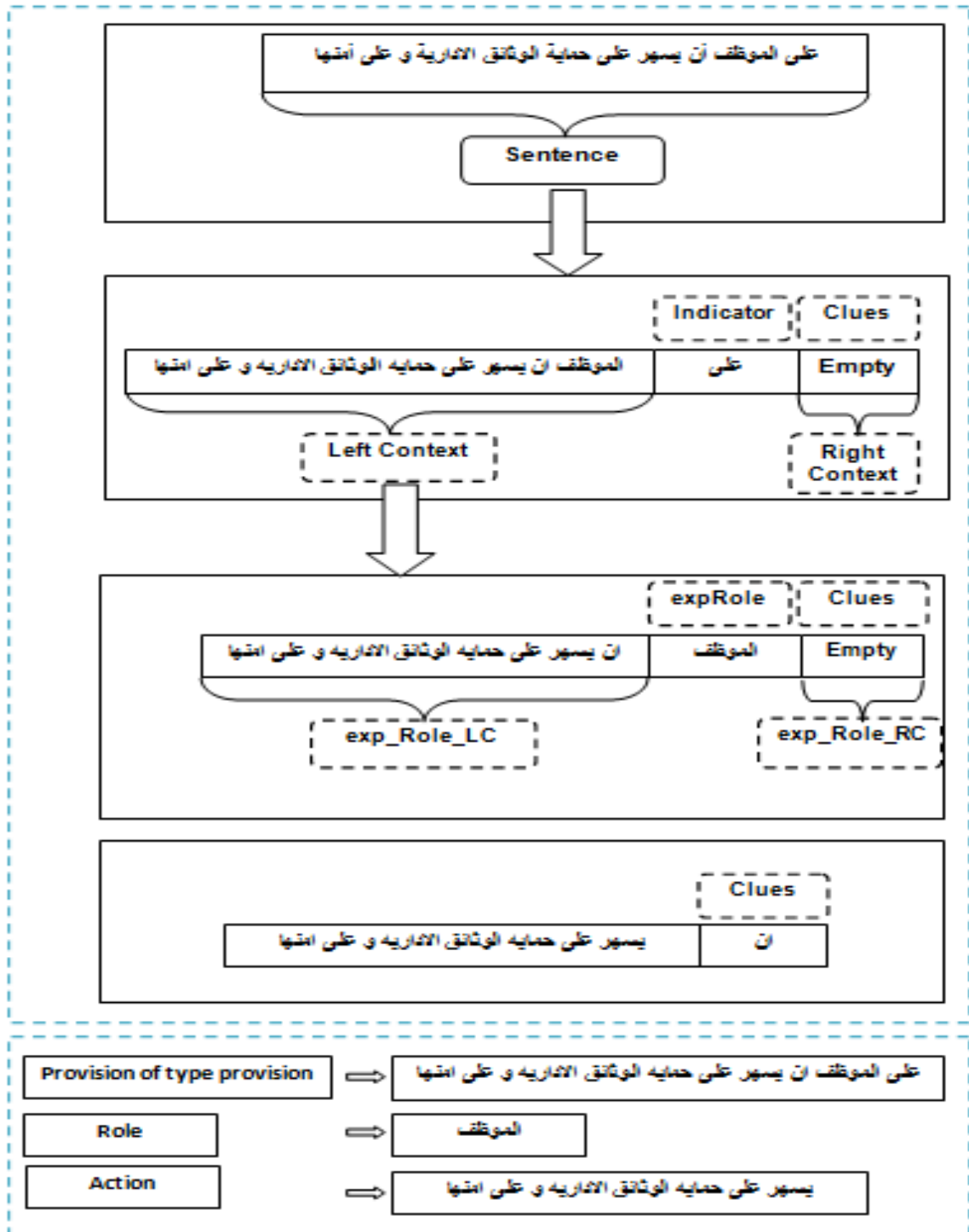
Fig.3. 9. Specification of semantic information of Provision that includes indicator "على".

The following example represents the rule used to annotate the example shown above:

Rule:

For each sentence S in the corpus

    If indicator ∈ S then

        Let LC be the left context of indicator and RC be the right context of indicator

        If LC is empty or ∃ word ∈ list_clue_beforeInd in the LC then

            If ∃ expRole in RC then

                Let exp_Role_Rc be the right context of expRole

                if exp_Role_Rc is empty or ∃ word ∈ list_clue_BeforeRole in exp_Role_Rc then

                sentence ← provision of type obligation

                expRole ← Role

                    If ∃ clue «أن» in exp_Role_Lc then

                        the part that follows "أن" ← Action

                    Else

                        the part that follows Role ← Action

                    End

                End

            End

        End

    End

End for

3.4. Population

Population is the task of enriching the knowledge base with new instances. We performed the population task manually. Firstly, we populated the extracted information,

65

which is the output of the semantic annotation process. Then, we enrich the other information into the ontology.

3.5. Arabic ontology model

This phase shows the conceptual modelling of legal texts. We introduce our model for Algerian Arabic legal texts conceptualization. The development of models that concern the legal contents helps the individuals complying with legislation. Where, the proposed ontology permits to add the semantics on legal texts to make them more understandable and to facilitate the research and the access processes.

The author Hoekstra (2007) argued that there are five main roles of ontologies which are: 1. organise and structure information; 2. reasoning and problem solving; 4. semantic indexing and search; 5. semantics integration and interoperation; and 6. understanding the domain.

The ontology allows storing information concerning the structure and the semantic information of legal texts. The modelling phase is made up of the following steps: first, defining the relevant terms in the legal texts and mine knowledge about the legal texts. Next, organizing and structuring this knowledge in order to build the legal ontology.

The LKIF represents the fundamental concepts used in legal domain, where it is the most widely used in the previous research. "LKIF-Core should cover "basic concepts of law". This means that the terms selected should be both highly abstract as to cover all domains of law, and be relevant for law" (Hoekstra, 2007). In the modelling phase we partially reused the LKIF ontology concepts.

We reused the top level ontology concepts of LKIF. The LKIF ontology has adopted the most general classes of the LRI Core. It is obvious that in order to model any domain, the task involves the basic conceptualization of the context. The provided definitions of the basic concepts help for the definition of the legal concepts in the ontology.

The Fig.3.10 presents the Top level LKIF ontology. This level includes the main concepts which are mental concept, physical concept, abstract concepts and occurrences.
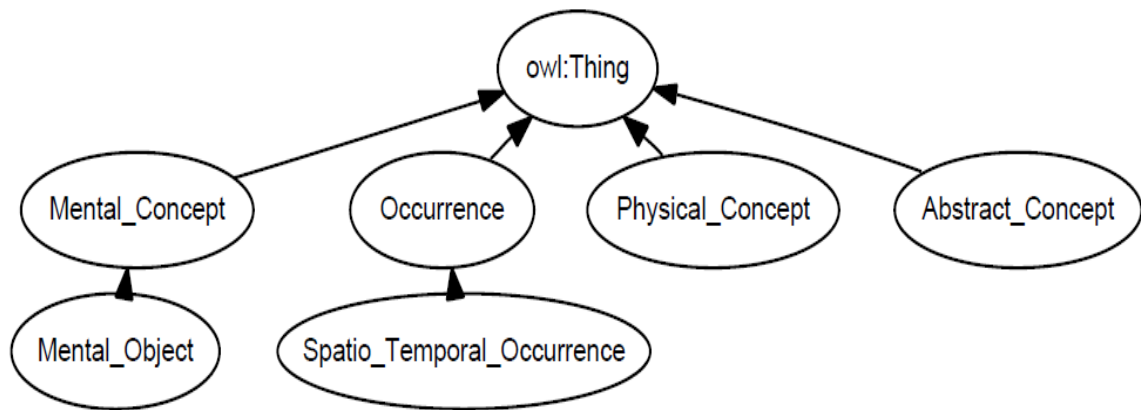
Fig.3.10 1 Concepts defined in the Top module (Hoekstra, 2007).

Mental_Concept

 "Mental entities reside in the human mind, and only have a temporal extension" (Hoekstra, 2009). Mental_Concept includes the Mental_Object which includes in turn the concepts Norm and Role.

Occurrence

The occurrence concept includes the main abstract concepts which are the place and time. In the occurrence concept the concern was in the representation of duration and location of changes; where it "captures those strictly temporal aspects related to the execution of scenarios involving objects and processes. This means that events are occurrences, but processes are not" (Hoekstra, 2007).

Physical_Concept

Here two main concepts are presented: the physical objects and the physical processes. The physical processes include the process concept which "is often used as synonymous to action and activity" (Breuker et al., 2007). The process concept depends on the definition of time and place. The Physical_Processes operate on Physical_Objects. The physical objects contain the concepts Natural_Object, Agent and Document.

Abstract_Concept

Abstract Concept represents the concepts: part and whole, part of, component of, containment, etc.

The development of our ontology can be divided into two main phases; the focus of the first phase is on the general structure of legal document by the representation of their metadata (e.g. source of document, date of publication). Where, we were interested in defining the different component that constitutes the legal document. Next, we focused on identifying and representing the semantic information that concerns legal document content. The concept names are created in Arabic language.

3.5.1. Representation of legal document structure
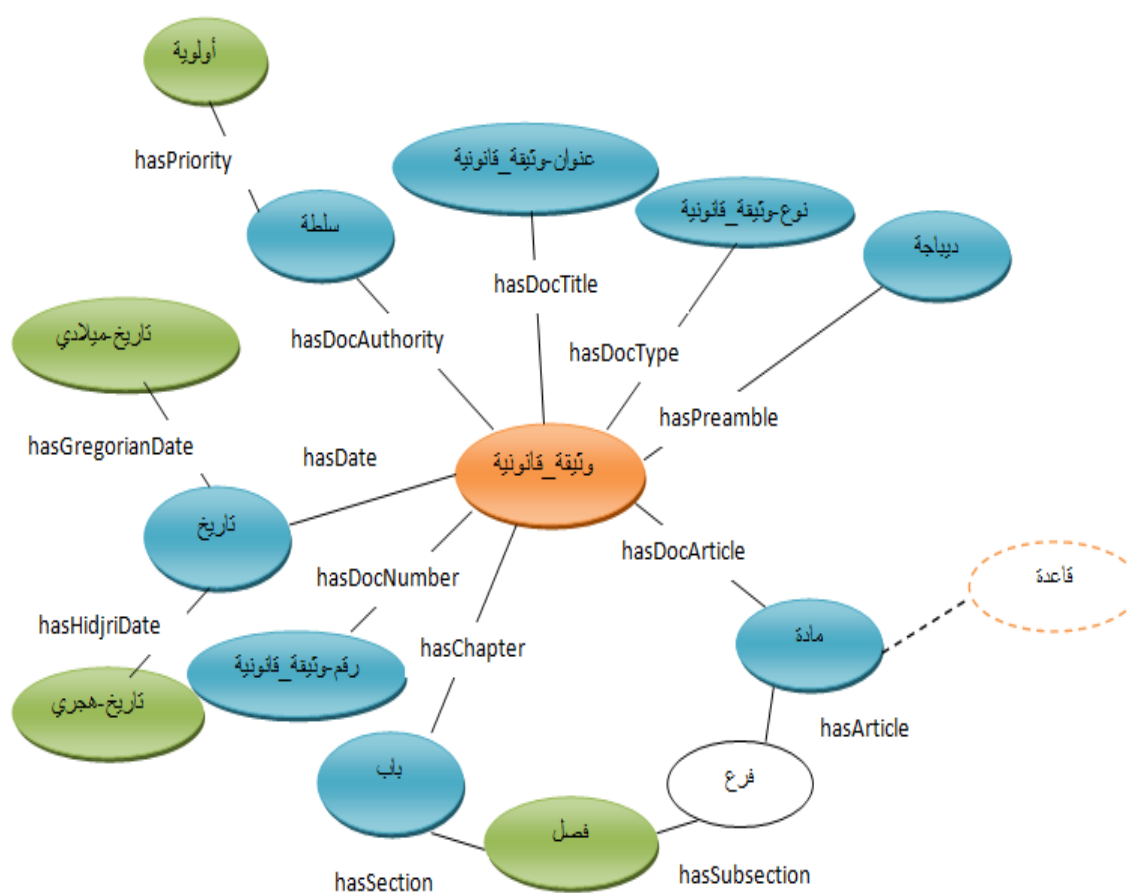
Fig.3.11 Arabic Ontology of legal document structure

The proposed ontology models the structure of legal documents. We relied on the concepts that exist and the relations that interlink the document structure concepts with each other. In this step we depended on the explicit information presented at the level of the legal documents. Where, this information is about the types of legal documents and their properties. The ontology contains concepts that identify metadata of the document that

allows their definition and description, the different relations that exist between them and the different parts that includes. Fig.3.11 depicts the main concepts.

- **وثيقة_قانونية:** Legal Document represents the text drafted by author state.
- **نوع_وثيقة_قانونية:** Legal Document Type is the category of the document which could be: مرسوم,قرار ,أمر, etc.
- **أولوية:** Document Priority is a feature that designate the ranking of the rule. This priority is defined according to Document Authority; whether is issued or created by president, minister, etc.
- **رقم_وثيقة_قانونية:** Legal Document Number is the number associated for the Legal Document which is composed by two numbers separated by (-), the first one indicates the year of the construction of the norm and the second number specifies the number associated to the norm; for example (15-112).
- **تاريخ:** Document Date correspond to the construction day of the norm, it is a succession of two dates :
  - ➢ **تاريخ-هجري** Hijric date (e.g. 20 Chaban 1436).
  - ➢ **تاريخ-ميلادي** Gregorian date (e.g. 8 June 2015).
- **عنوان_وثيقة_قانونية** : Legal Document Title represents the heading of the document where it defines the legal document context.
- **سلطة** : Legal Document Authority corresponds to the authority that constitutes or creates or participate in the construction of the legal Document.
- **ديباجة:** Preamble expresses for what the Author was based on during the drafting phase.
- **باب:** Legal Document Chapter defines the distinct parts of document. The Chapter has the legal document chapter title and it can be constituted directly by Article or contains Section.
- **فصل:** Legal Document Section represents the different parts of document chapter where it is associated with section title, beside it is constituted by Subsection or Article.
- **فرع:** Legal Document Subsection contains Subsection title and it has Article.
- **مادة:** Legal Document Article represents the Rule drafted by the legal document Authority at the level of the document, where the article can have more than one

Rule. Each Article has article number that represents the number associated to Article at the level of the legal document.

3.5.2. Representation of the content of legal documents

This phase aims for modelling the semantic information that exists in the law text which underlay the university organization. In the above subsection we relied on the concepts that exist in the structure of the legal documents; here we depended on the implicit information contained in legal documents. This requires first of all, an analysis by the help of domain experts to understand the meaning of terms, and to build a glossary that contains all the relevant terms of the domain (concepts, attributes, relations, etc.). After the analysis we built the ontology.

Representation of legal texts requires the understanding of the legal rules, and that involves a deep understanding of the relation that interlinks the different legal rules with each other.

The identified concepts concern the type of legislation, and concepts about the identified information from the legal rules. Besides, it models the relation that may interlink the legal rules with each other.
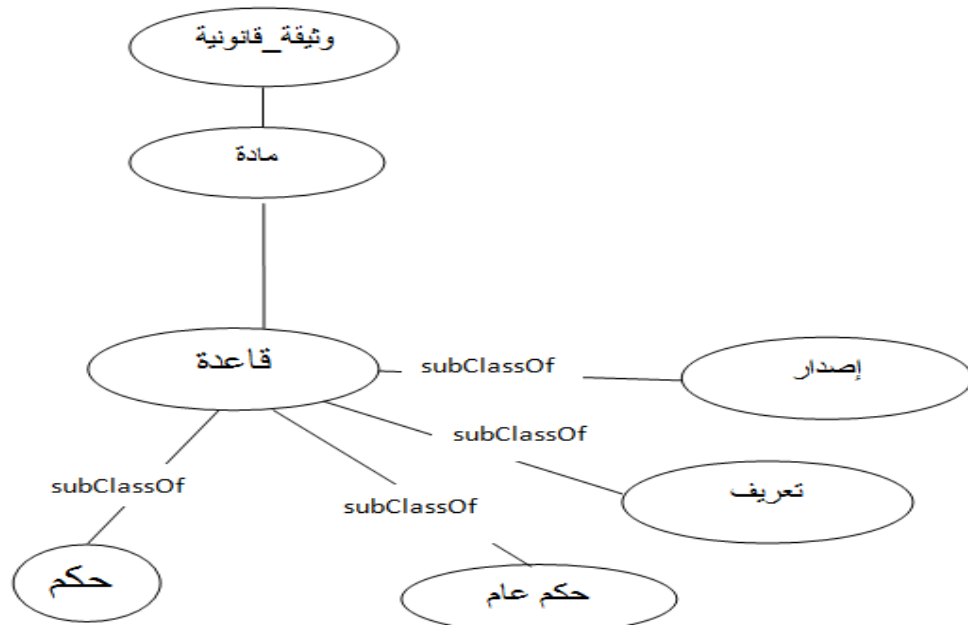
Fig.3. 12 Legal rule classes.

The Fig.3.12 represents the classes of legal rules (types of legal rules), where here we have depended on the organization of the different legal rules. In the following we provide a description of the main concepts used.

قاعدة : Legal Rule expresses the norm or the regulation that is a part of Article, the latter can include one rule or more. The legal rule may has different categories or types which are "حكم", "إصدار", "تعريف" and "حكم عام".

The provisions are considered as the main legal rules, where the most of legal rules are provisions. The provision includes the types: obligation "إلزام", permission "جواز", prohibition "حظر" and right "حق". They are represented as subclasses of the concept provision as illustrated in Fig.3.13.
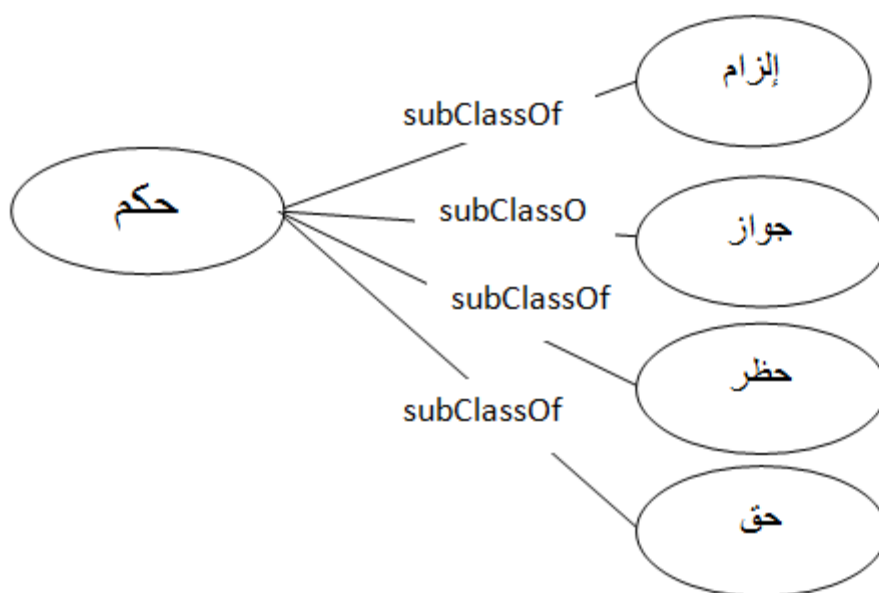
Fig.3.13 Types of provision concept.

There are a set of relation that may interlink the legal rules with each other. The relation that interconnect the modificatory rules with each other (e.g. addition, modification and deletion); that rules can make a network of relations between the different legal rules. The relation between the roles, interlink the legal rules with each other; where it concerns the hierarchy between the different roles of the organization and the process of the activities.

The provision includes a set of element (arguments) which are Role **"دور"**, Action " فعل",
Case description **"وصف-الحالة"**, Exception "إستثناء", Condition "شرط". In addition of these
elements we represented the arguments **"مستفيد"** Beneficiary, "نشاط" Activity and "مدة"
Periode. In the following we describe the different concepts. Fig.3.14 illustrates the main
concepts at the level of the provision and the different relations that interlink the different
concepts.

Fig.3.14 Semantic representation of Arabic legal rules.

- **مستفيد** : Beneficiary expresses the Agent that is the receiver of the action at the
  level of the provision.

- **فعل** : Action in this ontology represents the part that addresses the action at the level of the provision. The action is performed by an agent where it designates the act attributed to the Role.

- **نشاط**: Activity concept represents a set of actions that can be performed by one agent or more. In another word the Activity is the process that is performed by the execution of a set of action by one Role or more.

- **إستثناء**: Exception represents what is excluded.

- **شرط**: Condition expresses that the Action happens if a particular conditions exist.

- **وصف-الحالة**: Case description addresses when the rule is applicable.

- **مدة**: Period represents the duration required for the execution of the legal rule in another word it is the duration that must be respected in order to fulfill the obligation.

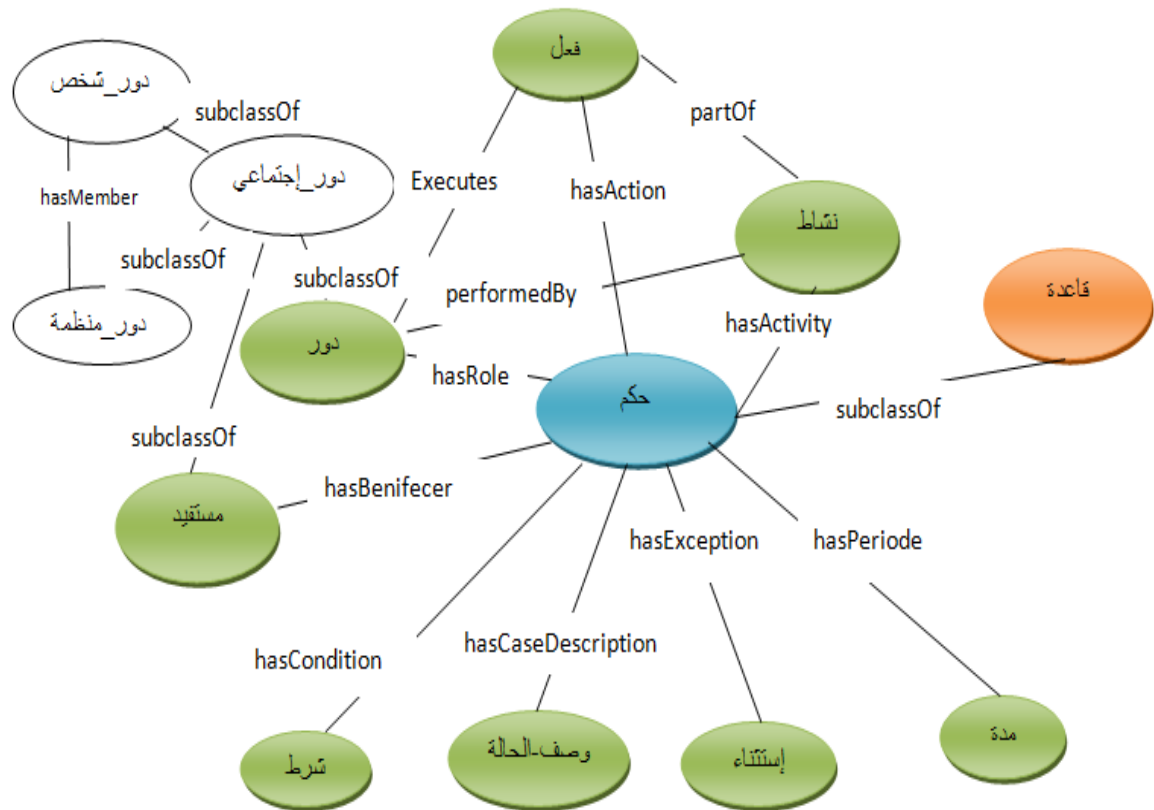- **دور**: Role represents the legal role at the level of the provision, "Indeed, roles and actions are closely related concepts: a role defines some set of actions that can be performed by an agent, but is conversely defined by those actions" (Hoekstra, 2009). The role represents the actions that can be played by agents (Hoekstra, 2009). The concept Role includes the concept social role. The latter includes the two concepts Person Role (دور_شخص) and Organization Role (دور_منظمة) which is a set of person role or a set of organization role; the organization role is a group that can be represented as one. Each role can perform one Action or more.

In this chapter we provide a description of the phases followed in the methodology starting by the identification of semantic information and ending by the ontology that allows storing information concerning the structure and the semantic information of legal texts. We provided a model that helps to facilitate the access and research of the relevant information. In the next chapter we will describe the first phase in detail, where we will show how we built our corpus and we present the results and the evaluations over the constructed corpus of Algerian Arabic legal texts and the built ontology.

Chapter 04

Results & Discussion

4.   Results & discussion

In this chapter we first illustrate the steps followed for the construction of the corpus and the different   results   obtained and then we present the evaluation of the semantic annotation of information (provision type and the Arguments Role and Action).Next we present the evaluation of the built ontology.

4.1.   Semantic annotation

  We build our corpus of regulatory text written in Arabic language from the official portal, which is the main source of the legislation on the Web in Algeria. It publishes all legislation texts (law, decree, order, etc.). We dealt with the regulation that covers the university, which is a public institution of administrative nature. The size of the constructed corpus is about (180 Ko).

We developed our system with Java language under NetBeans environment. We applied an algorithm for the segmentation of documents to articles and the latter into sentences. Taking into account the complexity of the segmentation and limits of time, we added a manual segmentation or correction to the results obtained from the segmentation phase. After that, we called the rules created for the semantic annotation; the number of the created rules for the identification of provision type is eighteen (18).

With the help of an expert of legal domain, and with the study of more than 4000 Articles that cover variant topics, we have arrived for the collection of information that concerns the three types: obligation, permission, and prohibition. We have defined lists of indicators and clues that cover several domains; the total number of indicators is fifty-seven (57) and the total number of clues is thirty-five (35), while we have collected the roles that cover the domain of the university; the number of collected roles is six hundred and three (603). Non exhaustive list of roles is illustrated in Table.4.1.

Role:
الطالب
الأستاذ
الموظف
العميد

العون

المترشح

المشرف

رئيس المؤسسة

المجلس العلمي

المجلس البيداغوجي

الهيئات الإدارية المعنية

المصالح الإدارية

الإدارة

Table.4. 1 : Sample of list of roles.

To evaluate how our system performs, we have examined it, using the test corpus that contains 778 provisions of various documents.

	A	B	C	D	E
Obligation	308	315	311	309	326
Permission	311	322	317	319	308
Prohibition	119	121	117	105	118
Do not belong	40	20	33	45	26
Total number	778				

Table.4. 2: The results of the manually annotated corpus.

We have prepared a manually annotated corpus, annotated by three law experts (Annotator A, Annotator B, and Annotator C) and two linguists (Annotator D and Annotator E). They associated the corresponding type to each sentence in the documents with the specification whether the sentence belongs to the types (obligation, permission, and prohibition) or does not belong to the selected categories. The result is illustrated in Table.4. 2.

75

The annotators have annotated some provisions differently, and this is due to several reasons, including:

- Some annotators have classified the rights as obligations, and other annotators have classified them as permissions.

"المادة 38: للموظف الحق في التكوين وتحسين المستوى والترقية في الرتبة خلال حياته المهنية"[17]

"المادة 167: يحق للموظف الذي تعرض لإجراء تأديبي أن يبلغ بالأخطاء المنسوبة إليه وأن يطلع على كامل ملفه التأديبي في أجل خمسة عشر (15) يوما ابتداء من تحريك الدعوى التأديبية"[18].

- Some indicators were categorized differently by the annotators as the indicator: "ينبغي", where some of the annotators have annotated it as an obligation and other annotators have annotated it as permission. Example:

"المادة 39: ينبغي أن يدمج المترشح الذي يحضر مذكرة شهادة الماجستير في مجموعة أو فريق بحث ..."[19]

Here where are some provisions that have been annotated differently by the annotators:

- "كل مترشح ناجح نهائيا لم يلتحق بمنصب عمله في أجل خمسة عشرة (15) يوما ابتداء من تاريخ تبليغه كتابيا بنجاحه يفقد حق التوظيف .."[20].
- "...و ينطبق هذا المنع أيضا على الأشخاص المحرومين من حقوقهم المدنية."[21]
- ..."وكذا أسماء و ألقاب و مؤهلات الأساتذة أو الباحثين الذين بإمكانهم المشاركة في تأطير التكوين المنشود." [10]

---

[17] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

[18] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

[19] Executive Decree No. 98-254 of 24 Rabie Ethani 1419 corresponding to 17 August 1998 on doctoral formation about the postgraduate specialized and the academic empowerment.

[20] Decree of 30 Rabie El Aouel 1429 corresponding to April 7, 2008 specifying the composition of the administrative file, the organizational arrangements for the recruitment of contractual agents and the advertising procedure

21Order of 14 Joumada Ethania 1429 corresponding to June 18, 2008 fixing the specifications to issue the authorization to create a private institution of higher education.

A comparison was made between the different annotations. Concerning the provisions that are labeled in conflict; we chose the highest number indicated by the annotators, and we eliminated the provisions that cannot be selected as distinct.

Evaluation of provision type annotation

To evaluate the results, we measured the precision and the recall of the annotation of the provision types and the arguments: role and action. Precision and recall are measured corresponding to Goutte and Gaussier (2005) rule:

TP: True Positive.

FP: False Positive.

FP: False Negative.

$$ Precesion = \frac{TP}{TP + FP} $$

$$ Recall = \frac{TP}{TP + FN} $$

We also measured the F-score. The results are illustrated in Table.4. 3.

	Provision	Precision	Recall	F-Score
Obligation	305	0.96	0.98	0.97
Permission	318	0.93	0.95	0.94
Prohibition	119	0.97	0.97	0.97

Table.4. 3: The results of the annotation of provision types.

As shown in Table.4.4, a comparison was made with a rule-based approach adopted by Zeni et al.(2015). Which they depend on a set of heuristic rules for the annotation of the provisions according to the following categories: obligation, anti-obligation, right, anti-right. The GaiusT collected 140 rights and obligations. They have achieved high result

100% on precision and 33% on recall of right annotation, 76% on precision and 42% on recall of obligation annotation, and 100% on precision, 50% on recall of anti-obligation.

Dataset		GaiusT		SALEM	Our System
		140		49	742
Obligation	Precision	76%		95%	96%
	Recall	42%		95%	98%
Permission	Precision	right	100%	83%	93%
	Recall	right	33%	100%	95%
prohibition	Precision	anti-obligation	100%	93%	97%
	Recall	anti-obligation	50%	93%	97%

Table.4. 4: Comparison with evaluations of GaiusT and SALEM.

Another comparison was made with a pure machine learning approach adopted by SALEM (Soria et al.(2007)) which worked for the classification of provisions. SALEM experiments were conducted over a dataset made of 473 provisions where only 49 provisions from them concern the obligation, permission and prohibition types. They have achieved 95% on precision and 95% on recall of obligation classification, 83% on precision and one on recall of permission classification, and 93% on precision and 93% on recall of prohibition classification.

We conducted our experiments on provisions of a total number greater than Zeni et al.(2015) and Soria et al.(2007). We presented a more granular study on the Arabic provisions that allowed us to perform a better categorization. We have achieved promising results with the semantic annotation of the provision types. Our research has achieved better results in comparison with other researchers.

	precision	recall
Role	0.71	0.81
Action	0.81	0.87

Table.4. 5 The results of the annotation of the arguments.

Evaluation of role annotation

As illustrated in Fig. 4.5, in the annotation of the argument role, the experimental result got 0.71 on precision and 0.81 on recall. We have faced problems related to the agglutination. In addition, some roles do not occur explicitly in the provisions where it refers to the role mentioned in the titles or in the previous provisions.

Evaluation of action annotation

Finally, as illustrated in Fig.4.5, the experimental result got 0.81 on precision and 0.87 on recall of action annotation. In this study, we consider the action all the part that exists after the clues indicating the existence of the action according to the indicator. In the case of the action argument clues are absent then it is difficult to specify the action. In this case, we assume that the action is the part that follows the role (if the role exists in the sentence) or the part that follows the indicator.

Some of wrongly classified provisions

Here we mention some of the wrongly annotated provisions:

1. When the indicator is preceded by a definition clause, our system categorizes the sentence following the indicator, however, in this case, the definition takes the priority and the indicator is ignored.

   "وضعية خارج الإطار هي الحالة التي يمكن أن يوضع فيها الموظف بطلب منه، بعـد استنفاد حقوقـه في الانتداب، في إطار أحكام المادة 135 أعلاه، في وظيفة لايحكمها هذا القانون الأساسي"[22]

2. In some cases, when the indicator exists in the argument Case description or the argument Condition, our system categorizes the sentence. However, the

---

[22] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service

indicator does not take the meaning of the indicators addressed in our research. Example: in the following provisions, the indicator "يمكن" does not take the meaning of permission.

"المادة 65: لكل عون متعاقد صدر في حقه إجراء تأديبي **يمكن** أن يترتب عليه فسخ عقده,الحق في الإطلاع على ملفه التأديبي."[23]

"المادة 173: في حالة إرتكاب الموظف خطأ جسيما، يمكن أن يؤدي إلى عقوبة من الدرجة الرابعة، تقوم السلطة التي لها صلاحيات التعيين بتوقيفه عن مهامه فورا."[24]

3. In the case of the role in the provision is addressed as a subject pronoun like (She/He) and the previous provision contains the role, but their type is not addressed in our research thesis. Here only the subject pronoun is detected, without the assignment of a particular role to the subject pronoun.

4.2.   Ontology construction

Semantic web is based on machine understandable languages: Extensible Markup Language (XML), Resource Description Framework (RDF), RDF Schema (RDFS) and Ontology Web Language (OWL). These technologies permit the interpretation and the processing of information. XML is used to structure the data in the web. The "RDF is typically applied to provide metadata about a resource on the web using a simple data model" (Neumann, 2015). The RDF is used to describe resources in the form of triples: subject, predicate and object (Barba-González, et al., 2019). RDF schema is a semantic extension of RDF (Cardoso, 2006). **"**RDF Schemas serve to define the relations between resources of the RDF documents" (Kumar and Dwivedi , 2011)**.**

OWL is considered the most prominent language for defining ontology in the Semantic Web.
 "OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web" (McGuinness and Van Harmelen, 2004). OWL provides a richer language for describing terms in vocabularies and the relationships between them.

---

[23]Presidential Decree No. 07-308 of 17 Ramadhan 1428 corresponding to September 29, 2007 setting the procedures for the recruitment of contractual agents their rights and obligations, the components of their remuneration, the rules relating to their management and the disciplinary regime that is theirs relevant

[24] Order No. 06-03 of 19 Joumada Ethania 1427 Corresponding to 15 July 2006 concerning the general status of the public service.

(Cardoso, 2006). It permits to provide constraints on the classes and the properties. (Cardoso, 2006). Where, in addition of describing concepts, OWL enables to add different logical operators for describing the classes and the properties. Such as intersection, union, equivalent, disjoint, etc.

Ontology editor is built in order to assist in the creation of ontologies. Several ontology editors are existed such as Protégé[25], NeOn Toolkit[26], Knoodl[27], and OWLGrEd[28], etc. Ontology editing environment that we used to build our ontology is Protégé. Protégé is a popular development editor. We used **Protege_4.3** (protege.stanford.edu). It is a free, open-source ontology editor and framework developed by Standford University for building intelligent systems.
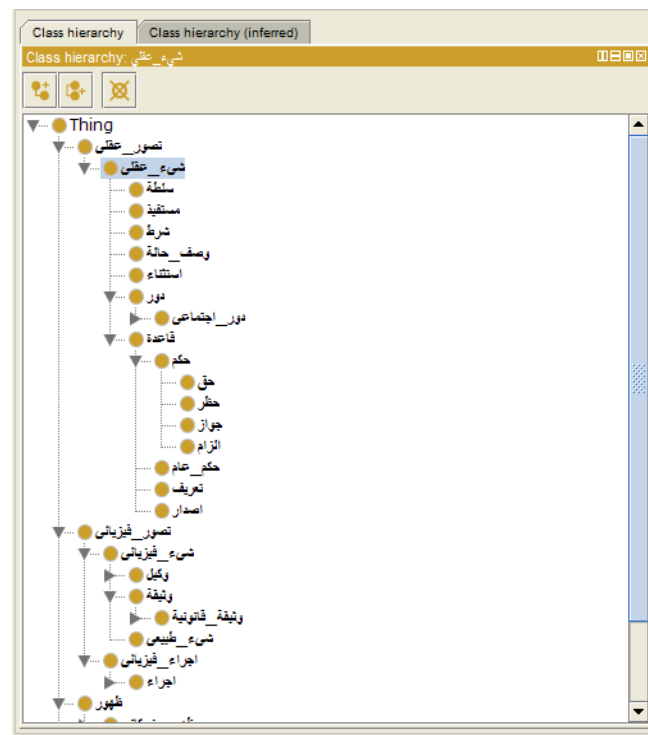
Fig.4.1 Arabic ontology model.

[25] Protégé. **Available at:** http://protege.stanford.edu/ (accessed on 10 Mars 2018)

[26] NeOn Toolkit. **Available at:** http://neon-toolkit.org/ (accessed on 10 Mars 2018)

[27] Knoodl. **Available at:** http://www.knoodl.com/ (accessed on 10 Mars 2018)

[28] OWLGrEd. **Available at:** http://owlgred.lumii.lv/ (accessed on 10 Mars 2018)

The proposed ontology does not cover all the Arabic legal domain entities. Where, we make an effort to build an ontology that contains some of the important concepts. The Fig.4.1 illustrates the developed legal Arabic Algerian ontology.

We have represented several concepts. In our research, in the step of semantic annotation we have annotated the provision and we have identified the elements role and action. The constructed ontology covers several concepts in addition of the above mentioned ones.

Concerning the Role, as we have seen in the previous chapter, the Role may include person role and organization role. We enriched our ontology by the different agents. The Table.4.6 illustrates the instances of Role. Where, we have populated all the roles which are considered the individuals or the instances of the predefined class Role. For example (الطالب) is an individual of the class (دور_شخص).

دور / Class Role	
دور_شخص / Class Person_Role	دور_منظمة / Class Organization_ Role
المدير	الجامعة
العميد	الادارة
نائب العميد	الكلية
الاستاذ	القسم
الموظف	اللجنة العلمية
الطالب	المجلس العلمي

Table.4. 6: Presentation of class (Role) and their individuals.

Querying the system

In this phase, our aim is querying the ontology. After the population of the extracted information in the corresponding classes, we enrich the other concepts of the ontology, after that we can retrieve the relevant information.

In order to retrieve the data we use specific query languages. Where, machine understandable languages such as RDF, OWL, etc. rely on specific protocols (e.g.

SPARQL, SWRL, etc.). In our research, we used the protocol SPARQL for the aim of querying the built ontology.

The queries enable users to find out for example:

- ✓ What are the different rules that concern a particular role?
- ✓ Or what are the actions of a particular role?
- ✓ Or what are the roles addressed in a particular legal document?
- ✓ Or what are the provisions that concern a particular role?
- ✓ Or what are the arguments of a particular rule?
- ✓ etc.

In our research we used the Apache Jena Fueski Server which permits the access into data by the use of SPARQL protocol over HTTP.

The Table 4.7 shows a sample of the SPARQL query language. In this query we are interested to know: What are the provisions that concern the Role الموظف؟

SPARQL	`SELECT ?s`
	`        WHERE { ?s AO:hasRole ?o.`
	`                ?o  AO:roleName "الـموظف "`
	
	`        }`

Table.4. 7: Sample of SPARQL query language.

The result of the query

The result shows a set of rule. Here we present some of them.

> يجب على الموظف، في إطار تأدية مهامه، احترام سلطة الدولة وفـرض احترامها وفقا للقوانين والتنظيمات المعمول بها
>
> يجب على الموظف أن يمارس مهامه بكل أمانـة وبدون تحيز
>
> يجب على الموظف تجنب كل فعل يتنافى مع طبيعة مهامه ولو كان ذلك خارج الخدمة

كما يجب عليه أن يتسم في كل الأحوال بسلوك لائق ومحترم

وفي هذه الحالة، لا يمكن الموظف ذكر صفته أو رتبته الإدارية بمناسبة نشر هذه الأعمال، إلا بعد موافقة السلطة التي لها صلاحيات التعين

يمنع على كل موظف، مهما كانت وضعيته في السلم الإداري، أن يمتلك داخل التراب الوطني أو خارجه، مباشرة أو بواسطة شخص آخر، بأية صفة من الصفات، مصالح من طبيعتها أن تؤثر على استقلاليته أو تشكل عائقا للقيام بمهمته بصفة عادية في مؤسسة تخضع إلى رقابة الإدارة التي ينتمي إليها أو لها صلة مع هذه الإدارة، وذلك تحت طائلة تعرضه للعقوبات التأديبية المنصوص عليها في هذا القانون الأساسي

إذا كان زوج الموظف يمارس، بصفة مهنية، نشاطا خاصا مربحا، وجب على الموظف التصريح بذلك للإدارة التي ينتمي إليها

يجب على الموظف الالتزام بالسر المهني

على الموظف أن يسهر على حماية الوثائق الإدارية وعلى أمنها

يجب على الموظف، ألا يستعمل، بأية حال، لأغراض شخصية أو لأغراض خارجة عن المصلحة، المحلات والتجهيزات ووسائل الإدارة

يجب على الموظف التعامل بأدب واحترام في علاقاته مع رؤسائه وزملائه ومرؤوسيه

يجب على الموظف التعامل مع مستعملي المرفق العام بلياقة ودون مماطلة

يمنع على الموظف تحت طائلة المتابعات الجزائية، طلب أو اشتراط أو استلام، هدايا أو هبات أو أية امتيازات من أي نوع كانت، بطريقة مباشرة أو بواسطة شخص آخر، مقابل تأدية خدمة في إطار مهامه

يمكن الموظف الذي قام أثناء تأديته مهامه بعمل شجاع مثبت قانونا أو قام بمجهودات استثنائية ساهمت في تحسين أداء المصلحة، أن يستفيد من أوسمة شرفية و/أو مكافآت، بعد استشارة لجنة خاصة تنشأ لدى السلطة الوزارية المختصة.

4.3. Concluding Remarks

Concerning the analysis phase carried out by the linguists and domain specialists; there are several categories of provisions and different indicators in the Arabic legal

domain. However, there is no defined categorization on these indicators, where we have encountered difficulties in their identification and categorization because some indicators are difficult to be distinguished by the domain specialists themselves. Some indicators are categorized differently by the expert annotators; one of the causes is the ambiguous nature of the legal domain.

We have integrated the right category in the obligation category; some annotators have classified the rights as an obligation, and other annotators have classified them as permissions; hence we conclude that it is better to define it as a category of provision.

We conclude that the definition of arguments contributes to the definition of the provision type. Furthermore, the identification of other types of provisions may contribute to a better annotation of the provisions.

Conclusion

Conclusion

This thesis deals with the identification of the semantic information of legal rules. The difficulties encountered when dealing with legislation motivate us as to do this research. The mastery of these texts plays crucially in the effective management of organization, in particular the public organizations. Due to the increasing number of regulations, accessing and retrieving the relative regulation has become a laboring task especially by non domain specialist, in addition the complex nature of legislation (contradictory, ambiguous, variety of legal sources, dynamic character, etc.) lead the actor to prefer to rely on the expert rather than resort to the search for legal rules in a collection of documents. Therefore, several research have tendency to process the information existed in legal texts.

The extraction of information from documents is an important task in several fields. However, the accuracy in the extraction of information is more important in some fields such as the legal domain. Improvement of methods and tools for identifying semantic information from legal texts plays a considerable role for both organizations and citizens.

In this study we are interested in the specification and semantic annotation of Algerian legal provisions written in Arabic language. We focused on this study on the provisions types: obligation, permission and prohibition, and their arguments role and action. In order to treat legal texts we created a corpus that concerns Algerian legal documents that concerns the university.

For the aim of specifying semantic information, we have argued that it is so important to disambiguate terms; thus we explored the contexts of the indicators. We have addressed the problems faced when dealing with Arabic legal domain. We presented a rule-based approach based on the contextual exploration method. We categorized the indicators into three classes that are "Indicator sufficient", "The indicator needs a clue" and "The indicator needs a role". The process takes the following steps: we segment the text into articles, then article into sentences, if the latter to segment is containing the indicator; then we call the appropriate class. Finally, if all

the rule conditions are satisfied, then we attribute the appropriate annotations to the sentence. In our view, the proposed solution is more precise than the previous research in the specification of semantic information. Experiments have shown that the use of clues occurring in the context of the rule in addition to the indicator, achieve better results on the detection of the provision types.

We built a legal Arabic ontology that models the Algerian Legal texts written in Arabic language which concerns the university. The constructed ontology permits storing information concerning the structure and the semantic information of legal texts for the aim of providing a model that helps to facilitate the access and research for relevant information. After the creation of the ontology, the output of the semantic annotation process is populated in the ontology.

In further research, we aim to extend our work to cover more provisions. We also aim to specify the other elements existing in the provisions like the case description, exception, condition…etc. We aim also for the definition (segmentation) of Arabic legal documents structure.

References

Al-Kharashi, I. A., Al-Sughaiyer, I. A. (2004) 'Arabic morphological analysis techniques: A comprehensive survey'. *Journal of the American Society for Information Science and Technology,* Vol. 55, No.3, pp.189–213.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

Alrahabi M. (2015) E-Quotes: Enunciative Modalities Analysis Tool for Direct Reported Speech in Arabic. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9041. Springer, Cham

Alrahabi, M., Ibrahim, A. H., Descles, J. P. (2006) 'Semantic Annotation of Reported Information in Arabic'. *Proceedings of FLAIRS 06*, Melbourne Beach, Florida, 11-13 May, pp. 263-268, aaai press.

Amin, A. (1998) 'Off-line Arabic character recognition: the state of the art'. *Pattern Recognition,* Vol. 31, No. 5, pp. 517–530.

Anusuya , M.A. ,Katti, M.A., (2009). 'Speech Recognition by Machine: A Review'. International Journal of Computer Science and Information Security. Vol. 06, No.03, pp.181–205.

Athan, T., Governatori, G., Palmirani, M., Paschke, A., & Wyner, A. (2015, July). LegalRuleML: Design principles and foundations. In Reasoning Web International Summer School(pp. 151-188). Springer, Cham.

Barba-González, C., García-Nieto, J., del Mar Roldán-García, M., Navas-Delgado, I., Nebro, A. J., & Aldana-Montes, J. F. (2019). BIGOWL: Knowledge centered Big Data analytics. Expert Systems with Applications, 115, 543-556.

Bartolini R., Lenci A., Montemagni S., Pirrelli V. (2002) 'Grammar and lexicon in the robust parsing of Italian: Towards a non-naıve interplay'. *Proceedings of* COLING-GEE, Taipei, 1st September, pp.1-7.

Bartolini, R., Lenci, A., Simonetta Montemagni, S., Pirrelli, V., Soria, C. (2004) 'Semantic mark-up of italian legal texts through nlp-based techniques'. *Proceedings of LREC 04*, 26-28 May, Lisbon, pp. 795–798.

Berrazega, I. ,Faiz, R. ,Bouhafs, A, Mourad, G. (2016) 'A Knowledge-Based Approach for Provisions' Categorization in Arabic Normative Texts'. *Proceedings of CSOC 16*, pp. 415-425, Springer International Publishing.

Bertin, M., Desclés, J. P., & Hacène, T. Comment extraire des définitions des textes?

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., Soria C. (2005) 'Automatic semantics extraction in law documents'. *Proceedings of ICAIL 05*,Bologna, 6-11 June, pp. 133–140, ACM, New York.

Biagioli, C., Grossi, D. (2008). Formal aspects of legislative meta-drafting. In Legal Knowledge and Information Systems: JURIX 2008: the Twentieth First Annual Conference (Vol. 21, p. 192). IOS Press.

Bikel, D., & Zitouni, I. (2012). Multilingual natural language processing applications: from theory to practice. IBM Press.

Biskri, I., Berrakem, F. Z., & Jebali, A. (2017). The Applicative Combinatory Categorial Analysis of Arabic. *Proceedings of ACLing 2017*, 05-06 November, pp. 199-207, Elsevier.

Boella, G., Caro, L.D., Humphreys, L. et al. (2016) 'Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law'. *Artificial Intelligence and Law,*Vol. 24, No. 3, pp. 245–283.

Boella, G., Di Caro, L., Humphreys, L., Robaldo, L., Rossi, P., & van der Torre, L. (2016). Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. Artificial Intelligence and Law, 24(3), 245-283.

Boella, G., Humphreys, L., & van der Torre, L. (2013). The Role of Roles in Eunomos, a Legal Document and Knowledge Management System for Regulatory Compliance. In Organizational Change and Information Systems (pp. 451-459). Springer, Berlin, Heidelberg.

Boudelaa, S., Marslen-Wilson, WD. (2010) 'Aralex: a lexical database for Modern Standard Arabic'. *Behavior Research Methods*, Vol. 42, No. 2, pp. 481-487.

Bouhafs, A. (2005) Utilisation de la méthode d'exploration contextuelle pour une extraction d'informations sur le Web dédiées â la veille, réalisation du systéme informatique JAVA Veille, These de doctorat, Université Paris-Sorbonne.

Bousmaha, K.Z , Charef_Abdoun,S., Hadrich_Belguith, L., Rahmouni, M.K. (2013) 'Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil'. *La revue de l'Information Scientifique et Technique (RIST),* Vol. 20, No. 2, pp.32-46.

Brank, J., Leban, G., & Grobelnik, M. (2018). Semantic Annotation of Documents Based on Wikipedia Concepts. Informatica, 42(1).

Breaux, T.D., Vail, M.W., Anton, A.I. (2006) 'Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations'. *Proceedings of RE 2006*, Minneapolis/St.Paul, Minnesota, 11-15 September, pp. 46-55, IEEE.

Breuker, J., Hoekstra, R., Boer, A., van den Berg, K., Sartor, G., Rubino, R., ... & Bench-Capon, T. (2007). OWL Ontology of Basic Legal Concepts (LKIF-Core), Deliverable 1.4. Information Society Technologies.

Buey, M. G., Garrido, A. L., Bobed, C., & Ilarri, S. (2016, February). The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies. In ICAART (2) (pp. 438-445).

Bui, T. D., Nguyen, S. T., & Ho, Q. B. (2014, November). Towards a Conceptual Search for Vietnamese Legal Text. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 175-185). Springer, Berlin, Heidelberg.

Bui, T. D., Nguyen, S. T., & Ho, Q. B. (2015, September). Towards a Conceptual Search for Vietnamese Legal Text. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 175-185). Springer, Berlin, Heidelberg.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), 48-57.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. IEEE Intelligent Systems, 32(6), 74-80.

Cardellino, C., Teruel, M., Alemany, L. A., & Villata, S. (2017, May). Learning Slowly To Learn Better: Curriculum Learning for Legal Ontology Population. In The Thirtieth International Flairs Conference.

Cardoso, J. (2006). Developing an owl ontology for e-tourism. In Semantic web services, processes and applications (pp. 247-282). Springer, Boston, MA.

Casellas, N. (2011). Legal ontology engineering: Methodologies, modelling trends, and the ontology of professional judicial knowledge (Vol. 3). Springer Science & Business Media.

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

Ciravegna, F., Dingli, A., Petrelli, D., & Wilks, Y. (2002, October). User-system cooperation in document annotation based on information extraction. In International Conference on Knowledge Engineering and Knowledge Management (pp. 122-137). Springer, Berlin, Heidelberg.

Cunningham, H. (2002). GATE, a general architecture for text engineering. Computers and the Humanities, 36(2), 223-254

Cunningham, H., Tablan, V., Roberts, I., Greenwood, M. A., & Aswani, N. (2011). Information extraction and semantic annotation for multi-paradigm information management. In Current Challenges in Patent Information Retrieval (pp. 307-327). Springer Berlin Heidelberg.

De Maat, E., Krabben, K., & Winkels, R. (2010, December). Machine Learning versus Knowledge Based Classification of Legal Texts. In JURIX (pp. 87-96).

Descles, J. P. (1997) 'Systèmes d'exploration contextuelle'. *Cotexte et calcul du sens, Presses universitaires de Caen*, pp. 215-232.

Descles, J., Alrahabi, M., & Descles, J. P. (2009). BioExcom: Automatic Annotation and categorization of speculative sentences in biological literature by a Contextual Exploration processing. *Proceedings of the 4th Language & Technology Conference*, pp. 32-40.

Descles, J.P. (2006) 'Contextual Exploration Processing for Discourse and Automatic Annotations of Texts'.*Proceedings of FLAIRS 06*, Melbourne Beach, Florida, 11-13 May, pp. 281-284, aaai press.

Dhouib, K., & Gargouri, F. (2015). An applied legal ontology in Arabic for the jurisprudence decision-structuring. International Journal of Knowledge Society Research (IJKSR), 6(1), 43-54.

Douzidia, F., G. Lapalme. (2004) 'Lakhas, an Arabic summarization system'. *Proceedings of DUC 04,* pp. 128–135, NIST, Boston.

El Kassas, D., Kahane, S. (2004) 'Modélisation de l'ordre des mots en arabe standard'. Atelier sur le Traitement Automatique de la Langue Arabe Ecrite et Parlée, pp.259-264.

Elkhlifi, A. and Faiz, R. (2009) 'Automatic annotation approach of events in news articles', International Journal of Computing & Information Sciences, Vol. 7, No. 1, pp.40–50.

Fadili, H., & Jouis, C. (2016). Towards an automatic analyze and standardization of unstructured data in the context of big and linked data. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, pp. 223-230. ACM.

Farghaly, A., Shaalan, K. (2009) 'Arabic Natural Language Processing: Challenges and Solutions'. *ACM Transactions on Asian Language Information Processing (TALIP),* Vol. 8, No. 4 ,pp.14.

Firth, J. R. (1935). The Technique of Semantics. Transactions of the philological society, 34(1), 36-73.

Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (2010) *'Semantic Processing of Legal Texts: where the language of law meets the law of language*', Springer, Heidelberg.

Francesconi, E., Passerini, A. (2007) 'Automatic Classification of Provisions in Legislative Texts'. *Artificial Intelligence and Law,* Vol. 15, No. 1, pp.1-17.

Franssen, M. (2007) 'Automated Detection of Norm Sentences in Laws'. *Proceedings of the Twente Student Conference on IT*.

Garcia-Flores, J. (2007) Annotation sémantique des spécifications informatiques de besoins par la méthode d'Exploration Contextuelle: une contribution des méthodes linguistiques aux conceptions de logiciels, These de doctorat, Université ParisSorbonne.

Gordon, T. F., Governatori, G., & Rotolo, A. (2009, November). Rules and norms: Requirements for rule interchange languages in the legal domain. In International Workshop on Rules and Rule Markup Languages for the Semantic Web (pp. 282-296). Springer, Berlin, Heidelberg.

Goutte, C., Gaussier, E. (2005) 'A probabilistic interpretation of precision, recall and f-score, with implication for evaluation'. *In: Proceedings of the 27th European Conference on Advances in Information Retrieval Research (ECIR),* Santiago de Compostela, Spain, pp. 345–359.

Gruber, Thomas R. "A translation approach to portable ontology specifications." Knowledge acquisition 5.2 (1993): 199-220.

Harmanani, H. M., Keirouz, W., Raheel, S. (2006) 'A Rule-Based Extensible Stemmer for Information Retrieval with Application to Arabic'. *The international Arab journal of Information Technology,* Vol. *3*, No.3 , pp. 265-272.

Hashmi, M. (2015) 'A Methodology for Extracting Legal Norms from Regulatory Documents'. *Proceedings of EDOCW 15*, Adelaide, 21-25, pp. 41 – 50 September, IEEE Computer Society.

Hindriks, F. (2009). Constitutive rules, language, and ontology. Erkenntnis, 71(2), 253-275.

Hoekstra, R. J. (2009). Ontology Representation : design patterns and ontologies that make sense Amsterdam: IOS Press

Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF Core Ontology of Basic Legal Concepts. LOAIT, 321, 43-63.

Hoekstra, R., Breuker, J., Di Bello, M., Boer, A. (2007) 'The LKIF Core Ontology of Basic Legal Concepts'. *Proceedings of LOAIT 07,*Stanford, California , pp. 43-63.

Jackiewicz, A. (1996) 'La notion de cause pour le filtrage de phrases importantes d'un texte'. *Natural Language Processing and Industrial Applications*, pp.136-141.

Jiang, J. (2012). Information extraction from text. In Mining text data (pp. 11-41). Springer, Boston, MA.

Jiang, J., Aldewereld, H., Dignum, V., Tan, Y. H. (2012). 'Norm Contextualization'. *Proceedings of COIN 12*, Valencia, pp. 141-157, Springer, Berlin.

Khorsheed, M. S. (2002) 'Off-line Arabic character recognition--a review'. Pattern analysis & applications, Vol. 5, No. 1, pp. 31-45.

Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J.R., Mylopoulos, J. (2006) 'Text mining through semi automatic semantic annotation'. *Proceedings of PAKM 06*,Vienna, 30 November – 1 December, pp. 143–154, Springer, Berlin.

Kiyavitskaya,N., Zeni, N., Breaux, T. D., Antón, A. I., Cordy, J. R., Mich, L., Mylopoulos, J. (2008) 'Automating the extraction of rights and obligations for regulatory compliance'. *Proceedings of ER 08*,Barcelona, 20-24 October, pp. 154–168, Springer, Berlin.

Kumar, A., & Dwivedi, S. K. (2011, December). Ontology exemplification for aSPOCMS in the Semantic Web. In 2011 World Congress on Information and Communication Technologies (pp. 473-478). IEEE.

Lauser, B., & Hotho, A. (2003, August). Automatic multi-label subject indexing in a multilingual environment. In International Conference on Theory and Practice of Digital Libraries (pp. 140-151). Springer, Berlin, Heidelberg.

Law, L. C. (2008). Automatic classification of sentences in dutch laws. In Legal Knowledge and Information Systems: JURIX 2008: the Twentieth First Annual Conference (Vol. 21, p. 207). IOS Press.

Le Pirol, F. 2011. 'Image and Text Mining Based on Contextual Exploration from Multiple Points of View'.*Proceedings of FLAIRS 11*. AAAI Press, California.

Le Priol, F. (2007) 'Automatic annotation of discourse and semantic relations supplemented by terminology extraction for domain ontology building and information retrieval', Proceedings of the Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida, pp.494–499.

Lesmo, L., Mazzei, A. , Palmirani, M. , Radicioni, D. P. (2013) 'TULSI: an NLP system for extracting legal modificatory provisions'. *Artificial Intelligence and Law,* Vol 21, No. 2, pp.139-172.

Liddy, E. D. (2001). Natural language processing

Maybury, M. T. (1995). Generating summaries from event data. Information Processing & Management, 31(5), 735-751.

McCarthy, N. (2018) The worlds most spoken languages, https://www.statista.com/chart/12868/the-worlds-most-spoken-languages/. mai 10, 2019

McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. W3C recommendation, 10(10), 2004.

Mezghanni, I. B., & Gargouri, F. (2017). CrimAr: A Criminal Arabic Ontology for a Benchmark Based Evaluation. Procedia Computer Science, 112, 653-662.

Minel, J. L. (2002) *Filtrage sémantique: du résumé automatique à la fouille de textes*. Hermès Science Publ, Paris.

Mourad, G. (2002) 'La segmentation de textes par Exploration Contextuelle automatique, presentation du module SegATex', Paper presented at Inscription Spatiale du Langage: structure et processus ISLsp, Toulouse.

Muslea, I. (1999, July). Extraction patterns for information extraction tasks: A survey. In The AAAI-99 workshop on machine learning for information extraction (Vol. 2, No. 2).

Neill, J. O., Buitelaar, P., Robin, C., & Brien, L. O. (2017, June). Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law (pp. 159-168). ACM

Neumann, F. (2015). Analyzing and Modeling Interdisciplinary Product Development: A Framework for the Analysis of Knowledge Characteristics and Design Support. Springer.

Otto, P. N., & Antón, A. I. (2007, October). Addressing legal requirements in requirements engineering. In Requirements Engineering Conference, 2007. RE'07. 15th IEEE International(pp. 5-14). IEEE.

Plisson, J., Lavrac, N., & Mladenić, D. (2004). A rule based approach to word lemmatization.

Rammal, M., Bahsoun, Z., Al Achkar Jabbour, M. (2015) 'Keyword extraction from Arabic legal texts'. *Interactive Technology and Smart Education,* Vol. 12, No. 1, pp.62-71.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Shaalan, K. (2010). 'Rule-based Approach in Arabic Natural Language Processing'. *The International Journal on Information and Communication Technologies,* Vol. 3, No. 3, pp.11-19.

Soria,C., Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. (2007). 'Automatic Extraction of Semantics in Law Documents'. *Proceedings of the V Legislative XML Workshop*, Florence, 14-16 June pp. 253–266, European Press Academic Publishing, Florence.

Spinosa, P., Giardiello, G., Cherubini, M., Marchi, S., Venturi, G., Montemagni, S. (2009). 'Nlp-based metadata extraction for legal text consolidation'. *Proceedings of ICAIL 09*, Barcelona, 08–12 June, pp. 40–49, ACM, New York.

Stevenson, M., & Wilks, Y. (2003). Word sense disambiguation. The Oxford Handbook of Comp. Linguistics, 249-265

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. Data & knowledge engineering, 25(1-2), 161-197.

Teissedre, C. (2007) La définition, Etude linguistique, Utilisation dans un systéme de Recherche d'Information, Comparaison avec le «Define» de Google, Master Informatique et Ingénierie de la langue pour la gestion de l'Information, Université de la Sorbonne, France.

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. Proceedings of InSciT, 4, 354-358

Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94). Association for Computational Linguistics.

Wu, L., Hoi, S. C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. IEEE Transactions on Image Processing, 19(7), 1908-1920.

Wyner, A., Peters, W. (2011) 'Legal Knowledge and Information Systems On Rule Extraction from Regulations'. *Proceedings of JURIX 11,* University of Vienna, 14-16 December, pp. 113-122, IOS Press.

Zaidi, S., Laskri, M. T., & Bechkoum, K. (2005, November). A cross-language information retrieval based on an Arabic ontology in the legal domain. In Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05) (pp. 86-91).

Zeni, N., Kiyavitskaya, N., Mich, L., Cordy, J. R., & Mylopoulos, J. (2015) 'GaiusT: supporting the extraction of rights and obligations for regulatory compliance'. *Requirements engineering*, Vol. 20,No. 1 , pp.1-22.

Zeni, N., Mich, L., & Mylopoulos, J. (2016, November). GaiusT 2.0: Evolution of a Framework for Annotating Legal Documents. In Research Conference on Metadata and Semantics Research(pp. 43-54). Springer, Cham.

Zeni, N., Mich, L., & Mylopoulos, J. (2017). Annotating legal documents with GaiusT 2.0. International Journal of Metadata, Semantics and Ontologies, 12(1), 47-58.

Zitouni, I. (2011). Book Reviews: Introduction to Arabic Natural Language Processing by Nizar Y. Habash. Computational Linguistics, 37(3).

جعفون, م. س. (2007). مدخل الى العلوم القانونية.الجزائر: دار هومة.

زعلاني, ع. ا. (2007). المدخل لدراسة القانون النظرية العامة للقانون.الجزائر: دار هومة.

دودين م. س. (2009). الترجمة القانونية والمصطلحات والنماذج والصياغة القانونية. مكتبة المجتمع العربي.

سمير عبد السيد تناغو (1984). النظرية العامة للقانون. منشأة المعارف.

صبره ع. م. م. (2008) مشروع تعزيز قدرات مجلس النواب العراقي. ورشة تدريبية حول الصياغة التشريعية. لأعضاء المجلس والمستشارين القانونيين.

السباعي ك. أ. محاضرات في الترجمة القانونية. الجمعية الدولية للمترجمين واللغويين العرب.