Système de reconnaissance des entités nommées amazighes

Meryem TALHA[1], Siham Boulaknadel[2], Driss Aboutajdine[1]
[1] LRIT – CNRST URAC29, Université Mohammed V-Agdal Rabat
4, Avenue Ibn Battouta, B.P. 1014 RP, 10006 Rabat, Maroc
meriem.talha@gmail.com, aboutaj@fsr.ac.ma
[2] IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
boulaknadel@ircam.ma

Abstract

Named Entity Recognition (NER) for Amazigh language is a potentially useful pretreatment for many processing applications for the Amazigh language. However, this task represents a tough challenge, given the specificities of this language. In this paper, we present (NERAM) the first named entity system for the Amazigh language based on a symbolic approach that uses linguistic rules built manually by using an information extraction tool available within the platform GATE.

Résumé

La reconnaissance des Entités Nommées (REN) en langue amazighe est un prétraitement potentiellement utile pour de nombreuses applications du traitement de la langue amazighe. Cette tâche représente toutefois un sévère challenge, compte tenu des particularités de cette langue. Dans cet article, nous présentons le premier système d'extraction d'entités nommées amazighes (RENAM) fondé sur une approche symbolique qui utilise le principe de transducteur à états finis disponible sous la plateforme GATE.

Mots-clés : reconnaissance des entités nommées (REN), langue amazighe, gazetteers, règles linguistiques, JAPE, GATE.

1. Introduction

Au Maroc, la langue berbère ou amazighe (ⵜⴰⵎⴰⵣⵉⵖⵜ), qui fait partie des langues chamito-sémitiques ou encore appelé afro-asiatiques (Cohen, 2007 ; Chaker, 1989), a toujours disposé d'un statut restreint, bien qu'elle soit considérée comme la plus ancienne langue d'Afrique du Nord ; elle est fortement employée sur une aire très vaste par environ 50% de la population marocaine (Boukouss, 1995). Grâce aux efforts de revitalisation de la langue, la langue amazighe est maintenant une langue qui en possède tous les 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 riche.

L'augmentation des flux d'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 (REN) dans l'étape de prétraitement. La tâche REN est une sous-tâche du domaine de l'extraction d'information consistant à identifier et à catégoriser certaines expressions linguistiques autonomes et mono-référentielles (Ehrmann, 2008).

Certaines langues ont éveillé beaucoup d'intérêt, notamment à travers les campagnes d'évaluations telles que CONLL (Tjong Kim Sang, 2002) pour l'espagnol et l'allemand, MUC (Grishman et Sundheim, 1996) pour l'anglais et le japonais, et ESTER (Galliano et al., 2009) pour le français. Toutefois, l'amazighe étant une langue peu dotée en terme de ressources linguistiques informatisées, les travaux sur la Reconnaissance des Entités Nommées sont une opportunité pour la valorisation de cette langue dans la société de l'information.

C'est dans cette optique que se situe le travail que nous présentons dont le but est de détecter et d'extraire, dans des textes amazighes, les entités nommées pertinentes et ceci en développant le premier système de reconnaissance d'entités nommées (RENAM) qui est une étape d'outillage de l'analyse qui servira à des applications plus spécifiques dans le cadre d'une démarche incrémentale. Notre système est fondé sur une approche symbolique où l'extraction s'effectue en se basant sur un ensemble de *gazetteers* et de règles construites manuellement en exploitant l'outil d'extraction des entités nommées disponible sous la plateforme GATE[1].

Notre article est structuré comme suit : la première section présente un état de l'art des approches d'extraction des entités nommées ; la deuxième examine les difficultés entravant l'extraction des entités nommées en amazighe ; la troisième aborde les particularités de la plateforme choisie ainsi que la méthodologie utilisée, tandis que la description des résultats obtenus est l'objet d'étude de la quatrième section. Finalement, la dernière section est consacrée à la conclusion et aux perspectives.

2. Approches d'extraction d'entités nommées

La REN constitue un champ de recherche très actif depuis de nombreuses années dans plusieurs langues. Des approches fondamentales existent : l'extraction fondée sur des démarches linguistiques ou encore nommées symboliques, des approches statistiques ou à base d'apprentissage, des approches hybrides qui combinent les deux précédentes.

La première approche exploite les avancées en TAL et s'appuie spécialement sur l'utilisation de grammaires formelles construites à la main par un expert-linguiste. Elle se fonde particulièrement sur la description des entités nommées (désormais EN) grâce à des règles qui exploitent des marqueurs lexicaux, des dictionnaires de noms propres et parfois un étiquetage syntaxique.

En ce qui concerne les marqueurs lexicaux qui sont aussi nommés des *mots déclencheurs,* il s'agit de signes ou d'indices qui encadrent, soit à gauche ou à droite, l'entité nommée et qui permettent souvent de dévoiler sa présence. D'autre part, les dictionnaires de noms propres regroupent généralement une liste des noms et des prénoms les plus fréquents, des noms de localisations (villes, pays, fleuves, etc.) et parfois des noms d'organisations (organismes, compagnies, etc.). Ces dictionnaires sont fréquemment utilisés dans les systèmes à base de règles, ils peuvent aussi être utilisés par des systèmes à base d'apprentissage. Les ressources qu'on a mentionnées renforcent l'élaboration de règles et de patrons linguistiques qui spécifient les contextes d'apparition de telles entités. Pour les langues peu dotées, cette approche semble la plus adéquate. Dans le cadre des approches linguistiques, nous rappelons quelques systèmes de REN arabes : les travaux de (Shaalan et Raza, 2008) qui ont développé le système NERA permettant d'extraire dix types d'EN. Ce système repose sur l'utilisation d'un ensemble de dictionnaires d'EN et sur une grammaire sous forme d'expressions régulières pour la reconnaissance des EN. Le meilleur taux F-mesure acquis par ce système est de 98.6%. Suivant le même principe, les travaux de (Zaghouani et al., 2010) ont présenté un module de repérage des EN à base de règles pour la langue arabe, la seule différence est

---

[1] http://gate.ac.uk

qu’ils ont procédé à une première étape de prétraitement lexical qui prépare le texte pour son analyse linguistique, ce module a été évalué sur un corpus de presse. La valeur de F-mesure apportée par ce module est de 47.35% pour le type *organisation* et 95.10% pour le type *date*.

La seconde démarche fait usage de techniques statistiques ou encore dites à base d’apprentissage pour apprendre des spécifiés sur de larges corpus de textes (nommés ainsi corpus d’apprentissage) où les entités-cibles ont été auparavant étiquetées ; elle utilise par la suite un algorithme d’apprentissage qui va permettre d’élaborer automatiquement une base de connaissances à l'aide de plusieurs modèles numériques (CRF, SVM, HMM…). Cette méthode a été envisagée pour avoir une certaine intelligence lors de la prise des décisions, ce sont principalement certains paramètres qui peuvent être manipulés dans le but d’améliorer les résultats du système, ce qui n’est pas le cas pour les approches symboliques qui n’appliquent que les règles préalablement injectées. (Benajiba et al., 2009) ont conçu une technique d’apprentissage SVM pour la mise en œuvre de leur système de REN en se servant d’un ensemble de particularités de la langue arabe. Ce système a produit une F-mesure globale de 82.71%. Pour des langues peu dotées, le Telugu par exemple, (Srikanth et Kavi, 2008) ont utilisé la technique CRF pour extraire les entités nommées et ils ont obtenu un score de f-mesure qui est égal à 92% ; pour le Bengali en utilisant le SVM, (Asif et Sivaji, 2008) ont réussi à avoir un résultat dont la valeur de f-mesure est de 91,8%. Similairement au Telugu, (Vijayakrishna et Sobha, 2008) ont conçu leur propre système pour le Tamil en se servant du CRF et ils ont obtenu un résultat satisfaisant de 80,44% comme valeur de f-mesure.

Il apparaît que les deux approches citées précédemment sont complémentaires, ce qui a conduit à la mise en œuvre d’une troisième approche (une combinaison de deux) qui utilise des règles écrites manuellement mais qui construit aussi une partie de ses règles en se basant sur des informations syntaxiques et sur des informations extraites des données grâce à des algorithmes d'apprentissage, des arbres de décisions. (Abuleil, 2006) a adopté une approche hybride pour faire l’extraction des entités arabes, en tirant profit des approches symboliques et à base d’apprentissage.

3. Difficultés entravant l’extraction des entités nommées amazighes

La langue amazighe fait partie des langues chamito-sémitiques (Cohen 2007 ; Chaker 1989). Elle est composée de 27 consonnes, 2 semi-consonnes, 3 voyelles pleines et une voyelle neutre. Elle présente une morphologie riche et complexe. Les mots peuvent être classés en trois catégories morphosyntaxiques: Nom, Verbe et Particules (Boukhris et al., 2008).

Cependant, les travaux liés à la REN pour la langue amazighe sont encore limités pour les raisons suivantes :

- 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 rencontrent pas, ni au début ni à l’initiale des noms propres amazighes.
- 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 étiqueteurs, les analyseurs morphologiques.

- La langue amazighe 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.

Similairement à d'autres langues naturelles, l'amazighe présente des incertitudes au niveau des classes grammaticales. En effet, une même forme peut relever de nombreuses catégories grammaticales, suivant le contexte de la phrase. Par exemple, (ⵉⵍⵍⵉ) peut être considéré comme un verbe à l'accompli positif (il signifie «il existe») ou comme un nom de parenté («ma fille»).

Le nombre de mots fréquemment utilisés comme noms communs et qui peuvent être également utilisés comme noms propres est très grand. Comme par exemple Ait Larbi (ⴰⵢⵜ ⵍⵄⴰⵔⴱⵉ) qui désigne à la fois un nom de personne et un nom de société.

Les noms propres dans la langue amazighe sont extrêmement nombreux, ils ont de nombreuses variantes et ils sont difficiles à détecter sans la présence d'un lexique ; 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.

Les particularités que nous avons décrites dans cette partie vont nous guider lors de l'établissement d'un système de repérage des entités nommées.

4. Système de reconnaissance d'entités nommées amazighes (RENAM)

La tâche de reconnaissance des entités nommées amazighes apparaît en effet comme fondamentale pour diverses applications participant à l'analyse du contenu des textes amazighes. Dans cette contribution, nous nous intéressons à développer un système d'analyse de textes amazighes permettant le repérage et la catégorisation des entités nommées en fonction de types sémantiques prédéfinis à savoir le type *personne*, *organisation*, et *localisation* en exploitant la plate-forme GATE.

4.1. Plate-forme GATE

La plateforme d'ingénierie textuelle GATE (*General Architecture for Text Engineering*) (Cunningham et al., 2002) est une infrastructure de développement de traitement du langage humain développée par l'Université de Sheffield et est exploitée dans une vaste variété de recherche et de projets de développement incluant l'extraction de connaissances pour l'anglais, l'espagnol, le chinois, l'arabe, le français, l'allemand, l'hindi, le cebuano, le roumain, le russe.

Nous avons choisi cet environnement car il permet de fournir un *framework* permettant d'implémenter une architecture de développement ; il peut être utilisé pour l'exploitation des traitements linguistiques dans diverses applications. En outre, il dispose d'une grande communauté d'utilisateurs : on dispose ainsi d'un ensemble de solutions d'aide et de support (forum, liste de diffusion, tutoriels, etc.), point indispensable lorsque l'on débute avec un tel outil.

La plateforme repose sur le principe d'une chaine de traitement composée de plusieurs modules dédiés à l'analyse textuelle appliqués successivement sur un ou plusieurs textes. Le module d'extraction proposé dans GATE est réalisé selon une approche symbolique basée sur le formalisme JAPE (*Java Annotation Patterns Engine*) qui est un transducteur à états finis permettant de définir les contextes d'apparition des unités à extraire dans le but de les repérer et les annoter. Le principe est de réunir différentes annotations «basiques» (*tokens*, syntagmes, etc.) pour en constituer de nouvelles plus complexes (entités nommées, relations, etc.).

4.2. Architecture du système

Notre système d'extraction d'entités nommées est un système de détection et de typage des entités d'intérêt dans un texte. Notre but se limite à l'extraction des entités de type *personne*, *organisation* et *localisation*, dans des textes écrits en amazighe. Il comporte deux phases (figure 1) : la préparation des données et la reconnaissance des entités nommées.

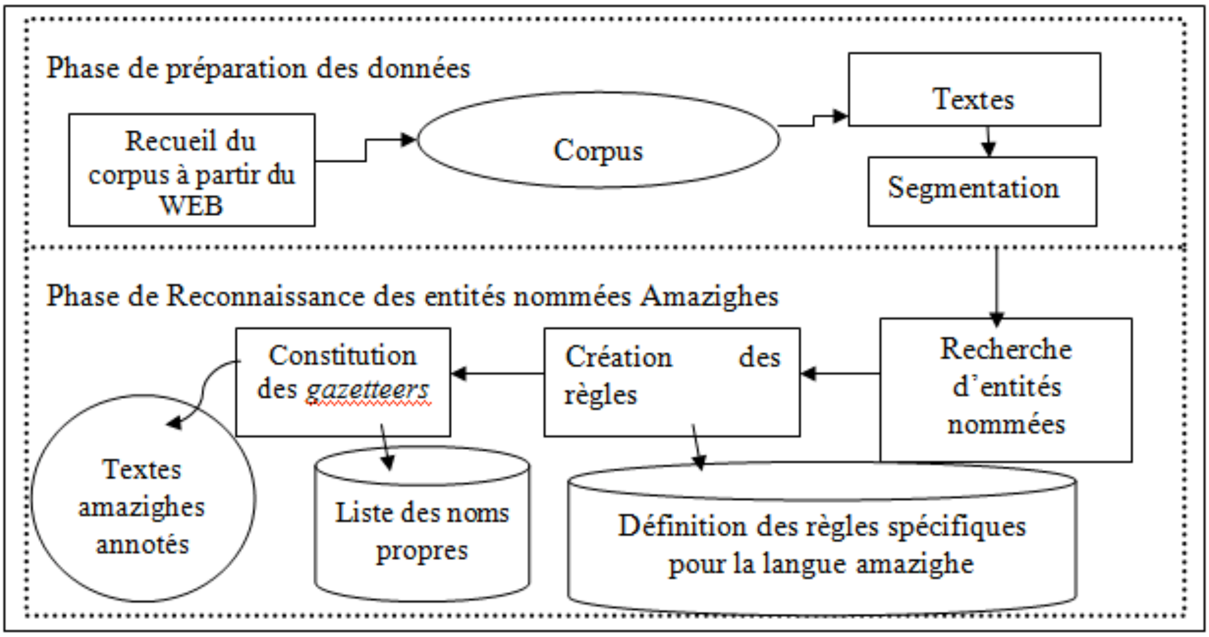

Figure 1. Architecture du système d'extraction des entités nommées amazighes

4.2.1 Phase de préparation du corpus

Recueil du corpus

La construction d'un système d'extraction d'entités nommées exige tout d'abord de rassembler un nombre suffisant de textes qui serviront non seulement de corpus d'observation (pour constituer les règles) mais également de corpus de test. Le contexte de cette contribution nous a naturellement guidés vers la collecte de textes amazighes journalistiques à partir du site web « mapamazighe[2] », le portail d'informations amazighes de l'Agence Maghreb Arabe Presse (*MAP*). Le corpus contient toute l'actualité sur les activités royales de SM le Roi Mohammed VI ; on dispose de 200 articles au format HTML écrits en amazighe, cumulant un total de 6 512 *tokens*, nous signalons aussi qu'outre les unités écrites en amazighe ce corpus inclut 482 771 séquences de chiffres.

Segmentation du corpus

Dans cette phase, nous segmentons le texte amazighe en phrases puis en mots. Les phrases sont fragmentées sur la base de signes de ponctuation (y compris le retour à la ligne), comme

---

[2] http://www.mapamazighe.ma/am/

pour le français ou l'anglais. La phrase est ensuite segmentée en mots en s'appuyant sur des signes comme : espace, virgule… Nous présentons à titre d'exemple le texte suivant :

> ⴱⴰⵎⴰⴽⵓ – ⵉⵍⴽⵎ ⴱⴰⴱ ⵏ ⵡⴰⴷⴷⵓⵔ ⴰⴳⵍⵉⵉⴷ ⵎⵓⵃⵎⵎⴷ ⵡⵉⵙⵙ ⵚⴹⵉⵚ, ⴰⴷ ⵜ ⵉⵏⵚⵕ ⵕⴱⴱⵉ, ⴳ ⵜⵣⴷⵡⵉⵜ ⵏ ⵓⵙⵉⵏⴰⵙ ⴳ ⴱⴰⵎⴰⴽⵓ, ⴰⵎⵎⴰⵙ ⵏ ⵢⴰⵏ ⵓⵔⵣⴰⴼ ⵉ ⵜⴳⴷⵓⴷⴰⵏⵜ ⵏ ⵎⴰⵍⵉ, ⵎⴰⵢⴰⴷ ⵉⴳⴰⵏ ⵜⵉⴽⵍⵉ ⵜⴰⵎⵣⵡⴰⵔⵓⵜ ⵏ ⵢⴰⵜ ⵜⵙⵙⵓⵜⵍⵜ ⵜⴰⴼⵔⵉⴽⵉⵜ ⵍⵍⵉ ⵔⴰⴷ ⵢⴰⵡⵉ ⴱⴰⴱ ⵏ ⵡⴰⴷⴷⵓⵔ ⴰⴳⵍⵍⵉⴷ ⵙ ⵍⴽⵓⵜ ⴷⵉⴼⵡⴰⵔ ⴷ ⵖⵉⵏⵢⴰ ⵜⴰⴽⵓⵏⴰⴽⵔⵉⵜⴷ ⵍⴳⴰⴱⵓⵏ.
>
> Bamaku – ilkm bab n waddur agllid muhmmd wiss sdis, ad t insr rbbi, g tzdwit n usinas g bamaku, ammas n yan urzaf i tgdudant n mali, mayad igan tikli tamzwarut n yat tssutlt tafrikit lli rad yawi bab n waddur agllid s lkut difwar d ghinya takunakritd lgabun.
>
> Bamako - Sa Majesté le Roi Mohammed VI, que Dieu L'assiste, est arrivé, mardi en fin d'après-midi à Bamako, pour une visite officielle en République du Mali, première étape d'une tournée africaine qui conduira le Souverain en Côte d'Ivoire, en Guinée Conakry et au Gabon.

4.3 Phase d'extraction des entités nommées

Constitution des listes des noms propres (gazetteers)

Dans le but de pouvoir reconnaître automatiquement les entités nommées de type *personne*, *organisation* et *localisation*, nous avons commencé par l'élaboration d'un lexique ou *gazetteers* à partir du Web.

Pour les entités nommées de type *personne*, nous avons construit une liste d'environ 1 380 entrées de noms amazighes et noms étrangers transcrits en amazighe.

Pour les entités nommées de type *localisation*, nous nous sommes inspirés de la classification faite par (Piton et Maurel, 2004) qui considère comme type *localisation* ou toponyme : les pays, villes, fleuves, montagnes, océans et mers. Ainsi, nous avons élaboré un lexique qui contient 1 830 entités de type *localisation*.

A l'instar de la procédure d'élaboration de lexiques des deux entités nommées précédentes, l'identification des entités nommées de type *organisation* a commencé par le développement d'un lexique qui contient 64 de noms d'organisation.

Développement des règles linguistiques

Pour la construction de nos règles linguistiques pour le traitement de l'amazighe, nous nous sommes inspirés des travaux de (McDonald, 1996) dans le domaine qui consiste à déterminer les indices d'apparition des entités à extraire. Ainsi, ces indices font partie intégrante de l'entité nommée comme le prénom pour une entité de type *personne* ou correspondent au contexte textuel de l'entité pouvant indiquer son type (dits «mots déclencheurs »). La découverte de ces indices s'est faite par observation du corpus.

Règles de type « personne » :

Les règles écrites pour la détermination de l'entité *personne* prend appui sur la présence :

- Des civilités, à titre d'exemple : (ⵎⴰⵙⵙ (Mass, Monsieur), ⵎⴰⵙⵜ (Mast, Mme), ⵎⴰⵙⵙⴰ (Massa, Mme), ⵍⴰⵍⵍⴰ (Lalla, Mme), ⵎⵓⵍⴰⵢ(Moulay) …
- Les titres de toute sortes tels que les titres politiques ⴰⵎⴰⵡⴰⵙ (**amawas**, ministre), les titres honorifiques ⴱⴰⴱ ⵏ ⵡⴰⴷⴷⵓⵔ (bab n waddur, Sa Majesté), etc

De ce fait, les noms qui viennent directement après les titres de personnes, vont être annotés comme entités *personne*.

<TitrePersonne> <Nom> ➔ Nom= « entité personne ».

Règles de type « localisation » :

Les règles linguistiques pour la reconnaissance des entités de type *localisation* sont écrites en se basant sur des préfixes de localisation et des prépositions qui accompagnent toujours les noms de lieux comme par exemple (‘ⴷⴰⵔ (dar, en direction de)’, ‘ⵖⵔ (ghr, vers)’, ‘ⵊⴰⵔ…ⴷ…’ (jar…d, entre… et…)…).

En outre, nous avons construit une liste de mots déclencheurs comme par exemple, « ⵜⵎⴷⵉⵏⵜ » (tmdint,ville).

Règles de type « organisation » :

Nous avons élaboré une liste de mots déclencheurs comme par exemple ⵜⴰⵎⴰⵡⴰⵙⵜ **(tamawast,** Ministère) pour la description des règles de reconnaissance des entités de type *organisation*.

Titre	Exemple	Nombre de règles construites
Personne	ⵎⴰⵙⵙ (mass, Monsieur), ⵎⴰⵙⵜ (mast, Madame), ⴱⴰⴱ ⵏ ⵡⴰⴷⴷⵓⵔ (bab n waddur, Sa majesté)…	9
Organisation	ⵜⵉⵏⵎⵍ (tinml, Ecole), ⵜⴰⵎⴰⵡⴰⵙⵜ (tamawast, Ministère) …	14
Lieux	ⴰⵙⵉⴼ (asif, Fleuve), ⵜⵎⴷⵉⵏⵜ(tmdint, Ville) …	22

Table1. Tableau récapitulatif

5. Evaluation des résultats

Une dernière étape de cette contribution est l'évaluation des résultats de notre outil d’extraction des entités nommées amazighes. Ainsi, nous présentons, dans cette section, les différentes étapes de notre évaluation et les conclusions que nous avons pu faire au vu des résultats obtenus.

5.1. Protocole d’évaluation

L'évaluation de notre système a permis d’établir plusieurs choix visant le type d'évaluation à mettre en place. Deux solutions se sont alors présentées : exploiter les données d'une campagne d'évaluation existante ou créer notre propre système d'évaluation.

Le premier cas n’étant pas possible puisqu’il n’existe pas un corpus de la langue amazighe, où les entités nommées ont été préalablement annotées. En conséquence, nous avons envisagé la deuxième solution, qui est le développement de notre système d'évaluation.

De ce fait, nous avons développé un corpus journalistique de la langue amazighe. Ce corpus a, par la suite, été annoté grâce à notre outil de repérage d'entités nommées.

Nous avons réalisé deux évaluations sur le corpus en mesurant tout d’abord la reconnaissance des entités nommées en exploitant seulement les *gazetteers*, puis en combinant les règles linguistiques avec les *gazetteers*.

5.2. Analyse des résultats

Le but de l'extraction des entités nommées est de trouver/rechercher les informations et uniquement celles qui sont pertinentes par rapport à une catégorie considérée à partir d'un ensemble de textes.

Dans une deuxième étape de notre évaluation, nous avons déterminé les métriques qui nous permettent d'évaluer les résultats de nos travaux et ainsi mesurer les performances du système proposé. La précision (mesure de qualité), le rappel (mesure de quantité) et la F-mesure (synthèse de Rappel et de précision) (Rijsbergen, 1979), ont été choisis pour leur exploitation fréquente dans le domaine du TAL.

Les résultats de notre reconnaissance d'entités nommées en amazighe sont présentés ci-dessous (cf. tableaux 2 et 3). Les trois métriques précédentes sont calculées pour chaque type d'entité (LOC, ORG, PERS) mais également pour l'ensemble des entités nommées (EN).

Nous avons effectué une évaluation sur notre corpus « Activités Royales » qui donne les résultats suivants :

	Précision	Rappel	F-mesure
PERS	64%	63%	64%
LOC	27%	71%	40%
ORG	82%	81%	82%

Table 2. Evaluation de notre système de reconnaissance des entités nommées

	Précision	Rappel	F-mesure
PERS	Iban Roumaine Amazighe	50% 67% 64%	Iban Roumaine Amazighe
LOC	Iban Roumaine Amazighe	50% 86% 40%	Iban Roumaine Amazighe
ORG	Iban Roumaine Amazighe	60% 85% 82%	Iban Roumaine Amazighe

Table 3. Comparaison des résultats de notre système avec d'autres langues peu dotées

D'après la table 2, les résultats que nous avons obtenus sont plus ou moins satisfaisants et encourageants. Cependant notre outil fonctionne légèrement moins bien pour l'entité nommée de type «organisation». Ceci est dû principalement à la procédure de reconnaissance des entités nommées :

La prise en compte des variantes orthographiques des noms propres transcrits en l'absence de conventions pour leurs écritures (notamment pour les noms de lieux). En amazighe, la translittération et la transcription des noms propres étrangers n'obéissent pas à des règles d'écritures par exemple (ⴰⴱⵓ ⴷⴰⴱⵉ (abu dabi, Abu Dhabi) ou ⴰⴱⵓⴷⴰⴱⵉ (abudabi, Abudhabi)).

Le problème de délimitation de l'entité nommée est dû à l'absence d'informations morphosyntaxiques qui pourraient aider à l'amélioration de leur détection : le mot déclencheur ⵜⴰⵎⴰⵡⴰⵙⵜ (tamawast, Ministère) indique que la séquence qui va suivre représente une entité nommée de type *organisation*, néanmoins l'absence d'informations morphologiques nécessaires rend la tâche de délimitation de l'entité en question plus difficile.

Si on considère la séquence : ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵓⵙⴳⵎⵉ ⴰⵏⴰⵎⵓⵔ (tamawast n usgmi anamur, Ministère de l'éducation nationale) ; notre outil ne va reconnaitre que l'entité « ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵓⵙⴳⵎⵉ », parce qu'il n y a pas de critère d'arrêt selon chaque entité. Nous suggérons d'effectuer un traitement syntaxique additionnel dans l'espoir d'effectuer une désambigüisation morphosyntaxique avant de passer à la reconnaissance des entités nommées.

Nous avons aussi comparé notre système avec ceux développés par (Oana et al., 2003) et (Fong et al., 2011). Les résultats de la comparaison sont synthétisés dans le tableau 3. Bien que les trois systèmes travaillent sur des corpus différents pour leurs évaluations, ils utilisent le même outil GATE ; nous pouvons admettre que les résultats obtenus par notre système sont, dans certains cas, assez encourageants.

6. Conclusion

L'objectif principal de cette contribution est de reconnaître les entités nommées de types (personne, localisations, organisations) dans les textes amazighes. Pour ce faire, nous avons élaboré dans un premier temps, un système de reconnaissance d'entités nommées pour l'amazighe, fondé sur une approche symbolique (utilisant des règles linguistiques construites manuellement et sur l'élaboration des *gazetteers*) en utilisant la plateforme GATE. Ceci nous a permis d'obtenir un taux de reconnaissance assez encourageant malgré les problèmes décrits dans ce papier. Dans un futur immédiat, nous tentons d'ajouter d'autres règles d'extraction d'entités nommées et d'enrichir la structure de nos *gazetteers* pour aboutir à un taux de reconnaissance élevé.

Références

Asif E. et Sivaji B. (2008). Bengali Named Entity Recognition Using Support Vector Machine. *International Joint Conference on Natural Language Processing(IJCNLP):* 51-58.

Benajiba Y. et Rosso P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. *In Proceedings of Workshop on HLT and NLP within the Arabic World*, *LREC*.

Bickel M., Miller S., Schwartz R. et Weischedel R. (1997). "Nymble: a high-performance learning name-finder". *In Proceedings of the ANLP 97.*

Borthwick A., Sterling J., Agichtein E. et Grishman R. (1998). "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition". *In Proceedings of the WVLC 98.*

Boukhris F., Boumalk A., Elmoujahid E. et Souifi H. (2008). La nouvelle grammaire de l'amazighe. Rabat, Maroc: IRCAM.

Chaker S. (1984). Textes en linguistique berbère - introduction au domaine berbère, *éditions du CNRS*, P 232-242.

Cohen D. (2007). Chamito-sémitiques (langues). *In Encyclopædia Universalis*.

Cunningham H., Maynard D., Bontcheva K., Tablan V. et Ursu C. (2002), The GATE User Guide.

Cunningham H. (1999), Information Extraction: a User Guide (revised version), *Research Memorandum*, Department of Computer Science, University of Sheffield.

Ehrmann M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. *PhD thesis*, Université Paris 7.

Fong Y., Bali R. et Wee A. (2011). NERSIL - the Named-Entity Recognition System for Iban Language. *PACLIC 2011:* 549-558

Gahbiche B S., Bonneau H., Maynard D., Lavergne T. et Yvon F. (2012), Repérage des entités nommées pour l'arabe, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 487–494.*

Grishman R. (1995). The NYU system for MUC-6 or Where's the Syntax. *In the proceedings of Sixth Message Understanding Conference (MUC-6)* (167-195). Fairfax, Virginia

Li W. et McCallum A. (2004). A Note on Semi-supervised Learning using Markov Random Fields. *Technical Note*.

Maynard D. (2011). Developing Language Processing Components with GATE, Version 6, *http://gate.ac.uk/sale/tao/tao.pdf.*

McDonald D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In B. Boguraev and J. Pustejovsky, editors, Corpus Processing for Lexical Acquisition,* 21–39.

Oana H., Bontcheva K., Maynard D., Tablan V. et Cunningham H. (2003). Named entity Recognition in Romanian using Gate. *RANLP 2003 Workshop on information Extraction for slavonic and other central and eastern European languages*, Borovets : Bulgaria.

Rijsbergen V., C.J. (1979). Information Retrieval, *2nd edition*, London: Butterworths.

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

Shaalan K. et Raza H. (2009). NERA : Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(9):1652–1663.

Srikanth P. et Kavi N. M. (2008). Named Entity Recognition for Telugu. *In the proceedings of the IJCNLP-2008 Workshop on NERSSEAL (Named Entity Recognition in South and South East Asian Languages),* Hyderabad, India.

Takeuchi K. et Collier N. (2002). Use of support vector machines in extended named entity. *In*: *Proc. CoNLL-2002.*

Tjong K. S. et Erik. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *In Proc. Conference on Natural Language Learning*. Taipei, Taiwan

Tjong K. S. et DeMeulder F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, *in: Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada, 142–147.

Vijayakrishna R. et Sobha L. (2008). Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields. *In Proceedings of International Joint Conference on Natural Language Processing Workshop on NER for South and South East Asian Languages*, pp. 59 - 66.

Wakao T., Gaizauskas R. et Wilks Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. *In Proceedings of COLING-96*.

Zaghouani W., Pouliquen B., Ebrahim M. et Steinberger R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. *Proceedings of the Seventh conference on International Language Resources and Evaluation* (*LREC'10*), 563–567.