MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

RAPPORT DE MINI-PROJET

Option : **Ingénierie des Systèmes d'Information**

THEME :

Extraction des entités nommées pour la construction ontologie linguistique

Etudiant : « **Mahieddine Mohamed Bettahar** »

« **Ketroussi Ahmed** »

Encadrante : « Kenniche Ahlem »

Année Universitaire 2020-2021

Résumé

Ces dernières décennies, le développement considérable des technologies de l'information et de la communication a modifié en profondeur la manière dont nous avons accès aux connaissances. Face à l'afflux de données et à leur diversité, il est nécessaire de mettre au point des technologies performantes et robustes pour y rechercher des informations. Les entités nommées (personnes, lieux, organisations, dates, expressions numériques, marques, fonctions, etc.) sont sollicitées afin de catégoriser, indexer ou, plus généralement, manipuler des contenus. Notre travail porte sur leur reconnaissance et leur annotation. En première partie, nous abordons la problématique de la reconnaissance automatique des entités nommées. Et le deuxième chapitre, nous avons présente une partie des travaux d'exploitation de Wikipédia extraction et la reconnaissance.

Mots-clés : les entités nommées, reconnaissance des entités nommées, Extraction de l'information.

Abstract

In recent decades, the considerable development of information and communication technologies has profoundly changed the way we access knowledge. Faced with the influx of data and its diversity, it is necessary to develop high-performance and robust technologies to search for information. The named entities (people, places, organizations, dates, numerical expressions, brands, functions, etc.) are requested in order to categorize, index or, more generally, manipulate content. Our work focuses on their recognition and annotation. In the first part, we address the issue of automatic recognition of named entities. and the second chapter we have presents some of the operating work of Wikipedia extraction and recognition.

Keywords: Named entities, Knowledge discovery, natural language.

Liste des figures

Liste des abréviations

Table des matières

Introduction Générale

Depuis un quart de siècle, la croissance ininterrompue du web met à notre disposition une somme chaque jour plus considérable de connaissances, dispersées dans des milliards de pages. Mais enfermées dans des documents rédigés à l'attention d'êtres humains, ces connaissances ne peuvent être traitées que superficiellement par une machine, incapable de comprendre le langage naturel. Certains types de mots, toutefois, peuvent être isolés et extraits automatiquement d'un texte. L'opération fait l'objet d'une discipline informatique à part entière nommée « extraction d'informations », dont la tâche principale consiste à remplir automatiquement des formulaires ou une base de données à partir de textes non structurés [Moens 06, p. 2]. Sur l'ensemble des informations susceptibles d'être extraites d'un texte, une catégorie particulière nourrit les recherches depuis deux décennies. Sous le terme « entités nommées » (Named Entities), forgé en 1996, des informaticiens ont regroupé un ensemble flou de concepts qui se rapprochent, sans se limiter à lui, de celui de « noms propres ». Il s'agit par exemple des noms de personnes, de lieux, d'organisations, mais aussi parfois les dates, voire les noms de composants chimiques mentionnés dans un article scientifique.

La technologie en général et l'informatique en particulier font aujourd'hui un part de notre vie quotidienne. Ainsi, l'utilisation des médias est devenue très exigeante surtout à l'arrivé de l'Internet et des réseaux sociaux. C'est pourquoi le nombre de personnes qui gèrent leurs propres données est trop augmenté. Et ce qui explique aussi la croissance de volume des données (surtout les données non structurées).

L'exploitation et l'accès à ces données deviennent très difficiles et nécessitent des méthodes pour résoudre ces difficultés. Tous ces phénomènes étaient parmi les facteurs qui font l'apparition de la tâche de l'extraction d'information (EI).

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

La reconnaissance d'entités nommées est la tâche primordiale de l'extraction d'information qui a pour but de détecter, d'extraire et de catégoriser les noms propres (nom de personne, organisation, localisation . . . etc.).

Différents systèmes sont développés à propos de cette tâche mais ses difficultés liées à l'identification et la désambiguïsation d'entités nommées sont toujours présentes et nécessitent la recherche d'autres méthodes pour les résoudre. Nous concentrons dans ce travail sur le problème d'extraction des entités nommées et le développement d'un système capable de recherche dans Wikipédia et détecter un grand nombre d'entités.

Le premier chapitre présente une étude sur l'état de l'art sur les entités nommées où nous citons les différentes définitions, les typologies, la recherche, ...etc. et Le deuxième chapitre appelle la reconnaissance des entités nommées à partir Wikipédia.

Chapitre 1

Les entités nommées

1.1 Introduction

Le concept de la recherche d'entités a pour but d'exploiter la richesse du web afin d'en tirer les données enfouies dans les pages non structurées. La recherche d'entités deviendra une des meilleures techniques d'exploitation du contenu du web. Dans ce qui suit, nous présentons un exemple représentatif de ce qui est considéré comme entité. Dans ce chapitre, nous présentons les travaux les plus importants sur la recherche d'entités. La section suivante portera sur la recherche d'entités, nous avons jugé qu'il était nécessaire de détailler mieux ce que c'est une entité avant de présenter les travaux relatifs à la recherche d'entités.

Comme il le note Ehrmann dans (Ehrmann 2008), le traitement des entités nommées fait actuellement figure d'incontournable en Traitement Automatique des Langues. Apparue au milieu des années 1990 à la faveur des dernières conférences MUC (Message Understanding Conferences), la tâche de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. apparaît en effet comme fondamentale pour diverses applications participant de l'analyse de contenu et nombreux sont les travaux se consacrant à sa mise en œuvre, obtenant des résultats plus qu'honorables. Fort de ce succès, le traitement des entités nommées s'oriente désormais vers de nouvelles perspectives avec, entre autres, la désambiguïsation et une annotation enrichie de ces unités.

1.2 De quoi s'agit-il :

Le cours de l'histoire, ou plutôt de la recherche, a voulu que l'on désigne un certain nombre des unités linguistiques de niveaux différents sous le nom d'Entités Nommées (EN) (named entities en anglais).

Ces dernières correspondent traditionnellement à l'ensemble de noms propres présents dans un texte qu'il s'agisse de noms de personnes, de lieux ou d'organisation, ensemble auquel sont souvent ajoutées d'autres expressions comme les dates, les unités monétaires, les pourcentages et autres. Contemporain des travaux en Extraction d'Information initiés au début des années 1990, le traitement des entités nommées s'articule en deux processus :

- Identification ou reconnaissance de ces unités dans les articles Wikipédia tout d'abord.
- Catégorisation ou typage selon des catégories sémantiques larges prédéfinies ensuite.

1.3 Les Entités nommées :

1.3.1 Définition et approches historiques :

La tâche de reconnaissance des entités nommées (Named Entity Recognition NER) est apparue à la faveur du développement de la tâche d'extraction d'information. Elle consiste à identifier et catégoriser les entités nommées.[1]

Définitions des entités nommées :

Une entité nommée est une expression linguistique référentielle, souvent associée aux nomes propres et aux description définies.

Selon [Le Meur et al., 2004] aucune définition standard d'entité nommée ne s'est pas encore imposée. Le NIST (National Institute of Standards and Technology) propose la définition suivante:

"Named Entity: a named object of interest such as a person, organization, or location"

Une autre définition, considérée plus « complète » pour [FOTSOH TAWOFAING, 2018] est celle proposée dans l'encyclopédie Atalapédia de l'ATALA (Association pour le Traitement Automatique des Langues) :

"Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprises, etc... Contenues dans un texte.

On ajoute souvent à ces éléments les dates et d'autres données chiffrées. [...]. Ces séquences référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes."

Comme nous avons mentionné précédemment, les entités nommées sont des objets catégorisables. Parmi les classes (catégories) de ces objets nous distinguons : les noms de personne, les noms de lieu, les noms d'organisation, les dates, les chiffres... etc.

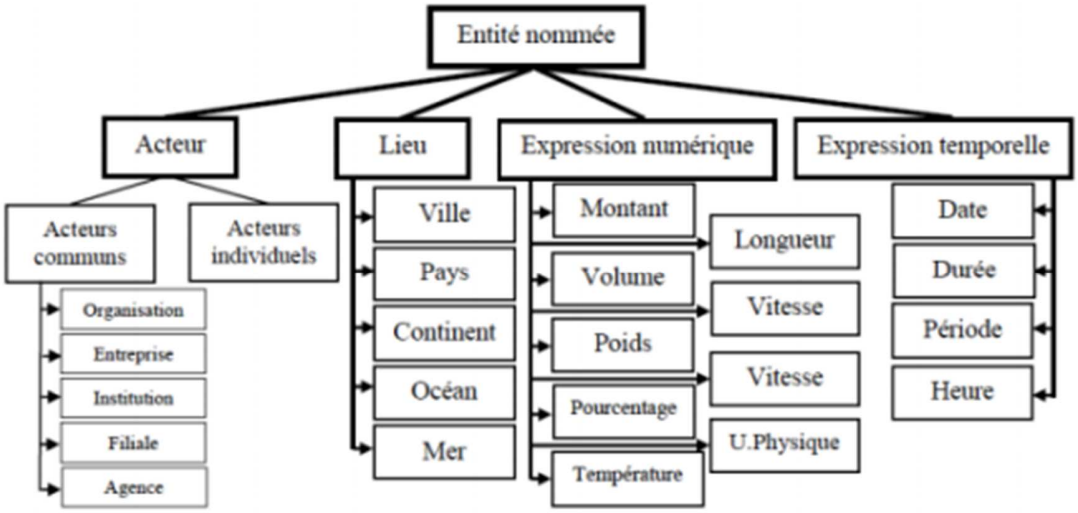


Figure 01 : les catégories des entités nommées.

Discussion linguistique :

En grammaire, le nom propre est en général considéré comme une sous-catégorie du nom et se distingue du nom commun. Ainsi, un nom commun est un nom employé pour designer tous les éléments d'un même ensemble. Par exemple, animal, poème, pièce de théâtre. Le nom commun dispose d'une définition et d'une signification et il est utilisé en fonction de cette signification. Par exemple, le nom commun cuillère dispose d'une définition ; et le fait d'évoquer cette définition permet à chacun d'imaginer à quoi ressemble une cuillère (W. Zaghouani 2009).[2]

Concernant les noms propres, Jonasson dans (Jonasson 1994) propose trois définitions de leurs sens :

- Un nom propre est un prédicat de dénomination : il ne décrit pas l'objet dénoté, mais lui colle une étiquette, par exemple telle fille est nommée Anissa.

- Le nom propre est vide de sens puisqu'il permet de référer sans désigner.

- Le sens du nom propre est une description du référent, soit il a un sens réduit à des traits sémantiques généraux comme la distinction féminin / masculin, animé / non animé, soit il dispose d'un sens fort et il permet d'identifier clairement un référent.

Enfin Boulanger et Cormier dans (Boulanger and Cormier 2001), proposent la définition suivante : le nom propre fait partie des éléments de nature langagière auxquels recourent les locuteurs pour produire des discours et pour construire leur image du monde ainsi que celle des réalités qui les entourent. Ainsi, le nom propre réfère principalement à une entité unique que cela soit pour représenter des objets, des personnes, des lieux géographiques, des marques déposées ou même des événements (W. Zaghouani 2009).

Du point de vue conceptuel sémantique, les noms propres s'appuient sur les réflexions qui établissent un lien entre le langage comme ensemble de symboles signifiants et les objets ou concepts du monde réel que le langage référence. De ce point de vue, diverses théories évoquent un lien reposant, selon, sur le sens, la dénotation, la référence, la désignation, etc.

D. Nouvel dans (Nouvel 2012) reprend la thèse du mathématicien Frege qui est le premier à établir une distinction claire entre le sens et la référence. La référence pointe vers un concept, qui peut correspondre à un objet du monde réel. De manière plus abstraite, le sens est un mécanisme par lequel un signe (symbole, nom propre par exemple) peut désigner une ou plusieurs références. Il peut y avoir plusieurs sens désignant une même référence ou a contrario certains sens ne désignant aucune référence.

En outre, Frege tient également compte du fait que le sens est nécessairement lié à une représentation individuelle, chaque humain interprétant les signes selon son expérience personnelle. Il doit donc exister une convention permettant à plusieurs individus d'attribuer un sens similaire à des expressions complexes du langage naturel.

Historiques :

La notion d'entité nommée (EN) est apparue en 1996. Ce type de noms ont un grand intérêt parce qu'ils sont présents dans tout type de texte. Ainsi, ils constituent un point de passage obligé pour tout système cherchant à rendre compte de l'information contenue dans un texte. La tâche de reconnaissance d'entités nommées a commencé aux États-Unis avec les campagnes MUC. [3]

Le but était d'évaluer cette tâche qui se focalise sur trois types d'entités :

- **Enamex :** pour les noms de personne, d'organisation et de lieu (Person, Organization et Location) ;
- **Timex :** pour les expressions temporelles (Date et Time) ;
- **Numex :** pour les expressions numériques, de monnaie et de pourcentage (Money et Percent).

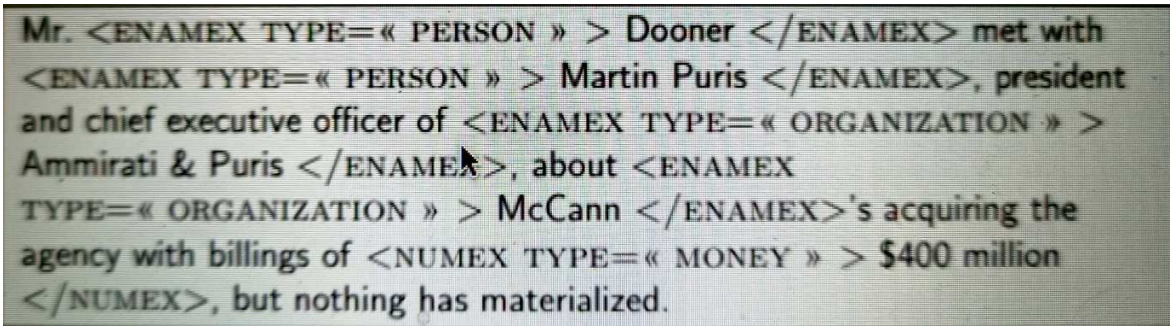


Figure 02 : Exemple d'annotation d'entités nommées (MUC-6). [Ehrmann, 2008].

1.4 Ambiguïtés :

Une entité nommée est dite ambiguë quand elle est susceptible d'avoir plusieurs interprétations concernant sa délimitation ou sa typologie. Cette ambiguïté a différents phénomènes qui complexifient la tâche d'annotation des entités nommées [HATMI, 2014] : [4]

- Ambiguïtés graphiques : La majuscule est un indicateur pour le repérage et la délimitation des entités nommées. Cependant cet indicateur n'est pas simple à manipuler pour plusieurs raisons :

  - ✓ Une entité nommée peut contenir des formes commençant par une minuscule (exemple : le château de Versailles). - La première forme d'une phrase comporte aussi une majuscule, que ce soit une entité nommée ou non.

  - ✓ L'emploi de la majuscule pour les noms propres n'est pas une règle dans toutes les langues.

  - ✓ La transcription automatique de la parole est souvent bruitée. La majuscule n'est pas toujours respectée.

- Ambiguïtés sémantiques : Comme les noms communs, les entités nommées n'échappent pas à la polysémie, à l'homonymie et à la métonymie. Considérons les énoncés suivants :

  - Orange a invité M. Dupont.

  - Leclerc a fermé ses magasins en Rhône-Alpes.

  - La France a signé le traité de Kyoto.

La difficulté qui existe ici est d'identifier la catégorie de ces entités. Est-il question de la ville d'Orange ou bien de la société de téléphonie ? De la personne Michel Edouard Leclerc ou de la chaîne de supermarché ? Faut-il préférer une annotation de France en tant que «organisation» ou « gouvernement » ou en tant que « lieu » ou « pays » ?

1.5 Entité, c'est quoi au juste :

Les travaux en recherche d'information ont porté une attention particulière aux noms propres de personnes, de lieux et d'organisations, appelés entités nommées. Les entités nommées sont des séquences lexicales qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (humain, social, politique, économique, géographique, etc.).

1.6 Les typologies des EN :

La typologie des EN tente de prédéfinir des catégories et des types dans lesquels on classifie les unités textuelles concernées par le processus de la REN. Le spectre des types d'EN va de la classification peu détaillée des conférences MUC dans les années 90 jusqu'aux typologies

adoptées par les projets et campagnes d'évaluation REN les plus récents18. Ces derniers sont caractérisés par une décomposition plus fine des catégories d'EN grossièrement définies au début et de l'ajout de nouvelles catégories. Dans ce qui suit dans cette section on énumère les typologies trouvées dans la littérature pendant l'accomplissement des travaux de cette thèse. [5]

1.6.1 Typologie des conférences MUC :

La conférence MUC a été créée dans le but de promouvoir la recherche en invitant les chercheurs à venir participer avec leurs outils et leurs systèmes à une compétition annuelle d'extraction de l'information.

Les participants étaient alors invités à développer un système qui permet l'extraction du plus grand nombre d'informations possibles sur des entités bien précises. Par la suite une évaluation est conduite en suivant la même procédure pour l'ensemble des participants. Les systèmes d'extraction participants ont été évalués sur des domaines tels que le terrorisme en Amérique Latine lors de MUC-3 (MUC 1991) et de MUC-4 (MUC 1992). Lors de MUC-5 (MUC 1993), le domaine était la fusion d'entreprises et la fabrication de circuits électroniques.

Lors de MUC-6, le domaine était les changements de dirigeants des entreprises (MUC 1995).

Enfin, MUC-7 a porté sur les accidents d'avion.

À partir de la sixième édition de MUC, baptisée MUC-6, la tâche d'extraction des EN a été créée et par la même occasion la notion d'entités nommées a été introduite.

La conférence MUC-7 a distingué trois types d'entités à reconnaître et à catégoriser, soit ENAMEX, NUMEX et TIMEX. La Figure 3 montre ces trois grandes catégories et leur limite par rapport à la grande famille des EN. On remarque clairement qu'il y a une majeure partie d'EN qui n'est pas couverte par la classification MUC.

- Les entités de type ENAMEX sont composées des noms propres, des sigles et des abréviations. Les entités ENAMEX se divisent en trois catégories :
  - **personnes :** les noms de personnes ou de familles,
  - **noms de lieux :** ce sont des lieux définis géographiquement ou politiquement comme les villes, provinces, rivières, montagnes,
  - **organisations :** cette catégorie inclut les noms de gouvernements, sociétés, et autres entités organisationnelles.
- Les entités NUMEX rassemblent les nombres et les pourcentages, les unités de mesures, les devises,
- Enfin, les entités TIMEX couvrent les expressions de temps et les dates.

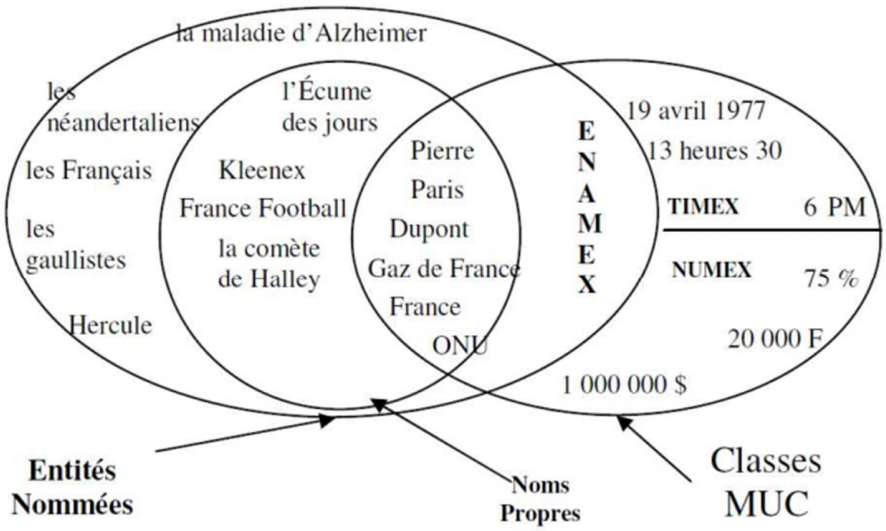


Figure 03 : Les entités nommées vs la classification MUC (Daille et al. 2000).

1.6.2 Typologie de Paik

La classification de Paik et al citée par (Maurel et al. 2011) regroupe ensemble les entités nommées et les entités temporelles. Cette approche a été mise au point à la suite de l'analyse

d'un corpus du Wall Street Journal. Elle comporte 30 catégories divisées en 9 classes (W. Zaghouani 2009) :

- Géographique : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques.
- Affiliation : religions, nationalités.
- Organisation : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations.
- Humain : personnes, fonctions.
- Document : documents.
- Équipement : logiciels, matériels, machines.
- Scientifique : maladies, drogues, médicaments.
- Temporelle : dates et heures.
- Divers : autres noms d'entités nommées.

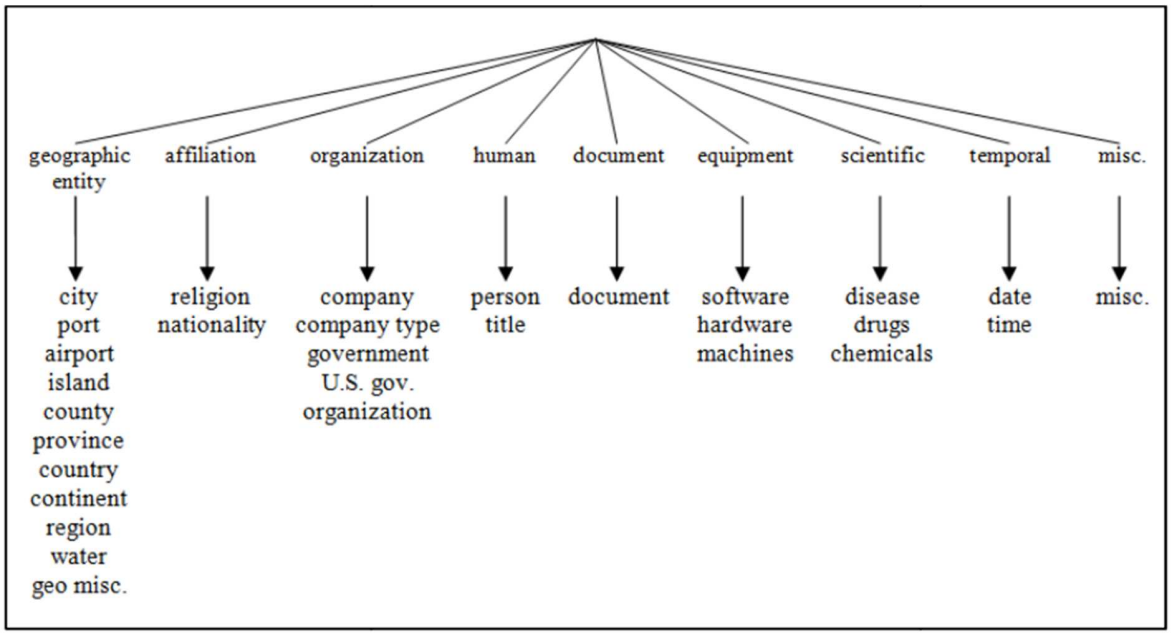


Figure 04 : Typologie de paik et al. Citée par (Maurel et al.2011)

1.6.3 Typologie de Bauer

Bauer (1985) cité par (Grass, 2000) a présenté une autre catégorisation du nom propre dans le cadre de ses recherches sur la traduction. Sa classification n'inclut pas les entités temporelles et se divise en six classes et chaque classe comporte plusieurs catégories :

- Anthroponymes : les personnes individuelles ou les groupes :
  - Patronymes,
  - Prénoms,
  - Pseudonymes, o gentilés,
  - Hypocrites,
  - Ethnonymes,
  - Groupes musicaux
  - Ensembles artistiques et or
  - Partis
  - Organisations

- Toponymes : les noms de lieux :
  - Pays, villes
  - Microtoponymes
  - Hydronymes,
  - Oronymes,
  - installations militaires

- Ergonymes : les objets et les produits manufactur entreprises, établissements d de publications, d'œuvre d'art.

1.6.4 Typologie du projet ReNom

ReNom est un projet français piloté par le laboratoire d'informatique à l'Université de Tours en collaboration avec le laboratoire ligérien de linguistique de l'université d'Orléans. Son objectif est d'enrichir les textes de renaissance par des informations sémantique telles que les étiquettes des catégories des EN (Maurel et al.2013) ont adoptée dans ce projet une typologie majoritairement inspiré de la norme TEI les entités reconnues sont réparties en quatre types, les lieux géographiques (geogName), les lieux administratifs (placeName), les organisations (orgName) et, enfin, les personnes ou personnages(persName).

1.7 Méthodes de reconnaissance

A l'image de la plupart des applications TAL [Poibeau 11, p. 24], les systèmes de reconnaissance d'entités nommées peuvent se diviser globalement en trois grandes catégories : les systèmes à base de règles, les systèmes fondés sur l'apprentissage automatique et les méthodes hybrides [Poibeau 01]. [6]

1.7.1 Les systèmes à base de règles

Ces méthodes dites aussi « linguistiques » ou « symboliques » impliquent de confectionner à la main des patrons d'extraction, à la manière d'une expression régulière. Ceux-ci peuvent être fondés sur des preuves internes (l'abréviation inc dans Micro- soft inc en tant que nom d'entreprise) ou externes (les termes « Monsieur » ou « Madame » devant des noms de personnes) [Friburger 02, p. 16]. Certains de ces ensembles de règles sont couplés à des lexiques, notamment des dictionnaires de noms propres.

L'on peut citer comme exemple de ces méthodes le système CasEN [Maurel 11] pour la reconnaissance d'entités nommées en français, qui fait appel à une cascade de transducteurs dans le logiciel CasSys de la plateforme Unitex définir le patron d'un seul sous-type d'entité (ici les organisations politiques) peut réclamer une quantité considérable de travail et d'expertise linguistique.

Dominants jusqu'au début des années 90, les systèmes à base de règles ont peu à peu cédé du terrain face aux méthodes probabilistes, sans pour autant devenir obsolètes. Plus ardus à concevoir et à maintenir, « ils prédominent pour les langues ou les typologies pour lesquelles il n'existe pas de corpus de données an- noté de taille suffisante », soulignent [Nouvel 15, p. 91]. Généralement très précis, ils permettent en outre un contrôle plus fin sur les règles de détection. Ils ont toutefois l'inconvénient d'offrir un rappel plus faible. Chaque type de détection devant être prévu à l'avance et codifié, ils s'adaptent mal à un contexte nouveau.

1.7.2 Les systèmes à apprentissage automatique

Issues des travaux en IA, les méthodes par apprentissage automatique (*Machine Learning*) reposent sur un principe aussi simple que séduisant : plutôt que d'écrire soi-même un programme qui effectue une action quelconque, par exemple annoter un texte, il est plus avantageux d'en écrire un qui apprend à le faire tout seul à partir d'exemples [Tellier 10, p. 95]. Autrement dit, un programme qui *s'améliore* à mesure qu'il reçoit de nouvelles données. Pour reprendre la définition formelle de l'un des pionniers de ces systèmes :

« A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. » [Mitchell 97, p. 2].

Ces méthodes, parfois appelées « probabilistes » ou « fondées sur les données », peuvent elles-mêmes se subdiviser en trois grandes catégories [Pustejovsky 13, p. 141] :

— **Les méthodes non supervisées**, qui visent à découvrir une structure dans un jeu de données non annoté. La classification automatique [Sebastiani 02] et le regroupement (*clustering)* en constituent deux applications courantes. [Etzioni 05] ou encore [Nadeau 06] fournissent des exemples d'applications à la classification d'entités nommées ;

— **Les méthodes supervisées**, dans lesquelles il s'agit de reproduire automatiquement un schéma d'annotation appris d'un corpus annoté manuellement. Les systèmes qui y ont recours peuvent faire appel à une large gamme de modèles d'apprentissage automatique, tels que les champs conditionnels aléatoires (CRF), les modèles de Markov cachés (HMM), les machines à vecteurs de support (SVM), les classifieurs d'entropie maximale (MaxEnt), etc. ;

— **Les méthodes semi-supervisées**, qui combinent les deux précédentes en ingérant à la fois des données annotées et des données brutes. Pour un exemple concret, voir [Nadeau 07a].

En matière de reconnaissance d'entités nommées, les systèmes à apprentissage automatique supervisé se sont principalement développés grâce à la publication de vastes corpus annotés, issus notamment des campagnes d'évaluation déjà citées.

Pour les langues peu dotées, en revanche, la confection de tels corpus *ex nihilo* peut s'avérer rédhibitoire. Bien que l'annotation réclame moins de compétences que la confection de règles linguistiques, obtenir un F-score (voir Section 1.6) satisfaisant impose des volumes de données considérables. [Poibeau 01] cite plusieurs exemples de systèmes qui nécessitaient plus d'un million de mots annotés afin obtenir un taux de rappel/précision avoisinant 0.90. Par ailleurs, un système entraîné sur un corpus précis offrira de piètres performances sur un autre. [Ciaramita 05], par exemple, a constaté une chute de performances considérable dans un système entraîné sur des dépêches *Reuters* (ConLL 2003) appliqué à un corpus d'articles du *Wall Street Journal,* le F1-score étant passé de 0.908 à 0.643.

1.7.3 Les systèmes mixtes

Enfin, rien n'interdit qu'un ensemble de règles apprises automatiquement soit ensuite affiné par un linguiste, ou l'inverse. [Béchet 11], par exemple, témoigne de la complémentarité de deux démarches en associant un système à base de règles, conçu sur des dépêches de l'Agence France-Presse, avec un système probabiliste entraîné sur des transcriptions de l'oral issues du corpus ESTER. Selon leur conclusion, cette adaptation leur a permis de disposer de deux systèmes, l'un privilégiant la précision, l'autre le rappel, « sans nécessiter aucune annotation supplémentaire, illustrant la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques ».

Ce type de métissage semble constituer la voie la plus prometteuse. Loin de se cantonner à la reconnaissance d'entités nommées, le croisement des méthodes à base de règles et des systèmes probabilistes est d'ailleurs devenu une tendance centrale dans l'ensemble du TAL [Tellier 09, Watrin 06].

1.8 La recherche d'entités :

Le concept de la recherche d'entités a pour but d'exploiter la richesse du web afin d'en tirer les données enfouies dans les pages non structurées. La recherche d'entités deviendra une des meilleures techniques d'exploitation du contenu du web. Dans ce qui suit, nous présentons un exemple représentatif de ce qui est considéré comme entité.

Dans ce chapitre, nous présentons les travaux les plus importants sur la recherche d'entités. La section suivante portera sur la recherche d'entités, nous avons jugé qu'il était nécessaire de détailler mieux ce que c'est une entité avant de présenter les travaux relatifs à la recherche d'entités.

La recherche d'entités sur le Web (Wikipédia) est un nouveau groupe de recherche qui va au-delà de la recherche d'un document classique.

Alors que pour les tâches de recherche d'information la recherche de document peut donner des résultats satisfaisants pour l'utilisateur, différentes approches doivent être suivies lorsque l'utilisateur doit rechercher des entités spécifiques. Par exemple (voir la figure 03.), lorsque l'utilisateur veut trouver une liste des "Pays d'Afrique du Nord il est facile pour un moteur de recherche classique de retourner des documents sur l'Afrique du nord, et c'est à l'utilisateur d'extraire l'information sur les entités demandées dans les résultats fournis.

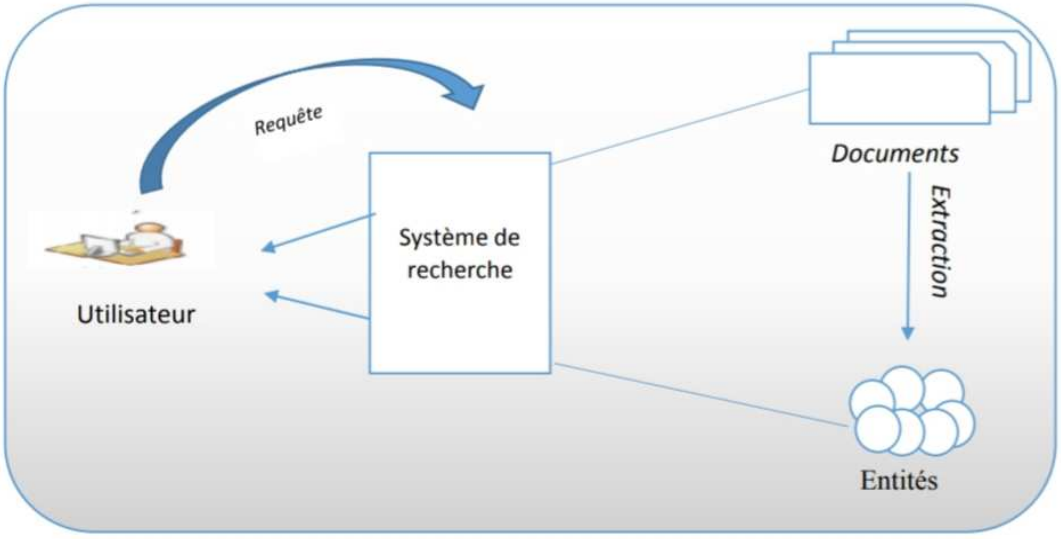


Figure 05 : Recherche entités vs recherche documents.

Être en mesure de trouver les entités sur le Web en général et en particulier sur Wikipédia, peut devenir une nouvelle caractéristique importante des moteurs de recherche actuels. Il peut permettre aux utilisateurs de trouver plus que des pages Web (Wikipédia), mais aussi des gens, des numéros de téléphone, livres, films, voitures, etc. La recherche d'entités dans une collection de documents n'est pas une tâche facile. Par conséquent, afin de trouver des entités, il est nécessaire de faire une étape de prétraitement d'identification des entités dans les documents. En outre, nous avons besoin de construire des descriptions de ces entités pour permettre aux pages Wikipédia (les article Wikipédia) de classer et retrouver les entités pertinentes pour une requête. L'application des méthodes de recherche RI classique pour la recherche d'entités peuvent mener à une faible efficacité.

C'est parce que la recherche d'entité, est une tâche différente de celle de la recherche de documents. Un exemple d'une requête est "Aéroports" en Allemagne où un résultat pertinent est, par exemple, "l'aéroport de FrankfurtHahn".

Il est intéressant de noter qu'un recensement des différents types de requêtes réelles du web a été effectué. Les auteurs de ce travail classifient les requêtes en quatre types et présentent le pourcentage de chaque type, comme suit :

- Requête d'entité « Entity query » (40 %), exemple : « 1978 cj5 jeep » (marque de voiture).
- Requête de type « Type query » (12 %), exemple : « doctors in barcelona ».
- Requête d'attribut « Attribute query » (5 %), exemple : « zip code atlanta ».
- Autre « Other query » (36 %), même si (14 %) de ces (36 %) contiennent un contexte d'entité ou un type.

Exemple de motivation :

Les entités recherchées pouvant être connues ou inconnues aux utilisateurs, Cela signifie qu'il existe différents choix pour poser sa requête : recherche par une seule entité (R1E), recherche par plusieurs entités (RPE) et recherche par mots clés (RMC).

Remarque : si la requête est un mélange d'entités et de mot clés, elle est alors considérée comme une requête de mots clés (RMC), car nous supposons que lorsque l'utilisateur forme une requête d'entités (RPE), c'est qu'il cherche un lien ou veut faire une comparaison (ex., Renault ou Peugeot, infection et tumeur, etc.).

Nous présentons dans ce qui suit un exemple de motivation. Cet exemple est tiré d'un contexte d'informations politiques.

Supposons qu'un utilisateur souhaite avoir des informations sur la politique algérienne. L'utilisateur peut poser différentes requêtes.

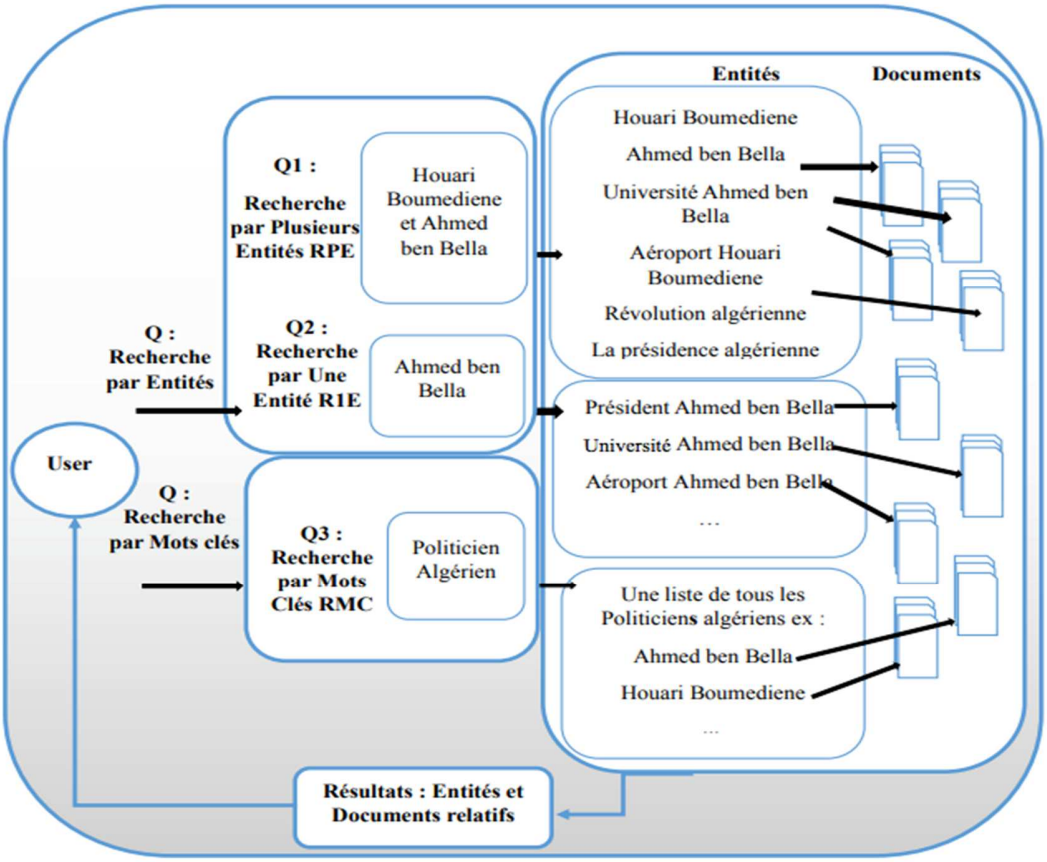


Figure 06 : Exemple de motivation

Dans la première requête, l'utilisateur veut avoir des informations sur deux personnes, il s'intéresse aux informations communes entre ces deux personnes. Ce cas représente la recherche par plusieurs entités (RPE). L'utilisateur saisit les entités (connues) : "Houari Boumediene et Ahmed ben Bella ", il aura en résultat les entités relatives à sa recherche ainsi que leurs documents. L'utilisateur pourra explorer les documents relatifs aux entités trouvées.

La deuxième requête est un cas spécial de la recherche précédente RPE. Ce cas consiste en la recherche d'une seule entité (R1E). L'utilisateur voudrait des informations sur une entité particulière. Pour cela, il saisit l'entité (connue) qu'il veut rechercher, par exemple : « Ahmed ben Bella ». Il est intéressant de retourner à l'utilisateur, en plus de sa requête : « Ahmed ben Bella », les entités composées par cette dernière (par exemple, Ahmed ben Bella aéroport, Ahmed ben Bella université, etc.). L'utilisateur pourra alors explorer leurs documents relatifs.

Dans la troisième requête, l'utilisateur veut avoir des informations sur une requête formée de mots clés, par exemple : « président de l'Algérie ». Ce cas représente la recherche d'entités (inconnues) par mots clés (RMC). Les résultats sont les entités relatives (pertinentes et contextuelles) à cette requête et les documents répondant à chaque entité.

1.9 Conclusion :

Dans ce chapitre, nous avons présenté une théorie complète des entités nommées, nous avons également discuté des définitions, des approches historiques des entités nommées, les différentes ambiguïtés qui complexifient cette tâche, de leur classification et de leur méthode pour les identifier et les rechercher.

Chapitre 2

Extraction des entités nommées à partir Wikipédia

2.1 Introduction

La reconnaissance d'entité nommée est une sous-tâche primordiale dans l'activité d'extraction d'information.

Elle a fait son apparition dans les conférences MUC. L'objectif était d'extraire les entités qui désignent les noms de personnes, d'organisations et des lieus à partir des textes politiques (domaine militaire).

Cette tâche a évolué au fils des années est devenue une composante essentielle dans plusieurs domaines de traitement automatique des langues naturelles et elle s'attache à extraire d'autres types d'entité nommée tels que les expressions numériques et temporelles.

Les ressources libres sont des recueils de données informatisées de productions langagières écrites ou parlées. Ces ressources sont généralement utilisées pour l'enrichissement d'un contenu textuel. Elles fournissent l'exploitation automatique des informations linguistiques. Il faut assurer alors la qualité et la cohérence de ces informations afin de permettre la bonne exploitation. Parmi les ressources libres, nous citons la Wikipédia qui est une encyclopédie multilingue créée par Jimmy Wales et Larry Sanger le 15 janvier 2001. Cette encyclopédie apporte des liens explicatifs et offre un contenu librement réutilisable, objectif et vérifiable. Elle fournit aussi l'accessibilité et la reconnaissance automatique des sujets mentionnés dans des textes non structurés à travers des liens.

2.2 Travaux exploitant la Wikipédia

Parmi les domaines qui ont choisi la ressource libre Wikipédia pour élaborer des systèmes puissants, nous citons celui de la traduction automatique. Dans ce contexte, [Sellami at al., 2013] ont proposé un système de traduction automatique statistique à partir de corpus comparables. Dans ce système, les auteurs ont choisi la paire de langues arabe-français. Le travail effectué consiste à exploiter les liens inter-langues qui relient les articles en arabe à ceux en français. Ce type de lien facilite l'extraction entre les termes (simples ou composés) arabes et leurs traductions en français et vice versa. Plusieurs travaux s'inspiraient de cette ressource pour réaliser des enrichissements de leurs systèmes. En fait, la ressource Wikipédia est très utilisée pour construire des ressources linguistiques et les enrichir également.

C'est dans ce cadre que s'inscrit le travail de [Sellami et al., 2012] consistant à construire un lexique bilingue à partir de la Wikipédia en profitant de son aspect multilingue.

La richesse de la Wikipédia en termes d'EN et son aspect multilingue ont joué un rôle important pour la proposition des systèmes de REN. Parmi ces travaux, nous citons le travail de [Biltawi et al., 2016] qui se concentre sur la création des dictionnaires en exploitant cette ressource libre. Les dictionnaires créés sont classés en trois principales catégories : *nom de personne, nom de lieu et nom d'organisation*. Cette création de différents types de dictionnaires est une initiation à un processus de REN.

2.3 Reconnaissance d'entités nommées :

La reconnaissance d'entités nommées est la tâche primordiale de l'extraction d'information qui a pour but de détecter, d'extraire et de catégoriser les noms propres (nom de personne, organisation, localisation . . . etc.).

2.3.1 Extraction d'information :

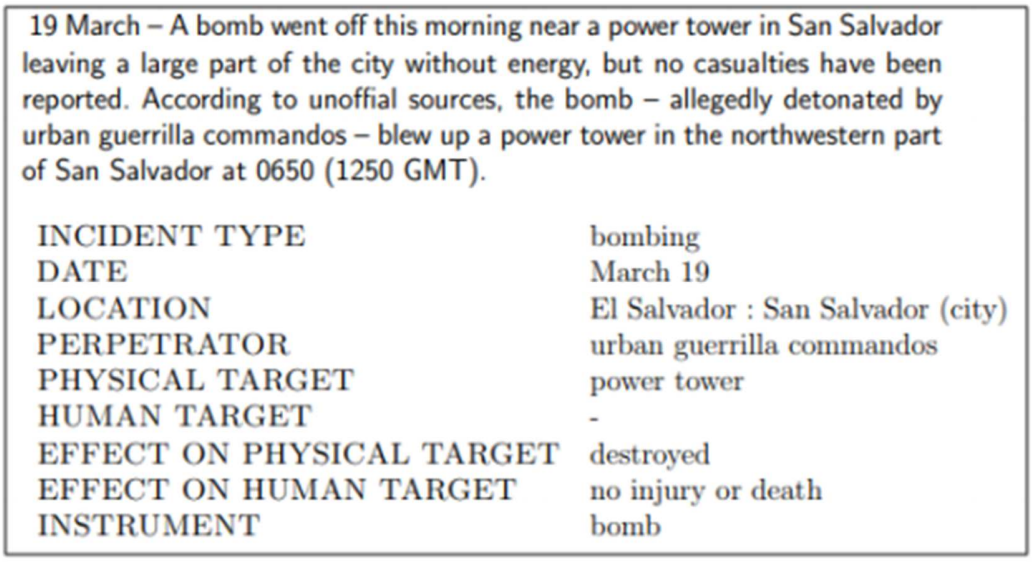
L'extraction d'information, considérée comme l'un des domaines essentiels du traitement automatique des langues naturelles TALN (Natural Language Processing NLP), est la tâche qui vise à extraire des informations pertinentes à partir de collections de documents non structurés sans chercher à comprendre les textes dans leur ensemble. [Bogers, 2004].
Dans la littérature, différentes définitions ont été proposées à propos de cette tâche [AIT RADI and IGGUI, 2013] : [7]

> **Définition 1.1 :** L'Extraction d'Information consiste à remplir automatiquement des formulaires ou une banque de données à partir de textes écrits en langue naturelle. [Pazienza, 1997].
>
> **Définition 1.2 :** L'Extraction d'Information consiste à remplir une source de données structurées (base de données) à partir d'une source de données non structurées (texte libre). [Gaizauskas and Wilks, 1998].
>
> **Définition 1.3 :** L'Extraction d'Information consiste à analyser des textes écrits en langage naturel dans le but d'obtenir des informations en vue d'une application précise. [AIT RADI and IGGUI, 2013].
>
> **Définition 1.4 :** L'Extraction d'Information est l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle [Poibeau, 2003]

La première idée de la tâche d'extraction d'information avait été introduit par le linguiste Américain Zellig Harris en 1950 mais elle est lancée officiellement dans la série des conférences MUC1 (Message Understanding Conferences) en 1987. Ce cycle de conférences, organisé par diverses institutions américaines et financé par la DARPA s'est déroulé de 1987 à 1998, motivant de la sorte de nombreuses équipes de recherche pendant plus d'une décennie.

Comme leur nom l'indique, l'objectif de ces conférences était à l'origine d'encourager la recherche autour de la compréhension automatique de messages militaires. La figure 07 montre un exemple de formulaire à remplir pour MUC-3 : à partir d'une dépêche sur un acte terroriste. Le champ de l'EI est souvent décomposé en plusieurs sous problèmes qui sont :

- L'extraction d'entités nommées ;
- L'extraction de descripteurs thématiques (libres ou normalisés) ;
- L'extraction de phrases importantes sous un point de vue donné ;
- L'extraction d'attributs ;
- L'extraction d'associations entre entités nommées et descripteurs ;
- L'extraction de correspondances multilingues.



19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unoffial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

Figure 07 : Exemple de formulaire d'extraction d'information pour des actes terroristes (MUC-3). [Ehrmann, 2008]

La reconnaissance des entités nommées (REN) apparaît comme une composante essentielle dans plusieurs domaines du Traitement Automatique des Langues Naturelles (TALN) : analyse syntaxique, résolution de coréférence, traduction automatique, recherche d'information, etc. Cette tâche relativement récente s'est complexifiée au fil du temps sur différents axes : taxonomie, modalité, langue, etc.

Les travaux menés en traitement automatique des langues (TAL) ont porté une attention particulière aux noms propres de personnes, de lieux et d'organisations, appelée des entités nommées. Ces éléments semblant être utiles à diverses tâches comme, par exemple, la recherche d'information.

La reconnaissance des entités nommées est une sous tache de l'extraction d'informations qui prend en entrée un bloc de texte non annoté et produit un bloc de texte annoté contenant les entités nommées trouvées. Chaque entité reçoit une étiquette en fonction de son type sémantique.

La reconnaissance des entités nommées consiste à :

- Identifier des unités lexicales dans un article Wikipédia.
- Les catégoriser.
- Eventuellement, les normaliser.

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. Plusieurs systèmes de REN sont présentés en détails dans Poibeau (2001), Friburger (2002), Nadeau et Sekine (2007), Fourour (2004). Nous ne présenterons ici qu'un seul système pour chaque type d'approche.

2.3.2 Type des approches :

a. 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. Par exemple, si le mot déclencheur « Monsieur » (connaissance issue d'un lexique) 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 les textes bien formés (Galliano et al. 2009). Plusieurs systèmes utilisant cette approche ont été développés pour différentes langues telles que l'anglais (McDonald 1996, Wacholder et al. 1997, Kim et Woodland 2000) et le français (Brun et Ehrmann 2010, Stern et Sagot 2010, Maurel et al. 2011). Nous présentons ici le système français Nemesis. [8]

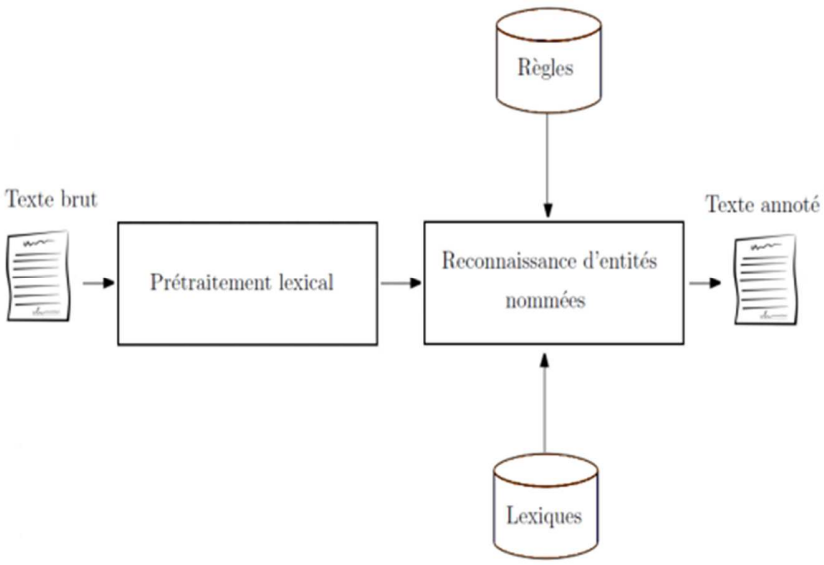


Figure 08 : Architecture général de Nemesis

Nemesis : 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 développé 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.

- Prétraitement lexical : segmentation du texte en occurrences de formes et de phrases, puis association des sigles à leur forme étendue.

- Projection des lexiques : les lexiques ont été construits soit manuellement, soit automatiquement à partir du Web. Les éléments composant ces lexiques (79 476 éléments) sont répartis en 45 listes selon les catégories dans lesquelles ils sont utilisés : prénom connu, mot déclencheur d'un nom d'organisation (l'élément fait partie de l'entité nommée : « Fédération française de handball »), contexte d'un nom de personne (l'élément appartient au contexte gauche immédiat de l'entité nommée, mais ne fait pas partie de celle-ci : « philosophe Emmanuel Kant »), fin d'un nom d'organisation (l'élément est la dernière forme composant l'entité nommée : « Conseil régional », « Coupe du monde de football »), etc. La projection des lexiques consiste à associer les étiquettes liées aux lexiques aux différentes formes du texte. Une forme peut avoir plusieurs étiquettes (Washington|prénom-connu|lieu-connu).

- Application des règles : les règles de réécriture permettent l'annotation du texte par des balises identifiant les entités nommées (délimitation et catégorisation). Elles sont basées sur des étiquettes sémantiques référant à une forme capitalisée ou à une forme appartenant à un lexique. En tout, Nemesis utilise 93 règles qui s'exécutent dans un ordre prédéfini. Lorsque plusieurs règles s'appliquent, Nemesis opte pour la règle ayant la priorité la plus élevée. Voici un exemple d'une règle de réécriture :

$$\$Clé\text{-}oronyme\ \$Article\text{-}min\ [\$Forme\text{-}capitalisée+] \rightarrow ORONYME$$

Et le résultat de son application :

« Montagne du Mont-Blanc »

Pour améliorer les performances de Nemesis, de nouveaux lexiques sont appris automatiquement en se basant sur un ensemble d'heuristiques (22 pour les patronymes et 3 pour les toponymes).

L'évaluation de Nemesis a été réalisée sur un corpus composé de textes issus du journal Le Monde et du Web (31 000 mots). Les performances sur
L'ensemble des entités nommées montrent un rappel de 79 % et une précision de 91 % (Fourour et Morin 2003).

b. Approches orientées données:

Les approches orientées données visent à apprendre les règles d'extraction de manière autonome. L'acquisition de ces règles ainsi que de certaines ressources de connaissances s'effectue à partir d'un corpus de grande taille de manière supervisée ou semi-supervisée.
L'apprentissage supervisé consiste à apprendre les règles à partir d'un corpus préalablement annoté. La supervision concerne l'intervention humaine, par le biais d'étiquetage d'une base d'exemples, afin de guider le système lors du processus d'apprentissage. Une méthode d'apprentissage est appliquée pour entraîner le système à exploiter les différents traits singularisant les entités nommées. Ensuite le système d'apprentissage généralise le processus afin de produire un modèle permettant d'extraire les entités nommées dans de nouveaux documents. La performance des systèmes orientés données augmente proportionnellement avec la quantité et la qualité du corpus d'apprentissage. Les différents systèmes de ce type se basent notamment sur les méthodes d'apprentissage suivantes : machines à vecteurs de support (SVM) (Isozaki et Kazawa 2002), modèle de Markov à états cachés (HMM) (Bikel et al. 1997), modèle de l'entropie maximale (EM) (Andrew et al. 1998), modèle de champs conditionnels aléatoires (CRF) (Béchet et Charton 2010, Dinarelli et Rosset 2011, Raymond 2013, Hatmi et al. 2013) et arbres de décision (Isozaki 2001). Nouvel et al. (2013) proposent une méthode de fouille de données séquentielle hiérarchique permettant d'extraire des motifs d'extraction d'entités nommées à partir d'un corpus annoté.

Nous détaillons ci-dessous un exemple de système orienté données.

Le système de Raymond et Fayolle (2010) : un système orienté données de REN pour le français

Raymond et Fayolle (2010) utilisent différents algorithmes d'apprentissage automatique (CRF, SVM, et transducteurs à états finis (FST)) pour la reconnaissance d'entités nommées dans les transcriptions de la parole. Outre les mots de la transcription, trois traits ont été exploités afin de mieux caractériser les entités nommées : les étiquettes morphosyntaxiques issues d'un processus d'étiquetage morpho-syntaxique (par exemple, nom propre, verbe, etc.), les connaissances issues des dictionnaires d'entités nommées et de mots déclencheurs, et l'importance du mot (un mot important est un mot dont l'information mutuelle partagée avec son étiquette d'entité nommée est supérieure à zéro et qui apparaît au moins trente fois dans le corpus d'apprentissage). Un corpus d'entraînement composé de 66h d'émissions radiophoniques francophones, dont 60h représentant le corpus d'apprentissage Ester 1 et 6h représentant le corpus de développement Ester 2, a été utilisé pour l'entraînement des différents classifieurs. Ce corpus a été transcrit et annoté manuellement. Le corpus de test représente 6h de transcriptions automatiques avec un taux d'erreur de mots de 26,09 % (corpus de test Ester 2). Les performances pour la reconnaissance d'entités nommées sont évaluées en termes de SER et de F-mesure.

Les résultats d'évaluation montrent que les méthodes discriminantes (28,90 % de SER et 81 % de F-mesure pour le modèle SVM et 28,10 % de SER et 80 % de F-mesure pour le modèle CRF) obtiennent des performances meilleures que les méthodes génératives (30,90 % de SER et 78 % de F-mesure pour le modèle FST).

Du fait que le corpus utilisé pour l'apprentissage est la fusion de deux corpus ayant une annotation légèrement différente (Ester 1 et Ester 2), les performances des différents classifieurs se voient affectées. Les auteurs proposent une méthode permettant de rendre les deux annotations cohérentes. Cela consiste à apprendre un modèle, en utilisant le système le plus performant (CRF), sur la concaténation des deux corpus. Ce modèle est utilisé ensuite pour réannoter le corpus d'entraînement.

C'est ce nouveau corpus qui va servir de référence pour entraîner les différents classifieurs. Les résultats d'évaluation montrent une amélioration significative : 26,60 % de SER pour le modèle FST, 26,60 % de SER pour le modèle SVM et 22,80 % de SER pour le modèle CRF.

L'apprentissage semi-supervisé (ou légèrement supervisé) permet de réduire le travail manuel et laborieux exigé par les approches à base d'apprentissage supervisé pour annoter un corpus d'apprentissage. L'apprentissage semi-supervisé demande un corpus non annoté de grande taille et un petit degré de supervision sous la forme d'un ensemble d'amorces des règles ou de mots. La technique principale utilisée pour l'apprentissage est connue sous le nom d'amorçage qui consiste à apprendre automatiquement les patrons d'extraction en se basant sur un ensemble d'amorces pertinents par rapport à la classe sémantique des entités à extraire. Par exemple, afin d'extraire les noms de pays ou de villes, le système demande à l'utilisateur d'en introduire quelques-uns (par exemple, France, Tunisie, Paris, Nantes, etc.).

Il analyse les phrases contenant les entités nommées introduites et retient des indices contextuels communs (par exemple, marqueurs lexicales, typographiques, etc.). Il essaye ensuite de trouver d'autres entités nommées qui se produisent dans le même contexte. En réappliquant ce processus sur de nouveaux textes, le système apprend des nouveaux contextes pertinents et des nouveaux noms de pays ou de villes. Les systèmes basés sur cette approche affichent des résultats sensiblement identiques à ceux obtenus par les systèmes utilisant une approche à base d'apprentissage supervisé.

Michael et Yoram (1999) exploitent un corpus non étiqueté du New York Times et un certain nombre d'amorces des règles pour repérer et catégoriser les entités nommées. Les résultats d'évaluation de leur système affichent 91,10 % de rappel et 83,10 % de précision. Richard (2003) propose un système de reconnaissance des entités nommées à domaine ouvert (indépendant du domaine). La typologie des entités nommées n'est pas spécifiée a priori mais elle est identifiée automatiquement en fonction du contexte des documents. Pour chaque mot ou groupe de mots commençant par une majuscule, le système soumet une requête à un moteur de recherche (Google) afin d'identifier ses hyperonymes potentiels.

Cette identification se base sur la méthode de Hearst (1992) concernant l'extraction d'hyponymes. Un exemple d'une telle requête est : such as X, avec X représentant une entité nommée commençant par une majuscule. Une fois collecté l'ensemble de ces hyperonymes, le système procède à leur classification afin d'identifier les catégories d'entités nommées qui apparaissent dans les documents. Le thésaurus WordNet est utilisé afin d'attribuer des noms des catégories aux classes résultantes.

L'évaluation de cette méthode affiche un rappel égal à 67,39 % et une précision égale à 46,97 % en utilisant un corpus en anglais composé de 18 852 mots dont 1 051 entités nommées. Dans un travail récent, Tahmasebi et al. (2012) proposent une méthode non-supervisée pour la détection des entités nommées qui évoluent dans le temps. Leur méthode se base sur le fait que les entités nommées qui appartiennent à une même catégorie sont susceptibles de se produire dans des contextes proches lorsque l'écart entre les périodes de changement n'est pas grand. Au début, les périodes de changement concernant un terme donné sont identifiées sur toute la collection. Ensuite, un ensemble des mots contextuels est généré pour chaque période afin de détecter les mentions coréférences qui se produisent dans le même contexte. L'évaluation de cette méthode montre un rappel important de 90 %.

Certains systèmes tirent profit des avantages respectifs des méthodes orientées connaissances et celles orientées données. Les règles sont, soient apprises automatiquement puis révisées

manuellement (Nagesh et al. 2012), soient écrites manuellement puis corrigées et améliorées automatiquement (Andrei et al. 1998, Oudah et Shaalan 2012). Nous présentons ici un système basé sur cette approche.

LTG : un système hybride de REN pour l'anglais

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

- Passage des règles les plus sûres (sure-fire rules) : ces règles sont appliquées seulement si les indices internes et externes permettent de classer le candidat sans ambiguïté. Elles se présentent sous forme d'une liste de mots déclencheurs et un ensemble de règles contextuelles utilisant un étiquetage en parties du discours.

- Première reconnaissance partielle (partial match 1) : une fois les règles les plus sûres appliquées, le système génère des variantes d'entités nommées déjà reconnues en changeant l'ordre des mots ou en en supprimant. Ensuite, un algorithme probabiliste fondé sur le modèle de maximisation de l'entropie est utilisé pour l'étiquetage des noms propres.

- Passage des règles plus lâches (rule relaxation) : des règles plus lâches en termes de contraintes contextuelles sont appliquées. Cette étape permet aussi de résoudre le problème des conjonctions et celui des entités en début de phrases.

- Deuxième reconnaissance partielle (partial match 2) : cette reconnaissance partielle annote les noms propres en utilisant le modèle d'entropie maximale.

- Traitement des titres des articles (title assignment) : des règles et un algorithme probabiliste sont utilisés pour la désambiguïsation des entités nommées situées dans les titres car ces derniers sont entièrement écrits en majuscules.

Les résultats obtenus par ce système à la compétition MUC-7 montrent une F-mesure de 93,39 %.

2.4 Plateformes de reconnaissance des entités nommées :

- **GATE** prend en charge REN dans de nombreux langages et domaines prêts à l'emploi, utilisable via une interface graphique et AOI Java.

- **OpenNLP** inclut la reconnaissance d'entités nommées statistiques et basées sur desrègles.

- **SpaCy propose** une REN statique rapide ainsi qu'un visualiseur d'entités nommées open source.

2.5 Conclusion :

Cette séparation nous a permis une partie des travaux d'exploitation de Wikipédia qui est la tâche de reconnaître une entité nommée. Et nous avons donné un aperçu général sur l'extraction

Conclusion Générale

La reconnaissance d'entités nommées est une tâche dont l'objectif est d'extraire et de typer des éléments informationnels à partir d'un texte donné. Des systèmes de reconnaissance de noms propres à base de ressources linguistiques.

L'objectif principal de cette contribution est de reconnaître les entités nommées de types (personne, localisations, organisations,), dans les articles Wikipédia.

La Recherche d'Information a pour objectif de fournir à un utilisateur un accès facile à l'information qui l'intéresse, cette information étant située dans une masse de documents textuels (les articles Wikipédia).

Dans le premier chapitre nous avons abordé les entité nommée (définition et quelques exemples) ensuit nous avons expliqué son Ambiguïtés et leurs différentes formes, puis notre intérêt qu'il est les Méthodes de reconnaissance et la recherche des entités nommées.

Ensuite, en passant au deuxième chapitre dont Nous avons donc abordé certains Travaux exploitant la Wikipédia, parmi eux la reconnaissance d'entités nommées à partir Wikipédia (REN), ensuite nous somme passer à expliquer le fonctionnement de système d'extraction et la recherche d'information, puis nous avons cité Type des approches et Plateformes de reconnaissance des entités nommées.

Bibliographie

[1]: Dbpedia. http ://dbpedia.org/About

[2]: Geonames. http ://www.geonames.org/

[3]: Unitex. http ://www-igm.univ-mlv.fr/ unitex/

[4] : Abbes, R. (2004). La conception et la réalisation d'un concordancier électronique pour l'arabe. Thèse de Doctorat, L'institut national des sciences appliquées de Lyon

[5]: Pustejovsky, 1995] Pustejovsky, J. (1995). The generative lexicon. The MIT Press, Cambridge

[6]: Abdelrahman, S., El-Arnaoty, M., Magdy, M., & Fahmy, A. (2010). Integrated Machine Learning Techniques for Arabic Named Entity Recognition. International Journal of Computer Science Issues,

[7] : Meryem Talha, Siham Boulaknadel, Driss Aboutajdine, « Système de reconnaissance des entités nommées amazighes », université Mohammed V-Agdal Rabat 4, Avenue Ibn Battouta, B.P. 1014 RP, 10006 Rabat, 21ème Traitement Automatique des Langues Naturelles, Marseille, France, 2014.

[8] : Abd Elkrim. Bouramoul, « recherche d'information contextuelle et sémantique sur le web », thèse de doctorat, université Mentouri Constantine, 2011.