Analyse syntactico-sémantique de la langue amazighe

Siham Boulaknadel
IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
E-mail: siham_06@yahoo.fr

Meryem TALHA
LRIT – CNRST URAC29, Université Mohammed V-Agdal Rabat
4, Avenue Ibn Battouta, B.P. 1014 RP, 10006 Rabat, Maroc
E-mail: meriem.talha@gmail.com

Abstract—L'augmentation des flux de l'information numérique nécessite l'extraction, le filtrage et la classification des informations pertinentes à partir de grands volumes de textes. Toutes ces tâches bénéficient amplement de l'implication de la Reconnaissance des entités nommées (NER) dans l'étape de prétraitement. La tâche NER implique non seulement l'identification des noms propres (entités nommées) dans le texte en langage naturel, mais aussi leur catégorisation en un ensemble de catégories prédéfinies, telles que les noms de personnes, les organisations (entreprises, organismes gouvernementaux, comités, etc), Lieux (villes, pays, fleuves, etc) et divers (titres de films, événements sportifs, etc.) Dans les langues anglaises et françaises, beaucoup de travaux on été effectués dans ce domaine, où la capitalisation est un indice important pour identifier les entités nommées, alors que la langue amazighe n'a pas cette particularité. Elle souffre d'un manque de disponibilité de ressources et de répertoires toponymiques, se caractérise par sa nature agglutinante, et sa morphologie flexionnelle, Ce qui rend la tâche de la reconnaissance plus difficile pour la langue amazighe. Les travaux sur les outils linguistiques automatiques sont une opportunité pour la valorisation de l'amazighe dans la société de l'information. C'est dans ce contexte que s'inscrit cette thèse qui consiste à construire des outils d'analyse syntactico-sémantique automatique indispensables pour l'analyse de textes amazighs. Dans cet article, nous décrivons les caractéristiques de la langue amazighe, nous présentons une étude des différentes approches pour la détection des entités nommées (NE).

Keywords: Reconnaissance des entités nommées; langue amazighe; approches NER;

I. INTRODUCTION

L'extraction d'entités nommées est une sous tâches du domaine d'extraction d'information. Concrètement, la notion de la reconnaissance d'entités nommées (NER) est apparue comme un outil de prétraitement important pour nombreuses applications du traitement de langage naturelle, telles que l'Extraction d'information (IE), la recherche d'information multilingue (IR), la traduction automatique (MT), systèmes de recherche de réponses précises à des questions « question-réponse » (QA), résumé automatique (AS), indexation automatique des documents (AI) entre autres applications de traitement de texte, parce que les entités nommées fournissent principalement des indices importants pour l'identification des informations pertinentes dans le texte. Elle vise à localiser (par exemple, « Le Chef du gouvernement, M. Abdel-Ilah Benkiran »), et classer les éléments de texte dans des catégories prédéfinies, c'est-à-dire le système va affecter l'étiquette personne au NE « Le Chef du gouvernement, M. Abdel-Ilah Benkiran » en utilisant des méthodes statistiques, fondées sur des règles ou hybrides.

Les entités nommées désignent traditionnellement les noms de personnes, de lieux, d'organisations mais aussi les expressions de dates ou les unités monétaires, les pourcentages, les fonctions et autres mais peuvent aussi se rapporter à des notions plus techniques comme les maladies. (Ehrmann, 2008) analyse la problématique des EN du point de vue théorique des difficultés définitoires et catégorielles, proposant la définition suivante : *« étant donné un modèle applicatif et un corpus, on appelle EN toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ».*

A. Caractéristiques de la langue amazighe :

Au Maroc, la langue berbère ou amazighe (ⵜⴰⵎⴰⵣⵉⵖⵜ), fait partie des langues chamito-sémitiques ou encore appelé afro-asiatiques (Cohen 2007, Chaker 1989) a toujours disposé d'un statut restreint et triplement minoré, bien qu'elle est considérée comme la plus ancienne langue d'Afrique du Nord, et fortement employé sur une aire très vaste par environ 50% (Boukouss, 1995),de la population marocaine, mais soumis à une pression conjuguée de l'arabe et du français, autrement dit, elle est régie par l'arabe standard (la langue officielle), l'arabe dialectal marocain et le français.

Sur le plan linguistique, la langue est caractérisée par la prolifération (Outahajala 2012) des dialectes en raison de facteurs historiques, géographiques et sociolinguistiques. Au Maroc, on peut classifier la langue berbère en trois grandes zones dialectales capitales dont les trois principales langues vernaculaires peuvent être identifiées: Tarifit (Nord du Maroc), Tamazight (centrale) et Tachelhit (Sud du Maroc). Pour pallier cette situation, vint la création de l'Institut Royal de la Culture Amazighe en 2001 (IRCAM), dans le but était de standardiser, redonner aux culture et langue amazighes la place qu'elles méritent, ainsi que pour uniformiser les structures et à adoucir les pluralités qui représentent des difficultés au niveau de l'intercompréhension.

L'Amazighe est maintenant une langue qui possède tous ses attributs: dotée d'une graphie officielle, un codage propre dans le standard Unicode, une grammaire, une orthographe ainsi qu'un vocabulaire très riche et une littérature orale fabuleusement fortuné. Elle est actuellement une langue nationale et depuis 2003, elle a été introduite dans le système éducatif comme matière inévitable dans les écoles primaires, quelque soit la langue maternelle des élèves, ainsi que dans l'administration et les médias.

• **Système d'écriture :**

-Le libyque : Elle est la plus ancienne, a deux formes : l'oriental et l'occidental. L'alphabet libyque est strictement consonantique.

-Le tifinaghe saharien : Libyco-berbère, Elle contient des signes additionnelles en comparaison avec le libyque, à savoir un trait vertical pour noter la voyelle finale.

-Le tifinaghe touareg : Il y a au niveau du tifinaghe touareg des pluralités dans la valeur accordée aux signes qui conviennent aux variations dialectales touarègues.

-Le néotifinaghe : Le néotifinaghe désigne les systèmes d'écriture développés pour représenter les parlers berbères (amazighes) du Maghreb.

En se référent au système d'origine, et grâce à l'IRCAM qui a développé un système d'alphabet nommé **Tifinaghe-IRCAM**, l'amazighe dispose de sa propre graphie, qui s'écrit horizontalement de gauche à droite.

L'IRCAM a gardé seulement les phonèmes pertinents pour le tamazight, de sorte que le nombre des entités alphabétiques phonétiques est de 33, à savoir **27 consonnes**, **2 semi-consonnes**, **3 voyelles pleines** et **une voyelle neutre**, mais les codes Unicode seulement 31 lettres, plus un modificateur de lettre pour former les deux entités phonétiques: ⴳ ⵯ (gw^w) et ⴽ ⵯ (kw^w). Elle a également préconisé l'utilisation des signes de ponctuation usuelle, qu'on utilise déjà dans les écritures latines, et a adopté pour ces alphabets le tri suivant :
ⴰ< ⴱ < ⴳ< ⴷ< ⴹ< ⴻ< ⴼ< ⴽ< ⵀ<ⵃ< ⵄ< ⵅ< ⵇ< ⵉ< ⵊ< ⵍ< ⵎ< ⵏ< ⵓ< ⵔ< ⵕ< ⵖ< ⵙ< ⵚ< ⵛ< ⵜ< ⵟ< ⵡ< ⵢ< ⵣ< ⵥ.

• **Encodage Unicode :**

L'ensemble des lettres tifinagh comprend quatre sous-ensembles à savoir: le jeu de base de l'IRCAM, le jeu étendu de l'IRCAM, d'autres lettres néotifinaghe en usage, des lettres touarègues modernes dont l'usage est attesté. Le nombre atteigne 55 caractères. L'ensemble de caractères choisis par l'IRCAM se présentent dans les deux premiers sous-ensembles.

Avant de pouvoir commencer les démarches pour la création d'un système d'extraction d'entités nommées pour l'amazighe, plusieurs particularités de cette dernière doivent être prises en compte vu leurs impacts sur la création du système.

• **Morphologie :**

La langue amazighe, dotée d'une morphologie riches, peut être considérée comme une langue complexe, dont les mots peuvent être répertoriés en trois catégories morpho-syntaxiques (Boukhris et al., 2008).

Nom :

Le nom est une unité lexicale formée d'une racine et d'un schème, Il peut être représenté par diverses formes, soit une forme simple (ⴰⵔⴳⴰⵣ "homme"), une forme composée (ⴱⵓⵀⵢⵢⵓⴼ "la famine") ou dérivée (ⴰⵎⵙⴰⵡⴰⴹ "la communication"). Il varie en genre (féminin ⵜⴰⵎⵃⴹⴰⵕⵜ (une élève), masculin ⴰⵎⵃⴹⴰⵕ (un élève)), en nombre (singulier ⴰⵎⵃⴹⴰⵕ, pluriel ⵉⵎⵃⴹⴰⵕⵏ) et en état (libre (EL) ⴰⵎⵃⴹⴰⵕ, annexion (EA) ⵓⵎⵃⴹⴰⵕ).

(i) Genre : L'amazighe connaît deux genres : **le masculin** et **le féminin.**

Les noms masculins : Le nom masculin commence en général par une des voyelles initiales a(a), i (i), ou u(u). Les noms commençant par la voyelle a sont de loin les plus nombreux. (Par exemple : ⴰⴼⵓⵙ "main", ⵉⵍⵙ "langue", ⵓⴷⵎ "visage", ⴰⵔⴳⴰⵣ "homme").

Les noms féminins : La marque du féminin en amazighe est le morphème discontinu ⵜ....ⵜ. En général, on peut obtenir le féminin à partir du radical d'un nom masculin. (Par exemple : ⴰⴳⵎⴰⵔ "cheval" → ⵜⴰⴳⵎⴰⵔⵜ "jument", ⵉⵙⵍⵉ "marié" → ⵜⵉⵙⵍⵉⵜ "mariée"). Mais il ya toujours des exceptions, pour certains noms, le féminin n'est pas formé directement sur le même radical que le masculin, donnant quelques exemples :

ⴰⵔⴳⴰⵣ "homme" → ⵜⴰⵎⵟⵟⵓⵜ "femme"
ⴰⵣⴳⵔ "taureau" → ⵜⴰⴼⵓⵏⴰⵙⵜ "vache"

(ii) Nombre : L'amazighe a un singulier et un pluriel. Ce dernier peut prendre plusieurs formes à savoir externe, interne, mixte et le pluriel en ⵉⴷ [id].

Le pluriel externe (ajout d'un suffixe) : Le nom ne subit aucune modification interne, seule la voyelle initiale ⴰ se transforme en ⵉ. En outre, le pluriel externe est obtenu par l'addition du suffixe ⵏ pour les noms masculins et ⵉⵏ pour les noms féminins. Par exemple : ⴰⵅⵅⴰⵎ → ⵉⵅⵅⴰⵎⵏ "maisons", ⵜⴰⵔⴱⴰⵜ → ⵜⵉⵔⴱⴰⵜⵉⵏ "filles".

Le pluriel interne (ou brisé) : En plus de l'alternance vocalique initiale, le pluriel se forme par un changement de voyelles internes. Aucun suffixe ne vient s'adjoindre à la forme de base (singulier). Par exemple : ⴰⴷⵔⴰⵔ → ⵉⴷⵓⵔⴰⵔ "montagnes".

Le pluriel mixte: soit il est formé par une alternance d'une voyelle interne et/ou d'une consonne plus une suffixation par n => (ⵉⵍⵉ "part" → ⵉⵍⴰⵏ "parts"). Ou bien par une alternance vocalique initiale accompagné d'un changement vocalique final ⴰ [a] plus une alternance interne => (ⴰⵎⴳⴳⴰⵔⵓ → ⵉⵎⴳⴳⵓⵔⴰ "derniers").

Pluriel en ⵉⴷ : Une catégorie de noms forme son pluriel par la préfixation de ⵉⴷ au singulier. Ce procédé est utilisé pour des noms à initiale consonantique, les noms de parenté, des noms propres, les noms composés en ⴱⵓ - / ⵎⵎⵓ - ou ⴱⴰⴱ / ⵍⴰⵍ et les emprunts non intégrés.

Par exemple : ⵅⴰⵍⵉ "oncle (mon)" → ⵉⴷ ⵅⴰⵍⵉ, ⴱⵓ ⵉⵙⵍⵎⴰⵏ "poissonnier" → ⵉⴷ ⴱⵓ ⵉⵙⵍⵎ, ⵍⴽⴰⵎⵢⵓ "camion" → ⵉⴷ ⵍⴽⴰⵎⵢⵓ.

Pluriel des noms empruntés et intégrés : Les emprunts intégrés forment leur pluriel comme les noms amazighes.

Par exemple : ⵜⴰⵡⵔⵉⵇⵜ → ⵜⵉⵡⵔⵉⵇⵉⵏ "feuilles", ⴰⴼⵔⵎⵍⵉⵢ → ⵉⴼⵔⵎⵍⵉⵢⵏ "infirmiers", ⴰⴱⵓⵇⴰⴷⵢⵓ → ⵉⴱⵓⵇⴰⴷⵢⵓⵜⵏ "sandwichs".

(iii) Etat : On distingue deux états en amazighe : l'état libre et l'état d'annexion (l'état construit).

Etat Libre (EL) : La voyelle initiale du nom ne subit aucune modification. Le nom est en état libre lorsqu'il est:

- Un mot isolé de tout contexte syntaxique ⴰⵜⴱⵉⵔ "pigeon".

- Un complément d’objet direct « ⵉⵟⵟⴼ ⴰⵙⵍⵎ ⴳ ⵓⴼⵓⵙ » il tient un poisson à la main.
- Un complément de la particule prédictive ⴷ = c’est : ⴷ ⴰⵙⵍⵎ "c'est un poisson".

Etat lié (EA) : qui se manifeste par une variation des noms à initiale vocalique.

- alternance vocalique ⴰ / ⵓ pour les noms masculins => ⴰⵔⴳⴰⵣ → ⵓⵔⴳⴰⵣ.
- chute de la voyelle initiale pour les noms féminins => ⵜⴰⵎⵖⴰⵔⵜ → ⵜⵎⵖⴰⵔⵜ
- addition d’un ⵡ ou ⵢ aux noms à voyelle ⴰ ou ⵉ => ⴰⵙⵉⴼ → ⵡⴰⵙⵉⴼ
- maintien de la voyelle initiale ⴰ avec apparition de la semi-consonne ⵡ pour le masculin, les noms féminins restent sans modification. Par exemple :
  Masculin => ⴰⵙⵙ "jour" → ⵡⴰⵙⵙ
  Féminin => ⵜⴰⵙⴰ "foie" → ⵜⴰⵙⴰ

L’état d’annexion se réalise dans les contextes syntaxiques suivants :

- Le sujet lexical suit le verbe => ⵢⵓⵙ ⴷ ⵓⵙⵍⵎⴰⴷ → "Le professeur est venu".
- Après une préposition => ⵙⵉⵡⵍⵖ ⵉ ⵓⵙⵍⵎⴰⴷ → "J'ai parlé au professeur".
- Après un coordonnant => ⴰⵏⵥⴰⵕ ⴷ ⵓⵚⵎⵎⵉⴹ → "la pluie et le froid".

Verbe :

Le verbe peut prendre deux aspects : une forme simple ou dérivée, qu'il soit simple ou dérivé, il se conjugue à l'un des quatre thèmes: l’aoriste, l’inaccompli, l’accompli positif et l’accompli négatif.

Le verbe simple est composé d’une racine et d’un radical. Par contre Les verbes dérivés sont obtenus à partir des verbes simples par la préfixation de l'un des morphèmes suivants : ⵙ/ⵙⵙ, ⵜⵜ et ⵎ / ⵎⵎ.

- **Les dérivés en ⵙ/ⵙⵙ** : Ils expriment la causativité, autrement dit l’idée de "faire faire" une action à un patient ou celle de "faire devenir". Par exemple :

ⵎⵓⵏ "accompagner" → ⵙⵎⵓⵏ "faire accompagner", ⴽⵛⵎ "entrer" → ⵙⵙⴽⵛⵎ "faire entrer".

- **Les dérivés en ⵜⵜ** : on procède à la préfixation de ⵜⵜ au verbe simple pour exprimer la forme passive. Ce morphème peut se présenter sous forme de ⵜⵜⵓ, ⵜⵜⵡⴰ ou ⵜⵜⵢⴰ. Par exemple :
  ⴽⵔⴼ "attacher" → ⵜⵜⵡⴰⴽⵔⴰⴼ/ⵜⵜⵓⴽⵔⴰⴼ "être attaché"
  ⴰⵎⵥ "tenir" → ⵜⵜⵡⴰⵎⵥ / ⵜⵜⵢⴰⵎⴰⵥ "être tenu"
- **Les dérivés en mm** : Un verbe dérivé par la préfixation de ⵎ ou ses variantes (ⵎⵎ, ⵎⵢ) exprime la réciprocité **:** ⵥⵕ "voir" → ⵎⵎⵥⵕ "se voir".

Les particules :

Les particules sont un ensemble de mots bien évidement amazighe, en général, assez courts qui jouent le rôle d'indicateurs grammaticaux au sein d’une phrase.

Cet ensemble contient plusieurs formes: les particules d’aspect, d’orientation et de négation; les pronoms indéfinis, démonstratifs, possessifs et interrogatifs; les pronoms personnels autonomes, affixes sujet, affixes d’objet direct et indirect, compléments du nom ordinaire et de parenté, compléments de prépositions; les adverbes de lieu, de temps, de quantité et de manière; les prépositions; les subordonnants et les conjonctions.

Généralement, les particules sont invariables. Or, dans le cas de l’Amazighe, similairement au cas du français, il existe des particules flexionnelles telles que les pronoms possessifs (ⵡⵉⵏⵏⵙ “le sien” → ⵡⵉⵏⵏⵙⵏ “le leur”).

B. Difficultés entravant le traitement automatique de la langue amazighe :

Les travaux liés à NER dans la langue amazighe sont encore à l'état naissant en raison des faits potentiels:

(i). L’absence de la distinction majuscule/minuscule : c’est un obstacle majeur pour la langue amazighe. En fait, la REN pour certaines langues comme les langues indo-européennes se base principalement sur la présence des lettres majuscules qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l’alphabet latin. Les lettres majuscules, néanmoins, ne se produisent pas, ni au début ni à l’initiale des noms propres amazighe.

(ii). Il est un fait que la langue amazighe est très agglutinative ayant une morphologie dérivationnelle et flexionnelle assez complexe et riche, les noms peuvent avoir plusieurs formes fléchies et dérivées, la simple suppression des suffixes ne peut suffire à regrouper des familles de mots. En effet, dans la pratique les affixes peuvent altérer le sens d’un mot. Donc, le fait de les éliminer peut engendrer une perte d’information.

(iii). Similairement à d’autres langues naturelles, l’amazighe présentent des incertitudes au niveau des classes grammaticales. En effet, la même forme convient à nombreuses catégories grammaticales, cela dépend du contexte dans la phrase. Par exemple, (ⵉⵍⵍⵉ) peut être considéré comme verbe à l’accompli négatif, il signifie «il existe», ou comme nom de parenté « ma fille ».

(iv). L’amazighe se caractérise par le manque de ressources dictionnairiques (de noms etc.), de répertoires toponymiques, de ressources langagières et outils du TAL, à savoir les analyseurs morphologiques, POS taggers.

(v). Manque de corpus annotés qui ont une très grande importance dans diverses tâches du traitement automatique des langues.

(vi). Le nombre de mots fréquemment utilisés comme noms communs et qui peuvent être, également, utilisés comme noms propres est très grand.

(vi). Les substantifs, les adjectifs, les verbes, les pronoms, les adverbes, prépositions, les particules… etc, consistent en un seul mot se produisant entre deux blancs ou des signes de ponctuation (Ameur et al. 2006). Cependant, si une préposition ou un nom de parenté est suivie par un pronom personnel, à la fois la préposition/nom de parenté et le pronom qui suit, forment une chaîne unique délimitée par des espaces ou des signes de ponctuation. Par exemple: (ⴳⵔ) signifiant « pour, au », (ⵉ) qui signifie « moi » (pronom personnel), donnent (ⴳⴰⵔⵉ ou ⴳⵓⵔⵉ) ».

(vii). L’état d’annexion :

(viii). Les noms propres au niveau de la langue amazighe sont extrêmement nombreux, ont de nombreuses variantes ainsi qu’ils sont difficiles à détecter sans la présence d’un lexique, et c’est également le cas pour les noms d’organisations ou de produits, les noms de lieux même si ces derniers sont relativement stables par rapport aux précédents qui subissent fréquemment des changements. L’amazighe a une orthographe instable, par exemple : Mohammed peut s’écrire: ⵎⵓⵃⴰⵎⵎⴰⴷ, ⵎⵓⵃⴰⵎⵎⵓⴷ, ⵎⵓⵃⴰⵎⴰⴷ, ⵎⵃⴰⵎⴰⴷ, ⵎⵃⴰⵎⴷ, ⵎⵃⵎⵎⴷ, ⵎⵃⴰⵎⵎⴷ…

(ix). Les noms propres, de lieux ou d’organisations ne figurent pas tous dans des corpus, (notamment pour dans le cas d’apprentissage automatique), ces derniers ne sont pas capables de décrire tout le domaine des noms propres existants qui se développe sans cesse, augmente et qui est extrêmement étendu, les noms propres sont nommée la catégorie ouverte.

(x). les mots ambigus (‘ⵟⵉⵣⵜⵓⵍ’ peut-être une personne, un lieu ou une organisation) introduiront méthodiquement un taux d’erreur directement lié à leur degré d’ambiguïté.

Les particularités que nous avons décrites dans cette partie et le fait d'avoir étudié quelques traits linguistiques de la langue amazighe, ainsi que les difficultés posées par l'orthographe et la morphologie, vont servir directement lors de l’établissement d’un système de repérage des entités nommées. Le fait d'avoir étudié quelques traits linguistiques de la langue amazighe, ainsi que les difficultés posées par l'orthographe et la morphologie.

Nous citerons, une liste de fonction utilisée pour identifier les entités nommées :

Suffixes, Préfixes des mots, les prénoms, noms de familles, deuxième prénom, chiffes, noms de lieux, mois, jour, année, longueur, mots numériques, les mots vides, caractéristiques numériques, racines de mots, étiquetage dynamiques des entités nommées, listes des répertoires toponymiques, Parts of Speech Information (POS) etc.

II. APPROCHES D’EXTRACTION D’ENTITÉS NOMMÉES

La REN constitue un champ de recherche très actif depuis de nombreuses années dans d’autres langues. Les approches automatiques, initialement élaborées par une analyse à l'aide de connaissances humaines, sont aujourd'hui également guidés par les données, généralement à l'aide de modèles numérique automatiquement paramétrés (maximum d'entropie **[1]**, HMM, CRF **[2]**, SVM). Couramment, des dictionnaires (listes) de noms propres, des prétraitements linguistiques (morphosyntaxiques) sont utilisés pour enrichir les textes en entrée afin d'y reconnaître les entités.

Deux approches fondamentales coulent, à savoir l’extraction fondée sur des démarches statistiques à base d’apprentissage/ statistiques d’une part et sur une fondée sur des connaissances explicites appelée démarche linguistique, aussi nommée symbolique/à base de règle, d’une autre part, avec une apparition d’une nouvelle approche hybride qui combinent les avantages des deux autres approches.

A. Approches à base de règles :

La première approche exploite les avancées en TAL et s’appuie spécialement sur l'utilisation de grammaires formelles construites par la main d'un expert-linguiste. Ils se fondent particulièrement sur la description des EN grâce à des règles qui exploitent un étiquetage syntaxique, des marqueurs lexicaux et des dictionnaires de noms propres, L'étiqueteur syntaxique retient en entrée un texte et réalise automatiquement en sortie une version étiquetée de ce même texte.

L’annotation syntaxique consiste à produire les propriétés grammaticales d'un mot ou d'un groupe de mots dans une phrase donnée (ex. noms, verbes, conjonctions) souvent accompagnée d'informations morphologiques (ex. : genre, nombre, personne). Les marqueurs lexicaux sont aussi appelés mots amorces ou mots déclencheurs ou parfois preuves contextuelles. Il s'agit de mots ou d'indices qui entourent (à gauche ou à droite) le nom propre et qui permettent souvent de prédire sa présence.

Par exemple, les deux règles suivantes sont présentes dans beaucoup de systèmes à base de règles, pour la langue française.

- Toute entité nommée de type personne sera présenté comme suit :
  Article défini + titre de civilité (ex. : M. Mme, Mlle. Mr., Dr.) + Mot majuscule.
  Ex. : Le Pr. Amrani déclare sa démission.
- Un second exemple reflète que toute entité nommée de type lieu sera présenté comme suit:
  Article défini + Marqueur de lieu (ex. ville) + préposition / de / + mot majuscule.
  Cet exemple concerne l'utilisation d'un article défini (ex. le), d'un marqueur lexical (ex. ville) et d'une information de l'étiqueteur syntaxique (ex. prépositions).
  Ex. : Ma famille habite la ville de Rabat.

Finalement, les dictionnaires de noms propres assemblent généralement d'une liste des noms et des prénoms les plus courants, des noms de lieux (villes, pays, etc.) et parfois des noms d'organisations (organismes, compagnies, etc.). Ces dictionnaires sont souvent usés dans les systèmes symboliques comme ils peuvent nous servir dans les systèmes automatiques.

Les prétraitements cités plus haut servent de base à la construction de règles et patrons linguistiques qui définissent les contextes d'apparition de telle entité ou relation.

Plusieurs systèmes ont été réalisés dans le cadre de ces approches à savoir, celui de Senellart (1998), Poibeau (1999).

B. Approches statistiques :

La seconde approche utilise des techniques statistiques pour « apprendre » des régularités sur de larges corpus de textes où les entités-cibles ont été préalablement annotées. Cette approche a été conçue pour avoir une certaine intelligence lors de la prise des décisions, contrairement aux approches linguistiques qui n’appliquent que les règles injectées auparavant.

Pour entraîner un système, il faut essentiellement avant tout lui procurer des données, nommées souvent corpus d’apprentissage, et par la suite adapter un algorithme d’apprentissage qui va permettre d’élaborer automatiquement une base de connaissances. Dans ce cas le système serait capable de prendre une décision tout seul d’une manière autonome.

Les chercheurs, de nos jours, penchent populairement vers l’utilisation des méthodes d'apprentissage automatique pour les REN car ces dernières sont facilement adaptables à différents domaines et langues ainsi que leur maintenance est aussi moins coûteuse.

Plusieurs systèmes d'apprentissage automatique ont été traités depuis quelques années. Nous distinguons les méthodes d'apprentissage selon qu'elles soient supervisées, semi-supervisé ou non-supervisé.

- **Méthodes supervisées :**

Il s'agit d'une méthode d'apprentissage qui nécessite une intervention humaine importante, et requiert tout de même d’une manière globale, un volume considérable de données pour les nécessités de l’apprentissage. Elles Consistent à faire un apprentissage à partir d'un corpus d'entraînement, préalablement préparé, parmi ces méthodes on peut citer :

- **HMM (Hidden Markov Model) :**

Un modèle HMM (Hidden Markov Model) est un modèle stochastique particulier, génératif on modélise comment obtenir x alors que les données x sont disponibles, proposé pour la première fois à la fin des années 60 début des années 70 (Baum. 1970, Baker. 1975), Il représente un objet donné par deux suites de variables aléatoires : l’une dite caché et l’autre observable. Le premier processus génère la suite caché correspond à la suite d’états q1, q2, ..., qT, le deuxième HMM est responsable de la génération de la suite observable correspond à la suite d’observations o1, o2,..., oT à partir de la séquence d'états. Un HMM se définit comme une structure composée d’un ensemble d’états, de graphes de transitions et de distributions de probabilités sur les transitions. En effet, Les HMMs garantissent une fusion cohérente de différents niveaux de modélisation à savoir celle morphologique, lexicale et syntaxique.

Sans s’approfondir dans les détails des formules mathématiques derrière cette méthode, nous présenterons succinctement le principe de cette approche.

Supposant qu’on a choisit au départ un modèle pour chaque catégorie d’entités nommées, sous entendu l’apprentissage se fait à partir d’un corpus d’apprentissage, où chaque élément du texte, inclusivement les entités nommées, a subit une annotation et catégorisation.

La première étape de l'apprentissage consiste à sauvegarder en mémoire un modèle pour chaque mot dans le corpus d'apprentissage. Prenons le cas où un modèle HMM rencontre la phrase suivante dans un corpus de test :

ⵉⴷⴷⴰ ⴰⵔ ⴰⵎⵣⵎⵉⵣ = « il est allé jusqu’à Amzmiz. »

Le mot « ⴰⵎⵣⵎⵉⵣ » qui suit immédiatement la préposition « ⴰⵔ » n’a pas été reconnu par le modèle HMM, vu que ce dernier n'était pas observé dans le corpus d'apprentissage.

Dans ce cas, le modèle HMM va se servir du contexte antécédent afin d'analyser le mot inconnu.

Le contexte antécédent dans notre cas consiste en la préposition du lieu « ⴰⵔ ». Supposant que dans notre corpus d'entraînement, la préposition « ⴰⵔ » est souvent suivie d'un nom de lieu. Dans ce cas, le système va affecter la catégorie nom de lieu au mot « ⴰⵎⵣⵎⵉⵣ » en se référant sur le modèle probabiliste du contexte d'apparition de la préposition « ⴰⵔ » dans le corpus d'apprentissage.

- **CRF (Conditional Random Field) :**

Conditional Random Fields (CRF) **[3]** ou Champs Markoviens Conditionnels sont une famille de modèles statistiques graphiques non orientés non génératif, également connu sous le nom de champ aléatoire, un cas particulier de ce qui est une chaîne linéaire qui correspond à une condition formés de machine à états finis. En se basant sur les valeurs affectées à d'autres nœuds d'entrée assignées, ce modèle est utilisé pour calculer la probabilité conditionnelle des valeurs sur les nœuds de sortie affectés. CRF modèle est très utile dans la reconnaissance d'entités nommées.

Champ aléatoire: Soit G = (Y, E) un graphe où chaque sommet YV est une variable aléatoire. Supposons que P (Yv | tout autre Y) = P (Yv | voisins (Yv)), alors Y est un champ aléatoire.

L’un des travaux pionniers de la NER pour les langues indiennes a été réalisé par Li and McCallum (2004) **[4]**, où ils ont utilisé le CRF. Dans leur étude, la taille du corpus d’apprentissage était de 340K mots avec 15,063 entités nommées appartenant à trois types, nommés personne, location et organisation. Cette méthode (CRF) a été introduite par d’autres chercheurs (Settles, 2004; Tsai et al., 2006; Leaman and Gonzalez, 2008).

- **Maximum Entropy (ME) :**

Introduite par (Berger et al., 1996) communément appelé MaxEnt, Avec les contraintes basées sur les attributs choisis fj(x, y), la méthode du maximum d’entropie va chercher à maximiser la probabilité conditionnelle de fj(x, y). Des contraintes sont obtenues à partir des données d’apprentissage, il exprime des relations entre caractéristiques et des résultats de sorties. La probabilité de distribution, qui est satisfaite des propriétés au-dessus, est la plus haute entropie. Il fait l’hypothèse qu’elle suit une loi exponentielle.

- **SVM (Support Vector Machine) :**

Les machines à vecteurs de support représentent une famille d'algorithmes d'apprentissage qui permet de remédier à des problèmes de classification, ou de régression, en se référant aux résultats de la théorie de l'apprentissage statistique introduite par Vladimir Vapnik 1995. SVM est une méthode de classification binaire par apprentissage supervisé c'est-à-dire qu’on connait à l’avance les classes auxquelles appartiennent les échantillons d’apprentissage, et la classification se fait en se basant sur les bases prédéfinies. L’utilisation des SVM est très fréquente pour la tâche de reconnaissance d’entités nommées ainsi que d’autres tâches de traitement du langage naturel. Plusieurs recherches on été entamés dans ce sens, Takeuchi et Collier (2002) **[5]** ont utilisé le SVM dans la tâche NER avec une représentation binaire des caractéristiques. Isozaki et Kazawa (2002) **[6]** ont également utilisé le SVM, ils ont proposé quelques techniques comme la suppression de caractéristiques inutiles pour rendre le classificateur efficace en termes de temps d’apprentissage et de performance.

- **Decision Tree (DT) :**

Il s'agit des méthodes les plus célèbres dans le domaine d'apprentissage supervisé, ils représentent graphiquement sous forme d'arbre, un ensemble de règles et sont aisément interprétables, Pour les arbres de grande taille, la procédure globale peut être difficile à appréhender, cependant, la classification d'un élément particulier est toujours compréhensible. Chaque branche de l'arbre en question convient à une décision et dispose d'un certain poids et d'une certaine probabilité pour prendre une décision donnée, elle réclame l'utilisation d'un corpus d'apprentissage. Cette méthode a été utilisée par Sekine 1998 **[7]** pour la langue japonaise.

Un des avantages de cette méthode c'est que leur structure arborescente permet la génération automatique d'un ensemble de règles facilement lisibles et compréhensibles par un linguiste, contrairement à d'autres approches.

- **Méthodes non-supervisées :**

Cette méthode se distingue des autres méthodes d'apprentissage supervisé par le fait que l'algorithme procède à un apprentissage avec une intervention humaine minimale. Les techniques se reposent fondamentalement sur des ressources lexicales (par exemple, WordNet), abordé par E. Alfonseca and Manandhar (2002), sur des modèles lexicaux traité par R. Evans (2003) et sur les statistiques obtenues sur un large corpus non annoté. Autres recherches ont été effectués dans le même cadre par Y. Shinyama and Sekine (2004), et O. Etzioni et al. (2005). La méthode non supervisée utilise, entre autres, la technique du « Clustering » qui consiste à regrouper automatiquement les entités similaires. Selon Candillier (2006), l'apprentissage non supervisé « consiste à former différents groupes à partir d'un ensemble de données, de telle manière que les données considérées comme les plus similaires soient associées au même groupe et qu'au contraire les données considérées comme différentes se retrouvent dans des groupes distincts, permettant ainsi d'extraire de la connaissance à partir de ces données ». Les EN seront réunies automatiquement en sous-ensembles selon la ressemblance de leur contexte.

- **Apprentissage paresseux :**

Cette méthode a été populaire lors de la MUC-6, elle consiste à l'enregistrement des données du corpus d'entraînement dans la mémoire et fait par la suite à la comparaison des nouvelles requêtes avec celles qu'il vient de conserver en mémoire, ensuite, à l'aide d'une formule mathématique, il calcule le degré de similarité des lettres composant les deux requêtes. Finalement le système va affecter à la nouvelle requête la classe de l'exemple d'apprentissage le plus proche.

Prenons un exemple inventé qui un corpus d'apprentissage qui serait restreint juste à deux éléments:

Liste des éléments dans le corpus	Catégorie attribuée d'entraînement manuellement
ⴰⵎⵣⵎⵉⵣ	Ville
ⵎⵉⵎⵓⵏ	Personne

Voyons donc les nouvelles requêtes suivantes:
Requête 1 : ⵓⵎⵣⵎⵉⵣ
Requête 2 : ⴰⵎⵣⵎⵉⵣ
Requête 3 : ⵙⵉⵎⵃⵓⵏ
Requête 3 : ⵎⵓⵖⴰ

Suite au calcul de la distance entre ces 4 requêtes et les 2 éléments dans le corpus d'apprentissage, le système va affecter aux nouvelles requêtes, les catégories les plus probables en respectant un degré de ressemblance bien précis.

- Résultat de la première requête : « ⵓⵎⵣⵎⵉⵣ », mot inconnu, touche un degré de ressemblance orthographique avec « ⴰⵎⵣⵎⵉⵣ ». Il serait affecté dans la catégorie ville.
- Résultat de la deuxième requête : « ⴰⵎⵣⵎⵉⵣ », mot existant dans le corpus d'entraînement, il va être attribué dans la classe ville.
- Résultat de la troisième requête : « ⵙⵉⵎⵃⵓⵏ », mot n'est pas connu, et pourtant il est proche d'un point de vue orthographique de « ⵎⵉⵎⵓⵏ », il reçoit donc la classe personne.
- Résultat de la quatrième requête : « ⵎⵓⵖⴰ », mot n'est pas reconnu, en outre il n'est pas proche d'aucun éléments du corpus, alors il ne reçoit aucune catégorie.

Après l'examen des résultats de sortie, on constate que la méthode de calcul de ressemblance entre les exemples précédents est assez rigoureuse, puisqu'elle ne permet de classer que les mots inconnus qui sont très proches des mots existants dans le corpus.

- **Apprentissage semi-supervisé** :

Cette méthode est parfois nommée *apprentissage limité* **[8]**. Parmi les principales techniques employées, nous mentionnons le ***bootstrapping*** Borthwick (1999). Cette méthode ne nécessite qu'un nombre restreint de données injectées (***appelée semence***) auparavant pour marcher convenablement. Elle peut être sous forme d'une liste de noms de personnes par exemple. Le système procède à l'analyse des phrases qui sont dotées d'un certain type d'entités nommées, par la suite, le système garde les marqueurs lexicaux de ces entités. En répétant ce processus, un grand nombre de noms de personne peut être détecté. Ce principe a été utilisé par S. Brin (1998), Riloff et Jones (1999), Collins et Singer (1999), Yanarber et al. (2000), Nadeau et al. (2006) et M. Pasca et al. (2006).

Les règles abordées sont très facile.

« **Lausanne** » est un lieu; « **Rouen** » est un lieu; « **France** » est un lieu; tout nom qui contient « **Mr.** » est un nom de personne; tout nom qui contient « **Incorporated** » ou « **Inc**. » est un nom d'organisation; « **IBM.** » est une organisation; « **Microsoft** » est une organisation;

Ces règles précédentes, vont permettre de déduire de nouvelles règles. Prenons l'exemple suivant :

Mr. Ahmed.

Le système infère que Ahmed est un nom de personne, car il est précédé par le titre « **Mr.** » qui prédit un nom de personne dans les règles et ainsi et suite.

C. Approches Hybrides

Enfin, il y a les approches dites mixtes, appelées aussi hybrides **[9, 10, 11]**, combinant l'apprentissage automatique (à partir d'un corpus) et l'écriture des règles à la main (création de règles à la main). Les limites de chacune des approches, ont amené les experts du domaine à fusionner les méthodes existantes pour augmenter les performances de leurs outils. Les approches symboliques, souffrent également du coût de leur développement manuel et de la nécessité d'un expert en linguistique pour pouvoir les modifier et les adapter. Pour remédier à ce problème, les experts se penchent vers des méthodes d'apprentissage automatique de patrons linguistiques. Pourtant, les méthodes statistiques nécessitent, lors de la phase d'entraînement, une grande quantité de textes pré-étiquetés autrement dit un large corpus annoté, vu qu'on ne dispose pas toujours de ces données, cela constitue un réel problème. Ce qui a engendré l'apparition des systèmes hybrides qui tentent de palier problèmes de deux autres approches.

Ces systèmes ont été utilisés par le système suédois Swenam fait par Dalianas et Astrom et bien d'autres.

D. Mesures de performances

Le but de l'extraction des entités nommées et un peu analogue à ce lui de la classification, il s'agit à partir d'un ensemble de textes de *trouver/rechercher* les informations et uniquement celles qui sont *pertinentes* par rapport à une *catégorie considérée*.

Dans le cadre de la détection des entités nommées, deux objets vont être évalués par entité :

La première des choses, **le repérage de l'entité nommée** à travers des frontières fixées par le système (bornes de l'entité). Deuxièmement **la classe attribuée** à cette entité c.à.d. le type de l'entité.

Dans ce cas, nous cherchons à évaluer la capacité du système à trouver les informations pertinentes et uniquement celles là. Pour évaluer les sorties plusieurs critères de mesure des performances ont été proposé le **rappel**, *la* **précision** *et le* **F-mesure.** *(Van Rijsbergen, 1979)* Celles-ci mesurent le nombre d'objets pertinents effectivement récupérés parmi un ensemble d'objets.

E. Conclusion :

Cet article décrit la tâche de la reconnaissance d'entités nommées qui a une double finalité : elle consiste à « **baliser** » des documents textes pour l'identification et la « **catégorisation** » des mots dans des entités nommées correspondantes. Il vise ainsi à accéder aux informations contenues dans des textes dans la perspective de répondre à des questions basiques : Qui? Quoi? Où? Quand? Comment? Pourquoi?

Les conférences MUC, CoNLL ou ACE ont proposé un ensemble limité de catégories (par exemple, une personne, une organisation, l'emplacement et les expressions numériques comme l'argent ou des expressions pour cent), mais il ya beaucoup plus d'étiquettes possibles proposées illustré dans une hiérarchie étendu d'entités nommées**[12]** qui contient environ 150 types NE développé par Sekine, et elle atteint encore actuellement 200 types.

REFERENCES

[1] R. Grishman 1995 The NYU system for MUC-6 or Where's the Syntax. In the proceedings of Sixth Message Understanding Conference (MUC-6) (pp167-195). Fairfax, Virginia.

[2] S. Sekine, Description of the Japanese NE system used for MET-2, In: MUC-7, Fairfax, Virginia, 1998.

[3] CRF++:http://crfpp.sourceforge.net/ Yet Another CRF toolkit (accessed on 3 rd may 2010)

[4] Li, W., McCallum, A., 2004. Rapid development of Hindi named entity recognition using conditional random fields and feature induction. ACM Tran. on Asian Language Information Processing (TALIP) 2 (3), 290–294.

[5] Takeuchi, K., Collier, N., 2002. Use of support vector machines in extended named entity. In: Proc. CoNLL-2002.

[6] Isozaki, H., Kazawa, H., 2002. Efficient support vector classifiers for named entity recognition. In: Proc. COLING-2002, pp. 390–396.

[7] Nadeau, David and Satoshi Sekine (2007) A survey of named entity recognition and classification. Linguisticae Investigationes 30(1):3–26.

[8] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, Exploiting diverse knowledge sources via maximum entropy in named entity recognition, in: Proceedings of the 6th Workshop on Very Large Corpora, WVLC-98, Montreal, Canada, 1998.

[9] A. Mikheev, C. Grover, M. Moens, Description of the LTG System used for MUC-7, In: MUC-7, Fairfax, Virginia, 1998.

[10] R. Srihari, C. Niu, W. Li, A hybrid approach for named entity and sub-type tagging, In: Proceedings of Sixth Conference on Applied Natural Language Processing (ANLP), 2002, pp. 247–254.

[11] X. YU, Chinese named entity recognition with cascaded hybrid model, In: Proceedings of NAACL HLT 2007, Prague, 2007, pp. 197–200.

[12] S. Sekine, K. Sudo, C. Nobata, Extended named entity hierarchy, in: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002, Las Palmas, Canary Islands, Spain, 2002.

[13] Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E.M., Soui_ H., (2004). Initiation à la langue amazighe, IRCAM, Rabat, Maroc.

[14] Ali Rachidi & Driss Mammass, Informatisation de La Langue Amazighe : Méthodes et Mises En Œuvre. 3rd International Conference: Sciences of Electronic Technologies of Information and Telecommunications, March 27-31, 2005 - TUNISIA