

**République Algérienne Démocratique et populaire**

Ministère de l’Enseignement Supérieur et de la Recherche

Scientifique

**Université des Sciences et de la Technologie Houari Boumediene**

Faculté d’Electronique et Informatique

Département Informatique

Master Systèmes Informatiques Intelligents  
  
Rapport de projet

Extraction des entités nommées (Named Entity Extraction) pour la langue arabe

*Module : Traitement Automatique du Langage Naturel*

Réalisé par :

SELMANE Fayçal

Année scolaire :

2019/2020

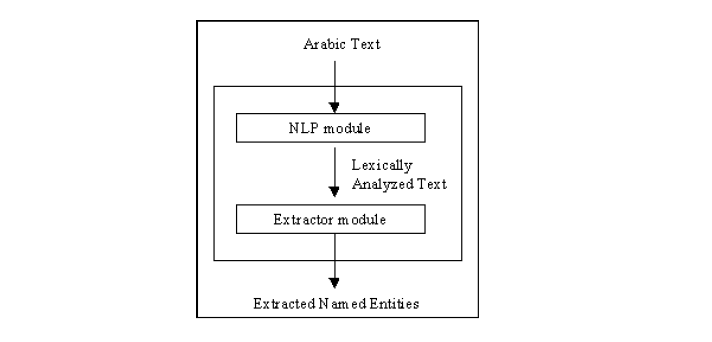
**Sommaire**

1. **Introduction** …………………………………………………………………………………………………………………………………1
2. Problématique ……………………………………………………………………………………………………………………………………

**Table des figures :**

*Figure1 : …………………………………………………………………………………………………5*

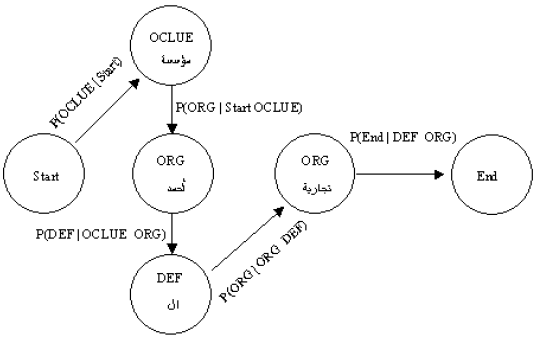
*Figure2 : …………………………………………………………………………………………………6*

1. **Introduction**Depuis sa création, l’être humain a réalisé la nécessité de communiquer avec les autres et d’interagir avec son environnement, parce que l’homme est un être social qui tire l’essentiel de sa force du groupe auquel il appartient. L’échange tient dans son existence une place primordiale. Il en est ainsi depuis l’origine et il en sera ainsi pout toujours[1].   
   A cet effet, l’être humain utilisait différentes façons pour s’exprimer, en s’appuyant sur la communication par signes, l’utilisation d’un ensemble d’entités sonores, et arrivant à l’établissement de langage humaine naturel de nos jours. Ce dernier s’avère être le moyen de communication le plus répandu chez l’homme et est considérée comme étant l’élixir de la vie pour lui.  
   La langue humaine à nos jours reste aussi si précieuse, pas pour l’humain mais pour surtout pour les machines. Le volume contextuel électronique immense qui évolue exponentiellement de plus en plus fut une mine d’or pour les machines qui s’appuient sur l’intelligence artificielle afin de comprendre et analyser le langage humain. Ce qui donné naissance un vaste domaine de recherche connu par le traitement automatique du langage naturel (TALN) ou NLP (Natural Language Processing) en anglais.  
   Le traitement automatique de la langue naturelle1 est au fait un domaine multidisciplinaire qui regroupe la linguistique, l'informatique et l'intelligence artificielle. L’intention est de créer des outils de traitement de la langue naturelle pour diverses applications [2].  
   Ce domaine de recherche a affronter et affronte plusieurs défis, non seulement vis-à-vis de l’évolution massive des flux de données, mais surtout par rapport à l’émetteur de ces sources d’informations : L’humain ! En effets, l’homme est la structure la plus complexe qui fait un pont entre un état intérieur implicite et un environnement très variant qui influe sur son comportement. Ceci complique de plus en plus la tâche et montre l’extraction de telles connaissances comme étant le mirage d’un oasis qui est très loin d’être atteint.  
   L’intelligence artificielle devant ce challenge continu tente rigoureusement à automatiser ce traitement d’une manière la plus précise possible dans l’intention de concrétiser un dialogue entre l’homme et la machine dont l’être humain a toujours rêver de le réaliser.  
   Ce domaine de recherche a finalement pu voir la lumière, grâce à des recherches et expérimentations intensives. Les chercheurs ont atteint un très grand progrès concrétisé par l’analyseur « Alibaba » qui a pu réaliser un meilleur score qu’un humain au sein de l’université de Stanford [3]. De nouvel horizons sont apparus et la course de l’innovation et de la recherche continue a été fortement lancée par les chercheurs  
   afin de booster ce domaine et enrichir les corpus des différentes langues et faire apprendre nos future voisins.  
   L’intelligence artificielle, étant l’axe centrale de toute science au monde actuel, se trouve devant la nécessité de satisfaire régulièrement les besoins humains, et nous -par conséquences- sommes aussi en face de ce défi continu dans l’intention à toujours promouvoir la recherche et l’innovation.
2. **Problématique**
3. **Etude de l’existant : SAIE**
   1. **Introduction**  
      Le SAIE (Statistical Arabic Information Extraction) est un système d’extraction d’information arabe, développé Par Pr. Guessoum and Dr. Al Shamsi en 2005 [X1]. Le système est fondu sur des méthodes statistiques et linguistiques : une approche d’extraction à base règle, une deuxième basé sur des modèles probabilistes stochastiques, et une troisième statistique aussi mais basée sur les chaines de Markov cachées ou HMM (Hidden Markov Model) en anglais[X2]. Ceci en réponse aux besoins d’un modèle plus robuste pour l’étiquetage morphosyntaxique des phrases arabes, et l’extraction des entités nommées. La tâche étant plus complexe vu les variations et les irrégularités des entités arabes, elle était donc implémenté dans un système composé pour y remédier.
   2. **Architecture de SAIE**  
      Le système, afin de pouvoir appliquer son modèle probabiliste (HMM) pour l’extraction des entité nommées, a été construit en deux modules : Un modèle NLP conçu pour l’analyse et l’extraction lexicale et syntaxique en effectuant un « POS-Tagging » dans une phase hors-lignes d’entrainement, et un module d’extraction des entités nommées s’appuyant sur le premier module [X3]. L’architecture globale de ce système est récapitulée dans le schéma suivant :  
        
      

*Figure 3.1 : Structure générale de SAIE[X3].*

* 1. **Le module d’extraction des entités nommées**  
     Le seconde module du système développé est dédié à la reconnaissance des entités nommées. Un ensemble d’étiquettes des entités nommé (Named Entities tag set) a été pris comme base de segmentation. Cet ensemble contient 12 étiquettes en plus de l’étiquette OTHER qui représente une entité autre que celles définies (elle peut être un verbe, adjectif, une entité nommé non inclue dans l’ensemble, etc…).  
       
     A partir de cet ensemble, le modèle en question est entrainé sur un corpus lexical correspondant, auquel le premier module peut extraire les différentes étiquettes. L’apprentissage consiste à construire une table de la couche sous-jacente appelée la Matrice de transition, et une autre table relative aux observations des entités nommées appelée la matrice d’émission ou de vraisemblance d’observations. Le principe consiste à implémenter un modèle de Markov qui exploite les deux matrices afin de déduire la séquence d'états sous-jacents à partir des évènements observés [X4]. La séquence en sortie est une liste des étiquettes où chacune a la probabilité conditionnelle la plus élevée.  
       
     Le modèle implémenté dans l’extracteur des NE dans le SAIE est à base de trigrammes. Il s’ensuit qu’il est possible que certaines trigrammes ne figurait pas dans le Dataset de l’apprentissage, et donc un modèle de couverture/soutien (back-off) à base de bigrammes pour ne pas avoir une probabilité nulle qui engendra une succession nulle. De même, certaines bigrammes peuvent ne pas exister dans le corpus d’apprentissage et à nouveau, un modèle à base unigrammes est introduit comme un dernier back-off.

1. **Le modèle de Markov Caché (HMM)**  
   Comme mentionné précédemment, le SAIE se basait un modèle statistique génératif nommé HMM.
   1. **Définition**« Un HMM est un modèle statistique dans lequel le système modélisé est supposé être un processus markovien de paramètres inconnus. Contrairement à une chaîne de Markov classique, où les transitions prises sont inconnues de l'utilisateur mais où les états d'une exécution sont connus, dans un modèle de Markov caché, les états d'une exécution sont inconnus de l'utilisateur »[X5].
   2. **Matrice de transition et Matrice d’émission**  
      Un HMM est définit par [X6]:
      * + Un ensemble d’états Q=q1,q2,….,qN.
        + Un ensemble d’observations O=o1,o2,…,oM
        + Une matrice de probabilitéde transition A={aij}, tq *aij=P(qt=j|qt-1=i) 1≤i,j≤N*
        + *Une matrice de probabilté de vraissemblance d’observation B={bi(k)},tq bi(k)=P(Xt=ok|qt=i).*
        + *Un vecteur de probabilité initial ӆ ӆi=P(q1=i) 1≤i≤N.*

La matrice de transition A est une matrice carrée de taille NxN contenant les probabilités de transition d’un état à un autre, en d’autres termes, la probabilité de passer d’une étiquette à une autre. Par exemple : la probabilité qu’un nom suit un verbe, ou qu’un adjectif suit un nom suivi d’un verbe, etc…  
La matrice d’émission B est une matrice de taille NxM représentant les probabilités d’émettre chacune des observations par chacun des états.  
  
La figure suivante illustre la manière dont le modèle de Markov caché combine coordonne entre la couche sous-jacente (les probabilités de transition) et la couche surfacique (les probabilités d’observations) []:  
   
 

* 1. **Algorithme de Viterbi**  
     Le calcul de la séquence la plus probable t1,….,tj se divise en deux parties : calculer la probabilité de la meilleure s séquence d’étiquettes jusqu’à l’état j-1, puis calculer la meilleure probabilité de transition de vers l'état j, et le produit des deux est multiplié par la probabilité du mot observé.  
     Un calcul intuitif et intensif exploitant toutes les combinaisons possibles et déterminant le chemin avec la meilleure probabilité peut conduire facilement à une explosion combinatoire lorsque le corpus contient beaucoup d’états.  
     L’algorithme de *Viterbi* propose une solution efficace en s’appuyant sur la programmation dynamique. Il permet de sauvegarder le meilleur résultat qui mène à l’état courant, et effectuer les combinaisons de transition depuis cet état, tout en gardant la trace des meilleurs étiquettes parcourues.  
     L’algorithme de *Vitervi* est assez simple et facile à implémenter. L’ossature de cet algorithme est associée dans l’annexe de ce rapport **(voir annexe )**.

**Références et bibliographie**

[1]. Jean-Paul Pigasse ,« Les révolutions de la communication », 3ème-4ème trimestre 1981. pp. 9-20.  
 <https://www.persee.fr/doc/colan_0336-1500_1981_num_50_1_3478>

[2]. « Traitement automatique du langage naturel », sur Wikipédia.  
 <https://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel>

[3]. Matthew Levinson,« Alibaba's AI Outguns Humans in Reading Test»

<http://news.fintech.io/post/102eoew/alibabas-ai-outguns-humans-in-reading-test>

[X1].   
[X2]. « A hidden Markov model-based POS tagger for Arabic», article de conférence, par Ahmed Guessoum et Fatma Al Shamsi.

<https://www.researchgate.net/publication/228666837_A_hidden_Markov_model-based_POS_tagger_for_Arabic>

[X3]. « SAIE: Statistical Arabic Information Extraction System», p.58, thèse de doctorat, par Fatma Nasser Al Shamsi

[X4]. « Modèle de Markov caché », sur Wikipédia.

<https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_Markov_cach%C3%A9>

[X5]. « Modèle de Markov caché », sur Wikipédia.

<https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_Markov_cach%C3%A9>

[6] .« Etiquetage morphosyntaxique », cours TALN, Chapitre 6, *A.Guessoum*, basé sur « Speech and Language Processing» par Jurafsky and Martin