وزارة التعليم العالي والبحث العلمي

Centre Universitaire
Abdelhafid Boussouf - Mila

Abdelhafid Boussouf
University Center of Mila

المركز الجامعي

عبد الحفيظ بوالصوف – ميلة –

Institut Science et Technologie

Année 2022

Filière : Informatique

Thèse

Présentée

En vue de l'obtention du Diplôme de Master

Un Système d'Annotation des Rôles Sémantiques dans la Langue Arabe Basé sur une Méthode d'Intelligence Artificielle

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Sciences et Technologies de l'Information et de la Communication (STIC)

Par : ABBAS Kahina et AZIZ Zineb

Devant le Jury

Président :	TALAI Meriem	MAA	C.U. Mila
Examinateur :	GUEMRI Oualid	MAB	C.U. Mila
Rapporteur :	MEGUEHOUT Hamza	MCB	C.U. Mila

Certes, la science guide, dirige et sauve;
l'ignorance égare, trompe et ruine

Imâm Ali ibn Abi Talib

إهداء

الحمد لله الذي بنعمته تتم الصالحات الحمد لله الذي وفقنا وسدد خطانا إلى أن وصلنا وحققنا مبتغانا وبوصولنا هذا لا ننسى من كانوا سندا وعونا لنا في مشوارنا بعد الله عز وجل أهدي هذا العمل إلى من كانت سبب في وجودي وأعطتني من عمرها وزهرة شبابها حبا ودافعا لمستقبل أجمل، الغالية التي دائما نرى الحب والحنان في عينيها أمي الحبيبة وأغلي من سكنت قلبي، إلى من سعى وشقا لأنعم بالراحة والهناء الذي لم يبخل بشيء من أجل دفعي في طريق النجاح أبي العزيز.

إلى إخوتي الأعزاء وأخواتي الكريمات (عامر، ياسين، فيصل، راضي، شيماء، ملاك) أتمنى لكم مستقبًلا مشرقا، إلي من ساندني ووقف بجانبي دائما خطيبي العزيز' محسن 'حفضه الله، إلى من جمعتنا الصدفة لنكون سويا في هذا العمل صديقتي و زميلتي 'كاهينة' حفضها الله، إلى من يتجدد بهم الأمل وسرنا سويا نحو طريق النجاح صديقاتي وزميلاتي وخاصة (نهال، مها، أسماء، رميسة، ملاك، نسرين)، إلى أساتذتي الكرام الذين رافقونا في مشوارنا التعليمي ولم يبخلوا علينا بتوجيههم ونصائحهم من أجل إتمام هذا العمل ، إلى كل من حملهم قلبي ونسيهم قلمي، إلى كل من أمد لي يد العون والمساعدة ، شكرا لكم جميعا.

راجية من المولى عزوجل أن يوفقنا لما يحب ويرضى.

زينب

إهداء

أهدي هذا العمل إلى من قال فيهما الرب: ' وَاخْفِضْ لَهُمَا جَنَاحَ الذُّلِّ مِنَ الرَّحْمَةِ وَقُل رَّبِّ ارْحَمْهُمَا كَمَا رَبَّيَانِي صَغِيرًا'.

إلى أمي الغالية التي وقفت بجانبي طوال مشواري الدراسي وكرست ليلها ونهارها لأبلغ العلى، إلى أبي الغالي الذي تعب وشقا من أجلي منذ نعومة أظافري، هذا العمل أنجز بفضلكما بعد الله، إلى إخوتي وأخواتي الأعزاء (لمياء، مهدي، يونس، فريدة، يزيد، سمير) الدين وقفو بجانبي وشجعوني أتمنى لكم التوفيق حيث ما كنتم ، إلى زوجي الغالي 'بلال' ادامك الله تاجا فوق رأسي فقد سهرت معي وساندتني، إلى من تقاسمت معهم الأفراح والأحزان وشققنا طريق النجاح صديقاتي العزيزات اللواتي شاركنني مشواري الدراسي (إيمان، أمينة، منال، أسيا، رميسة، خولة)،واخص بالذكر زميلتي وصديقتي 'زينب' التي تقاسمت وتشاركت معها هذا العمل.

إلى كل أساتذتي الذين علموني الأخلاق قبل كل حرف وإلى زملائي وزميلاتي دفعة 2017.

كاهينة

شكر وعرفان

نشكر في المقام الاول الله تعالى على فضله ونعمه التي أنعم بها علينا ومنحنا القوة والإرادة والصبر للقيام بهذا العمل المتواضع. فإن أصبنا فيه فمن الله وإن أخطأنا فمن أنفسينا ومن الشيطان.

ونتقدم بجزيل الشكر والتقدير لصاحب الفضل في إنجاز هذا العمل الأستاذ مقحوت حمزة على توجيهاته ونصائحه القيمة التي ساعدتنا كثيرًا ومكنتنا من تحسين جودة العمل. مهما قدمنا من كلمات شكر لا نستطيع أن نوافيك حقك، بارك الله فيك.

كما نتقدم بالشكر الجزيل لأعضاء لجنة المناقشة قمري وليد مناقشا وطلعي مريم رئيسا على قبولهما مناقشة هذا العمل.

ويمتد شكرنا إلى أهلنا الأعزاء وخاصة والدينا وإخواننا وأخواتنا على دعمهم المعنوي والمادي.

في الاخير، نود أن نشكر جميع أساتذتنا وزملائنا وكل من ساهم من قريب أو من بعيد في تطوير هذا العمل. نتمنى أن يقبلوا شكرنا المتواضع.

Résumé

Au cours de la dernière décennie, l'intérêt pour l'intelligence artificielle s'est accru en raison de ses nombreux avantages et de sa Incursion dans de nombreux secteurs et domaines, y compris le traitement du langage naturel .

La langue est l'un des moyens de communication les plus importants entre les humains, il est donc nécessaire d'explorer les contributions de l'intelligence artificielle au traitement du langage naturel, nous nous soucions donc de la langue arabe en raison du nombre de pays qui la parlent.

Motivés par son importance et le manque de travail dans le domaine de l'annotation automatique des rôles sémantiques de la langue arabe, nous nous sommes engagés dans notre dernière thèse à contribuer à la réalisation du système d'annotation des rôles sémantiques arabes basé sur l'intelligence artificielle.

Pour cela, nous avons utilisé la plus grande base de données annotée en rôles sémantiques pour la langue arabe (OntoNotes 5.0), et des méthodes d'intelligence artificielle.

Mots-clés : annotation, Deep Learning, intelligence artificielle, langue arabe, OntoNotes 5.0, PropBank, rôles sémantiques.

Abstract

Over the past decade, interest in artificial intelligence has grown due to its many benefits and its penetration into many industries and fields, including natural language processing.

Language is one of the most important means of communication between humans, so it is necessary to explore the contributions of artificial intelligence to natural language processing, so we care about the Arabic language due to the number of countries that speak it.

Motivated by its importance and lack of work in the field of automatic annotation of the semantic roles of the Arabic language, we are committed in our last thesis to contribute to the realization of the system of annotation the Arabic semantic roles based on artificial intelligence.

For this, we used the largest database annotated in semantic roles for the Arabic language (OntoNotes 5.0), and artificial intelligence methods.

Keywords: annotation, Deep Learning, artificial intelligence, Arabic language, OntoNotes 5.0, PropBank, semantic roles.

الملخص

على مدار العقد الماضي كان الاهتمام بالذكاء الاصطناعي متزايد بسبب إيجابياته الكثيرة وتوغله في العديد من القطاعات والمجالات، من بينها معالجة اللغة الطبيعية.

تعتبر اللغة من أهم وسائل الاتصال بين البشر، لذلك من الضروري استكشاف إسهامات الذكاء الاصطناعي في معالجة اللغة الطبيعية، لهدا نهتم باللغة العربية نظرا لعدد الدول المتحدثون بها.

وبدافع من أهميتها وعدم عملها في مجال الشرح التلقائي للأدوار الدلالية للغة العربية، فإننا ملتزمون في أطروحتنا الأخيرة بالمساهمة في تحقيق نظام شرح الأدوار الدلالية العربية القائم على الذكاء الاصطناعي.

لهذا الغرض، استخدمنا أكبر قاعدة بيانات مشروحة في الأدوار الدلالية للغة العربية وطرق الذكاء الاصطناعي.

الكلمات المفتاحية: شرح، التعلم العميق، الذكاء الاصطناعي، اللغة العربية، OntoNotes 5.0، PropBank، أدوار دلالية.

Liste des figures

Liste des tableaux

Table des matières

Introduction

Introduction

La communication est une opération étudiée par de nombreux domaines et parmi eux l'informatique. Il n'est pas nécessaire d'approuver l'importance du langage pour la communication, dès lors ce dernier est l'un des premiers domaines de recherche en informatique, sous le nom de traitement automatique du langage naturel.

Ce domaine d'étude contient de nombreuses contributions surtout dans des langues importantes comme l'anglais. Nul ne peut nier l'importance de la langue arabe d'un point de vue historique, scientifique, politique, etc. Pour cela, on s'intéresse au traitement automatique de la langue arabe.

Notre contribution dans cette thèse se porte sur la réalisation d'un système d'annotation des rôles sémantiques pour la langue arabe. Le manque de recherches qui traitent ce sujet, particulièrement dans le monde arabe, impose des apports de recherche par la communauté scientifique arabophone.

À la suite d'une étude bibliographique, on a remarqué le manque de travaux sur l'annotation des rôles sémantiques arabes, surtout après la construction d'une ressource textuelle, plus riches que celle déjà utilisée dans d'autres systèmes d'annotation des rôles sémantiques arabes.

Ce système est établi en s'appuyant sur des méthodes d'intelligence artificielle, dont une est l'apprentissage profond (Deep Learning). L'objectif est de classifier les constituants de la phrase en plusieurs rôles sémantiques.

Organisation de mémoire

Cette thèse est structurée de la manière suivante :

- **Chapitre 1 : Annotation des rôles sémantiques**

Cette partie est consacrée au sujet principal de notre étude. On montrera les rôles sémantiques, on donnera une description détaillée de leurs différents aspects et les formes utilisées. Les travaux sur cette approche démontrent son importance et justifient notre intérêt à l'étudier.

- **Chapitre 2 : Annotation des rôles sémantiques pour la langue arabe**

Dans ce chapitre, Nous avons jugé qu'il était nécessaire de présenter les rôles sémantiques dans la langue arabe et exposer les quelques travaux et ressources disponibles pour travailler sur l'annotation des rôles sémantiques arabes.

- **Chapitre 3 : Méthodes d'apprentissage automatique**

Il est important de passer par l'intelligence artificielle, pour cela, on montrera : l'apprentissage automatique, ses différents types, les algorithmes utilisés et les différents domaines. Également, on présentera le rôle qu'il a joué dans l'annotation des rôles sémantiques.

- **Chapitre 4 : Implémentation, résultats et discussion**

C'est le plus important chapitre, car on présentera les outils utilisés pour mettre en œuvre le système proposé. On décrira également l'approche utilisée pour construire notre système, ainsi que l'ensemble de données utilisé pour évaluer le modèle. On expliquera également les résultats des expérimentations d'algorithmes d'apprentissage automatique utilisés.

À la fin nous terminerons par une conclusion et les perspectives de recherches.

Chapitre 1

Annotation des Rôles Sémantiques

Chapitre 1
Annotation des Rôles Sémantiques

1- Introduction

Dans un premier temps, il est important de parcourir d'abord le sujet principal de notre thèse. Donc, nous commencerons par la définition des rôles sémantiques et l'annotation, puis nous montrerons les différents formalismes de représentation du sens. Enfin, nous aborderons les différentes approches d'annotation et l'impact de cette dernière sur le traitement du langage naturel.

2- Rôles sémantiques

Le domaine du traitement automatique du langage (TAL) comme son nom l'indique, c'est la relation entre la langue, l'informatique et l'intelligence artificielle. Il est basé sur la conception de systèmes et de technologies qui traitent le langage humain [1, p. 2].

Les langues naturelles sont caractérisées par l'ambiguïté, ce qui a conduit à des tâches complexes (désambiguïsation lexicale, analyse syntaxique, correction orthographique, etc.) pour les informaticiens et les chercheurs dans le domaine du traitement automatique du langage naturel (TALN) [2].

Le terme rôle est apparu, il y a longtemps par *Panini* qui a défini les rôles *Karak* du langage *Sanskrit*, afin de catégoriser les participants d'un événement. Puis, ils ont été réabordés plus tard par[3], [4] et jusqu'à nos jours la communauté scientifique continue les discussions autour des rôles, sous plusieurs termes rôles sémantiques, cas profonds, relations thématiques, rôles thématiques, thêta-rôles, etc. [5, p. 8].

Wrihatnala (2016) [6] comme référence les auteurs dans [7, p. 59] expliquent que l'objectif des rôles sémantiques est l'étude du sens du verbe en analysant les constituants autour de lui. On note plusieurs rôles sémantiques [8, p. 374]:

- Agent : celui qui cause ;
- Expérience: l'expérience tirée d'un événement ;
- Force : responsable intentionnel ;
- Thème : participant touché directement;
- Résultat : produit final ;
- Bénéficiaire : l'acquéreur ;

- Source : l'origine (si un événement de transfert) ;
- But : la destination (si un événement de transfert) ;
- Accompagnement : quelque chose qui participe avec un agent, etc. ;
- Locatif : montre l'emplacement.

Par exemple dans la phrase « *Cliquez sur le bouton* » (ci-dessous) :

[Agent Vous] CLIQUEZ sur [Patient le bouton]

Les constituants VOUS et le BOUTON sont les arguments du verbe CLIQUER. Le pronom personnel (VOUS) a le rôle sémantique Agent, et le (BOUTON) a le rôle sémantique Patient [9, p. 3].

3- Annotation des rôles sémantiques

Lors de l'annotation des rôles sémantiques, les constituants de la phrase (arguments) reçoivent des étiquettes de rôles (Agent, Destination, Instrument, etc.). Cette annotation est conditionnée par la disposition de riches ressources lexicales ou des phrases annotées (corpus annoté), ces deux derniers sont la base de méthodes statistiques ou d'apprentissage machine .[10, p. iii]

On emploie l'annotation des rôles sémantique pour répondre à des questions comme : qui a fait ? Quoi ? À qui ? etc. Puis à la fin, on disposera de plusieurs informations par exemple [11, p. 5]:

- Le verbe (prédicat) du contexte actuel ;
- Le cadre en question (Frame) ;
- Rôles annotés.

3.1 Terminologie

De nombreux termes qui ont la même signification qu'annotation en rôles sémantiques coexistent :

- Annotation syntaxico-sémantique des actants [10];
- Étiquetage en rôles sémantiques [12];
- Étiquetage de rôles sémantiques [13];

- Prédiction de la structure sémantique [14].

Pour l'anglais généralement deux termes sont utilisés, si le lexique utilisé est PropBank, on emploi (Semantic Role Labeling), sinon (Frame Semantic Parsing) pour FrameNet (ces les deux lexiques sons défini à la suite (4.2,4.4)).

L'utilisation d'annotations PropBank consiste à l'identification des arguments de chaque verbe (ARG0, ARG1, Arg2, etc.) [15]. Dans FrameNet l'identification consiste à trouver le prédicat (verbe, nom, adjectif et adverbe), puis la frame et enfin les rôles sémantiques des arguments [11, p. 36].

3.2 Le système Semantic Roles Labeling (SRL)

Le processus SRL consiste à l'identification des prédicats et leurs arguments [16, p. 1]. Il passe par quatre étapes [17, p. 620]:

- Identification du prédicat ;
- Extraction des arguments ;
- Catégorisation des arguments ;
- Faire face aux limitations linguistiques.

Les informations contenues dans FrameNet [18], PropBank [19] et le développement de l'apprentissage automatique ont aidé au développement de méthodes puissantes de SRL [15] ,[16, p. 1]. Mais malheureusement, la plupart des systèmes SRL sont dédiés à l'anglais, tandis que les autres langues essayent de suivre les systèmes réussis en anglais.[20, p. 95].

4- Formalismes de représentation de sens

En traitement du langage naturel, de nombreuses ressources lexicales qui contiennent les rôles sémantiques ont été créées [10, p. 38], nous citons:

4.1- WordNet

Dans [11, p. 9] les auteurs disent que l'élaboration commence dans les années 80. Il représente le lexique comme un graphe [21] et il est basé sur la psycholinguistique. Quatre graphes sur quatre parties du discours sont fournis : nom, verbe, adjectif et adverbe.

Chaque groupe de synonymes représente un nœud qui contient mots, définition et exemples. Ces groupes sont appelés aussi *synset* et ils sont liés entre eux par des relations : hyperonymie, méronymie, antonymie, etc.

L'entrainement de systèmes basés sur les données (supervisé) a nécessité l'annotation de plusieurs corpus à partir de WordNet [22] et il a été employé dans plusieurs évaluations [23] .

4.2- FrameNet

FrameNet a été développé en 1997 à l'Université de Berkeley [24, p. 2] ,il est une base lexicale basée sur la Frame Semantics [18].

Il y a plusieurs Semantic Frames qui peuvent contenir le même mot, mais avec des sens différents. À titre d'exemple les deux frames *Activity-stop* et *Abandonnement* contiennent le verbe (abandon) [10, p. 18.19]

Cette base de données lexicale est composée de trois parties :

- **La description des cadres sémantiques**

Dans le langage naturel, il y a plusieurs contextes. Ainsi, la sémantique des cadres de Fillmore est utilisée dans FrameNet avec plus de 1000 cadres pour englober plus de contextes que possible.

- **Le dictionnaire d'unités lexicales**

Chaque unité lexicale inclue une définition et des exemples annotés et il y a plus de 10000 unités (prédicats des cadres).

- **Le corpus annoté**

Il contient 175000 exemples annotés de British National et des annotations de texte complètement annoté [24, p. 2.5].

4.3- VerbNet

Selon [11, p. 9.12], on peut dire que VerbNet [25] est un perfectionnement des classes de Levin et une version électronique de celle-ci :

- À partir de [26] d'autres classes sont ajoutées, ainsi que des verbes [27];

- Le projet SemLink [28] lie des verbes à FrameNet, PropBank, WordNet et OntoNotes ;
- L'ajout de plusieurs verbes est prévu par le traitement de grand corpus [29].

4.4- PropBank

L'ajout d'annotations sémantiques sur les annotations syntaxiques de Penn TreeBank fait de PropBank le lexique le plus utilisé pour l'annotation sémantique.

PropBank a une certaine abstraction des arguments (rôles sémantiques) qui sont autour d'un prédicat (verbe). Cette abstraction fait du PropBank une ressource idéale pour l'apprentissage automatique [24, p. 10] .

L'agent est souvent représenté par le rôle sémantique ARG0, patient par ARG1 et il y a d'autres rôles comme ARG2, ARG3, etc., ainsi que des rôles moins fréquents qui représentent le temps, la location, etc. Le PropBank est basé sur cette présentation générale des rôles afin que les systèmes d'annotation des rôles sémantiques profitent de la puissance de l'apprentissage automatique [11, p. 32].

Tableau 1.1 : Signification et arguments PropBank

Tag	Description
ARG0	Agent,operator
ARG1	Thing operated
ARG2	Explicit patient (thing operated on)
ARG3	Explicit argument
ARG4	Explicit instrument

Dans la figure 1.1, on montre un arbre syntaxique avec des annotations PropBank.

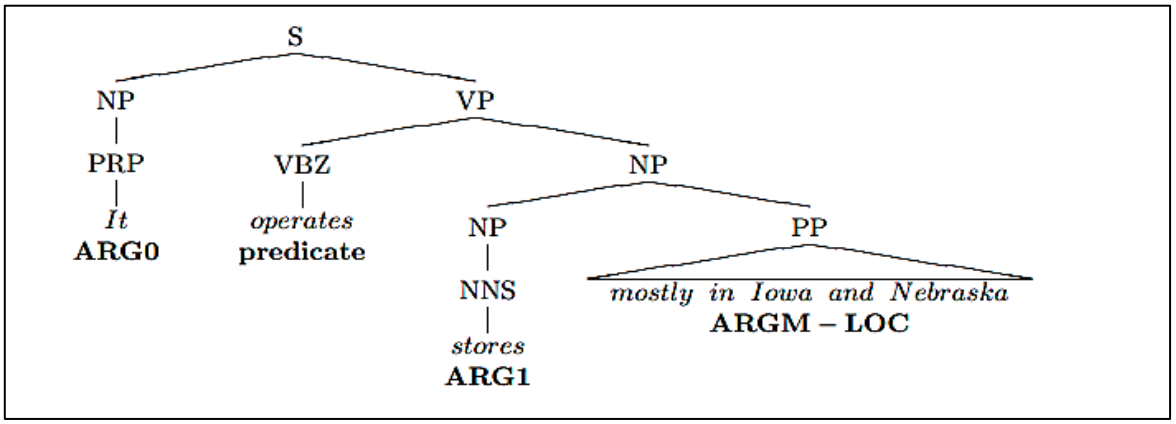


Figure 1.1 : Annotation PropBank sur un arbre syntaxique [31, p. 3.4]

Tableau 1.2 : Les rôles sémantiques dans VerbNet, PropBank et FrameNet [32, p. 3]

VerbNet	PropBank	FrameNet
Agent	Arg0, Arg1	Agent, Speaker, Cognizer, Communicator, Ingestor, Deformer, etc.
Actor	Arg0	Avenger, Communicator, Item, Participants, Partners, Wrongdoer
Actor1	Arg0	Arguer1, Avenger, Communicator, Interlocutor1, Participant 1, etc.
Actor2	Arg1, Arg2	Addressee, Arguer2, Injured Party, Participant2, Partner2
Attribute	Arg1, Arg2	Attribute, Dimension, Extent, Feature, etc.
Beneficiary	Arg1, Arg2, Arg3, Arg4	Audience, Beneficiary, Benefitted party, Goal, Purpose, Reason, Studio
Cause	Arg0, Arg1, Arg2, Arg3	Addressee, Agent, Cause, Communicator, etc.
Destination	Arg1, Arg2, Arg5	Addressee, Body part, Context, Goal, etc.
Experiencer	Arg0, Arg1	Cognizer, Experiencer, Perceiver, etc.
Extent	Arg2	Difference, Size change
Instrument	Arg2	Agent, Fastener, Heating instrument, Hot Cold source, etc.
Location	Arg1, Arg2, Arg3, Arg4, Arg5	Action, Area, Fixed location, etc
Material	Arg1, Arg2, Arg3	Components, Ingredients, Initial entity, Original, Resource, Undergoer
Patient	Arg0, Arg1, Arg2	Addressee, Affliction, Dryee, Employee, Entity, Executed, etc.
Patient1	Arg0, Arg1	Concept 1, Connector, Fastener, Item, Item 1, Part 1, Whole patient
Patient2	Arg2, Arg3	Concept 2, Containing object, Item 2, Part 2
Predicate	Arg1, Arg2	Action, Category, Containing event, etc.
Product	Arg1, Arg2, Arg4	Category, Copy, Created entity, etc.
Proposition	Arg1, Arg2,	Act, Action, Assailant, Attribute, etc.
Recipient	Arg1, Arg2, Arg3	Addressee, Audience, Authorities, Recipient
Stimulus	Arg1	Emotion, Emotional state, Phenomenon, Text
Theme	Arg0, Arg1, Arg2	Accused, Action, Co-participant, Co-resident, Content, Cotheme, etc.
Theme1	Arg0, Arg1	Cause, Container, Phenomenon 1, Profiled item, Theme
Theme2	Arg1, Arg2, Arg3	Containing object, Contents, Cotheme, etc.
Time	ArgM TMP	Time
Topic	Arg1, Arg2	Act, Behavior, Communication, Content, etc.
Asset	Arg1, Arg3	Asset, Category, Measurement, Result, Value
Value	Arg1	Measurement, Result, Value, Asset, Category
Source	Arg2, Arg3	Role, Victim, Patient, Source, Path start, etc.
-	-	Setting, ContainingEvent
-	-	Means
-	ArgM _Manner	Manner
-	ArgM _Purpose	Purpose

5- Approches d'annotation

Dans la section suivante, les auteurs survolent les approches d'annotation [11, p. 33.35] .Ils disent que les inventaires et les corpus annotés sont les ressources utilisées pour l'annotation des rôles sémantiques en suivant quatre approches :

- **Fondées sur la connaissance**

Certaines approches utilisent un inventaire de sens (WordNet, VerbNet, PropBank, etc.), en d'autres termes, ils recueillent des connaissances sans exemples annotés. Ainsi, dans ce type d'approche, la taille de corpus n'a pas d'importance.

- **Supervisées**

Les approches dites supervisées se basent sur les corpus annotés selon un inventaire X et utilisent les méthodes d'apprentissage automatique pour connaitre le sens des constituants de la phrase. Les systèmes d'annotation basés sur cette approche sont généralement découpés sur plusieurs parties :

- Détermination de l'élément principal (Prédicat) ;
- Connaitre la Frame ;
- Détermination des autres éléments (arguments) autour du prédicat ;
- Classification des arguments, selon leur rôle dans la phrase.

- **Semi-supervisées**

Ce type d'approche peut présenter des obstacles comme la correction manuelle, résultat faible, utilisation directe des données des autres langues, etc. Mais, c'est un axe intéressant pour la génération de corpus annotés.

À titre d'exemple des auteurs de [33]avec un système basé PropBank ont annoté le corpus Europarl anglais, puis ils ont aligné les deux corpus anglais/français. Ainsi, ils obtiennent un corpus français annoté PropBank.

- **Non supervisées**

Par exemple, en utilisant le clustering de sens d'un corpus, ces approches construisent leur inventaire. Car, il n'utilise pas des connaissances préalables comme un inventaire ou un corpus annoté.

6- L'utilisation des rôles sémantiques dans le traitement automatique des langues

Dans cette partie, on présente l'importance des rôles sémantiques dans le TALN à partir de travaux intéressants qui profitent des avantages des rôles sémantiques.

Dans [9, p. 4] les auteurs citent un travail de Fliedner [34][1] dans lequel l'auteur montre l'intérêt de l'annotation automatique sur la performance des systèmes d'extraction d'information, traduction automatique, etc. Les mêmes auteurs citent d'autres travaux : le résumé automatique [35], la traduction automatique [36], l'extraction d'informations [37] et [38] et les systèmes de questions/réponses [39] .

Dans le même contexte [40, p. 63] cite plusieurs travaux qui utilisent les rôles sémantiques : résumé automatique [10], [41], [35] et [42], d'extraction des réponses [43], [39], [44] et [45] ou encore la traduction automatique [46], [36], [47] et [48].

La facilité d'intégration des rôles sémantiques à d'autres tâches a permis son introduction à des travaux moins répandus tels que : prédiction de bourse, recommandation, détection de plagiat et comparaison de produits, évaluation de la traduction automatique, etc. [11, p. 6].

7- Conclusion

Ce chapitre a permis au lecteur de notre thèse de se familiariser avec la notion de rôle sémantique. Car nous avons des notions de base sur les rôles sémantiques et leurs annotations dans le langage naturel de façon générale. Dans le chapitre qui suit, nous aborderons cette notion dans la langue arabe en particulier.

[1] Référence citée par notre référence [8], non disponible pour être vérifiée.

Chapitre 2

Annotation des Rôles Sémantiques pour la langue Arabe

Chapitre 2
Annotation des Rôles Sémantiques pour la langue Arabe

1- Introduction

Nous avons décidé de consacrer un chapitre séparé aux formalismes de représentation de sens dans la langue arabe et les contributions scientifiques sur les systèmes d'annotation des rôles sémantiques arabes.

L'arabe est une langue très répondue auparavant, elle régna huit siècles en Espagne et aussi sur un grand territoire qui commence en lnde et se termine devant l'océan atlantique au Maroc. Elle est l'une des langues sémitiques et répondues jusqu'à maintenant avec certaines différences régionales comme dans notre pays et le nord de l'Afrique en général, la Syrie, l'Arabie Saoudite, etc. [49].

Dans l'arabe le verbe exprime plusieurs traits tels que : le temps, la voix, la personne, l'indicatif, etc., de même les nominaux expriment d'autres marques et certaines choses sont exprimées par les voyelles, ainsi tout cela fait que cette langue soit riche morphologiquement [50, p. 2].

2- Formalismes de représentation de sens pour l'arabe

2.1- WordNet

Les gains de traitement automatique du langage qui sont acquis par le développement de Princeton WordNet (PWN) pour l'anglais ont ouvert la voie à la réalisation de WordNet pour d'autres langues [51, p. 29].

Il est évident que des efforts sont effectués pour le développement d'un arabe WordNet basé sur Princeton WordNet et l'Euro WordNet [52] et [53].

2.2- VerbNet

Si on considère le nombre de verbes, pour la langue anglaise VerbNet [54] est la plus grande ressource. Des travaux sont réalisés pour la construction d'un arabe VerbNet par Jaouad, M dans [55], on se basant et adaptant le VerbNet anglais [54]. La figure 2.2 montre un exemple de VerbNet en arabe :

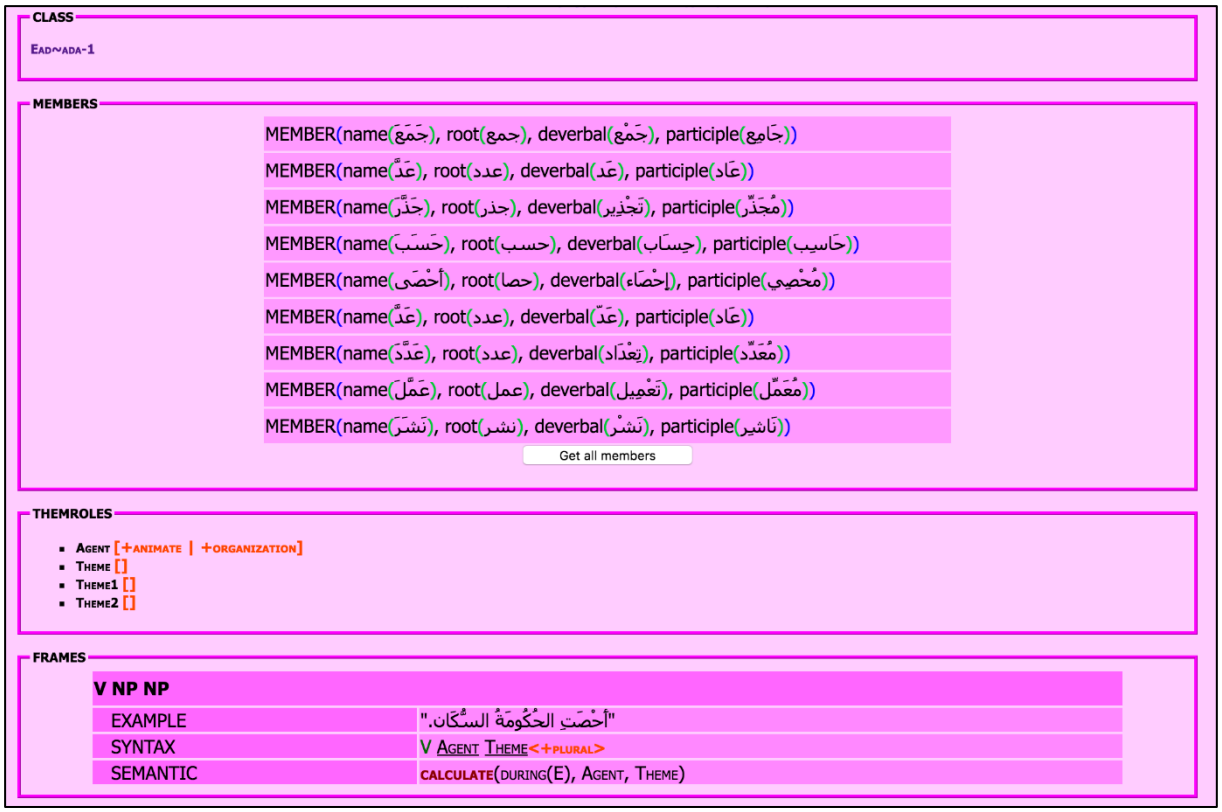


Figure 2.1 : Classe VerbNet en arabe [56]

2.3- FrameNet

Pour l'anglais, les travaux de ce projet ont commencé en 1997 [57], [58] et [59]. Il est basé sur les frames sémantiques et les traits syntaxiques et sémantiques dans le corpus British National Corpus [60]. Certainement, cela donne naissance à des travaux similaires dans d'autres langues [61, p. 1].

Dans le site officiel de FrameNet,[2] on constate qu'il y a des ressources pour le français, chinois, allemand, etc., mais aucune ressource n'est disponible pour l'arabe. Selon [40, p. 69], il y a des efforts pour la construction d'un FrameNet arabe pour les verbes du Coran [62], aussi qu'une méthodologie de construction dans [63]. Cependant et selon le même auteur, ils restent des travaux minimes comparés au FrameNet anglais et les attentes de la communauté scientifique internationale.

2.4- TreeBank

C'est une ressource fondamentale pour les chercheurs en traitement du langage naturel. En 2001 des travaux commencent pour l'arabe TreeBank avec comme objectif

[2] https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages

l'annotation (automatique et manuelle) d'un large corpus arabe utilisable par la machine et il se compose de deux phases : la première "Part-of-Speech" et la deuxième le "Tree banking" [64, p. 1].

Il y a aussi un arabe TreeBank qui représente l'arabe traditionnel du saint Coran [65, p. 1]. Construit afin de pouvoir l'utiliser par la machine dans d'autres recherches sur la grammaire du saint Coran [66].

Le corpus utilisé dans nos recherches (OntoNotes 5.0) est basé sur l'arabe TreeBank et sur l'arabe PropBank présenté dans le prochain titre.

2.5- PropBank

Comme les autres formalismes de représentation de sens, PropBank existe pour l'anglais et certaines autres langues [67], [68] et [69, p. 1]. Selon [70, p. 78] sur la base de l'arabe TreeBank [71], seulement un unique travail existe pour l'arabe PropBank [72] et [73] et contient 4.5K verbes et qui reste modeste.

sa création est semblable à celle dans d'autres langues, elle inclue des étapes comme : création des framesets, utilisation de ces derniers par les annotateurs, identification des prédicats, etc. [69, p. 1].

La phrase arabe,

بدا رئيس الوزراء الصيني زو رونجي زيارة رسمية للهند الاحد الماضي

est annotée dans l'arabe PropBank comme le montre la figure 2.2.

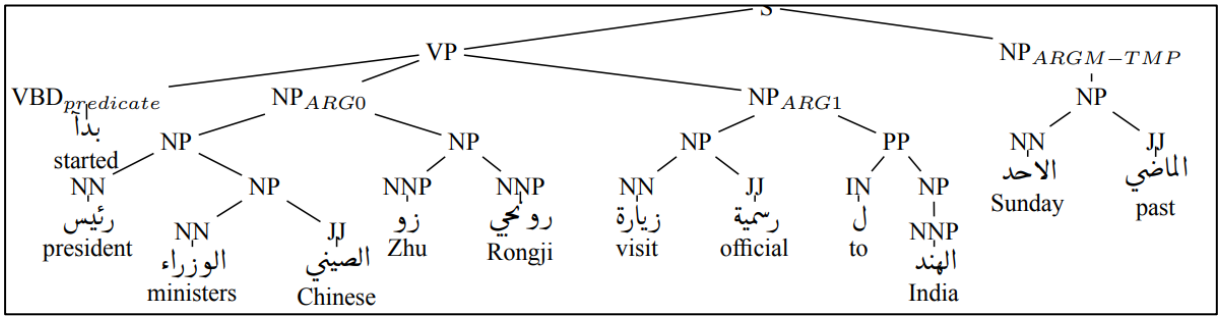


Figure 2.2 : Annotation de PropBank arabe sur la base d'un TreeBank [73, p. 800]

3 Système d'annotation pour l'arabe

Mona Diab et autres [74, p. 134] élaborent le premier système d'annotation pour une langue sémitique et la langue arabe. Principalement, ils se basent sur les travaux des systèmes d'annotation en anglais, en les adaptant aux caractéristiques de langue arabe.

C'est un système qui utilise les machines à vecteur de support (Support vector machine) pour la détection et la classification des arguments et classé dans la catégorie des systèmes supervisés, car il utilise le corpus annoté dans SEMEVAL 2007 Task 18[3]. Ce corpus caractérisé par :

- Basé sur l'arabe PropBank et TreeBank;
- Couvre 95 verbes de TreeBank ;
- Absence de voyelles ;
- 886 phrases pour le développement (1.725 arguments), 902 pour le test (1.661 arguments) et 8.402 pour l'entrainement (21.194 arguments)

Dans cette première contribution, les résultats sont 94,06% pour la détection et 81,43% pour la classification.

Les mêmes auteurs proposent une autre contribution dans [75], ils exploitent la richesse morphologique de la langue arabe et on utilisant les SVM et astuce de noyau (Kernel Methods).

Un autre travail [76] utilise un autre corpus (OntoNotes 5.0) plus récent, riche et important que les précédentes contributions, elle est aussi basée sur un système supervisé employant, le raisonnement à partir de cas (RàPC), les K plus proches voisins et il obtient un résultat de 62,42% sur la classification des arguments. Les mêmes auteurs donnent d'autres contributions dans une thèse et ils obtiennent un résultat de 88,66% d'une hybridation entre RàPC, les K plus proches voisins et le Deep Learning, mais ses résultats ne sont pas publiés pour le moment.

[3] http://web.archive.org/web/20080727062358/http://nlp.cs.swarthmore.edu/semeval/index.php

4- Conclusion

Dans cette partie, nous avons abordé le problème d'annotation des rôles sémantiques dans la langue arabe. Pour donner une idée sur les travaux précédents et montrer la problématique de notre thèse. Le chapitre suivant sera consacré aux méthodes d'intelligence artificielle.

Chapitre 3

Méthodes d'Apprentissage Automatique

Chapitre 3
Méthodes d'Apprentissage Automatique

1- Introduction

Pour résoudre la problématique de notre thèse, nous aurons eu recours à l'intelligence artificielle. Pour cela, dans le présent chapitre, nous présenterons l'apprentissage automatique et ses types. Puis, une description générale de certaines méthodes, leurs secteurs d'application et enfin leurs applications dans l'annotation des rôles sémantiques.

2- Historique

Les recherches basées sur les connaissances et les systèmes experts ont été plus répondues dans les années 70, mais dès les années 80, ils ont montré leurs limites. Cela a donné retour aux nouveaux algorithmes d'apprentissage. C'est un retour, puisque les études sur l'apprentissage automatique ont commencé une soixantaine années avant (en 1920).

Comme tout autre domaine, les recherches ont eu des hauts et des bas (les limites du perceptron simples), mais dans les années 80 un nouvel air et une nouvelle naissance commence. Ce domaine est dans le croisement de plusieurs autres domaines ou disciplines comme les mathématiques, la physique, l'informatique, l'intelligence artificielle, la biologie, l'économie, etc. [77].

3- Quelques mots sur l'apprentissage artificiel

Des compétences comme : parler, lire, écrire, faire des maths ou encore les capacités physiques pour faire certaines choses spéciales ou non sont acquises par l'apprentissage (l'exercice). L'être humain est connu par cette possibilité d'apprentissage et cela a permis son évolution à travers les cycles, cette compétence englobe plusieurs mécanismes ou processus pas toujours faciles à comprendre [78, p. 14].

On veut dire par l'apprentissage automatique ou artificiel la possibilité que la machine apprend à partir d'exemples comme l'humain avec ses vécus. Pour rendre cet apprentissage possible, il est nécessaire de mettre en œuvre des systèmes susceptibles d'apprendre à partir de données précédentes et utiliser cela pour résoudre de nouveau cas [79, p. 4.5].

3.1- Modèle de fonctionnement global

L'efficacité d'un système d'apprentissage est assignée par une mesure P sur des expériences réelles données au préalable, qui peuvent accroitre avec le temps [78, p. 15].

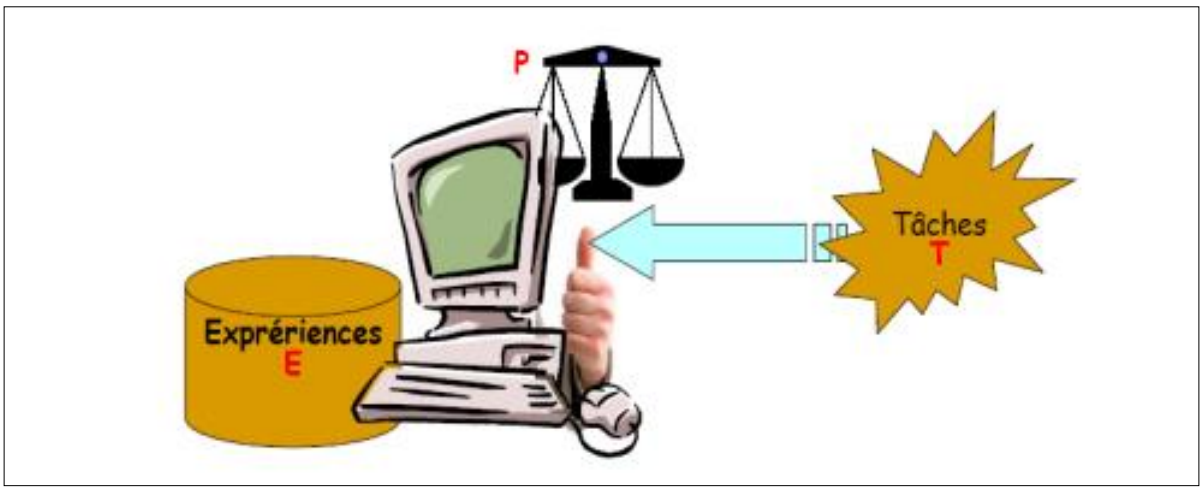


Figure 3.1 : Fonctionnement général d'un système d'apprentissage [78, p. 15]

4- Types de l'apprentissage automatique

Principalement, il y a trois types dans l'apprentissage artificiel :

4.1- Apprentissage supervisé

Pour bien expliquer l'apprentissage supervisé, on note qu'il y a une base d'exemples (cas) tirés du monde réel et approuvée par des experts du domaine ou autres. Un exemple ou un cas est représenté par :

$$E\_cas_i \rightarrow X_n, Y_n$$

Où:

- $X_{n=} = \{X_1, X_2, X_3, \dots X_n\}$: un ensemble d'attributs qui décrivent le cas.
- $Y_{n=} = \{Y_1, Y_2, Y_3, \dots Y_n\}$: un ou plusieurs outputs (classes).

Ces cas constituent la base d'apprentissage d'un algorithme supervisé, afin d'extraire des relations entre l'ensemble X_n et Y_n, puis utiliser ces relations, connaissances ou cet apprentissage pour prédire les Y_n des nouveaux X_n [80, p. 65].

La figure 3.2 montre un ensemble de photos X et leur nominations Y. Après un apprentissage machine, un nouveau (photo d'un chat à drotie) cas arrive, puis a base de connaissances d'apprentissage le système détermine le nom de l'animal sur la photo.

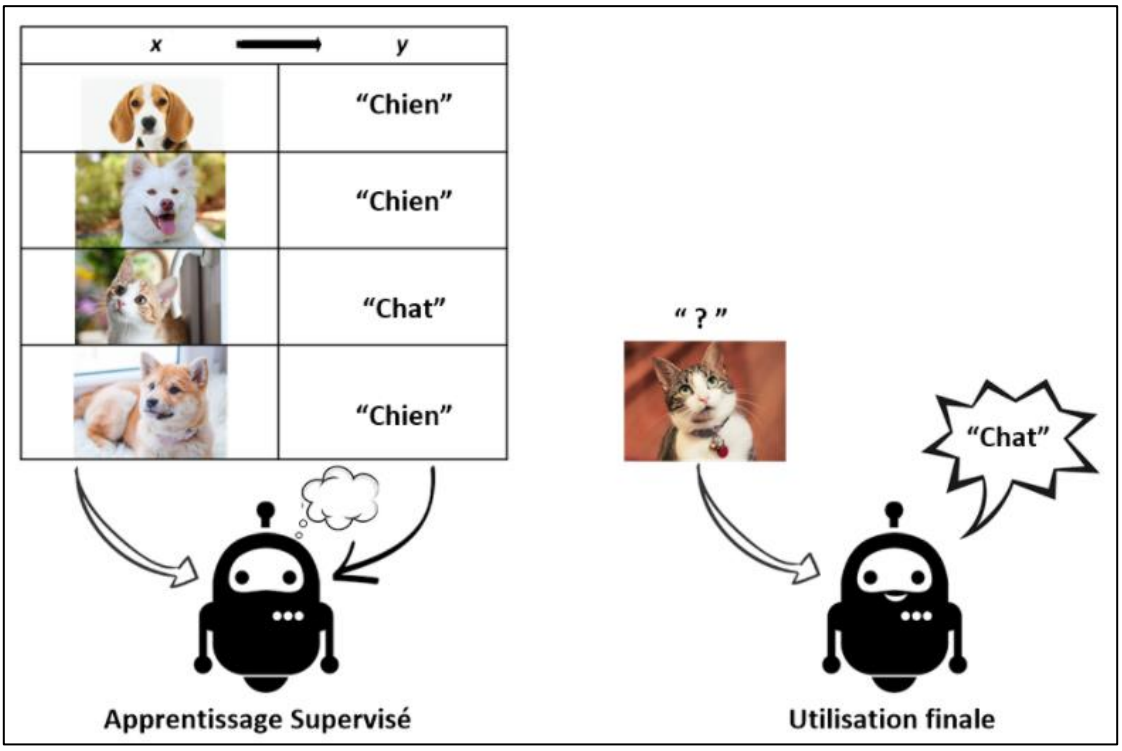


Figure 3.2 : Illustration de l'apprentissage à partir d'exemples [81]

Les auteurs dans[77] distinguent deux catégories dans l'apprentissage supervisé :

- **Symbolique**

Il est basé sur des modèles logiques, rendre les données binaires, etc., cela est issu des concepts de l'intelligence artificielle.

- **Numérique**

Ce groupe de méthodes où les données sont représentés dans la plupart des cas par des vecteurs de réels, utilisent des approches d'algèbre, optimisation et de la probabilité, car c'est un groupe ses idées sont issu de la statistique [82, p. 10].

4.2- Apprentissage non supervisé

Généralement, il est utilisé pour le regroupement des données en groupes, par des mesures de similarités comme la distance. Différemment de l'apprentissage supervisé dans ce type, la base est un ensemble de données sans variables cibles [80, p. 67].

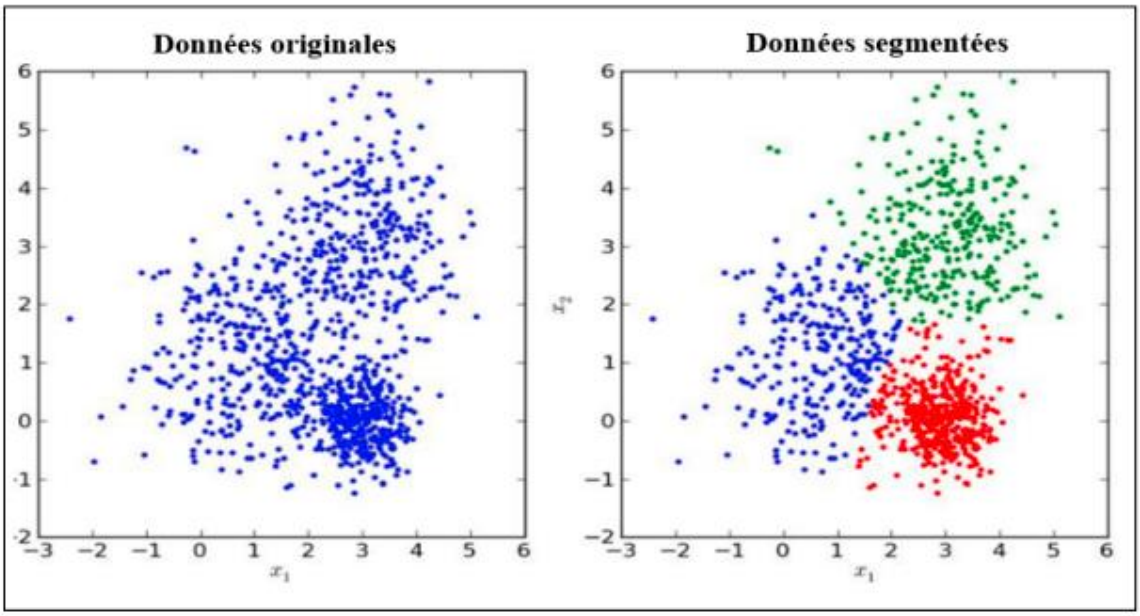


Figure 3.3 : Avant et après un apprentissage non supervisé [80, p. 67]

4.3 -Apprentissage par renforcement

C'est un apprentissage similaire au dressage des animaux, car il se base sur le système d'observations / récompense. Après les récompenses et leur maximisation, on obtient la décision ou la conduite optimale. La figure ci-dessous montre l'architecture de fonctionnement de l'apprentissage par renforcement [80, p. 68].

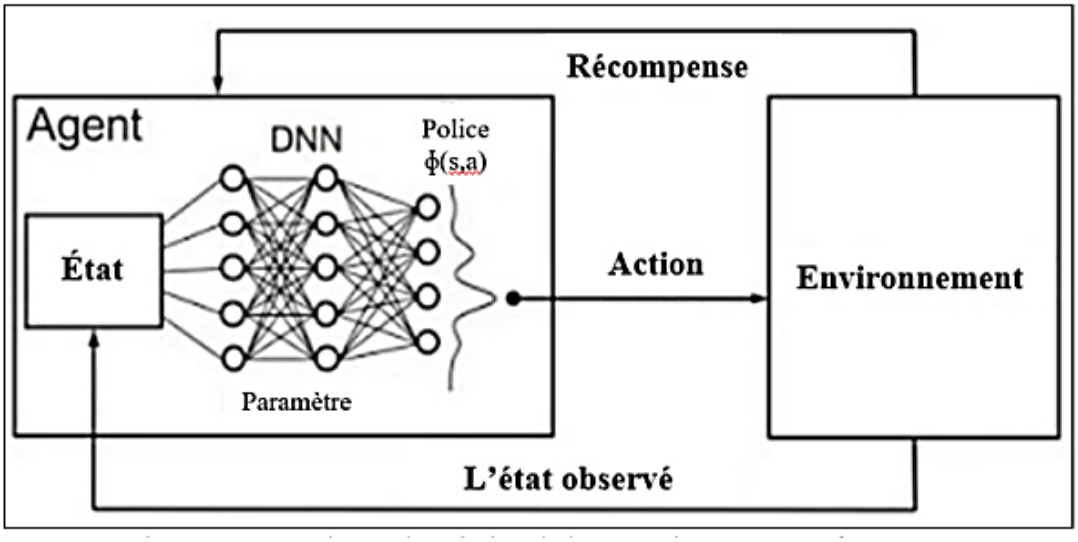


Figure 3.4 : Architecture générale de l'apprentissage par renforcement.

5- Algorithmes utilisés dans l'apprentissage automatique

Dans cette partie, on survolera quelques méthodes d'apprentissage qui sont à notre connaissance les plus utilisées.

5.1- Machine à vecteurs support (Support-Vector Machine - SVM)

Les données sont représentées par des attributs/valeurs, le nombre d'attributs est égal au nombre de dimensions. Le but de cet algorithme est de trouver l'hyperplan d'une plus petite distance (plan séparateur/cas d'entrainement). La résolution des problèmes de classification et de régression peut employer les SVMs et c'est l'un des algorithmes supervisés basés sur les données d'apprentissage[80, p. 73].

5.2- Réseau de neurones artificiels (Artificial Neural Network - ANN) et apprentissage profond (Deep Learning)

Dans [83] les réseaux de neurones sont définis par " *Un réseau de neurones formel, couramment appelé "réseau de neurones" est un calcul (ou algorithme), généralement réalisé à l'aide d'un ordinateur, dont le résultat reproduit ou prévoit aussi fidèlement que possible, le comportement de n'importe quel processus en fonction des facteurs qui déterminent ce comportement.* "

L'apprentissage profond qui est mentionné même par les francophones le Deep Learning constitue une partie de l'apprentissage automatique, en d'autres termes c'est un type d'intelligence artificielle, dans lequel la machine peut apprendre. Il est basé sur un réseau de neurones, contenant jusqu'à une centaine de couches [84].

5.3- K plus proches voisins (K-Nearest Neighbors – K-NN)

Selon [78, p. 28], le mode de fonctionnement de cette méthode est assez simple. Imaginant un espace contenant un ensemble de cas **N** et nouveau cas $\mathbf{New_n}$,. Dans le but de deviner la classe de $\mathbf{New_n}$, on cherche dans l'ensemble précédent les **K** (nombre) cas les plus proches au $\mathbf{New_n}$.
Par rapport aux autres méthodes, celle-ci ne propose pas un modèle à partir des cas précédemment stockés en mémoire.[4]

Le déroulement de cette méthode nécessite principalement le paramètre **K** représentant le nombre de cas les plus similaires au nouveau cas et ces derniers sont déterminés par une fonction de similarité.

5.4- Arbres de décision (Decision Tree)

[4] Cette référence prend la définition depuis une référence non disponible actuellement. Donc nous n'avons pas pu vérifier les informations dans la référence originale.

C'est un arbre normal au sens inverse, donc on débute par la racine, puis les branches et leurs intersections et en fin les feuilles. Dans ce domaine d'apprentissage artificiel, une intersection (nœud) contient une règle (mesures et seuil) destinée à la séparation au corpus initial. Les feuilles de cet arbre représentent les classes, étant donné que cette méthode est utilisée pour des problèmes de classification [85].

5.5- Forêt aléatoire (Random Forest)

Plus simplement, c'est une multitude d'Arabes de décision. Dans le but d'améliorer de la performance. [86]

5.6- Régression logistique (Logistic Regression)

Pour faire court, on peut dire que c'est une adaptation non linéaire d'une régression linéaire, dans le but de remplacer une valeur numérique continue par la prédiction d'une classe, car c'est une méthode prédictive [80, p. 72].

5.7- Algorithmes et programmation génétique (Genetic Algorithm - GA)

Le but global de ces algorithmes est d'avoir le point fort de chaque classificateur non optimal. Donc au début, on a un ensemble (population) initial de classificateurs non optimaux, ces derniers servent comme entrée aux algorithmes. Puis en sortie en reçoit un classificateur plus optimal. Elle a la possibilité d'être une seconde phase d'une autre méthode d'apprentissage [78, p. 33.34].

5.8- Naïve bayésienne (Naive Bayes)

C'est classificateur où ses origines commencent dans la probabilité et les statistiques. Basé sur les liaisons (présence ou non) entre les caractéristiques. [87]

6- Secteurs de recherche et d'application

Selon les auteurs dans[78, p. 15], [88, p. 4] l'automatisation de la prise de décision est un domaine très exploité par l'apprentissage automatique, la fouille de données et intelligence artificielle :

- Les reconnaissances automatisées de la parole, la forme, le texte, etc. ;
- Extraction de données en vue d'apprentissage ;
- Les diagnostics médicaux, industriels, etc. ;
- Globalement dans des recherches dans la robotique, imagerie, etc.,
- En général, il essaye d'automatiser les tâches humaines dans le but de les rendre plus faciles et idéaux.

7- Application de l'apprentissage automatique pour les rôles sémantiques

F. Hadouche et d'autres disent dans [9, p. 4] que l'annotation des rôles sémantiques a utilisée à la fois de l'apprentissage supervisé exploitant par ressources comme PropBank et FrameNet et non supervisé basé sur des connaissances lexicales.

À partir de notre impression basée sur la littérature, ainsi que les travaux cités dans [10, p. 42.43], les recherches d'annotation automatique se basent particulièrement sur l'apprentissage supervisé. Par exemple dans[37], ils utilisent les arbres de décision, la grammaire catégorielle combinatoire [89], machine à vecteurs support [90],[91] ,réseau de neurones [92] etc.

Dans le chapitre précédent, on a cité les travaux de Mona Diab qui a exploité l'apprentissage automatique dans le domaine d'annotation des rôles sémantiques pour la langue arabe.

8- Conclusion

À travers ce chapitre, nous avons montré une vue d'ensemble sur quelques méthodes d'apprentissage automatiques. Dans le chapitre suivant, nous décrirons nos réalisations pour un système d'annotation des rôles sémantiques basé sur l'intelligence artificielle.

Chapitre 4

Implémentation, résultats et discussion

Chapitre 4
Implémentation, résultats et discussion

1- Introduction

C'est le dernier et le principal chapitre de notre thèse. L'objectif sera de défendre nos choix de langage de programmation, surtout qu'il est un langage non étudié dans notre cursus universitaire, chose qui a rendu la tâche plus difficile, mais surement bénéfique.

Puis nous montrerons le corpus utilisé qui est à notre connaissance le plus grand corpus de rôles sémantiques arabes annotés. Ensuite, nous détaillerons un panorama sur nos expériences.

2- Outils de développement

Dans notre travail, on a utilisé Python comme langage de programmation, Anaconda comme distribution et Spyder comme environnement de développement. On va les présenter dans ce qui suit :

2.1- Langage Python

On a utilisé Python comme langage de programmation, malgré que nous ne l'avons pas étudié ou utilisé au centre universitaire de Mila. Pour plusieurs raisons :

- L'une des Langage les plus utilisées actuellement ;
- Apprendre et se familiariser avec une autre Langage avant d'obtenir le Master;
- C'est l'une des Langage les plus utilisées dans l'intelligence artificielle.

Selon un article web de "Bastien L" dans le site «lebigdata.fr» [93], Python est créé en 1991 par Guido van Rossum et son nom vient la série télévisée Monty Python Flying Circus. Il dit " *Python permet aux programmeurs de se concentrer sur ce qu'ils font plutôt que sur la façon dont ils le font*".

Elle contient plusieurs bibliothèques : Bokeh, Numpy, Scipy, Panda, Scikit-Learn, TensorFlow, etc. Cela et d'autres avantages comme open source, libre, compatibilité avec les plateformes, moins de codage, etc., ont permis une large utilisation dans la data science, machine learning et big data [94]

Dans un rapport intitulé « Where Programming, Ops, AI, and the Cloud are Headed in 2021 » basé sur les données de la plate-forme d'apprentissage O'Reilly, l'auteur Mike

Loukides montre dans de la figure 4.1 des statistiques sur les Langage les plus populaires. Concernant l'utilisation (usage), on voit clairement une large utilisation de Python. Pour la croissance (% usage growth), il y a une croissance d'utilisation et pour les requêtes de recherche (Queries), il est clair que Python dépasse largement les autres Langage [95]

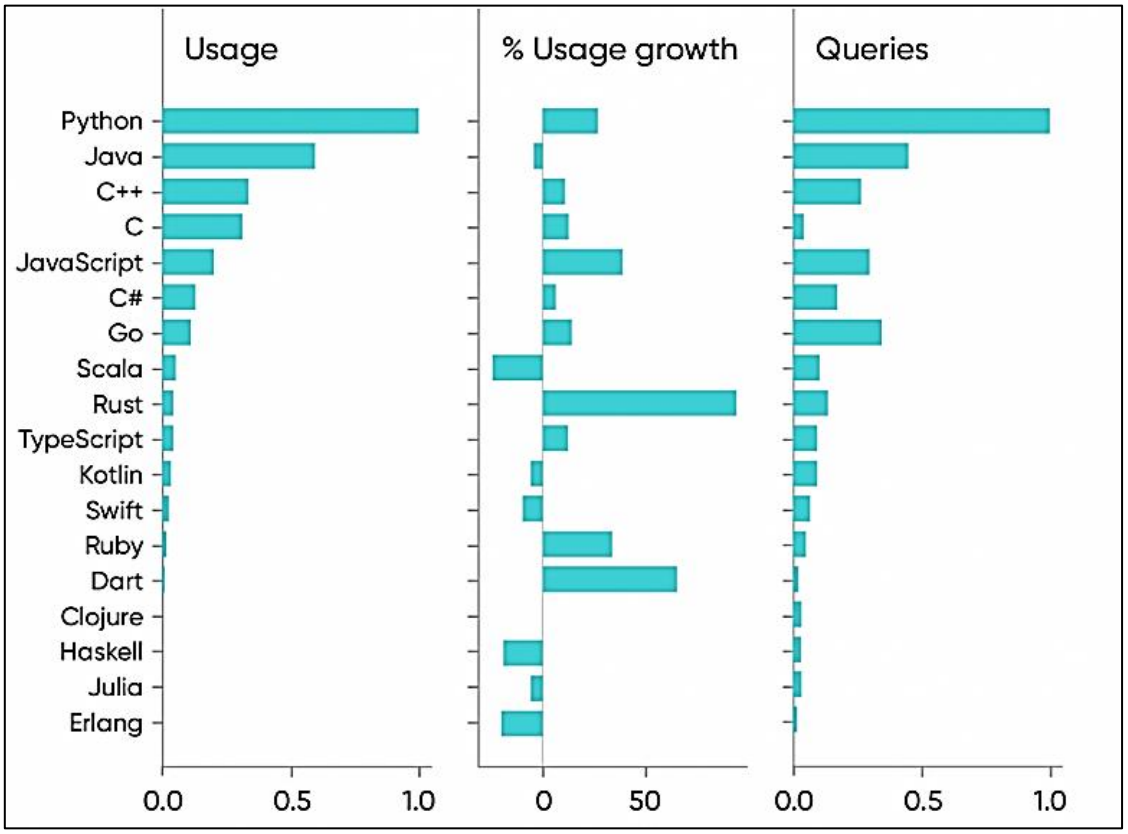


Figure 4.1 : comparaisons entre les Langage les plus populaires [95]

Des statistiques récentes ce mois-ci (septembre 2022) [96] de l'indice TIOBE Programming Community qui se base sur cours, ingénieurs qualifiés, moteurs de recherche pour le calcul des notes, etc., et qui montre les langues les plus populaires. Il classe Python dans la première place en septembre 2022 (Figure 4.2). Aussi en remarque une progression de Python depuis 20 ans (Figure 4.3).

sept. 2022	sept. 2021	Changer	Langage de programmation	Notes	Changer
1	2	^	Python	15,74 %	+4,07%
2	1	v	C	13,96 %	+2,13%
3	3		Java	11,72 %	+0,60%
4	4		C++	9,76 %	+2,63%
5	5		C#	4,88 %	-0,89%

Figure 4.2 : Popularité d'une Langage [96]

Langage de programmation	2022	2017	2012	2007	2002
Python	1	5	8	sept	12
C	2	2	1	2	2
Java	3	1	2	1	1
C++	4	3	3	3	3
C#	5	4	4	8	14

Figure 4.3 : Historique des Langage [96]

On ne tarde pas sur les justifications d'utilisation de cette langue, car une recherche simple sur internet montrera que Python est une Langage prometteuse et en large utilisation par la communauté scientifique ou professionnelle.

2.2- Anaconda

Utilisé souvent dans le data science, car il contient 300 libraires et plusieurs outils dédiés au data science. Employé aussi dans d'autres branches : machine learning, Deep Learning, etc. et c'est une distribution pour Python et R. [97]

2.3- Spyder

C'est un environnement du langage Python et contient plusieurs fonctions[98]:

- Édition avancée ;
- Débogage ;
- Introspection ;
- Tests interactifs ;
- Environnement pour le calcul numérique ;
- Support IPython.

2.4- Bibliothèques TensorFlow et Keras

❖ **TensorFlow**

Boite à outils open source très populaire dans la communauté scientifique, conçue pour les réseaux de neurones et l'apprentissage profond, en 2011 par Google Brain [99].

❖ **Keras**

Créée pour le Deep Learning, son écriture en Python le rend facile d'utilisation et extensible, c'est un Framework open source avec la possible utilisation au-dessus de TensorFlow [100].

Figure 4.4 : Bibliothèques Tensorflow et Keras[101]

3- Corpus utilisé

Nous nous sommes basés sur le corpus annoté OntoNotes 5.0 et son prétraitement dans [40].

3.1- OntoNotes 5.0[5]

Dans l'objectif de construire une grande ressource de nombreuses institutions américaines collaborent financièrement et scientifiquement : Département de la défense américain, BBN Technologies, les universités du Colorado, Pennsylvanie, Southern California, etc. Elle contient 1.5 million de mots anglais, 800 mille Chinois et 300 mille de mots arabes. Les données de cette dernière viennent du journal *An-Nahar* [102].

Dans le dossier des annotations arabes dans OntoNotes 5.0, il y a sept (07) extensions qui représentent les niveaux d'annotation (`coref, lemma, name, onf, parse, prop, sens et source`) distribuées dans six (06) dossiers (Figure 4.5). La figure 4.6 montre l'annotation d'une phrase, depuis un fichier d'extension (.onf) [40].

[5] https://catalog.ldc.upenn.edu/LDC2013T19

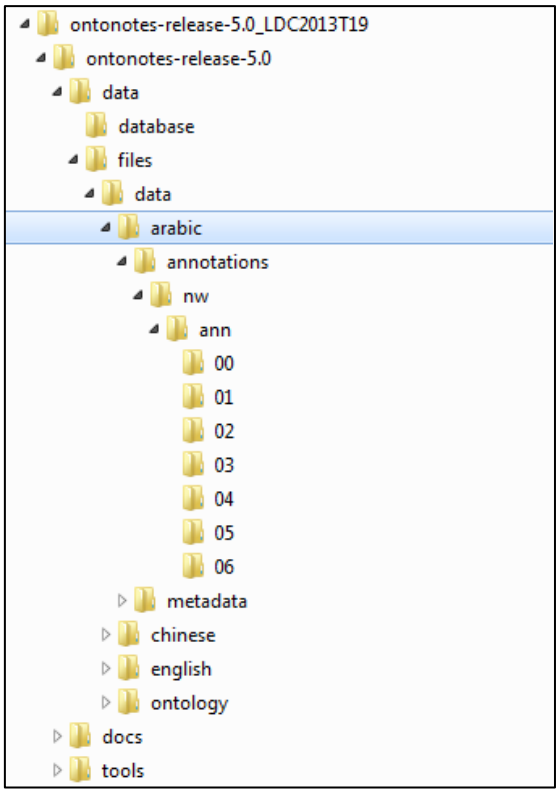


Figure 4.5 : Dossiers arabes dans OntoNotes 5.0 [40]

Figure 4.6 : Annotation d'une phrase arabe [40]

3.2- Données traitées

Les données sont celles de la conférence CoNLL-2012 (tirées d'OntoNotes 5.0) avec un format spécial où les lignes et les colonnes représentent respectivement les constituants et les informations (Tableau 4.1) [103] [40, p. 111]

Tableau 4.1 : Colonne format conll [40, p. 111]

Numéro de colonne	Représentation	Description
1	Identification dossier	Les dossiers du fichier
2	Numéro fichier	Numéro du fichier
3	Numéro du constituant	Numéro de ce constituant dans la phrase
4	Constituant	Le constituant lui-même
5	Partie de discours	Partie de discours du constituant
6	Parse bit	Partie de l'analyse après la première parenthèse
7	Lemma	C'est comme une racine du constituant
8	Frameset du prédicat	Identification du frameset du Lemma selon PropBank, s'il a un frameset.
9	Sens constituant	Le sens de ce constituant
10	Orateur/auteur	Le nom de l'orateur ou l'auteur, il n'est pas disponible pour la langue arabe
11	Entité nommée	Représentation des entités nommées
12 : N	Argument du prédicat	Le nombre de colonnes n'est pas fixe, car chaque colonne représente un seul prédicat (Lemma) et ses arguments.
N+1	Coréférence	Information sur les coréférences

La figure 4.7 montre un exemple d'un fichier en format conll [40]. Selon le même auteur, ces données ont plusieurs points forts comme :

- Validées par la communauté du TALN ;
- Un nombre important de rôles sémantiques annotés (la plus grande) ;
- Élaborées par de grandes institutions américaines.

1	2		3	4	5
nw/ann/02/ann	0290	0	0	كَ-#l`kin~a#lkn#l`kin~a-	PSEUDO_VERB
nw/ann/02/ann	0290	0	1	-ـُ#clitics#h#-hu	PRON_3MS
nw/ann/02/ann	0290	0	2	لَمْ#lam#lm#lam	NEG_PART
nw/ann/02/ann	0290	0	3	يُشِرْ#>a$Ar#y$r#yu+$ir+o	IV3MS+IV+IVSUFF_MOOD:J
nw/ann/02/ann	0290	0	4	إِلَى#<ilaY#AlY#<ilaY	PREP
nw/ann/02/ann	0290	0	5	إِنْفَالِ#<iqofAl#AqfAl#<iqofAl+i	NOUN+CASE_DEF_GEN
nw/ann/02/ann	0290	0	6	التُّرْقَاتِ#Tariyq#AlTrqAt#Al+Turuq+At+i	DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN
nw/ann/02/ann	0290	0	7	الْيَوْمَ#yawom#Alywm#Al+yawom+a	DET+NOUN+CASE_DEF_ACC
nw/ann/02/ann	0290	0	8	وَ-#clitics#w#wa-	CONJ
nw/ann/02/ann	0290	0	9	-ًا#gad#gdA#-gad+AF	NOUN+CASE_INDEF_ACC
nw/ann/02/ann	0290	0	10	كَ-#clitics#k#ka-	PREP
nw/ann/02/ann	0290	0	11	-ـما#kamA#mA#-mA	SUB_CONJ
nw/ann/02/ann	0290	0	12	تَعَدَّ#taEah~ad#tEhd#taEah~ad+a	PV+PVSUFF_SUBJ:3MS
nw/ann/02/ann	0290	0	13	سابِقاً#sAbiq#sAbqA#sAbiq+AF	NOUN+CASE_INDEF_ACC
nw/ann/02/ann	0290	0	14	.#DEFAULT#.#.	PUNC

6	7	8	9	10	11	12 : N	N+1	
(TOP(S(VP*	l`kin~a	-	-	-	*	*	*	-
(NP*)	clitics	-	-	-	*	(ARG0*)	(ARG0*)	(6)
(S(VP(PRT*)	lam	-	-	-	*	*	*	-
*	>a$Ar	01	-	-	*	(V*)	*	-
(PP*	<ilaY	-	-	-	*	(ARG1*	*	-
(NP*	<iqofAl	-	-	-	*	*	*	-
(NP*)))	Tariyq	-	-	-	*	*)	*	-
(NP*	yawom	-	4	-	*	(ARGM-TMP*	*	-
*	clitics	-	-	-	*	*	*	-
*)	gad	-	-	-	*	*)	*	-
(PP*	clitics	-	-	-	*	(ARGM-MNR*	*	-
(SBAR*	kamA	-	-	-	*	*	*	-
(S(VP*	taEah~ad	01	-	-	*	*	(V*)	-
(NP*))))))))	sAbiq	-	-	-	*	*)	(ARGM-TMP*)	-
*))	DEFAULT	-	-	-	*	*	*	-

Figure 4.7 : Exemple d'un fichier en format conll [40]

L'auteur dans [40] a élaboré des données pour les utilisées dans un système d'annotation des rôles sémantiques pour la langue arabe, on passent par ces étapes :

1. Téléchargement du corpus OntoNotes 5.0 ;
2. Téléchargement de données et scripts depuis CoNLL-2012 ;
3. Fusion entre OntoNotes 5.0 et CoNLL sous un système Ubuntu ;
4. Construction d'une base de données en format CSV.

Afin d'effectuer le processus de classification et d'accéder aux meilleurs résultats possible, on a utilisé les modèles suivants :

- Arbre décision ;
- Naive Baie ;
- Random Forest ;
- Deep Learning.

Pour notre travail de classification des constituants de la phrase en rôles sémantiques (29 rôles, tableau 4.2), on a utilisé une base qui contient un total de 55.716 lignes (Entrainement et test).

Selon [40], M. Diab de l'université de George Washington (travaux cités précédemment) a utilisé 24.561 et 24 rôles sémantiques et absence des rôles C-ARG et R-ARG. Donc notre base est plus grande que celle utilisée dans ses travaux sur les rôles sémantiques arabes.

Tableau 4.2 : Rôles sémantiques dans notre corpus [40]

Arguments simples	ARG-M		C-ARG et R-ARG	
ARG0	ARGM-ADV	ARGM-PRP	C-ARG0	C-ARGM- PRP
ARG3	ARGM-LOC	ARGM-TMP	C-ARG1	C-ARGM-TMP
ARG1	ARGM-MNR	ARGM-COM	C-ARG2	R-ARG0
ARG4	ARGM-NEG	ARGM-CAU	C-ARGM-ADV	R-ARG1
ARG2	ARGM-GOL	ARGM-EXT	C-ARGM-LOC	R-ARG2
			ARGMADV(C-ARG2	R-ARGM-TMP
			ARGMADV(C-ARG1	R-ARGM-LO

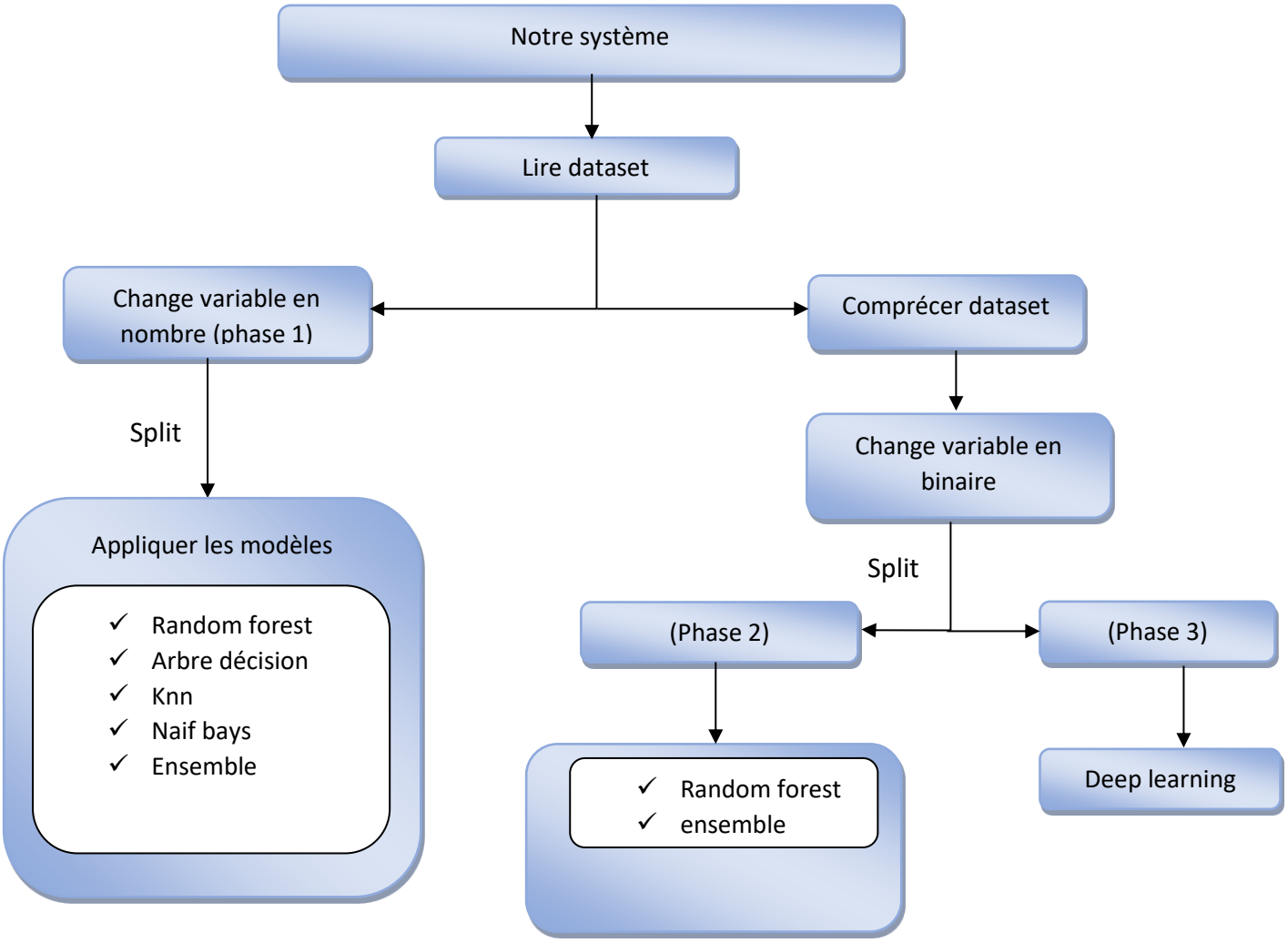
Dix colonnes importantes de la base de données ont été utilisées. Elles sont présentées dans ce tableau.

Tableau 4.3 : Les noms des colonnes de notre base de données

Colonne	Type
1	Attribut (verbe)
2	Lemma attribut frameset
3	ID frameset
4	1er constituant
5	Lemma 1er constituant
6	Position du 1er constituant par rapport à l'attribut
7	Endroit du 1er constituant par rapport à l'attribut
8	Parties de discours
9	Analyse
10	Argument

4- Expérimentations et interprétation

Pour l'expérimentation de notre système, on a fait plusieurs expériences. La première expérience dépend de la conversion des caractéristiques des variables en nombres. La deuxième expérience consiste à la conversion des variables en binaire, la dernière expérience est l'apprentissage profond, Tout est représenté dans ce graphique.



4.1- Première phase d'expérimentations

- **Arbres de décision**

Les arbres de décision sont utilisés après la fixation des valeurs pour `max_depth`[6]. Chaque fois que nous augmentons la valeur, on observe la différence et on adopte les deux fonctions de mesure de la qualité du `split` "*La fonction split () est l'opposé de la concaténation qui concatène de petites chaînes pour former une grande chaîne, tandis que split() est utilisé pour diviser une grande chaîne en sous-chaînes plus petites*" [104].

[6] Paramètre qui contrôle la taille de l'arbre en prenant des valeurs par défaut.

Les résultats obtenus sont présentés dans le tableau suivant :

Tableau 4.4 : Résultats arbres de décisions

Criterion \ Max_depth	5	10	20	50	75	85	150
Entropy	50.98%	62.25%	68.54%	68.66%	68.66%	68.66%	68.66%
Gini	49.46%	62.25%	69.34%	69.45%	69.45%	69.45%	69.45%

Dans ce tableau, si nous définissons des valeurs aléatoires pour `max_depth` avec le changement. Nous avons obtenu le pourcentage le plus faible **49,46 %** du `Criterion = gini` et `max_depth = 5`, ce qui nous amène à penser que ces caractéristiques n'aideront pas à atteindre une meilleure précision. Mais avec l'augmentation des valeurs `max_depth`, nous obtenons de meilleurs ratios **69,45%** à `max_depth = 50` et `Criterion = gini` avec des résultats stabilisants. Ainsi, ces propriétés peuvent aider à obtenir une meilleure précision.

❖ **Naïve Baie**

On a essayé le modèle de Naïve Baie sur notre système, on a obtenu une très petite précision de **2,79 %**. Ce très faible résultat montre que ce modèle n'est pas en cohérence avec notre travail. Parce qu'il fonctionne bien sur les petites données et non sur les grandes [105].

❖ **K plus proches voisins**

Les avantages du modèle K-PPV ont conduit à son utilisation dans à notre thèse. Un certain nombre de K-voisins a été utilisé, ainsi que l'utilisation de l'échelle de distance de Minkowski. Chaque fois qu'on augmente le nombre de voisins, une différence est constatée, comme indiqué dans le tableau suivant :

Tableau 4.5 : Résultats K plus proches voisins

Metric \ N_neighbors	1	2	5	10	20
Minkowski	60.63%	59.53%	58.38%	57.80%	56.16%

On remarque que plus le nombre de K voisins sera grand, plus la précision sera faible, donc la meilleure précision **60,63%** est atteinte avec K = 1.

❖ **Random Forest**

On a choisi d'utiliser ce modèle, puisqu'il est l'un des modèles les plus utilisés dans le cas de multiclasses. On a changé le nombre `n` d'arbres à proximité et à chaque fois on augmente les valeurs et on adapte les deux fonctions pour mesurer la qualité du split (gini, entropie), afin d'atteindre le meilleur résultat possible, cela est illustré dans le tableau suivant :

Tableau 4.6 : Résultats Random Forest

Criterion / N_estimators	Entropy	Gini
5	69.91%	69.18%
10	71.25%	71.42%
20	72.40%	72.37%
50	72.98%	73.22%
75	73.21%	73.23%
85	73.08%	73.28%
150	73.37%	73.54%
1000	-	73.85%
2000	-	73.90%
2500	-	73.90%
3000	-	73.86%

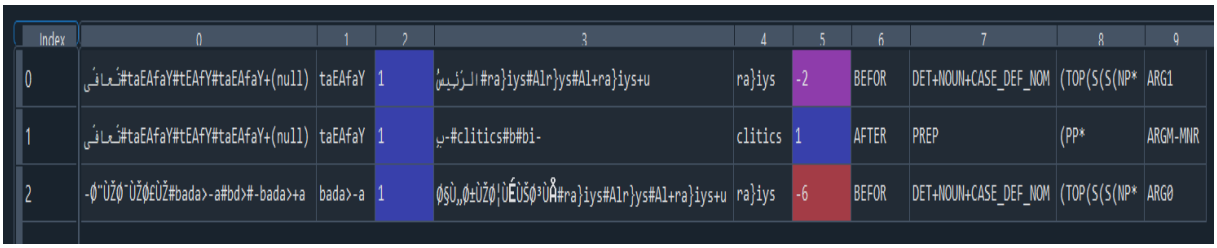
On remarque que l'augmentation de `n` engendre l'augmentation du résultat. Le meilleur rapport est **73,90%** avec n = 2500 et la fonction de division Gini. Cela montre que la forêt aléatoire est meilleure que les arbres de décision.

❖ **Hybrider un ensemble de modèles**

Dans ce travail, nous avons utilisé trois modèles ensemble pour l'apprentissage automatique (arbres de décision, k-PPV et forêt aléatoire) en tant qu'apprenants faibles. Nous avons créé chacun des trois modèles d'apprentissage automatique 3 fois, ce qui donne une combinaison d'un total de 27 apprenants faibles. Enfin, la méthode Max Voting Classifier est utilisée en combinant les résultats de chaque apprenant faible pour obtenir le résultat final, de sorte que le pourcentage obtenu était de **72,48 %**. Ensuite, nous avons moulé 5modèles pour chaque type (125 apprenants), le ratio était de **72,60 %**, puis nous avons supprimé K-PPV. Il était de **73,66 %**. [106]

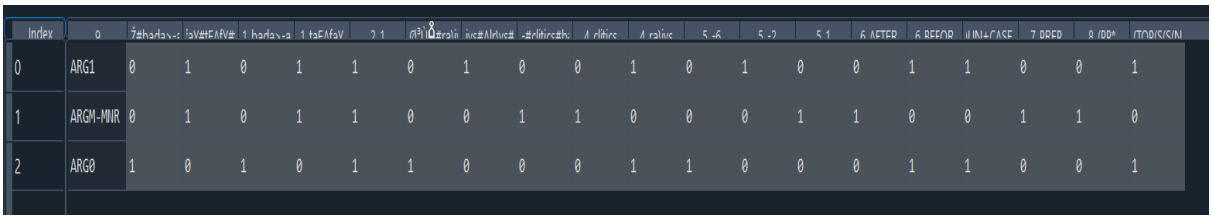
4.2- Deuxième phase d'expérimentations

À ce stade, on a converti les variables indépendantes (X) au binaire dans l'objectif d'obtenir de meilleur résultat. Comme indiqué dans les deux figures suivantes :



Index	0	1	2	3	4	5	6	7	8	9
0	ـنَـعـافَـى#taEAfaY#tEAfY#taEAfaY+(null)	taEAfaY	1	ٱلـرَّئِـيـسُ#ra}iys#Alr}ys#Al+ra}iys+u	ra}iys	-2	BEFOR	DET+NOUN+CASE_DEF_NOM	(TOP(S(S(NP*	ARG1
1	ـنَـعـافَـى#taEAfaY#tEAfY#taEAfaY+(null)	taEAfaY	1	بـ-#clitics#b#bi-	clitics	1	AFTER	PREP	(PP*	ARGM-MNR
2	-Ø"ÙŽØ¯ÙŽØ£ÙŽ#bada>-a#bd>#-bada>+a	bada>-a	1	Ø§Ù„Ø±Ø¦ÙŠØ³#ra}iys#Alr}ys#Al+ra}iys+u	ra}iys	-6	BEFOR	DET+NOUN+CASE_DEF_NOM	(TOP(S(S(NP*	ARG0

Figure 4.8 : Dataset avant le codage au binaire



Index	0	Ž#bada>-a	aY#tEAfY#	1 bada>-a	1 taEAfaY	2 1	Ø§Ù„Ø#rali	ys#Alrys#	-#clitics#b	4 clitics	4 ra}iys	5 -6	5 -2	5 1	6 AFTER	6 BEFOR	JUN+CASE	7 PREP	8 (PP*	(TOP(S(S(N
0	ARG1	0	1	0	1	1	0	1	0	0	1	0	1	0	0	1	1	0	0	1
1	ARGM-MNR	0	1	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0
2	ARG0	1	0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	1

Figure 4.9 : Dataset après le codage binaire

❖ **Random Forest**

On a noté dans la précédente expérience que lors de l'utilisation de la qualité de la division gini les résultats étaient meilleurs. Donc on a repris la même expérience avec la conversion des caractéristiques non approuvées au binaire. Les résultats sont les suivants :

Tableau 4.7 : Résultats Random Forest (binaire)

Criterion \ N_estimators	5	10	20	50	75	85
Gini	73.20%	75.02%	76.32%	77.16%	77.42%	77.37%

Le tableau a montré une amélioration remarquable du ratio, puisque notre meilleur résultat précédent était **73,66%**. Maintenant, on voit une augmentation du ratio de **77,42%** avec n = 75.

❖ **Faire un ensemble des modèles**

D'après les résultats de l'expérience précédente, on a constaté que lors du passage au binaire, il y avait une nette amélioration de la précision. Donc on décide de combiner à nouveau la forêt aléatoire et les arbres de décision où on a formé cinq modèles pour chaque espèce, mais malheureusement le ratio est tombé à **74,28%**.

4.3- Troisième phase d'expérimentation (Deep Learning)

Puisque le Deep Learning est une nouvelle approche de l'intelligence artificielle, nous allons l'expérimenter, sur un système d'annotation des rôles sémantiques en arabe.

Le modèle d'apprentissage profond qu'on utilise contient six couches (6 couches) : une couche d'entrée, quatre couches cachées et une couche de sortie. Paramètres tels que : fonction d'activation, nombre d'époques, taille du lot (batch size : le nombre d'échantillons qui seront propagés via le réseau) sont déterminés lors des expérimentations.

Tableau 4.8 : Résultats Deep Learning (binaire)

Taille du lot	Nb époques	1ère couche	2ème couche	3ème couche	4ème couche	Précision
125	5	578	160	_	_	80.60%
150	5	578	160	_	_	80.03%
175	5	578	160	_	_	79.80%
125	100	578	160	_	_	78.77%
125	10	578	160	80	_	79.83%
125	5	578	160	80	_	80.56%
125	5	1010	579	261	_	80.24%
125	5	1010	1000	261	_	81.18%
125	5	1010	1000	993	_	80.91%
174	5	1010	1000	993	_	81.29%
174	5	1010	1000	993	703	81.10 %
174	5	1500	1000	993	750	81.31%

Avec l'augmentation du nombre d'époques, le Batch size et l'utilisation de deux couches cachées, l'exactitude est réduite.

Mais quand on a utilisé une troisième couche cachée et le changement de la valeur de taille, on a remarqué une légère augmentation de la précision. On prend ces paramètres : taille = 174 et Np_époques = 5 l'amélioration est de **81.29%**.

Les résultats obtenus dans le tableau montrent que l'augmentation du nombre de neurones et de nombreuses couches cachées a eu un effet sur l'amélioration de la précision, même légèrement. Puisque notre pourcentage le plus élevé est de **81,31%,** on constate que les résultats obtenus sont meilleurs que ceux obtenus précédemment et cela est presque évident compte tenu de la nouveauté de l'apprentissage profond en intelligence et de sa contribution significative à l'amélioration des systèmes d'apprentissage.

5- Discussions et comparaisons

On a mené plusieurs expériences sur le système d'annotation des rôles sémantiques en arabe, on utilise une gamme de modèles d'apprentissage automatique : les forêts aléatoires, les arbres de décision et les K plus proches voisins individuellement, puis dans une fusion et enfin l'utilisation de l'apprentissage profond.

La première expérience vise à tester la performance globale d'un ensemble de modèles d'apprentissage automatiques en convertissant les données en nombres, puis en les combinant pour atteindre le meilleur pourcentage possible. Le résultat était de **73,66%,** en créant un ensemble d'arbres de décision et de forêts aléatoires.

La deuxième expérience a converti les fonctionnalités non approuvées au binaire et on a utilisé la forêt aléatoire seule. Le résultat était de **77.42%**. En raison de l'amélioration des résultats lors de la combinaison de la forêt aléatoire et les arbres de décision, on a fait la même chose. Mais malheureusement cette fois, il a donné un pourcentage inférieur **74.28%**.

Dans la troisième expérience, on a utilisé l'apprentissage profond, il a donné un résultat de **81.31%** et c'est le meilleur pourcentage obtenu.

Selon nos connaissances sur l'annotation des rôles sémantiques dans la langue, nous trouvons les travaux de M. Diab et H. Meguehout (présenté dans les chapitres précédents). On ne peut pas comparer avec les travaux du premier auteur, car on utilise un corpus plus récent et plus grand que celui utilisé dans ses travaux.

Pour les travaux du deuxième auteur, on a travaillé avec l'un de ses corpus élaborés. On a essayé un nombre plus grand d'algorithmes que cet auteur et certains algorithmes n'ont jamais été utilisés pour l'annotation des rôles sémantiques arabe. Il a obtenu une précision plus grande que la nôtre de **9%** pour l'algorithme des K plus proches voisins, cependant la comparaison est un peu difficile à cause de certaines différences dans son corpus et ses attributs. Pour le Deep Learning, c'est le même corpus dans une partie de ses tests et on note une légère augmentation dans nos résultats par rapport à lui.

6- conclusion

Ce dernier chapitre a permis de montrer notre contribution dans le traitement du langage naturel et l'annotation des rôles sémantiques en particulier, nos expériences et résultats encourageants d'autres auteurs à s'approfondir dans l'utilisation de nouvelles méthodes d'intelligence artificielle.

Conclusion et Perspectives

Nous avons abordé un sujet très peu abordable dans les travaux de recherches en informatique dans notre pays qui est le traitement automatique du langage naturel écrit et le sujet choisi « annotation des rôles sémantiques pour la langue arabe » présente peu de contribution. Cela engendrera des difficultés comme le manque de ressources annotées et un nombre restreint de chercheurs autour de sujet, etc.

Notre contribution dans l'annotation des rôles sémantiques arabes avec l'utilisation de méthodes d'intelligence artificielle montre la possible utilisation de méthodes à notre connaissance jamais utilisée pour l'annotation des rôles sémantiques arabes, comme les arbres de décision et la forêt aléatoire.

Cette contribution ouvre la voie pour l'utilisation de d'autres méthodes d'IA, surtout que ce champ de recherche a connu de nombreux progrès ces dernières années.

Pour les travaux futurs, nous comptons raffiner les paramètres des algorithmes qui ont eu de bons résultats et faire une hybridation de ces algorithmes. Aussi, nous essaierons de rédiger un article pour éventuelle conférence.

Références

[1] "TAL :Traitement Automatique des Langues," 2012 2011.

[2] F.-R. Chaumartin and P. Lemberger, *Le traitement automatique des langues: comprendre les textes grâce à l'intelligence artificielle*. Malakoff: Dunod, 2020.

[3] J. S. Gruber, "Studies in Lexical Relations," Thèse de Doctorat, Massachusetts Institute of Technlogy, États-Unis, 1965. [Online]. Available: https://books.google.dz/books?id=WRzPswEACAAJ

[4] C. J. Fillmore, "The case for case," États-Unis, 1968.

[5] M. Djemaa, "Stratégie domaine par domaine pour la création d'un FrameNet du français : annotations en corpus de cadres et rôles sémantiques," Thèse de Doctorat, Université Sorbonne Paris Cité, France, 2017. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01661689

[6] I. Kardana, I. Wrihatnala, and M. Satyawati, "Category Of Complement And Semantic Role Of Single Argument In Balinese Syntactic Constructions," *RETORIKA J. Ilmu Bhs.*, vol. 2, p. 384, Feb. 2017, doi: 10.22225/jr.2.2.67.384-393.

[7] M. D. Sidabutar and Z. Zakrimal, "Semantic Roles in Joko Widodo Re-Elected as President of BBC Online News," *Linguist. Engl. Educ. Art LEEA J.*, vol. 4, no. 1, pp. 56–65, Jul. 2020, doi: 10.31539/leea.v4i1.1362.

[8] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. Third Edition draft. 2017.

[9] F. Hadouche, G. Lapalme, and Marie-Claude L Homme, "Attribution de rôles sémantiques à des actants," 2011, doi: 10.13140/2.1.1847.1682.

[10] F. Hadouche, "Annotation syntaxico-sémantique des actants en corpus spécialisé," Thèse de Doctorat, Université de Montréal, Québec, Canada, 2011.

[11] Q. Pradet, "Annotation en rôles sémantiques du français en domaine spécifique," Thèse de Doctorat, Université Paris Diderot (Paris 7), France, 2015. [Online]. Available: https://hal.inria.fr/tel-01182711

[12] E. Boros, R. Besançon, O. Ferret, and B. Grau, "Étiquetage en rôles événementiels fondé sur l'utilisation d'un modèle neuronal," in *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France, Juillet 2014, pp. 25–35. [Online]. Available: http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-003

[13] W. Léchelle and P. Langlais, "Utilisation de représentations de mots pour l'étiquetage de rôles sémantiques suivant FrameNet," in *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France, Juillet 2014, pp. 36–45. [Online]. Available: http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-004

[14] O. Michalon, "Modélisation probabiliste de l'interface syntaxe sémantique à l'aide de grammaires hors contexte probabilistes Expériences avec FrameNet," in *Actes des 16e*

Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Marseille, France, Juillet 2014, pp. 1–12. [Online]. Available: http://www.atala.org/taln_archives/RECITAL/RECITAL-2014/recital-2014-long-001

[15] X. Carreras and L. Màrquez, "Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling," in *Proceedings of the Ninth Conference on Computational Natural Language Learning*, Stroudsburg, Pennsylvanie, États-Unis, 2005, pp. 152–164. [Online]. Available: http://dl.acm.org/citation.cfm?id=1706543.1706571

[16] M. Diab and A. Moschitti, "Semantic parsing of modern standard Arabic," in *International Conference Recent advances in natural language processing*, Borovets, Bulgaria, Sep. 2007, pp. 162–166.

[17] FRC CSC RAS / Moscow, Russia *et al.*, "Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates," in *Proceedings - Natural Language Processing in a Deep Learning World*, Oct. 2019, pp. 619–628. doi: 10.26615/978-954-452-056-4_073.

[18] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING-ACL '98: Proceedings of the Conference*, Montreal, Canada, 1998, pp. 86–90.

[19] P. Kingsbury and M. Palmer, "PropBank: the Next Level of TreeBank," p. 12.

[20] M. Diab, M. Alkhalifa, S. ElKateb, C. Fellbaum, A. Mansouri, and M. Palmer, "SemEval-2007 Task 18: Arabic Semantic Labeling," in *Proceedings of the Fourth International Workshop on Semantic Evaluations*, Prague, République tchèque, 2007, pp. 93–98. [Online]. Available: http://www.aclweb.org/anthology/S07-1017

[21] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press, 1998.

[22] T. Petrolito and F. Bond, "A Survey of WordNet Annotated Corpora," in *Proceedings of the Seventh Global Wordnet Conference*, Tartu, Estonia, Jan. 2014, pp. 236–245. Accessed: Sep. 01, 2022. [Online]. Available: https://aclanthology.org/W14-0132

[23] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.*, vol. 41, no. 2, Art. no. 2, Feb. 2009, doi: 10.1145/1459352.1459355.

[24] W. Léchelle, "Utilisation de représentations de mots pour l'étiquetage de rôles sémantiques suivant FrameNet," Thèse de Master, Université de Montréal, Québec, Canada, 2014.

[25] K. K. Schuler, "Verbnet: A Broad-coverage, Comprehensive Verb Lexicon," Thèse de Doctorat, Université de Pennsylvania, États-Unis, 2005. [Online]. Available: https://books.google.dz/books?id=VMY-OAAACAAJ

[26] A. Korhonen and T. Briscoe, "Extended Lexical-semantic Classification of English Verbs," in *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, Stroudsburg, Pennsylvanie, États-Unis, 2004, pp. 38–45. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596431.1596437

[27] B. J Dorr, M. Olsen, N. Habash, and S. Thomas, “LCS Database Documentation,” 2001. http://users.umiacs.umd.edu/~bonnie/Demos/LCS_Database_Documentation.html (accessed Aug. 10, 2022).

[28] M. Palmer, C. Bonial, and D. McCarthy, “SemLink+: FrameNet, VerbNet and Event Ontologies,” in *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, Baltimore, MD, USA, Jun. 2014, pp. 13–17. doi: 10.3115/v1/W14-3004.

[29] C. Bonial, O. Hargraves, and M. Palmer, “Expanding VerbNet with Sketch Engine,” in *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, Pisa, Italy, Sep. 2013, pp. 44–53. Accessed: Aug. 10, 2022. [Online]. Available: https://aclanthology.org/W13-5407

[30] W. Léchelle and P. Langlais, “Utilisation de représentations de mots pour l’étiquetage de rôles sémantiques suivant FrameNet,” in *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France, Juillet 2014, pp. 36–45. [Online]. Available: http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-004

[31] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky, “Support Vector Learning for Semantic Argument Classification,” *Mach. Learn.*, vol. 60, no. 1–3, Art. no. 1–3, Sep. 2005, doi: 10.1007/s10994-005-0912-2.

[32] V. Petukhova and H. Bunt, “LIRICS semantic role annotation: design and evaluation of a set of data categories,” p. 7.

[33] L. van der Plas and M. Apidianaki, “Cross-lingual Word Sense Disambiguation for Predicate Labelling of French,” in *Proceedings of TALN 2014 (Volume 1: Long Papers)*, Marseille, France, Jul. 2014, pp. 46–55. Accessed: Oct. 03, 2022. [Online]. Available: https://aclanthology.org/F14-1005

[34] G. Fliedner, *Tools for building a lexical semantic annotation*. Nancy, France, 2003. [Online]. Available: http://www.loria.fr/~duchier/Lorraine-Saarland/fliedner.pdf

[35] G. Melli *et al.*, “Description of squash, the sfu question answering summary handler for the duc-2005 summarization task,” presented at the DUC workshop 2005, Vancouver, Canada, 2005.

[36] H. C. Boas, “Bilingual FrameNet Dictionaries for Machine Translation,” in *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Espagne, 2002, pp. 1364–1371.

[37] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, “Using Predicate-Argument Structures for Information Extraction,” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, Jul. 2003, pp. 8–15. doi: 10.3115/1075096.1075098.

[38] R. Bunescu and R. Mooney, “A Shortest Path Dependency Kernel for Relation Extraction,” in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, Oct. 2005, pp. 724–731. Accessed: Aug. 10, 2022. [Online]. Available: https://aclanthology.org/H05-1091

[39] S. Narayanan and S. Harabagiu, “Question Answering Based on Semantic Structures,” in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, Aug. 2004, pp. 693–701. Accessed: Aug. 31, 2022. [Online]. Available: https://aclanthology.org/C04-1100

[40] H. MEGUEHOUT, “Un raisonnement à partir de cas (RàPC) pour la traduction automatique du langage naturel (de l’Arabe vers le Français),” Thèse de Doctorat, BADJI MOKHTAR, Annaba, Algérie, 2009.

[41] L. Suanmali, N. Salim, and M. S. Binwahlan, “SRL-GSM: A Hybrid Approach based on Semantic Role Labeling and General Statistic Method for Text Summarization,” *J. Appl. Sci.*, vol. 10, no. 3, Art. no. 3, Mar. 2010, doi: 10.3923/jas.2010.166.173.

[42] G. Melli, Z. Shi, Y. Wang, Y. Liu, A. Sarkar, and F. Popowich, “Description of squash, the sfu question answering summary handler for the duc-2006 summarization task,” presented at the DUC 2006, Brooklyn, Etat-Unis, 2006.

[43] P. Moreda, H. Llorens, E. Saquete, and M. Palomar, “The influence of Semantic Roles in QA: A comparative analysis,” *Proces. Leng. Nat.*, no. 41, Art. no. 41, 2008.

[44] R. Sun, J. Jiang, Y. F. Tan, H. Cui, T.-S. Chua, and M.-Y. Kan, “Using Syntactic and Semantic Relation Analysis in Question Answering,” presented at the TREC 2005, Gaithersburg, Maryland, États-Unis, 2005. [Online]. Available: http://trec.nist.gov/pubs/trec14/papers/nus.qa.pdf

[45] D. Shen and M. Lapata, “Using Semantic Roles to Improve Question Answering,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, République tchèque, Juin 2007, pp. 12–21. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1002

[46] A. Lakhfif and M. T. Laskri, “A frame-based approach for capturing semantics from Arabic text for text-to-sign language MT,” *Int. J. Speech Technol.*, vol. 19, no. 2, Art. no. 2, 2015, doi: 10.1007/s10772-015-9290-8.

[47] M. Bazrafshan and D. Gildea, “Semantic Roles for String to Tree Machine Translation,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, 2013, pp. 419–423. [Online]. Available: http://www.aclweb.org/anthology/P13-2074

[48] M. Bazrafshan and D. Gildea, “Comparing Representations of Semantic Roles for String-To-Tree Decoding,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1786–1791. doi: 10.3115/v1/D14-1188.

[49] odra Serge, “La langue arabe.,” *www.cosmovisions.com*, Sep. 01, 2022. https://www.cosmovisions.com/langueArabe.htm (accessed Sep. 01, 2022).

[50] M. Diab and A. Moschitti, “Semantic parsing of modern standard Arabic,” in *International Conference Recent advances in natural language processing*, Borovets, Bulgaria, Sep. 2007, pp. 162–166.

[51] S. Elkateb *et al.*, “Building a WordNet for Arabic,” presented at the LREC 2006, Genoa, Italy, May 2006. Accessed: Aug. 03, 2022. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/805_pdf.pdf

[52] P. Vossen, Ed., *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht, Pays-Bas: Springer Netherlands, 1998. doi: 10.1007/978-94-017-1491-4.

[53] S. Elkateb *et al.*, “Arabic WordNet and the Challenges of Arabic,” in *Proceedings of the International Conference on the Challenge of Arabic for NLP/MT*, London, UK, Oct. 2006, pp. 15–24. Accessed: Sep. 15, 2022. [Online]. Available: https://aclanthology.org/2006.bcs-1.2

[54] N. Loukil and K. Haddar, “Extracting HPSG Lexicon from Arabic VerbNet,” *Res. Comput. Sci.*, vol. 117, no. 1, Art. no. 1, Dec. 2016, doi: 10.13053/rcs-117-1-3.

[55] J. Mousser, “Classifying Arabic Verbs Using Sibling Classes,” 2011. Accessed: Sep. 16, 2022. [Online]. Available: https://aclanthology.org/W11-0142

[56] “VerbNet.” https://verbs.colorado.edu/verbnet/ (accessed Sep. 15, 2022).

[57] J. Ruppenhofer, M. Ellsworth, M. R. Petruck, C. R. Johnson, and J. Scheffczyk, *FrameNet II: Extended theory and practice*. Berkeley, California, États-Unis: Institut für Deutsche Sprache, Bibliothek, 2016.

[58] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The Berkeley FrameNet Project,” presented at the COLING 1998, 1998. Accessed: Aug. 10, 2022. [Online]. Available: https://aclanthology.org/C98-1013

[59] C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck, “Background to Framenet,” *Int. J. Lexicogr.*, vol. 16, no. 3, Art. no. 3, Sep. 2003, doi: 10.1093/ijl/16.3.235.

[60] G. Aston and L. Burnard, *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, 1998.

[61] E. Atwell and A. Sharaf, “A Corpus-based Computational Model for Knowledge Representation of the Quran,” Jan. 2009.

[62] A.-B. M. Sharaf and E. Atwell, “Knowledge representation of the Quran through frame semantics: A corpus-based approach,” Liverpool, Royaume-Uni, Jul. 2009. [Online]. Available: http://www.comp.leeds.ac.uk/scsams/papers/sharaf2009-cl2009.pdf

[63] N. Ghneim, E. Karhely, and W. Sa, “First Step of Building an Arabic FrameNet (AFN),” presented at the 13th International Business Information Management Conference, Marrakech, Maroc, 2009. [Online]. Available: https://www.researchgate.net/publication/284324899_First_Step_of_Building_an_Arabic_FrameNet_AFN

[64] Mohamed Maamouri (project head), Bies, Ann, Buckwalter, Tim, Jin, Hubert, and Mekki, Wigdan, “Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis).” Linguistic Data Consortium, Jun. 15, 2005. doi: 10.35111/GHRM-VT27.

[65] S. Al-Ghamdi, H. Al-Khalifa, and A. Al-Salman, “A Dependency Treebank for Classical Arabic Poetry,” p. 9.

[66] K. Dukes, E. Atwell, and A.-B. M. Sharaf, “Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank,” p. 6.

[67] K. Erk and S. Padó, “SHALMANESER - A Toolchain For Shallow Semantic Parsing,” presented at the LREC 06, Genoa, Italy, May 2006.

[68] H. Sun and D. Jurafsky, “Shallow Semantic Parsing of Chinese,” in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, Massachusetts, USA, May 2004, pp. 249–256. Accessed: Sep. 02, 2022. [Online]. Available: https://aclanthology.org/N04-1032

[69] M. Palmer *et al.*, “A Pilot Arabic Propbank,” Marrakech, Maroc, May 2008. [Online]. Available: http://www.aclweb.org/anthology/L08-1461

[70] W. Zaghouani, A. Hawwari, and M. Diab, “A Pilot PropBank Annotation for Quranic Arabic,” in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Montrèal, Canada, Jun. 2012, pp. 78–83. [Online]. Available: http://www.aclweb.org/anthology/W12-2511

[71] Maamouri, Mohamed *et al.*, “Arabic Treebank: Part 2 v 3.1.” Linguistic Data Consortium, p. 266240 KB, Aug. 15, 2011. doi: 10.35111/BWN1-3F08.

[72] W. Zaghouani, M. Diab, A. Mansouri, S. Pradhan, and M. Palmer, “The Revised Arabic PropBank,” in *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Suède, Jul. 2010, pp. 222–226. [Online]. Available: http://www.aclweb.org/anthology/W10-1836

[73] M. Diab, A. Moschitti, and D. Pighin, “Semantic Role Labeling Systems for Arabic using Kernel Methods,” in *Proceedings of ACL-08: HLT*, Columbus, Ohio, États-Unis, 2008, pp. 798–806. [Online]. Available: http://www.aclweb.org/anthology/P08-1091

[74] M. Diab, A. Moschitti, and D. Pighin, “CUNIT: a semantic role labeling system for modern standard Arabic,” in *Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval ’07*, Prague, Czech Republic, 2007, pp. 133–136. doi: 10.3115/1621474.1621500.

[75] M. Diab, A. Moschitti, and D. Pighin, “Semantic Role Labeling Systems for Arabic using Kernel Methods,” in *Proceedings of ACL-08: HLT*, Columbus, Ohio, États-Unis, 2008, pp. 798–806. [Online]. Available: http://www.aclweb.org/anthology/P08-1091

[76] H. Meguehout, T. Bouhadada, and M. T. Laskri, “Semantic Role Labeling for Arabic Language Using Case-based Reasoning Approach,” *Int. J. Speech Technol.*, vol. 20, no. 2, Art. no. 2, Jun. 2017, doi: 10.1007/s10772-017-9412-6.

[77] A. Cornuéjols, L. Miclet, and Y. Kodratoff, *Apprentissage artificiel : Concepts et algorithmes*, 2e éd. édition. Paris: Eyrolles, 2002.

[78] N. Marref, “Apprentissage Incrémental & Machines à Vecteurs Supports,” Thèse de Magister, HADJ LAKHDAR, Batna ,Algérie, 2013. [Online]. Available: Http://eprints.univ-batna2.dz/624/1/sce%20Marref%20Nadia.pdf

[79] P. Vincent, “Modèles à noyaux à structure locale,” Thèse de Doctorat, Université de Montréal, Montréal,canada, 2003.

[80] R. Mifdal, “Application des techniques d’apprentissage automatique pour la prédiction de la tendance des titres financiers,” Thèse de Magister, ÉCOLE DE TECHNOLOGIE SUPÉRIEURE UNIVERSITÉ DU QUÉBEC, Québec, Canada, 2019.

[81] “Apprentissage Supervisé : Introduction,” *Machine Learnia*, Jul. 02, 2019. https://machinelearnia.com/apprentissage-supervise-4-etapes/ (accessed Aug. 09, 2022).

[82] J. Ah-Pine, “Apprentissage automatique,” p. 90.

[83] E. Universalis, “RÉSEAUX DE NEURONES FORMELS,” *Encyclopædia Universalis*. https://www.universalis.fr/encyclopedie/reseaux-de-neurones-formels/ (accessed Sep. 01, 2022).

[84] “Définition | Deep Learning - Apprentissage profond | Futura Tech,” Aug. 31, 2022. https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/ (accessed Aug. 31, 2022).

[85] Gaël, “8 Algorithmes de Machine Learning expliqués en Language Humain,” *Datakeen*, Nov. 06, 2017. https://www.datakeen.co/8-machine-learning-algorithms-explained-in-human-language/ (accessed Aug. 09, 2022).

[86] “Panorama des algorithmes de machine learning.” https://www.journaldunet.com/solutions/dsi/1209100-panorama-des-algorithmes-de-machine-learning/ (accessed Sep. 19, 2022).

[87] “Naive Bayes classifier.” [Online]. Available: https://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf

[88] M. Taffar, “Initiation à l’apprentissage Automatique,” p. 82.

[89] D. Gildea and J. Hockenmaier, “Identifying Semantic Roles Using Combinatory Categorial Grammar,” in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 57–64. Accessed: Aug. 31, 2022. [Online]. Available: https://aclanthology.org/W03-1008

[90] S. S. Pradhan, W. H. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky, “Shallow Semantic Parsing using Support Vector Machines,” in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, Massachusetts, USA, May 2004, pp. 233–240. Accessed: Aug. 31, 2022. [Online]. Available: https://aclanthology.org/N04-1030

[91] A. Moschitti, D. Pighin, and R. Basili, “Tree Kernels for Semantic Role Labeling,” *Comput. Linguist.*, vol. 34, no. 2, pp. 193–224, 2008, doi: 10.1162/coli.2008.34.2.193.

[92] R. Collobert and J. Weston, “Fast Semantic Extraction Using a Novel Neural Network Architecture,” in *Proceedings of the 45th Annual Meeting of the Association of*

Computational Linguistics, Prague, Czech Republic, Jun. 2007, pp. 560–567. Accessed: Aug. 31, 2022. [Online]. Available: https://aclanthology.org/P07-1071

[93] B. L, “Python : tout savoir sur le principal langage Big Data et Machine Learning,” *LeBigData.fr*, Sep. 10, 2022. https://www.lebigdata.fr/python-langage-definition (accessed Sep. 19, 2022).

[94] la rédaction de Futura, “Définition | Python | Futura Tech,” *Futura*. https://www.futura-sciences.com/tech/definitions/informatique-python-19349/ (accessed Sep. 19, 2022).

[95] M. Loukides, “Where Programming, Ops, AI, and the Cloud are Headed in 2021,” *O’Reilly Media*, Jan. 25, 2021. https://www.oreilly.com/radar/where-programming-ops-ai-and-the-cloud-are-headed-in-2021/ (accessed Sep. 19, 2022).

[96] “TIOBE Index,” *TIOBE*. https://www.tiobe.com/tiobe-index/ (accessed Sep. 19, 2022).

[97] “Anaconda Python Tutorial: Everything You Need to Know - DZone Big Data,” *dzone.com*. https://dzone.com/articles/python-anaconda-tutorial-everything-you-need-to-kn (accessed Sep. 05, 2022).

[98] D. A, “Spyder, un puissant environnement de développement interactif pour Python,” *Ubunlog*, Dec. 09, 2017. https://ubunlog.com/fr/environnement-de-d%C3%A9veloppement-spyder-python/ (accessed Sep. 05, 2022).

[99] “TensorFlow : Deep Learning sous Python.” https://blent.ai/tensorflow-deep-learning-python/ (accessed Sep. 05, 2022).

[100] “Quels sont les frameworks utilisés en Deep Learning ?,” *Mobiskill*, May 25, 2021. https://mobiskill.fr/blog/conseils-emploi-tech/quels-sont-les-frameworks-utilises-en-deep-learning/ (accessed Sep. 05, 2022).

[101] Pierre, “Installation de Anaconda - Rodeo - TensorFlow et Keras sous Windows 10,” *Anakeyn*, Sep. 15, 2018. https://www.anakeyn.com/2018/09/15/install-anaconda-rodeo-tensorflow-keras-windows-10/ (accessed Sep. 05, 2022).

[102] W. Ralph *et al.*, “OntoNotes Release 5.0 with OntoNotes DB Tool v0.999 beta.” Sep. 28, 2012.

[103] P. Sameer, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, “CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes,” in *Joint Conference on EMNLP and CoNLL - Shared Task*, Jeju Island, Corée, Jul. 2012, pp. 1–40. [Online]. Available: http://www.aclweb.org/anthology/W12-4501

[104] R. Hajji, “Comment utiliser Split en Python,” *apcpedagogie*, Apr. 04, 2021. https://apcpedagogie.com/comment-utiliser-split-en-python/ (accessed Sep. 20, 2022).

[105] “خوارزمية Naive Bayes المبرمج العربي - مع مزايا وعيوب الخوارزميات الشائعة في التعلم الآلي.” https://www.arabicprogrammer.com/article/3851956356/ (accessed Sep. 19, 2022).

[106] “A Hands-on Guide To Hybrid Ensemble Learning Models, With Python.” https://analyticsindiamag.com/a-hands-on-guide-to-hybrid-ensemble-learning-models-with-python-code/ (accessed Sep. 20, 2022).

Annexes

Annexe A : Portions du code Random Forest

```
# -*- coding: utf-8 -*-
import pandas as pd
#read the dataset
dataSet = pd.read_csv('Data_Big_New_v11.0.csv',header=None)

# check the memory usage of our original dataset
BYTES_TO_MB_DIV = 0.000001
def print_memory_usage_of_data_frame(df):
    mem = round(df.memory_usage().sum() * BYTES_TO_MB_DIV, 3)
    print("Memory usage is " + str(mem) + " MB")

print_memory_usage_of_data_frame(dataSet)

from IPython.display import display
data_one_hot = pd.get_dummies(dataSet, columns=[0,1,2,3,4,5,6,7,8], sparse=True)
display(data_one_hot.head())
print_memory_usage_of_data_frame(data_one_hot)

#split the dataset into x(indibanded variable),y(dibanded variable)
X = data_one_hot.iloc[:, 1:].values
y = data_one_hot.iloc[:, 0].values

#encoding dataset
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
    #encoding y
y = le.fit_transform(y)

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

#forest
# Training the Random Forest Classification model on the Training set
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators =2500, criterion = 'gini', random_state = 0)
classifier.fit(X_train, y_train)
# Predicting the Test set results
y_pred = classifier.predict(X_test)

from sklearn.metrics import accuracy_score
asm=accuracy_score(y_test,y_pred)
print(asm)
```

Annexe B : Portions du code Deep Learning

```
   # build a model
model = Sequential()

model.add(Dense(1500, input_shape=(X_train.shape[1],), activation='relu')) # inpu
model.add(Dense(1000,kernel_initializer='glorot_uniform' , activation='relu'))
model.add(Dense(993,kernel_initializer='glorot_uniform' , activation='relu'))
model.add(Dense(750,kernel_initializer='glorot_uniform' , activation='relu'))
model.add(Dense(29,kernel_initializer='glorot_uniform', activation='softmax'))
model.summary()

   # compile the model
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy', # this is different instead
              metrics=['accuracy'])

model.fit(X_train,y_train,batch_size=175,epochs=5)
```

```
 Layer (type)                 Output Shape              Param #
=================================================================
 dense (Dense)                (None, 1500)              42972000

 dense_1 (Dense)              (None, 1000)              1501000

 dense_2 (Dense)              (None, 993)               993993

 dense_3 (Dense)              (None, 750)               745500

 dense_4 (Dense)              (None, 29)                21779

=================================================================
Total params: 46,234,272
Trainable params: 46,234,272
Non-trainable params: 0
```

```
Epoch 1/5
257/257 [==============================] - 163s 629ms/step - loss: 0.9896 -
accuracy: 0.6869
Epoch 2/5
257/257 [==============================] - 165s 640ms/step - loss: 0.4107 -
accuracy: 0.8753
Epoch 3/5
257/257 [==============================] - 179s 698ms/step - loss: 0.1952 -
accuracy: 0.9394
Epoch 4/5
257/257 [==============================] - 173s 675ms/step - loss: 0.1101 -
accuracy: 0.9673
Epoch 5/5
257/257 [==============================] - 171s 665ms/step - loss: 0.0711 -
accuracy: 0.9788
```

Annexe C : Portions du code Arbres de décision

```
# -*- coding: utf-8 -*-
import numpy as np
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
# from keras.wrappers.scikit_learn import KerasClassifier
# from keras.utils import np_utils

import pandas as pd
#read the dataset
dataSet = pd.read_csv('Data_Big_New_v11.0.csv',header=None)
# split the dataset into x(indibanded variable),y(dibanded variable)
X = dataSet.iloc[:, :-1].values
y = dataSet.iloc[:, -1].values
#encoding dataset
le = LabelEncoder()
    #encoding y
y = le.fit_transform(y)
    #encoding X
X[:,0]=le.fit_transform(X[:,0])
X[:,1]=le.fit_transform(X[:,1])
X[:,3]=le.fit_transform(X[:,3])
X[:,4]=le.fit_transform(X[:,4])
X[:,6]=le.fit_transform(X[:,6])
X[:,7]=le.fit_transform(X[:,7])
X[:,8]=le.fit_transform(X[:,8])
# Feature Scaling
from sklearn.preprocessing import MinMaxScaler
sc_X = MinMaxScaler()
X = sc_X.fit_transform(X)
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)


#ARBRE D
# Training the Decision Tree Classification model on the Training set
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(max_depth =5,criterion = 'gini', random_state = 0)

classifier.fit(X_train, y_train)
# Predicting the Test set results
y_pred = classifier.predict(X_test)

from sklearn.metrics import accuracy_score
asm=accuracy_score(y_test,y_pred)
```