L'étiquetage grammatical de l'amazighe en utilisant les propriétés n-grammes et un prétraitement de segmentation

Mohamed Outahajala,

Laboratoire Electronique et Communication- Ecole Mohammadia d'Ingénieurs- Institut Royal de la Culture Amazighe, Maroc, outahajala@ircam.ma

Yassine Benajiba,

Philips Research North America, Briacliff Manor, USA, yassine.benajiba@philips.com

Paolo Rosso,

Natural Language Engineering Lab - ElíRF, DSIC, Universidad Politécnica de Valencia, Spain, prosso@dsic.upv.es

Lahbib Zenkouar,

Laboratoire Electronique et Communication- Ecole Mohammadia d'Ingénieurs, Rabat, Maroc, zenkouar@emi.ac.ma

Résumé

L'objectif de cet article est de présenter le premier étiqueteur grammatical amazighe. Très peu de ressources ont été développées pour l'amazighe et nous croyons que le développement d'un outil d'étiquetage grammatical est une étape préalable au traitement automatique de textes. Afin d'atteindre cet objectif, nous avons formé deux modèles de classification de séquences en utilisant les SVMs, séparateurs à vaste marge (*Support Vector Machines*) et les CRFs, champs markoviens conditionnels (*Conditional Random Fields*) en utilisant une phase de segmentation. Nous avons utilisé la technique de 10 fois la validation croisée pour évaluer notre approche. Les résultats montrent que les performances des SVMs et des CRFs sont très comparables. Dans l'ensemble, les SVMs ont légèrement dépassé les CRFs au niveau des échantillons (92,58% contre 92,14%) et la moyenne de précision des CRFs dépasse celle des SVMs (89,48% contre 89,29%). Ces résultats sont très prometteurs étant donné que nous avons utilisé un corpus de seulement ~ 20k mots.

Abstract

The aim of this paper is to present the first amazigh POS tagger. Very few linguistic resources have been developed so far for amazigh and we believe that the development of a POS tagger tool is the first step needed for automatic text processing. In order to achieve this endeavor, we have trained two sequence classification models using Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) after using a tokenization step. We have used the 10-fold technique to evaluate our approach. Results show that the performance of SVMs and CRFs are very comparable. Across the board, SVMs outperformed CRFs on the fold level (92.58% vs. 92.14%) and CRFs outperformed SVMs on the 10 folds average level (89.48% vs. 89.29%). These results are very promising considering that we have used a corpus of only ~20k tokens.

Mots-clés

étiquetage grammatical automatique, langue amazighe, TAL, apprentissage supervisé, segmentation

Keywords

automatic POS tagging, amazigh language, NLP, supervised learning, segmentation

1. Introduction

L'étiquetage grammatical[1], appelé également étiquetage morpho-syntaxique, consiste en l'annotation de chaque mot d'une phrase avec sa catégorie grammaticale. Il s'agit de la première couche au-dessus du niveau lexical et le niveau le plus bas de l'analyse syntaxique. Ainsi, le POS tagging est utilisé pour toutes les tâches du traitement automatique des langues (TAL) traitant des niveaux linguistiques supérieurs utilisent telles que l'analyse partielle, la désambiguïsation de sémantiques des mots, l'affectation des fonctions grammaticales (Cutting et al, 1992) et la reconnaissance d'entités nommées (Benajiba *et al.*, 2010a; Benajiba *et al.* 2010b). Conjointement avec l'analyse partielle, l'étiquetage grammatical est utilisé dans des tâches plus complexes (Manning et Schütze, 1999) telles que l'acquisition lexicale, l'extraction des informations, la recherche des termes d'indexation dans la récupération de l'information et les questions réponses.

Dans la littérature, il a été montré que les approches fondées sur l'apprentissage supervisé (cf. section 2) sont les plus efficaces pour construire les étiqueteurs grammaticaux, en s'appuyant sur un corpus annoté manuellement et souvent d'autres ressources, telles que des dictionnaires et des outils de segmentation. Pour construire notre étiqueteur grammatical, nous utilisons des techniques de classification des séquences supervisée fondées sur deux approches de ces techniques, à savoir les SVMs (Support Vector Machines) et les CRFs (Conditional Random Fields). Nous nous appuyons sur un corpus de ~20k mots annotés manuellement (Outahajala *et al.*, 2011) pour former nos modèles et les caractéristiques n-grammes lexicales en vue d'une amélioration de la performance.

Le reste de l'article est organisé comme suit : à la section 2, nous présentons les travaux connexes sur les techniques d'étiquetage grammatical dans d'autres langues. Puis, dans la section 3, nous fournissons un aperçu sur la langue amazighe et le jeu d'étiquettes employé dans l'étiquetage du corpus Amazighe collecté. Dans la section 4, nous présentons les deux approches d'apprentissage supervisées utilisant les SVMs et les CRFs et qui ont été employées pour le marquage grammatical. Dans la section 5, nous décrivons les expériences et nous discutons les résultats. Enfin, dans la section 6, nous dressons quelques conclusions et nous présentons les travaux à effectuer dans le futur proche.

2. Etat de l'art

De nombreux systèmes pour l'étiquetage automatique des parties du discours ont été développés pour un large éventail de langues. Parmi ces systèmes, certains s'appuient sur les règles linguistiques et d'autres sur des techniques d'apprentissage automatique (Manning & Schütze, 1999). Les premiers POS taggeurs étaient principalement à base de règles. La construction de tels systèmes nécessitent un travail considérable afin d'écrire manuellement les règles et de coder les connaissances linguistiques qui régissent l'ordre de leur application. Un exemple d'étiqueteur à base de règles est TAGGIT, développé par Green et Robin (Greene, Rubin, 1971) et contenant environ 3300 règles. Ce système atteint une précision de 77%. Par la suite, l'apprentissage automatique des étiqueteurs s'est avéré à la fois moins pénible et plus efficace que ceux à base de règles. Dans la littérature, de nombreuses méthodes d'apprentissage automatique ont été appliquées avec succès pour réaliser des POS taggers. Nous les citons dans ce qui suit.

- Les Modèles de Markov Cachés (HMM) (Charniak, 1993) dont les états sont des balises ou des tuples de balises. Pour un tagueur bigrammes par exemple, les états de l'HMM sont les

[1] La tradition anglophone utilise le terme *part-of-speech tagging (POS tagging),* c'est-à-dire étiquetage des parties du discours.

balises, les probabilités de transition sont les probabilités d'une étiquette donnant l'étiquette précédente et les probabilités d'émission sont les probabilités d'un mot sachant une étiquette donnée.

- La transformation système fondée sur la réduction du taux d'erreur (Brill, 1995) consiste en l'affectation de l'étiquette la plus fréquente d'un mot donné en utilisant un corpus de référence. Elle procède par la suite en sélectionnant la règle qui donne la plus grande erreur. Ce processus est itéré tant que les résultats d'annotation ne sont pas suffisamment proches de celles du corpus de référence.

- Les arbres de décision permettent de construire un outil d'aide à la décision (Schmid, 1999) sur la base d'un corpus de référence qui utilise ce modèle. La meilleure étiquette attribuée à un mot donné est celle qui donne la plus forte probabilité conditionnelle pour le nœud courant de cet arbre.

- Le modèle d'entropie maximale (Ratnaparkhi, 1996) permet la combinaison de diverses formes d'informations contextuelles sans imposer aucune hypothèse sur les données d'entraînement. Son but est de maximiser l'entropie de la distribution d'un mot à certaines contraintes contenues dans le corpus de référence.

- Les méthodes d'apprentissage automatique permettent de construire des modèles complexes (comportant de très nombreux paramètres), difficile à faire manuellement. La qualité des modèles est souvent liée à la quantité de données utilisées dans l'apprentissage. Ainsi à partir d'exemples appris précédemment, les programmes s'appuyant sur ces méthodes affectent l'étiquette aux mots selon le contexte (Schmid, 1994; Kudo, Matsumoto, 2000; Lafferty *et al.* 2001).

- Les méthodes hybrides qui utilisent à la fois des règles à base de connaissances linguistiques codées manuellement et les méthodes d'apprentissage automatique.

Les résultats annoncés utilisant ces méthodes sont supérieurs à 95%. Bien que ces méthodes aient une bonne performance, la précision des mots inconnus est beaucoup plus faible que celle des mots connus, ce qui est problématique lorsque le corpus d'apprentissage est de petite taille. La taille du jeu d'étiquettes varie considérablement selon la langue et la finalité de la tâche d'étiquetage voulue. Leech (1997) montre que le nombre d'étiquettes varie de 32 à 270 dans les principaux corpus anglais. Dans la pratique, la plupart des analyseurs limitent le nombre d'étiquettes en ignorant certaines distinctions difficiles à désambiguïser automatiquement, ou sujettes à discussion du point de vue linguistique.

3. L'amazighe

Dans cette section, nous présenterons un bref aperçu sur la langue amazighe avant de décrire le jeu de balises adoptées dans nos expérimentations.

3.1 La langue amazighe

La langue amazighe est parlée au Maroc, en Algérie, en Libye, en Tunisie et en Egypte (oasis de Siwa). Elle est également parlée par d'autres communautés dans certaines régions du Niger et du Mali. L'amazighe est un composite de dialectes dont aucun n'a été considéré comme la norme nationale dans aucun des pays déjà susmentionnés. Avec l'émergence de la revendication identitaire, les locuteurs natifs militent pour la sauvegarde et la promotion de leur langue et de leur culture. Pour atteindre un tel objectif, certains Etats du Maghreb ont créé des institutions spécialisées, telles que l'Institut Royal de la Culture Amazighe (IRCAM, désormais) au Maroc et le Haut-commissariat de l'Amazighité en Algérie. Au Maroc,

l'amazighe a été introduite dans les médias et dans le système éducatif. En conséquence, l'Alphabet Tifinaghe a été reconnu officiellement par le consortium Unicode le 05/07/2004, une nouvelle chaîne de télévision amazighe a été lancée le 1e mars 2010, un peu plus de 3 000 écoles primaires marocaines enseignent l'Amazighe à plus de 600 000 élèves. Au niveau de l'enseignement supérieur, des filières d'études amazighes et des masters ont été créés. Le 01 Juillet 2011, les Marocains ont voté favorablement pour la nouvelle constitution du pays qui octroie le statut langue officielle à l'amazighe.

Nous remarquons une croissance notable du nombre de publications traitant de la langue et de la culture amazighes. Cependant, en TAL, la langue amazighe, comme la plupart des langues non européennes, souffre encore de la pénurie d'outils et des ressources pour son traitement automatique.

Sur le plan linguistique, la langue est caractérisée par la prolifération des dialectes en raison de facteurs historiques, géographiques et sociolinguistiques. Au Maroc, par exemple, on peut distinguer trois principaux dialectes : le tarifite dans le Nord, le tamazighte dans le centre et le tachlhite dans le sud du pays.

En raison de sa morphologie complexe (Chafiq, 1991 ; Boukhris *et al.* 2008) ainsi que de l'utilisation des différents dialectes dans sa normalisation, la langue amazighe présente des défis intéressants à prendre en compte pour les chercheurs en Traitement Automatique des Langues (TAL). Nous en citons dans ce qui suit quelques uns.

- L'amazighe dispose de sa propre graphie, le Tifinaghe qui s'écrit de gauche à droite.
- Le Tifinaghe ne contient pas de majuscules.
- Les noms, les noms de qualité (adjectifs), les verbes, les pronoms, les adverbes, les prépositions, les focaliseurs, les 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. Toutefois, si une préposition ou un nom de parenté est suivi d'un pronom personnel, l'ensemble préposition/nom de parenté et le pronom qui suit forment une chaîne unique délimitée par des espaces ou des signes de ponctuation. Par exemple: □□ (ɣr) signifiant « pour, au » + □ (i) qui signifie « moi » (pronom personnel) donnent «□□□□/□□□□ (ɣari/ɣuri) ».
- Les signes de ponctuation amazighe sont semblables aux signes de ponctuation adoptés au niveau international et ont les mêmes fonctions.
- L'amazighe, à l'instar d'autres langues naturelles, 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 un verbe à l'accompli négatif signifiant «il n'existe pas » ou comme nom de parenté signifiant « ma fille ». Quelques mots tels que « □ » (d) peuvent fonctionner comme préposition ou comme conjonction de coordination ou comme particule de prédication ou d'orientation.
- L'amazighe est peu dotée en ressources langagières et outils du TAL au même titre que la majorité des langues dont les recherches en TAL ont récemment commencé.

3.2 Jeu d'étiquettes

Définir le jeu de balises adéquat est une tâche essentielle dans la construction d'un POS tagger automatiquement. Il vise à définir un ensemble de balises traitables qui ne soit ni grand et nuire à la performance de l'apprentissage automatique, ni petit et n'offrir pas suffisamment d'informations pour être utilisé par les systèmes fédérateurs. Dans (Outahajala *et al.*, 2010),

un ensemble contenant 13 balises (verbe, nom, adverbe, etc.) a été conçu. Pour chaque élément, les caractéristiques morphosyntaxiques et deux attributs communs ont été définis : «wd» pour mot et «lem» pour lemme, dont les valeurs dépendent de l'item lexical en question. L'ensemble des balises utilisées comprend 13 étiquettes représentant les principales parties du discours dans la langue amazighe, comme il est résumé dans le tableau 1. Ce jeu de balises est un sous-ensemble du jeu d'étiquettes présentée dans (Outahajala *et al.*, 2010). Le genre, la personne, l'aspect et d'autres informations n'ont pas été inclus dans ce jeu d'étiquettes et ont été considérées comme une piste de recherche à poursuivre dans l'avenir.

Tableau 1. Jeu d'étiquettes

Classe	Désignation
V	Verbe
N	Nom
A	Nom de qualité/Adjectif
AD	Adverbe
C	Conjonction
D	Déterminant
S	Préposition
FOC	Focalisateur
I	Interjection
P	Pronom
PR	Particule
R	Résiduel (nom étranger, nombre, date, monnaie, signe mathématique, etc.)
F	Ponctuation

4. Apprentissage supervisé pour l'étiquetage grammatical

Dans cette section, nous décrivons les fondements théoriques de l'apprentissage supervisé en général, et des SVMs et des CRFs en particulier. Les bonnes performances de ces derniers ont été prouvées dans les problèmes de classification des séquences.

4.1. Apprentissage supervisé

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

$$h : X \rightarrow Y \quad (1)$$

où $x \in X$ sont les entrées et $y \in Y$ sont les sorties.

Les objets d'entrée appelés instances ou exemples peuvent être de tout type, selon la tâche d'apprentissage particulière. En TAL, une entrée peut être un classement des documents, des chaînes de mots à étiqueter avec une séquence d'étiquettes (ce qui est notre cas).

Selon la nature de l'espace Y de sortie, les tâches d'apprentissage peuvent être classées en plusieurs types :

- Classification binaire avec Y = {−1, +1}
- Multi classification avec Y = {1, …, K} (ensemble fini de balises)
- Régression avec Y = R
- Prédiction structurée dans le cas où les sorties sont complexes. Par exemple, dans une tâche d'étiquetage des séquences telles que le POS-tagging, Y = {1 ..., K} signifie que la sortie est une séquence d'étiquettes de longueur égale à la longueur de la chaîne d'entrée.

4.2. Les séparateurs à vaste marge

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

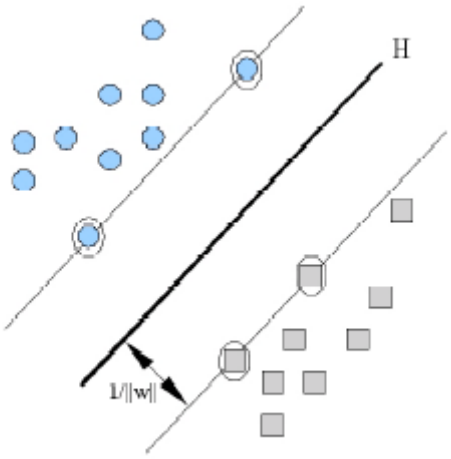

Figure.1. Séparation des régions par un hyperplan

L'objectif des SVMs étant de trouver un hyperplan optimal, on les appelle « 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. Le problème est de trouver cette frontière séparatrice optimale, à partir d'un ensemble d'apprentissage. Ceci est fait en formulant le problème comme un problème d'optimisation.

En TAL, les SVMs sont appliquées à la catégorisation de textes (Kudu, Matsomoto, 2000), l'analyse partielle (Diab *et al.*, 2007), la reconnaissance d'entités nommées (Benajiba *et al.*, 2010), etc. Plusieurs étiqueteurs grammaticaux ont été réalisés sur la base des SVMs. C'est le cas de l'arabe (Diab *et al.*, 2004; Diab *et al.*, 2007), du bengali (Ekbal, Bandyopadhyay, 2008), etc. Les SVMs atteignent des performances élevées sans apprentissage même en utilisant plusieurs caractéristiques. Ils réagissent également bien avec les données éparses et bruitées.

En ce qui concerne la tâche de l'étiquetage grammatical de l'amazighe, le processus d'apprentissage a été conduit en utilisant YamCha[2], un outil fondé sur les SVMs. Pour la

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

classification, nous avons utilisé TinySVM[3], un outil public pour la reconnaissance des motifs.

4.3. Les champs markoviens conditionnels

Les CRFs ou champs markoviens conditionnels sont des modèles probabilistes discriminants introduits par (Lafferty *et al.*, 2001) pour l'annotation séquentielle. Il s'agit de graphes non orientés. Etant donnée une séquence d'observation, le modèle conditionnel indique les probabilités de séquences d'étiquettes possibles. En plus des avantages des MEMMs, les CRFs peuvent être imaginés comme un modèle à états finis avec des probabilités de transition non normalisée. Ils sont appliqués dans de nombreuses tâches du TAL, telles que la reconnaissance d'entités nommées (Benajiba *et al.*, 2010), l'analyse syntaxique partielle (Sha, Pereira, 2003) et l'extraction d'information à partir de tables (Pinto *et al.*, 2003). En relation avec l'étiquetage grammatical, les CRFs ont été utilisés pour de nombreuses langues, telles que l'Amharique (Adafre, 2005) et le Tamoul (Lakshmana, Geetha, 2009).

Les CRFs sont définis par des champs aléatoires X et Y 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= 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.

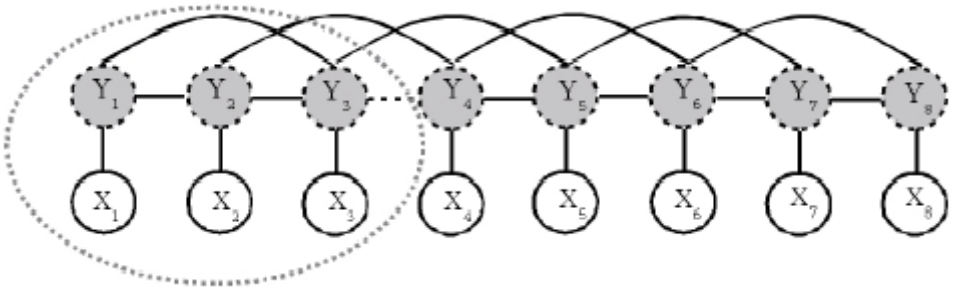

Figure.2. Exemple d'un graphe des CRFs

Dans un CRF, étant donnée une observation x et une annotation y on a la relation suivante (Lafferty et al, 2001) :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \exp\left(\sum_{k} \lambda_k f_k(y_c, x, c) \right) \qquad (2)$$

où

- C est l'ensemble des cliques -sous-graphes complètement connectés- de G
- y_c est l'ensemble des valeurs prises par les variables de Y sur la clique c
- Z(x) est un coefficient de normalisation défini de telle sorte que la somme sur y de toutes les probabilités p(y|x) pour une donnée x fixée soit égale à 1

[3] http://chasen.org/~taku/software/TinySVM/

- f_k sont des fonctions caractéristiques définies à l'intérieur de chaque clique c et sont à valeurs réelles. La valeur de ces fonctions peut dépendre des étiquettes présentes dans une certaine clique c ainsi que de la valeur de l'observation x

- λ_k sont des poids à valeurs réelles. Ils permettent de fixer l'importance à chaque fonction f_k. Leurs valeurs sont fixées lors de la phase d'apprentissage en cherchant à maximiser la log-vraisemblance sur des exemples déjà annotés formant le corpus de référence.

Nous avons utilisé l'outil CRF++4, 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.

5. Expérimentations et analyse des erreurs

5.1. Corpus

Notre corpus se compose de textes extraits d'une variété de sources telles que la version amazighe du site Web de l'IRCAM[5], le périodique «Inghmisn n usinag » (bulletin d'nformation de IRCAM) et des manuels scolaires. Les textes ont été annotés en utilisant l'outil AnCoraPipe (Bertran *et al.*, 2008). La vitesse d'annotation de ce corpus est de 80 et 120 mots par heure. Afin de calculer l'accord entre les annotateurs, des textes ont été choisis aléatoirement et révisés par des annotateurs différents. Le résultat de cette expérience a donné un accord de 94,98%. Les remarques et les corrections communes ont été généralisées à l'ensemble du corpus dans la deuxième validation par un annotateur autre que celui qui a fait l'annotation. Le format d'entrée pour Yamcha et CRF++ est le même (voir Figure 3 et Figure 4 ci-dessous). La première colonne de l'entrée est formée des mots de notre corpus, la dernière colonne est réservée à la classe grammaticale, et le cas échéant les caractéristiques lexicales n-grammes sont mises dans les colonnes entre les deux (voir Figure 4). i.e., la première et la dernière colonnes.

Nous décrivons ci-dessous un exemple dans lequel nous n'utilisons pas de segmentation du format d'entrée pour la phrase « ar as ttHyyaln i tmGra ann sg usggwas lli izrin ». [En Français : Ils se préparaient au mariage depuis l'année dernière]:

ar	PR
as	S_P
ttHyyaln	V
i	S
tmGra	N
ann	D
sg	S
usggwas	N
lli	P
izrin	V
.	F

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

[5] http://www.ircam.ma/

Figure 3. Un extrait à partir du corpus d'apprentissage

ar	a	-	-	-	r	-	-	-	PR
as	a	-	-	-	s	-	-	-	S_P
ttHyyaln	t	tt	ttH	ttHy	n	ln	aln	yaln	V
i	-	-	-	-	-	-	-	-	S
tmGra	t	tm	tmG	tmGr	a	ra	Gra	mGra	N
ann	a	an	-	-	n	nn	-	-	D
sg	s	-	-	-	g	-	-	-	S
usggwas	u	us	usg	usgg	s	as	was	gwas	N
lli	l	ll	-	-	i	li	-	-	P
izrin	i	iz	izr	izri	n	in	rin	zrin	V
.	-	-	-	-	-	-	-	-	F

Figure 4. Un extrait à partir du corpus d'apprentissage en utilisant les propriétés lexicales

Dans cet article, nous explorons deux groupes d'expériences, basés sur les SVMs et les CRFs, avec ou sans propriétés lexicales (voir paragraphe 5.2). Dans la première expérimentation, nous ne segmentons pas les mots contenant plusieurs morphèmes lexicaux, et nous utilisons "S_P" et "N_P" pour désigner respectivement les prépositions et les noms de parenté lorsqu'ils sont suivis des pronoms personnels. Dans la deuxième expérimentation, nous segmentons les prépositions et les noms de parenté quand ils sont utilisés avec les pronoms personnels. Toutefois, ceci est un cas problématique car la fonction inverse n'est pas déterministe. En effet, si par exemple l'union des deux morphèmes «dg» et «s» peut donner soit "digs" ou "dags", signifiant [en lui/elle]. Ainsi, une fois nous avons subdivisé le mot composé en ses morphèmes constituants, il n'est plus possible de calculer la forme originale après l'étiquetage grammatical en ne se basant que sur la sortie du POS tagger. Une solution à cette question est de prendre la forme la plus utilisée dans le corpus en se basant sur la fréquence des mots, ainsi on prendra pour ce cas la forme "digs". Dans toutes nos expériences, nous avons utilisé deux ensembles de balises celui présenté ci-dessus (voir le tableau 1) et ce même ensemble plus les deux balises "S_P" et "N_P" correspondant aux prépositions et aux noms de parenté lorsqu'ils sont employés avec les pronoms.

5.2. Caractéristiques pour l'étiquetage

Dans cet article, nous explorons deux ensembles de caractéristiques basées sur le texte réel et faciles à extraire. Dans le premier sous-ensemble d'expérimentations, nous utilisons les mots qui entourent le mot à étiqueter ainsi que leurs étiquettes dans une fenêtre de +/- 2 mots (Figure 3). Dans le deuxième sous-ensemble, (illustré dans la Figure 4), nous ajoutons aux premières caractéristiques les propriétés lexicales n-grammes du mot à étiqueter et des mots entourant ce mot, avec la même fenêtre de +/- 2. Les propriétés n-grammes se composent des i premiers et i derniers n-grammes caractères, avec i variant de 1 à 4.

5.3. Les expérimentations

Dans nos premières expériences sur l'étiquetage grammatical (Outahajala *et al.* 2011b), nous avons montré que la courbe d'apprentissage croît avec l'augmentation de la taille du corpus d'entraînement.

Tableau 2. Résultats de la validation croisée en 10 des parties

Partie#	SVMs	SVMs (avec les propriétés lexicales)	CRFs	CRFs (avec les propriétés lexicales)
0	81,01	86,86	83,19	86,95
1	76,02	83,86	80,7	84,98
2	85,64	91,66	87	90,86
3	82,56	88,34	86,45	88,58
4	83,55	88,24	85,8	88,87
5	83,28	89,99	86,24	90,48
6	76,59	85,38	79,98	85,38
7	79,07	86,6	81,79	87,96
8	87,35	91,38	88,88	91,14
9	84,64	90,41	86,79	91,35
Moyenne	81,97	88,27	84,68	88,66

Nous avons mené et évalué nos expériences en utilisant une 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.

Tableau 3. Résultats de la validation croisée en 10 parties après une phase de segmentation

Partie #	SVMs	SVMs (avec les propriétés lexicales)	CRFs	CRFs (avec les propriétés lexicales)
0	82,85	87,94	84,46	87,31
1	78,27	85,06	81,55	85,9
2	87,59	92,58	87,9	91,42
3	83,95	89,62	87,39	89,22
4	85,06	89,02	86,93	89,26
5	86,08	91,38	87,6	91,62
6	79,27	86,42	82,9	87,18
7	81,34	86,96	83,69	88,96
8	88,54	92,47	89,32	91,79
9	86,45	91,49	88,65	92,14
Moyenne	83,93	89,29	86,01	89,48

Vu l'impact positif de la segmentation, comme prétraitement, à l'instar de ce qui se fait pour d'autres langues comme l'arabe (Diab *et al*., 2004), nous avons subdivisé les prépositions et les noms de parenté lorsqu'ils sont utilisés conjointement avec les pronoms personnels, et nous avons obtenu de meilleurs résultats, comme il est indiqué dans le tableau 4. Par la suite, vu l'amélioration des résultats obtenus en utilisant cette phase de segmentation comme prétraitement à l'étiquetage, nous avons décidé de réaliser un étiqueteur, en s'appuyant sur les méthodes d'apprentissage automatique. Pour ce faire, nous avons entrainé deux modèles à base de séquences d'étiquettes en utilisant cinq étiquettes : {B-WORD, I-WORD, B-SUFF, I-SUFF, O}. Le corpus utilisé pour l'entrainement de ce segmenteur a été construit de manière semi-automatique. Sur la base d'expérimentations, nous avons constaté que les performances des SVMs et des CRFs sont très comparables. Le modèle à base des SVMs a légèrement dépassé celui basé sur les CRFs (99.95% contre 99,89%).

5.4. Expérimentations et discussion des résultats des étiqueteurs

Pour une meilleure compréhension du comportement du système, nous avons examiné la matrice de confusion pour l'expérience qui a donné la plus grande précision. L'analyse de la matrice de confusion présente toutes les étiquettes erronées comme le montrent les tableaux 4 et 5.

Tableau 4. La matrice de confusion en pourcentage en utilisant les SVMs avec les caractéristiques lexicales

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	93,1	0,3	1,8	0,6	3,9	0	0	0	0	0	0,3	0	0
A	18,2	63,6	18,2	0	0	0	0	0	0	0	0	0	0
V	5,4	0,3	93	0	0	0,7	0	0	0,7	0	0	0	0
P	0,7	0	0,7	91	5,5	0,7	0,7	0	0,7	0	0	0	0
D	3,3	0	1,1	9,9	84,6	0	0	1,1	0	0	0	0	0
S	0,5	0	1	0,5	0	94	2,1	0,5	1,6	0	0	0	0
C	0	0	0	2,1	2,1	2,1	83,3	4,2	4,2	2,1	0	0	0
AD	23,2	0	7,1	1,8	1,8	3,6	1,8	60,7	0	0	0	0	0
PR	0	0	0	0	1,9	0,6	0	0,6	96,8	0	0	0	0
FOC	0	0	0	0	40	0	0	0	0	60	0	0	0
F	0	0	0	0,2	0	0	0	0	0	0	99,8	0	0
I	36,4	0	0	0	0	0	0	0	18,2	0	0	45,4	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0

En analysant les erreurs les plus fréquentes dans les deux matrices de confusion, nous avons

constaté que les adjectifs sont souvent étiquetés comme substantifs. Ceci est dû au fait que les adjectifs peuvent être utilisés comme des noms. En optant pour une non distinction entre les noms et les adjectifs, nous obtenons une amélioration de 0,73 et un meilleur score de 90,02% en utilisant la validation croisée en 10 parties avec les SVMs. La même expérience avec les CRFs permet une amélioration de 0,77 et un meilleur score de 90,25%.

Tableau 5. La matrice de confusion en pourcentage en utilsant les CRFs avec les caractéristiques lexicales.

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	94,6	2,4	2,1	0,2	0,3	0,1	0,2	0	0	0	0	0	0,1
A	12,6	82,3	4,6	0	0	0	0	0	0	0,6	0	0	0
V	2,6	1,5	93,3	0	0,4	1,5	0	0,4	0,4	0	0	0	0
P	3,7	0	0	75	13,9	0,9	0,9	0,9	3,7	0,9	0	0	0
D	2,4	0	0	4,8	82,5	0	0	2,4	7,9	0	0	0	0
S	0	0	0,2	0,3	0	99	0,5	0	0	0	0	0	0
C	1,7	0	0,6	0	0,6	2,9	91,4	0	2,9	0	0	0	0
AD	23,8	0	0	9,5	0	9,5	14,3	42,9	0	0	0	0	0
PR	0	0	1,1	1,1	2,3	1,1	9,2	1,1	83,9	0	0	0	0
FOC	0	0	0	0	0	0	0	0	50	50	0	0	0
F	0	0	0	0	0	0	0	0	0	0	100	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0
R	3,1	0	1,6	1,6	0	4,7	0	0	4,7	0	6,3	0	78,1

Le taux d'erreur des pronoms est également élevé en raison du chevauchement important entre eux et les déterminants. Les verbes constituent une autre source commune d'erreurs. Comme le montre le Tableau 5, le POS tagger basé sur les CRFs a étiqueté 4,1% des verbes comme des noms et des adjectifs et de 1,6% comme des prépositions, alors que le POS-tagger basé sur les SVMs a étiqueté 5,7% des verbes comme des noms et des adjectifs. Pour les autres classes le POS-tagger basé sur les SVMs a de meilleurs résultats dans l'étiquetage des pronoms, des déterminants, des adverbes, des focalisateurs et des particules.

Certaines particules ont un niveau d'ambigüité important, par exemple la particule ⵙ(d) qui a plusieurs balises possibles, en fonction du contexte. Par exemple, le mot «d» pourrait être :

- Une conjonction de coordination: «tamaziGt d tiknulujiyin timaynutin" [l'amazighe et les nouvelles technologies],
- Une préposition: "iman d ubrid" [il a pris le chemin],
- Une particule de prédication: «d argaz » [il est un homme]
- Une particule d'orientation: «asi d tikint tamjahdit » [prend le grand bol].

En analysant les corpus d'entrainement et de test, nous avons observé que les mots inconnus sont importants. Les mots mal classés du corpus de test sont invisibles dans le corpus

d'entrainement. Aussi, certaines erreurs existent-elles encore dans notre corpus manuellement annoté.

6. Conclusions et travaux futurs

Dans cet article, nous avons essayé de décrire les caractéristiques morphosyntaxiques de la langue amazighe. Nous avons abordé la conception de deux jeux d'étiquettes et deux étiqueteurs grammaticaux basés sur les SVMs et les CRFs. La précision obtenue a atteint 92,58%. Pour valider nos résultats, nous avons utilisé un petit corpus annoté manuellement d'environ ~20k mots, les caractéristiques lexicales et une phase de prétraitement de segmentation et la technique de 10 fois la validation croisée. Le tagueur grammatical utilisant les CRFs a atteint 89,48% alors que celui basé sur les SVMs a atteint une précision de 89,29%.

Comme perspective, nous visons l'amélioration des performances de l'étiqueteur en élargissant le jeu d'étiquettes en utilisant plus de données étiquetées et obtenues avec les techniques d'apprentissage semi-supervisé et l'apprentissage actif.

Remerciements : nous tenons à remercier tous les chercheurs linguistes de l'IRCAM de leur aide précieuse. Les travaux du troisième auteur ont été financés par le projet de recherche EU FP7 Marie Curie PEOPLE-IRSES 269180 WiQ-Ei, MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i), VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Références

Adafre, S. F. (2005), Part of Speech tagging for Amharic using Conditional Random Fields. *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 47-54.

Benajiba Y., Diab M., Rosso P. (2010a), Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. *Special Issue on Processing Morphologically Rich Languages*, pp. 926-934. DOI: 10.1109/TASL.2009.2019927.

Benajiba Y., Zitouni I., Diab M., Rosso P. (2010b), Arabic Named Entity Recognition: Using Features Extracted from Noisy Data. *In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL-2010, Uppsala, Sweden, July 11-16, pp. 281-285.

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.

Brants, T. (2000), TnT - A Statistical Part-of-Speech Tagger. *In Proceedings. Of the 6th Applied Natural Language Proces- sing Conference*. Seattle, USA.

Brill, E. (1995), Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*. vol 21,pp. 543—565.

Chafiq, M. (1991), *الأمازيغية في درسا وأربعون أربعة* . éd. Arabo-africaines.

Charniak, E. (1993), *Statistical Language Learning.* MIT Press, Cambridge

Cutting, D., Kupiec, J., Jan Pedersen, J. Sibun, P. (1992), Practical Part-of-Speech Tagger. Xerox Palo Alto Research Center.

Diab, M., Hacioglu, K., Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL).*

Diab, M., Hacioglu, K., Jurafsky, D. (2007), Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter 9. Springer.

Ekbal, A.; Bandyopadhyay, S. (2008), Part of Speech Tagging in Bengali Using Support Vector Machine. *In Information Technology*, ICIT '08, pp. 106-111.

Greene, B.B., and Rubin, G.M. (1971), Automatic Grammatical Tagging of English. Department of Linguistics, Brown University, Providence, R.I.

Kudo, T., Yuji Matsumoto, Y. 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.

Lakshmana Pandian S., Geetha T. V. (2009), CRF Models for Tamil Part of Speech Tagging and Chunking. *In: Proc. ICCPOL '09*. Springer-Verlag Berlin, Heidelberg

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

Leech, G. (1997). Introduction to corpus annotation. In R. Garside, G., Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora*. pp. 1-18. London: Longman.

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. (2011a), Building an annotated corpus for Amazighe. *Will appear In Proc. of 4th International Conference on Amazigh and ICT*. Rabat, Morocco.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2011b), POS tagging in Amazighe using Support Vector Machines and Conditional Random Fields. *In. Proc. of 16th International Conference on Applications of Natural Language to Information Systems*, NLDB 2011, LNCS(6716), Springer-Verlag, pp. 238-241.

Pinto, D., McCallum, A., Wei, X., Croft. W. B. (2003), Table extraction using conditional random fields. *In SIGIR '03: Proceedings of the 26th annual international*, pp. 235-242, New York, USA.

Ratnaparkhi, A. (1996), A Maximum Entropy Model for Part-Of-Speech Tagging. *In proc. of EMNLP*, Philadelphia, USA.

Schmid, H. (1999), Improvements in Part-of-Speech Tagging with an Application to German. *In Proc. of the ACL SIGDAT-Workshop*. Academic Publishers, Dordrecht, 13-26.

Sha, F. and Pereira F. (2003), Shallow Parsing with Conditional Random Fields. *In Proc. of Human Language Technology*.

Schmid, H. (1994). Part-of-speech tagging with neural networks. *In Proc. of international conference on Computational Linguistics*, Kyoto, Japan.

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

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.