# L'apport d'une approche symbolique pour le repérage des entités nommées en langue amazighe

Meryem Talha*, Siham Boulaknadel**, Driss Aboutajdine*

*LRIT, Unité Associée au CNRST
Faculté Des Sciences, Université Mohammed V - Agdal Rabat, Maroc
meriem.talha@gmail.com
aboutaj@fsr.ac.ma
**Institut Royal de la Culture Amazighe
Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
boulaknadel@ircam.ma

**Résumé.** Le repérage des Entités Nommées (REN) en langue amazighe est un prétraitement éventuellement essentiel pour de nombreuses applications du traitement automatique des langues (TAL), en particulier pour la traduction automatique. Dans cet article, nous présentons une chaîne de repérage des entités nommées en amazighe fondée sur une étude synthétique des spécificités de la langue et des entités nommées en amazighe. L'article met l'accent sur les choix méthodologiques à résoudre les ambiguïtés dues à la langue, en exploitant les technologies existantes pour d'autres langues.

## 1 Introduction

Depuis les débuts du TALN, la compréhension de texte fait l'objet d'un suivi particulier de plusieurs recherches. C'est en faveur du développement rapide de la tâche d'extraction d'information que la tâche de REN s'est manifestée. Elle consiste à rechercher les expressions référentielles (Ehrmann (2008)), qui recouvrent classiquement les noms désignant des personnes, des lieux, des organisations, des expressions temporelles et celles numériques, mais peuvent aussi se rapporter à des notions plus techniques comme les maladies. Dès la campagne MUC-6, la tâche de REN s'est ainsi polarisée sur trois types d'entités (Grishman et Sundheim (1996)), à savoir : ENAMEX (personnes, organisations et lieux), TIMEX (expressions temporelles), NUMEX (expressions numériques). Cette première définition a été étendue dans la campagne CoNLL (Tjong Kim Sang et De Meulder, 2003) où 4 classes ont été normalisées : personnes, organisations, lieux, Divers. Dans les campagnes d'évaluation ESTER2 (Galliano et al., 2009) 8 catégories ont été normalisées à savoir personnes, fonctions,organisations, lieux, productions humaines, dates, montants et événements.

## 2 Travaux Connexes

Auparavant, la visibilité de la langue amazighe au Maroc était quasiment nulle. Récemment, et grâce aux revendications qui se sont faites à l'aide de l'IRCAM [1], elle a été soumise à un processus de codification et de standardisation. Face à l'augmentation vertigineuse des informations en langue amazighe, disponibles librement sur le Web, plusieurs recherches ont été entamées dans ce sens. Il y en a celles qui se concentrent sur la reconnaissance optique des caractères (OCR) (Es-Saady et al., 2012) et celles qui se focalisent sur le TALN que nous pouvons classer en deux grandes catégories : (1) ressources informatiques, y compris des études sur la construction des corpus amazighe (Boulaknadel et Ataa Allah, 2011) et (2) les outils du TAL qui ont été réalisés comme le concordancier (Boulaknadel et Ataa Allah, 2010), l'analyseur morphologique (Nejme et al., 2013a,b) et (Ataa Allah et Boulaknadel, 2010). Quant au domaine de la REN, il a acquis un certain intérêt à travers les travaux réalisés de Talha et Boulaknadel (Talha et al., 2014b,a; Boulaknadel et al., 2014).

## 3 Aperçu général de notre approche

On distingue traditionnellement trois grandes approches : Approches symboliques qui reposent sur l'utilisation de grammaire formelle construite par la main. Approches statistiques qui permettent d'apprendre, des modèles d'analyse de textes sur de large corpus annoté auparavant, et ensuite établir automatiquement une base de connaissances à l'aide de plusieurs modèles numériques comme le CRF, SVM, etc. Au-delà de ces deux approches, il existe une autre qualifiée d'hybride qui représente un arrangement entre ses antécédents. Dans notre contribution, nous proposons un système fondé sur une approche symbolique, vu la non disponibilité d'un large corpus, où le repérage s'effectue en se basant sur un ensemble de gazetteers et de règles qu'on a construit manuellement tout en exploitant le principe de transducteurs à états finis disponibles sous GATE.

## 4 Architecture Logicielle

### 4.1 Plateforme GATE

La plateforme GATE [2](Cunningham et al., 2002) est une architecture modulaire qui inclut un système d'extraction d'information appelé ANNIE contenant un module de repérage d'entités nommées qui est réalisé selon une approche symbolique, basé sur le formalisme JAPE [3].

### 4.2 Architecture du système

Notre système de repérage d'entités nommées permet l'identification des bornes des EN, ainsi que leur catégorisation dans des classes prédéfinies. Son architecture, détaillée sur la figure 1, comporte 3 modules qui effectuent un traitement séquentiel immédiat des données :

1. http ://www.ircam.ma/
2. http ://gate.ac.uk
3. Transducteur à états finis permettant de définir les contextes d'apparition des unités à repérer

un pré-traitement morphologique du texte, le repérage des entités nommées en se basant sur la consultation directe des gazetteers et le repérage des entités nommées à partir des règles.

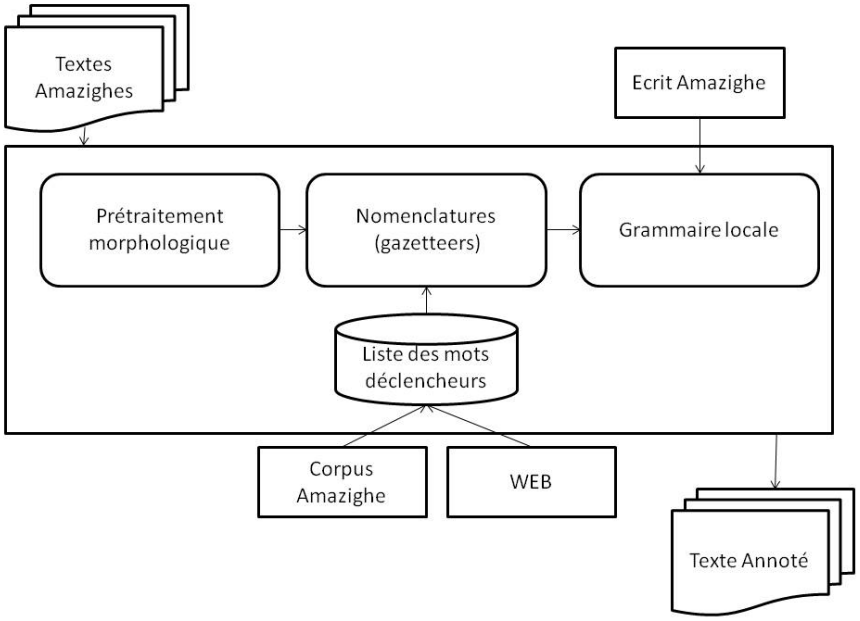


FIG. 1 – *Architecture générale du Système de Repérage des Entités Nommées Amazighes.*

### 4.2.1 Prétraitement Morphologique

Manipuler des textes écrits en langue amazighe nécessite une analyse préliminaire qui consiste à : La suppression des espaces supplémentaires existants entre les mots et l'élimination de tous les mots non-amazighes figurant dans le corpus. Notre analyse comprend deux phases :

- La segmentation du texte amazighe en des phrases.
- L'identification des entités linguistiques de base « tokenisation ».

Ces deux phases citées au dessus sont implémentées en utilisant, respectivement, les modules de GATE : le « Sentence Splitter » et le « Tokeniser ».

### 4.2.2 Constitution des gazetteers

- Noms de personnes : Nous avons élaboré une liste de 2200 entrées des noms amazighes et des noms étrangers transcrits en amazighe, acquise à partir de notre corpus et du web.
- Noms de lieux : contient 2222 entités de type «Localisation». Nous nous sommes inspirés de la classification faite par Piton et Maurel (Piton et Maurel, 2004) qui considère comme toponyme : pays, villes, fleuves, montagnes, océans.

- Noms d'organisations : Nous avons créé une liste qui comprend 449 entrées de types « Organisation », qui regroupe les noms des associations, universités, ministères, etc.
- Expressions numériques : contient 152 entrées de numéros transcrits en amazighe.
- Expressions temporelles : nous avons inséré en total 170 entrées.

#### 4.2.3 Grammaires locales de nom propre

Nous avons développé manuellement de nouvelles grammaires locales, et nous nous sommes inspirés des travaux de McDonald (McDonald, 1996) dans le domaine qui consiste à déterminer les indices d'apparition des entités à repérer.

L'entité nommée peut être parfaitement annotée quand elle contienne un mot déclencheur ou une entrée de nos gazetteers. Pour évaluer notre système nous avons élaboré un certain nombre de règles linguistiques (tableau 1) :

| Entités | PERS | LOC | ORG | DATE | NUM |
|---|---|---|---|---|---|
| Nombre de règles linguistiques | 20 | 12 | 13 | 23 | 8 |

TAB. 1 – *Nombre de règles linguistiques contruites pour la REN amazighe.*

# 5 Validation expérimentale sur les entités nommées

## 5.1 Données expérimentales

### 5.1.1 Corpus

Nous avons fait une conversion des textes vers un format textuel brut en gardant leur structure physique. Pour les différentes expérimentations, nous avons construit notre corpus, contenant l'ensemble de 867 articles extraits à partir de « mapamazighe [4] ». Le corpus traite divers thèmes de toute l'actualité sur les activités royales de SM le Roi Mohammed VI (395), Activités princières (93), Régionales (31), Économiques (58), Sociales (60), Politiques (61), Sport (61), Activités mondiales (52) et les nouveautés générales (56), il cumule un total de 173 480 tokens, 19 102 chiffres. Le recours à une grande diversité de thèmes a pour finalité d'avoir une large couverture des mots. Pour notre expérimentation, nous avons sélectionné 430 articles parmi celles dont nous disposons.

## 5.2 Évaluation

A l'image des campagnes d'évaluation sur lesquelles nous reportons nos résultats, nos métriques d'évaluation seront la Précision, le Rappel, La F-mesure. Le tableau 2 montre les résultats de notre système de REN amazighes :

4. http ://www.mapamazighe.ma/am/

| Entités nommées | PERS | LOC | ORG | DATE | NUM |
|---|---|---|---|---|---|
| Précision | 74% | 76% | 75.5% | 76.5% | 75.5% |
| Rappel | 95% | 100% | 100% | 92% | 99% |
| F-Mesure | 81.5% | 87.75% | 84% | 80% | 83.5% |

TAB. 2 – *Performance de notre système de repérage d'entités nommées amazighes.*

# 6 Discussion

Les entités nommées qui n'ont pas été identifiées, correspondent soit à des entités qui ne font pas partie de nos ressources, soit à des entités qui font partie de nos ressources, mais sont ambiguës. Certaines entités sont cependant ambiguës pour cause d'homographie, ou encore le cas d'entités poly-référentielles, une même entité nommée peut convenir à plusieurs classes. La prépondérance des entités mal classées implique un manque d'information que ce soit au niveau du contexte syntaxique ou de la présence des indices externes, qui, en plus de détermination des mots d'arrêt qui permettent de décider ou s'arrêter, augmente les probabilités d'erreurs de délimitation. Une analyse approfondie a conduit aux constats suivants :

- Enrichir nos gazetteers (Personne, Organisation, Localisation, DATE, NUM).
- Effectuer un traitement syntaxique supplémentaire afin de mieux saisir la structure syntaxique des phrases amazighes avant d'effectuer le repérage des entités nommées.
- Étendre le nombre de règles linguistiques pour chaque classe d'entité nommée.

# 7 Conclusions et perspectives

Dans cet article nous avons proposé un système de repérage des entités nommées amazighes à base de règle. L'évaluation du système montre que les résultats obtenus sont assez encourageants et nous invitent à explorer de nouveaux modes de repérage d'entités nommées, afin de tirer le meilleur parti de notre approche et affiner le repérage des entités nommées amazighes.

# Références

Ataa Allah, F. et S. Boulaknadel (2010). Light morphology processing for amazighe language. In *proceeding of the Workshop on Language Resources and Human Language Technology for Semitic Languages*, Volume 17.

Boulaknadel, S. et F. Ataa Allah (2010). Online amazigh concordancer. In *I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on*, pp. 1–4. IEEE.

Boulaknadel, S. et F. Ataa Allah (2011). Building a standard amazighe corpus. In *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic*, pp. 91–98. Springer Berlin Heidelberg.

Boulaknadel, S., M. Talha, et D. Aboutajdine (2014). Amazighe named entity recognition using a rule based approach. In *International Conference on Computer Systems and Applications (AICCSA)*, Doha, Qatar.

Cunningham, H., D. Maynard, K. Bontcheva, et V. Tablan (2002). A framework and graphical development environment for robust nlp tools and applications. In *Association for Computational Linguistics*, pp. 168–175.

Ehrmann, M. (2008). *Les entités nommées, de la linguistique au TAL*. Ph. D. thesis, Paris 7.

Es-Saady, Y., M. Amrouch, A. Rachidi, M. El Yassa, et D. Mammass (2012). Réalisation d un ocr pour l écriture amazighe imprimée. In *Conférence Internationale sur les Technologies d'Information et de Communication pour l'AMazighe*.

Galliano, S., G. Gravier, et L. Chaubard (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, Volume 9, pp. 2583–2586.

Grishman, R. et B. Sundheim (1996). Design of the muc-6 evaluation. In *Proceedings of a workshop on held at Vienna, Virginia : May 6-8, 1996*, pp. 413–422. Association for Computational Linguistics.

McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. pp. 21–39. Cambridge, MA : MIT Press.

Nejme, F. Z., S. Boulaknadel, et D. Aboutajdine (2013a). Analyse automatique de la morphologie nominale amazighe. pp. 5.

Nejme, F. Z., S. Boulaknadel, et D. Aboutajdine (2013b). Finite state morphology for amazigh language. In *14th International Conference, CICLing Conference on Intelligent Text Processing and Computational Linguistics*.

Piton, O. et D. Maurel (2004). Les noms propres géographiques et le dictionnaire prolintex, les lieux situés hors de france. Number 1, pp. 53. Presses Univ. Franche-Comté.

Talha, M., S. Boulaknadel, et D. Aboutajdine (2014a). "neram : Named entity recognition for amazighe language". In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, Marseille, France, pp. 517–524. Association pour le Traitement Automatique des Langues.

Talha, M., S. Boulaknadel, et D. Aboutajdine (2014b). Système de reconnaissance des entités nommées amazighes). In *Journées internationales d Analyse statistique des Données Textuelles (JADT ) 2014*, Paris, France, pp. 629–638.

Tjong Kim Sang, E. F. et F. De Meulder (2003). Introduction to the conll-2003 shared task. In *Proceedings of the seventh CONLL at HLT-NAACL 2003-Volume 4*, pp. 142–147. Association for Computational Linguistics.

## Summary

Amazighe Named Entity Recognition (NER) system is a potentially vital pretreatment for so many Natural Language preprocessing (NLP) applications, and more specifically: Machine Translation (MT). In this paper, we present an Amazighe named entity recognition system based on a synthetic study of the language-specificity and the amazighe named entities. The article focuses on the methodological choices to resolve ambiguities of this language, by exploiting existing ressources in a related languages.