الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

جامعة وهران للعلوم و التكنولوجيا محمد بوضياف

THÈSE

En vue de l'obtention du

Diplôme de Doctorat en Sciences

Présenté par : LAKEL Kheira

Intitulé

Les annotations sémantiques dans les documents Web : application aux textes psychologiques en langue arabe

Faculté ***: Mathématiques et Informatique***

Département ***:Informatique***

Spécialité ***: Informatique***

Option ***: Informatique***

Devant le Jury Composé de :

Membres de Jury	*Grade*	*Qualité*	*Domiciliation*
RAHAL Sid Ahmed	*Pr*	*Président*	*USTO-MB*
BENDELLA Fatima	*Pr*	*Encadrant*	*USTO-MB*
BENAISSA Moussa	*Pr*	*Examinateurs*	*U-Oran1*
BARIGOU Fatiha	*MCA*		*U-Oran1*
MEKKAKIA MAAZA Zoulikha	*MCA*		*USTO-MB*
TLEMSANI Radouane	*MCA*		*INTIIC Oran*

Année Universitaire : 2017-2018

Résumé : La reconnaissance d'entités nommées est une composante essentielle du traitement (bio)médical du langage naturel, permettant l'extraction d'informations et la découverte de connaissances à partir de textes. Généralement, les études réalisées concernant l'extraction de l'information (bio)médicale ont été développées en anglais et dans certaines langues. Cependant, aucune étude n'a été développée en langue arabe. Pour cela, la langue arabe doit effectuer plus de recherches dans ce domaine et par conséquent nous avons introduit une approche d'extraction d'informations psychologiques. Cette recherche consiste en la reconnaissance des entités psychologiques et l'extraction des relations à partir du texte. Deux techniques ont été appliquées pour le processus de reconnaissance : la première condition préalable à la technique dépendait entièrement de l'identification directe avec l'utilisation des Gazetteers et la deuxième technique est un modèle basé sur des règles dans lequel les techniques sont construites sur la base des nomenclatures. Les expériences donnent F-mesure globales de 86, 407%. Et pour lier les NERs psychologique, nous avons intégré une formulation de programmation linéaire. Au meilleur de nos connaissances, c'est la première approche sur l'extraction d'information psychologique pour inclure dans l'état de l'art des travaux effectués en la langue arabe.

Les mots clés: Programmation Linéaire en Nombre Entier, Reconnaissance des Entités Nommées en Arabe, Extraction des Relations, Extraction des Informations Psychologiques.

Abstract: Named entity recognition is a crucial component of (bio)medical natural language processing, enabling information extraction and knowledge discovery from text. Generally, the achieved studies concerning the (bio)medical information extraction were developed in English and some languages. However, there is no study that was developed in the Arabic language. For this, the Arabic Language needs to perform more researches in this area and hence we introduced a Psychological information extraction approach. This research consists of psychological entities recognition and Relation Extraction from the text. Two techniques were applied for the recognition processes: the first requirement prior to the technique was completely dependent on direct identification with the utilization of gazetteers and the second technique is a rule-based model in which rules are techniques were put constructed on the basis of a gazetteers list. Experiments yield the overall F–measure values of 86,407%. And for joint Psychological NER we integrate a Linear Programming Formulation. On the best of our knowledge, this is the first approach on psychological information Extraction to include the state of the work done for Arabic language.

Keywords: Integer Linear Programming, Arabic Named Entity Recognition, Relation Extraction, Psychological Information Extraction.

التلخيص: يعتبر التعرف على الكيانات المسماة مكونا أساسيا في معالجة اللغة الطبيعية الطبية (الحيوية) ، مما يتيح استخراج المعلومات واكتشاف المعرفة من النص. عموما، تم تطوير الدراسات التي تم التوصل إليها بشأن استخراج المعلومات الطبية (الحيوية) باللغة الإنجليزية وبعض اللغات الأخرى. ومع ذلك، لا توجد أي دراسة تم تطويرها باللغة العربية. لهذا، يجب أن تكون اللغة العربية أكثر فاعلية في هذا المجال، ومن ثم فإننا أدرجنا نهج استخراج المعلومات النفسية. هذا البحث يهتم بالتعرف على الكيانات النفسية واستخراج العلاقات من النص. طُبقت اثنين من التقنيات لعملية الاعتراف: فالمتطلب الأول هو أن التقنية تعتمد بشكل كامل على التحديد المباشر لاستخدام قوائم مصطلحات والتقنية الثانية هي نموذج قائم على القواعد حيث تم فيه وضع القواعد على أساس قائمة المعاجم. تنتج عن التجارب القيمة الإجمالية 86,407%. و من أجل وصل المصطلحات النفسية، قمنا بإدماج البرمجة الخطية. على حد علمنا، هذا هو النهج الأول الذي يهتم باستخراج المعلومات النفسية. و ذلك من أجل ضمه إلى ما تم إجراؤه للغة العربية.

الكلمات المفتاحية: البرمجة الخطية الصحيحة، التعرف على الكيانات العربية ، استخلاص العلاقة ، استخراج المعلومات النفسية.

« Une thèse est une sorte d'égoïsme en soi, eh oui, la plume transformé en clavier d'ordinateur de nos jours, demande silence et solitude pour la faire danser, sur un papier blanc, qu'on ne touche jamais, mais on a l'illusion de le voir monter et descendre dans l'écran, je remercie donc les personnes qui ont compris le besoin de ma solitude, et non rien obtenu en échange. » Leoncio JIMÉNEZ CANDIA

Remerciements

TOut d'abord, merci à *Allah* pour m'avoir donné le pouvoir et l'aide accomplir cette recherche. Sans la grâce d'Allah, je n'ai pas pu accomplir ce travail.

Je tiens tout d'abord à remercier ma directrice de recherche, Mme **Fatima BENDELLA**, professeur à l'USTO-MB d'avoir dirigé cette thèse, pour son soutien, sa disponibilité, son aide, son encouragement dans les moments difficiles et durant toute la période de ce travail et les nombreuses discussions qui m'a permis d'y voir plus clair au sujet.

Mes vifs remerciements vont également aux membres du jury, Mr **RAHAL Sid Ahmed**, Professeur à l'USTO-MB, Mr **BENAISSA Moussa**, Professeur à l'Université d'Oran1, Mme **BARIGOU Fatiha**, Maitre de Conférences A à l'Université d'Oran1, Mme **MEKKAKIA MAAZA Zoulikha** Maitre de Conférences A à l'USTO-MB et Mr **TLEMSANI Redouane** Maitre de Conférences A à l'INTIIC Oran pour l'intérêt qu'ils ont porté à mon travail en acceptant d'examiner cette thèse et de l'enrichir par leurs propositions.

Je tiens tout spécialement à remercier mon Mari pour son soutien et sa patience tout au long de cette thèse.

J'adresse aussi mes remerciements et ma profonde gratitude à ma mère qui n'a jamais cessé de me soutenir pour que je puisse finir mes études. Merci aussi à toute ma famille.

Personnellement, je ne saurais exprimer assez de reconnaissance pour mon père «Que Dieu lui apporte paix et miséricorde » pour son amour et soutien qui n'ont cessé de jalonner ma vie. Merci d'avoir cru en moi, je vous dédie ma vie et humblement ce modeste effort. Les mots ne peuvent exprimer combien je suis reconnaissante envers mon père.

Enfin, un remerciement spécial à mes enfants **Alaa** et **Abdel rahmane** qui sont ma source de bonheur. Merci à tous ceux qui, de près ou de loin, m'ont encouragé et soutenu tout au long de ce travail.

Résumé

La reconnaissance d'entités nommées est une composante essentielle du traitement (bio)médical du langage naturel, permettant l'extraction d'informations et la découverte de connaissances à partir de textes. Généralement, les études réalisées concernant l'extraction de l'information (bio)médicale ont été développées en anglais et dans certaines langues. Cependant, aucune étude n'a été développée en langue arabe. Pour cela, la langue arabe doit effectuer plus de recherches dans ce domaine et par conséquent nous avons introduit une approche d'extraction d'informations psychologiques. Cette recherche consiste en la reconnaissance des entités psychologiques et l'extraction des relations à partir du texte. Deux techniques ont été appliquées pour le processus de reconnaissance : la première condition préalable à la technique dépendait entièrement de l'identification directe avec l'utilisation des Gazetteers et la deuxième technique est un modèle basé sur des règles dans lequel les techniques sont construites sur la base des nomenclatures. Les expériences donnent F-mesure globales de 86, 407%. Et pour lier les NERs psychologiques, nous avons intégré une formulation de programmation linéaire. Au meilleur de nos connaissances, c'est la première approche sur l'extraction d'information psychologique pour inclure dans l'état de l'art des travaux effectués en la langue arabe.

Les mots clés : Programmation Linéaire en Nombre Entier, Reconnaissance des Entités Nommées en Arabe, Extraction des Relations, Extraction des Informations Psychologiques.

Abstract

Named entity recognition is a crucial component of (bio)medical natural language processing, enabling information extraction and knowledge discovery from text. Generally, the achieved studies concerning the (bio)medical information extraction were developed in English and some languages. However, there is no study that was developed in the Arabic language. For this, the Arabic Language needs to perform more researches in this area and hence we introduced a Psychological information extraction approach. This research consists of psychological entities recognition and Relation Extraction from the text. Two techniques were applied for the recognition processes : the first requirement prior to the technique was completely dependent on direct identification with the utilization of gazetteers and the second technique is a rule-based model in which rules are techniques were put constructed on the basis of a gazetteers list. Experiments yield the overall F–measure values of 86,407%. And for joint Psychological NER we integrate a Linear Programming Formulation. On the best of our knowledge, this is the first approach on psychological information Extraction to include the state of the work done for Arabic language.

Keywords : Integer Linear Programming, Arabic Named Entity Recognition, Relation Extraction, Psychological Information Extraction.

التلخيص

يعتبر التعرف على الكيانات المسماة مكونا أساسيا في معالجة اللغة الطبيعية الطبية (الحيوية) ، مما يتيح استخراج المعلومات واكتشاف المعرفة من النص. عموما ، تم تطوير الدراسات التي تم التوصل إليها بشأن استخراج المعلومات الطبية (الحيوية) باللغة الإنجليزية وبعض اللغات الأخرى . ومع ذلك ، لا توجد أي دراسة تم تطويرها باللغة العربية. لهذا ، يجب أن تكون اللغة العربية أكثر فاعلية في هذا المجال ، ومن ثم فإننا أدرجنا نهج إستخراج المعلومات النفسية. هذا البحث يهتم بالتعرف على الكيانات النفسية واستخراج العلاقات من النص. طُبقت اثنين من التقنيات لعملية الاعتراف : فالمتطلب الأول هو أن التقنية تعتمد بشكل كامل على التحديد المباشر لاستخدام قوائم مصطلحات والتقنية الثانية هي نموذج قائم على القواعد حيث تم فيه وضع القواعد على أساس قائمة المعاجم. تنتج عن التجارب القيمة الإجمالية 86,407%. و من أجل وصل المصطلحات النفسية ، قمنا بإدماج البرمجة الخطية. على حد علمنا ، هذا هو النهج الأول الذي يهتم باستخراج المعلومات النفسية. و ذلك من أجل ضمه إلى ما تم إجراؤه للغة العربية.

الكلمات المفتاحية : البرمجة الخطية الصحيحة، التعرف على الكيانات العربية ، استخلاص العلاقة ، استخراج المعلومات النفسية.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Introduction

Dans cette thèse, nous nous intéressons à l'une des sous-tâches de l'Extraction d'Information (El) qui est la reconnaissance des entités nommées. Cette dernière est devenue très utile pour la recherche d'information et les applications de Traitement Automatique de la Langue Naturel (TALN), notamment pour la Traduction Automatique (TA) et l'annotation sémantique. Les ENs (Entités Nommées) sont des mots particuliers qui peuvent désigner les noms propres (noms de personne, noms de lieu et noms d'organisation), les expressions numériques et les expressions temporelles.

Dans l'ère islamique au 7ème siècle toutes les sciences arabes avaient atteint leur visière spécialement la pharmacologie et la médecine arabe, plus spécifiquement dans l'ère Umayyade et abbasside, où les mouvements de la traduction en arabe ont prospéré, suivi d'une période des contributions arabes. L'histoire de la médecine arabe étendue à partir du VIIIe siècle, quand les intellectuels arabes ont commencé à émerger multiples sciences vers l'est. Cette balise de sciences y resta jusqu'au début du XIIIe siècle. L'histoire de la médecine arabe peut être scindée en trois phases principales : phase de la traduction, phase de la contribution arabe originale et phase de déclin et la transmission vers l'Europe [1].

L'arabe est l'une des six langues officielles de l'Organisation des Nations Unies. C'est un langage sémitique avec plus que 300 millions orateurs dans 23 pays. Le traitement du langage arabe est considéré complexe à cause des caractéristiques structurelles et morphologiques telle que l'inflexion, polysémie et les formes irrégulières des mots ; et pour des raisons diverses, aujourd'hui la médecine arabe souffre de l'inconsistance terminologique et plusieurs tentatives pour mettre la langue arabe officielle dans plusieurs instituts médicales reste vaine.

C'est dans cette optique que se situe le travail que nous présentons dont le but est de

détecter et d'extraire les informations pertinentes dans des textes psychologiques en langue arabe, et ceci en développant notre système de reconnaissance d'entités nommées psychologiques qui est une étape d'outillage de l'analyse qui servira à des applications plus spécifiques dans le cadre d'une démarche incrémentale. Notre système est fondé sur une approche symbolique où l'extraction s'effectue en se basant sur un ensemble de Gazetteers et de règles construites manuellement en exploitant l'outil d'extraction des entités nommées disponible sous la plateforme GATE .

Objectifs

La détection des entités nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour l'annotation. Cette tâche représente un sérieux défi, compte tenu des spécificités de l'arabe. Dans cette thèse, nous présentons une étude détaillée des entités nommées en arabe dans le cadre d'une application d'annotation des pages web. Les objectifs principaux de cette thèse sont :

- construction des monocultures (Gazetteers) psychologiques bilingues (arabe- anglais) à partir des ressources psychologiques.
- La création des règles JAPE (Java Annotation Pattern Engine) pour les différents types des ENs psychologiques.
- L'addition de module pour la traduction des entités nommées identifie dans chaque règle.
- L'intégration des Gazetteers d'ENs construits et les règles dans la plate forme GATE.
- L'extraction des ENs psychologiques en arabe à partir le web.
- L'extraire des relations explicites et implicites entre les ENs psychologiques

Contribution

La tache de reconnaissance des entités nommées a fait cette dernière décennie l'objet d'une attention plus soutenue et suscite aujourd'hui un intérêt certain, elle apparait en effet comme fondamentale pour diverses applications de TALN participant de l'analyse de contenu, à l'instar de la recherche et l'extraction d'information, la tache de question-réponse, le résume automatique ou encore le fonctionnement des moteurs de recherche,

et nombreux sont les travaux se consacrant à cette tache, obtenant des résultats plus que probants, et ce pour diverses langues. Aussi, il est désormais possible d'affirmer qu'il s'agit d'un des incontournables du traitement automatique des textes.
Nos contributions portent sur trois aspects distincts. Le premier aspect est l'extraction des informations psychologiques en arabe à partir des pages web.
Dans le domaine psychologique, les informations ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, des méthodes de TALN ont été développées avec succès afin d'extraire des informations pertinentes des textes libres et de les convertir en représentations formelles exploitables par l'homme et par la machine.
Dans cette contribution, nous nous intéressons à l'extraction d'informations à partir de textes psychologiques. La reconnaissance d'entités nommées et l'extraction de relations sont deux tâches fondamentales pour l'extraction de l'information. La première tache, La reconnaissance d'entité psychologique (PsyNER) [2][3] est fondée sur une approche symbolique où l'extraction s'effectue en se basant sur un ensemble de Gazetteers et de règles construites manuellement en exploitant l'outil d'extraction des entités nommées disponibles sous la plateforme GATE. La deuxième tâche consiste à représenter l'interaction entre ces entités. Pour cette raison, nous utilisons la Programmation Linéaire (Linear Programming - LP) pour extraire des relations explicites et implicites. Autant que nous sachions, l'extraction d'informations psychologiques en langue arabe n'a toujours pas été tentée par aucun chercheur.

L'extraction d'information psychologique est importante pour de nombreuses applications, y compris le soutien à la décision clinique, la biologie intégrative et la pharmacovigilance, et a donc fait l'objet d'une recherche active.

Le deuxième aspect est la traduction des ENs identifie de la langue arabe vers la langue anglais. La traduction des EN d'une langue à une autre ouvre de nouvelles perspectives car elle peut être à la base de nouvelles applications notamment dans les domaines de l'accès multilingue aux informations, l'annotation/l'indexation des documents et l'enseignement à distance. Le troisième aspect est de donner une démarche pour les chercheurs désirant entamer le domaine de reconnaissance des entités psychologiques.

Organisation de thèse

Ce document de thèse comporte trois parties. Après, la première partie, l'introduction, les objectifs et la contribution, nous présentons, dans ce qui suit, le reste de la struc-

ture de cette thèse. La deuxième partie, expose les notions de base de notre travail. Elle comporte trois chapitres, dans le premier chapitre nous étudions la particularité et l'importance de langue arabe en Web sémantique, ensuite, nous avons révélé un certain nombre de problèmes qui suggèrent des raisons pour le manque de recherche arabe. Le deuxième chapitre présente les différentes notions liées à la compréhension du domaine de l'extraction des ENs. Nous dressons leurs méthodes d'évaluation à travers les principales campagnes d'évaluation. Tout d'abord, nous commençons par décrire les campagnes d'évaluation de la tâche de REN et les mesures utilisées pour son évaluation.

Le troisième chapitre est un état de l'art sur les travaux d'extraction d'EN. où nous décrivons les principes génériques des différentes techniques utilisées pour l'extraction des entités nommées et des relations, tout en énonçant pour chaque approche les travaux de recherche réalisés pour la langue arabe dans le domaine général et le domaine Biomédicale. Enfin, nous faisons un tour d'horizon sur les plateformes TALN existantes dédiées à cette langue pour extraction EN.

La troisième partie, elle comporte quatre chapitres, le quatrième chapitre décrit notre méthodologie pour la reconnaissance des ENs à partir des page web psychologique. Dans le cinquième chapitre, nous présentons notre méthode pour l'acquisition des relations binaire avec la programmation linéaire en nombre entier. Par la suite, dans le sixième chapitre, nous présentons les différentes évaluations qui ont été faites avec une discussion des résultats obtenus. Le septième chapitre est une conclusion où nous résumons les travaux présentés dans cette thèse ainsi que des perspectives pour la suite de nos travaux.

Première partie

Fondements théoriques

Chapitre 1

La langue Arabe et le Web Sémantique

Sommaire

1.1 Introduction

Ce chapitre expose un ensemble de concepts de base qui sont nécessaires pour la compréhension de cette recherche, nous commençons par l'illustration de quelques caractéristiques de la langue arabe. Ensuite, nous avons révélé un certain nombre de problèmes qui suggèrent des raisons pour le manque de recherche arabe et ensuite nous étudions l'importance de langue arabe en Web sémantique.

1.2 Caractéristiques de la langue arabe

Appliquer des tâches de TALN en général et de la tâche de REN est en particulier très provocante quand il s'agit d'arabe en raison de ses particularités et sa nature unique. Les caractéristiques principales de l'arabe qui lancent des défis non triviaux pour la tâche de REN sont comme suit :

- Aucune capitalisation – la capitalisation n'est pas une caractéristique de script arabe, à la différence des langues latines où un EN commence habituellement par une majuscule. Par conséquent, l'utilisation de la caractéristique de capitalisation n'est pas une option dans REN arabe. Cependant, la traduction en anglais des mots arabes peut être exploitée à cet égard [4].

- La nature agglutinative – l'arabe a une nature agglutinative élevée en laquelle un mot peut se composer des préfixes, du lemme et des suffixes dans différentes combinaisons, qui a conduit à une morphologie très compliquée [5].

- Voyelles courtes facultatives – dans la théorie, les voyelles courtes, ou les signes diacritiques, sont nécessaires pour la prononciation et la désambigüisation. L'absence des voyelles dans les textes arabes engendre une certaine ambiguïté en ce qui concerne le sens du mot d'une part, et augmente la difficulté à identifier sa fonction dans la phrase d'autre part. Cependant, les textes arabes modernes n'incluent pas des signes diacritiques, donc, un mot arabe peut se rapporter deux ou plus différents significations selon le contexte qu'il apparaît dedans.

- Variantes d'orthographe – en script arabe, le mot peut être orthographié différemment et se réfère toujours au même mot avec le même sens. Par exemple, le mot « جرام », jrAm1, « gramme », peut également être écrit comme « غرام», gramme, avec la même signification.

- Le manque de ressources linguistiques – il y a généralement une limitation sur le nombre de ressources linguistiques arabes qui sont gratuits à des fins de recherche. En particulier, plusieurs des corpus disponibles ni l'un ni l'autre sont annotés avec ENs ni incluez le nombre suffisant de ENs arabe, qui les rend inappropriés pour la tâche de REN en arabe. Les Gazetteers arabes sont rares et limité dans la taille. Par conséquent, les chercheurs ont tendance à construire leurs propres ressources linguistiques en arabe afin de former et examiner leurs systèmes arabes proposés de REN.

Dans notre recherche nous avons donné beaucoup de temps pour élaborer nos Gazetteers afin de pouvoir évoluer notre système proposé.

1.3 Défis et opportunités

L'enquête préliminaire dans le domaine du Web Sémantiques (WS) a révélé un certain nombre de problèmes qui suggèrent des raisons pour le manque de recherche arabe, à savoir le manque de soutien technologique et de ressources adéquates.

1) **Manque de support arabe dans les outils Web Sémantiques existants**

Un problème spécifique avec les outils de traitement de texte en langue arabe concerne le codage. L'encodage différent de l'écriture arabe existe sur le Web ; codages dominants comprennent UTF-8, Windows-1256 et ISO-8859-6[6]. De plus, la plupart des outils WS que nous avons rencontrés ont été construits en utilisant Java, qui supporte l'internationalisation. Par conséquent, il existe un fort besoin de consolider les différents codages arabes ou simplement d'adhérer à un schéma de codage lors de la représentation d'un texte arabe (c'est-à-dire, Unicode). Le WS typique les outils des développeurs utilisent Unicode partout[7] ; par conséquent, cela pourrait résoudre une partie du problème de support.

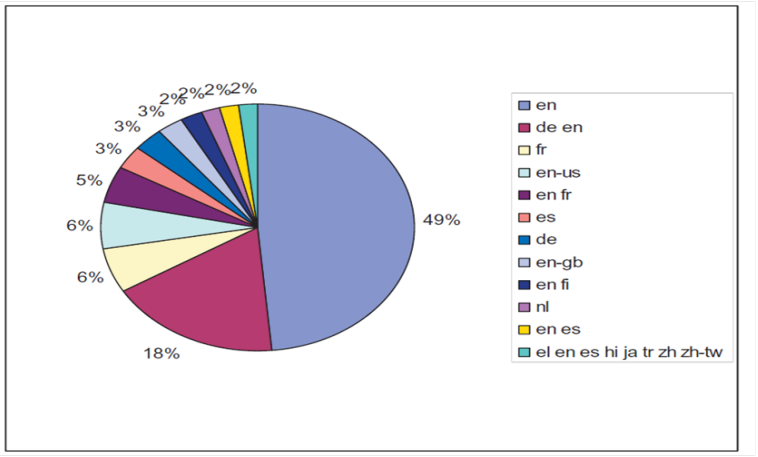
2) **Manque d'applications Web sémantique en arabe**

Une autre preuve du manque d'arabe dans le monde du Web sémantique est une statistique récente fournie par la bibliothèque d'ontologies OntoSelect[8], qui montre que 49% des ontologies de la bibliothèque sont créées en anglais. La distribution des langues utilisées dans la création d'ontologies, illustrée dans la figure 1.1, met en évidence le problème d'applications WS arabes limitées. Ce problème peut être attribué au manque d'outils et d'environnements de développement de logiciels qui traitent le script arabe à toutes les étapes du processus d'annotation sémantique. Peu de moyens sont disponibles pour générer des annotations WS de manière routinière et sans effort au moment de l'utilisation ou de la création de contenu.

3) **Support limité pour la recherche arabe dans le domaine des technologies du Web sémantique**

La plupart des recherches WS est le résultat de l'investissement des organismes subventionnaires et des centres de recherche universitaires. Outils WS, tels que Protégé[9], OntoSelect[8], GATE[10] et Jena[11] pour n'en nommer que quelques-uns sont tous les produits d'un investissement réussi dans le domaine WS. Dans le cas particulier de l'arabe, le problème limité de la recherche peut être attribué au manque de ressources adéquates en termes de compétences, de financement et d'intérêt dans ce domaine

FIGURE 1.1 : Diagramme à secteurs montrant la distribution des langages utilisés dans la création d'ontologies stockées dans la bibliothèque OntoSelect.



émergent de la recherche sur le Web. La répartition du financement de la recherche, la fourniture de ressources et l'intérêt d'une communauté de pratique engagée sont essentiels si nous voulons surmonter ce problème.

Une expérience avec des technologies Web similaires suggère qu'une communauté ciblée d'adeptes précoces enthousiastes, aussi petite soit-elle, est une condition préalable essentielle au succès. Principaux groupes d'intérêt sur l'arabisation et réseaux de développeurs, tels que Arabeyes et Unicode consortium, peuvent être en mesure de contribuer à cet effort. En ce qui concerne les ressources financières, le coût du développement et de la maintenance des applications WS est une préoccupation majeure, en particulier pour les communautés ayant une faible base d'utilisateurs, comme les utilisateurs arabes. Dans certains domaines bien structurés tels que les applications scientifiques, commerciales et gouvernementales, les avantages potentiels le gain de productivité et le profit l'emporteront sur le coût de développement et de maintenance d'une ontologie / application. Le coût, en termes de temps et d'effort requis, diminuera à mesure que la base d'utilisateurs augmentera.

Il y a beaucoup de défis et d'opportunités potentiels que les utilisateurs de la langue arabe peuvent rencontrer avec le commencement des solutions WS.
Certains de ces défis sont les suivants :

1) La nécessité de développer des vocabulaires contrôlés et des ontologies pour définir explicitement la sémantique traitable par la machine. Ces vocabulaires contrôlés et les ontologies aideraient les gens et les machines à communiquer de façon

concise et supporte l'échange sémantique et syntaxique entre eux. De tels efforts prometteurs sont en cours de développement ; ils comprennent la construction d'un Arabic WordNet [12].

2) Le traitement informatique du texte arabe diffère de son homologue anglais. La langue arabe englobe des aspects morphologiques, grammaticaux et sémantiques plus complexes que les algorithmes existants de traitement du langage naturel (TAL) utilisés pour la langue anglaise, ils ne peuvent pas être directement les réutilisés pour la langue arabe. Ce problème a aussi des racines dans le domaine WS et Extraction d'information, l'un des principaux processus du WS, repose largement sur l'extraction de concepts à partir de documents Web ; ce processus nécessite l'analyse du contenu du document - soit morphologiquement, grammaticalement ou sémantiquement - pour pouvoir relier les instances extraites à un concept d'ontologie prédéfini. Ainsi, des outils TAL plus adaptés tels que GATE [10] sont nécessaires pour traiter Langue arabe en conséquence.

En ce qui concerne les opportunités, le WS fournit de riches opportunités aux utilisateurs de la langue arabe pour traiter les données à plusieurs niveaux. En termes de données, il existe sur le Web actuel de grands répertoires de contenus arabes dans les affaires, la science, les documents gouvernementaux et les courriels. Le WS fournit aux utilisateurs un cadre commun pour intégrer et dériver une nouvelle signification, valeur et intelligence de ces référentiels arabes existants.

1.4 Langue arabe et Web sémantique

De nombreux efforts ont été consacrés au traitement automatique de la langue arabe, mais les différents problèmes posés par cette langue et leurs spécificités graphiques et morphologiques ont ralenti le processus du développement des outils dans ce domaine.

1.4.1 Importance de la langue arabe

L'arabe est l'une des six langues officielles de l'Organisation des Nations Unies. C'est un langage sémitique avec plus que 300 millions orateurs dans 23 pays, et est la langue religieuse de tous les musulmans de différentes ethnies à travers le monde. Elle appartient au groupe des langues sémitiques avec 28 lettres de l'alphabet. Son orientation d'écriture est de droite à gauche. Le script arabe constitue 8,9% des langues du monde, comme le montre le tableau 1.1. Quels que soient les chiffres, il est essentiel de

comprendre que ce pourcentage élevé de la langue arabe soulève la question suivante : pourquoi trouvons-nous des applications pour le script Farsi et nous ne trouvons pas la même application pour le script arabe (comme le farsi et l'arabe viennent de la même famille de script) ?

TABLE 1.1 : Pourcentage de script de langues du monde [13].

Script	**Latin**	**Cyrillic**	**Arabic**	**Hanzi**	**Indic**	**Others**
Million users	2,238	451	462	1,085	807	129
% of total	43,3%	8,7%	8,9%	21,0%	15,6%	2,5%
Key languages	Romance (European) Slavic (some) Vietnamese Malay	Rusian Slavic (some) Kazakh Uzbek	Arabic Urdu Persian Pashtu	Chinese Japanse Korean	Hindi Tamil Bengali Punjabi Sanskrit Thai	Greek Hebrew Georgian Assyrian Amenian

Donc, l'importance de la langue arabe est résumée dans ce qui suit :

1. La langue arabe est la langue religieuse pour les musulmans du monde entier.
2. L'arabe est la langue du Coran.
3. C'est la langue maternelle de vingt trois pays.
4. Il y a 300 millions de personnes qui parlent l'arabe comme langue principale.
5. C'est l'une des six langues officielles des Nations Unies.
6. L'arabe est l'un des langages sémantiques les plus riches qui ont des mots spécifiques pour décrire une chose spécifique.
7. Le scripte arabe représente 8,9% des langues du monde, comme le montre le tableau 1.1 qui explique que la classification des langues du monde dépend de leur famille [14].

1.4.2 Langue arabe et la recherche web sémantique

Les technologies du Web sémantique ont prouvé leur succès dans plusieurs domaines, tels que la médecine, e-Commerce, e-Learning et biologie. Pour étendre ce succès à la langue arabe, un ensemble d'outils et d'applications WS doit être créé pour répondre aux exigences de la langue arabe.
Le Web sémantique vise à améliorer le Web existant avec une couche de métadonnées interprétables par machine (c'est-à-dire des données sur les données) afin qu'un programme informatique puisse comprendre le contenu d'une page Web et tirer des conclusions sur la page Web. Le WS, tel que défini par son créateur Tim Berners-Lee [15], implique "... Le Web sémantique est une extension du Web actuel, dans lequel l'information reçoit une signification bien définie, améliorant les possibilités de travail collaboratif entre les ordinateurs et les personnes."

Il existe diverses études menées sur la langue arabe dans Web sémantique. Zaidi, Laskri et Bechkoum [16] ont proposé d'améliorer la récupération de l'information arabe sur le Web dans le domaine juridique par un moteur de recherche arabe supportant la traduction des requêtes arabes en requêtes anglaises ou françaises. L'objectif était de rendre les documents écrits en arabe, en français ou en anglais. Vossen, Pease et Fellbaum[17] ont travaillé sur Arabic WordNet (AWN) en basant sur les méthodes développées pour EuroWordNet (EWN).

Ce ne sont que quelques études menées directement ou indirectement dans le web sémantique en langue arabe. Sur la base des informations recueillies, on peut conclure que le travail en langue arabe pour le Web sémantique est encore à l'enfance.

1.4.3 Pénétration de l'internet en langue arabe

On s'aperçoit que le l'arabe qui était en septième position en 2010 est passé en quatrième position. L'anglais, chinois, l'espagnol et l'arabe sont maintenant devant. Cela voudrait donc dire qu'il y a progression a explosé pour la langue arabe ?

Le tableau qui suit (Les dix premières langues de l'Internet) montre les dix premières langues utilisées sur l'Internet. Par exemple, on compte pour le japonais 99,1 millions de personnes parlant cette langue, ce qui représente 4,7% de tous les utilisateurs d'Internet dans le monde. Mais la population totale estimée utilisant le japonais est de 126,4 millions, alors que 78,4 % des locuteurs du japonais ont accès à l'Internet. En une dé-

FIGURE 1.2 : Les 10 langues les plus utilisées sur l'Internet - 30 Juin 2017.

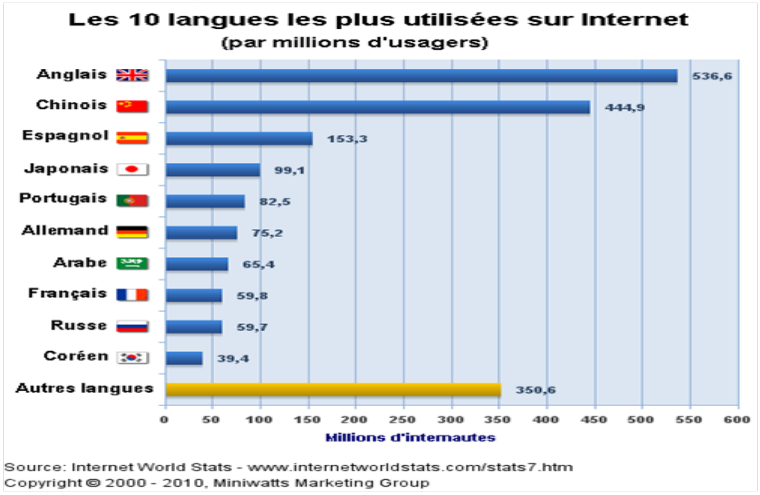
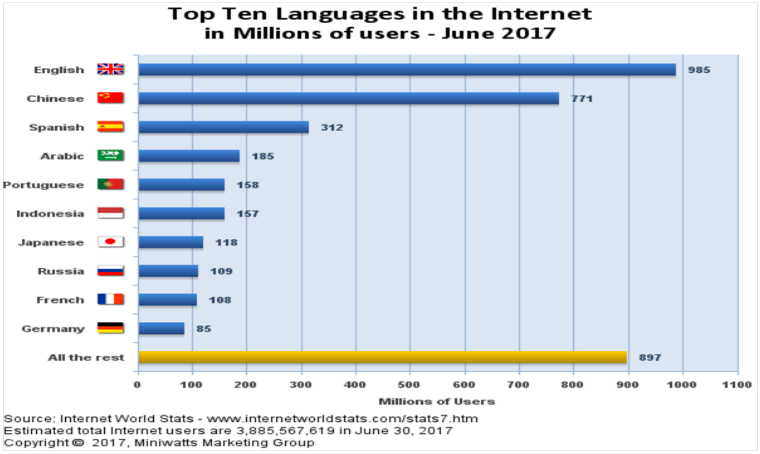


FIGURE 1.3 : Les 10 langues les plus utilisées sur l'Internet - 30 Juin 2017.



cennie (2000-2017), le nombre des utilisateurs du japonais sur l'Internet a augmenté de 110,7 %. Par comparaison, la progression de l'anglais (301,4 %) et du français (398,2 %) est trois fois plus considérable. Cette progression a explosé pour le chinois (1478,7 %), l'arabe 2501,2 % et le russe (1825,8 %), sans oublier les « autres langues » avec une augmentation de 588,5 %, soit davantage que les dix premières langues réunies.

TABLE 1.2 : Les dix premières langues de l'Internet 30 Juin 2017.

Les dix premières langues de l'Internet - 30 Juin 2017(Nombre d'internautes par langue)					
Les 10 langues les plus utilisées sur Internet	**Nombre d'internautes par langue**	**Taux de pénétration par langue**	**Progression de l'Internet (2000 2017)**	**Percentage des internautes**	**Population mondiale de la langue utilisée (estimation(2017)**
Anglais	984,703,501	68.6 %	599.6%	25.3 %	1,434,937,438
Chinois	770,797,306	54.1 %	2,286.1 %	19.8 %	1,425,430,865
Espagnol	312,069,111	61.1 %	1,616.4%	8.0 %	510,380,423
Arabe	184,631,496	43.8 %	7,247.3 %	4.8 %	421,345,425
Portugais	158,399,082	56.2 %	1,990.8 %	4.1 %	281,603,515
Indonesian / Malaysian	157,580,091	53.4 %	2,650.1 %	4.1 %	295,108,771
Japonais	118,453,595	94.0 %	151.6 %	3.0 %	126,045,211
Russe	109,552,842	76.4 %	3,434.0 %	2.8 %	143,375,006
Français	108,014,564	26.6 %	800.2 %	2.8 %	405,644,599
Allemand	84,700,419	89.2 %	207.8 %	2.2 %	94,943,848
TOTAL des 10 premières langues	2,988,902,008	58.2 %	907.2 %	76.9%	5,138,815,101
Autres langues	896,665,611	37.7 %	1,296.1 %	23.1 %	2,380,213,869
TOTAL DES LANGUES DU MONDE	3,885,567,619	51.7 %	976.4 %	100.0 %	7,519,028,970

1.5 Annotation sémantique

En général, le terme annotation réfère à une note, une critique, une remarque ou encore un commentaire qui accompagne un texte afin d'y apporter plus de précision et d'explication. Autrement dit, l'annotation est l'action d'apposer une note sur une partie de document ou de texte de document.

Dans le cadre du web sémantique, une annotation descriptive est le plus souvent appelée annotation sémantique. L'objectif des annotations sémantiques est d'exprimer la sémantique du contenu d'une ressource afin d'en améliorer la compréhension, la recherche par des agents logiciels et la réutilisation par les utilisateurs. On peut définir l'annotation sémantique comme une représentation formelle des étiquettes sémantiques attachées au contenu textuel d'une ressource, exprimées à l'aide de concepts, relations et instances décrits par un des standards formels du web sémantique tels que les ontologies (OWL), RDF, XML.

S'agissant de la langue Arabe, des recherches intéressantes dans le développement de certains systèmes ont été effectuées, particulièrement dans la morphologie[18][19], dans les systèmes de la Recherche d'Information (RI) [20] et de Q/R [21], etc. Cependant, ces recherches n'ont pas encore atteint le même niveau d'avancement que celles concernant les langues latines.

1.6 Extraction d'entités

De nombreux chercheurs ont attaqué le problème de l'identification des ENs dans une variété de langues, mais quelques efforts de recherche limités ont porté sur la reconnaissance des entités nommées pour le script arabe. Cela est dû au manque de ressources pour les entités nommées en arabe et au nombre limité de progrès réalisés dans le traitement de la langue arabe en général.

Les premiers travaux sur la reconnaissance des EN pour l'arabe datent de 1998 et reposent sur des méthodes à base de règles [22], voir également les travaux de [23] ou de Zaghouani et al. [24]. Samy et al.[25] utilisent un corpus parallèle pour extraire des EN en arabe. Ils utilisent un étiqueteur à base de règles enrichies avec un lexique monolingue espagnol pour extraire les EN en espagnol qui sont, par la suite, translittérées vers l'arabe.

1.7 Conclusion

Dans ce chapitre, nous avons abordé, d'une part, l'importance de la langue arabe dans le web et de l'autre part nous avons révélé un certain nombre de problèmes qui suggèrent des raisons pour le manque de recherche arabe et en fin nous avons pris une vue

générale sur l'annotation sémantique, le Web Sémantique et les entités nommées.

Nous présentons dans le prochain chapitre un état de l'art sur les ENs et les conférences MUC. En suite les applications TAL et reconnaissance de l'entité nommée biomédicale.

Chapitre 2

Reconnaissance de l'Entité Nommée - Etat de l'art

Sommaire

2.1 Introduction

Dans ce chapitre, les concepts fondamentaux qui constituent la base de la compréhension de notre recherche sont présentés. D'abord, nous parcourons brièvement l'historique des principales campagnes d'évaluation qui se sont intéressées à la problématique de la reconnaissance d'entité nommées (REN), afin de mieux comprendre les raisons de son succès. Après, nous décrivons la relation entre les applications de Traitement Automatique du Langage Naturel (TALN) et la reconnaissance d'entités nommées. Enfin, nous présentons quelques difficultés de la reconnaissance des entités nommées arabe dans le domaine biomédicale

2.2 L'extraction d'information

L'Extraction d'Information (EI) est utilisée pour extraire automatiquement des informations structurées à partir de documents non structurés ou semi-structurés. L'extraction d'information est un sous-domaine du TALN [26].

Le traitement du langage naturel est le domaine de l'Intelligence Artificielle (IA) qui concerne les interactions entre les ordinateurs et les langages humains naturels. Au cours des dernières décennies, de nombreuses applications ont été développées pour traiter le domaine de la TALN telles que la Recherche d'Information, les traductions automatiques, les systèmes questions/réponses, les résumés automatiques, le Web sémantique, et la biomédicale. Le but principal de TALN ou de l'EI en général est la REN [27][28].

2.3 Origine de la reconnaissance de l'entité nommée

Le terme EN (Entité Nommée) est apparu au cours de la sixième conférence MUC (Conférences sur la compréhension de messages, en anglais, Message Understanding Conference) en 1996. Une EN désigne les noms de tous les personnes, organisations et lieux dans un texte.

La REN est également appelé extraction d'entité ou identification d'entité. C'est une sous-tâche très importante de l'extraction de l'information qui vise à trouver et classifier le nom dans un texte.

Plusieurs chercheurs ont proposé des définitions différentes pour ce terme. Notez que la définition de EN est relié au domaine intéressé.

Pour le domaine général, Le Meur et al [29] ont donné la définition suivante : «les ENs sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)». Goldman et Scherrer[30] ont définit l'EN comme un mot ou un groupe de mots désignant une personne, une organisation ou entreprise, un lieu, une date ou encore une expression numérique.

Pour le domaine biomédical, Zhong Huang and Xiaohua Hu [31] ont défini l'EN comme un terme de mot unique ou une expression de plusieurs mots qui désigne un objet biomédical, par exemple une protéine, un gène, une maladie ou un médicament.

Nous avons cité ces définitions comme des exemples, mais non pas en exclusivité.

2.4 Les conférences MUC

De 1987 à 1998, sept conférences MUC ont eu lieu pour traiter du problème de l'El[32][33]. Les données des participants étaient sous forme de messages et elles étaient évaluées sur des sujets particuliers.
Les conférences MUC ont été organisées et financées par DARPA (en anglais, Defense Advanced Research Projects Agency) et NOSC (en anglais, Naval Ocean System Center) dans le but d'encourager la recherche et le développement en El [32]. Nous pouvons distinguer trois phases différentes dans le déroulement de la conférence MUC selon la nature de la tâche visée.

2.4.1 Phase exploratoire

Cette phase se caractérise par l'utilisation exclusive de données provenant des messages de la marine américaine (US Navy), et par l'absence de tâches et de procédures d'évaluation bien définies. Elle comporte deux éditions MUC-1 et MUC-2.
MUC-1 (1987) : le but de cette édition a été de faire un état de l'art des systèmes de compréhension de textes et de population de base de connaissances. Lors de cette première édition les participants choisissaient le format des sorties de leurs systèmes, aucune évaluation formelle n'ayant été mise en place. Cette conférence a le mérite d'avoir réuni

des chercheurs et des développeurs pour discuter de la nature des informations utiles à extraire à partir des messages de la marine américaine et de développer les premières approches pour aborder une telle problématique.

MUC-2 (1989) : dans cette d édition, les participants ont exigé d'avoir une définition de la tâche d'extraction précise avec un format bien défini afin de pouvoir, par la suite, évaluer la qualité des systèmes abordant une même tâche. Cette tâche consistait à repérer des événements dans des textes et à remplir, pour chaque événement, un formulaire. Chaque formulaire contenait des champs relatifs à un incident (l'opération) qui étaient les acteurs, la date, l'heure, le lieu et les résultats (d'autres informations à extraire existaient mais n'étaient pas bien définies). La précision (P) et le rappel (R) étaient les métriques adoptées pour mesurer les performances des systèmes, dans MUC-2 P et R étaient définis comme suit :

$$\mathcal{R} = \frac{Nombre\ de\ reponses\ corrects}{Nombre\ total\ de\ reponses\ recherchees}$$

$$\mathcal{P} = \frac{Nombre\ de\ reponses\ corrects}{Nombre\ de\ reponses\ corrects + Nombre\ de\ reponses\ incorrects}$$

L'inconvénient majeur de cette conférence était l'absence d'une procédure d'évaluation unifiée et automatisée. Ceci a conduit à une grande variation dans la manière dont chaque participant évaluait les sorties de son système. Ceci a mis en évidence la nécessité de mettre en place un paradigme d'évaluation unifié et automatisé.

2.4.2 Remplir des formulaires

Cette phase est distinguée de la précédente par l'utilisation de données plus diversifiés et par l'adoption d'une tâche axée sur la compréhension du texte pour remplir des formulaires. Des procédures d'évaluation automatisées ont été mises en place et de nouvelles métriques d'évaluation fondées sur le taux d'erreur ont été introduites.

MUC-3 (1991) : la troisième édition inclus une diversité d'articles journalistiques portant sur les activités terroristes en Amérique latine. Ceci avait permis d'avoir plus de données pour les tests et pour l'entraînement des systèmes. La tâche consistait à détecter les événements se référant à des actes terroristes en Amérique latine. Cette tâche est devenue un peu plus complexe, avec plus d'informations à extraire et des données plus

riches et plus diversifiées à analyser. Il est important de remarquer aussi que les informations à extraire nécessitaient plus d'analyse et de compréhension du texte ce qui a augmenté la complexité de la tâche[34].

Liste non exhaustive des informations à extraire :

- ⋆ Déterminer si c'est un acte criminel ou terroriste ;
- ⋆ Déterminer si les informations (dans le texte) sont précises ou vagues ;
- ⋆ Déterminer s'il s'agit d'une menace ou d'un véritable acte ;
- ⋆ Déterminer la cause de l'acte ;
- ⋆ Déterminer l'origine des acteurs ;
- ⋆ Déterminer le lieu, la cible, la date, les instruments utilisés, etc.

MUC-3 s'est aussi distingué par la mise en place d'une vraie procédure d'évaluation automatisée qui consiste à comparer les formulaires remplis automatiquement par les systèmes aux formulaires de référence remplis manuellement par les participants, puis, ensuite, à comptabiliser les bonnes et les mauvaises réponses pour calculer le rappel et la précision pour chaque système. Mais, durant cette campagne, ils ont remarqué que l'utilisation de deux métriques d'évaluation (P et R) a crée une certaine confusion tel que, dans certains cas, il était difficile de décider quel système avait de meilleures performances par rapport aux autres.

MUC-4 (1992) : la même tâche que celle de MUC-3 a été proposée dans MUC-4, avec améliorations apportées aux définitions de certains slots, qui sont passé de dix huit (dans MUC-3) à vingt-quatre (dans MUC-4)[35]. MUC-4 s'est distingué des conférences précédentes par son utilisation d'une nouvelle métrique d'évaluation la « F-mesure » qui était définie comme la moyenne harmonique entre P et R [36]. La F-mesure a donné la possibilité de classer les systèmes de REN selon une mesure unique.
MUC-5 (1993) : durant cette cinquième édition, il y avait deux types d'événements à traiter : les coentreprises internationales et la fabrication de circuits électroniques, et ce, dans deux langues, l'anglais et le japonais. Contrairement aux autres conférences où il n'y avait qu'un seul type de formulaire, dans MUC- 5, il y en avait onze avec un total de quarante-sept types de champs distincts. Ainsi, la tâche avait encore gagné en complexité par rapport aux éditions précédentes. Cette conférence a permis d'enregistrer l'utilisation d'une nouvelle métrique qui permettait d'évaluer les performances des

systèmes en termes de taux d'erreur. Cette métrique était appelée « the error per response » (ERR) et elle était la mesure officielle durant MUC-5[36].

$$\mathcal{ERR} = \frac{Nombre\ de\ fausses\ reponses}{Nombre\ total\ de\ reponses + Nombre\ d'omissions}$$

2.4.3 Entités nommées

Durant les cinq premières campagnes (MUC-1 à MUC-5) qui se sont déroulées entre 1987 et 1993, la tâche mise en place a consisté à extraire des informations se trouvant dans des documents textuels afin de remplir des formulaires. Même si les performances obtenues par les systèmes étaient encourageantes, 57 % pour le rappel et 64 % pour la précision sur l'ensemble des données dans MUC-5, la tâche de remplissage des formulaires n'a pas cessé de gagner en complexité d'une campagne à l'autre. Elle nécessitait des niveaux d'analyse et de compréhension de textes de plus en plus avancés. En effet, pour remplir un formulaire, le système doit effectuer des traitements à plusieurs niveaux : détecter les entités, catégoriser les entités détectées, extraire les événements, repérer les relations pouvant exister entre les entités ou entre les entités et les événements et déterminer la nature des relations. La complexité de cette tâche rend les diagnostics et la compréhension des provenances des erreurs très difficiles, puisque tous les modules sont étroitement liés les uns aux autres.

MUC-6 (1995) : les objectifs dans MUC-6 (1995) étaient d'améliorer les performances des systèmes d'extraction d'informations. L'hypothèse était que ceci n'était possible que si l'on améliorait l'analyse sémantique effectuée par les systèmes. Ainsi, une nouvelle tâche de coréférence fut introduite pour encourager l'amélioration du traitement sémantique. Elle a consisté à marquer des relations comme l'anaphore, la métonymie, et d'autres relations.

Mais, comme il s'agissait de l'avant-dernière édition de la série, il y avait dans MUC-6 une volonté de prouver que la technologie d'extraction d'informations existante était exploitable rapidement avec des performances élevées et qu'elle pouvait être indépendante du domaine. Pour atteindre ce but, l'idée était d'identifier, parmi la technologie développée, la brique (le composant) qui satisfaisait au mieux ces critères. C'est dans cette logique que la tâche (Named Entity) de reconnaissance d'entités nommées a été introduite pour la première fois durant MUC-6. La tâche de la REN a consisté à utiliser des marqueurs SGML pour identifier des noms propres dans les textes (noms de

personnes, noms d'organisations ou noms de lieux), des expressions temporelles et des expressions numériques (monétaires ou pourcentages) [36].
Le tableau 2.1 montre ces trois classes avec leurs sous-classes

TABLE 2.1 : Classes d'EN d'après la campagne MUC.

Classe	Sous-classe
ENAMEX	Les noms propres (noms de personne, noms de lieu et noms d'organisation)
NUMEX	Les expressions numériques (limité aux expressions monétaires et de pourcentages).
TIMEX	Expressions temporelles (date, temps)

La tâche de REN a été un vrai succès. Les résultats des évaluations ont été très satisfaisants. Presque la moitié des participants ont obtenu des valeurs de rappel et de précision supérieures à 90%. Le meilleur système avait 96% de rappel, 97% de précision. La métrique officielle adoptée pour classer les systèmes était ERR et le meilleur système affichait 5% d'erreurs[36].

MUC 7 (1997) : après son succès dans la version précédente, la tâche de REN a été également adoptée dans cette septième et dernière édition de MUC. La définition de la tâche était la même que celle introduite dans MUC-6, avec cependant de légères différences. En effet, dans MUC-7 il était question d'annoter certaines expressions temporelles relatives[38], des règles d'annotations distinctes de celles de MUC-6 ayant été mises en place pour traiter les cas des conjonctions de coordination. Cette édition s'est distinguée des précédentes par le fait que les données d'apprentissage et de test ne portaient pas sur le même domaine (apprentissage : crash des avions, test : lancement de fusée). Ceci avait pour but d'encourager le développement de systèmes flexibles et génériques.

2.4.4 Les autres conférences CoNLL et ACE

A) **La Conférence sur l'apprentissage des langues naturelles (The Conference on Natural Language Learning (CoNLL)**

À la suite de CoNLL2002 et CoNLL2003, quatre catégories de NE sont définies, y compris personne (PERS), localisation (LOC), organisation (ORG), et autres (MISC). Divers fait référence à d'autres éléments de réseau n'appartenant pas à une classe de personne, d'emplacement ou d'organisation. CoNLL suit le format BIO (Begin, Inside, Outside)

afin de marquer les entités nommées dans un ensemble de données. Les annotations CoNLL sont les suivantes :

- ⋆ B-PERS indique le début d'une entité nommée Personne ;
- ⋆ I-PERS désigne un mot à l'intérieur d'une entité nommée Personne ;
- ⋆ B-LOC indique le début d'une entité nommée Location ;
- ⋆ I-LOC désigne un mot à l'intérieur d'une entité nommée Location ;
- ⋆ B-ORG indique le début d'une entité nommée Organisation ;
- ⋆ I-ORG indique un mot à l'intérieur d'une entité nommée Organisation ;
- ⋆ B-MISC indique le début d'une entité nommée Divers ;
- ⋆ I-MISC désigne un mot à l'intérieur d'une entité nommée Divers ;
- ⋆ O indique que le mot n'appartient à aucune des classes précédentes ;

Par exemple, la phrase "L'auteur est John Smith" est annotée comme ci-dessous :

- O Le
- O auteur
- O est
- B-PERS John
- I-PERS Smith

B) **Le programme d'extraction automatique de contenu (The Automatic Content Extraction program (ACE))**

Trois catégories d'entités nommées ont été définies par ACE2003 : personne, installation, organisation et GPE (entités géographiques et politiques). Plus tard dans ACE 2004 et 2005, deux autres catégories ont été ajoutées à cet ensemble d'étiquettes : les véhicules et les armes. Les expressions temporelles, qui suivent les spécifications de TIMEX2, et les expressions numériques, y compris l'argent, le numéro de téléphone et le pourcentage, ont été couvertes par le Corpus de formation multilingue ACE 2005. Un ensemble de données ACE est fourni avec plusieurs fichiers de types différents dans le langage

de balisage généralisé standard ; chaque fichier de données a un fichier XML correspondant qui représente les informations d'entité (c'est-à-dire des annotations d'entités nommées dans un fichier de données). La figure 2.1 illustre un exemple d'information d'entité dans le corpus de formation multilingue ACE 2005 à partir de l'ensemble de données arabe.

FIGURE 2.1 : Exemple d'information sur l'entité ACE 2005.

```
▼<entity ID="ALH20001028.1300.0072-E14" TYPE="GPE" SUBTYPE="Nation" CLASS="SPC">
  ▼<entity_mention ID="ALH20001028.1300.0072-E14-15" TYPE="NAM" LDCTYPE="NAM" ROLE="LOC">
    ▼<extent>
       <charseq START="606" END="611">المغرب</charseq>
     </extent>
    ▼<head>
       <charseq START="606" END="611">المغرب</charseq>
     </head>
   </entity_mention>
  ▼<entity_mention ID="ALH20001028.1300.0072-E14-21" TYPE="NAM" LDCTYPE="NAM" ROLE="LOC">
    ▼<extent>
       <charseq START="166" END="171">المغرب</charseq>
     </extent>
    ▼<head>
       <charseq START="166" END="171">المغرب</charseq>
     </head>
   </entity_mention>
▼<value ID="ALH20001028.1300.0072-V2" TYPE="Numeric" SUBTYPE="Percent">
  ▼<value_mention ID="ALH20001028.1300.0072-V2-1">
    ▼<extent>
       <charseq START="342" END="357">نحو ٨٤٣ في المئة</charseq>
     </extent>
   </value_mention>
 </value>
▼<value ID="ALH20001028.1300.0072-V3" TYPE="Numeric" SUBTYPE="Percent">
  ▼<value_mention ID="ALH20001028.1300.0072-V3-1">
    ▼<extent>
       <charseq START="371" END="383">٢،١١ في المئة</charseq>
     </extent>
   </value_mention>
 </value>
```

Dans notre recherche, nous suivons un ensemble d'étiquettes comprenant les troubles mentaux désigné par DSM-IV (Diagnostic and Statistical Manual of the American Psychiatric Association) [39], substances psycho-actives, symptômes, maladies et les médicaments.

2.5 La reconnaissance d'entité nommée et les applications TAL

La reconnaissance d'entité nommée est considérée comme l'une des tâches cruciales d'extraction d'information dans laquelle de nombreuses applications de TALN s'appuient comme une étape importante du prétraitement. Le REN est la tâche d'extraire les entités nommées (c'est-à-dire les noms propres) des textes structurés et non structurés, puis les classées en classes prédéfinies (par exemple, personne, lieu et organisation)[43],[44]. La performance des systèmes REN utilisés par différents systèmes TAL a un impact significatif sur la performance globale de ces systèmes TAL ce qui rend la qualité des systèmes NER hautement nécessaire. Le rôle de REN dans les applications TAL diffère d'une application à l'autre. Exemple des applications TALN qui trouvent les fonctionnalités de REN utiles pour leurs fins sont : la Recherche d'Information (RI), l'Annotation Sémantique (AS), la Traduction Automatique (TA), les systèmes de Questions Réponde (QR) et le Clustiring de Textes.

2.5.1 Recherche d'information

La RI est la tâche d'identifier et de récupérer des documents pertinents à partir d'une base de données de documents en fonction d'une requête d'entrée. Les RI peuvent bénéficier du REN en deux phases : premièrement, reconnaître les NEs dans la requête ; deuxièmement, reconnaître les NEs dans les documents de la base de données et extraire les documents pertinents en tenant compte leurs classification des NEs et de la manière dont ils sont liés à la requête. Par exemple, si la requête contient le mot مايكروسوفت maAykruwsuwft "Microsoft" qui représente une entité nommée Organisation, les documents relatifs à Microsoft Corporation seront considérés comme pertinents et seront récupérés.

2.5.2 Annotation sémantique

L'AS de texte consiste à apposer sur le texte des informations ou métadonnées dont la sémantique est portée par un modèle (par exemple, langage d'indexation, thesaurus, ontologie). L'AS peut bénéficier du REN pour associer au texte une représentation sémantique, premièrement, reconnaître les NEs dans le texte ; deuxièmement, reconnaître leurs classifications et de la manière dont ils sont liés entre eux.

2.5.3 Traduction automatique

La TA est la tâche de traduire un texte d'une langue source (originale) vers une langue cible, en utilisant des programmes informatiques sans l'intervention humaine. Les NEs ont besoin d'approches spécifiques pour être correctement traduites, il est donc important d'avoir un système REN faisant partie du système TA afin d'améliorer les performances du système global [44]. Dans le cas de la langue arabe, les noms d'organes peuvent être trouvés comme des mots réguliers dans la langue et aucune caractéristique orthographique ne peut distinguer les deux formes, par exemple, le mot سِنّ sin peut être lu comme un organe qui signifie dent, et il peut aussi être un âge.

2.5.4 Le système de Question/ Réponse

Le système de Question/Réponse (QR) peut être considéré comme l'une des applications de Recherche d'Information, mais avec des résultats plus sophistiqués. Les systèmes de QR prennent les requêtes de l'utilisateur comme intrants et donnent en retour des réponses concises et précises. La tâche REN peut être utilisée dans la phase d'analyse de la requête afin de reconnaître les NE dans la requête car cela aidera plus tard à identifier les documents pertinents dans la base de données et à extraire la bonne réponse[46]. Par exemple, l'entité nommée النهارُ "le jour" peut être classée en tant qu'organisation (c'est-à-dire journal) ou en tant que espace de temps en fonction du contexte. Par conséquent, la classification appropriée pour l'entité nommée النهارُ aidera à décider quel groupe de documents doit être ciblé et recherché pour trouver la bonne réponse.

2.5.5 Clustering de texte

Clustering de texte peut exploiter REN pour classer les clusters résultants en fonction du ratio d'entités que contient chaque cluster. Cela améliorera le processus d'analyse de la nature de chaque cluster et améliorera également l'approche de Clustering en termes de caractéristiques sélectionnées.

2.6 Traitement automatique de la langue naturelle

Le domaine de la TALN vise à comprendre et manipuler automatiquement le langage humain en utilisant des systèmes informatiques intelligents. Les progrès dans ce domaine remontent aux années 1960, lorsque les premiers systèmes simples de TAL ont été publiés, par exemple. ELIZA[47]. Les fondements de ce domaine de recherche sont

informatique, intelligence artificielle et linguistique. Les applications possibles du langage non linéaire sont la traduction automatique, l'interaction homme-machine et la recherche d'information[48].

Même si la TAL peut sembler simple à première vue, il s'agit en fait d'une tâche plutôt difficile. Les humains habituellement ne sait pas combien de connaissances sont nécessaires pour comprendre le langage naturel, mais même les bébés doivent d'abord l'apprendre après leur naissance. La difficulté à comprendre le langage naturel tient à sa grande variété, à son expressivité, à son ambiguïté et à son imprécision[49].

2.7 Traitement Automatique de la Langue naturelle Médicale (TALNM)

Il y a vingt ans, le TALNM était considérée comme l'une des tâches les plus difficiles dans la recherche d'informations médicales. Aujourd'hui, la plupart des informations médicales sont sous la forme de texte libre, par ex. lettres de médecins, lettres de décharge.

Même si les avantages du stockage de l'information sous une forme structurée comme les bases de données sont évidentes, la plus grande quantité de documentation médicale reste dans la nature texte. Une explication simple serait que le langage naturel est simplement le moyen le plus facile de communiquer des informations complexes. Cependant, cela augmente la difficulté d'accéder automatiquement aux informations importantes. Le TALN essaie de s'attaquer exactement à ce problème et un aspect crucial est l'étape d'extraction et de reconnaissance des entités.

2.8 Reconnaissance de l'entité nommée biomédicale et la langue arabe

Historiquement, REN s'était concentré sur identifier des noms apparaissant en journaux (nous l'appelons de ce fait le newswire REN). Depuis fin 1990 s, la communauté de REN a prêté plus d'attention à NEs dans le domaine biomédical, probablement dans faire face au besoin de plus en plus fort du mien du texte biomédical.

Dans le domaine de la biologie et de la médecine, le NEs incluent les gènes, les protéines, les cellules, les drogues, les produits chimiques, les maladies, etc., qui sont fré-

quemment employés en texte biomédical et intéressant aux chercheurs biomédicaux. En conséquence, BioNER est la tâche qui vise à identifier automatiquement tout un tel NEs en texte biomédical. Il peut être simplement regardé comme NER appliqué au domaine biomédical.

Pour diverses raisons, l'arabe médical souffre aujourd'hui d'inconsistance terminologique, la question qui frustre les tentatives visant à faire de la langue arabe le moyen officiel d'enseignement dans les facultés de médecine du monde arabe. Par exemple, les variantes orthographiques de traduction ou de translation (par exemple, un nom de syndrome comme le syndrome de Turner peut être écrit de deux manières différentes متلازمة ترنر ou متلازمة تيرنر). Avoir différents types d'équivalence et différents termes médicaux arabes pour le même terme médical anglais étranger peut sembler inévitable en raison de différents facteurs. Tous ces phénomènes rendent plus difficile la tâche de reconnaissance psychologique de l'entité nommée (PsyNER) dans le domaine de la santé mentale que dans les domaines ouverts. Dans cette thèse, nous présentons un système de reconnaissance d'entités nommées spécialisé dans le domaine psychologique. À notre connaissance, il s'agit de la première étude à entamer l'entité nommée dans le domaine psychologique pour la langue arabe.

2.9 Conclusion

Récemment, Biomedical Named Entity Recognition (BNER) a reçu beaucoup d'attention dans le domaine de la recherche TAL pour différentes langues, comme l'anglais, l'allemand, le chinois, en raison du rôle important joué par REN dans différents systèmes TAL qui conduit à l'optimisation de la performance globale de ces systèmes. Le chapitre suivant sera consacré à la présentation des approches et des méthodes les plus utilisées dans le développement des systèmes de reconnaissance d'entités nommées et les Outils TAL pour la langue arabe en suite nous présentons quelques travaux connexes dans le domaine ouvert et le domaine médical.

Chapitre 3

Approches d'extraction d'information et outils

Sommaire

3.1 Introduction

La tâche de reconnaissance d'entités nommées (REN) a été introduite à la fin des années 90, elle a suscité un grand intérêt dans les communautés scientifiques et industrielles vu son importance dans les traitements automatiques des documents. En effet, les premiers systèmes étaient une adaptation des systèmes de reconnaissance de noms propres à base de ressources linguistiques. Ces systèmes sont classés en deux grandes familles : systèmes à base d'approches symboliques et systèmes à base d'approches statistiques. Par la suite, des systèmes combinant les deux types d'approches sont apparus sous le nom de systèmes hybrides. L'approche symbolique se base sur la connaissance humaine pour la construction manuelle des règles d'analyse sous forme de règles contextuelles. En revanche, les systèmes à base d'approches statistiques utilisent des volumes importants de données pour la mise au point automatique de modèles d'analyse.
Les approches par apprentissage pour la tâche de REN étaient peu présentées dans les premières campagnes MUC, seulement deux des cinq systèmes proposés en MUC-7 étaient à base d'apprentissage. Cette tendance s'est renversée au cours des années au point de voir naître une série de conférences dédiées aux systèmes par apprentissage CoNLL.
Dans ce chapitre nous commençons par définir le concept de l'EN. Nous présentons les principes génériques des différentes techniques utilisées pour l'extraction des entités nommées et des relations, tout en énonçant pour chaque approche les travaux de recherche réalisés pour la langue arabe dans le domaine générale et le domaine Biomédicale. Enfin, nous faisons un tour d'horizon sur les plateformes TALN existantes dédiées à cette langue pour extraction EN.

3.2 Techniques d'identification d'entités nommées

Les systèmes de reconnaissance des entités nommées reposent sur des indices qui permettent de les aider pour analyser ce texte afin de reconnaitre et catégoriser des entités nommées. Les principaux indices utilisés pour la reconnaissance des entités nommées sont divisé on deux types, internes et externes selon [50] :

1) **Les indices internes (la structure des entités)**

Concernent toutes les informations se trouvant à l'intérieur de la structure de l'entité nommée. Elles peuvent être contenues dans des listes de mots déclencheurs ou de noms propres appelées Gazetteers. Les indices internes peuvent prendre la forme d'un ou plusieurs mots ou d'une abréviation connue pour faire partie d'un nom propre.

La majuscule est une marque typographique tel que chaque mot (ou séquence de mots) commence par une lettre majuscule est considéré comme entité nommée. McDonald [50] et [51] s'appuient seulement sur cet indice pour l'identification et la délimitation des entités nommées pour l'anglais. Mots ou affixes de type classifiant (lieux et organisations) : peuvent prendre la forme d'un ou plusieurs mots ou d'une abréviation connue pour faire partie d'un nom propre (ex. : **Organisation** des Nations Unies, la **Banque** centrale, **université** d'USTO).

1) **Les indices externes (le contexte des entités))**

Sont le contexte dans lequel une entité nommée apparaît dans la phrase. Elles sont des informations complémentaires ou propriétés spécifiques sur les personnes, lieux, organisations. Ces informations peuvent aider, dans un processus automatique, à déterminer le type d'un nom propre. Les déclencheurs sont généralement une liste de fonctions ou préfixes de type M. ou Mme pour les noms de personnes et des indices de position pour les localisations. (ex. : la **ville** d'Oran, le **professeur** Bendella, le **groupe** Vivendi, Derrick, l'**inspecteur** allemand).

3.3 Approches de reconnaissance de l'entité nommée

A la fin des années 80 sous l'impulsion des conférences MUC, les systèmes d'extraction d'information sont classés en deux grandes familles : systèmes à base d'approches symboliques ou linguistique (à base de règles) et systèmes à base d'approches statistiques (ou à base d'apprentissage). Par la suite, des systèmes combinant les deux types d'approches sont apparus sous le nom de systèmes hybrides. Dans cette section, nous présentons ces trois approches tout en s'appuyant sur quelques exemples représentatifs des systèmes de reconnaissance d'EN.

3.3.1 Approche symbolique (à base de règles)

Appelée aussi approche linguistique, elle est utilisée par la majorité des systèmes de reconnaissance d'entités nommées. Son principe de base est d'utiliser des connaissances linguistiques (les indices internes ou les indices externes) et ainsi que des dictionnaires de noms propres pour établir une liste de règles de connaissance. Ces règles sont écrites manuellement par des experts du domaine (en anglais handcrafted rules).
Les règles sont généralement présentées sous la forme d'expressions régulières ou de transducteurs à états finis [52]. La maintenance de systèmes basés sur des règles n'est pas une tâche simple, car des linguistes expérimentés doivent être disponibles pour

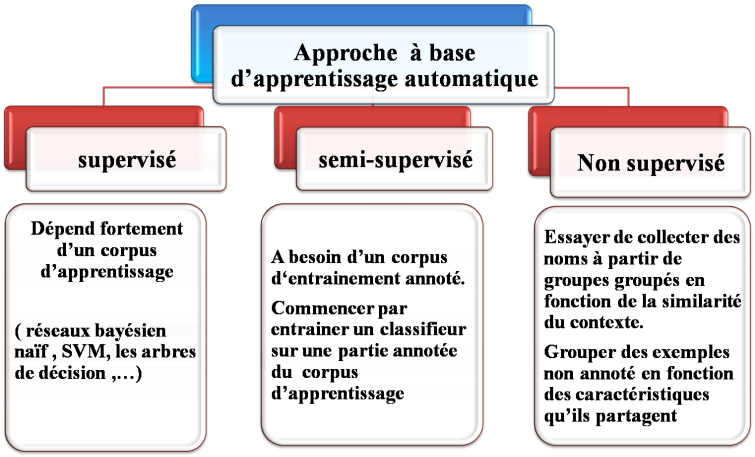
fournir les ajustements appropriés du système [53]. Ainsi, tout ajustement à de tels systèmes nécessiterait beaucoup de travail et de temps.
Les systèmes de NE basés sur des règles manquent de capacité de portabilité et de robustesse, et en outre le coût élevé de la règle maintient les augmentations même si les données sont légèrement modifiées. Ces types d'approches sont souvent spécifiques au domaine et à la langue et ne s'adaptent pas nécessairement bien aux nouveaux domaines et langages de plus la capacité du système dépend de la couverture de la base de données car le système ne peut reconnaître que les s ENs qui se trouvent dans sa base de données.

3.3.2 Approche par apprentissage machine

L'approche par apprentissage machine est très utilisée dans plusieurs domaines tels que la biomédicale, la finance. Les systèmes d'apprentissage automatique utilisent certaines données pour apprendre des régularités ou des modèles, qui peuvent être exploités pour identifient et classent les entités en classes particulières telles que les personnes, les lieux, les heures, etc. Ils peuvent être divisés en fonction du type d'apprentissage machine qu'ils utilisent. Il y a trois types d'apprentissage machine : supervisé, semi-supervisé et non supervisé.

Si le système a besoin d'un corpus avec des entités déjà étiquetées, le système utilise un apprentissage supervisé. Le système utilise un apprentissage non supervisé, s'il n'utilise aucun exemple de sortie désirée, l'apprentissage semi-supervisé est une classe spéciale d'apprentissage supervisé, où le système utilise des données étiquetées, mais il peut aussi exploiter des données non étiquetées. Nous présentons dans les prochaines sous-sections les types d'apprentissage machine appliqués en particulier dans le domaine de reconnaissance des ENs.

FIGURE 3.1 : Les types d'approches statistiques pour l'EI



Apprentissage supervisé

La technique d'apprentissage supervisé consiste à utiliser un corpus préalablement annotés pour réaliser la tâche d'extraction des ENs. Elle se déroule en deux étapes : la première étape est 1' apprentissage qui consiste à construire un processus automatique d'extraction des entités nommées pour un corpus d'entraînement annoté. La deuxième étape consiste à généralise le processus afin concevoir des règles permettant d'extraire les entités nommées dans de nouveaux documents[55].
Les résultats sont dépond du modèle qu'il a appris. Parmi les algorithmes d'apprentissage supervisé les plus utilisés, on trouve les chaînes de Markov cachée [56], les arbres de décision(en anglais, Dicision Tree(DT))[57], l'entropie maximale (en anglais, Maximum Entropy(ME))[58] et machine à vecteurs de support (en anglais, Support Vector Machines(SVM))[59].

Apprentissage semi-supervisé

Cette technique d'apprentissage combine des données étiquetées et des données non étiquetées. Donc elle se situe entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées. Cette combinaison permet d'améliorer la qualité de 1'apprentissage, car l'intervention humaine est nécessaire pour l'annotation des données non annotées [60], ce qui rend le coût d'apprentissage de cette technique élevé [61].

Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, la technique d'apprentissage non supervisé consiste à apprendre à classer sans supervision. Elle vise à construire des groupes (clusters) d'objets similaires à partir d'un ensemble hétérogène d'objets.
Cette approche repose sur une mesure précise de la similarité basée sur les ressources lexicales comme par exemple WordNet, sur les schémas lexicaux et sur des statistiques calculées à partir d'un corpus large non annoté pour construire les clusters, [62].

3.3.3 Approche hybride

L'approche hybride consiste à combiner 1' approche à base de règles et 1' approche d'apprentissage, et créer une nouvelle méthode pour l'extraction des ENs. Cette methode permet de profiter des avantages de l'utilisation des deux approches : symbolique et statistique. Ce qui de produire un système idéal pour la reconnaissance des ENs. Bien que ce type d'approche peut obtenir de meilleurs résultats que d'autres approches, mais les points faible de l'approche à base de règles reste la même lorsqu'il est nécessaire de changer le domaine des données.

3.4 Travaux connexes

Nous nous intéresserons plus particulièrement aux travaux effectués pour la reconnaissance des entités nommées arabe. La première tentative sur la reconnaissance des EN pour l'arabe était le système TAGARAB en 1998 [63]. Ensuite le travail de Shaalan et Raza [23] en 2009 et de Zaghouani et al[64] en 2010. Ces Travaux reposent sur des méthodes à base de règles.

3.4.1 NER dans le domaine général

Systèmes NER basés sur des règles

Dans l'approche basée sur des règles, le NER est effectué en utilisant des règles linguistiques faites à la main et le lexique existant. Il a été largement utilisé dans de nombreuses études qui ont adapté la méthode basée sur des règles pour extraire des entités nommées prédéfinies. Dans ce cadre, nous rappelons quelques systèmes de REN arabes :

Shaalan et Raza [23] ont développé un système de reconnaissance des EN arabes (NERA) en utilisant une approche fondée sur des règles. Ce système repose sur l'utilisation d'un

ensemble de dictionnaires d'EN et sur une grammaire sous forme d'expressions régulières pour la reconnaissance des EN. Le meilleur taux F-mesure acquis par ce système est de 98.6%. Suivant le même principe, les travaux de Zaghouani et al.,[64] ont présenté un module de repérage des EN à base de règles pour la langue arabe, la seule différence est qu'ils ont procédé à une première étape de prétraitement lexical qui prépare le texte pour son analyse linguistique, ce module a été évalué sur un corpus de presse. La valeur de F-mesure apportée par ce module est de 47.35% pour le type organisation et 95.10% pour le type date.
Fehri, Haddar et Ben Hamadou [65] ont adopté une approche basée sur des règles pour identifier, extraire et combiner les informations spatiales et temporelles des documents de texte en arabe en utilisant les techniques de Traitement Naturel du Langage (TAL), de Recherche d'Information Géographique (Geographic Information Retrieval (GIR)) et d'Information Temporelle (Temporal Information Retrieval (TIR)). Un ensemble d'étapes a été utilisé pour développer ce système, à partir de la création de répertoires spatiaux et temporels arabes, jusqu'au traitement de texte. À cette étape, cette approche utilise deux composants principaux : l'analyseur morphologique arabe (Standard Arabic Morphological Analyzer (SAMA) et la bibliothèque de règles qui consiste en un ensemble de règles grammaticales.

Al-Ahmari et Al-Johar [66] ont introduit une approche basée sur des règles pour la reconnaissance des NE, qui inclut les noms de personnes, de lieux et d'organisations en texte arabe. Le système fonctionne sur la base d'un ensemble de règles grammaticales indépendantes du domaine ainsi que d'un tagueur de partie arabe du langage en plus des nomenclatures et des listes de mots déclencheurs. Ce système comprend trois étapes principales : prétraitement, (catégorisation morphosyntaxique) POS tagging et algorithme de reconnaissance. La méthode a été appliquée à deux corpus arabes de domaines différents. La précision du système a été mesurée en termes de précision comme suit pour la personne 80,7%, pour l'emplacement 93,2% et pour l'organisation 75,4%.

En résumé, le tableau 3.1 illustre les caractéristiques des systèmes de reconnaissance d'entités nommées qui ont été développés pour la langue arabe mentionnés ci-dessus.

TABLE 3.1 : Une comparaison entre les systèmes symbolique appliqué au texte arabe

Système	Corpus	Entités	Stemming	Gazetteers	POS	domaine
TAGARAB(1998) [22]	✗	Nombre, Temps, lieu et les noms de personnes	✔	✔	✔	Politique /MSA
NERA : Named entity recognition for Arabic(2009) [23]	De nombreuses ressources pour construire leurs propres corpus	les noms de Personnes, Lieu, Organisation, Date, Temps, ISBN, Mesure, les noms de fichier,les numéro de téléphone et Prix	✔	✔	✗	Politique, économique /MSA
Adapting a resource-light highly multilingual named entity recognition system to Arabic(2010) [24]	sources de nouvelles en ligne (le journal tunisien Assabah et le journal libanais Alanwar)	les noms de Personnes, Lieu, Organisation, Dates, expression numérique	✔	✔	✗	Générale / MSA
Automatic extraction of spatio-temporal information from arabic text documents(2015) [65]	textes de journaux	informations spatiales et temporelles	✔	✔	✗	Générale / MSAv
Cross domains Arabic named entity recognition System (2016) CH318	ANERcorp et Alsulieti corpus	les noms de personnes, de lieux et d'organisations	✔	✔	✔	Général/ MSA

Systèmes NER basés sur le ML

Dans l'approche ML, le problème NER est converti en un problème de classification et donc les techniques de ML sont utilisées comme une solution. Les approches d'apprentissage automatique appliquées dans la littérature aux textes arabes sont l'approche d'apprentissage supervisé, l'approche semi-supervisée et l'approche non supervisée. Les techniques qui appartiennent à l'apprentissage supervisé (SL) sont les modèles d'entropie maximale, les champs aléatoires conditionnels (Conditional Random Fields CRFs (CRFs)), les machines vectorielles de support (SVM) et les réseaux de neurones (NN) ; ceux-ci ont été appliqués aux textes arabes.

L'entropie maximale a été appliquée par Benajiba et al. [67], alors que les CRF ont été appliqués par Benajiba et Rosso [68], Abdul-Hamid et Darwish [69], Bidhendi et al. [70], Morsi et Rafea [71] et Alotaibi (2015)[72] . Benajiba et al. [73] et Koulali et Abdelouafi [74] ont mis en œuvre des SVM. Benajiba et al. [75] ont étudié les ramifications de l'utilisation de différentes caractéristiques avec des modèles tels que SVM, ME et CRF et ont conclu que les SVM et les CRF surpassaient le modèle ME. Ils ont également expliqué que le choix des fonctionnalités appropriées est une phase très importante de tout système à base de ML. Mohammed et Omar [76] ont adapté les réseaux de neurones dans leur approche.

Les auteurs Mohammed et Omar [76] ont proposé un système NER arabe basé sur le NN arabe (en anglais, Arabic Neural Networks (ANN)) qui vise à classer les NE arabes. Leur système se compose de trois étapes. Dans la première étape, le texte est pré-traité afin de nettoyer les données collectées. Dans la deuxième étape, les lettres arabes sont converties en alphabet romain. Enfin, dans la troisième étape, les données sont classées en utilisant les RNA. La précision de leur système a atteint 928%. Ce résultat a été comparé au résultat obtenu par les arbres de décision (DT) qui ont atteint 87% lorsqu'ils ont été appliqués sur les mêmes données.

En 2013, Morsi et Rafea [71] ont utilisé une approche ML supervisée pour évaluer l'effet de différentes caractéristiques sur la performance de la RN arabe effectué via des modèles de CRF. Le meilleur résultat des diverses combinaisons de caractéristiques était avec une F-mesure de 68,05%. Un apprentissage semi-supervisé du NER arabe connu sous le nom d'ASemiNER a été proposé par l'auteur Althobaiti (2016) [77]. Cette approche, en utilisant une combinaison de techniques d'apprentissage semi-supervisés et à distance. Dans l'algorithme semi-supervisé pour l'identification des éléments de réseau, il n'est pas nécessaire d'avoir des données d'apprentissage annotées ou des no-

menclatures. Il nécessite seulement, pour chaque type d'élément de réseau, une liste de départ de quelques instances pour lancer le processus d'apprentissage. Les nouveaux aspects de cet algorithme comprennent (i) une nouvelle façon de produire et de généraliser les modèles d'extraction (ii) un nouveau critère de filtrage pour éliminer les modèles bruyants (iii) une comparaison de deux mesures de classement pour déterminer les NE candidats les plus fiables. Divers schémas de combinaison plus classiques sont utilisés pour combiner des méthodes de supervision minimales. En particulier, elle a utilisé une variété de schémas de combinaison de classificateurs, y compris la procédure Bayesian Classifier Combination (BCC), récemment proposée pour l'analyse des sentiments.

Notre étude de la littérature n'a révélé aucun système NER arabe employant du ML non supervisé mais il a été appliqué par de nombreux chercheurs NER pour d'autres langues.

En résumé, le tableau 3.2 illustre les caractéristiques des systèmes de reconnaissance d'entités nommées symbolique qui ont été développés pour la langue arabe mentionnés ci-dessus.

TABLE 3.2 : Une comparaison entre les systèmes d'apprentissage appliqué au texte arabe

Système	Corpus	Méthode	Entités	Stemming	Gazetteers	POS	Domaine
Arabic Named Entity Recognition Using Artificial Neural Network (2012) [76]	ANERcorp	NN (Réseau de neurones) ,DT (arbre de décision)	les noms de Personnes, Lieu, Organisation, Divers.	✗	✗	✗	Général /MSA
Studying the impact of various features on the performance of Conditional Random Field-based Arabic Named Entity Recognition(2016) [71]	ANERcorp AQMAR corpus	CRF	les noms de Personnes, Lieu, Organisation,	✔	✔	✔	Général /MSA
Minimally-supervised methods for Arabic Named Entity Recognition (2016) [77]	Wikipedia	Bayesian Classifier Combination (BCC)	les noms de Personnes, Lieu, Organisation	✔	✗	✔	Général /MSA

Systèmes NER hybrides

Au cours des dernières années, certains systèmes hybrides ont été mis en place afin d'améliorer les performances des systèmes basés sur des règles et des ML. L'approche hybride consiste à combiner l'approche basée sur des règles et l'approche ML afin de profiter des avantages des deux. Abdallah et al. [78] ont utilisé une méthode hybride pour analyser les corpus construits entre les données ANERcorp et ACE 2003 extrait de sources d'information. Ils ont appliqué un classificateur d'arbre de décision J48 à une méthode basée sur des règles pour extraire les ENs personnes, lieux et d'organisations. Les expériences initiales ont montré une amélioration de F-mesures entre 8% et 14% en comparaison avec le système basé sur des règles et l'approche d'apprentissage automatique.

En 2014, Meselhi et al. [79] ont présenté une nouvelle approche hybride pour la REN en arabe. Ce système a été présenté avec la tâche d'extraire les ENs de type personne, lieu et organisation à partir d'un corpus ANERcorp extrait de newswires et d'autres sources Web. L'intégration d'une approche basée sur des règles avec une approche ML a été combinée avec la sélection et la correction des étiquettes pour identifier les faux négatifs. L'extraction des entités-personnes a atteint une F-mesure de 96,65% tandis que les autres entités, la localisation et l'organisation ont atteint, respectivement, de F -mesure 94,8% et 92,9%.

Les auteurs Shaalan et al. [80] ont proposé un système basé sur une approche hybride. Ce système est composé de deux composants qui coopèrent pour produire l'annotation finale et chacun d'entre eux peut être utilisé indépendamment : (i) un composant NER basé sur des règles qui produit des étiquettes NE basées sur des listes de NE / mots clés et de règles contextuelles ; et (ii) un post-processeur à base de ML destiné à utiliser des caractéristiques basées sur des règles extraites de jeux de données annotés avec des éléments de réseau qui sont reconnus par l'autre composant visant à améliorer la performance globale de la tâche NER. Ces auteurs ont réalisé une F -mesure de 90%. Leur système surpasse l'état de l'art pour NER arabe en termes de précision lorsqu'il est appliqué à l'ensemble de données standard ANERCorp.

Notre système diffère de cette approche dans le domaine appliqué : Cette approche présente un système pour reconnaître l'entité nommée dans le domaine général. Notre approche présente un système de reconnaissance d'entité nommée spécialisé dans le domaine médical. Autant que nous sachions, l'extraction d'informations psychologiques en langue arabe n'a toujours pas été tentée par aucun chercheur. Les résultats empi-

riques indiquent que l'approche hybride surpasse à la fois la approche basée sur des règles et celle basée sur la ML lorsqu'elles sont traitées indépendamment.

En résumé, le tableau 3.3 illustre les caractéristiques des systèmes de reconnaissance d'entités nommées hybride qui ont été développés pour la langue arabe mentionnés ci-dessus.

TABLE 3.3 : Une comparaison entre les systèmes d'hybride appliqué au texte arabe

Système	Corpus	Méthode	Entités	Domaine
Integrating rule-based system with classification for Arabic named entity recognition (2012) [78]	ANERcorp et ACE 2003	un classificateur d'arbre de décision J48 appliqué à une méthode basée sur des règles	les noms de Personnes, Lieu, Organisation.	Général /MSA
A Novel Hybrid Approach to Arabic Named Entity Recognition (2014) [79]	ANERcorp	SVM appliqué à une méthode basée sur des règles	Les noms personne, lieuet organisation	Général /MSA
A hybrid approach to Arabic named entity recognition (2014) [80]	ACE corpora, ATB part1 version 2.0, ANERcorp	Arbres de décision, SVM et régression logistique appliqués à une méthode basée sur des règles	les noms de Personnes, Lieu, Organisation, Date, Temps, ISBN, Mesure, les noms de fichier,les numéro de téléphone et Prix	Général /MSA
Notre approche Psychological Named Entity Recognition from psychological Arabic texts (2017) [2]	le corpus est recueilli à partir du web (WBTEB, TBEEB)	SVM appliqué à une méthode basée sur des règles	8 différents types d'entités nommées, y compris les troubles mentaux désigné par DSM-IV (Trouble psychologique, phobie, Syndrome), substances psychoactives, symptômes, maladies, Organe et les médicaments	Psychologique/ MSA

3.4.2 NER dans le domaine médical

Au meilleur de nos connaissances, les recherches sur le REN se concentrant sur la langue arabe se limitent aux domaines de la politique, de l'économie et de la criminalité, et les textes utilisés étaient principalement basés sur les ressources newswire. De plus, les auteurs Alanazi et al. [81] sont les seuls qui analysent des documents médicaux arabes. Mais l'analyse de document psychologique n'a pas encore été explorée. Cela a fourni une autre motivation pour tester notre approche de ce domaine inexploré. De plus, ce travail soutient et souligne le manque de travail similaire dans le domaine du REN médical en arabe.
Le NER (bio) médical est plus difficile que le NER général en raison des situations complexes telles que l'expression irrégulière, les frontières à peine distinguées et les termes changeant quotidiennement.

- Inconsistance terminologique en traduction médicale de l'anglais vers l'arabe.
- Variantes orthographiques pour la traduction.
- Néologismes : En raison de nombre immense des médicaments de marque émergents, d'éléments chimiques et d'acides et de principes actifs, l'utilisation de néologismes s'est avérée être un caractère inhérent à l'arabe médical.

Les auteurs Alanazi et al. [81] ont développé un système de reconnaissance d'entités nommées appliqué au texte arabe dans le domaine médical (NAMERAMA), ce nouveau système NER basé sur le Bayesian Belief Network (BBN). Ce système comprend quatre étapes ; prétraitement, analyse de données, extraction de caractéristiques et classification. Il extrait les noms de maladies, les symptômes, les méthodes de traitement et les méthodes de diagnostic à partir du texte arabe moderne dans le domaine médical.

3.5 Méthodes d'extraction des relations entre ENs

L'extraction de relations entre entités nommées est une opération importante pour beaucoup d'applications et de nombreuses études ont été proposées dans différents cadres de travail. Ces méthodes peuvent essentiellement être classées en trois grandes catégories : approche basé sur des règles, approche basé sur ML et approche hybride.

L'objet de la tâche d'extraction de relations n-aires ou binaires est d'établir des liens sémantiques, des liens d'hyperonymie, de synonymie, de méronymie, etc. entre entités à partir de textes. Divers travaux [82], [83] et [84] ont été effectués pour l'anglais surtout à bases d'approches statistiques, des approches syntaxiques, des approches utilisant des

marqueurs [85], ou encore des approches utilisant la programmation linéaire pour l'extraction des informations[86].
Peu de travaux se sont intéressés à l'acquisition de relations sémantiques entre termes à partir de textes spécialisés en MSA (Modern Standard Arabic).

Dans la première approche, les règles sont généralement implémentées sous la forme d'expressions régulières ou de transducteurs à états finis. A partir des études réalisées en langue arabe, nous mentionnons Ben Hamadou et al. [93] et de Boujelben et al. [94]. Ces auteurs ont extrait un ensemble de modèles linguistiques d'un corpus d'apprentissage. Par la suite, ils ont réécrit ces modèles en transducteurs à état fini au sein de la plate-forme linguistique NooJ, en utilisant des grammaires spécifiquement locales. Cette approche utilise une représentation des règles linguistiques au moyen de transducteurs. [93] ont rapporté une F-mesure de 70%, tandis que [94] a obtenu une F-mesure de 60%.

Ce résultat est significatif car Ben Hamadou et al. [93] sont limités,seulement, aux relations fonctionnelles entre les paires NE (PERS-ORG). Ainsi, ils se concentrent uniquement sur une paire NE, ce qui leur permet de construire des règles plus précises et plus concises. En revanche, Boujelben et al.[94] sont intéressés à extraire plus de relations entre cinq paires de NE (PERS-LOC, PERS PERS, PERS-ORG, ORG-LOC et LOC-LOC). Pour extraire les relations entre ces paires NE, les auteurs ont élaboré cinq sous-grammaires. Chaque grammaire contient le modèle des relations entre chaque paire.

Pour automatiser complètement la tâche d'extraction des relations, certaines études de recherche ont été orientées vers les méthodes d'apprentissage, y compris les techniques d'apprentissage non supervisées, semi-supervisées et supervisées.

Les méthodes non supervisées utilisent des quantités massives de texte non étiqueté et reposent presque entièrement sur des techniques de regroupement et des similarités entre des caractéristiques ou des mots contextuels. Par exemple, Hasegawa et al.[61] se sont concentrés sur le regroupement des paires de EN en fonction de la similarité des mots de contexte qui interviennent entre les ENs.

Pour remédier aux problèmes de l'approche non supervisée, Certaines études ont été orientées vers des approches d'apprentissage semi-supervisées ou des méthodes bootstrap. Cette approche repose sur un petit ensemble de graines initiales. Un échantillon de modèles linguistiques ou de certaines instances de relations cibles peut être utilisé pour acquérir des relations plus fondamentales jusqu'à la découverte de toutes les relations

cibles, comme dans [96] et [97].

Une dernière approche sous les techniques ML est la méthode supervisée , qui repose sur un corpus entièrement étiqueté. Cette approche considère l'extraction de relation comme une tâche de classification. Parmi les techniques supervisées les plus utilisées, nous citons les machines vectorielles de support (SVM), les champs aléatoires conditionnels (CRF), l'arbre de décision et l'entropie maximale (MaxEnt). Une tentative récente d'extraire les relations entre les NE arabes a été faite par Alotayq [98], qui a utilisé un classificateur basé sur MaxEnt. Basé uniquement sur les informations morphologiques et part of-speech (POS), ce système obtient des résultats satisfaisants lorsqu'il est appliqué au corpus ACE.

Les deux catégories d'approches décrites ci-dessus peuvent être combinées pour obtenir une approche mixte. Récemment, des études de recherche ont été orientées vers l'utilisation d'approches hybrides car cette approche permet d'obtenir une performance meilleure que l'approche fondée sur des règles ou l'approche basée sur le ML uniquement. Certaines études ont été réalisées sur un domaine spécifique, comme dans le domaine biomédical. A titre d'exemple, Ben Abacha et Zweigenbaum [99] proposent une approche hybride pour extraire les relations entre maladies et traitements. Ces auteurs ont combiné une méthode d'apprentissage supervisé avec une technique basée sur des règles. Pour la méthode linguistique, un ensemble de motifs est construit manuellement à partir du corpus d'apprentissage et d'autres corpus MEDLINE ; dans cet ensemble, un poids est associé à chaque motif. Ce poids sert à choisir le motif le plus commode dans le cas de plusieurs candidats d'extraction. Pour la méthode ML, les auteurs ont examiné le classificateur SVM, en utilisant des caractéristiques lexicales, morpho-syntaxiques et sémantiques. Les résultats obtenus de cette approche hybride montrent une amélioration par rapport aux méthodes à base de ML et de pattern.

Boujelben et al. [100] ont proposé une méthode hybride pour extraire les relations entre entités nommées arabe, ils ont utilisé les méthodes d'apprentissage pour identifier des règles d'association caractéristiques des relations entre paires d'entités nommées. Afin de l'améliorer, ils ont utilisé des algorithmes génétiques lors de l'étape de filtrage. La méthode basée sur des règles offre une analyse significative du contexte de chaque EN et de ses relations avec les autres ENs.

Dans notre approche, nous avons adapté la Programmation Linéaire (PL) pour l'identification des relations entre NEs pour la langue arabe, similaire à l'approche de [86]. On appelle Programmation Linéaire, le problème mathématique qui consiste à optimiser

(maximiser ou minimiser) une fonction linéaire de plusieurs variables qui sont reliées par des relations linéaires appelées contraintes.

Dans cette thèse, nous étudions une instanciation spécifique de ce problème dans le contexte de l'identification des entités nommées et des relations entre eux dans un texte psychologique de forme libre. Donc, nous développons une formulation de programmation linéaire pour traiter ce problème d'inférence globale et l'évaluer dans le contexte de l'apprentissage simultané d'entités et de relations nommées.
En résumé, le tableau 3.4 illustre les caractéristiques des systèmes d'extraction de relations qui ont été développés pour la langue arabe mentionnés ci-dessus.

TABLE 3.4 : Une comparaison entre les systèmes d'extraction de relations systèmes d'extraction de relations appliqués au texte arabe

Système	Méthode	Relation entre Entités
Multilingual extraction of functional relations between arabic named entities using Nooj platform (2010a) [93]	Symbolique	PERS-ORG
Rules based approach for semantic relations extraction between Arabic named entities (2012) [94]	Symbolique	PERS-LOC, PERS-PERS, PERS-ORG et ORG-LOC
Discovering relations among named entities from large corpora (2004) [95]	Apprentissage (méthode non supervisées)	PERSON-GPE (PER-GPE). [1] et COMPANY-COMPANY (COM-COM)
Extracting relations between Arabic named entities.(2013) [98]	Apprentissage (méthode supervisé)	**Physical :** Tels que situés dans ou à proximité d'une entité physique, **Part-Whole :** peut être un lien géographique, subsidiaire ou artefact entre une partie et une entité entière, **Personal-Social :** Décrit la relation entre les personnes, **ORG-Affiliation :** Décrit l'emploi, la propriété, l'adhésion, etc, **Agent-Artefact :** Tels que l'utilisateur, le propriétaire ou le fabricant (d'un artefact), **GEN-Affiliation :** liens tels que : Citoyen de, Résident de, Religion, Origine etc.
Genetic algorithm for extracting relation between named entities (2013b) [99]	Hybride	PERS-LOC, PERS-PERS, PERS-ORG, LOC-LOC et ORG-LOC
Notre approche	Apprentissage (méthode non supervisée) ILP	**Symptom_of :** décrit la relation entre ENs (Symptom, Mental Disorder), **Has_Symptom :** décrit la relation entre ENs (Mental Disorder, Symptom), **Due_to :** décrit la relation entre ENs (Mental Disorder, Psychoactive substance) **Induced :** décrit la relation entre ENs (Psychoactive substance , Mental Disorder), **Is_a :** décrit la relation entre ENs (Mental Disorder, Disease), **Causes :** décrit la relation entre ENs Mental Disorder, Mental Disorder) et (Disease , Disease) Voir la table 5.1

52

3.6 Approches de traduction des EN

La traduction automatique des ENs est une tâche pour laquelle la reconnaissance des EN constitue également une amorce importante. Il existe globalement deux grandes approches de base de TA : l'approche experte et l'approche empirique [101]. La première est fondée sur les connaissances d'experts humains. Elle renferme trois méthodes dérivées : la TA directe, la TA à base de règles de transfert et la TA fondée sur une inter-langue. La deuxième est fondée sur l'extraction des connaissances à partir des quantités importantes de données textuelles. Elle peut être subdivisée en deux grandes familles : l'approche par analogie (ou à base d'exemples) et l'approche statistique. Ces deux dernières, contrairement aux méthodes de l'approche experte, ne nécessitent aucune connaissance a priori pour développer un système de traduction.

La TA des EN est effectuée dans la majorité des travaux en suivant l'approche statistique. Parmi ces travaux, nous citons par exemple celui de Al-Onaizan et Knight [102]. Ce travail consiste à traduire les EN arabes vers l'anglais en utilisant un algorithme basé sur des ressources monolingues et bilingues. En effet, étant donné une EN dans la langue source, l'algorithme de traduction génère d'abord une liste de classement des candidats de traduction en utilisant les ressources bilingues et monolingues. Ensuite, la liste des candidats est recalculée à l'aide de différents indices monolingues. Dans le même contexte, nous trouvons le travail de Ling et al.[103] qui permet de récupérer une liste de documents web dans la langue cible, d'extraire les textes d'accroche (anchor textes) à partir des liens de ces documents et de recenser la bonne traduction de l'EN à partir de ces textes en utilisant une combinaison de traits dont certains sont spécifiques aux textes d'accroche.

Les travaux basés sur une approche experte pour la traduction des EN sont rares. Nous pouvons citer comme exemple celui de Gornostay et Skadina[104] qui permet la traduction des toponymes de l'anglais vers le letton. Ce travail se base sur l'utilisation d'un dictionnaire, d'une part, et sur l'utilisation des patrons syntaxiques, d'autre part. Ces patrons consistent à translittérer le toponyme ou le translittérer et le traduire ou le translittérer et lui ajouter une nomenclature.

En résumé, dans l'approche experte de la TA, les traductions sont construites en utilisant d'importants dictionnaires et des règles linguistiques sophistiquées. Les systèmes de TA basés sur cette approche fournissent un bon niveau de qualité de traduction dans des domaines non spécifiques. Leur adaptation à un domaine spécifique est possible mais elle représente un coût important en termes de temps et de moyens.

Quant à l'approche empirique, les traductions sont produites à partir d'événements rencontrés dans des corpus bilingues et à partir desquels le système va puiser ses connaissances. Pour fournir des traductions de bonne qualité, les systèmes de TA basés sur cette approche nécessitent une quantité importante de corpus bilingue. L'apprentissage des modèles dans cette approche est un processus rapide, automatique et peu coûteux.

Dans notre travail, nous avons utilisé une approche statistique qui consiste à traduire les EN arabes vers l'anglais en utilisant un algorithme basé sur les Gazetteers bilingues. En effet, étant reconnu une EN et sa catégorie dans la langue source (arabe), l'algorithme de traduction parcours le Gazetteer est généré la traduction dans la langue cible (anglais).

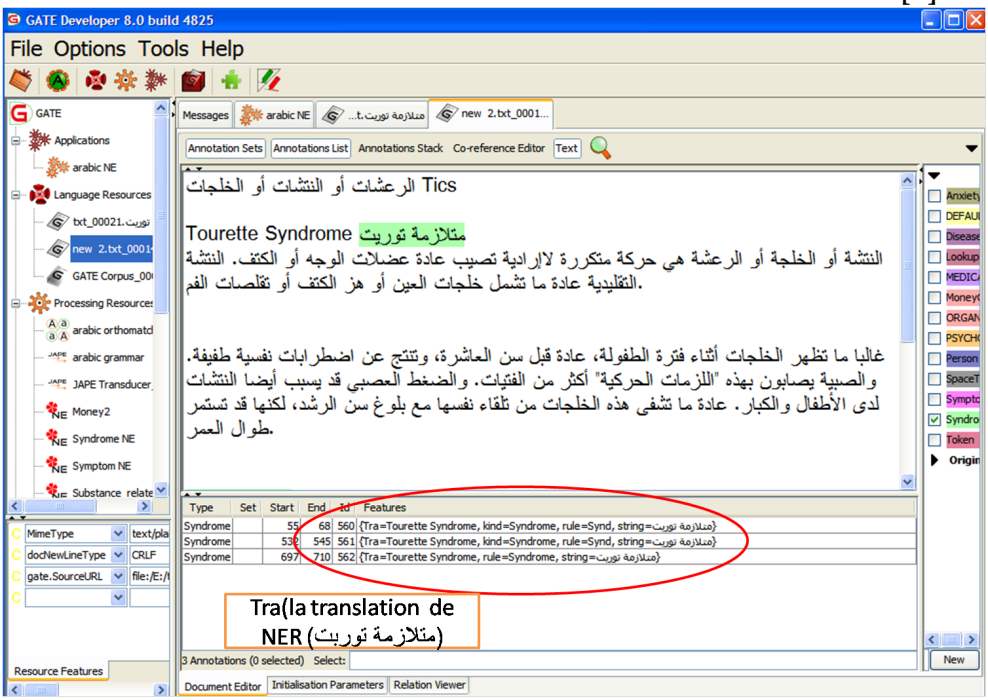
3.7 Outils TALN

Pour générer du NE à partir du texte, un outil peut être utilisé cet outil appelé environnements de développement intégré, les environnements communs sont :

3.7.1 GATE

1. **GATE** (**G**énérale **A**rchitecture pour l'ingénierie de **TE**xte) est une application libre en open source qui permet aux utilisateurs de construire et d'évaluer les applications pour diverses tâches de NLP en utilisant les différentes ressources intégrées et les composants en plusieurs langues et domaines développée depuis 1995 à l'Université de Sheffield. Quand il s'agit de NER, GATE facilite le développement de systèmes NER, tel que il fournit à l'utilisateur a capacité de mettre en œuvre des règles de grammaire comme transducteur d'états finis à l'aide de JAPE. Ce qui suit résume les principales composantes du GATE. GATE offre des fonctionnalités très complètes, mais en contrepartie se révèle assez complexe à prendre en main. Il dispose d'une API, GATE Embedded, qui permet son intégration dans d'autres applications.

FIGURE 3.2 : Extraction d'entités nommées arabes avec Gate[2]



GATE peut être utilisé de deux façons différentes : environnement de développement ou bibliothèque. L'utilisation la plus simple est comme environnement de développement au travers des ressources développées par ses concepteurs.

2. **CREOLE** (Collection of REusable Objects for Language Engineering) Englobe divers composants, réutilisables indépendamment les uns des autres pour le traitement automatique de la langue naturel, Nous pouvons définir dans CREOLE trois types de ressources :

   ⋆ **Ressources langagières** (LRs : Language Resources) Il s'agit d'un certain nombre de données linguistiques tels que des documents, des corpus, des lexiques ou des ontologies. A l'heure actuelle toutes les LRs sont basées sur le texte mais le modèle peut être étendu pour manipuler des données multimédias.

   ⋆ **Ressources de traitement (Algorithmique)** (PRs : Processing Resources) représentent les ressources de caractère algorithmique tels que les segmenteurs, les étiqueteurs, les analyseurs etc. Dans la majorité des cas les PRs sont utilisées pour traiter les données fournies par les LRs.

   ⋆ **Ressources de visualisation** (VRs : Visual Resources) Ce sont des composants graphiques, permettent la présentation des résultats à l'intérieur de l'environnement de développement GATE et l'édition d'autres types de ressources.

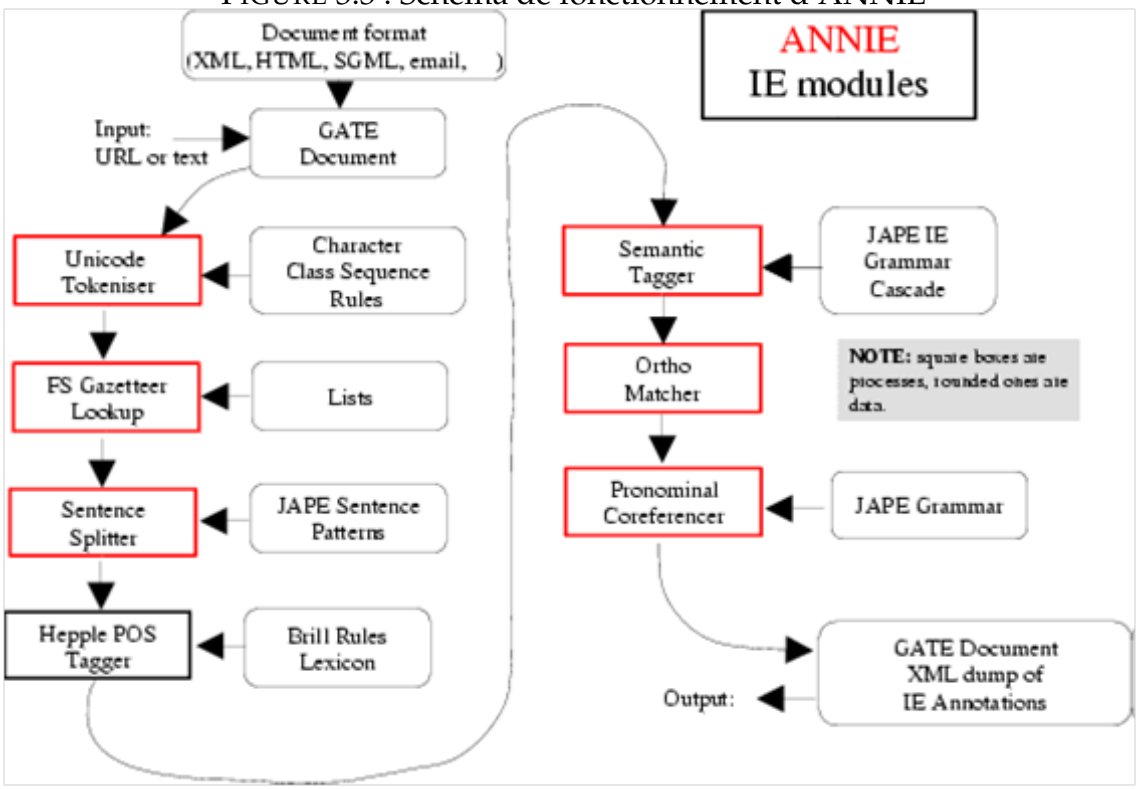
Parmi les ressources algorithmiques fournies par GATE le plugin ANNIE.

3. **ANNIE** (**A** **N**early-**N**ew **I**nformation **E**xtraction System, pour système quasi nouveau pour l'extraction d'information), est un composant de GATE, formé de plusieurs modules parmi lesquels un analyseur lexical, un Gazetteer, un segmenteur de phrases, un étiqueteur, un module d'extraction d'entités nommées et un module de détection de coréférences. ANNIE offre toute la gamme de Processing Ressources nécessaires au dépistage d'information sur les textes (Information Extraction). Et il offre aussi entre autre les outils pour le traitement de phrases, pour la détection des entités et pour la détection de références entre les sections d'un texte.

   ANNIE se compose :

   - ⋆ d'un découpeur de « tokens » (tokenizer), dont le rôle est de diviser le texte en jetons simples (ponctuations, nombres, mots, etc.),
   - ⋆ d'un gazetteer, qui est un ensemble de listes. Le rôle du module de Gazetteer est d'identifier les noms d'entités dans le texte en fonction des listes. Ces listes contiennent, par exemple, tous les noms de villes, ou de pays. Chaque liste représente un ensemble de noms, tels que les noms de villes, d'organisations, les jours de la semaine, etc.
   - ⋆ d'un séparateur de phrase (sentence splitter), qui comme son nom l'indique, sépare le texte qui lui est fourni en entrée en un ensemble de phrases en fonction de la ponctuation. En effet, l'extraction d'information s'effectue phrase par phrase. Aucune information hors de cette phrase ne peut être utilisée pendant le processus. ce module est une cascade de transducteurs à états finis qui segmentent le texte en phrases. Il est requis pour le Tagger Part of Speech (PoS).
   - ⋆ d'un POS-Tagger ,qui se charge d'étiqueter grammaticalement le texte. Le POS-tagger employé par GATE est une modification du tagger de Brill[111]. Il produit une étiquette de partie du discours sous la forme d'une annotation pour chaque mot ou symbole.
   - ⋆ d'un Named-Entity transducer (NE transducer), c'est la partie de l'algorithme qui va utiliser toutes les informations précédentes pour essayer de trouver les entités nommées. Le transducer va utiliser les règles par défaut de GATE ou des règles écrites par l'utilisateur. Les règles utilisées sont écrites en JAPE (Java annotation Patterns Engine).

FIGURE 3.3 : Schéma de fonctionnement d'ANNIE



GATE a été construit à l'origine pour l'extraction d'entités nommées en anglais, mais par la suite des améliorations ont permis de l'adapter pour de multiples fonctions dans plusieurs langues, nous présentons notre contribution dans le chapitre système proposé concernant l'extraction de entités à partir de textes psychologiques arabe.

3.7.2 NooJ

NooJ est un outil gratuit pour le développement linguistique dans un environnement multilingue. NooJ permet au développeur de construire, tester et maintenir des ressources lexicales a large couverture, ainsi que des outils morpho-syntaxiques appliqués pour le traitement automatique pour la langue arabe. Cependant, de nombreux chercheurs utilisent des outils NooJ tels que W Brini et al [105], Mesfa[106] et d'autres.

3.7.3 LingPipe

LingPipe est une bibliothèque commerciale de traitement de langage naturel implémentée en Java. LingPipe fournit un ensemble d outils pour le traitement de texte, il est utilisé pour effectuer des tâches telles que[107] : Trouvez les noms des personnes, des

organisations ou des lieux. LingPipe est multilingue comme l'arabe, l'anglais, le chinois [108][109].

De nombreux chercheurs ont utilisé l'outil LingPipe tel que S Abdel Rahman et al [108] et d'autres. Cependant, dans cette recherche, nous avons utilisé l'outil GATE pour extraire l'entité nommée à partir d'un texte arabe non structuré, car il a de nombreuses fonctionnalités telles que :

- Facile à utiliser.
- Avoir la plupart des outils tels que gazetteer, tokenizer, etc.
- Facile à construire des règles JAPE.
- Avoir beaucoup de plugins utilisés pour la langue arabe.

3.8 Conclusion

Dans ce chapitre, nous avons commencé par la description des différentes approches d'extraction des ENs et relation. En citant quelque approche dans les deux domaines (générale et biomédicale) lies à la langue arabe. En suite nous avons présenté quelques outils pour le TALN dédiés à cette langue. Dans la troisième partie, Contribution, nous avons décrit notre méthodologie pour 1' extraction des ENs à partir de corpus en suite nous avons concentré sur les évaluations de notre travail. En premier lieu, nous décrivons les données utilisées. Par la suite, nous présentons les différentes évaluations qui ont été faites avec une discussion des résultats obtenus.

Deuxième partie

Contribution - l'extraction d'information psychologique à partir des textes psychologiques arabe

Chapitre 4

La reconnaissance d'entité nommée psychologique

Sommaire

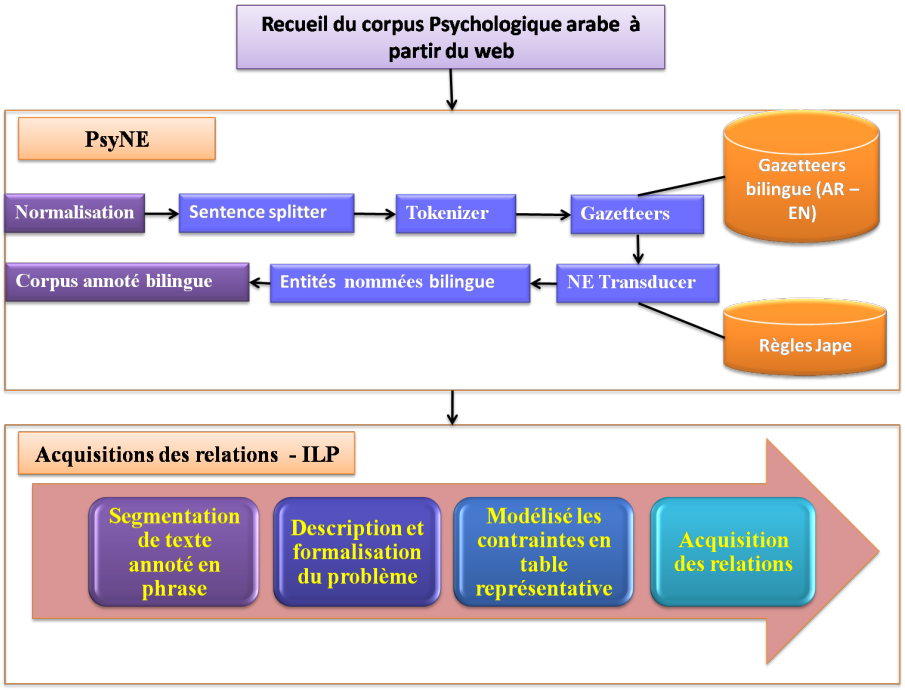
4.1 Introduction

Dans ce chapitre, nous nous intéressons à La reconnaissance d'entité psychologique (PsyNER)[2][3] qui est fondée sur une approche symbolique où l'extraction s'effectue en se basant sur un ensemble de Gazetteers et de règles construites manuellement en exploitant l'outil d'extraction des entités nommées disponibles sous la plateforme GATE. Ce chapitre décrit l'architecture générale de l'approche PsyNER proposée. Les composants du système sont décrits dans la section 4.2. La mise en œuvre des Gazetteers (nomenclatures) et les règles dans le cadre de GATE est expliquée en détails dans la section 4.3 et la section 4.6. La formalisation des contraintes pour l'extraction des relations est démontrée et discutée en détails dans la section 4.8. La dernière section, qui conclut ce chapitre, est un bref résumé des principaux points soulevés.

4.2 L'approche proposée

Notre approche comporte deux étapes : Dans la première étape, nous extrayons les entités psychologiques et déterminons leurs catégories. Dans la deuxième étape, nous identifions les relations entre ces entités extraites. La figure suivante figure 4.1 montre l'architecture détaillée de notre système.

FIGURE 4.1 : L'architecture de l'approche proposée.



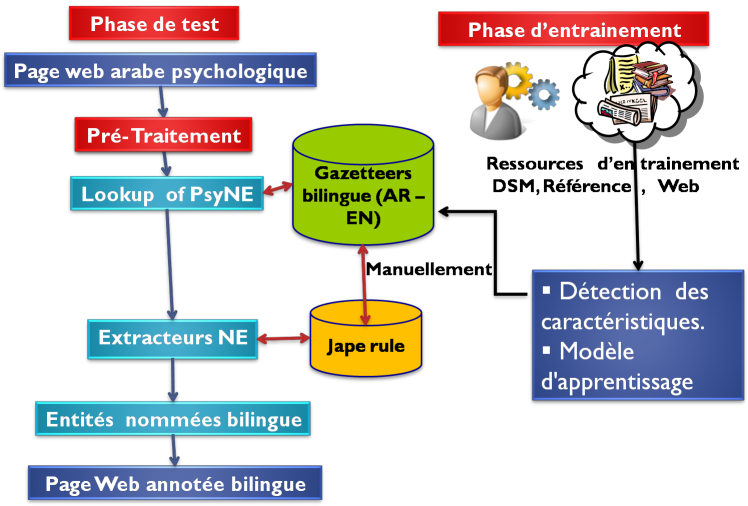
Dans le domaine de la santé, en particulier de la psychologique, l'extraction d'information concerne des entités telles que les noms de syndrome ou de phobie, par exemple. Ces données permettent notamment de peupler des bases de connaissances, utilisées en psychologique. Il faut souligner que la nature spécifique de ces données, d'un type différent de celles sur lesquelles l'intérêt des conférences traitant de l'extraction d'information – MUC en particulier. Généralement, un processus d'extraction d'informations englobe les tâches suivantes : Identification / reconnaissance des entités nommées ; Extraction des relations entres les entités nommées. Dans ce chapitre, on va présenter la tâche d'identification des entités nommées psychologiques et dans le chapitre suivant, on va étudier la tâche d'extraction des relations entres ces entités nommées.

4.3 Reconnaissance d'entité psychologique

La reconnaissance d'entité psychologique (PsyER) comprend deux étapes principales : (i) la détection et la délimitation d'informations se référant à des entités psychologiques dans des corpus textuels (par exemple : (متلازمة توريت، الرعشة، الرهاب) et (ii) l'identification de la catégorie sémantique des entités localisées (e.g. syndrome, maladies, traitement).
Notre système de NER comporte des modules pour le prétraitement linguistique, l'iden-

tification des entités nommées, la classification et la translation. Le système est basé sur la création de corpus arabes psychologiques à partir du Web et Gazetteers des sciences psychologiques, et est basé sur les règles de JAPE pour extraire des entités psychologiques. Le composant à base de règles est construit avec la capacité de reconnaître 08 types différents d'entités nommées, y compris les troubles mentaux, la phobie, les syndromes, les substances psycho-actives, les symptômes, les organes et les traitements, dans des textes écrits en arabe. La figure 4.2 montre l'architecture détaillée pour la reconnaissance d'entités psychologiques.

FIGURE 4.2 : L'architecture de PsyNER.



4.3.1 Description des composants du système PsyNER

La figure montre les différentes étapes proposées pour l'identification et l'extraction de PsyER, nous décrivons dans ce qui suit les principaux composants du système.

A. **La normalisation**

Habituellement, ce processus utilisé avant d'appliquer les techniques d'extraction de texte afin d'éviter ou réduire l'éparpillement de données dans les données en cours de traitement. En arabe, il est possible d'écrire trouble de deux manières différentes " اضطراب " Alef sans Hamza ci-dessous ou " إضطراب" Alef Hamza ci-dessous. Par conséquent, pour rendre les données plus cohérentes, ce processus est appliqué.

Le processus de normalisation comprend les étapes suivantes :

⁕. **Remplacer certains caractères :**

- Comme les voyelles : Alef avec Hamza (أ) ,(إ) avec Hamza ci-dessous ou madda (آ), devient simplement Alef (ا).
- Le second caractère est (ؤ) qui peut être écrit(ؤ) ou (و) ; ce sera normalisé à (و) .
- Le troisième caractère est (ة) qui peut être écrit à la fin des mots comme (ه) ou (ة) ; ce sera normalisé à (ه) .
- Le quatrième caractère est (ى) qui peut être écrit à la fin des mots comme (ي) ou (ى) ; ce sera normalisé (ي) .

Cependant, ce procédé rend l'ensemble plus cohérent.

B. **Séparateur de phrase (sentence splitter)**

Comme son nom l'indique, il sépare le texte qui lui est fourni en entrée en un ensemble de phrases en fonction de la ponctuation. En effet, l'extraction d'informations s'effectue phrase par phrase. Aucune information hors de cette phrase ne peut être utilisée pendant le processus.

C. **Tokenizer**

Dont le rôle est de diviser le texte en jetons simples (ponctuations, nombres, mots, etc.), Les annotations engendrées par le tokeniser sont de type Token et SpaceToken.

4.4 Gazetteers du système

Dans le but de reconnaitre les entités nommées de type psychologique, nous avons étendu GATE avec des listes supplémentaires de Gazetteer ainsi des règles d'extraction pour aider à identifier les entités concernées tels que syndromes, maladies,. . . , nous avons crée 8 Gazetteers (monocultures).

Le rôle d'un Gazetteer est d'identifier les entités nommées dans le texte. Il utilise un ensemble de listes de noms et d'abréviations tels que les noms de villes, d'organisations, jours de semaine, prénoms … etc. Le Gazetteer va annoter toutes les entités nommées qu'il trouve dans le texte et qui existent dans ses listes avec une annotation de type Lookup.

Nous avons commencé par l'élaboration manuelle d'un lexique bilingue (AR-EN) provenant de diverses ressources psychologiques telles que les dictionnaires psychologiques [41], DSM-IV-TR [40], DSM-IV [39]. Les 8 Gazetteers construites pour l'extracteur NE sont exprimées dans les sous-sections suivantes.

4.4.1 Gazetteers pour Extracteurs de troubles mentaux désignés par DSM-IV (Diagnostic and Statistical Manual of the American Psychiatric Association)

TABLE 4.1 : Les Gazetteers préparés pour les extracteurs de Syndrome, Trouble psychologique et Phobie.

Liste	Description	Taille
Syndrome	• le niveau supérieur d'analyse. • ce terme est appliqué à une constellation de symptômes qui se produisent en même temps ou covariant dans le temps. • Le terme ne porte pas de conséquences directes en termes de pathologie sous-adjacente. • Que ce soit, en fait, certains ensembles de symptômes covariant avec l'autre est une question empirique.	374
Trouble psychologique	• comme un syndrome, se réfère à un ensemble de symptômes, • mais le concept inclut l'idée que l'ensemble des symptômes ne sont pas comptabilisés par une condition plus envahissante. • Comme symptôme et le syndrome, il n'y a aucune incidence sur l'étiologie.	78
Phobie	• Une phobie est une peur démesurée et irrationnelle d'un objet ou d'une situation précise.	184

Les Gazetteers construits pour les extracteurs Syndromes, Trouble psychologique, Phobie sont répertoriés dans le tableau 4.1 avec la description et la taille en mots pour

chaque Gazetteer. Des exemples pour le contenu de chaque Gazetteer mentionné dans le tableau 4.1 sont illustrés dans le tableau 4.2 avec la traduction en anglais.

TABLE 4.2 : Exemples d'entrées pour les Gazetteers Syndrome, Trouble psychologique, Phobie.

Nom de Gazetteer	**Exemples d'entrées**	
	NE (en arabe)	Traduction en anglais
Syndrome	متلازمة توريت متلازمة داون متلازمة كانر	Tourette Syndrome down syndrome Kanner's syndrome
Trouble psychologique	اضطراب في التوازن اضطراب تمييز الألوان اضطراب ايضي	Ataxia Dyschromatopsia Metabolic Disorder
Phobie	رهاب الابر رهاب الدم رهاب الضوء	Belonephobia Pediophobia Odontophobia

4.4.2 Gazetteers pour Extracteur de maladies

Pour la reconnaissance d'entité nommée des noms de Maladies, nous utilisons le Gazetteer Disease.

TABLE 4.3 : Le Gazetteer préparé pour l'extracteur de Maladie.

Liste	**Description**	**Taille**
Syndrome	• Un trouble dans lequel la sous-jacente étiologie est connue. • Il est le plus haut niveau de compréhension conceptuelle.	59

Le Gazetteer construit pour l'extracteur Maladie est répertorié dans le tableau 4.3 avec la description et la taille en mots. Des exemples pour le contenu de ce Gazetteer mentionné dans le tableau 4.3 est illustré dans le tableau 4.4 avec la traduction en anglais.

TABLE 4.4 : Exemples d'entrées pour Gazetteer Maladie.

Nom de Gazetteer	**Exemples d'entrées**	
	NE (en arabe)	Traduction en anglais
Maladie	مرض كروزون مرض كوشينغ مرض نيمان بيک	Crouson disease Cusching disease Niemann-Pick disease

4.4.3 Gazetteer pour Extracteur de symptôme

Pour la reconnaissance d'entité nommée des noms de symptômes, nous utilisons le Gazetteer symptom.

TABLE 4.5 : Le Gazetteer préparé pour l'extracteur symptôme.

Liste	**Description**	**Taille**
Symptôme	• se réfère à un comportement observable ou de l'État. • Il n'y a pas d'incidence qu'un problème sous-adjacent existe nécessairement ou qu'il y a une étiologie physique. • le plus simple niveau de l'analyse d'un problème de présentation.	23

Le Gazetteer construit pour l'extracteur symptôme est répertorié dans le tableau 4.5.avec la description et la taille en mots. Des exemples pour le contenu de ce Gazetteer mentionné dans le tableau 4.5 est illustré dans le tableau 4.6 avec la traduction en anglais.

TABLE 4.6 : Exemples d'entrées pour Gazetteer symptôme.

Nom de Gazetteer	**Exemples d'entrées**	
	NE (en arabe)	Traduction en anglais
Symptôme	عارض كاسبر الخلجة عارض تحولي	Casper symptom Tic Conversion symptom

4.4.4 Gazetteer pour Extracteur de substances psycho-actives

Pour la reconnaissance d'entité nommée des noms de Substances psycho-actives, nous utilisons le Gazetteer Psychoactive Substances.

TABLE 4.7 : Le Gazetteer préparé pour Substances psycho-actives.

Liste	Description	Taille
Psycho-actives	Les substances psycho-actives à risque de dépendance (alcool, tabac, drogues, etc.), agissent sur le circuit de récompense du cerveau.	71

Le Gazetteer construit pour l'extracteur Substances psycho-actives est répertoriée dans le tableau 4.7 avec la description et la taille en mots.

Des exemples pour le contenu de ce Gazetteer mentionné dans le tableau 4.7 est illustré dans le tableau 4.8 avec la traduction en anglais.

TABLE 4.8 : Exemples d'entrées pour Gazetteer Substances psycho-actives.

Nom de Gazetteer	Exemples d'entrées	
	NE (en arabe)	Traduction en anglais
Substances psychoactives	ادمان التبغ	Tabagism
	ادمان المنومات	Narcomania
	ادمان الكوكيين	Cocainomania

4.4.5 Gazetteer pour l'extracteur d'Organe

Pour la reconnaissance d'entité nommée des noms d'Organe, nous utilisons le Gazetteer Organe.

TABLE 4.9 : Le Gazetteer préparé pour l'extracteur d'Organe.

Liste	Description	Taille
Organe	• Les organes du corps humain. • Un organe est un ensemble de tissus spécifiques capable de remplir une (ou plusieurs) fonction déterminée.	375

Le Gazetteer construit pour l'extracteur Organe est répertoriée dans le tableau 4.9 avec la description et la taille en mots. Des exemples pour le contenu de ce Gazetteer mentionné dans le tableau 4.9 est illustré dans le tableau 4.10 avec la traduction en anglais.

TABLE 4.10 : Exemples d'entrées pour Gazetteer Organe.

Nom de Gazetteer	Exemples d'entrées	
	NE (en arabe)	Traduction en anglais
Organe	قلب مخ كتف	Heart brain shoulder

4.4.6 Gazetteer pour Extracteur de Traitement

Pour la reconnaissance d'entité nommée des noms de Traitement, nous utilisons le Gazetteer Traitement.

TABLE 4.11 : Le Gazetteer préparé pour extracteur Traitement.

Liste	Description	Taille
Traitements	Il existe plusieurs types de traitements et ceux-ci varient grandement selon les troubles	31

Le Gazetteer construit pour l'extracteur Traitement est répertorié dans le tableau 4.11 avec la description et la taille en mots. Des exemples pour le contenu de ce Gazetteer mentionné dans le tableau 4.11 est illustré dans le tableau 4.12 avec la traduction en anglais.

TABLE 4.12 : Exemples d'entrées pour Gazetteer Traitements

Nom de Gazetteer	**Exemples d'entrées**	
	NE (en arabe)	Traduction en anglais
Traitements	بنزو ديازيبين كلونيدين سيكلوسبورين	Benzodiazepine Clonidine Cyclosporine

Ces Gazetteers sont repartis en cinq catégories illustrées dans le tableau 4.13. Ces catégories seront les classes principales pour notre ontologie psychologique.

TABLE 4.13 : Les classes principales.

Catégories	Description
1	Inclus la classification des troubles mentaux dans les 16 chapitres du groupe système DSM-IV
2	Symptôme
3	Organe
4	Substances psycho-actives
5	Traitement

4.5 La mise en œuvre des extracteurs NE

La phase d'extraction d'entités nommées consiste à mettre en place un système de détection et de catégorisation des entités d'intérêt dans un texte. Notre objectif, ici, se limite à extraire les 08 types des entités citées dans la sous-section 2.1, dans des textes écrits en arabe. Nous détaillons par la suite notre façon de procéder pour réaliser cet objectif.

4.5.1 Les ressources utilisées dans GATE

Dans notre contribution, GATE est adoptée comme une plateforme pour mettre en œuvre notre système NER pour les sciences psychologiques.

Afin de construire un système NERA à l'aide de GATE, un ensemble de ressources de traitement peut être utilisé comme des composants du système, notamment tokenizer pour la langue arabe, transducteurs JAPE pour des règles grammaticales et les Gazetteers.

Les ressources de traitement de la langue arabe que nous avons utilisé sont :

A) **Arabic Gazetteer :** les listes Gazetteers sont des fichiers plain texte avec une entrée par ligne, chaque liste représente un ensemble de noms tels que les noms des maladies, des symptômes, des organes etc. Ce type de nomenclature est construit manuellement.

B) **Arabic inferred-Gazetteer :** Cette Gazetteer est déduite automatiquement.

C) **Arabic grammar :** nous permet d'utiliser des fichiers contenant des diverses règles.

D) **Arabic tokeniser :** Segmente le texte arabe en jetons simples tels que les nombres, la ponctuation et mots de différents types.

4.6 Extracteurs d'entités nommées basées sur des règles

Dans cette thèse, l'implémentation de toutes les règles grammaticales est faite par le langage JAPE afin de permettre la reconnaissance des huit types d'entités nommées mentionnées précédemment.

4.6.1 JAPE (Java Annotation Pattern Engine)

JAPE est un langage dérivé de CPSL (Common Pattern Specification Language) [112], il fournit des transducteurs à états finis basés sur des expressions régulières. La grammaire JAPE consiste en un ensemble de phases, chacune d'elle est un ensemble de règles. Les phases sont exécutées séquentiellement constituant une cascade de transducteurs à états finis pour les annotations. Ces règles sont divisées en deux blocs : une partie gauche (‘‘ Left Hand Side‘‘ ou LHS) définissant un motif d'annotations à repérer et une partie droite (‘‘ Right Hand Side‘‘ ou RHS) contenant les opérations à effectuer sur ce motif.

Donc la partie gauche (LHS) de la règle contient le patron de l'annotation pouvant contenir des opérateurs d'expressions régulières (*, ?, +). La partie droite de la règle (RHS) donne le label de l'annotation, attachée au patron : l'action à entreprendre si le patron est détecté [113].

Une description de chaque extracteur d'entité nommée est donnée dans les sous-sections suivantes.

4.6.2 Extracteur entité nommée Syndrome

La règle de l'extracteur NE Syndrome a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Synd. La règle indique que la séquence de mots dans le texte ciblé est annotée en tant que syndrome, si les mots sont trouvés dans le Gazetteer des noms pour syndrome. Un exemple d'implémentation de règle dans l'extracteur NE Synd est donné ci-dessous.

La règle Synd est sous forme d'expression régulière :

((ال | ل | ك | ب | ف)?+Syndrome)

Description

L'utilisation de la règle Synd nous permet d'extraire les entités nommées syndromes et d'identifier les tokens qui contiennent ces entités comme suit :

- Chaque mot peut être constitué d'un ou plusieurs préfixes, une racine et un ou plusieurs suffixes dans différentes combinaisons, ayant pour résultat une morphologie très systématique, mais compliquée. Clitiques, qui dans d'autres langues telles que l'anglais seraient traités comme des mots séparés, agglutinent aux mots ;
- L'arabe a un ensemble de clitiques qui sont attachés à un NE, y compris les conjonctions telles que و (Waw, and) et ف(fa, then) et prépositions telles que ل (Ly, for/to), ك (ka, as) et ب (by, by /with) et l'article défini ال (al, the) ;
- Si l'expression commence par un clitique, puis suivie d'un syndrome (c.-à-d appartient au répertoire toponymique des syndromes) avec, l'expression est identifiée comme syndrome nommé entité. Parmi les caractéristiques de cette règle, elle nous donne la translation automatique de chaque entité nommée trouvée.

Exemples d'expression de syndrome correspondant à la Synd :

```
Rule :Synd
(
({Token within Lookup.majorType==Syndrome}|
{Lookup.majorType==Syndrome})
) :tag
–>
{
try {

if(s.contains("بمتلازمة")||s.contains("بتناذر")||s.contains("فمتلازمة")||
s.contains("فتناذر") ||s.contains("كتناذر")|| s.contains("كمتلازمة")||
s.contains("لتناذر")|| s.contains("لمتلازمة")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", "Syndrome");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+1,c) );
outputAS.add(offset+1, endOffset, "Syndrome", features);
}else {if(s.contains("المتلازمة")||s.contains("التناذر")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", "Syndrome");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+2,c) );
outputAS.add(offset+2, endOffset, "Syndrome", features);
}else {
FeatureMap features = Factory.newFeatureMap();
Object obj = features.get("tr");
features.put("rule", "Synd");
features.put("kind", "Syndrome");
String typeDoc=(String) doc.getFeatures().get("tr");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b,c) );
outputAS.add(offset, endOffset, "Syndrome", features);
}
}} catch (InvalidOffsetException e) {
throw new LuckyException(e);
}}
```

4.6.3 Extracteur entité nommée Trouble psychologique

La règle de l'extracteur NE Trouble psychologique a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Disorders. La règle indique que la séquence de mots dans le texte ciblé est annotée en tant que Trouble psychologique, si les mots sont trouvés dans le Gazetteer des Troubles psychologiques.

La règle Disorders sous forme d'expression régulière :

((ال | ل | ك | ب | ف)?+Disorders)

La règle Disorders dans le langage JAPE (implémentée dans GATE) :

```
Rule : Disorders ( ({Token within Lookup.majorType== Disorders }|
{Lookup.majorType== Disorders })
) :tag
–>
{
try { if(s.contains("باضطراب")||s.contains("فاضطراب")||
s.contains("كاضطراب")|| s.contains("لاضطراب")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", " Disorders ");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+1,c) );
outputAS.add(offset+1, endOffset, " Disorders ", features);
}else {if(s.contains("الاضطراب")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", " Disorders ");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+2,c) );
outputAS.add(offset+2, endOffset, " Disorders ", features);
}else{
FeatureMap features = Factory.newFeatureMap();
Object obj = features.get("tr");
features.put("rule", " Disorders ");
features.put("kind", " Disorders ");
String typeDoc=(String) doc.getFeatures().get("tr");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b,c) );
outputAS.add(offset, endOffset, " Disorders ", features);
} }}} catch (InvalidOffsetException e) {throw new LuckyException(e); } }
```

4.6.4 Extracteur entité nommée Phobie

La règle de l'extracteur NE phobie a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Phobia. La règle

indique que la séquence de mots dans le texte ciblé est annotée en tant que Phobia, si les mots sont trouvés dans le Gazetteer Phobia.
La règle Phobia sous forme d'expression régulière :

((ال | ل | ك | ب | ف)?+Phobia)

La règle Phobia dans le langage JAPE (implémentée dans GATE) :

```
Rule :Phobia
(
({Token within Lookup.majorType==Phobia}|
{Lookup.majorType== Phobia})
) :tag
–>
{
try {
if(s.contains("برهاب")||s.contains("فرهاب")||s.contains("كرهاب")||
s.contains("لرهاب")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", "Phobia");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+1,c) );
outputAS.add(offset+1, endOffset, " Phobia ", features);
}else {if(s.contains("الرهاب")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", " Phobia ");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+2,c) );
outputAS.add(offset+2, endOffset, " Phobia ", features);
}else{
FeatureMap features = Factory.newFeatureMap();
Object obj = features.get("tr");
features.put("rule", " Phobia ");
features.put("kind", " Phobia ");
String typeDoc=(String) doc.getFeatures().get("tr");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b,c) );
outputAS.add(offset, endOffset, " Phobia ", features);

} }}} catch (InvalidOffsetException e) {throw new LuckyException(e); }
}
```

4.6.5 Extracteur entité nommée Maladie

La règle de l'extracteur NE Maladie a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Maladie. La règle indique que la séquence de mots dans le texte cible est annotée en tant que Disease, si les mots sont trouvés dans le Gazetteer Disease.
La règle Disease sous forme d'expression régulière :

((ال | ل | ك | ب | ف)?+Disease)

La règle Disease dans le langage JAPE (implémentée dans GATE) :

```
Rule : Disease
(
({Token within Lookup.majorType== Disease }|
{Lookup.majorType== Disease})
) :tag
–>
{
try {
if(s.contains("بمرض")||s.contains("فمرض")||s.contains("كمرض")||
s.contains("لمرض")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", " Disease ");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+1,c) );
outputAS.add(offset+1, endOffset, " Phobia ", features);
}else {if(s.contains("المرض")){
FeatureMap features = Factory.newFeatureMap();
features.put("rule", " Disease ");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b+2,c) );
outputAS.add(offset+2, endOffset, " Disease ", features);
}else{
FeatureMap features = Factory.newFeatureMap();
Object obj = features.get("tr");
features.put("rule", " Disease ");
features.put("kind", " Disease ");
String typeDoc=(String) doc.getFeatures().get("tr");
features.put("Tra",annotation.getFeatures().get("tr"));
features.put("string",mydoc.substring(b,c) );
outputAS.add(offset, endOffset, " Disease ", features);

} }}} catch (InvalidOffsetException e) I{throw new LuckyException(e); }
}
```

4.6.6 Extracteur entité nommée Symptôme

La règle de l'extracteur NE Symptôme a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Symptom. La règle indique que la séquence de mots dans le texte ciblé est annotée en tant que Symptôme, si les mots sont trouvés dans le Gazetteer Symptom.
La règle Symptom sous forme d'expression régulière :

$(Token \in Symptom \mid \text{Symptom})$

La règle Symptom dans le langage JAPE (implémentée dans GATE) :

```
Rule : Symptom
(
({Token within Lookup.majorType== Symptom } |
{Lookup.majorType== Symptom })
) :tag
–>
{AnnotationSet anno=bindings.get("tag") ;
Annotation annotation = anno.iterator().next() ;
long offset = anno.firstNode().getOffset() ;
long endOffset = anno.lastNode().getOffset() ;
try {
if(anno !=null && anno.size()>0){
int b=anno.firstNode().getOffset().intValue() ;
int c=anno.lastNode().getOffset().intValue() ;
String mydoc=doc.getContent().toString() ;
String s=mydoc.substring(b,c) ;
FeatureMap features = Factory.newFeatureMap() ;
Object obj = features.get("tr") ;
features.put("rule", " Symptom ") ;
features.put("kind", " Symptom ") ;
String typeDoc=(String) doc.getFeatures().get("tr") ;
features.put("Tra",annotation.getFeatures().get("tr")) ;
features.put("string",mydoc.substring(b,c) ) ;
outputAS.add(offset, endOffset, " Symptom ", features) ;
}catch (InvalidOffsetException e) {throw new LuckyException(e) ; } }
```

4.6.7 Extracteur entité nommée Substances psycho-actives

La règle de l'extracteur NE Substance psycho-active a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Substance_related_disorders. La règle indique que la séquence de mots dans le texte cible est annotée en tant que Substance psycho-active, si les mots sont trouvés dans le Gazetteer Substance_related_disorders.

La règle Substance_related_disorders sous forme d'expression régulière :

$(Token \in Substance\_related\_disorders|Substance\_related\_disorders\ )$

La règle Substance_related_disorders dans le langage JAPE (implémentée dans GATE) :

```
Rule : Substance_related_disorders
(
({Token within Lookup.majorType== Substance_related_disorders } |
{Lookup.majorType== Substance_related_disorders })
) :tag
–>
{AnnotationSet anno=bindings.get("tag") ;
Annotation annotation = anno.iterator().next() ;
long offset = anno.firstNode().getOffset() ;
long endOffset = anno.lastNode().getOffset() ;
try {
if(anno !=null && anno.size()>0){
int b=anno.firstNode().getOffset().intValue() ;
int c=anno.lastNode().getOffset().intValue() ;
String mydoc=doc.getContent().toString() ;
String s=mydoc.substring(b,c) ;
FeatureMap features = Factory.newFeatureMap() ;
Object obj = features.get("tr") ;
features.put("rule", " Substance_related_disorders ") ;
features.put("kind", " Substance_related_disorders ") ;
String typeDoc=(String) doc.getFeatures().get("tr") ;
features.put("Tra",annotation.getFeatures().get("tr")) ;
features.put("string",mydoc.substring(b,c) ) ;
outputAS.add(offset, endOffset, " Substance_related_disorders ", features) ;
}catch (InvalidOffsetException e) {throw new LuckyException(e) ; } }
```

4.6.8 Extracteur entité nommée Organe

La règle de l'extracteur NE Organe a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Organ. La règle indique que la séquence de mots dans le texte cible est annotée en tant que Organe, si

les mots sont trouvés dans le Gazetteer Organ.

La règle Organ sous forme d'expression régulière :

$((Token \in Organ|Organ))$

La règle Organ dans le langage JAPE (implémentée dans GATE) :

```
Rule : Organ
(
({Token within Lookup.majorType== Organ } |
{Lookup.majorType== Organ })
) :tag
–>
{AnnotationSet anno=bindings.get("tag") ;
Annotation annotation = anno.iterator().next() ;
long offset = anno.firstNode().getOffset() ;
long endOffset = anno.lastNode().getOffset() ;
try {
if(anno !=null && anno.size()>0){
int b=anno.firstNode().getOffset().intValue() ;
int c=anno.lastNode().getOffset().intValue() ;
String mydoc=doc.getContent().toString() ;
String s=mydoc.substring(b,c) ;
FeatureMap features = Factory.newFeatureMap() ;
Object obj = features.get("tr") ;
features.put("rule", " Organ ") ;
features.put("kind", " Organ ") ;
String typeDoc=(String) doc.getFeatures().get("tr") ;
features.put("Tra",annotation.getFeatures().get("tr")) ;
features.put("string",mydoc.substring(b,c) ) ;
outputAS.add(offset, endOffset, " Organ ", features) ;
}catch (InvalidOffsetException e) {throw new LuckyException(e) ; } }
```

4.6.9 Extracteur entité nommée Traitement

La règle de l'extracteur NE Traitement a été implémentée dans le langage JAPE sur notre analyse du texte psychologique arabe qui a conduit à extraire la règle Traitment. La règle indique que la séquence de mots dans le texte ciblé est annotée en tant que Traitements, si les mots sont trouvés dans le Gazetteer Traitment.
La règle Traitment sous forme d'expression régulière :

$$((Token \in Traitment|Traitment))$$

La règle Traitment dans le langage JAPE (implémentée dans GATE) :

```
Rule : Traitment
(
({Token within Lookup.majorType== Traitment } |
{Lookup.majorType== Traitment })
) :tag
–>
{AnnotationSet anno=bindings.get("tag") ;
Annotation annotation = anno.iterator().next() ;
long offset = anno.firstNode().getOffset() ;
long endOffset = anno.lastNode().getOffset() ;
try {
if(anno !=null && anno.size()>0){
int b=anno.firstNode().getOffset().intValue() ;
int c=anno.lastNode().getOffset().intValue() ;
String mydoc=doc.getContent().toString() ;
String s=mydoc.substring(b,c) ;
FeatureMap features = Factory.newFeatureMap() ;
Object obj = features.get("tr") ;
features.put("rule", " Traitment ") ;
features.put("kind", " Traitment ") ;
String typeDoc=(String) doc.getFeatures().get("tr") ;
features.put("Tra",annotation.getFeatures().get("tr")) ;
features.put("string",mydoc.substring(b,c) ) ;
outputAS.add(offset, endOffset, " Traitment ", features) ;
} catch (InvalidOffsetException e) {throw new LuckyException(e) ; } }
```

4.7 La translation automatique des entités nommées

La translation des entités nommées (NE), telles que les maladies, les syndromes et les troubles mentaux est très importante pour plusieurs applications de traitement du langage naturel. Elle joue un rôle essentiel dans des applications telles que la recherche d'informations multilingues et la traduction automatique. Dans cette approche, nous introduisons dans chaque extracteur un module pour la translation d'entités nommées identifiées. Le module contient des instructions pour exploiter les Gazetteers bilingue (AR-FR) développé dans la section 3 afin de générer la translation d'entités nommées identifiées de la langue Arabe vers la langue anglaise.

Un exemple du module intégré dans la règle Syndrome

```
Object obj = features.get("tr") ;
features.put("rule", " Syndrome ") ;
features.put("kind", " Syndrome ") ;
String typeDoc=(String) doc.getFeatures().get("tr") ;
features.put("Tra",annotation.getFeatures().get("tr")) ;
```

4.8 Conclusion

Ce travail de recherche consiste en la reconnaissance des entités psychologiques. Deux techniques ont été appliquées pour les processus de reconnaissance des PsyNR : la première technique dépend entièrement de l'identification directe avec l'utilisation des divers ensembles de Gazetteers, y compris les Gazetteers de syndrome, des troubles mentaux, de phobie, des substances psycho-actives, de symptôme, d'organe et de traitement et la deuxième technique est un module basé sur des règles construites sur la base de ces Gazetteers.

Chapitre 5

Acquisition des relations

Sommaire

5.1 Introduction

L'extraction de relation est un sujet de recherche de longue date dans le traitement du langage naturel, et a été utilisé pour aider, entre autres, l'acquisition des connaissances, l'extraction d'information, et les systèmes de questions-réponses . Il a également reçu beaucoup d'attention dans les domaines médical et biomédical.

Les travaux menés dans le cadre d'identification des relations ont montré des résultats à la fois très encourageants mais qui sont toujours à améliorer. La problématique d'identification des relations ne cesse d'évoluer avec les nouveaux champs d'application qui s'imposent, ce qui augmente les contraintes liées à cette tâche.

Dans ce chapitre, nous nous intéressons à l'extraction des relations entre les entités nommées. Cette tâche consiste à représenter l'interaction entre les entités nommées. Pour cette raison, nous utilisons la Programmation Linéaire en nombre entier (Integer Linear Programming - ILP) pour extraire des relations explicites et implicites. Autant que nous sachions, l'extraction d'informations psychologiques en langue arabe n'a toujours pas été tentée par aucun chercheur. Ce chapitre décrit et discute en détails la formalisation des contraintes pour l'extraction des relations.

5.2 La Programmation linéaire en nombre entier dans le traitement du langage naturel

Les ILP sont des problèmes d'optimisation contraints où la fonction objective et les contraintes sont des équations linéaires avec des variables entières. Au cours des dernières années, en commençant par quelques articles écrits par Roth et Yih [86] [114].
des douzaines d'articles utilisaient la formulation de la programmation linéaire en nombre entier (ILP), y compris plusieurs articles primés [87] [88] [89]. Les techniques ILP ont été appliquées à plusieurs tâches TAL, notamment, l'extraction de relations[115], l'étiquetage de rôle sémantique [114], l'analyse syntaxique [116]. Récemment appliquées en, résumé abstractif [90], extraction d'information [91], Systèmes de questions-réponses[92].
Un programme linéaire en nombres entiers se présente sous la forme générique suivante :

$$\max \sum_{i\in\mathcal{E},k\in\mathcal{l}_{\mathcal{E}}} c_{i,k}\cdot x_{i,k} + \sum_{i\in\mathcal{R},k\in\mathcal{l}_{\mathcal{R}}} c_{i,k}\cdot x_{i,k} + \sum_{r\in\mathcal{R},e\in\mathcal{E},u\in\mathcal{L}_{\mathcal{R}},v\in\mathcal{l}_{\mathcal{E}}} c_{r,u,e,v}\cdot x_{r,u,e,v} \tag{5.1}$$

Sujet à

$$\mathcal{X}_{E,e}\in\{0,1\} \qquad \forall E\in\mathcal{E}, e\in\mathcal{l}_{\mathcal{E}} \tag{5.2}$$

$$\mathcal{X}_{R,r}\in\{0,1\} \qquad \forall R\in\mathcal{R}, r\in\mathcal{l}_{\mathcal{R}} \tag{5.3}$$

$$\mathcal{X}_{R,r,E,e}\in\{0,1\} \qquad \forall R\in\mathcal{R}\ r\in\mathcal{l}_{\mathcal{R}}, E\in\mathcal{E}, e\in\mathcal{l}_{\mathcal{E}} \tag{5.4}$$

$$\sum_{e\in\mathcal{L}_{\mathcal{E}}} \mathcal{X}_{\{E,e\}} = 1 \quad \forall E\in\mathcal{E} \tag{5.5}$$

$$\sum_{r\in\mathcal{L}_{\mathcal{R}}} \mathcal{X}_{\{R,r\}} = 1 \quad \forall R\in\mathcal{R} \tag{5.6}$$

$$\mathcal{X}_{E,e} = \sum_{r\in\mathcal{L}_{\mathcal{R}}} \mathcal{X}_{R,r,E,e} \quad \forall E\in\mathcal{E} \quad et \quad R\in\mathcal{N}(\mathcal{E}) \tag{5.7}$$

$$\mathcal{X}_{R,r} = \sum_{r \in \mathcal{L}_{\mathcal{R}}} \mathcal{X}_{R,r,E,e} \quad \forall R \in \mathcal{R} \quad et \quad E \in \mathcal{N}(\mathcal{R}) \tag{5.8}$$

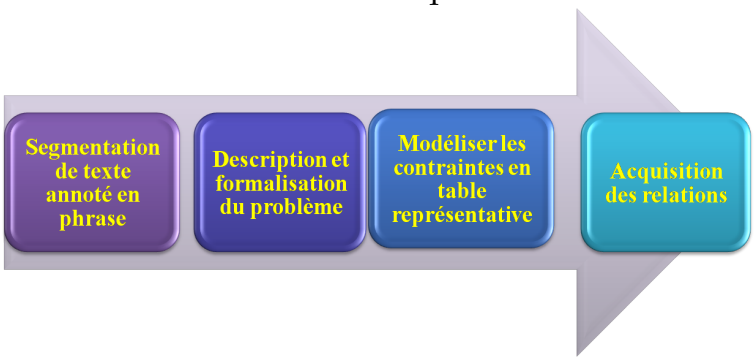
5.3 Identification des Relations

Les relations sont des objets où deux entités nommées ou plus sont liées selon une catégorie sémantique précise, comme la relation Has_Symptom entre une entité de type trouble mentale (MD) et une entité de type Symptom .

Dans le cadre de cette thèse, nous nous sommes intéressées à l'extraction de relations binaires dans un contexte psychologique. La tâche d'extraction de relations est posée comme une tâche de classification de couples d'entités, supposés connus : étant donné un couple d'entités (e1 ; e2), il s'agit de déterminer si ces entités sont reliées par l'une des relations données, et laquelle le cas échéant (les cas de non-relation peuvent correspondre à deux entités qui ne sont pas en relation, ou qui sont reliées par une autre relation que celles de la classification donnée).

Dans cet axe nous avons développé une approche pour l'acquisition des relations sémantique en utilisant la programmation par contraintes pour l'apprentissage non supervisé et pour le traitement automatique des langues. Récemment l'intérêt des approches déclaratives (programmation linéaire en nombres entiers, programmation par contraintes) pour des problèmes d'apprentissage et de traitement automatique des langues a été montré. Dans ce cadre, nous avons choisi d'utiliser une approche de programmation linéaire en nombres entiers (ILP) pour effectuer cette acquisition.

FIGURE 5.1 : Processus d'acquisition de relations.



.

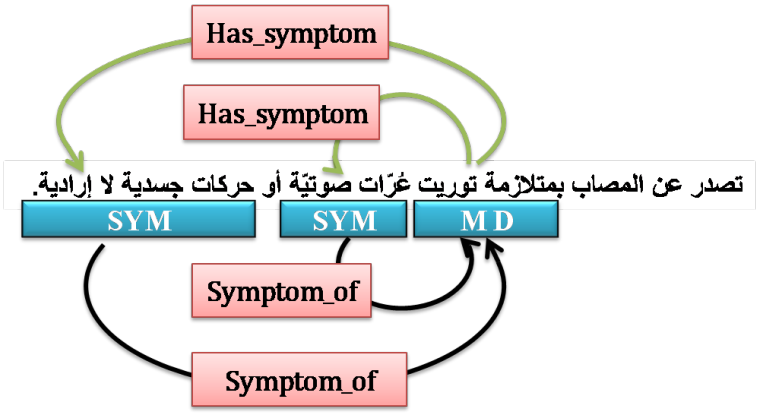
5.3.1 Segmentation de texte annoté en phrase

L'extraction des relations entre les entités nommée se fait phrase par phrase

5.3.2 Description et formalisation du problème

La méthode d'identification des relations utilisée dans cette thèse est la Programmation Linéaire en nombre entier (Integer Linear Programming -ILP), similaire à l'approche de[115]. En ILP, nous créons une variable indicatrice pour chaque affectation possible de chaque entité et relation dans une phrase.
Une phrase **S** est une liste chaînée constituée de mots **W** et d'entités **E** . Une entité peut être un mot unique ou un ensemble de mots consécutifs avec une limite prédéfinie. Les entités dans une phrase sont étiquetées comme $\mathcal{E} = \{E1, E2, ..., En\}$ selon leur ordre, et elles prennent des valeurs (c'est-à-dire, des étiquettes) qui s'étendent sur un ensemble de types d'entités. La valeur assignée à $E_i \in \mathcal{E}$ est notée $fEi \in \mathcal{L}_{\mathcal{E}}$

FIGURE 5.2 : Exemple d'entité et de relation. Les entités Troubles mentaux (MD) et Symptômes (SYM) sont connectées par les relations Has_symptom et Symptom_of.



Soit $\mathcal{E}$ l'ensemble des entités dans une phrase et soit $\mathcal{L}_{\mathcal{E}}$ la liste des étiquettes que les entités peuvent prendre. De même, soit $\mathcal{R}$ l'ensemble des relations dans une phrase et soit $\mathcal{L}_{\mathcal{R}}$ la liste des étiquettes possibles pour les relations.

Exemple :

La phrase de la figure 5.2 a trois entités et quatre relations : $\mathcal{E} = E1, E2, E3$ et $\mathcal{R} = R12, R13, R21, R31$, $\mathcal{L}_{\mathcal{E}} = MD, SYM$ et $\mathcal{L}_{\mathcal{R}} = Symptom\_of, Has\_symptom$. Pour les entités de la figure 5.2, E1 appartient à MD et E2 et E3 appartiennent à SYM. De plus, la

relation R_{12} et R_{13} sont Has_symptom, R_{21} et R_{31} sont Symptom_of.

$(R_{12} = Has\_symptom) \rightarrow (e_1 = MD) \bigwedge (e_2 = SYM)$

Avec ces définitions, le problème d'optimisation est défini avec ILP comme suit :

$$\mathcal{X}_{rij,lr} = \mathcal{X}_{ei,le} \bigwedge \mathcal{X}_{ej,le} \qquad \forall ei, ej \in \mathcal{L} \quad et \quad i \neq j \tag{5.9}$$

$$\mathcal{X}_{rji,lr} = \mathcal{X}_{ej,le} \bigwedge \mathcal{X}_{ei,le} \qquad \forall ei, ej \in \mathcal{L} \quad et \quad i \neq j \tag{5.10}$$

$$\sum_{le \in \mathcal{L}_E} \mathcal{X}_{ej,le} = 1 \quad \forall e \in \mathcal{L}_E \tag{5.11}$$

$$\sum_{lr \in \mathcal{L}_R} \mathcal{X}_{r,lr} = 1 \quad \forall r \in \mathcal{L}_R \tag{5.12}$$

$$\mathcal{X}_{e,le} \in \{0, 1\} \tag{5.13}$$

$$\mathcal{X}_{r,lr} \in \{0, 1\} \tag{5.14}$$

Les équations (5.11) et (5.12) exigent que chaque entité ou variable de relation doit appartenir à une et une seule étiquette. Les équations (5.13) et (5.14) sont les contraintes intégrales sur ces variables binaires. L'équation (5.9) garantit que l'affectation à chaque variable d'entité ou de relation est cohérente avec l'affectation à ses variables voisines. Si nous essayons d'étiqueter une relation has_symptom, nous devons assurer que la première entité dans cette relation est un MD et la seconde un SYM. Dans cette définition, la relation R est dirigée de ei vers ej.
Pour déduire la relation inverse R^{-1} entre ei et ej, nous avons ajouté la contrainte (5.14), cette contrainte supplémentaire peut être facilement incorporée dans notre système d'inférence.

En fin, nous avons formulé le problème d'extraction des relations entre les entités nommées psychologiques à l'aide de programmation linéaire entier pour chaque phrase en utilisant les résultats de la phase précédente (identification des entités).

5.3.3 Table de représentation des contraintes

En général, l'identification des relations consistent à trouver dans un premier temps les paires ou les couples de termes qui forment les arguments d'une relation (E1 et E2), et dans un deuxième temps l'identification de l'étiquette pour la relation sémantique qui relie les termes arguments de la relation. La table 5.1 illustre les entités nommées et les

relations entre ces dernières.

TABLE 5.1 : Les relations entre les entités nommées.

Relation		**Entre les entités (E1,E2)**	
Arabe	Anglais	Arabe	Anglais
عرض_له	Symptom_of	(عرض،اضطراب عقلي)	(Symptom, Mental Disorder)
له_عرض	Has_Symptom	(اضطراب عقلي ، عرض)	(Mental Disorder, Symptom)
له_علاج	Has_cure	(اضطراب عقلي، علاج)	(Mental Disorder, Treatment)
يعالج	cures	(علاج ،اضطراب عقلي)	(Treatment, Mental Disorder)
الناجم_عن	Due_to	(اضطراب عقلي، مادة ذات تأثير نفساني)	(Mental Disorder, Psychoactive substance)
يؤدي_إلى	Induced	(مادة ذات تأثير نفساني،اضطراب عقلي)	(Psychoactive substance , Mental Disorder)
عبارة_عن	Is_a	(اضطراب عقلي، مرض)	(Mental Disorder, Disease)
الأسباب	causes	(اضطراب عقلي،اضطراب عقلي) (مرض، مرض)	(Mental Disorder, Mental Disorder) (Disease , Disease)

Dans cette section, nous décrivons d'abord notre modèle de la table représentative. Sur la base de table 5.1, nous proposons une table d'entités et de relations qui représente conjointement des entités et des relations dans une phrase S. dénote $\perp$ une relation non définie. La table est *nxn* éléments, où *n* est le nombre de classes (type d'entités nommées).
La première colonne et la première ligne contiennent les classes, une classe peut avoir plusieurs entités. La cellule de l'intersection de cette colonne avec cette ligne contient la relation qui liés ces entités.

5.3.4 Acquisition des relations

Dans la littérature, il n'existe pas d'approche d'extraction de relation pour l'arabe standard moderne dans le domaine médical qui soutient la comparaison de nos résultats

avec d'autres approches. Les systèmes actuels d'extraction de Relation Arabe se concentrent sur la détection et la classification des relations entre différents ensembles d'entités tels que le nom des personnes, organisations et lieux. Notre approche permet de capturer des relations implicites telles que les relations détectées via une relation inverse R^{-1}. Nous pouvons donc extraire de nombreuses relations à partir de la phrase de la figure 5.2 :

تصدر عن المصاب بمتلازمة توريت عرات صوتية أو حركات جسدية لا إرادية .
" Le patient atteint du syndrome de Tourette produit des tics acoustiques ou des mouvements physiques involontaires."
" Issued by Tourette's syndrome patient sound or involuntary movements ".
Nous pouvons extraire Has_symptom (**متلازمة توريت** / syndrome de Tourette, **عرات صوتية** / tics acoustiques) et Has_symptom (**متلازمة توريت** / syndrome de Tourette, **حركات جسدية لا إرادية** / mouvements physiques involontaires)). Symptom_of (**عرات صوتية** / tics acoustiques,**متلازمة توريت** / syndrome de Tourette) et Symptom_of (**حركات جسدية لا إرادية** / mouvements physiques involontaires,**متلازمة توريت** / syndrome de Tourette).

5.4 Conclusion

Ce chapitre aborde le problème d'acquisition des relations sémantiques entre les ENs psychologiques en utilisant les contraints. Ce nouveau cadre modélise l'identification de relation comme un problème d'optimisation formulé comme un Programme Linéaire en Nombres Entiers (PLNE). La formulation proposée vise à trouver les relations implicites et explicites entre ces entités.

Chapitre 6

Analyse expérimentale

Sommaire

6.1 Introduction

Dans ce chapitre, nous présentons les métriques d'évaluation et les résultats obtenus par nos approches décrites dans le chapitre précédent. Pour évaluer et ajuster les propriétés de notre système, nous devons exploiter un corpus psychologique arabe. Vue qu'il n'existe pas de corpus psychologique arabe, nous avons collecté plusieurs textes à partir le Web. Les étapes suivantes ont été utilisées pour construire : Corpus, ressources de données, prétraitement de textes arabes psychologiques.

6.2 Ressources de données

La construction du corpus provenait de plusieurs textes du Web. Dans cette recherche, nous avons choisi les pages Web de WBTEB[120], TBEEB[121].

Prétraitement de données : Parmi les étapes les plus importantes durant le développement d'une application de TALN, il y a l'étape de prétraitement de données. L'objectif de cette étape est de traiter les données avec des processus unifiés et non une multitude de processus adaptés à tous les cas possibles. Pour obtenir du texte pur, nous avons supprimé les textes, images et signes indésirables ... et ainsi de suite.

6.3 Métriques de performance

Il existe plusieurs métriques de performance telles que : l'efficacité et l'évolutivité, et de nombreuses mesures de classification telles que : l'exactitude, la précision, le rappel et la F-mesure. Les métriques de performance sont une mesure de la performance du système. Ils seront utilisés plus tard pour évaluer l'efficacité de notre approche proposée.

6.4 Matrice de confusion

La matrice de confusion [122] est l'un des outils populaires pour évaluer la performance d'un système dans des tâches de classification ou de prédiction. La matrice de confusion est représentée par une matrice avec chaque ligne représentant les instances dans une classe prédite, tandis que chaque colonne représente dans une classe réelle comme indiqué dans le tableau 6.1.

TABLE 6.1 : Matrice de confusion.

		True Class	
		Positive	**Negative**
Predicted class	Positive	True Positive (TP)	False Positive (FP)
	Negative	Faux Négatif (FN)	True Negative (TN)

- ⋆ **True Positive (TP) :** fait référence au nombre d'instances positives correctement étiquetées par le classificateur.

- ⋆ **True Negative (TN) :** fait référence au nombre d'instances négatives correctement étiquetées par le classificateur.
- ⋆ **False Positive (FP) :** fait référence au nombre d'instances positives incorrectement étiquetées par le classificateur.
- ⋆ **Faux Négatif (FN) :** fait référence au nombre d'instances négatives incorrectement étiquetées par le classificateur.

6.4.1 La Précision et Le Rappel

Les résultats de nos extracteurs sont exprimés en précision et rappel (équations 6.1 et 6.2). La précision va nous permettre de savoir si ce que nous extrayons est correct. Le rappel va nous donner une indication sur le fait que nous extrayons suffisamment d'éléments ou pas. La précision, le rappel et F-mesure, peuvent être extraites de la matrice de confusion.
Les mesures IE standards sont calculées comme suit :

Précision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{6.1}$$

Rappel

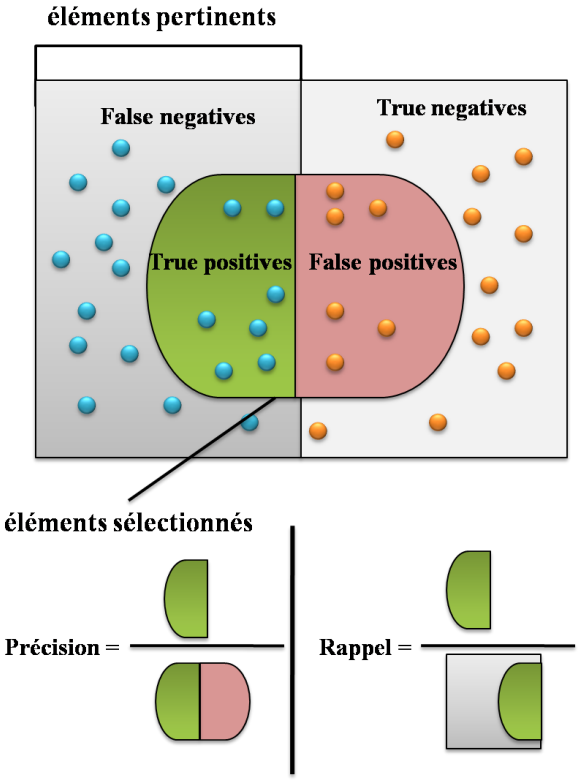
$$Rappel = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{6.2}$$

6.4.2 F-mesure

Les deux mesures de performance précédente sont combinées pour former la mesure de la performance, F-mesure, qui est calculée par la moyenne harmonique pondérée de précision et de rappel.

$$F - mesure = \frac{2XPrecisionXRappel}{Precision + Rappel} \tag{6.3}$$

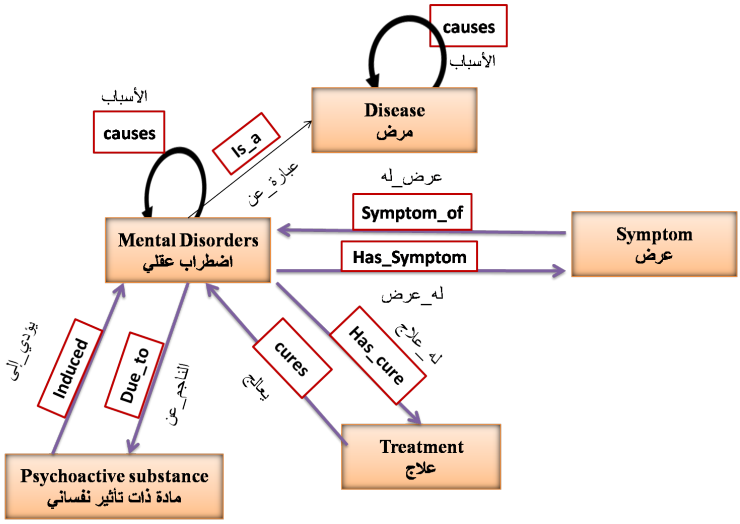
FIGURE 6.1 : Une visualisation de précision et de rappel.



6.5 Identification des relations

Dans le reste de cette section, nous analysons chacune des catégories de relations définies dans la figure 6.2.

FIGURE 6.2 : Le schéma de relation.



L'extraction de relation se concentre sur les relations possibles entre un couple de NE ("الأعراض" (Symptômes), "الأمراض" (Maladies), "الاضطرابات العقلية" (Trouble mental), "المواد المسببة للاضطرابات" (Substances psychoactives) et "العلاج" (Traitements)).

Trouble mental (DM) - Traitements (Trait)

$$X_{rij,Has\_cure} = X_{ei,DM} \bigwedge X_{ej,Trait} \tag{6.4}$$

Traitements - Trouble mental

$$X_{rij,Cures} = X_{ei,Trait} \bigwedge X_{ej,DM} \tag{6.5}$$

Trouble mental – symptômes

$$X_{rij,Has\_symptom} = X_{ei,DM} \bigwedge X_{ej,SYM} \tag{6.6}$$

Symptômes - Trouble mental

$$X_{rij,Symptom\_of} = X_{ei,SYM} \bigwedge X_{ej,DM} \tag{6.7}$$

Trouble mental – Maladies

$$X_{rij,is\_a} = X_{ei,DM} \bigwedge X_{ej,Diseases} \tag{6.8}$$

Trouble mental - Substances psychoactives (Psy_subs)

$$X_{rij,Due\_to} = X_{ei,DM} \bigwedge X_{ej,Psy\_subs} \tag{6.9}$$

Substances psychoactives (Psy_subs) - Trouble mental

$$\mathcal{X}_{rij,induced} = \mathcal{X}_{ei,Psy\_subs} \bigwedge \mathcal{X}_{ej,DM} \tag{6.10}$$

Trouble mental - Trouble mental

$$\mathcal{X}_{rij,causes} = \mathcal{X}_{ei,DM} \bigwedge \mathcal{X}_{ej,DM} \tag{6.11}$$

Maladies - Maladies

$$\mathcal{X}_{rij,causes} = \mathcal{X}_{ei,Diseases} \bigwedge \mathcal{X}_{ej,Diseases} \tag{6.12}$$

Sur la base de ces contraintes, nous proposons une table d'entités et de relations qui représente conjointement des entités et des relations dans une phrase **S**. $\perp$ dénote une relation non définie.

TABLE 6.2 : Modélisation de l'extraction d'une entité conjointe et d'une relation avec une table représentative.

	MD e1,e2,...	SYM e1,e2,...	Diseases e1,e2,...	Psy_subs e1,e2,...	Treatments e1,e2,...
MD e1,e2,...	causes	Has_symptom	Is_a	Due_to	Has_cure
SYM e1,e2,...	Symptom_of	$\perp$	$\perp$	$\perp$	$\perp$
Diseases e1,e2,...	$\perp$	$\perp$	causes	$\perp$	$\perp$
Psy_subs e1,e2,...	induced	$\perp$	$\perp$	$\perp$	$\perp$
Treatments e1,e2,...	cures	$\perp$	$\perp$	$\perp$	$\perp$

Soit ei la reconnaissance des entités nommées dans **S**. Nous mettons chaque entité nommée de la phrase **S** identifiée dans sa classe **C**. Une classe $\mathbf{C_i}$ peut avoir une ou plusieurs entités nommées. La table est **nxn** éléments, où **n** est le nombre de classes. Dans notre cas, nous avons 5 classes (troubles mentaux, symptôme, organe, substances psychoactives et traitement). Soit $ei \in Ck$ et $ej \in Cl$, tel que $ei \neq ej$ et $Ck \wedge Cl \neq null$.

Par conséquent, l'intersection de la colonne qui contient la classe Cl avec la ligne qui contient la classe Ck donne la cellule qui contient la relation R et l'intersection de la

colonne qui contient la classe *Ck* avec la ligne qui contient la classe *Cl* donne la relation R^{-1}.

Exemple
غالبا ما تظهر الخلجات أثناء فترة الطفولة، عادة قبل سن العاشرة، وتنتج عن اضطرابات نفسية طفيفة.

Les tics apparaissent souvent pendant l'enfance, habituellement avant l'âge de dix ans, et résultent d'un trouble psychologique mineur.

Dans l'exemple ci-dessus, nous avons deux entités et deux classes. "**الخلجات** / tics" appartient à la classe SYM et " **اضطرابات نفسية طفيفة**/ un trouble psychologique mineur" Appartient à la classe MD. Nous concluons que R = Has_symptom et R^{-1} = Symptom_of.

TABLE 6.3 : Modélisation de l'exemple précédent avec une table représentative.

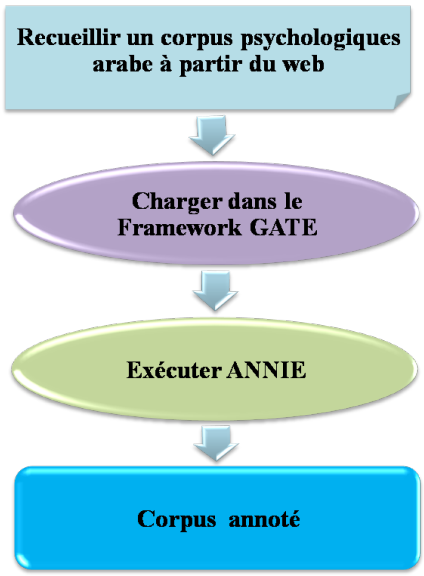
	MD **اضطرابات نفسية طفيفة**	SYM **الخلجات**	Diseases null	Psy_subs null	Treatments null
MD **اضطرابات نفسية طفيفة**	causes	Has_symptom	Is_a	Due_to	Has_cure
SYM **الخلجات**	Symptom_of	⊥	⊥	⊥	⊥
Diseases null	⊥	⊥	causes	⊥	⊥
Psy_subs null	induced	⊥	⊥	⊥	⊥
Treatments null	cures	⊥	⊥	⊥	⊥

Les mesures d'évaluation standard dans l'extraction de l'information (F-mesure, la précision et le rappel) sont utilisées pour évaluer la méthode proposée (voir la section 3).

6.6 Expériences et résultats

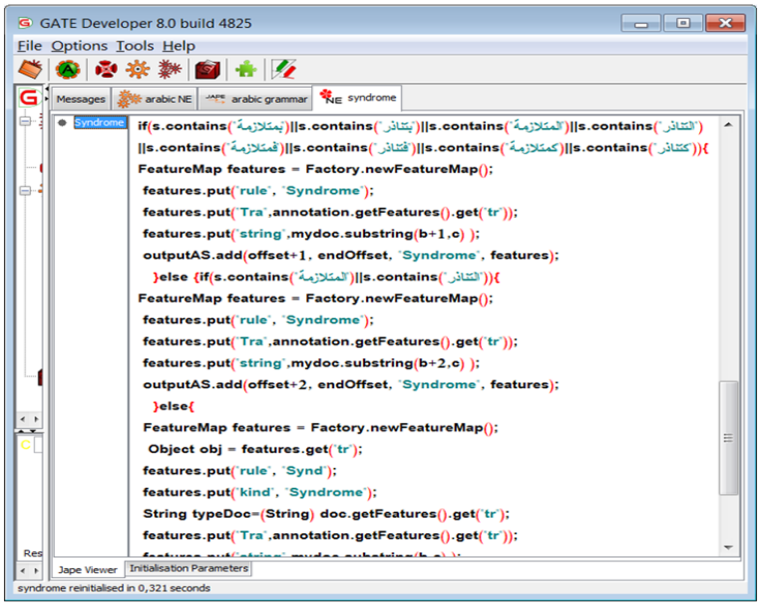
Le processus commence par le chargement du corpus dans le framework d'application avec les grammaires JAPE et les Gazatteers pour permettre l'annotation des concepts du corpus. La figure 6.3 illustre le processus d'annotation du corpus.

FIGURE 6.3 : Le processus d'annotation



6.7 Intégration de différents extracteurs de NE

FIGURE 6.4 : Règle JAPE pour l'extracteur syndrome NE



Un certain nombre d'expériences ont été menées afin d'évaluer la performance de notre système NERPSY lorsqu'elle est appliquée à différents ensembles de données pour d'extraire les entités nommées désirées. Et à chaque fois nous améliorons les Gazetteers et les règles d'extraction.

FIGURE 6.5 : Résultats de l'annotation des syndromes

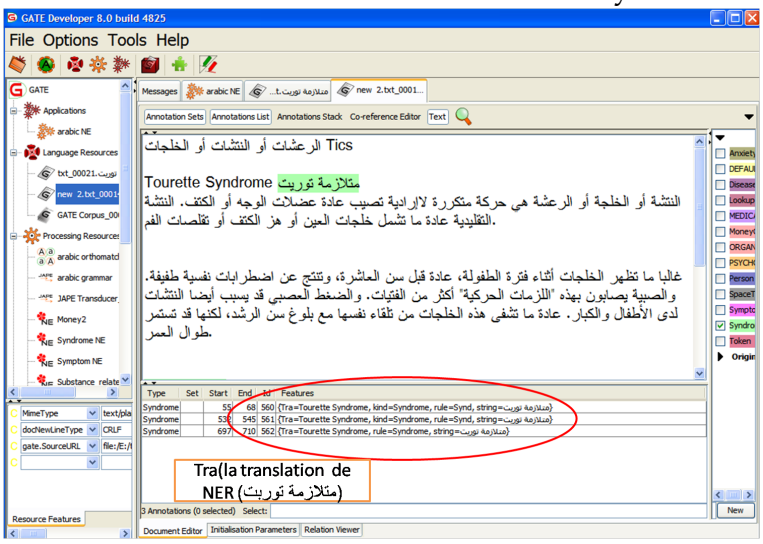
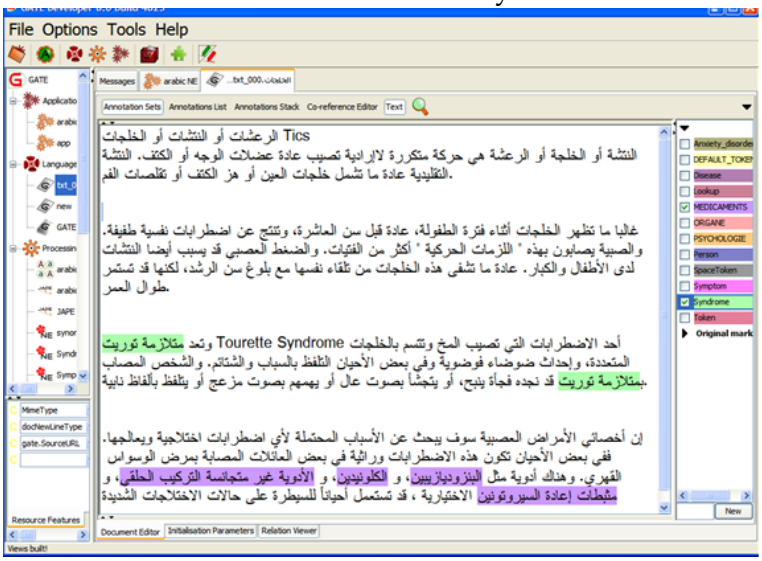


FIGURE 6.6 : Résultats de l'annotation des syndromes et médicaments



La figure 6.6 montre un exemple du résultat à l'aide des annotateurs écrites en JAPE, l'exemple montre les types d'annotation Syndrome et médicaments.

TABLE 6.4 : Annotation manuelle vs automatique.

Type d'annotation	Total d'entités reconnu	Correcte	Incorrecte
Syndromes	12	12	0
Trouble psychologique	18	14	4
Phobie	8	8	0
Maladies	4	4	2
symptômes	30	30	12
Organes	18	14	04
Substances	06	06	02
Traitements	08	08	04

Le tableau 6.7 montre les valeurs calculées de Précision et Rappel pour chaque classe des entités nommées ("متلازمات" (syndromes), "الأعراض" (symptômes), "الأمراض" (Maladies), "الأعضاء (organes)", "الاضطرابات النفسية". (Trouble psychologique) ","المواد المسببة للاضطرابات "(substances psychoactives)," العلاج "(traitements),"الرهاب "(Phobie)).

Les résultats sont calculés sur la base des équations 6.1, 6.2 et 6.3. Selon ces résultats, on note que, généralement, la précision moyenne est 97% et dans la plupart des cas, elle est meilleure que le rappel où la précision se réfère au nombre d'EN correctement

prédites rapportées au nombre total d'EN identifiés pour une EN donnée. À partir du tableau 6.5, on peut dire que le résultat de notre système est satisfaisant lorsqu'on le compare à l'évaluation humaine où le rappel moyen est de 79,90%.

Nous avons vue que la meilleure performance de notre système hybride proposé est atteint lorsque toutes les caractéristiques (features) des différents types sont pris en compte dans l'ensemble des caractéristiques représentant l'ensemble de données.
Nous remarquons aussi que le niveau de rappel moyen est modéré (80%). La raison principale est : Des attributs manquants dans les listes de Gazetteer, ajouter des attributs et des synonymes manquants dans le Gazetteer peut résoudre ce problème et a cause de l'ambigüité sémantique par exemple, lors du traitement de texte par la machine, l'entité mentionne "sin /سن " peuvent être liés plus de 7 entités différentes qui existent en arabe almaany[123] base de connaissances. Par exemple, "سن / âge", et "سن / dent". Un outil désambiguïsation peut résoudre le problème.

6.8 Comparaison avec les résultats existants

Dans la seconde expérimentation, nous comparons notre processus de reconnaissance d'entités nommées avec les résultats obtenus par NAMERAMA [123] testés sur des données obtenues sur le site Web de KAAHE [124].

TABLE 6.6 : Comparaison avec les résultats existants

	NAMERAMA			**Notre approche**		
Annotation type	Rappel	Précision	F-measure	Rappel	Précision	F-measure
Maladies	96.29%	100%	98.10%	100%	66,70%	80%
Symptomes	68.18%	30%	41.66%	100%	71,40%	83,30%
Traitements	55.73%	91.89%	69.38%	100%	66,70%	80%

Les résultats ont clairement montré que les performances des différents systèmes variaient significativement selon le type d'entité et le corpus. Avec notre système, les traitements ont été correctement identifiés dans 80% des cas, ce qui est encore mieux que les systèmes NAMERAMA (69,4%). Pour la catégorie des symptômes, notre approche a obtenu la meilleure F -mesure (83, 30%). Cela peut s'expliquer par la bonne couverture de la nomenclature des symptômes et la méthode de recherche sûre (règle de jape). Pour les maladies, NAMERAMA a obtenu la meilleure F -mesure (98,10%) et notre système s'est classé deuxième avec une F -mesure de 80%. De plus, notre outil n'a pas pu identifier correctement plusieurs entrées. Une analyse plus approfondie a révélé de nombreux cas

de catégorisation incorrecte en raison de l'ambiguïté de certains mots arabes. En outre, l'absence de normes terminologiques pour la saisie de texte en arabe a conduit à des incohérences dans l'orthographe de certains mots et a donc influencé nos résultats. En outre, problème de terme scientifique, représenté par le manque de termes scientifiques en arabe clair, facile et accepté. Certaines erreurs ont également été causées par les variantes orthographiques d'entités traduites ou translittérées qui n'étaient pas présentes dans notre répertoire géographique (par exemple, comme le syndrome de Turner pourrait être écrit de deux manières différentes comme متلازمة ترنر ou متلازمة تيرنر).

Pour l'extraction de la relation, nous avons sélectionné 50 phrases de notre corpus. Dans ces phrases, nous avons trouvé 123 relations. le tableau 6.8 illustre les valeurs globales de rappel, de précision et de F-mesure sur toutes les relations extraites. Cette expérience est une fonction écrite de la sortie du système conçue et comparée à une analyse humaine détaillée pour chaque type de relations extraites.

TABLE 6.7 : Résultats de précision, rappel et F-mesure.

Type de Relation	**Rappel**	**Précision**	**F-measure**
Has_symptom	66,67%	100%	80%
Symptom_of	66,67%	100%	80%
Induce	100%	100%	100%
Due_to	100%	100%	100%
Has_cure	50%	100%	66,67%
Cures	50%	100%	66,67%
Is_a	75,00%	100%	85,70%
Overall	100%	100%	100%

Dans la littérature, il n'y avait pas d'approche d'extraction de relation pour l'arabe standard moderne dans le domaine médical qui a soutenu la comparaison de nos résultats avec d'autres approches. Les systèmes actuels d'extraction de Relation Arabe se concentrent sur la détection et la classification des relations entre différents ensembles d'entités tels que la personne, l'organisation et la localisation. Cette approche permet de capturer des relations implicites telles que les relations détectées via une relation inverse R_{-1}. Nous pouvons donc extraire de nombreuses relations en une phrase (par exemple à partir de la phrase de la figure 5.2 :

" تصدر عن المصاب بمتلازمة توريت عراتٌ صوتية أو حركات جسدية لا إرادية ".
"Émis par le syndrome du patient de Tourette ou des mouvements involontaires.".

Nous pouvons extraire Has_symptom (متلازمة توريت / syndrome de Tourette /, عرات صوتية/ son) et Has_symptom (متلازمة توريت / syndrome de Tourette, حركات جسدية لا إرادية / mouvements involontaires)). Symptom_of (عرات صوتية / son, متلازمة توريت / syndrome de Tourette) et Symptom_of (حركات جسدية لا إرادية / mouvements involontaires, متلازمة توريت / syndrome de Tourette).

Basé sur des contraintes extraites automatiquement par la technique d'apprentissage supervisé ou ajoutées manuellement, notre méthode permet d'obtenir des résultats encourageants. Bien qu'il présente des performances prometteuses en termes de précision et de rappel, notre processus ne peut extraire certaines des relations présentes dans les phrases négatives.

6.9 Conclusion

Dans ce chapitre, nous avons parlé de la réalisation, des résultats expérimentaux et de l'évaluation du système proposé. Dans la première section, nous avons présenté les outils et les programmes utilisés pour la mise en œuvre du système. Dans la deuxième section, nous avons expliqué l'interface du système. Dans la troisième section, nous avons présenté les exemples expérimentaux réalisés pour certains types d'annotations. La quatrième section présentait les mesures d'évaluation de notre modèle. Dans la cinquième section, nous avons discuté des résultats.

Chapitre 7

Conclusion et futurs travaux

Cette thèse s'inscrit dans le domaine de traitement automatique de la langue naturel, plus particulièrement, le travail présenté ici s'intéresse à l'extraction des informations psychologique pour la langue arabe. Nous commençons par rappeler les principales contributions apportées par notre travail, avant de présenter quelques perspectives pour de futurs travaux de recherche qui s'inscrivent dans la continuité de cette thèse.

Sommaire

7.1 Conclusion générale

Avec l'explosion de l'information dans le domaine biomédical pose de vrais défis pour les chercheurs désireux d'analyser et d'organiser cette information. Donc, il existe une forte demande d'automatiser les techniques d'extraction d'informations biomédicales. Le but de l'extraction de l'information (IE) dans le domaine biomédical est de convertir un texte biomédical non structuré en information structurée de telle sorte que l'information puisse ensuite être analysée et agrégée.

La reconnaissance d'entités nommées et l'extraction de relations sont deux tâches fondamentales et essentielles dans l'extraction de l'information dans le traitement du langage naturel. Principalement pour deux raisons, le but de la première tâche est d'extraire une liste prédéfinie d'entités, la reconnaissance de ces entités représente la brique fondamentale pour construire une analyse sémantique. La deuxième tâche consiste à extraire

les relations explicites et implicites entre ces entités pour représenter l'interaction entre elles à partir du contenu du texte. L'information (bio)médicale joue un rôle important dans l'extraction d'information médicale, la description, l'exploration et la modélisation. De plus, les informations (bio)médicales dans un texte en langage naturel sont qualitatives, floues et compliquées.
NERPSY est développé pour la langue arabe dans le domaine psychologique. Les principales contributions de cette thèse sont comme suits :

- Extraction automatique d'entités nommées psychologiques à partir d'un véritable texte psychologique non structuré.
- Comporte un module de translation pour traduire les entités nommées extraites.
- Fournit des informations de base pour l'analyse psychologique et pour la construction des ontologies bilingue (AR-EN) du domaine psychologique.

Pour évaluer notre contribution, nous avons suivi les cinq critères d'évaluations proposées par [125], ces critères suggérés pour la planification et l'évaluation de systèmes d'informations et qui est été adapté au contexte de l'extraction d'information par [126].
Significativité
Ce critère, significativité (Significance), vise à mesurer l'importance du système ou de l'approche proposée sur le plan théorique et/ou pratique par rapport aux travaux existants. Par exemple, en point de vue des performances (système plus rapide, etc.) ou du point de vue des fonctionnalités (fonctions supplémentaires, etc.). Nous pouvons traduire ce critère par les questions suivantes :

- Quels sont les principales contributions du travail de recherche pour l'extraction d'information ?
- Est-ce que les approches proposées améliorent les performances par rapport aux approches existantes ?
- Quels sont les distinctions entre les approches proposées et les travaux précédents ?

Vis à vis de la première question, nos approches dans la partie II, portent les contributions suivantes :
Nous nous sommes intéressés à l'extraction d'informations à partir de textes psychologiques. Nos principaux objectifs sont l'identification, la classification automatiquement l'entité nommée en intégrant l'approche basée sur la Machine Learning avec l'approche symbolique pour former une approche hybride et pour la représentation d'interaction

entre ces entités,est formulé avec la programmation linéaire pour extraire des relations explicites et implicites. Autant que nous sachions, l'extraction d'informations psychologiques en langue arabe n'a toujours pas été tentée par aucun chercheur.
La première tâche, la reconnaissance des entités nommées psychologiques revient a localiser une sous-chaine dans la phrase et a lui attribuer une des catégories pré-définies. Le système hybride proposé est capable de reconnaître 8 différents types d'entités nommées, y compris les troubles mentaux désigné par DSM-IV (Trouble psychologique, phobie, Syndrome), substances psychoactives, symptômes, maladies, Organe et les médicaments. En suite, nous avons utilisé un module de translation pour traduire les entités nommées.
La deuxième tâche, Pour l'acquisition des relations, nous avons utilisé une méthode fondée sur la programmation linéaire en nombre entier tel que l'acquisition est effectué sous un ensemble de contraintes à fixer selon les relations médicales. Notre motivation est de fournir des informations de base pour l'analyse psychologique et pour la construction des ontologies du domaine psychologique.
Concernant nos performances, il n'est pas simple de juger de façon simple les résultats, en tout cas, pour la tache d'extraction d'information psychologique. Nos expérimentations ont été effectuées sur un corpus de petite taille provenait de plusieurs textes du Web en comparaison des corpus utilisés dans les autres approches tel que ACE 2004 Newswire ou encore ANERCorp. Aussi, Autant que nous sachions, l'extraction d'informations psychologiques en langue arabe n'a toujours pas été tentée par aucun chercheur. En revanche, concernant l'extraction des relations à large échelle, notre méthode permet d'obtenir des résultats encourageants. Bien qu'il présente des performances prometteuses en termes de précision et de rappel, notre processus ne peut extraire certaines des relations présentes dans les phrases négatives.
Pour ce qui est de la dernière question, notre système diffère des approches existantes, dans le domaine appliqué : Ces approches présentent des systèmes pour reconnaître l'entité nommée dans le domaine général, tandis que, notre approche présente un système de reconnaissance d'entité nommée spécialisée dans le domaine médical plus précisément dans le domaine psychologique.
Validité interne (Internal validity)
Ce critère fait référence à la crédibilité des arguments avancés. Les résultats de l'étude sont-ils logiques ou cohérents avec ces arguments ? Dans le travail expérimental et quasi-expérimental, la validité interne se réfère à la nature de la preuve et aux arguments pour les relations causales entre les construits. Ce critère nous a conduit à posé les questions suivantes :

- Les méthodes et procédures de l'étude sont-elles décrites explicitement et en dé-

tail ?

- Est-ce que la démarche proposée est adaptée à la problématique ? Atteint elle son objectif ?
- Si des hypothèses ont été faites sur les résultats, ont-elles été validées ? Des approches similaires (concurrentes) ont-elles été considérées ?

Nous avons vu que le premier objectif était la reconnaissance des entités nommées psychologiques en langue arabe, puis la translation de ces entités vers la langue anglaise, donc nous avons vu que pour atteindre cet objectif, nous avons adapté la plateforme Gate pour la langue arabe, en se basant sur les étapes suivantes : Premièrement, nous avons construit des Gazetteers psychologiques bilingues (arabe- anglais) à partir des ressources psychologiques. Puis nous avons crée des règles JAPE pour les différents types des ENs psychologiques. Ensuite, nous avons ajouté un module de traduction des entités nommées identifiées dans chaque règle. Enfin, nous avons intégré les Gazetteers d'ENs construits et les règles dans la plate forme GATE. Les évaluations ont montré que cette tache n'était pas totalement résolu et que une analyse plus approfondie a révélé de nombreux problème liées aux Gazetteers et aux textes. Certaines erreurs ont également été causées par les variantes orthographiques d'entités traduites ou translittérées qui n'étaient pas présentes dans nos nomenclateurs. Par ailleurs, pour l'acquisition des relations bio(medical), nous avons montré que notre méthode basée sur les contraintes permet de capturer les relations explicites via la relation directe et les relations implicites telles que les relations détectées via une relation inverse R^{-1}. Dans la littérature, il n'y avait pas d'approche d'extraction de relation pour l'arabe standard moderne dans le domaine (bio)médical qui a soutenu la comparaison de nos résultats avec d'autres approches. Notre méthode permet d'obtenir des résultats encourageants. Bien qu'elle présente des performances prometteuses en termes de précision et de rappel, notre processus ne peut extraire certaines des relations présentes dans les phrases négatives.

Validité externe (External validity)

Ce critère concerne le niveau de généralisation associe aux résultats trouvés. Par exemple, il peut s'agir d'une abstraction d'une approche existante ou l'application d'une même démarche dans des configurations différentes. Les questions liées à ce critère sont :

- Si les résultats sont obtenus à partir d'approches existantes, est-ce qu'ils sont cohérents avec ceux obtenus par d'autres travaux reposant sur ces mêmes approches ?
- Est-ce que les méthodes ou processus produits sont suffisamment génériques pour être transfèrés à d'autres domaines ?

- Quels sont les limitations à considérer afin d'appliquer les approches proposées à d'autres domaines ?

Concernant la transposition de notre approche à d'autres domaines, pour l'approche de reconnaissance d'entités nommées psychologiques, il s'agit de la langue et du domaine étudié. Notre approche est fortement liée à la langue et au domaine étudié, mais on peut la transposer vers d'autres domaines où il faut un changement total de ces composants (les Gazetteers et les règles). Enfin, le processus de reconnaissance d'entités reste lié à la langue et au domaine.
Pour l'acquisition des relations, l'approche s'appuie sur des contraintes indépendantes du domaine et de la langue, nous avons transposé ces contraintes de la langue anglaise en domaine général vers la langue arabe en domaine psychologique, donc il faut les adapter d'une langue à l'autre et d'un domaine à l'autre.
Confirmabilité (Objectivity/Confirmability)
Ce critère s'intéresse à la maniéré dont les ressources et les procédures sont décrites afin de pouvoir reproduire les résultats. La question liée à ce critère est :

- Les méthodes de l'approche sont-elles décrites en détail ?

Les méthodes et les ressources de notre approche sont détaillées dans la deuxième partie.
Fiabilité (Reliability/Dependability/Audibility)
Ce critère affirme que le processus de l'étude est cohérent, raisonnablement stable dans le temps et entre les chercheurs et les méthodes. La question est de contrôle de qualité. Est-ce que les choses ont été faites avec un soin ? Cela conduit à poser les questions suivantes :

- Les problématiques sont ils clairement définies ?
- Les constructions de base sont-elles clairement spécifiées ?

Nous avons entamé la problématique de l'extraction d'information psychologique à partir des textes arabe en deux chapitres (chapitre 4 La reconnaissance d'entité nommée psychologique et chapitre 5 Acquisition des relations) où les constructions de base sont clairement décrites.

7.2 Perspectives

Les travaux présentés dans cette thèse ouvrent de nombreuses perspectives cela a conduit à étendre notre travail. Ce qui suit est un résumé des futurs travaux :

(a) **Construire une ontologie psychologique**

Nous avons fourni dans cette thèse les éléments de base pour la construction d'une ontologie psychologique en langue arabe (les concepts et les relations). Le développement de l'ontologie comprend les étapes suivantes[127] :

- Définissez des concepts, c'est-à-dire des classes basées sur l'étude et l'analyse du domaine.
- Définissez des instances, c'est-à-dire des éléments réels dans notre domaine.
- Définir les relations entre les classes comme une exigence pour arriver à l'ontologie.
- Enrichissez l'ontologie avec les synonymes et les mots racines.
- Évaluation de l'ontologie.

(b) **Le Mapping de notre ontologie avec une autre ontologie écrite dans une autre langue avec MetaMap.**[2]

MetaMap [128] est un outil efficace et robuste pour mapper des termes aux concepts du Metathesaurus UMLS [129]. Il peut prendre une chaîne de texte comme entrée, le segmenter en phrases, puis mapper chaque phrase à plusieurs CUI UMLS (Concept Unique Identifiers Unified Medical Language System) avec des scores de confiance. Nous pouvons prévoir cette perspective comme suit :

Table de traduction

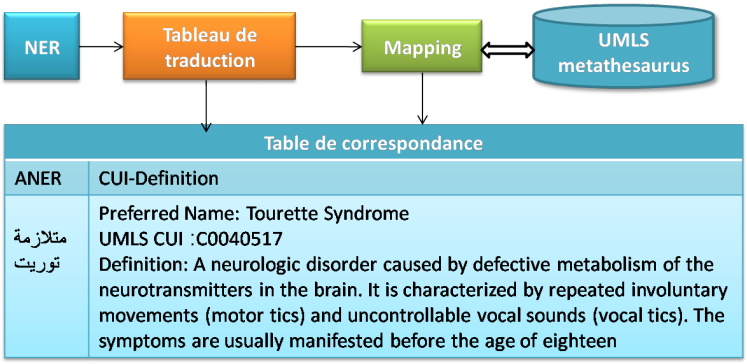
Construire une table de traduction arabe-anglais à partir des entités bilingues nommées extraites est un point principal pour utiliser le programme MetaMap, car il est basé sur la grammaire anglaise.

Exécution de MetaMap

Dans cette étape, les ENs seront traités avec MetaMap. La sortie de chaque traitement d'EN est une liste de termes candidats et leurs CUI UMLS. Enfin, en utilisant la table de traduction et les CUI UMLS, nous pouvons créer la table correspondante entre un ensemble de NER dans un concept médical arabe et des termes anglais dans les mêmes concepts UMLS.

Exemple

FIGURE 7.1 : CUI UMLS et les types sémantiques : Syndrome de la Tourette



Plus généralement, une perspective importante pour la suite des travaux concerne l'opérationnalisation de l'ontologie dans la recherche d'information et dans l'analyse psychologique pour aider les spécialiste du domaine à prendre des décisions plus précises.

Bibliographie

[1] Najjar, J.(2010). From Anesthetic Sponge to Nonsinking Skull Perforator. Unitary Work Neurosurgery in the Ancient Arabic and Islamic World, pp.587-594.

[2] Lakel, K. and Bendella, F .(2017).Psychological Named Entity Recognition from psychological Arabic texts. Internationaljournal of Metadata, Semantics and Ontologies,Inderscience publishers, Vol. 12, Nos. 2/3, pp.82-88.

[3] Lakel, K., Bendella, F. and Benkhedda, S. (2017). Named entity recognition for Psychological domain : Challenges in document annotation for the Arabic Language. IEEE, First International Conference on Embedded & Distributed Systems (EDiS),pp. 39-43, special issue on International Journal of Reasoning-based Intelligent Systems , Inderscience Publishers, 2018 .

[4] Farber, B., Freitag, D., Habash, N. and Rambow, O.(2008) Improving NER in Arabic using a morphological tagger. Proceedings of workshop on HLT & NLP within the Arabic world (LREC 2008), pp. 2509-2514.

[5] AbdelRahman, S., Elarnaoty, M., Magdy, M. and Fahmy, A.(2010). Integrated machine learning techniques for Arabic named entity recognition. International Journal of Computer Science Issues (IJCSI) , 7 : 27-36.

[6] For a full list of other Arabic legacy encoding the reader is referred to http ://www.i18nguy.com/unicode/codepages.html..

[7] Carroll, J.J. (2005). An Introduction to the Semantic Web, Considerations for building multilingual Semantic Web sites and applications. HPL-2005-67 technical report. http ://www.hpl.hp.com/techreports/2005/HPL-2005- 67.html

[8] OntoSelect - a multilingual ontology library and ontology selection service. http ://olp.dfki.de/OntoSelect/

[9] Protégé, http ://protege.stanford.edu/

[10] General Architecture for Text Engineering ,http ://gate.ac.uk/

[11] Java framework for building Semantic Web and Linked Data applications, http ://jena.sourceforge.net/

[12] Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C. (2006). Building a WordNet for Arabic. Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).

[13] Paolillo, J. (2005). Language Diversity on the Internet. Measuring Linguistic Diversity on the Internet, UNESCO Publications for the World Summit on the Information Society, pp. 43-89.

[14] Alkhalifa, H. and Alwabil, A.(2007). The Arabic language and the semantic web : Challenges and Opportunities. The 1st int. symposium on computer and Arabic language.

[15] Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web, Scientific American Magazine, Vol. 284, No. 5, pp. 34-43.

[16] Zaidi, S., Laskri, M.T. and Bechkoum,K.(2005). A Cross-language Information Retrieval : Based on an Arabic Ontology in the Legal Domain. Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05), pp. 86–91.

[17] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Introducing the Arabic WordNet Project. Proceedings of the Third International WordNet Conference, Fellbaum and Vossen (eds), pp. 295–300.

[18] Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Ould Abdallahi Ould Bebah, M. and Shoul, M. (2010).Alkhalil Morpho SYS1 : A Morphosyntactic Analysis System for Arabic Texts. Proceedings of the 11th International Arab Conference on Information Technology. Benghazi, Libya.

[19] Buckwalter, T. (2002) Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

[20] Abu ElKhair, I. (2007). Arabic information retrieval. Annual review of information science and technology, 41(1), pp. 505-533.

[21] Hammo, B., Abu-Salem, H. and Lytinen, S.(2002).QARAB : A question answering system to support the Arabic language . Proceedings of the ACL-02 workshop on Computational approaches to semitic languages. Association for Computational Linguistics. pp. 1-11.

[22] Maloney, J., Niv, M. (1998). TAGARAB : A Fast, Accurate Arabie Name Recognizer Using High-Precision Morphological Analysis. Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, pp.8-15.

[23] Shaalan, K. and Raza, H. (2009). NERA : Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60 :1652 -663.

[24] Zaghouani, W., Pouliquen, B., Ebrahim, M. and Steinberger, R. (2010). Adapting a resource-light highly multilingual named entity recognition system to Arabic. Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10), Malta. European Language Resources Association (ELRA).

[25] Samy, D., Moreno, A. and Guirao, J. M. (2005). A proposal for an Arabic named entity tagger leveraging a parallel corpus. Proceedings of the Recent Advances in Natural Language Processing RANLP, Borovets, Bulgaria, pp. 459-465.

[26] Aggarwal, C. C. and Zhai, C. (2012).A survey of text classification algorithms. In Mining Text Data. Vol. 9781461432234, Springer US, pp. 163-222).

[27] Elsebai, A.(2009). A rule based system for named entity recognition in modern standard Arabic, School of Computing , Science and Engineering University of Salford , Salford , UK, Ph.D Thesis.

[28] Sondhi, P. (2008). A Survey on named Entity Extraction in the Biomedical Domain. http ://sifaka.cs.uiuc.edu/ sondhi1/survey1.pdf

[29] Le Meur, C., Galliano, S. and Geoffrois, E. (2004). Conventions d'annotations en Entités Nommées - ESTER. http ://www.afcpparole.org/ester/docs/convention_en_old.pdf.

[30] Goldman. J.P. and Scherrer, Y. (2012). Création automatique de dictionnaires bilingues d'entités nommées grâce à Wikipédia. Cahiers de linguistique française, Vol. 30, No. 11, pp .213-227.

[31] Huang, Z. and Hu, X. (2013). Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. International Journal of Machine Learning and Computing. Vol. 3, No. 6.

[32] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference- 6 : A Brief History. Proceedings of the /6th conference on Computational linguistics (COLING), Copenhagen, Denmark, août 1996, pp.466-471.

[33] Chincor, N. (1997). MUC-7 Named Entity Task Definition Dry Run Version, Version 3.5, 17 September 1997. https ://catalog.ldc.upenn.edu/docs/LDC2001T02/guidelines.NEtaskdef.3.5.ps.

[34] Sundheim, B. M. (1991). Overview of the third message understanding evaluation and conference. Proceedings of the 3rd conference on Message understanding, Association for Computational Linguistics. pp. 3–16.

[35] Sundheim, B. M. (1992). Overview of the fourth message understanding evaluation and conference. Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics. pp. 3–21.

[36] Chinchor, N. and Sundheim, B. (1993). MUC-5 evaluation metrics. Proceedings of the 5th conference on Message understanding, Association for Computational Linguistics. pp. 69–78.

[37] Sundheim, B. M. (1996). Overview of results of the MUC-6 evaluation. Proceedings of a workshop on held at Vienna, Virginia.

[38] Chinchor, N. A. (1998). Overview of MUC-7/MET-2, Proceedings of the Seventh Message Understanding Conference (MUC-7), Virginia, USA.

[39] American Psychiatric Association. (1994). Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association.

[40] Hassoon, T. (2004). Quick Reference to the Diagnostic Criteria from DSM-IV-TR™. Translated by Dr : Tayseer Hassoun Psychiatrist, Ibn Sina Medical Center, Damascus, Syria.

[41] Turky, J. (2004). Psychological sciences dictionary Ar-Fr-En. WebPsySoft Arab Company. Webteb. Webteb is the largest Arabic-language health and wellness site.

[42] Nadeau, D. and Sekine S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes. 30(1) :3–26.

[43] Shaalan, K. and Raza, H. (2008). Arabic named entity recognition from diverse text types. Advances in Natural Language Processing, ed : Springer. pp. 440-451.

[44] Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools : Resources and Tools for Building MT, Association for Computational Linguistics. pp. 1–8.

[45] Hamadene, A., Shaheen, M. and Badawy O. (2011). ARQA : An intelligent Arabic question answering system. Proceedings of Arabic language technology international conference (ALTIC 2011).

[46] Weizenbaum, J. (1966). ELIZA- a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9 (1), pp.36–45.

[47] Chowdhury, G. G. (2005). Natural language processing. Annual Review of Information Science and Technology, pp.51–89.

[48] Friedman, C. and Hripcsak, G. (1999). Natural language processing and its future in medicine. Acad Med, 74 (8), pp.890–895.

[49] Spyns, P. (1996). Natural Language Processing in Medicine : An Overview. Methods of information in medicine, 35 (4), pp.285–301.

[50] McDonald, A. (1996) . Internal and external evidence in the identification and semantic categorization of proper names. B. Boguraev and J. Pustejovsky, editors, Corpus Processing for Lexical Aquisition, MIT Press, Cambridge, MA, chapter 2, pp. 21-39.

[51] Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. Proceedings of the 37th Annual Meeting of the ACL, , University of California, Maryland, pp. 159-166.

[52] Mesfar, S.(2007). Named entity recognition for Arabic using syntactic grammars. Proceedings of the 12th international conference on application of natural language to information systems. Berlin : Springer, pp. 305–316.

[53] Shaalan, K. (2010). Rule-based approach in Arabic natural language Processing. The International Journal on Information and Communication Technologies (IJICT) ; 3. pp.11–19.

[54] Mayfield, J., McNamee., P. and Piatko, C.(2003). Named entity recognition using hundreds of thousands of features. Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003 (CONLL 2003), pp. 184–187.

[55] Sun, B. (2010). Named entity recognition Evaluation of Existing Systems. Mémoire de maîtrise en système d'information. Université norvégienne de sciences et de technologie (NTNU).

[56] Bikel, D., Miller, S., Schwartz, R. and Weischedel, R .(1997). Nymble : a High- Performance Learning Name-finder. Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, USA, pp.194-201.

[57] Sekine, S. (1998). Nyu : Description of the Japanese NE System Used For Met-2. Proceedings of the Seventh Message Understanding Conference (MUC- 7),Fairfax, Virginia, USA, pp.l-6.

[58] Berger, A., Della Pietra , S. and Della Pietra, V. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, Vol. 22, No. 1, pp.39-71.

[59] Vapnik , V. (1999). An Overview of Statistical Learning Theory». IEEE Transactions on neural networks, Vol. 10, No. 5, pp.988-999.

[60] Blum, A. V. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. Proceedings of the Workshop on Computational Learning Theory (COLT), Morgan Kaufmann, pp.92-100.

[61] Larochelle, H. (2009). Étude des techniques d'apprentissage non-supervisé pour l'amélioration de l'entraînement supervisé de modèles connexionnistes. Thèse de doctorat en informatique, Université de Montréal.

[62] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. Linguisticae Investigaciones, Vol. 30, No. 1, pp.3-26.

[63] Maloney, J. and Niv, M. (1998). TAGARAB : a fast, accurate Arabic name recognizer using high-precision morphological analysis. Proceedings of the Workshop on Computational Approaches to Semitic Languages, Semitic '98, Stroudsburg, PA, USA, pp. 8-15.

[64] Zaghouani, W., Pouliquen, B., Ebrahim, M. and Steinberger, R. (2010). Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 563-567.

[65] Abdelkoui, F. and Kholladi, M.K. (2015). Automatic extraction of spatio-temporal information from Arabic text documents. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 7, No. 5, pp. 97-107.

[66] Al-Ahmari, S. S. and Al-Johar, B. A.(2016). Cross domains Arabic named entity recognition system, Proceedings of SPIE, First International Workshop on Pattern Recognition, vol. 10011, pp. 100111I-1 – 100111I-9.

[67] Benajiba, Y. and Rosso, P. (2007). ANERsys 2.0 : Conquering the NER task for the Arabie language by combining the maximum entropy with POS-tag information. Proceedings of the 3rd Indian International Conference on Artificial Intelligence, I/CAI-2007, Pune, India, pp.1814-1823.

[68] Benajiba, Y. and Rosso, P. (2008). Arabie named entity recognition using conditional random fields. Proceedings of the Conference on Language Resources and Evaluation (LREC), Marrakech Morocco.

[69] Ahmed, A. A. and Darwish, K. (2010). Simplified Feature Set for Arabic Named Entity Recognition. Proceedings of the 2010 Named Entities Workshop. NEWS '10. Stroudsburg, PA, USA : Association for Computational Linguistics, pp.110–115.

[70] Bidhendi, M., Behrouz, M. and Hosein, J. (2012) Extracting person names from ancient Islamic Arabic texts. Proceedings of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp.1–6.

[71] Morsi, A. and Rafea, A. (2013). Studying the impact of various features on the performance of Conditional Random Field-based Arabic Named Entity Recognition. Computer Systems and Applications (AICCSA), 2013 ACS International Conference, IEEE, pp. 1-5.

[72] Alotaibi, F. (2015). Fine-grained Arabic named entity recognition. Doctoral dissertation, University of Birmingham.

[73] Benajiba, Y., Diab, M. and Rosso, P. (2008). Arabie Named Entity Recognition using Optimized Feature Sets. Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Waikiki, Honolulu, Hawaii, pp.284-293.

[74] Koulali, R. and Abdelouafi, M. (2012) A contribution to Arabic named entity recognition. Proceedings of 10th International Conference on ICT and Knowledge Engineering, Morocco, pp.46–52.

[75] Benajiba, Y., Diab, D. and Paolo, R. (2009). Arabic named entity recognition : A feature-driven study. IEEE Transactions on Audio, Speech, and Language Processing, pp.926–934.

[76] Mohammed, N.F. and Omar, N. (2012). Arabic Named Entity Recognition Using Artificial Neural Network. Journal of Computer Science, vol. 8, No. 8, pp. 1285-1293.

[77] Althobaiti, M.J.(2016). Minimally-supervised methods for Arabic Named Entity Recognition, Ph.D. Thesis, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK.

[78] Abdallah, S., Shaalan, K. and Shoaib, M. (2012). Integrating rule-based system with classification for Arabic named entity recognition. A Gelbukh, ed. 2012 Computational Linguistics and Intelligent Text Processing, Berlin Heidelberg, (7181), pp.11–322.

[79] Meselhi, M. A., Bakr, H. M. A., Ziedan, I. and Shaalan, K. (2014). A Novel Hybrid Approach to Arabic Named Entity Recognition. China Workshop on Machine Translation. Springer Berlin Heidelberg, pp. 93-103

[80] Shaalan, K. and Mai, O. (2014). A hybrid approach to Arabic named entity recognition, Journal of Information Science, vol. 40, No. 1, pp. 67-87.

[81] Alanazi, S., Sharp, B. and Stanier, C.(2015). A Named Entity Recognition System Applied to Arabic Text in the Medical Domain. International Journal of Computer Science IJCSI , vol. 12, No. 3, pp.109-117.

[82] Conrath, J., Afantenos, S., Asher, N. and Muller, P. (2014). Extraction non supervisée de relations sémantiques lexicales. Traitement Automatique des Langues Naturelles-TALN . pp. 244- 255.

[83] Kumar, R., Jaloree, S. and Thakur, R. S. (2016). Developing Context Ontology using Information Extraction. International Journal of Computer Science and Information Security, Vol. 14, No. 3, pp. 358-363.

[84] Nebhi, K. (2013). Extraction de Relations basées sur une Ontologie pour le Web Sémantique. Congrès International des Linguistes. Genève (Suisse).

[85] Nguyen, V. T. (2008). Identification et extraction de relations n-aires à partir des textes. Rapport Equipe-projet EDELWEISS de l'INRIA Sophia Antipolis, France.

[86] Roth, D. and Yih, W. (2007). Global Inference for Entity and Relation Identification via a Linear Programming Formulation. L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning, The MIT Press, Cambridge, MA, pp.553-580.

[87] Berant, J., Srikumar, V., Chen, P.C., Linden, A. V., Harding, B., Huang, B., Clark, P. and D. Manning, C. (2014). Modeling biological processes for reading comprehension. EMNLP. pp. 499–1510.

[88] Berant, J., Dagan, I. and Goldberger, J. (2011). Global learning of typed entailment rules. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, Vol.1, Association for Computational Linguistics, pp. 610–619.

[89] Martins, A.F. T., Smith, N.A. and Xing, E.P. (2009). Concise Integer Linear Programming Formulations for Dependency Parsing, ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol.1, pp. 342-350.

[90] Liu, F., Flanigan, J., Thomson, S., Sadeh, N. and Smith, A. (2015). Toward abstractive summarization using semantic representations. In NAACL.

[91] Zeng, Y., Feng, Y., Ma, R., Wang, Z., Yan, R., Shi, C. and Zhao, D. (2017). Scale Up Event Extraction Learning via Automatic Training Data Generation. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). AAAI.

[92] Khashabi, D,. Khot, T., Sabharwal, A. and Roth, D.(2018). Question Answering as Global Reasoning over Semantic Abstractions. Conference of Association for the Advancement of Artificial Intelligence.

[93] Ben Hamadou, A., Piton, O. and Fehri, H. (2010a). Multilingual extraction of functional relations between arabic named entities using Nooj platform. hal-00547940, version 1.

[94] Boujelben, I., Jamoussi, S. and Ben Hamadou, A. (2012). Rules based approach for semantic relations extraction between Arabic named entities. NooJ2012. IN- ALCO, Paris, pp. 123–133.

[95] Hasegawa, T., Sekine, S. and Grishman, R. (2004). Discovering relations among named entities from large corpora. Association for Computational Linguistics. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA.

[96] Zhou, G., Qian, L. and Zhu, Q. (2009). Label propagation via bootstrapped support vectors for semantic relation extraction between named entities. Comput. Speech Lang. 23 (4), pp. 464–478.

[97] Zhang, Z. (2004). Weakly supervised relation classification for information extraction. Proceedings of ACM 13th Conference on Information and Knowledge Management CIKM2004, Washington D.C., USA.

[98] Alotayq, A.(2013). Extracting relations between Arabic named entities. TSD2013. Springer-Verlag, Berlin Heidelberg, Pilsen, pp. 265– 271.

[99] Ben Abacha, A. and Zweigenbaum, P.(2011). A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. 12th International Conference on Intelligent Text Processing and Computational Linguistics CICLING2011, Tokyo, Japan, pp. 139–150.

[100] Boujelben, I., Jamoussi, S. and Ben Hamadou, A.(2013b). Genetic algorithm for extracting relation between named entities. 6th Language and Technology Conference, LTC, Poznan, Poland, pp.484–488.

[101] Lavecchia, C. (2010). Les triggers inter-langues pour la Traduction Automatique Statistique. Thèse de doctorat en informatique, Université Nancy.

[102] Al-Onaizan, Y. and Knight, K. (2002). Machine transliteration of names in Arabic text. Proceedings of the ACL workshop on Computational approaches to semitic languages, Philadelphia, PA, USA, pp.1-13.

[103] Ling, W., Calado, P., Martins, B., Trancoso, I., Black, A. and Coheu, L.(2011). Named Entity Translation using Anchor Texts , Proceedings of the IWSLT, San Francisco, USA.

[104] Gornostay, T. and Skadina, I.(2009). Pattern-based English-Latvian Toponym Translation. European Association for Machine Translation conference.

[105] Brini.W ., Ellouze. M ., Trigui. O., Mesfar. S ., Belguith. H . and Rosso,.P .(2009). Factoid and definitional Arabic question answering system, Post-Proc. NOOJ-2009, Tozeur, Tunisia pp. 8-10.

[106] Mesfar. S.(2007), Named entity recognition for arabic using syntactic grammars. Natural Language Processing and Information Systems, ed : Springer, pp. 305-316.

[107] Alias-i, (2011). LingPipe. URL : http ://alia-i.com/lingpipe/

[108] Shaalan. K.(2014). A survey of Arabic named entity recognition and classification. Computational Linguistics, vol.40, pp. 469-510.

[109] Carpenter, B. (2006).Character language models for Chinese word segmentation and named entity recognition, Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 169-172.

[110] AbdelRahman, S., Elarnaoty ,M., Magdy, M. and Fahmy, A. (2010). Integrated machine learning techniques for Arabic named entity recognition, IJCSI, vol.7, pp. 27-36.

[111] Brill, E. (1992). A simple rule-based part of speech tagger. Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, pp.152-155.

[112] Plamondon, L. (2004). L'ingénierie de la langue avec GATE, Recherche Appliqué en Linguistique Informatique (RALI/DIRO), Université de Montréal.

[113] Thakker, D., Sman, T., Lakin, P. Thakker, D., Sman, T., Lakin, P. (2009). GATE JAPE Grammar Tutorial, Version 1.0, PA Photos, UK.

[114] Punyakanok, V., Roth, D., Yih, W. and Zimak, D. (2004). Semantic role labeling via integer linear programming inference. Proceedings of the International Conference on Computational Linguistics, Geneva, Switzerland, pp. 1346–1352.

[115] Roth, D. and Yih, W. (2004). A linear programming formulation for global inference in natural language tasks. Proceedings of the Annual Conference on Computational Natural Language Learning, Boston, MA, USA, pp. 1–8.

[116] Riedel, S. and Clarke, J. (2006). Incremental integer linear programming for non-projective dependency parsing. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, pp. 129–137.

[117] Berant, J., Dagan, I. and Goldberger, J. (2011). Global learning of typed entailment rules. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, Vol.1, Association for Computational Linguistics, pp. 610–619.

[118] Martins, A.F. T., Smith, N.A. and Xing, E.P. (2009). Concise Integer Linear Programming Formulations for Dependency Parsing. ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol.1, pp. 342-350.

[119] Liu, F., Flanigan, J,. Thomson, S,. Sadeh, N. and Smith ,N. (2015). Toward abstractive summarization using semantic representations. NAACL.

[120] Webteb. Webteb is the largest Arabic-language health and wellness site, Retrieved May 2, 2017, from https ://www.webteb.com/mental-health

[121] Tbeeb. Tbeeb is a personal site with a group of individuals and supervised by a group of doctors, from http ://www.tbeeb.net/t-11.htm

[122] De Sitter, A., Calders, T. and Daelemans, W. (2004). A Formal Framework for Evaluation of Information Extraction (Technical Report). Antwerp, Belgium : University of Antwerp, Department of Mathematics and Computer Science.

[123] Almaany. Almaany Arabic English and English Arabic dictionary, Retrieved May 1st 2017. from http ://www.almaany.com/ar/dict/arar/سَن/

[124] KAAHE.Encyclopédie de santé arabe du Roi Abdullah Bin Abdulaziz, https : //www.kaahe. org / ar /.

[125] Burstein, F. and Gregor, S. (1999). The Systems Development or Engineering Approach to Research in Information Systems : An Action Research Perspective System. Proceedings 10th Australasian Conference 122 on Information Systems, pp. 122–134.

[126] Ludovic, J.L. and Ferret, O. (2011). Approches supervisées et faiblement supervisées pour l'extraction d'événements complexes et le peuplement de bases de connaissances. thèse de doctorat en Sciences de l'Université de Paris 11 - Paris Sud

[127] Lakel, K. and Bendella, F. (2015).Dynamic Evaluation of Ontologies, Proceedings of the International Conference on Advanced Wireless Information and Communication Technologies, Procedia Computer Science, Elsevier, Tunisia, October 05-07, Vol.73, pp. 16-23.

[128] MetaMap. https ://metamap.nlm.nih.gov/

[129] Metathesaurus UMLS. https ://www.nlm.nih.gov/research/umls/.