Etiquetage grammatical de l'amazighe marocain en utilisant les techniques d'apprentissage supervisé

Par **Mohamed OUTAHAJALA, Lahbib ZENKOUAR**

Université Mohamed V, Rabat, Maroc,

Introduction

Avec l'émergence de la revendication identitaire due aux respects des droits de l'homme, les locuteurs natifs qui militent pour la sauvegarde et la promotion de leur langue et leur culture, aspirent à des actions décisives et concrètes pour l'émancipation de leur langue.

C'est ainsi qu'au Maroc, l'amazighe a été introduit dans les médias et dans le système éducatif. L'Alphabet Tifinaghe a été reconnu officiellement par le consortium Unicode le 05/07/2004 (Zenkouar, 2004). Une nouvelle chaîne de télévision amazighe a été lancée le premier mars 2010. La langue amazighe est enseignée dans diverses écoles marocaines : un peu plus de 3.000 écoles et plus de 600.000 élèves suivent cet enseignement dans les écoles du primaire. Au niveau de l'enseignement supérieur, des filières études amazighes et des masters ont été créés. Le 01 juillet 2011, les Marocains ont voté favorablement la nouvelle constitution du pays qui octroie le statut de langue officielle à la langue amazighe. Néanmoins, très peu de ressources ont été développées pour l'amazighe et nous croyons que la création d'un corpus annoté et le développement d'un outil d'étiquetage grammatical est une étape préalable pour le traitement automatique de textes de cette langue.

Dans la deuxième section nous présentons les caractéristiques du corpus annoté. Dans la section trois, nous donnerons un aperçu des méthodes d'apprentissage supervisé. Dans la dernière section nous donnerons quelques conclusions et nous présenterons les travaux en cours et futurs.

1. Caractéristiques du corpus annoté

La langue amazighe présente des défis intéressants pour les chercheurs en Traitement Automatique des Langues(TAL). Certaines de ses caractéristiques sont les suivantes:

1 L'Amazighe dispose de sa propre graphie: le Tifinaghe, qui s'écrit de gauche à

droite (Zenkouar, 2008);

2 Les lettres majuscules, néanmoins, ne se produisent pas, ni au début ni à l'initiale des noms propres ;

3 Les noms, les noms de qualité (adjectifs), les verbes, les pronoms, les adverbes, prépositions, focaliseurs, interjections, les conjonctions, les pronoms, les particules et les déterminants, consistent en un seul mot se produisant entre deux blancs ou des signes de ponctuation (Ameur et al., 2006 ; Boukhris et al., 2008). Toutefois, 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: □□ (ⵖr) signifiant « pour, au », + □ (i) qui signifie « moi » (pronom personnel), donnent «□□□□/□□□□ (ⵖari/ⵖuri) » selon les régions;

4 Les signes de ponctuation amazighe sont semblables aux signes de ponctuation adoptée au niveau international et ont les mêmes fonctions ;

5 A l'instar d'autres langues naturelles, l'amazighe peut présenter des ambigüités au niveau des classes grammaticales. En effet, la même forme de surface peut appartenir à plusieurs catégories grammaticales selon le contexte dans la phrase. Par exemple, □□□□ (illi) peut fonctionner comme verbe à l'accompli négatif, il signifie «il n'existe pas », ou comme nom de parenté « ma fille ». Quelques mots tel que « □ » (d) peuvent fonctionner comme préposition, ou une conjonction de coordination ou particule de prédication ou d'orientation ;

6 De même que la majorité des langues dont les recherches en TAL ont récemment commencé, l'amazighe est peu doté en ressources langagières et outils du TAL.

Les corpus annotés sont d'une grande utilité dans la réalisation des dictionnaires, ainsi que dans plusieurs tâches du traitement automatique des langues (Manning and Schütze, 1999). Dans ce sens, nous avons annoté morphologiquement 20,667 jetons de l'amazighe marocain choisis à partir de sources diverses (Outahajala et al., 2011).

Le tableau 1 donne une description des sources choisies et le tableau 2 le nombre d'occurrences de chacune des parties du discours.

Corpus description	Nombre de jetons	Nombre de phrases
Manuel 2	5079	372
Manuel 5	2319	179
Manuel 6	3773	253
IRCAM web site	4258	185
Inghmisn n usinag	4636	415
Textes Divers	602	34
Total	**20667**	**1438**

Tableau 1: Description du corpus description.

Etiquette de la classe	Désignation	Nombre d'occurrences
V	Verbe	3190
N	Nom	4993
A	Nom de qualité	503
AD	Adverbe	516
C	Conjonction	834
D	Déterminant	1076
S	Préposition	2775
FOC	Focalisateur	91
I	Interjection	40
P	Pronom	1496
PR	Particule	1593
R	Résiduel (nom étranger, nombre, date, monnaie, signe mathématique ou autre)	178
F	Ponctuation	3382
Total		**20667**

Tableau 2: Nombre d'occurrences des parties du discours.

Les textes choisis ainsi que leurs spécifications morphologiques, sont encodés en utilisant le langage XML. Chaque mot de chaque texte est marqué avec les attributs et les sous attributs présentés dans nos travaux relatifs à la conception d'un corpus annoté pour

l’Amazighe (Outahajala et al., 2010). L’annotation a été effectuée en utilisant l’outil d’annotation AncoraPipe (Bertran et al., 2008).

2. Apprentissage supervisé

Trois types d’apprentissage existent ; l’apprentissage supervisé, l’apprentissage semi-supervisé dans lequel on utilise des données étiquetées avec des données non étiquetées et l’apprentissage actif. Dans cette section, nous décrivons les fondements théoriques de l'apprentissage supervisé de manière générale, des SVMs (Kudo and Matsomoto, 2000) et des CRFs en particulier, qui ont donné de bons résultats dans les problèmes de classification des séquences.

En apprentissage supervisé, l'objectif est d'apprendre une fonction:

$h : X \rightarrow Y$ (1)

Où $x \in X$ sont les entrées et $y \in Y$ sont les sorties. Les objets d'entrée sont appelés instances, ou des exemples qui peuvent être de tout type selon la tâche d'apprentissage particulier. En TAL elle peut être le classement des documents, étiquetage des chaînes de mots avec une séquence d’étiquettes, ce qui est notre cas,…etc. Selon la nature de l'espace Y de sortie, les tâches d'apprentissage peuvent être classées en plusieurs types: classification binaire, multi classification, régression et prédiction structurée. Par exemple, dans une tâche d'étiquetage des séquences tel que le POS-tagging, Y = {1 ..., K}, c'est à dire la sortie est une séquence d'étiquettes de longueur n égale à la longueur de la chaîne d'entrée.

Les séparateurs à vaste marge

Les séparateurs à vaste marge (SVMs) ont été introduits par Vapnik (Vapnik, 1995). Ils sont connus pour leur performance et leur généralisation. Ils ont été utilisés pour des problèmes de reconnaissance différents et ont donné de bons résultats.

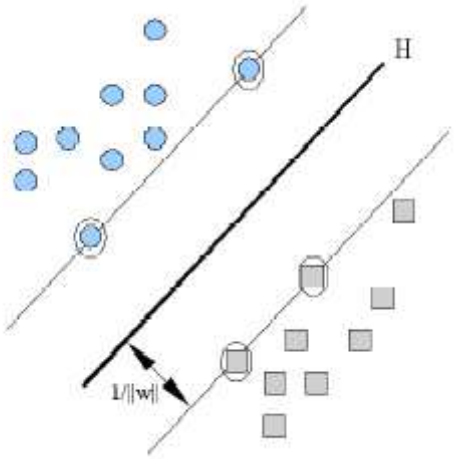

Figure 1: Séparation des régions par un hyperplan

L'objectif des SVMs est de trouver un hyperplan optimal, et c'est pour cette raison qu'on leur donne le nom de séparateurs à vaste marge. La marge est la distance entre la frontière de séparation et les échantillons les plus proches (Figure.1) qui sont appelés vecteurs supports. Dans les SVMs, la frontière de séparation est choisie de sorte à maximiser la marge.

En ce qui concerne la tâche de l'étiquetage grammatical de l'amazighe, le processus d'apprentissage a été conduit en utilisant YamCha[1], un outil basé sur les SVMs. Pour la classification, nous avons utilisé TinySVM[2], un outil public pour la reconnaissance des motifs.

Les champs markoviens conditionnels

Les champs markoviens conditionnels (CRFs) sont des modèles probabilistes discriminants introduits par (Lafferty et al. 2001) pour l'annotation séquentielle. Ce sont des graphes non orientés. Donnant une séquence d'observation, le modèle conditionnel indique les probabilités de séquences d'étiquettes possibles. En plus d'avoir les avantages des MEMMs ???, les CRFs peuvent être interprétés comme un modèle à états finis avec des probabilités de transition non normalisées. Les CRFs sont définis par X et Y, champs aléatoires décrivant respectivement chaque unité de l'observation et son annotation, et par un graphe G= (V, E) dont V est l'ensemble des nœuds et E l'ensemble des arcs, avec V=

[1] http://chasen.org/~taku/software/yamcha/
[2] http://chasen.org/~taku/software/TinySVM/

X U Y.

Deux variables sont reliées dans le graphe si elles dépendent l'une de l'autre. Chaque étiquette dépend des étiquettes précédentes et suivantes (Figure 2).

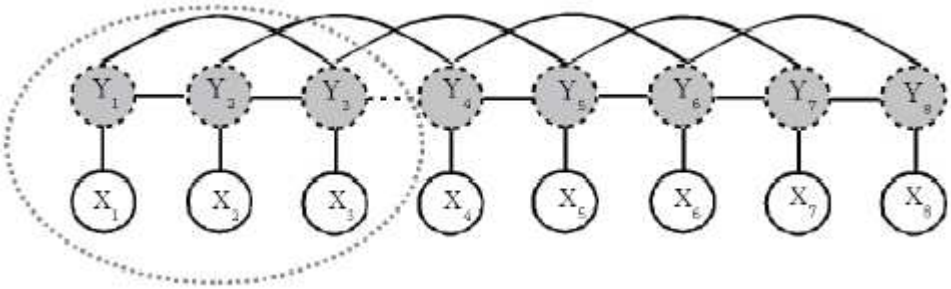

Figure 2 : Exemple d'un graphe des CRFs

Nous avons utilisé l'outil CRF++[3], une implémentation open source des CRFs pour la segmentation et l'étiquetage des données, en utilisant le même ensemble de données que celui utilisé avec Yamcha.

Nous avons utilisé la technique de « 10 fois validation croisée » pour évaluer notre approche et un jeu de 28 étiquettes. Dans cet ensemble d'expérimentation, nous avons mené et évalué nos expériences en utilisant la validation croisée en 10 parties, i.e. l'entraînement avec 90% du corpus de référence et l'utilisation de 10% pour le test, en répétant l'expérience dix fois et en prenant à chaque fois une tranche différente du corpus.

Les résultats montrent que les performances des SVMs et des CRFs sont comparables. Dans l'ensemble, les CRFs ont légèrement dépassé les SVMs au niveau des dossiers (91.18% contre 90.75%) et la moyenne de « 10 fois validation croisée » (89,48% contre 89,29%). La figure 3 synthétise les résultats des expérimentations de l'étiquetage grammatical.

[3] http://crfpp.sourceforge.net/

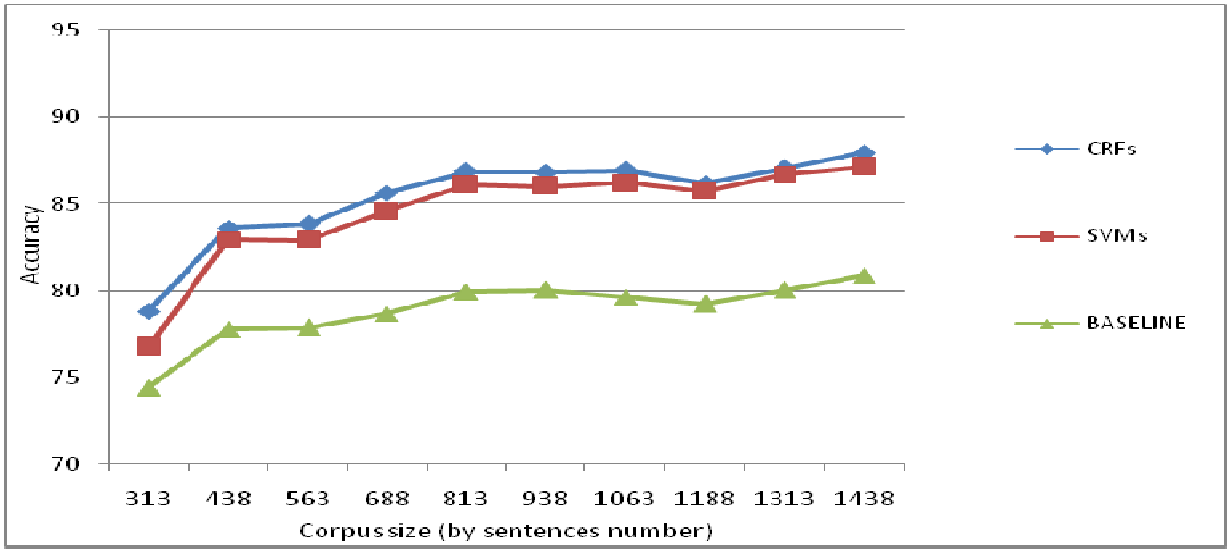

Figure 3 : Courbe d'apprentissage

Entre le septième et le huitième point de la courbe d'apprentissage de la figure3, la précision diminue légèrement. En effet le nombre des mots inconnus entre ces deux points a augmenté. *Baseline* est calculé en prenant pour chaque jeton l'étiquette la plus fréquente à partir du corpus d'entrainement et l'étiquette *nom commun* pour les mots inconnus.

3. Conclusion et travaux futurs

Dans ce papier, nous avons essayé de décrire les caractéristiques morphosyntaxiques de la langue amazighe. Nous avons abordé la réalisation d'un étiqueteur grammatical basé sur un jeu de 28 étiquettes, en adoptant les méthodes d'apprentissage supervisé, en particulier les SVMs et les CRFs. Ces dernières ont donné de bons résultats dans les tâches de classification des séquences pour les autres langues.

Ces résultats sont prometteurs, étant donné que nous avons utilisé un corpus de petite taille. Actuellement, nous avons collecté un corpus de diverses sources avec un total de 225.240 jetons. Nous essayons actuellement d'améliorer les performances de l'étiqueteur, en utilisant ces données non étiquetées obtenues à l'aide des techniques d'apprentissage semi-supervisé.

A court terme, nous étudierons l'apport en performance à cette tâche d'étiquetage grammatical de l'Amazighe, en utilisant les méthodes d'apprentissage actif.

Références bibliographiques :

-Ameur, M., Bouhjar, A., Boukhris, F. Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. (2006). *Graphie et orthographe de l'Amazighe*. Publications de l'IRCAM.

-Boukhris, F. Boumalk, A. El moujahid, E., Souifi, H. (2008). *La nouvelle grammaire de l'amazighe*. Publications de l'IRCAM.

-Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008), AnCoraPipe A tool for multilevel annotation. *Procesamiento del lenguaje Natural*, nº 41. Madrid, Spain.

-Kudo, T., Yuji Matsumoto, Y. (2000). Use of Support Vector Learning for Chunk Identification. *Proceeding of CoNLL-2000 and LLL-2000*.

-Lafferty, J. McCallum, A. Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proc. of ICML-01*, pp. 282-289. 2001.

-Manning, C., Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

-Outahajala M., Zenkouar L., Rosso P., Martí A. (2010), Tagging Amazighe with AncoraPipe. *In: Proc. Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation*, LREC-2010, Malta, May 17-23, pp. 52-56.

-Outahajala M., Zenkouar L., Rosso P. (2011), Building an annotated corpus for Amazighe. I*n Proc. of 4th International Conference on Amazigh and ICT*. Rabat, Morocco.

-Vapnik, Valdimir N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York, USA.

-Zenkouar L. (2004), « L'écriture Amazighe Tifinaghe et Unicode ». *revue Etudes et Documents Berbères n°22*, pp. 175-192.

-Zenkouar L. (2008), « Normes des technologies de l'information pour l'ancrage de l'écriture Amazighe ». *revue Etudes et Documents Berbères n°27*, pp. 159-172.