الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Nº Réf :……

Centre Universitaire

Abd Elhafid Boussouf Mila

Institut des Sciences et Technologie **Département de Mathématiques et Informatique**

Mémoire préparé en vue de l'obtention du diplôme de Master

En : Informatique

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

Thème
Une nouvelle approche pour l'extraction automatique des entités nommées

Préparé par :

Yacoub Rania
Melghid Nedjla

Soutenue devant le jury :

Encadré par	**HADJI Atmane**	**Grade MAA**
Président	**LALOUCI Ali**	**Grade MAA**
Examinateur	**ATTIA Mourad**	**Grade MAA**

Année Universitaire : 2020/2021

Remerciement

En préambule à ce modeste mémoire nous remercions ALLAH de nous avoir aidé et donné la patience et le courage durant toutes les années de notre étude.

Nous tenons à remercier sincèrement et respectueusement notre encadreur HADJI Atmane de nous avoir fait partager ses connaissances et les lignes directives qu'il a apporté à ce travail.

Nous adressons nos plus sincères remerciements aux membres du jury d'avoir accepté d'examiner ce modeste travail.

Ces remerciements vont également envers les corps professoral et administratif du département Informatique, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Nos sincères dévouements et profondes gratitudes à Nos parents, frères et sœurs pour leurs encouragement et soutient tout au long de notre cycle d'étude.

Nous adressons nos plus sincères remerciements à tous nos proches et amis qui nous ont toujours aidé et encouragés au cours de la réalisation de ce mémoire spécialement « Ben Messioud Houcine ».

Enfin, Nous remercions tous ceux qui nous ont soutenus de près ou de loin durant notre cycle d'étude.

Merci à tous et à toutes

Nedjla & Rania

Dédicace

Je dédie mon travail au pays d'un million et demi de martyrs, mon pays bien-aimé, l'Algérie

À la propriétaire du plus haut degré dans mon cœur, ma grand-mère **Hadda**, que Allah lui fasse miséricorde

Au cher fils du précieux, mon cousin **Salam Eddine**, que Allah lui fasse miséricorde

Au symbole de ma vie, mon soutien et ma force, ma fierté A mon cher père **Athmane** que Allah le préserve

À la lumière de mon cœur et purificateur de ma tristesse, à la source de la tendresse, ma mère **Fatima-Zahra**, que Allah la préserve

À ma sœur unique et mon guide, à un morceau de mon âme, ma **sœur Soumia** et ses enfants **Loudjain** et **Arkane**

À qui j'avais une force dans ma faiblesse et une lumière dans mes ténèbres, à mon deuxième père mon cher **frère Zin Eddin** et sa femme **Meriem** et ses enfants **Iless**, mes princesses **Sadja** et **Radja**

À celui qui m'a poussé par derrière, pour être au premier plan, à la personne la plus proche de mon cœur, mon cher **frère Taki Eddin** et sa femme **Chahinez**

À mon cher **oncle Mourad** et sa femme **Samira**, à mon cher **oncle AbelHak,** sa femme **Hanane**

À ma belle et unique cousine **Hadil**, à tous mes cousins, mes frères : **Mahmoud, Mounir** et sa femme **Mouna, Oussama, Djihad, Noufel, Anass, Nidal.**

À mes amies et à ce que j'ai de plus précieuse : **Imane, Khawla, Manal, Amina, Sarah, Sihem, Hiba**

À ma unique et compagne sur mon chemin, À qui est devenue ma sœur avant de devenir mon binôme, **Rania**

À mes oncles et tantes et leurs enfants et toute la famille **Merghid**

Nedjla

Dédicace

Je dédie ce travail à la femme au beau cœur, titre de ma vie et de ma fierté, à ma chère maman **Sonia**, que Allah la préserve.

A celui qui m'a appris les bases de la vie, au symbole de ma force et de ma fierté, à mon cher père **Abdeldjalil** que Allah le préserve

À celui qui a le plus grand rôle, à mon soutien et à mon honneur, à celui qui valorisait le savoir à sa juste valeur, à celui qui est misérable à cause de moi, mon deuxième père, mon cher **frèreMohammad**

Au sens de ma vie et ma joie à deux morceaux de mon âme, mon cher frère **Faize** et ma sœur choyée **Hiba**.

Au symbole de dignité et de prestige A la tendresse ma sœur aînée **Samiha** et son mari **Ammar** et sa fille **Israa**

A ma chère cousine **Aisha**, à ma chère amie **Noussaiba**, que Allah leur fasse miséricorde

À mes amies et à ce que j'ai de plus précieuse : **Imane**, **Khawla**, **Amina**, **Manal**, **Sarah**

À ma belle et compagne sur mon chemin, À qui est devenue ma sœur avant de devenir mon binôme, **Nedjla**

Pour mes amis du parcoure d'étude : **Mohammad**, **Aissa**, **Muhammad Amin**, **Houcine**, **Khaled**

A mes tantes et oncles et à toute la famille **Yacoub**

Rania

Résumé

Aujourd'hui, un grand nombre de sources d'information publiques (sites internet, presse, radio, télévision, etc.) a fait émergé le besoin de rechercher cette masse de documents afin d'en extraire des connaissances pertinentes dans un but donné. Pour pouvoir identifier correctement les informations qui doivent être présentes à l'utilisateur de tels systèmes nécessitent généralement l'intervention d'un processus de traitement automatique du langage naturel (TALN) et ce afin de déterminer l'intention derrière la demande de l'utilisateur et pour révéler les entités nommées présentes dans ce dernier.

L'extraction automatique de l'information est un processus important et nécessaire au moment où les informations sont devenues des dossiers électroniques, que ce soit dans des sites commerciaux, d'actualités ou de médias sociaux. Cependant, l'extraction des informations sans utilisation de moyens auxiliaires ou spécification du contenu de ce que l'utilisateur veut avec précision, ne lui donne pas des résultats satisfaisants et efficaces, car il est possible de se retrouver dans certains cas de similitude ou d'erreur complète.

Pour résoudre ce problème, des entités nommées pour réduire la portée des résultats et améliorer leur qualité sont impliquées. L'objectif d'utiliser les entités nommées dans l'extraction de l'information est de renforcer la performance des résultats données et améliorer leur qualité. Dans ce mémoire nous présenterons une notre approche proposée d'un système d'extraction d'information, des réseaux sociaux (Facebook) sur les accidents afin d'extraire des entités nommées temporelles et spatiales avec des événements.

Mots-clés : Extraction de l'information, entité nommée, traitement automatique du langage naturel.

Abstract

Today, a large number of public information sources (websites, press, radio, television, etc.) has given rise to the need to search this mass of documents in order to extract relevant knowledge for a given purpose. In order to correctly identify the information to be presented to the user, such systems usually require the intervention of an automatic natural language processing (NLP) process to determine the intention behind the user's request and to reveal the named entities present in it.

Automatic information extraction is an important and necessary process as information has become electronic records, whether in commercial, news or social media sites. However, extracting information without using auxiliary means or specifying the content of what the user wants precisely, does not give him satisfactory and effective results, as it is possible to end up in some cases of similarity or complete error.

To solve this problem, named entities to reduce the scope of the results and improve their quality are involved. The objective of using named entities in information extraction is to enhance the performance of the given results and improve their quality. In this thesis we will present our proposed approach of an information extraction system, from social networks (Facebook) on accidents in order to extract temporal and spatial named entities with events.

Keywords: Information extraction, named entity, automatic natural language processing.

الملخص

اليوم ، أدى عدد كبير من مصادر المعلومات العامة (مواقع الويب ، والصحافة ، والإذاعة ، والتلفزيون ، وما إلى ذلك) إلى الحاجة إلى البحث في هذا الكم الهائل من الوثائق لاستخراج المعرفة ذات الصلة لغرض معين. من أجل تحديد المعلومات التي سيتم تقديمها للمستخدم بشكل صحيح ، تتطلب هذه الأنظمة عادةً تدخل عملية معالجة اللغة الطبيعية التلقائية (NLP) لتحديد طلب المستخدم والكشف عن الكيانات المحددة الموجودة فيه.

يعتبر الاستخراج التلقائي للمعلومات عملية مهمة وضرورية حيث أصبحت المعلومات سجلات إلكترونية سواء في المواقع التجارية أو الإخبارية أو مواقع التواصل الاجتماعي. ومع ذلك ، فإن استخلاص المعلومات دون استخدام وسائل مساعدة أو تحديد محتوى ما يريده المستخدم بدقة ، لا يعطيه نتائج مرضية وفعالة ، حيث من الممكن أن ينتهي الأمر الي بعض حالات التشابه أو الخطأ الكامل.

لحل هذه المشكلة ، يتم إشراك الكيانات المسماة لتقليل نطاق النتائج وتحسين جودتها. الهدف من استخدام الكيانات المسماة في استخراج المعلومات هو تحسين أداء النتائج المعينة وتحسين جودتها. في هذه الأطروحة سوف نقدم نهجنا المقترح لنظام استخراج المعلومات ، من الشبكات الاجتماعية (Facebook) عن الحوادث من أجل استخراج الكيانات الزمنية والمكانية المسماة مع الأحداث.

الكلمات المفتاحية : استخراج المعلومات ، الكيان المسمى ، المعالجة التلقائية للغة الطبيعية.

Table des matières

Remerciement i

Dédicaces ii

Dédicaces iii

Résumé. iv

Table des matières vii

Liste des figures xii

Liste des tableaux xiii

Introduction générale 1

Chapitre I : Extraction de l'information 4

1. Introduction 4

2. Information 4

3. Extraction de l'information 4

3.1 Définition 4

3.2. Tâches d'extraction d'information 5

3.2.1 Identifier les entités nommées 5

3.2.2 Tâche d'extraire la relation 5

3.2.3 Résolution de référence 5

3.2.4 Motif de remplissage 6

3.2.5 Description des événements 6

3.3Extraction des Relations Sémantiques 6

3.3.1 Relations paradigmatiques 6

3.3.2 Relations syntagmatiques 6

3.4 Méthodes d'extraction d'information 7

3.4.1 Méthodes à base de règles 7
3.4.2 Méthodes d'apprentissage automatique 7
3.4.3 Méthodes hybrides 8
3.4.4 Méthodes à base d'ontologies 8
4. Architecture d'un système d'extraction d'information 9
4.1 Phase de prétraitements 9
4.2 Phase d'analyse linguistique 9
4.3 Phase d'installation des formulaires 9
5. Extraction d'Information et Recherche d'Information 10
5.1 Recherche d'Information (définition) 10
5.1.1 Modélisation des documents et des requêtes 11
5.1.2 Appariement 11
5.1.3 Production et mise en forme des résultats 11
5.2 Différences et liens avec l'Extraction d'Information 11
5.3 Combinaison de Recherche et Extraction d'information 12
5.3.1 Utiliser la Recherche d'Information en prétraitement de l'Extraction d'Information 12
5.3.2 Utiliser l'Extraction d'Information pour affiner les résultats d'un système de Recherche d'Information 13
6. Conclusion 13
Chapitre II Entités Nommées 14
1. Introduction 16
2. Entités nommées 16
2.1 Définition 16
2.2 Formes des entités nommées 16

Entités nommées simples :.....16
Entités nommées composées :.....16
2.3 Types des entités nommées 17
2.4 Classification des entités nommées.....17
2.4.1 Classification1 17
2.4.2 Classification 2 18
2.4.3 Classification 3 18
2.4.4 Classification 4 19
2.5 Utilisations des entités nommées 19
3. Traitement des entités nommées 19
3.1 Détection des entités nommées.....19
3.2 Identification des entités nommées 20
3.2.1 Indices internes (la structure des entités) 20
3.2.2 Indices externes (le contexte des entités).....21
3.3 Reconnaissance des entités nommées 21
3.3.1 Approches orientées connaissances 21
3.3.2 Approches orientées données 22
3.3.3 Approches hybrides 24
3.4 Extraction des entités nommées (EEN).....25
3.4.1 Extraction des informations du texte 26
3.4.2 Répondre automatiquement à des questions 26
4. Annotation manuelle de corpus des entités nommées 27
4.1 Guide d'annotation.....27
4.2 Outils d'annotation.....27

4.3 Mesures d'évaluation de la qualité des annotations 27
5. Métriques d'évaluation des entités nommées 27
6. Conclusion 28
Chapitre III : Conception d'un Système d'extraction d'information 29
1. Introduction 30
2. Travaux reliés 30
3. Notre Approche proposée 31
3.1 Indications spatiales 31
 3.1.1 Lieux dans le texte de la publication 31
 3.1.2 Coordonnées GPS jointes au poste 31
3.2 Indicateurs de temps 32
 3.2.1 Heure de publication 32
 3.2.2 Temps en poste 32
4. Architecture d'approche proposée 32
5. Phases de l'approche proposé 33
5.1 Collection de donnée 33
5.2 Segmentation des publications 34
5.3 Pré-Traitement 34
5.4 Analyse des informations 35
5.4.1 Tokeniser 35
5.4.2 Sentence Splitter 36
5.4.3 PartOfSpeech Tagger 36
5.4.4 Gazetteer 37
 Gazetter Personnel 37

- Gazetter Spatial 37
- Gazetter Evènement 37

5.4.5 Semantic Tagger 38

6. Conclusion 38

Chapitre IV : L'implémentation 39

1. Introduction 40

2. GATE 40

2.1 Définition 40

2.2 CREOLE 40

2.3 ANNIE 41

2.3.1 Découpeur de tokens (tokeniser) : 41

2.3.2 Gazetteer : 42

2.3.3 Séparateur de phrase (sentence splitter) : 42

2.3.4 D'un POS-Tagger : 42

2.3.5 D'un Named-Entity transducer (NE transducer): 42

2.4 Le formalisme JAPE 42

3. Application De L'approche Proposer 43

4. Conclusion 47

Conclusion générale 48

Bibliographie 49

Résumé

Liste des figures :

Figure 1.3 : Schéma générale de la tâche d'extraction d'information 5

Figure 1.4 : Architecture générale d'un système d'EI 10

Figure 1.5.1 : Architecture générale d'un système de recherche d'information 11

Figure 2.3.3.1 : Architecture générale de Nemesis.....22

Figure 2.3.3.2 : Types d'approches d'apprentissage automatique pour l'EI.....23

Figure 2.3.4.2 : Modèle du système de réponse automatique fourni des notes de concours 26

Figure 2.4 : Éléments d'un processus d'annotation.....27

Figure 3.4 : Architecture générale d'un système proposé.....33

Figure 4.2.1 : Exemple de l'interface dans GATE 41

Figure 4.2.4 : Exemple de règle avec le formalisme JAPE.....43

Figure 4.3.1 a : Extraction d'entités nommées temporelle avant les règles JAPE.....44

Figure 4.3.1 b : Extraction d'entités nommées temporelle après les règles JAPE.....44

Figure 4.3.2 a : Extraction d'entités nommées spatial avant les règles JAPE 45

Figure 4.3.2 b : Extraction d'entités nommées spatial après les règles JAPE 45

Figure 4.3.3 a : Extraction d'entités nommées d'évènement avant les règles JAPE 46

Figure 4.3.3 b : Extraction d'entités nommées d'évènement après les règles JAPE 46

Liste des tableaux :

Tableau 2.2.3 : Exemples d'entités nommées.....17

Tableau 2.4.1 : Classes d'EN d'après la campagne MUC 18

Tableau 2.2.4.3 : Classes d'EN 18

Tableau 2.2.4.4 : Classes des entités nommées 19

Tableau 2.3.3.3 : Résultats de certains systèmes de REN en termes de F-mesure.....25

Tableau 3.5.3 : Exemple d'application de la phase Prétraitement sur une publication 34

Introduction générale

Introduction générale

Face à l'augmentation de l'information disponible en ligne et à l'augmentation du nombre de documents électroniques rédigés en langage naturel, la classification automatique des textes est devenue de plus en plus une technologie clé pour la gestion intelligente interne de l'entreprise, plutôt qu'externe. La technologie d'extraction de l'information (IE) a été considérablement développée au cours de la dernière décennie : il consiste à extraire des informations précises des documents et à les construire sous une forme prédéfinie.

Lors d'extraction de l'information, il se produit également des représentations sémantiques externes au document. En se donnant la tâche de comprendre des parties extrêmement ciblées quant à la structure de l'information recherchée et quant aux formes linguistiques qui portent, mais non pas pour le texte dans son ensemble.

L'extraction automatique de l'information est un processus important et nécessaire au moment où les informations sont devenues des dossiers électroniques, que ce soit dans des sites commerciaux, d'actualités ou de médias sociaux. Cependant, l'extraction des informations sans utilisation de moyens auxiliaires ou spécification du contenu de ce que l'utilisateur veut avec précision, ne lui donne des résultats satisfaisants et efficaces, car il est possible de se retrouver dans certains cas de similitude ou d'erreur complète.

Pour résoudre ce problème, des entités nommées pour réduire la portée des résultats et améliorer leur qualité sont impliquées. L'objectif d'utiliser les entités nommées dans l'extraction de l'information est de renforcer la performance des résultats données et améliorer leur qualité.

Pour développer cet aspect nous avons subdivisé notre travail en quatre chapitres :

- Dans le premier chapitre, nous introduisons des notions générales sur l'extraction de l'information, les taches et l'extraction des relations sémantique. Nous définissons aussi les différentes méthodes d'extraction de l'information. Ensuite, nous expliquons une architecture générale d'un système d'extraction de l'information, une définition de la recherche de l'information et la différence entre la recherche et l'extraction de l'information ;

- Dans le deuxième chapitre, nous définissons les entités nommées, les formes essentielles, les types des entités, les différentes classifications et nous citons l'utilisation de ces entités nommées. Nous expliquons par la suit les traitements disponibles des entités nommées, la détection, l'identification, la reconnaissance avec des exemples illustratifs et l'extraction des entités

nommées. En fin, nous terminons par l'annotation manuelle de corpus des entités nommées et les métriques d'évaluation ;

- Dans le troisième chapitre, nous présentons des travaux reliés, en proposant et expliquant une approche en détaille. Ensuite, nous représentons une architecture d'approche proposée avec une explication de ses phases en donnant quelques exemples illustratifs ;

- Dans le quatrième chapitre, nous allons définir le GATE développer, CREOLE, ANNIE et le formalisme JAPE. Nous expliquons par la suite, les étapes de notre application sur GATE avec l'évaluation de chaque cas.

Chapitre I :

Extraction de l’information

1. Introduction

La recherche d'information est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées. La Recherche d'Information et l'Extraction d'Information poursuivent un but identique (trouver des informations dans un ensemble de textes) mais diffèrent dans leurs réponses et dans les moyens mis en œuvre. L'extraction d'information est une technologie récente et modern mais elle répond à un besoin très ancien : acquérir de la connaissance à partir du texte. L'extraction d'information devient nécessaire beaucoup plus avec l'essor considérable de la masse de document électronique (courrier électronique, Internet, réseaux sociaux, etc.).

Dans ce chapitre nous donnerons les concepts de base sur l'Extraction d'Information.

2. Information

L'information est un concept ayant plusieurs sens. Il est étroitement lié aux notions de contrainte, communication, contrôle, donnée, formulaire, instruction, connaissance, signification, perception et représentation. Au sens étymologique, l'information est ce qui donne une forme à l'esprit. Elle vient du verbe latin informer, qui signifie « donner forme à » ou « se former une idée de ».L'information désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire, elle utilise un code de signes porteurs de sens tels qu'un alphabet de lettres, une base de chiffres, des idéogrammes ou pictogrammes. Hors contexte, elle représente le véhicule des données comme dans la théorie de l'information et, hors support, elle représente un facteur d'organisation[1].

3. Extraction de l'information

3.1 Définition

L'extraction d'informations est une technologie moderne qui permet d'extraire et d'organiser automatiquement les informations pertinentes d'un ou plusieurs documents dans un langage naturel, et de les sauvegarder de manière organisée en créant une banque de données.

L'extraction s'effectue au moyen de formulaires prédéfini (Template). Ces modèles sont définis pour les informations requises à travers une structure prédéfinie, dans laquelle les entités,les relations entre elles et les événements impliqués dans ces entités sont décrits[2].

Ce n'est qu'avec l'avènement des MUC-7 (Message Understanding Conferences) que le domaine de l'exploration de l'information a fait des progrès significatifs[3].

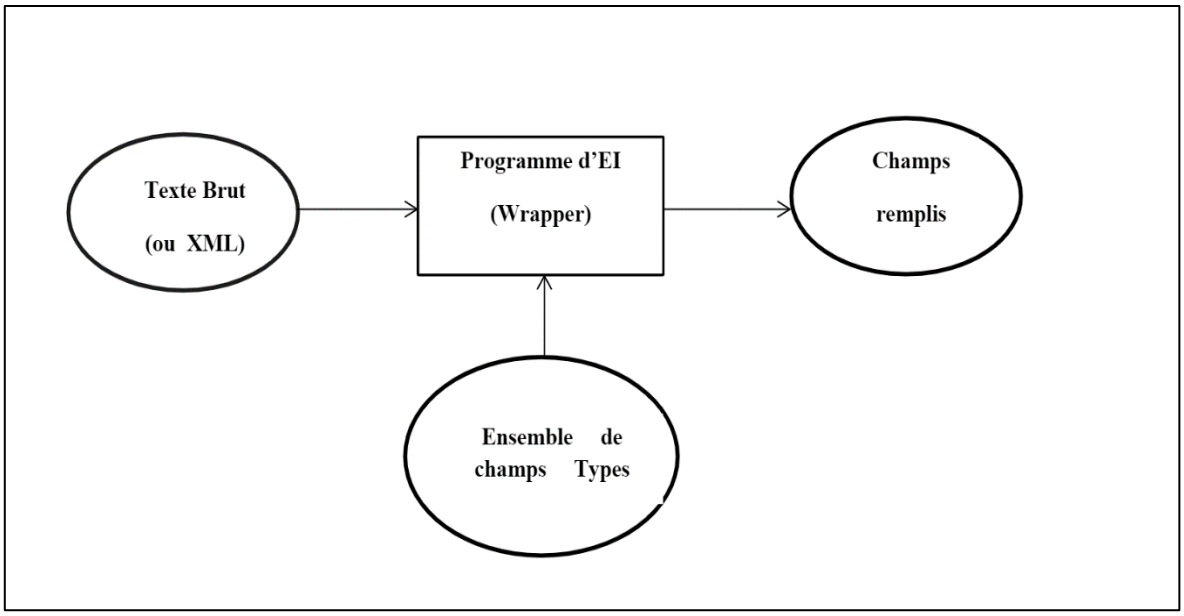


Figure 3 : Schéma générale de la tâche d'extraction d'information [4].

3.2. Tâches d'extraction d'information

L'évaluation de MUC-7 a spécifié cinq tâches. Ces tâches se concentrent sur l'extraction d'informations à partir d'enregistrements. Les tâches données ci-dessous sont adaptées de la définition de MUC-7 [3] **:**

3.2.1 Identifier les entités nommées

Exemple : noms d'organisations, de personnes et de lieux, dates et sommes d'argent. En développant autant que possible pour inclure toutes les choses abstraites, tangibles et nommées dans le texte.

3.2.2 Tâche d'extraire la relation

Elle consiste à détecter et à caractériser la relation sémantique entre les entités du texte.

3.2.3 Résolution de référence

Une tâche qui identifie les expressions linguistiques qui font référence à la même entité. La référence se compose de deux expressions linguistiques, des antécédents et des anagrammes. L'allitération est une expression dont l'interprétation dépend de l'interprétation de l'autre

expression (en l'associant à une entité réelle ou abstraite), tandis que les antécédents sont l'expression linguistique dont dépend l'allitération.

3.2.4 Motif de remplissage

Les entités, relations et événements à utiliser sont prédéfinis dans des structures définies par l'utilisateur appelées formulaires, chacune constituée d'un certain nombre d'attributs qui doivent être instanciés par un système pour extraire des informations lors du traitement de texte.

3.2.5 Description des événements

Afin de produire des relations possibles entre les modèles, les différents résultats sont liés, mais les événements liés à certains modèles résultant de l'information textuelle sont liés, ainsi une description des événements mentionnés dans les liens analysés est obtenue. Les types de relations et d'entités à l'aide de modèles de scénarios doivent être définis[3].

3.3Extraction des Relations Sémantiques

De nombreuses applications dans l'extraction d'information, la compréhension du langage naturel et la recherche d'information nécessitent une extraction de relations sémantiques entre les entités. L'extraction de relations est généralement précédée par la tâche de reconnaissance des entités. Les relations sémantiques sont regroupées en deux familles principales. Il s'agit des relations paradigmatiques et des relations syntagmatiques[3].

3.3.1 Relations paradigmatiques

Les relations paradigmatiques sont des relations fonctionnant principalement sur des concepts de la même classe. Habituellement, ce type de relations représente des relations hiérarchiques nommées liens verticaux. Elles sont utilisées pour organiser les concepts sous forme d'un arbre. Parmi ce type de relations, on peut citer les relations de l'antonymie, de la synonymie et de l'hyperonymie [3].

3.3.2 Relations syntagmatiques

Les relations syntagmatiques sont des liens sémantiques qui se produisent entre deux (ou plusieurs) unités linguistiques présentes dans une expression. Ils sont identifiés par l'étude des formes syntaxiques dans les textes, et par la présence d'un prédicat. Par exemple, on peut citer des relations spécifiques dans la sphère criminelle telles que : « X doit être attrapé par Y » ou « Y doit attraper X ». Les deux exemples précédents montrent qu'il existe une relation entre

l'infraction X et l'autorité chargée de l'arrêter Y. Il existe de nombreux autres exemples d'expressions pour extraire des relations temporelles entre les événements cliniques et les expressions temporelles [3].

3.4 Méthodes d'extraction d'information

Les systèmes d'extraction d'information reposent généralement sur deux approches principales : l'approche dite à base de règle et l'approche à base d'apprentissage automatique. Il existe des systèmes hybrides qui combinent ces deux approches et il y'a aussi des méthodes à base d'ontologie qui sont expliquées comme suit :

3.4.1 Méthodes à base de règles

Elles s'appuient sur l'utilisation des règles qui sont construites manuellement à partir des corpus d'un domaine donné pour identifier les entités.

Les méthodes à base de règles, également appelées des méthodes d'ingénierie de connaissances dans certaines sources, donnent de bons résultats, mais elles nécessitent cependant un grand effort et un temps considérable pour l'analyse des données et l'écriture des règles. Les systèmes basés sur la construction manuelle des règles sont plus intéressants dans des domaines fermés où l'intervention de l'être humain est à la fois essentielle et disponible.

Dans des domaines ouverts comme l'extraction d'opinions à partir de blogs, la souplesse des méthodes statistiques est plus appropriée [3].

3.4.2 Méthodes d'apprentissage automatique

Les méthodes d'apprentissage automatique ou encore les méthodes d'apprentissage statique, sont des techniques d'entraînement capables d'extraire automatiquement, par exemple, des entités étiquetées dans un jeu de données. Ces méthodes nécessitent un large corpus de texte étiquetés pour apprendre à identifier des entités.

Il existe différents types d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

3.4.2.1 Apprentissage supervisé

(Supervised Learning) est le paradigme d'apprentissage le plus populaire en Machine Learning et en Deep Learning. Comme son nom l'indique, cela consiste à superviser l'apprentissage de la machine en lui montrant des exemples (des données) de la tâche qu'elle doit réaliser. Les applications sont nombreuses : reconnaissance vocale, vision par ordinateur, régressions, classifications… La grande majorité des problèmes de Machine Learning et de Deep

Learning utilisent l'apprentissage supervisé. Il est donc essentiel de bien comprendre le fonctionnement de cette mécanique[5].

3.4.2.2 Apprentissage non supervisé

Dans le domaine de l'IA (Intelligence Artificielle), et plus précisément de laMachine Learning, la technique de l'apprentissage non supervisé consiste à entraîner des modèles, sans réaliser d'étiquetage manuel ou automatique des données au préalable. Les algorithmes doivent ici analyser et regrouper les données, sans aucune intervention humaine, en découvrant les patterns au sein des masses de données.

En cela, l'apprentissage non supervisé se distingue de l'apprentissage supervisé comme de l'apprentissage auto-supervisé (qui fait appel à des données étiquetées automatiquement)[6].

3.4.2.3Apprentissage par renforcement

L'apprentissage par renforcement, est un mode d'apprentissage statistique, inspiré de l'apprentissage humain et animal. Lorsqu'elles ont un résultat positif et induisent des récompenses, on conclut que ces expériences sont positives et qu'elles doivent être réitérées.

Inversement, si le résultat de l'expérience n'est pas concluant, on le mémorise pour ne plus faire la même erreur. Ainsi, dans le cadre d'un apprentissage par renforcement, l'IA veut maximiser ses récompenses[7].

3.4.3 Méthodes hybrides

Ces systèmes combinent les deux méthodes précédentes c'est-à-dire celles à base de règles et à base d'apprentissage automatique. Les méthodes hybrides utilisent des règles écrites à la main mais construisent aussi une partie de leurs règles à l'aide d'informations syntaxiques et d'informations sur le discours tirées de données d'entraînement grâce à des algorithmes d'apprentissage automatique.

La combinaison des deux méthodes peut se faire en commençant par exemple par la méthode manuelle à base de règles où l'hybridation est faite par transformation des résultats à base de règles puis considérer ces résultats comme des attributs pour le classifieur CRF, ou par la méthode d'apprentissage où l'hybridation est faite par utilisation de l'algorithme d'apprentissage des règles pour la détection des entités nommés puis les résultats obtenus sont validés par un ensemble de règles construites manuellement en calculant un score de validation pour chaque règle [3].

3.4.4 Méthodes à base d'ontologies

L'extraction d'informations basée sur l'ontologie (OBIE) est apparue comme un sous-domaine de l'extraction d'informations. Les ontologies sont utilisées par le processus d'extraction

d'informations et la sortie est généralement présentée via une ontologie. Il est à noter qu'une ontologie est définie comme une spécification formelle et explicite d'une conceptualisation partagée. Généralement, une ontologie est spécifiée pour un domaine particulier. Étant donné que l'extraction d'informations concerne essentiellement la tâche de récupération d'informations pour un domaine particulier, la spécification formelle et explicite des concepts de ce domaine via une ontologie peut être utile pour ce processus.

Par exemple, une ontologie géopolitique qui définit des concepts tels que pays, province et ville peut être utilisée pour guider le système d'extraction d'informations décrit précédemment. C'est l'idée générale derrière l'extraction d'informations basée sur l'ontologie [8].

4. Architecture d'un système d'extraction d'information

L'architecture générale présentée dans la figure 4 fait apparaître les principales composantes d'un système d'extraction d'information qui est divisée en trois phases [9] :

4.1 Phase de prétraitements

Consiste en un ensemble d'opérations «de surface » sur le matériau linguistique, permettant d'entreprendre l'analyse linguistique avec un texte « nettoyé » et « préparé ».

4.2 Phase d'analyse linguistique

Dans la seconde phase, l'analyse morphologique consiste à étiqueter (tagging) les mots selon leur classe (nom, verbe, adjectif…) et à repérer leur genre, nombre, personne, etc. Ensuite, L'analyse syntaxique doit produire (de manière plus ou moins exhaustive) les relations grammaticales : relations sujet-verbe ou verbe-COD, rattachements prépositionnels (compléments de nom ou compléments indirects des verbes), etc. L'analyse sémantique vise à construire à partir de chaque proposition une « représentation conceptuelle » sous forme d'expression logique ou de réseau sémantique. Enfin, l'analyse du discours doit établir les liens entre les différentes phrases, typiquement repérer les coréférences ou l'ordre temporel des énoncés. L'organisation de ces traitements, et bien sûr les méthodes linguistiques utilisées constituent des caractéristiques importantes des différents systèmes d'EI [9].

4.3 Phase d'installation des formulaires

À partir de cette représentation conceptuelle « abstraite », il s'agit de remplir les champs des formulaires. Ce qui suppose notamment d'affiner le calcul de l'identification des entités et des événements. Du point de vue des méthodes, une caractéristique de cette étape est d'être complètement. Orientée par le but, c'est à dire par la structure des formulaires, et de recourir à des inférences mettant en jeu des connaissances du domaine (« le monde financier », « le monde

de la route » …), alors que la phase d'analyse linguistique est plutôt orientée par la structure linguistique du texte.

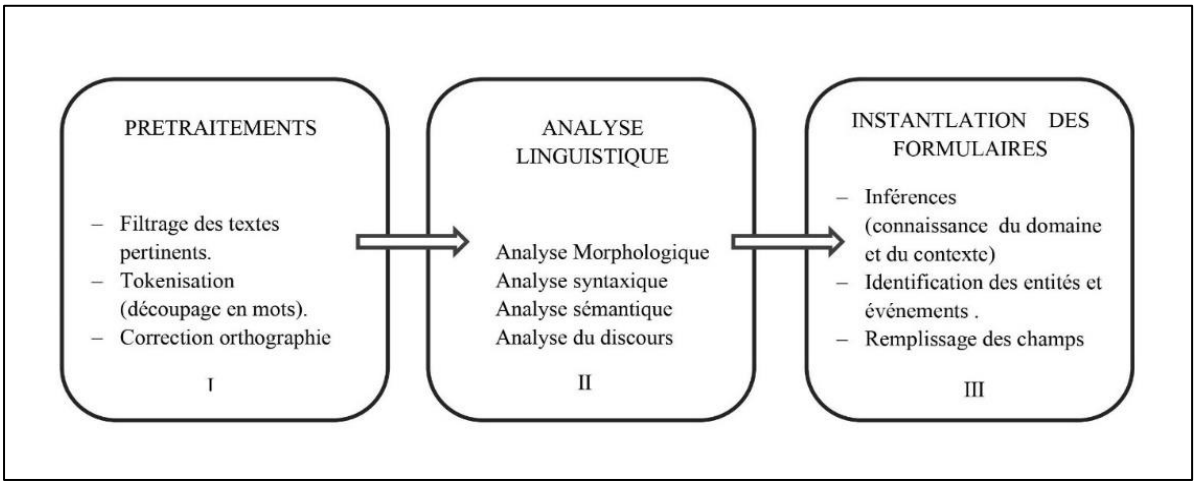


Figure 4 : Architecture générale d'un système d'EI [9].

5. Extraction d'Information et Recherche d'Information

5.1 Recherche d'Information (définition)

La recherche d'information est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus. Celui-ci est composé de documents d'une ou plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées. Les bases de données peuvent être relationnelles ou non structurées, telles celles mises en réseau par des liens hypertexte comme dans le World Wide Web, l'internet et les intranets. Le contenu des documents peut être du texte, des sons, ses images ou des données.

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui visent à représenter des documents dans le but d'en récupérer des informations, au moyen de la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information. Ainsi, la recherched'information est aujourd'hui un champ transdisciplinaire, intéressant même les sciences cognitives [11].

Classiquement, un processus de Recherche d'Information se déroule en trois phases comme présenter dans la figure 5.1:

5.1.1 Modélisation des documents et des requêtes

D'une part les documents de la collection de textes sont modélisés et d'autre part la requête de l'utilisateur est transformée en un modèle en accord avec la représentation choisie pour les documents.

5.1.2 Appariement

La modélisation de la requête est appariée avec celle des documents. Le but de cette étape est de déterminer la pertinence d'un document par rapport à la requête afin de sélectionner les documents les plus en adéquation avec celle-ci.

5.1.3 Production et mise en forme des résultats

En fonction de la tâche à effectuer ; renvoi de tous les documents ou d'une sélection de documents dans l'ordre décroissant de leur pertinence ; renvoi des documents de manière simple ou accompagnés d'un indice de pertinence ; mise en évidence de l'information via, par exemple, la mise en valeur de certains termes (coloration, soulignement, etc.) [2].

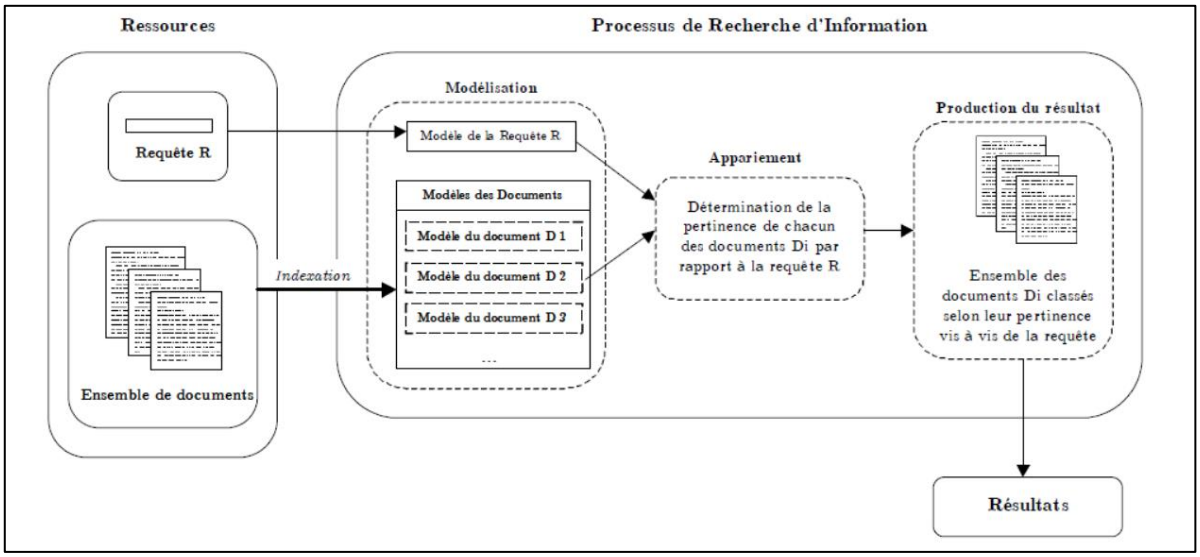


Figure5.1 : Architecture générale d'un système de recherche d'information[2].

5.2 Différences et liens avec l'Extraction d'Information

L'Extraction d'Information et la Recherche d'Information poursuivent un but identique (trouver des informations dans un ensemble de textes) mais diffèrent dans leurs réponses et dans les moyens mis en œuvre.

Leur différence fondamentale est la nature de l'information qu'ils renvoient. La Recherche d'Information modélise, de manière indépendante des informations à rechercher, les textes d'une

collection de documents, puis sélectionne ceux qui traitent d'un sujet donné (sujet exprimé par une requête), et les fournit à l'utilisateur. Un tel système est ouvert ; les requêtes ne sont pas fixées à priori. Les systèmes d'Extraction d'Information effectuent une analyse de documents bruts afin d'en extraire uniquement des informations précises qui intéresseront l'utilisateur, ces informations étant spécifiées à priori (il n'y a pas de requête en entrée du système)[2].

Ces deux méthodes utilisent des techniques différentes pour des raisons aussi bien pratiques qu'historiques. Les travaux sur les systèmes de Recherche d'Information ont été influencés par la théorie de l'information, les théories probabilistes et statistiques, alors que l'Extraction d'Information est issue de recherches en linguistique computationnelle et en TALN (**T**raitement **A**utomatique du **L**angage **N**aturel). Les systèmes de Recherche d'Information voient généralement le texte comme un ensemble non structuré de mots. Dans le cas contraire, les systèmes d'Extraction d'Information doivent s'intéresser à la structure grammaticale et aux propriétés syntagmatiques du texte pour éviter d'importantes erreurs de sens[2].

L'utilisation d'un système d'Extraction d'Information plutôt que de Recherche d'Information pour collecter des informations à partir de textes présente des avantages mais également des inconvénients : d'une part ils sont plus difficiles à mettre en œuvre et sont souvent liés à un domaine de connaissance particulier, ce qui les rend difficilement adaptables à d'autres domaines, et d'autre part les résultats renvoyés sont moins précis que ceux donnés par des lecteurs humains. Mais dans le cas de larges corpus, l'Extraction d'Information apparaît comme potentiellement beaucoup plus efficace que la Recherche d'Information en raison de la difficulté et du coût de la tâche que constitue alors la lecture et l'analyse manuelle de la masse de documents renvoyés par un système de Recherche d'Information, ces systèmes ne se révélant généralement pas assez discriminants. Les systèmes d'Extraction d'Information possèdent également l'atout de pouvoir extraire des faits précis et d'alimenter d'autres applications de traitement de l'information (bases de données, index) [2].

5.3 Combinaison de Recherche et Extraction d'information

Il existe plusieurs moyens de combiner ces deux systèmes :

5.3.1 Utiliser la Recherche d'Information en prétraitement de l'Extraction d'Information

Face à un très large volume de textes, elle peut fournir à un système d'Extraction d'Information une sous-collection ne regroupant que les documents les plus pertinents. Il existe plusieurs projets dans ce sens comme par exemple le programme TIPSTER.Ces projets sont encouragés par la masse de plus en plus grande de documents disponibles sur Internet et la

difficulté de faire traiter directement de grandes quantités de données textuelles par les systèmes d'Extraction d'Information, les temps et coûts de traitement devenant prohibitifs[2].

5.3.2 Utiliser l'Extraction d'Information pour affiner les résultats d'un système de Recherche d'Information

En améliorant la phase de modélisation des documents ; les informations extraites de chaque document via un formulaire par un processus d'Extraction d'Information peuvent être utilisées pour créer un index qui modélise le document. Par exemple, le projet Navilex se sert de formulaires d'Extraction d'Information pour indexer des documents légaux[2].

5.3.3 Compléter les approches classiques de recherche d'information par les techniques d'extraction d'information

Pour catégoriser, filtrer et ordonner les documents en fonction de leur pertinence. Un exemple de cette méthode est l'adaptation du système d'Extraction d'Information FASTUS par John Bear et ses collègues. Ce système d'Extraction d'Information attribue une note de pertinence à chacun des documents renvoyés par un système de Recherche d'Information pour un sujet donné, de manière à reclasser ceux-ci en plaçant en tête de liste les documents ayant les meilleures notes. Chaque note est donnée en fonction des informations extraites du texte par FASTUS[2].

6. Conclusion

Dans ce chapitre nous avons présenté les concepts de base de l'extraction d'information. Nous avons mentionné la définition de l'extraction d'information, les tâches et l'extraction des relations sémantiques, les méthodes et l'architecture d'un système d'extraction d'information. Nous avons également donné une petite comparaison entre la recherche d'information et l'extraction d'information. Principalement, l'extraction d'information est basée sur plusieurs descripteurs, parmi eux les entités nommées qui sont des unités textuelles particulières. Il s'agit de reconnaitre les noms propres (personne, lieu) mais aussi les expressions temporelles (dates, durées) et les noms de quantités (mesures, monétaire). Notre travail s'insère dans cette direction, à savoir l'utilisation des entités nommées que nous traiterons dans le chapitre suivant.

Chapitre II

Entités Nommées

1. Introduction

Dans le chapitre précédent nous avons défini les notions de base de l'extraction d'information. L'extraction d'information doit utiliser des indicateurs pour préciser l'extraction d'information comme les entités nommées. Les entités nommées constituent un champ de recherche très actif depuis plusieurs années. Elles sont depuis longtemps considérées comme un point central dans de multiples applications mettant en jeu des notions comme la compréhension, la recherche sémantique, etc.

Dans ce chapitre, nous allons exposer un ensemble de concepts de base qui sont nécessaires pour la compréhension du concept des entités nommées.

2. Entités nommées

2.1 Définition

Le concept d'Entité Nommée est apparu dans les années 90 au cours de la sixième conférence MUC (Message Understanding Conference), bien que l'EN (Entité Nommée) n'ait pas de définition standard, certains chercheurs ont proposé différentes définitions pour ce concept :

- « Les ENs sont des types d'unités lexicales particuliers quifont référence à une entité du monde concret dans certains domainesspécifiquesnotamment humains, sociaux, politiques, économiques ou géographiques et qui ont unnom (typiquement un nom propre ou un acronyme) »
- « L'ENest un mot ou un groupe de motsdésignant une personne, une organisation ou entreprise, un lieu, une date ou encore uneexpression numérique. » [11].

2.2 Formes des entités nommées

Il y a deux formes d'EN ; les ENs simples et les ENs composées, chaque forme a un traitement différent :

Entités nommées simples :

Est une EN qui est composée d'un seul mot, comme les noms de lieu « Mila » et «Algérie »ou le nom de personne « Faize ».

Entités nommées composées :

Est une EN qui est composée de deux ou plusieurs mots, comme par exemple le nom de personne « AbdelHafidBoussof » et le nom delieu « Afrique du Sud » [12].

2.3 Types des entités nommées

Les entités nommées y sont réparties en 7 types primaires et 32 sous-types, dont voici la liste de sept types :

- **Personnes :** personnes individuelles, personnes collectives.
- **Lieux :** lieux administratifs (ville, région, nations, surannations), lieuxphysiques (géographiques,hydrologiques, astrologiques).
- **Organisations :** entreprises, administrations.
- **Temps :** dates (absolues ou relatives) et horaires (absolus ou relatifs).
- **Montants :** quantités, durées.
- **Produits :** objets manufacturés, œuvres artistiques, œuvres médiatiques, produits financiers, logiciels, récompenses, voies, doctrines, lois.
- **Fonctions :** fonctions individuelles, fonctions collectives.

Cette typologie ajoute donc aux entités nommées traditionnelles les produits et les fonctions. Elle ajoute une granularité supplémentaire (sous-types) aux types principaux (ou primaires)[13].

Table 2.3 : Exemples d'entités nommées

Entité nommée	Type
Barack Obama	Nom de personne
USA	Nom de lieu
29 Décembre 2015	Expression temporelle
UNICEF	Nom d'organisation
18%	Expression Pourcentage

2.4 Classification des entités nommées

2.4.1 Classification1

Les conférences MUC ont été organisées et financées par DARPA (**D**efense **A**dvanced **R**esearch **P**rojects **A**gency) et NOSC(**N**aval **O**cean**S**ystem **C**enter) dans le but d'encourager la recherche et le développement ;les conférences MUC ont eu lieu pour traiter du problème de l'El. Les données des participants étaient sous forme de messages et elles étaient évaluées sur des sujets particuliers. La tâche d'extraction des ENs a été introduite dans la sixième conférence (MUC-6), puis une classification en trois classes a été proposée dans la dernière conférence MUC -7 [11].

Table 2.4.1 : Classes d'EN d'après la campagne MUC [11]

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

2.4.2 Classification 2

CoNLL 17 est la conférence annuelle organisée par SIGNLL (**S**pecial **I**nterest **G**roup on **N**atural **L**anguage **L**eaming). C'est une conférence internationale sur le langage naturel et 1' apprentissage machine. En 2002, le sujet principal de cette conférence était la reconnaissance des ENs, et une classification des ENs en quatre classes a été proposée. Ces classes sont les trois sous-classes de la classe ENAMEX (**E**ntity **Nam**e **Ex**traction) de MUC plus une quatrième classe qui regroupe toutes les entités qui n'appartiennent pas aux trois classes précédentes [14].

2.4.3 Classification 3

ESTER(Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophonique) est une campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques en langue française. La reconnaissance des ENs est l'une des tâches évaluées dans cette campagne, et uneclassification en sept types d'EN a été proposée[15].

Table 2.4.3 : Classes d'EN [11]

Classe	**Description**
Personne	Personne, humaine, animal, fonction et civilité.
Fonction	Politique, militaire, administrative, religieuse et aristocratique.
Organisation	Politique, éducative, commerciale, non commerciale, divertissement et média et géo administrative.
Lieu	Lieu géographique naturel, région administrative, axe de circulation, adresse (adresse postale, numéro de téléphone et fax, adresse électronique) et construction humaine.
Production	Humaine Moyen de transport, récompense, œuvre artistique et production documentaire.
Date et heure	Date, heure.
Montant	Age, durée, température, longueur, aire et surface, volume, poids, vitesse et valeur monétaire

2.4.4 Classification 4

Une classificationdes ENs a été introduite en neuf classes. Cette classification a été obtenue à partir d'une étude du corpus « Wall Street Journal » [11].

Table 2.4.4 : Classes des entités nommées [11]

Classe	**Description**
Entités	Géographiques Ville, port, aéroport, île, comté, province, pays, continent, région, mer et fleuves
Affiliation	Religion, nationalité
Organisation	Entreprise, types d'entreprises, administration, administration gouvernementale
Humain	Personne, fonction
Document	Document
Equipement	Logiciel, matériel, machine
Scientifique	Maladie, drogue, médicament.
Temporelle	Date, heure
Divers	Autres types d'EN

2.5 Utilisations des entités nommées

Les ENs sont utiles pour le développement des systèmes de questions/réponses, les résumés automatiques, la recherche d'information, la traduction automatique (TA), le Web sémantique, et la bio-informatique. Les ENs sont utilisées aussi pour la réduction du taux de Mots Hors Vocabulaire (MHV)[11].

3. Traitement des entités nommées

3.1 Détection des entités nommées

Certains moteurs de recherche comme Google, tentent maintenant de fournir directement, pour les requêtes auxquelles ce genre de procédé s'applique, l'information demandée par l'usager par le biais de sa requête. Pour être capable de bien répondre à ce besoin, ce genre de système doit être capable de découvrir les entités nommées contenues dans la requête émise par l'usager et d'en identifier leurs types. Cette tâche est habituellement référée par le terme « NamedEntity Recognition and Classification (NERC) ». En général, les types d'entités nommées à découvrir et à classifier diffèrent d'un problème à l'autre selon les besoins du système.

Dans le contexte de la recherche Web, un système de NERC (North-American Electric Reliability Corporation) peut être capable d'identifier des mentions d'entités nommées telles que, par exemple, des noms d'artistes, d'albums, de chansons, d'athlètes, de restaurants, de compagnies diverses ou de villes et villages.

De façon générale, deux grands types de technique peuvent être utilisés pour la conception d'un système de NERC. Le premier type regroupe les techniques basées sur des ensembles de règles grammaticales et syntaxiques qui ont été construites manuellement pour chaque type d'entité nommée considéré. Le deuxième type regroupe les techniques basées sur des modelés statistiques (Modelé de Markov Cache (MMC), Maximum Entropy Model (MEM) ou encore Conditional Random Field (CRF)) qui seront entraines avec un ensemble de textes dans lesquels les entités nommées à détecter ont déjà été identifiées et classées. Les modèles statistiques peuvent être utilisés pour identifier et classifier les entités nommées présentes dans un segment de texte donné.

Il existe également des systèmes hybrides qui utilisent à la fois un ensemble de règles grammaticales et syntaxiques et un ou des modelé(s) statistique(s) pour effectuer cette tâche [13].

3.2 Identification des 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 divisés en deux types, internes et externes selon :

3.2.1 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 [16].

La majuscule est une marque typographique tel que chaque mot (ou séquence de mots) commence par une lettre majuscule est considéré comme entité nommée. McDonalds'appuie 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 (Université de Sciences et Technologies Oran))[16].

3.2.2 Indices externes (le contexte des entités)

Ce sont le contexte dans lequel une entité nommée apparaît dans la phrase. Les indices externes 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) [16].

3.3 Reconnaissance des entités nommées

La plupart des systèmes de REN utilisent soit des approches orientées connaissances soit des approches orientées données. Les systèmes orientés connaissances sont fondés sur des lexiques (listes de prénom, de pays, etc.) et sur un ensemble de règles de réécriture. D'un autre côté, les systèmes orientés données sont basés sur un modèle appris à partir d'un corpus préalablement annoté. Afin de profiter des avantages de ces deux approches, d'autres systèmes combinent des techniques d'apprentissage automatique et des règles produites manuellement[17].

3.3.1 Approches orientées connaissances

Pour les approches orientées connaissances, les règles d'extraction sont produites manuellement par des experts en se reposant essentiellement sur des descriptions linguistiques, des indices et des dictionnaires de noms propres et de mots déclencheurs. Ces règles prennent la forme de patrons d'extraction permettant de repérer et de classifier les entités nommées (figure 3.3.1).

Exemple : le mot déclencheur « Monsieur » précède un mot inconnu commençant par une majuscule, alors le syntagme peut être étiqueté comme un nom de personne. Les systèmes orientés connaissances permettent d'obtenir de bons résultats sur des textes bien formés[18].

Exemple : Nemesis (Fourour 2002) **: un système orienté connaissances de REN pour le français.**

Nemesis (Fourour 2002) est un système qui permet la délimitation et la catégorisation des entités nommées développées pour le français et pour du texte bien formé. Il se base

essentiellement sur les indices internes et externes définis par McDonald (1996). L'architecture de Nemesis se compose principalement de trois modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles[17].

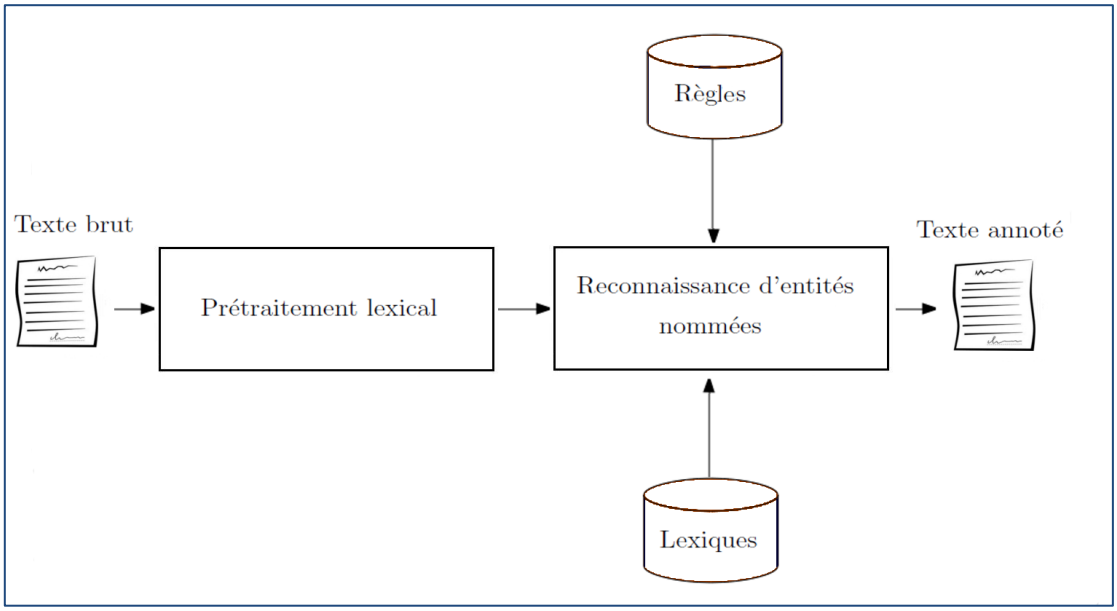


Figure 3.3.1 : Architecture générale de Nemesis [17].

3.3.2 Approches orientées données

L'approche apprentissage automatique (orientée donnée) 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. Lorsque 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[16].

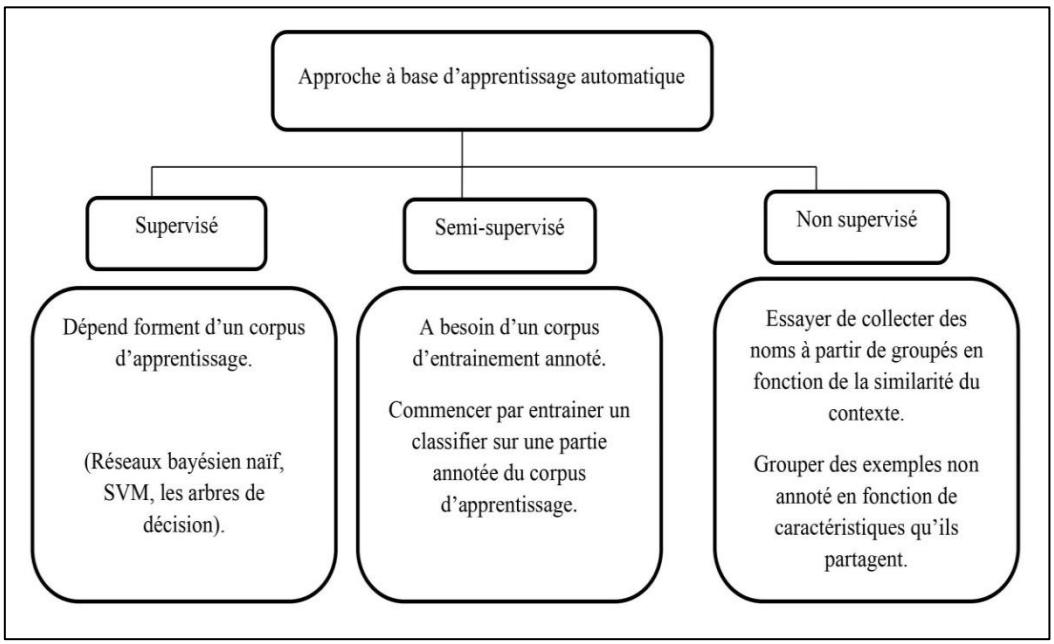


Figure 3.3.2: Types d'approches d'apprentissage automatique pour l'EI [16].

Les types d'apprentissage machine appliqués en particulier dans le domaine de reconnaissance des ENs sont présentés comme suit :

3.3.2.1 Apprentissage supervisé

La technique d'apprentissage supervisé consiste à utiliser un corpus préalablement annoté pour réaliser la tâche d'extraction des ENs. Elle se déroule en deux étapes : la première étape est l'apprentissage qui consiste à construire un processus automatique d'extraction des entités nommées pour un corpus d'entraînement annoté. La deuxième étape consiste à généraliser le processus afin de concevoir des règles permettant d'extraire les entités nommées dans de nouveaux documents [16].

3.3.2.2 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, ce qui rend le coût d'apprentissage de cette technique élevé [16].

3.3.2.3 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 [16].

Exemple : les systèmes d'apprentissage automatique

Des auteurs ont proposé un système NER(**N**ucleotide **E**xcision **R**epair) arabe basé sur le ANN (Arabic Neural Networks) qui vise à classer les NE arabes. Leur système se compose de trois étapes. Dans la première étape, le texte est prétraité afin de nettoyer les données collectées. Dans la deuxième étape, les lettres arabes sont converties en alphabet romain. Enfin, dans la troisième étape. La précision de leur système a atteint 92,8%. 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[16].

3.3.3 Approches hybrides

Certains systèmes tirent profit des avantages respectifs des méthodes orientées connaissances et celles orientées données. Les règles sontsoit apprises automatiquement puis révisées manuellement soit écrites manuellement puis corrigées et améliorées automatiquement. Les systèmes hybrides utilisent conjointementdes techniques orientées connaissances et des techniques orientées données. Un ensemble de règles est écrites par un expert puis enrichi automatiquement en utilisant des techniques d'apprentissage, ce qui permet d'obtenir progressivement une meilleure couverture [18].

Exemple : LTG ; un système hybride de REN pour l'Anglais

Andrei et al. (1998) présentent le système LTG (Language TechnologyGroup) lors de la compagne d'évaluation MUC-7. Ce système hybride aobtenu les meilleurs résultats lors de cette compétition. L'extraction des entités EnAmex par LTG est faite comme suit :

a) Passage des règles les plus sûres (sure-firerules)

Ces règles sont appliquéesseulement lorsque les indices internes et externes permettent declasser le candidat sans ambiguïté. Elles se présentent sous forme d'une liste de mots déclencheurs et un ensemble de règles contextuellesutilisant un étiquetage en parties du discours[17].

b)Première reconnaissance partielle (partial match 1)

Une fois les règlesles plus sûres appliquées, le système génère des variantes d'entitésnommées déjà reconnues en changeant l'ordre des mots ou en ensupprimant. Ensuite, un algorithme probabiliste fondé sur le modèlede maximisation de l'entropie est utilisé pourl'étiquetage des nomspropres[17].

c)Passage des règles plus lâches (Rule relaxation)

Des règles plus lâchesen termes de contraintes contextuelles sont appliquées. Cette étapepermet aussi de résoudre le problème des conjonctions et celui desentités en début de phrases[17].

d)Deuxième reconnaissance partielle (partial match 2)

Cette reconnaissancepartielle annote les noms propres en utilisant le modèle d'entropiemaximale [17].

e)Traitement des titres des articles (title assignment)

Des règles et unalgorithme probabiliste sont utilisés pour la désambiguïsation desentités nommées situées dans les titres car ces derniers sont entièrementécrits en majuscules[17].

Les résultats obtenus par ce système sont affichés dans la table 3.3.3.

Table 3.3.3 :Résultats de certains systèmes de REN (Reconnaissance d'Entité Nommée) sur le corpus MUC-7 en termes de F-mesure [17].

Système	**Approche**	**F-mesure (%)**
LTG	Hybride	93,39
MENE	Orienté données (supervisé)	92,20
IsoQuest	Orienté connaissances	91,60
BALIE	Orienté données (semi-supervisé)	77,71

3.4 Extraction des entités nommées (EEN)

L'Extraction des entités nommées (EEN) est la combinaison des méthodes pour indiquer des entités nommées dans des documents et les utilisé dans des objets différents. L'EEN a plusieurs applications dans la réalité. Dans cette partie, quelques applications dans des systèmes de traitement des informations automatiques par l'ordinateur sont présentées.

3.4.1 Extraction des informations du texte

Lors de l'extraction des informations dans des textes libres, à titre d'exemples les informations de la personne (nom, adresse, numéro de téléphone, lieu domicilié), l'utilisateur doit lire des documents et noter toutes ces informations dans un tableau. Mais le travail est morne notamment avec les grandes données. Le système d'extraction des entités nommées peut tirer automatiquement ces informations. La phase d'extraction d'entités nommées consiste à mettre en place un système de détection et de typage des entités d'intérêt dans un texte [19].

3.4.2 Répondre automatiquement à des questions

L'EEN (Extraction d'Entité Nommée) joue un rôle important dans le système de réponse automatique. Le système peut savoir le nom de la personne et fournit des services correspondants.

Nous exprimons ci-dessous le système de réponse des notes dans le concours à l'Université. Pour savoir les notes du concours, des personnes doivent préparer un message avec le format fixé. Le message contient le numéro d'identité. Le système obtient ce numéro d'identité et trouve des notes dans la base de données. Mais si l'utilisateur ne connaît pas le format par exemple, il a seulement le nom et la date de naissance ou autres informations, alors, il est impossible de réaliser la recherche[11].

Dans ce système, l'utilisateur peut envoyer un message avec n'importe quel format. Il contient des informations nécessaires (**par ex.** le nom, la date de naissance, etc.). Le système reconnaît des entités nommées et les utilise pour trouver des notes correspondantes. Cependant, l'EEN est une partie du système. Le système peut intégrer le module de reconnaissance des paroles ou autre pour améliorer des fonctions[19].

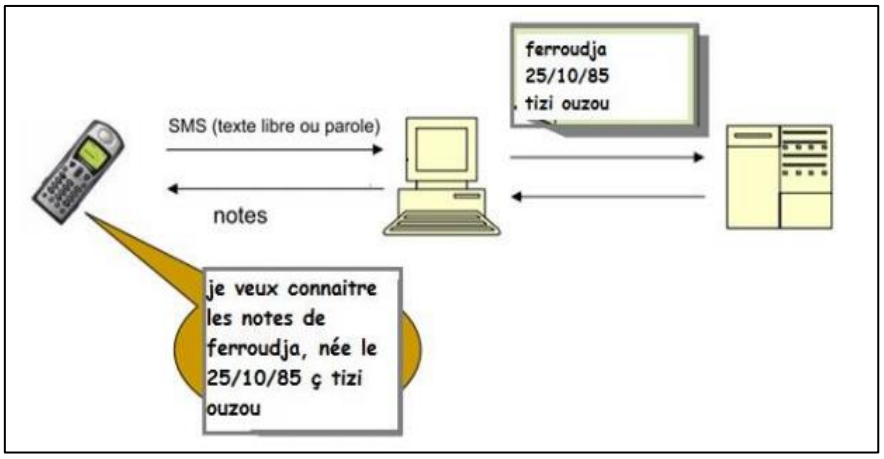


Figure 3.4.2 : Modèle du système de réponse automatique fourni des notes de concours [19].

4. Annotation manuelle de corpus des entités nommées

L'annotation de corpus est une thématique très active qui fait l'objet de nombreux travaux. Effectivement, celle-ci peut être plus ou moins assistée, guidée, automatisée. De plus, le travail nécessite une grande rigueur et beaucoup de préparation afin d'obtenir une annotation fiable. Dans l'essentiel, trois éléments paraissent indispensables :

4.1 Guide d'annotation

Il détaille les expressions linguistiques à annoter, selon des critères qui doivent laisser aussi peu de latitude que possible à la personne qui réalisera l'annotation.

4.2 Outils d'annotation

Logiciels servant à annoter, dont les interfaces doivent faciliter, mais sans biaiser, le travail de l'annotateur, en incluant éventuellement une phase de pré-annotation automatique.

4.3 Mesures d'évaluation de la qualité des annotations

Tests prévus afin de confirmer la fiabilité d'une annotation (accord inter-annotateurs) sur les parties annotées par plusieurs personnes (annotation croisée)[13].

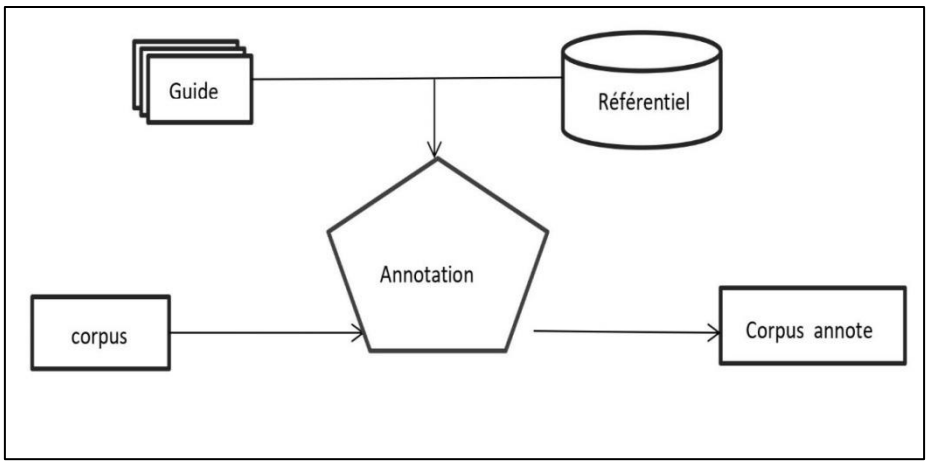


Figure 4 : Éléments d'un processus d'annotation [13]

5. Métriques d'évaluation des entités nommées

- Le rappel, la précision et la F-mesure sont des mesures largement utilisées dans les évaluations en TALN ;
- La précision est le pourcentage des résultats corrects parmi les résultats obtenus ;
- Le rappel est le pourcentage des résultats corrects parmi les résultats qu'on doit trouver ;

- Pour le domaine de l'extraction des ENs, les taux de la précision et du rappel sont calculés selon les formules suivantes :

$$\text{Précision} = \frac{\text{Nombre d'ENScorrectemnt reconnues}}{\text{Nombre d'ENS reconnues}}$$

$$\text{Rappel} = \frac{\text{Nombre d'ENS correctement reconnues}}{\text{Nombre d'ENS dans le corpus}}$$

La F-mesure est la combinaison de la précision et du rappel et leur pondération. La formule de la F-mesure est [11] :

$$\text{F\_mesure} = \frac{2(\text{précision} * \text{rappel})}{\text{précision} + \text{rappel}}$$

6. Conclusion

Nous avons défini les entités nommées et nous avons également mentionné les différentes formes, les types de façon générale, les différentes classifications des entités nommées basant sur plusieurs ressources et l'utilisation des entités nommées.

Nous avons précisé les traitements des entités nommées : la détection, la reconnaissance, l'extraction, et ainsi de suite. Enfin, nous terminons par les métriques et les compagnes d'évaluation qui vont nous aider à mieux comprendre les chapitres suivants.

Après avoir introduit à la fois l'extraction d'information et les entités nommées, nous sommes prêts pour commencer l'étape de conception d'un système d'extraction d'information utilisant les entités nommées

Chapitre III :

Conception d’un Système d’extraction d’information

1. Introduction

Dans les chapitres précédents nous avons défini l'extraction d'information, et nous avons détaillé ces méthodes et ces tâches avec une architecture générale d'un système d'extraction d'information. Ensuite, nous avons présenté précisément les entités nommées, les formes, les types et les traitements sur les entités nommées. Maintenant, dans ce chapitre nous présenterons une architecture générale de notre approche proposée d'un système d'extraction d'information, des réseaux sociaux (Facebook) sur les catastrophes naturelles et les accidents.

2. Etat de l'art

Dans la littérature en général et dans les médias sociaux en particulier, l'intérêt pour les relations internationales en langue arabe est dans une large mesure limitée par rapport à ce qui se fait dans d'autres langues.

Ces derniers jours, les médias sociaux ont joué un rôle important dans l'identification et l'extraction de diverses informations telles que les informations temporelles et spatiales.

Pour une utilisation dans le trafic, en 2013, Daly, Lecue & Bicer ont développé une application appelée Dub-STAR (Semantic Traffic Annotator de Dublin et Reasoner) qui combine à la fois les données de source de la ville et les données dynamiques des médias sociaux pour fournir des interprétations en temps réel des conditions de circulation.

Twitter a été utilisé pour appliquer la théorie d'extraction de l'informations spatiales et temporelles pour extraire des événements liés à la criminalité des tweets arabes sur la base d'informations spatiales et temporelles, ainsi que des informations de la page personnelle pour comparer les résultats et obtenir les plus précis.

Les informations sont extraites dans l'environnement Java à l'aide de l'interface de programmation d'application Gateway. Des techniques de traitement du langage naturel ont été appliquées après l'application des filtres, puis trois dictionnaires géographiques de base ont été créés (dictionnaires d'événements, spatiaux et temporels), et enfin un ensemble d'algorithmes développés et implémentés en Java ont été implémentés.

L'idée d'utiliser Twitter comme source valable d'informations géographiques est en cours de l'étude ; de nombreuses études se sont focalisées sur la découverte de risques naturels avec Twitter. En 2010, Sakaki, Okazaki et Matsuo ont tenté de découvrir les emplacements des tremblements de terre au Japon en temps réel à partir de Messages Twitter. Un modèle probabiliste spatio-temporel a été développé ; un ensemble de caractéristiques telles que les

mots-clés (tremblement de terre), puis ils ont essayé de centrer et de suivre l'emplacement de l'événement à l'aide de filtres bayésiens tels que les filtres de Kelman et de particules. [20]

3. Notre Approche proposée

Les réseaux sociaux en générale (Facebook)sont des plates-formes importantes pour échanger des informations et obtenir des nouvelles du monde, à cet égard, cette recherche s'est concentrée sur Facebook en premier lieu car il est considéré comme la première plate-forme qui touche la catégorie des jeunes.

Ce travail étudie les publications des réseaux sociaux basant sur Facebook pour extraire des informations personnelles et des indicateurs spatio-temporels pour les victimes des catastrophes naturelles ou des accidents dans les évènements.

3.1 Indications spatiales

Les indicateurs spatiaux sont des informations qui nous aident à déterminer l'emplacement d'une publication Facebook et les indicateurs spatiaux sont divisés en plusieurs types. Ils peuvent indiquer : l'emplacement dans le profil, les coordonnées GPS (Global Positioning System) attachées aux appareils mobiles, les liens vers des sites Web, etc. Dans ce travail nous fondons les informations quantitatives suivantes :

- **3.1.1 Lieux dans le texte de la publication**

Le texte du message est un message limité, la plus grande limite est 63206 caractères, il peut être écrit en langage naturel, mais la plupart des messages sont avec des symboles, un dialecte et des abréviations pour Facebook, donc la tâche d'extraire des informations spatiales est généralement difficile car elle est ambigüe, et des pseudonymes et des abréviations au lieu d'utiliser le nom d'origine.

- **3.1.2 Coordonnées GPS jointes au poste**

Dans les publications Facebook, l'utilisateur peut ajouter au texte de la publication : son état psychologique, une référence à des amis, ajouter un lieu, répondre à une question, etc. Par conséquent, la présence des coordonnées jointes à la publication est facultative et l'utilisateur peut l'inclure et peut s'en passer.

3.2 Indicateurs de temps

Comme les indicateurs spatiaux, les indicateurs temporels permettent de connaître l'heure réelle de la survenance de l'acte délictueux afin de faciliter et de réduire le champ de recherche, en fonction : de l'heure de publication et des indicateurs temporels mentionnés dans le texte de la publication.

➢ 3.2.1 Heure de publication

L'horodatage (heure de publication) est le moment où un utilisateur publie une publication sur sa page Facebook. Dans ce travail, nous prenons en compte l'horodatage pour nous aider à trouver l'heure exacte.

➢ 3.2.2 Temps en poste

Les indicateurs de temps peuvent être exprimés dans le texte de la publication. Ils peuvent être énoncés explicitement, tels que : jours, mois, saisons, etc., ou implicitement mentionnés dans le texte, tels que : six ans, trois jours, etc., et certains d'entre eux sont plus ambigus, tels que : les derniers matins, ce jour-là, etc. Cette ambiguïté rend l'extraction des informations temporelles plus difficile et nécessite le développement de nouveaux mécanismes et techniques.

4. Architecture d'approche proposée

L'architecture suivante représente la conception détaillée de notre système. L'approche proposée construit un prototype qui extrait les entités spatiale et temporelle, des individus(personnes) et des évènements dans les réseaux sociaux. Le système proposé se compose de plusieurs phases ;

- **Phase de préparation :** les données sont collectées à partir des réseaux sociaux (Facebook).
- **Phase de la segmentation :** les statuts sont choisis selon les conditions posées.
- **Phase de prétraitement** : dans cette phase le bruit est réduit, et tous les caractères non alphabétiques et signes, et les liens, etc. sont effacés.
- **Phase d'analyse des informations :** on applique les techniques NLP (Natural Language Processing) et les Gazetteers et on extrait les informations nécessaires ; comme le montre la figure suivante :

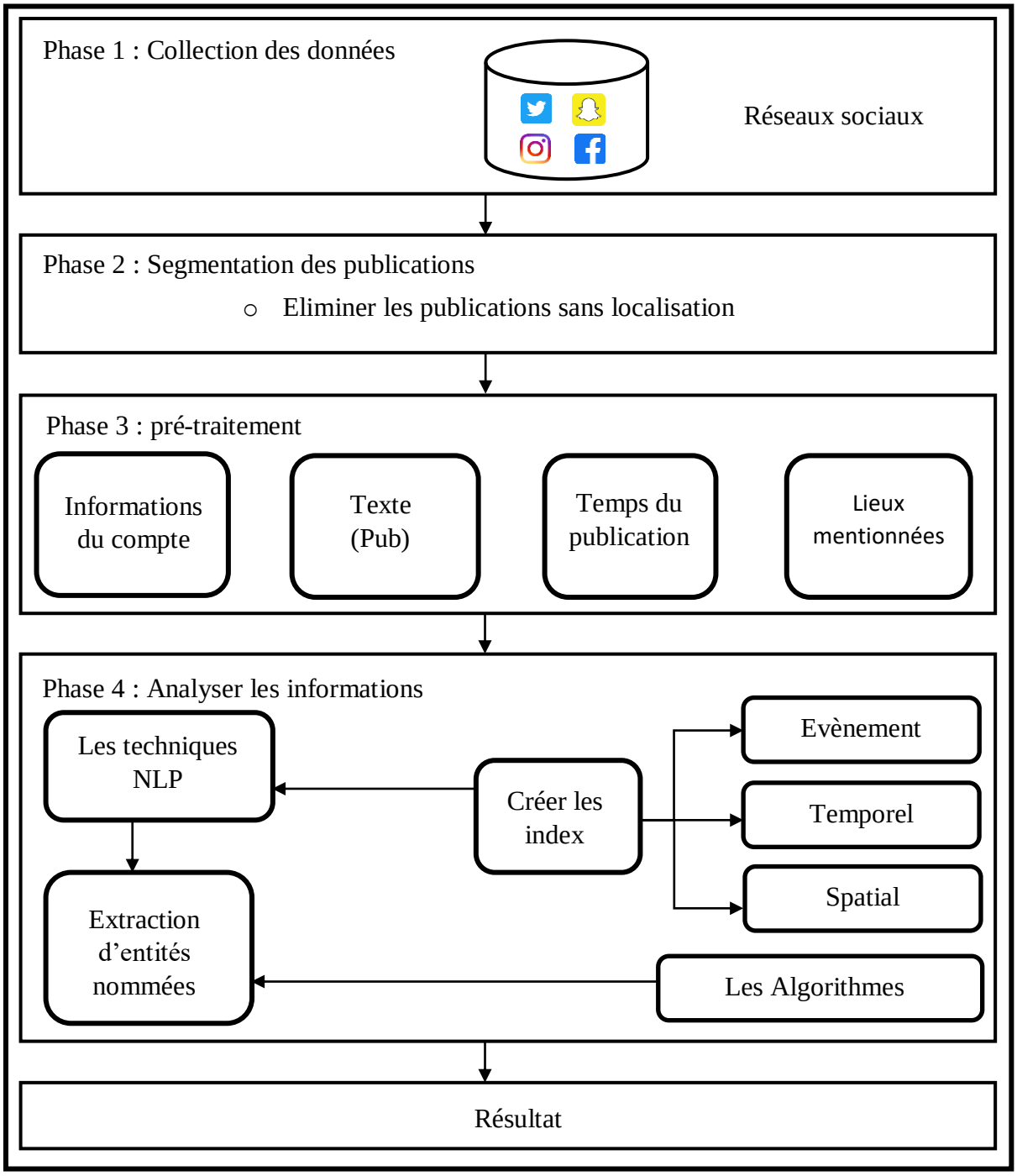


Figure 3.4 : Architecture générale du système proposé.

5. Phases de l'approche proposé

5.1 Collection de donnée

Nous obtenons les informations de réseaux sociaux (Facebook) online. Nous essayons de collecter manuellement les publications liées aux évènements qui laissent des victimes ;il s'agit bien d'accidents ou de catastrophes naturelles.

5.2 Segmentation des publications

Les bonnes et efficaces publications sont sélectionnées et les indésirables qui ne contiennent pas les conditions nécessaires sont supprimées. Nous vérifions que les publications sur Facebook contiennent une localisation géographique réelle et pas seulement des lieux imaginaires. (Explication personnel)

5.3 Pré-Traitement

Après avoir collecté les publications liées à ce que nous voulons, nous passons a l'application de la phase de Prétraitement qui vise à représenter les données et à les transformer sous forme analysable. Elle améliore également la qualité des données en réduisant le bruit du texte et les objets inutiles. Par conséquent, le traitement se fait en supprimant tous : les signes (@) et les liens URL de la publication, après quoi nous supprimons les espaces de début et de fin et les caractères non alphabétiques tels que : / * ; \ |, et les sauts de ligne.

Nous effectuons cette étape pour obtenir le texte contenant les caractères alphabétiques, les signes de ponctuation et les chiffres uniquement. Nous pouvons nous expliquer plus avec un exemple d'une publication et noter le résultat après y avoir appliqué l'étape de prétraitement. (Explication personnel)

Table 3.5.3 : Exemple d'application de la phase Prétraitement sur une publication

Source	Facebook
Publication original	#News News News #Urgent Un grave accident de la circulation à l'entrée de la commune de Hejjaj, Mostaganem, nous vous demandons d'être prudent..... 😢 Nous remercions la Protection Civile Algérienne pour l'intervention immédiate et le sauvetage des blessés 😍🙏🙏 @la protection_civile_Algérienne @GendarmerieAlgérien
Publication après Prétraitements	Un grave accident de la circulation à l'entrée de la commune de Hejjaj, Mostaganem, nous vous demandons d'être prudent..... Nous remercions la Protection Civile Algérienne pour l'intervention immédiate et le sauvetage des blessés
	Un grave accident de la circulation à l'entrée de la commune de Hejjaj, Mostaganem, nous vous demandons d'être prudent Nous remercions la Protection Civile Algérienne pour l'intervention immédiate et le sauvetage des blessés

5.4 Analyse des informations

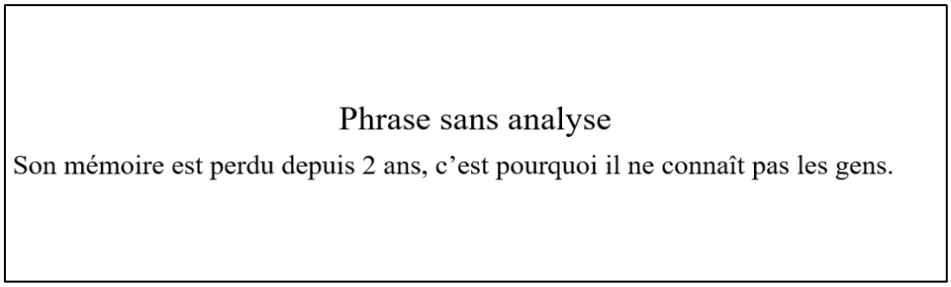
L'information est analysée en traitant le texte de l'illuminateur en suivant des techniques de traitement du langage naturel, et des index géographiques sont créés (événement, spatial) afin d'aider à faire correspondre le texte, la dernière étape consiste à extraire l'information en implémentant un ensemble des techniques d'analyse classer comme suit :

5.4.1 Tokeniser

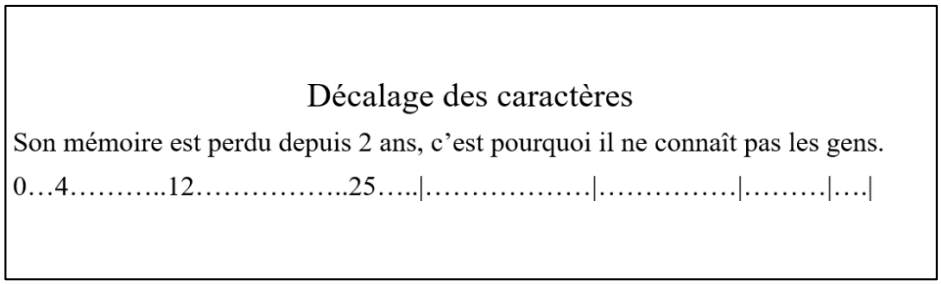
Le Tokeniser divise le texte en jetons très simples tels que des chiffres, des signes de ponctuation et des mots de différents types. Par exemple, nous distinguons les mots en Majuscule et en Minuscule, et entre certains types de ponctuation, etc. En ajoutant une annotation "Jeton" à chacun, il n'a pas besoin d'être modifié pour différentes applications ou types de texte. L'objectif est de limiter le travail du Tokeniser pour maximiser l'efficacité et permettre une plus grande flexibilité en mettant la charge sur les règles de grammaire, qui sont plus adaptables. [21]

Les étapes d'analyse des informations

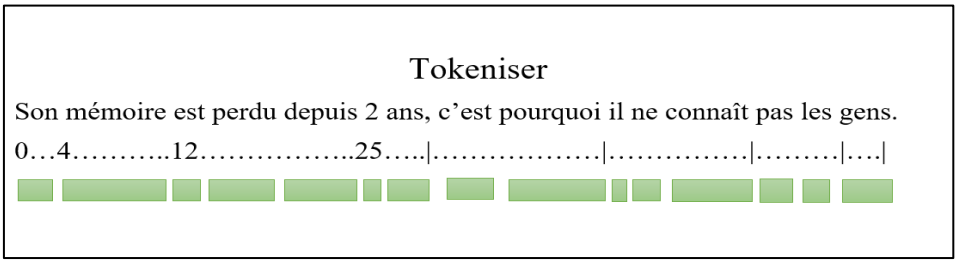
1)



2)



3)

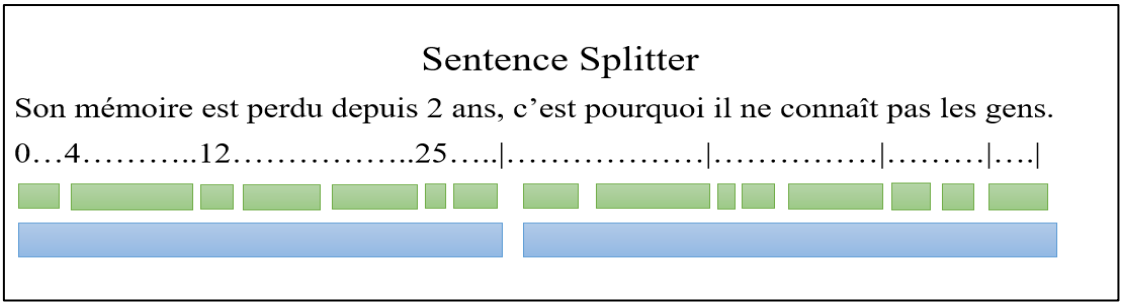


5.4.2 Sentence Splitter

Le séparateur de phrases est une cascade de transducteurs à états finis qui segmente le texte en phrases. Ce module est requis pour le tagueur. Le séparateur utilise une liste d'abréviations du répertoire géographique pour aider à distinguer les points de marquage des phrases des autres types. [21]

Chaque phrase est annotée avec le type « Phrase ». Chaque saut de phrase (comme un point) reçoit également une annotation « Split ». Il a une fonctionnalité « genre » avec deux valeurs possibles : « interne » pour toute combinaison d'exclamation et de point d'interrogation ou un à quatre points et « externe » pour une nouvelle ligne.

Le séparateur de phrases est indépendant du domaine et de l'application.[21]



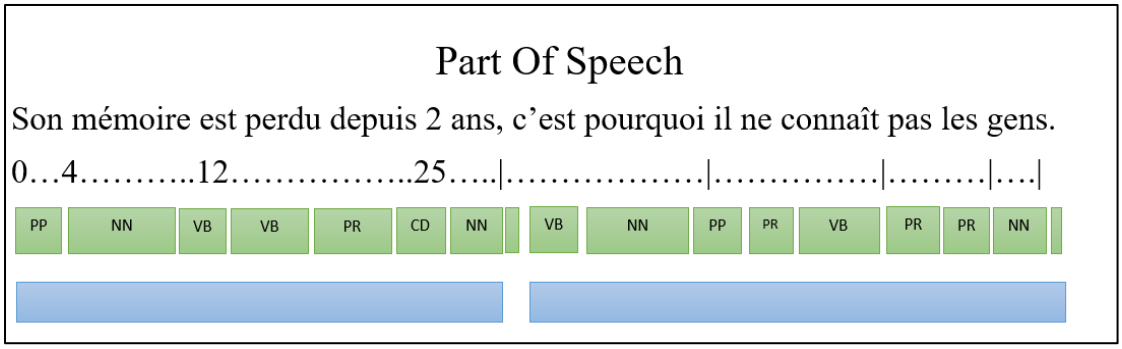
5.4.3 PartOfSpeech Tagger

Le taggueur est une version modifiée du marqueur Brill, qui produit une étiquette de partie de discours en tant qu'annotation sur chaque mot ou symbole. Le tagueur utilise un lexique et un ensemble de règles par défaut (résultat d'un apprentissage sur un large corpus tiré du Wall Street Journal). Ces deux éléments peuvent être modifiés manuellement si nécessaire. Deux lexiques supplémentaires existent ; un pour les textes tout en majuscules et un pour les textes tout en minuscules. Pour les utiliser, le lexique par défaut doit être remplacé par le lexique approprié au moment du chargement. L'ensemble de règles par défaut doit toujours être utilisé dans ce cas.[21]

Le tagueur ANNIE Part-of-Speech nécessite certains paramètres :

- Encoding
- LexiconURL
- RulesURL
- Document
- InputASName
- OutputASName

- BaseTokenAnnotationType
- BaseSentenceAnnotationType
- OutputAnnotationType
- PosTagAllTokens
- FailOnMissingInputAnnotations [21]



5.4.4 Gazetteer

Le rôle du Gazetteer est d'identifier les noms d'entités dans le texte à partir de listes.Ilse compose de listes telles que les villes, les organisations, les jours de la semaine, etc. Ces listes utilisées sont des fichiers en texte brut, avec une entrée par ligne.Le Gazetteer contient quelques entités, mais aussi des noms de mots clés utiles, tels que des désignations d'entreprise (par exemple "Ltd."), des titres (par exemple "Dr."), etc. Les listes sont compilées dans des machines à l'état finis, qui peuvent correspondre à des jetons de texte. [21]

- **Gazetter Personnel**

La plateforme GATE contient des plugins, mais ces derniers ne peuvent pas reconnaître certains noms Algériens, donc un dictionnaire spécial doit être créé qui contient une liste de noms berbères pour aider GATE à les identifier.

- **Gazetter Spatial**

Ce travail se concentre sur la recherche des personnes qui ont été touchées par des accidents ou bien par des catastrophes naturelles. En raison de l'absence du dictionnaire inclus dans GATE de la majorité des régions et villes Algériennes, en particulier les municipalités et les petites villes, la liste des villes algériens doit donc être incluse pour augmenter la précision des informations extraites.

- **Gazetter Evènement**

Comme le gazetteer personnel et spatial, un Gazetteer basé sur les événements des accidents ou des catastrophes naturelles a étédéveloppé. Pour créer ce Gazetteer, nous allons

collecter tous les mots, les expressions et les données de publication du Facebook qui ont une relation avec ces évènements.

5.4.5 Semantic Tagger

Le marqueur sémantique d'ANNIE (A Nearly-New Information Extraction System) est basé sur le langage JAPE (Java Annotation Patterns Engine.). Il contient des règles qui agissent sur des annotations, fonctionnalités et valeurs attribuées dans les phrases précédentes. Il contient aussi des règles qui agissent sur des annotations, fonctionnalités et valeurs qui doivent être attribuées manuellement.[21]

Les types annotations, les caractéristiques et les valeurs possibles par défaut produits par ANNIE sont basés sur les types d'entités MUC d'origine et sont les suivants :

- Person: genre: male, female
- Location: locType: region, airport, city, country, county, province, other
- Organization:orgType: company, department, government, newspaper, team, other
- Money
- Percent
- Date: kind: date, time, dateTime
- Address: kind: email, URL, phone, postcode, complete, ip, other
- Identifier
- Unknown [21]

6. Conclusion

Dans ce chapitre nous avons détaillé les étapes de conception du système d'extraction d'information, expliqué les phases de l'architecture. Dans la phase d'analyse, nous avons cité les différentes étapes comme Tokenisation, Sentence Splitter, et le Part Of Speech avec des exemples pour mieux expliquer.

Dans le chapitre suivant nous allons appliquer un modèle d'un Système d'extraction d'information de documents des réseaux sociaux et spécialement Facebook.

Chapitre IV :

L’implémentation

1. Introduction

Dans le chapitre précédent nous avons créés une conception d'un système d'extraction des entités nommées, suivant ce chemin nous avons choisi le GATE développer comme un environnement d'application de notre approche proposée, utilisant les règles du JAPE. Dans ce chapitre, nous allons appliquer les étapes de traitement sur un ou plusieurs documents Facebook importés en ligne et on va voire les résultats.

.

2. GATE

2.1 Définition

GATE (General Architecture for Text Engineering) est une application libre en open source (en Java) 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. GATE est largement utilisé par les experts en TAL et dispose d'une grande communauté d'utilisateurs. Quand il s'agit de NER, GATE facilite le développement de systèmes NER, tel qu'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. Par ailleurs, les créateurs de GATE proposent des formations pour améliorer son niveau ainsi que des certifications permettant de faire valoir ses compétences à un niveau professionnel.

2.2 CREOLE

CREOLE (Collection of REusableObjects 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 :

2.2.1 Ressources langage (LRs : Laguage 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.

2.2.2 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.

2.2.3 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.

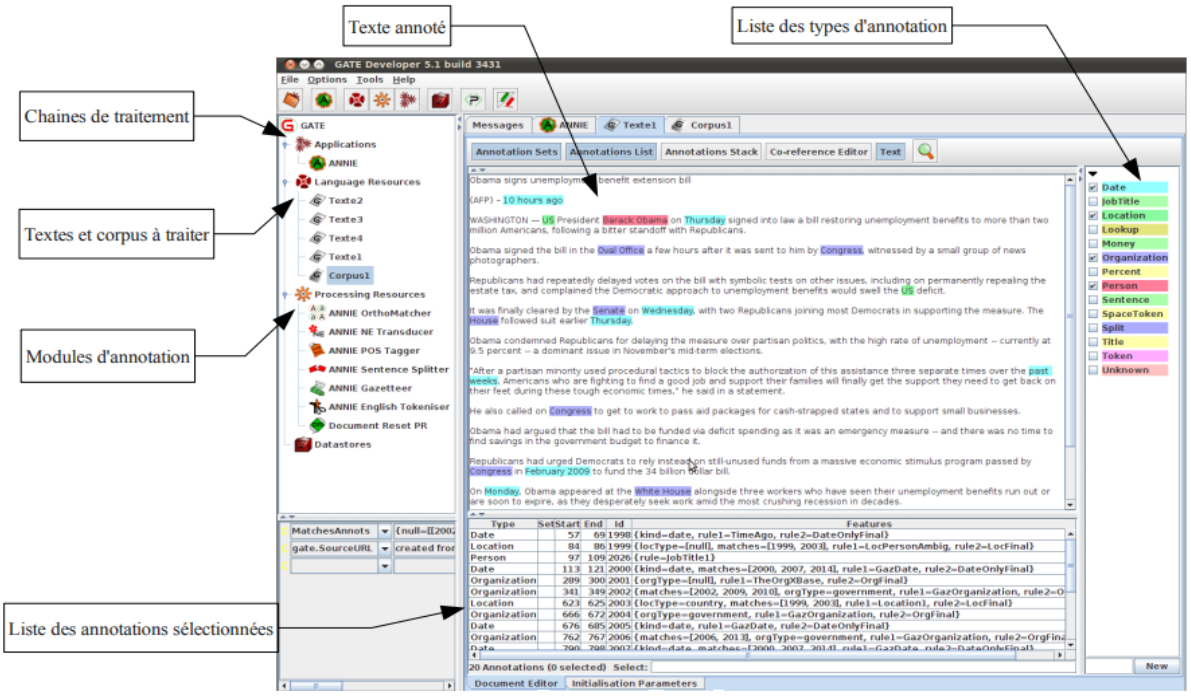


Figure 2.1 : Exemple de l'interface dans GATE

2.3 ANNIE

ANNIE (A Nearly-New Information Extraction 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 entreautres 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 :

2.3.1 Découpeur de tokens (tokeniser) :

Dont le rôle est de diviser le texte en jetons simples (ponctuations, nombres, mots, etc.),

2.3.2 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.

2.3.3 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).

2.3.4 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. Il produit une étiquette de partie du discours sous la forme d'une annotation pour chaque mot ou symbole.

2.3.5 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).

2.4 Le formalisme JAPE

Une partie des différents modules proposés dans GATE est basée sur JAPE (Java Annotation Patterns Engine), un transducteur à états finis permettant de reconnaître des expressions régulières sur les annotations. Ce système s'avère très utile en extraction d'informations car il permet de définir les contextes d'apparition des éléments à extraire pour ensuite les repérer et les annoter. Le principe est de combiner différentes annotations « basiques » (tokens, syntagmes, relations syntaxiques, etc.) pour en créer de nouvelles plus complexes (entités nommées, relations, évènements, etc.) : cela revient à l'écriture de règles de production et donc à l'élaboration d'une grammaire régulière. Une grammaire JAPE se décompose en plusieurs phases exécutées consécutivement et formant une cascade d'automates à états finis. Chaque phase

correspond à un fichier «. jape » et peut être constituée d'une ou plusieurs règle(s) écrite(s) selon le formalisme associé à JAPE. Classiquement, 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. Le lien entre ces deux parties se fait par l'attribution d'une étiquette au motif (ou à ses constituants) en LHS et par sa réutilisation en RHS pour y appliquer les opérations nécessaires. Pour plus de clarté, prenons l'exemple d'une règle simple :

```
1.  Rule: OrgAcronym
2.  ((
3.   {Organization}
4.   {Token.string == "("}
5.   ({Token.orth == "allCaps"}):org
6.   {Token.string == ")"}
7.  )
8.  -->
9.  :org.Organization = {rule = "OrgAcronym"}
```

Figure 2.4 : Exemple de règle avec le formalisme JAPE

3. Application De L'approche Proposer

L'approche d'extraction détaillée ci-dessus a été réalisée dans l'environnement GATE pour le traitement de textes. Notre chaine d'extraction de relations utilise pour base les différents modules dans l'extraction d'entités nommées. Nous y avons ajouté des règles :

3.1. Entité Nommée Temporelle (date)

a) Le document avant l'ajout des règles JAPE

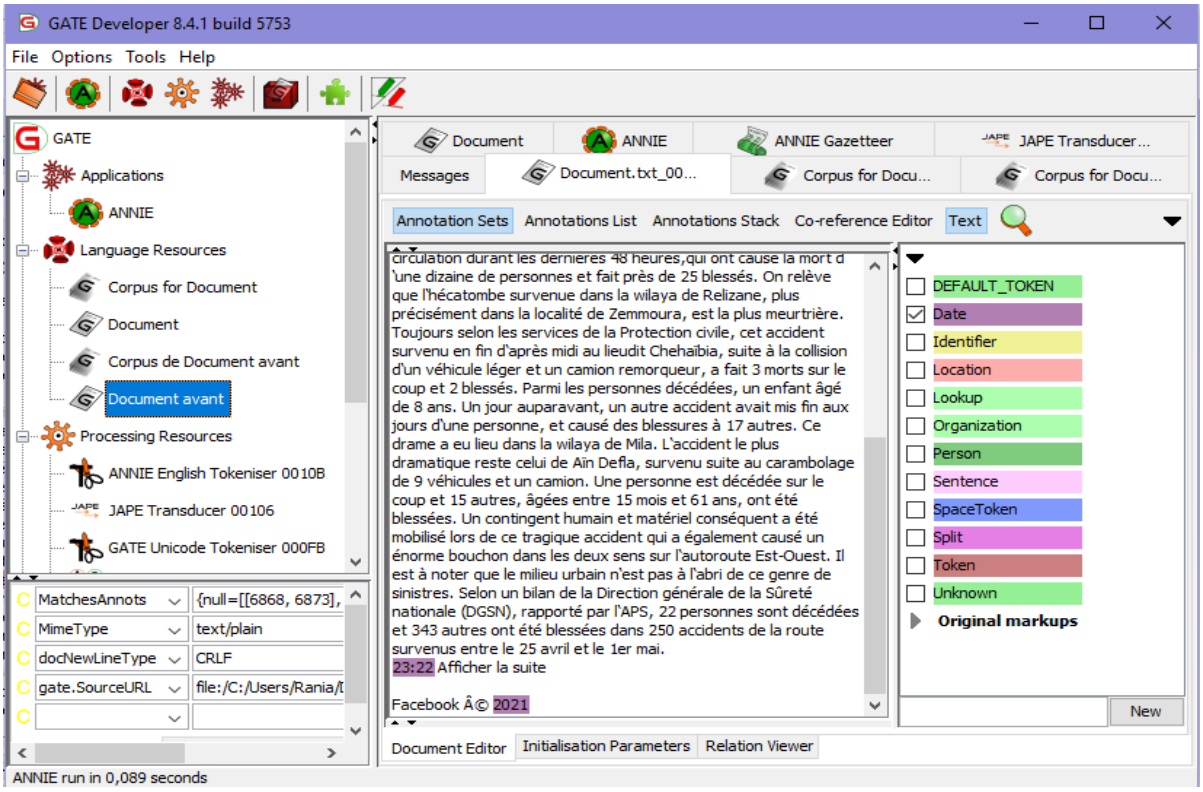


Figure 4.3.1 a : Extraction d'entités nommées temporelle avant les règles JAPE

b) Le document après l'ajout des règles JAPE

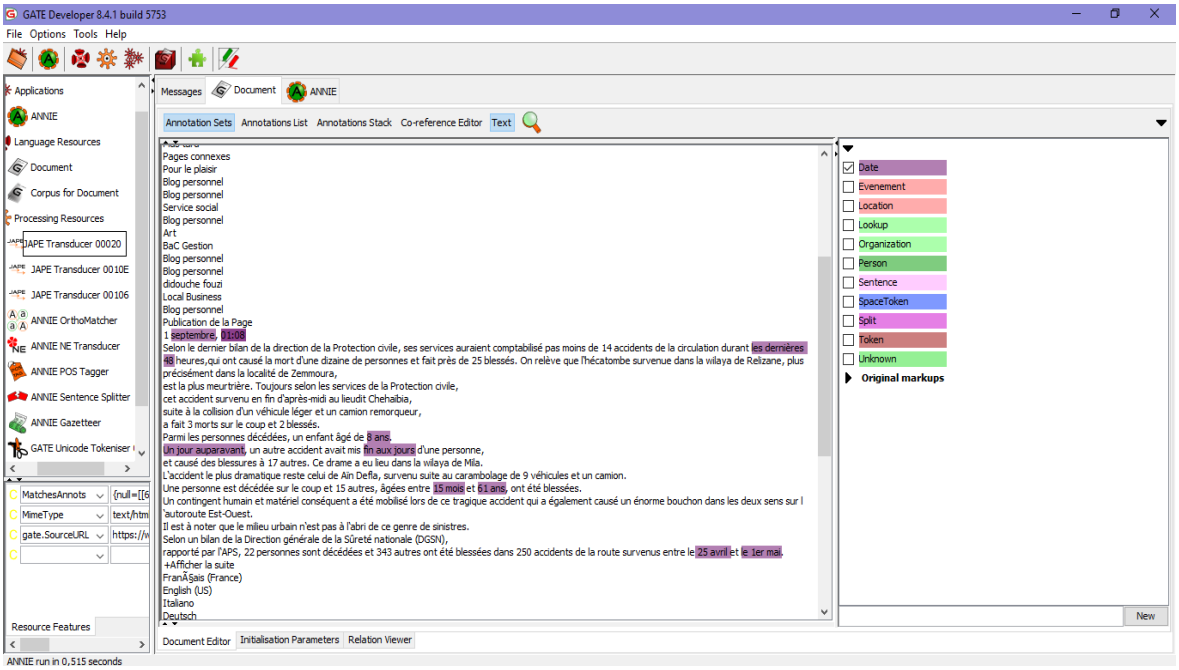


Figure 4.3.1 b : Extraction d'entités nommées temporelle après les règles JAPE

c) Evaluation :

Categorie	Précision	Rappel	F-mesure
Systéme par default	0,15	0,13	0,14
Système avec Notre Approche	0,76	0,66	0,7

3.2 Entité Nommée Spatiale (Location)

a) le document avant l'exécution l'ajout des règles JAPE

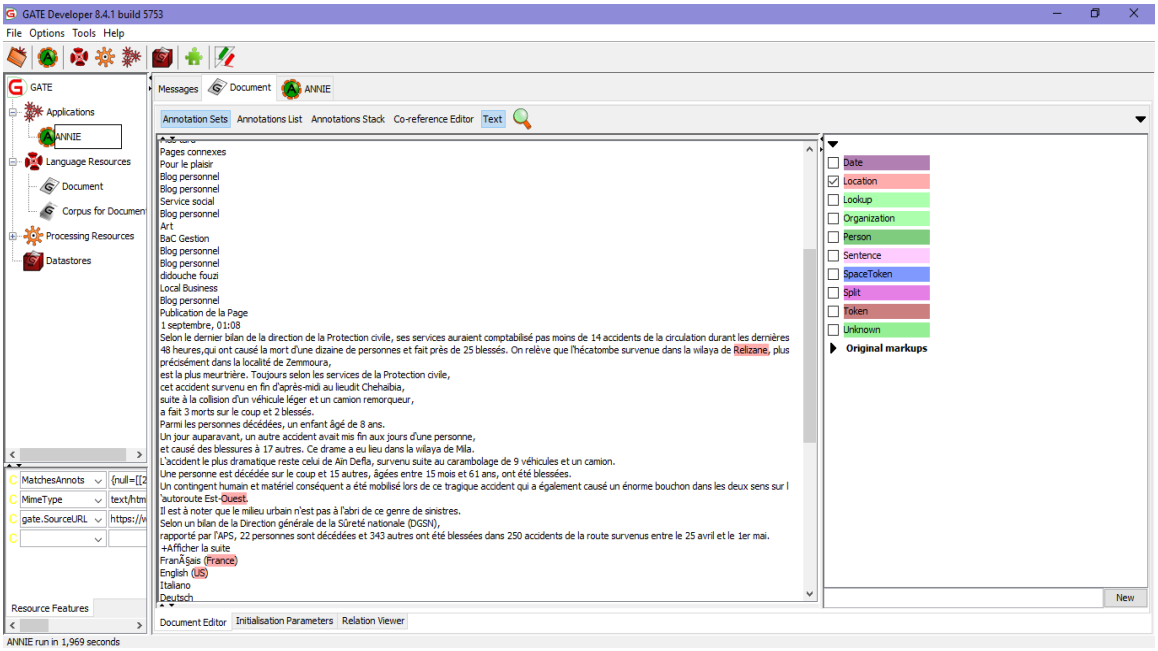


Figure 4.3.2 a : Extraction d'entités nommées spatial avant les règles JAPE

b) Le document après l'ajout des règles JAPE

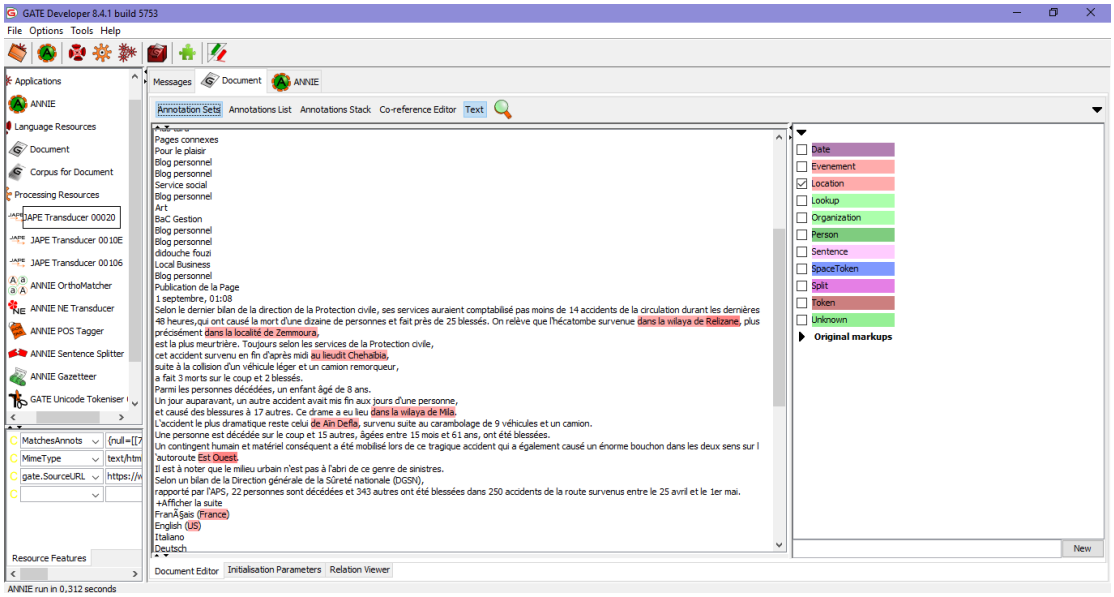


Figure 4.3.2 b : Extraction d'entités nommées spatial après les règles JAPE

c) Evaluation

Categorie	Précision	Rappel	F-mesure
Systéme par default	0,26	0,22	0,23
Système avec Notre Approche	0,86	0,72	0,78

3.3 Entité Nommée Evènement :

a) Le document avant l'exécution

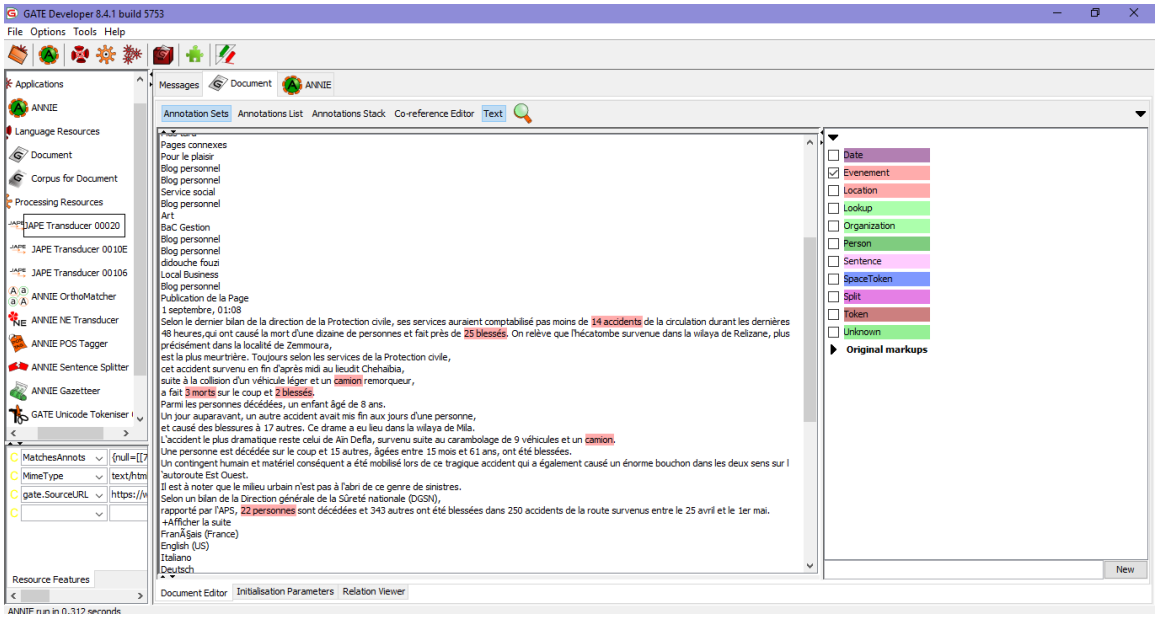


Figure 4.3.3 a : Extraction d'entités nommées d'évènement avant les règles JAPE

b) Le document après l'exécution

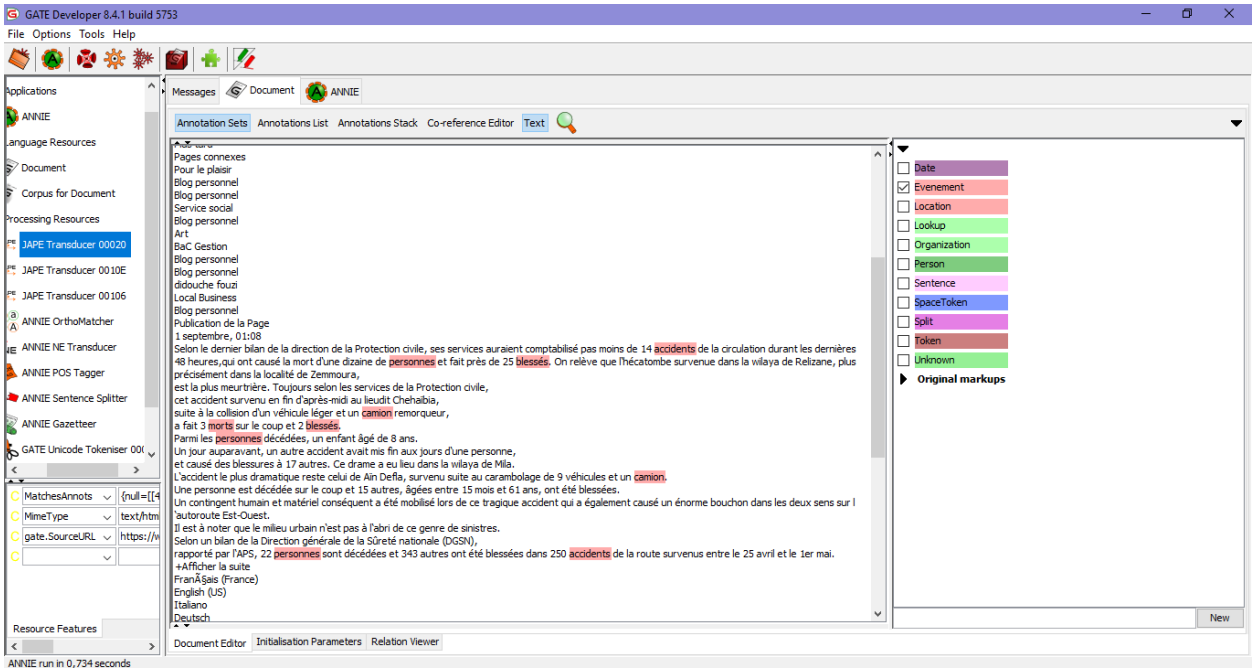


Figure 4.3.3 b : Extraction d'entités nommées d'évènement après les règles JAPE

c) Evaluation

Categorie	Précision	Rappel	F-mesure
Systéme par default	0,53	0,51	0,52
Système avec Notre Approche	0,46	0,42	0,44

4. Conclusion

Dans ce chapitre, Nous avons appliqué les règles du JAPE sur un document Facebook importé en ligne dans l'environnement du GATE pour extraire les entités nommées spatial, Temporelle, et d'évènement. Nous avons cité les résultats avant et après l'application des règles JAPE.

Conclusion générale

Notre travail a pour objectif de faire l'extraction de l'information à partir des entités nommées.

Dans un premier temps, nous avons mis l'accent d'une manière générale sur des thèmes de recherche qui sont en relation avec TALN, et particulier les thèmes d'extraction de l'information à partir des entités nommées pour apprendre une autre approche.

Dans les chapitres I et II, nous avons présenté le concept général d'extraction de l'information et les entités nommées, en commençant par introduire les différents concepts permettant de cerner la notion d'entité nommée.

Dans le chapitre III, nous avons mis en évidence les différentes étapes d'extraction de l'information à partir des documents dans les medias sociaux, puis présenté notre modèle d'extraction de l'information basé sur l'entité nommée.

Dans le dernier chapitre, nous avons réalisé notre approche qui a pour objectif de se limiter à l'extraction des entités de type « Date », « Location » et « Evènement » dans des textes importateurs d'après media social.

Notre objectif principal est donc l'extraction de toutes les informations existantes sur un document à l'aide de certains types d'entité nommé et évaluer cet objectif par rapport aux autres approches présents dans le système.

Pour une meilleure évaluation de notre approche et afin de compléter ce travail, il serait souhaitable de donner quelques perspectives qui s'inscrivent dans la continuité directe de notre travail qui est l'amélioration de l'extraction des entités nommées avec d'autres moyens sémantiques basées sur les ontologies.

Bibliographie

[1] : SEGHIRI N. « Détection et extraction d'information temporelles dans un entrepôts de données » . Tizi-Ouzou : Université Mouloud Mammeri, 2014, 88.

[2] : EVEN F. « Extraction d'information et modélisation de connaissance à partir de Naute de communication orale » . Informatique. Université de Naute, 2005, 253.

[3] : GHOULAM A. « L'extraction d'information pour la recherche dans un système médical à large échelle ». Informatique. Université d'Oran, 2018, 158.

[4] : TELLIR Isabelle. « TAL et extraction d'information ». France : Université Paris 3-Sorbonne Nouvelle, 2014, 49. (4-TELLIER)

[5] :Apprentissage Supervisé : Introduction – (Machine Learnia.com)

[6] :Apprentissage non-supervisé : définition et algorithmes populaires (journaldunet.fr)

[7] :Apprentissage par renforcement (reinforcmentlearning) (larevueia.fr)

[8]: WIMALASURIYA D-C. « Ontology Based Information Extraction ». Etats Unis : Université de l'Oregon, 2009, 37. (AREA-200903)

[9] : VICTORRI B, MATHET Y, ENJALBERT P. « Nouvelle perspective en extraction d'information » . Manuscrit auteur, 2002, 19.

[10] :Recherche d'information - Définition et Explications (techno-science.net)

[11] DEFFAF F, « Extraction des entités nommées par projection cross-linguistique et construction de lexique bilingues d'entités nommées pour la traduction automatique statique », mémoire présenté comme exigence partielle de la maitrise en informatique, Université du Québec à Montréal, 2015.

[12] FLITTI S, « Identification Automatique d'Entités Nommée », UNIVERSITE ABDELHAMID IBN BADIS – MOSTAGANEM,2016.

[13] DAMIEN N, « Reconnaissance des entités nommées par exploration de règles d'annotation Interpréter les marqueurs d'annotation comme instructions de structuration locale », thèse de

docteur en Informatique, l'École Doctorale MIPTIS, l'université François Rabelais de Tours, France, 2012.

[14] Tjong Kim Sang Erik F. « Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition ». In proceedings of the 6th conference on Naturallanguageeaming- COUNG-02, Vol.20 , 2002.P.1-4.

[15] GRAVIER G, BONASTRE J.-F, GEOFFROIS E, GALLIANO S, MCTAIT K et CHOUKRI K. 2004. « The ESTER evaluation campaign of rich transcription of French broadcast news ». In Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC) , Lisboa, Portugal, 2004, P.885-888.

[16] LAKEL K « Les annotations sémantiques dans les documents Web : application aux textes psychologiques en langue arabe » thèse de doctorat en science informatique, l'Université d'Oran Mohammed boudiafe, Oran, 2018.

[17] HATMI M, « Reconnaissance des entités nommées dans des documents multimodaux », thèse de doctorat en Informatique, Ecole doctorale sciences et technologies de l'information et mathématique, Université de Nantes, France, 2014.

[18] SAIDI I, « Contributions aux techniques de recherche d'informations », thèse de doctorat en l'informatique, l'Université d'Oran Ahmed Benbella, Oran – LITIO, 2014.

[19] BERDANE F, HOUALI N,REZZOUG F «Modélé de langue mixte combinant entitée nommées et mots simples»,Universite Mouloud Mammeri de Tizi-Ouzou,2013

[20] ABDELKOUI F. « Recherche d'information Géographique à l'aide des Ontologies Spatiales de Location ». Constantine : Université Abdelhamid Mehri-Constantine 2,2017,117.

[21] https://gate.ac.uk/sale/tao/splitch6.html#x9-1210006.2