ⴰⵙⵉⵏⴰⴳ-*Asinag,* 9, 2014, p. 91-104

Utilisation des CACs et des ressources externes pour l'amélioration des performances de l'étiquetage morphosyntaxique[1]

Mohamed Outahajala (1), Lahbib Zenkouar (1),
Yassine Benajiba (2), Paolo Rosso (3)

(1)[2] LEC–EMI, Université Mohammed V- Agdal, Maroc
(2) Thomson Reuters, New York, USA
(3)[3] Universidad Politécnica de Valencia, Spain

La langue amazighe, comme la plupart des langues de moindre diffusion, souffre encore de la pénurie d'outils et des ressources pour son traitement automatique, en particulier les corpus annotés. Ces derniers sont plus difficiles à construire que les corpus bruts qui à leur tour nécessitent, dans la majorité des cas, des prétraitements. L'objectif de cet article est de présenter une approche basée sur l'apprentissage semi-supervisé visant l'utilisation d'un corpus de textes bruts, sélectionnés sur la base de la mesure de confiance des Champs Aléatoires Conditionnels(CACs), conjointement avec un corpus annoté manuellement de 20k morphèmes. Les résultats des expérimentations préliminaires montrent une réduction du taux d'erreur de l'étiqueteur morphosyntaxique de 1,3%. De même, la réduction du taux d'erreur est-elle de 5,9%, entre 60% et 90% du corpus, lorsque le modèle est entrainé par les phrases du corpus brut annotées automatiquement.

Amazigh language, and like most of the languages which have only recently started being investigated for the Natural Language Processing (NLP) tasks, lacks annotated corpora and tools and still suffers from the scarcity of linguistic tools and resources and especially annotated corpora. Creating labeled data is a hard task. However, obtaining unlabeled data, although needing most time preprocessing for languages with scarce resources, is less difficult. The aim of

[1] Le premier auteur exprime sa gratitude à la CODESRIA. Les travaux du quatrième auteur ont été financés dans le cadre des projets de recherche: VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, la commission européenne WIQ-EI IRSES (no. 269180) et DIANA-APPLICATIONS (TIN2012-38603-C02-01).

[2] Laboratoire Electronique et Communication, Ecole Mohammadia d'Ingénieurs (EMI).
[3] Natural Language Engineering Lab – EliRF, DSIC.

this paper is to present a semi-supervised based approach using labeled and unlabeled data. Preliminary results show an error reduction of 1,3%, when training our POS tagger with Conditional Random Fields(CRFs), with chosen automatically annotated texts and a small manually annotated corpus of about 20k tokens. Also, when trained with automatically annotated data, the achieved improvement between 60% and 90% of the trained data is 5.9%.

1. Introduction

L'étiquetage morphosyntaxique consiste en l'annotation de chaque mot d'une phrase avec une étiquette récapitulant une information morphosyntaxique selon le contexte. Il augmente l'information des mots étiquetés pour les couches supérieures pour le traitement automatique des langues(TAL). Il s'agit de la première couche au-dessus du niveau lexical et le niveau le plus bas de l'analyse syntaxique. Ainsi, toutes les tâches traitant des niveaux linguistiques supérieurs, utilisent le POS tagging, par exemple : l'analyse partielle ; la désambiguïsation des sens des mots; l'affectation des fonctions grammaticales, la reconnaissance d'entités nommées, etc. (Manning & Schütze, 1999, Cutting et *al*., 1992, Benajiba et *al*., 2010).

Dans la littérature, il a été démontré que les approches basées sur l'apprentissage supervisé 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. Dans l'approche que nous proposons dans ce papier, nous utilisons des techniques de classification des séquences, basées sur les CACs et conjointement des données étiquetées et non étiquetées, pour construire notre étiqueteur grammatical. D'une part, nous utilisons un corpus de ~20k mots annoté manuellement (Outahajala et *al*., 2011a) pour former nos modèles et les caractéristiques n-grammes lexicales pour aider à augmenter la performance ainsi que des ressources externes qui consistent en un ensemble de textes bruts.

Le papier est organisé comme suit : en section 2, nous présenterons les travaux connexes sur les techniques d'étiquetage morphosyntaxique. Puis, dans la section 3 nous donnerons le cadre théorique des CACs. Dans la section 4, nous décrirons les expériences et nous discuterons les résultats. Enfin, dans la section 5, nous dresserons quelques conclusions et nous présenterons 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, Jurafsky & Martin, 2009). Les premiers étiqueteurs morphosyntaxiques étaient principalement à base de

règles. La construction de tels systèmes nécessite 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 ont été appliquées avec succès pour réaliser des POS taggeurs, tels que les Modèles de Markov Cachés (HMM) (Charniak, 1993), la transformation système basé sur la réduction du taux d'erreur (Brill, 1995), le modèle d'entropie maximale (Ratnaparkhi, 1996), les arbres de décision permettent de construire (Schmid, 1999), sur la base d'un corpus de référence, un outil d'aide à la décision qui utilise ce modèle. Les méthodes d'apprentissage automatique permettent de construire des modèles complexes (comportant de très nombreux paramètres), chose qui est 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. Parmi les travaux basés sur l'apprentissage qui ont donné de bon résultats, on cite ceux de Kudo & Matsumoto (2000) et de Lafferty et *al*. (2001).

Bien que ces méthodes aient une bonne performance, la précision des mots inconnus, mots hors vocabulaire du corpus de test par rapport au corpus d'apprentissage, est beaucoup plus faible que celle des mots connus, ce qui est problématique lorsque le corpus d'apprentissage est de petite taille.

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.

En raison de sa morphologie complexe (Chafiq, 1991 ; Ameur et *al*. 2004; Ameur et *al*. 2006; Boukhris et *al*. 2008) ainsi que l'utilisation des différents dialectes dans sa normalisation, la langue amazighe présente des défis intéressants, pour les chercheurs en TAL, qui doivent être pris en compte. Concernant la tâche d'étiquetage morphosyntaxique, certains défis du TAL pour l'amazighe sont les suivants :

1. L'amazighe dispose de sa propre graphie : le Tifinaghe, qui s'écrit de gauche à droite ;
2. Il ne contient pas de majuscules ;
3. Les noms, les noms de qualité, 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 entre deux blancs ou des signes de ponctuation. Toutefois, si une préposition ou un nom de parenté est suivi par un pronom personnel, à la fois la préposition/nom de parenté et le pronom qui suit, forment chaîne unique délimitée par des espaces ou des signes de ponctuation. Par exemple : ⵖ ⵔ (ɣr) signifiant « pour, au » + ⵉ (i) qui signifie « moi » (pronom personnel première personne du singulier) donnent «ⵖ ⴰ ⵔⵉ /ⵖ ⵉ ⵔⵉ (ɣari/ɣuri) » ;

4. Les signes de ponctuation amazighe sont semblables aux signes de ponctuation adoptés au niveau international et ont les mêmes fonctions[4]. Les lettres majuscules, néanmoins, ne se produisent ni au début ni à l'initiale des noms propres ;
5. A l'instar d'autres langues naturelles, l'amazighe peut présenter des ambiguïtés 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, ⵢⵉⵍⵍⵉ (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 « ⵏ » (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.

3. Les Champs Aléatoires conditionnels

Les CACs ou CRFs sont des processus stochastiques qui modélisent les dépendances entre un ensemble d'observations discrètes réalisées sur une séquence discrète et un ensemble d'étiquettes. Dans le cas de l'analyse morphosyntaxique la suite des mots est la séquence discrète. En comparaison avec les Modèles de Markov Cachés, un CAC ne repose pas sur l'hypothèse forte d'indépendance des observations entre elles conditionnellement aux états associés.

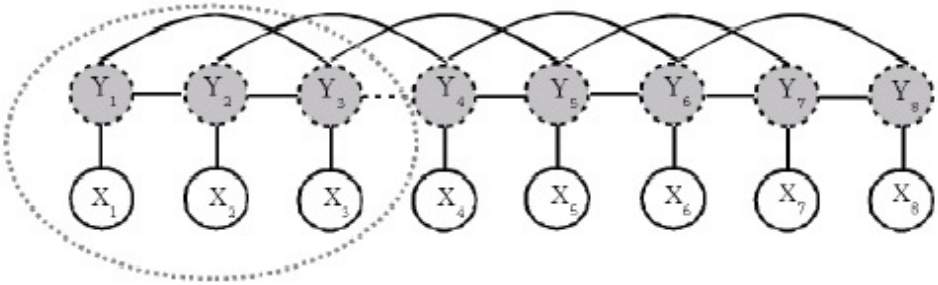

Figure 1 : Exemple d'un graphe des CACs, la partie encerclée est une clique.

Les CACs sont des modèles graphiques probabilistes se basant et sur la théorie des graphes et sur la théorie des probabilités. Ces deux théories permettent de modéliser le problème de classification des séquences : la théorie des graphes

[4] Les deux caractères : ⵯ (2D70) et ⵿ (2D7F) sont deux signes de ponctuation supplémentaires utilisés par les Touaregs. Ils font désormais suite à un amendement du standard Unicode, partie des caractères tifinaghes dont la liste actualisée est sur : http://www.unicode.org/charts/PDF/U2D30.pdf .

permet la modélisation des structures de séquence des étiquettes des phrases, quant à la théorie des probabilités, elle permet de gérer les ambiguïtés causées par les séquences des étiquettes. Les CACs sont avec les Modèles de Markov à Entropie Maximale(MMEMs) les deux principaux modèles discriminants. Bien que les MMEMs aient obtenu de bons résultats sur les tâches d'extraction d'information et de segmentation (MCallum, 2000), ils souffrent du problème du biais du label. En effet, si le graphe est tel qu'un nœud *i* n'a qu'un successeur *i+1*, alors la masse de probabilité est entièrement transmise à y_{i+1} indépendamment des observations x, appelé biais du label. Les CACs permettent de palier à ce problème et cela en calculant les poids de transition non normalisée et en calculant un facteur de normalisation sur l'ensemble de la séquence y conditionnellement à x.

Définition : Soit *G= (V, E)*, où *V* est l'ensemble des sommets et *E* l'ensemble des arcs, un graphe non orienté et soient *X* et *Y* deux champs aléatoires décrivant respectivement l'ensemble des étiquettes, de sorte que pour chaque nœud *i* appartenant à *V*, il existe une variable aléatoire y_i dans *Y*. Nous désignons *(X, Y)* comme étant un champ aléatoire conditionnel si chaque variable aléatoire Y_i respecte la propriété de Markov suivante : $p(Yi|X, Yj\ i \neq j) = p(Yi|X, Yj, i \sim j)$, où $i \sim j$ signifie que *i* et *j* sont voisins dans *G*. La figure 1 présente un exemple d'un graphe de CACs.

Cette propriété n'est par conséquent satisfaite que si chaque variable aléatoire ne dépend que de ses voisins : Y_i ne dépend que de *X* et des Y_j ses voisins dans le graphe d'indépendance.

D'après le théorème de Hammersely-Clifford (Hammersly et *al.*, 1971), la distribution de probabilité *p* d'un champ de Markov est décomposable comme un produit de fonctions φ_c définies sur cliques, sous graphes complets, maximales c de l'ensemble des cliques C de G. Ainsi, la probabilité d'un étiquetage y étant donnée une réalisation d'observations *x* s'écrit :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \varphi_c(y_c, x)$$

Où y_c est la réalisation des variables aléatoires de la clique *c* et *Z(x)* est un coefficient de normalisation défini comme suit :

$$Z(x) = \sum_{y} \prod_{c \in C} \varphi_c(y_c, x)$$

Le coefficient *Z(x)* est un coefficient de normalisation égal au produit des fonctions de potentiel de tous les étiquetages possibles sachant la séquence d'observation *x*.

Lafferty et ses co-auteurs (Lafferty et *al.*, 2001) ont proposé de définir la forme de la fonction φ_c comme l'exponentiel de sommes pondérées des fonctions caractéristiques f_k ayant des poids w_k.

$$\varphi_c(y_c, x, W) = \exp(\sum_{k=1}^{K} w_k f_k(y_c, x))$$

La forme de ces fonctions dépend du domaine d'application. Par exemple, dans le TAL, il s'agit généralement de fonctions binaires qui testent la présence ou l'absence de certaines caractéristiques. Concernant les poids w_k, ils permettent d'accorder plus ou moins d'importance à chacune des fonctions caractéristiques. Ils sont fixés lors de la phase d'apprentissage en cherchant à maximiser la log-vraisemblance sur un ensemble d'exemples déjà annotés formant le corpus de référence. La probabilité d'un étiquetage sachant une réalisation d'observations s'exprime ainsi par :

$$p(y|x) = \frac{1}{Z(x, W)} \exp(\sum_{c \in C} \sum_{k=1}^{K} w_k f_k(y_c, x))$$

Les CACs sont appliquées à de nombreuses tâches du TAL, à titre indicatif l'analyse syntaxique partielle (Sha, Pereira, 2003), l'extraction d'informations à partir des tables (Pinto et *al*., 2003), la reconnaissance d'entités nommées (Li & McCallum, 2003 ; Benajiba et *al*., 2010) et l'étiquetage morphosyntaxique (Outahajala et *al*., 2011b). Les CACs ont été utilisés pour de nombreuses langues pour l'étiquetage morphosyntaxique, tel que l'amharique (Adafre, 2005), le tamoul (Lakshmana & Geetha, 2009), etc.

Dans les expérimentations présentées dans la section suivante, nous avons utilisé l'outil CRF++[5], une implémentation open source des CACs pour la segmentation et l'étiquetage des données.

4. Expérimentations et résultats

Dans cette section, nous présentons une description du corpus brut ainsi que son prétraitement, ensuite les modèles de référence et enfin les expérimentations relatives à l'utilisation de la mesure de confiance, le choix des données aléatoirement pour un apprentissage semi-supèrvisé et l'apprentissage de notre étiqueteur morphosyntxique.

4.1. Description du corpus brut utilisé

Le corpus utilisé dans ces expérimentations a été puisé dans quelques romans amazighes, une partie des données collectées par le Linguistic Data Consortium en collaboration avec l'IRCAM (Cieri et Liberman, 2008), textes brut des sites web de l'IRCAM[6] et de l'Agence Marocaine de Presse[7]ainsi que certaines phrases traduites

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

[6] http://www.ircam.ma/amz/index.php

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

en amazighe de divers sources. Le corpus collecté a subi de multiples prétraitements, à savoir :

- révision des textes collectés selon les règles orthographiques adoptées par l'IRCAM. Aussi, la correction de certaines erreurs fréquentes telles que le mauvais placement du e muet "ⵄ". Dans ce sens, un script écrit en PERL a été réalisé afin de fixer cette erreur. En effet, l'utilisation du e muet s'impose dans les deux cas suivants :

- Succession de plus de trois consonnes radicales identiques à l'intérieur du même mot, par exemple ⵣⵎⵎⵄⵎ (zmmem) "inscrire", ⵜⵄⵜⵜⵄ (tettu) "elle a oublié" ;

- Radicaux verbaux se terminant par deux consonnes identiques, par exemple ⵎⵍⵄⵍ(mlel) "être blanc",

- Pour les textes rédigés en utilisant la police Tifinaghe-IRCAM (Tifinaghe-IRCAM fait usage de glyphes tifinaghes mais caractères latins), afin de corriger certains éléments comme le caractère "^" qui existe dans certains textes dû à une erreur en saisissant les lettres emphatiques ;

- Translittération des textes écrits en Tifinaghe-IRCAM et des textes écrits en utilisant la transcription officielle tifinaghe de la langue amazighe, vers le système d'écriture choisi ;

- Segmentation, en utilisant le segmenteur amazighe réalisé pour cet effet (Outahajala et *al*. 2013) ;

Le nombre total des morphèmes à partir du corpus recueilli est d'environ un quart de million.

4.2. Modèles de références

En ce qui concerne les modèles de références utilisés, nous avons choisi d'adopter deux lignes de base comme références dans ces expériences. En outre, nous avons utilisé le dernier jeu d'étiquettes disponible, composé de 28 étiquettes (Outahajala et *al*., 2013), et les CACs comme modèles de classification des séquences pour les générations des modèles de classification. Un jeu d'étiquettes de taille presque similaire a été utilisé pour l'étiquetage morphosyntaxique de l'arabe (Diab et *al*., 2004). Les modèles de références utilisés comme lignes de base dans les éxpérimentations des sous sections 4.3 et 4.4 sont :

1 – Modèle de référence basé sur la fréquence des mots (Freq-Base.) : il s'agit d'un algorithme basé sur la fréquence des étiquettes des mots. L'étiquette prévue pour un mot est tout simplement l'étiquette la plus fréquente qui a été associée dans les données de formation. Ainsi, cette base ignore totalement le contexte environnant et résout les cas ambigus utilisant uniquement les fréquences des étiquettes. Une telle référence a été utilisée dans la tâche de reconnaissance d'entités nommées

dans CoNLL. Le code source de ce modèle basé sur les fréquences est librement disponible[8].

2 – Modèle de référence du meilleur cas (Best-Base.) : pour étudier le meilleur des cas, on a commencé par la génération d'un modèle initial M_{init} à partir de 60% des données étiquetées. Les 30% des données étiquetées restantes on été subdivisées en blocks de 2k jetons. Ceci, dans le but d'étudier la performance des modèles générés à partir des données annotées automatiquement. Le choix des données pour la génération de M_{init} n'est pas aléatoire. En effet, nous avons effectué la validation croisée de 60% du corpus et nous avons pris le modèle qui a donné la meilleure précision.

Le choix de l'ensemble des caractéristiques a été obtenu suite à des résultats empiriques. Ils sont les mêmes que ceux employés dans (Outahajala et *al*., 2012) à savoir :

1. Le jeton actuel ;
2. Les propriétés lexicales n-grammes : consistant en les i premiers et dernier n-grammes du jeton, avec i variant de 1 à 4. Les caractéristiques n-grammes servent comme caractéristiques représentant les suffixes et les préfixes des jetons ;
3. Le contexte lexical : s'agissant des jetons voisinant plus leurs propriétés n-grammes définies dans le point 2 ci-dessus ;
4. Etiquettes de contexte qui consistent en les balises prévues pour les deux mots précédents.

4.3. Expérimentation : choix des données selon la mesure de confiance

Le but de ces expérimentations préliminaires est d'évaluer le critère de confiance dans la sélection des phrases pour l'auto-apprentissage de notre modèle. Nous avons part de l'hypothèse que notre modèle apprend plus quand la confiance est élevée. Pour évaluer notre approche, nous commenceons par un modèle initial M_{init} entraîné sur la base de Tr_1 (voir Figure 2), contenant l'équivalent de 60% du corpus de référence.

Pour ce faire, nous étudierons la corrélation entre la mesure de confiance et la probabilité d'obtenir un étiquetage correct. C'est l'estimation des chances d'assigner une étiquette correcte à un mot automatiquement quand la probabilité de l'étiquette affectée au mot par le système est élevée. Nous pensons que cette estimation est importante car lorsque la corrélation observée tend vers 1, la probabilité des données sélectionnées tend à améliorer le système et, lorsque cette probabilité tend vers 0,5, l'amélioration est aléatoire. D'un point de vue de filtrage du bruit, on peut dire que dans le cas d'absence de corrélation entre les deux termes en question, il n'est pas possible de filtrer le bruit en se basant sur la mesure de confiance générée par le système.

[8] http://www.outamed.com/downloads/baseline.txt

Afin d'obtenir l'information requise, nous avons automatiquement annoté 10% du corpus de test (nous n'avons intentionnellement pas utilisé le corpus de test lors du du calcul de la corrélation) utilisant M_{init}. Les étiquettes obtenues ont servi comme données de base dans le calcul de la corrélation.

La corrélation entre la mesure de confiance et la probabilité d'avoir un étiquetage correct est 0,78. On a ainsi une nette régression positive.

Pour étudier l'utilité de la mesure de confiance du système pour les mots dans la sélection des données, nous avons effectué des expérimentations utilisant M_{init} et les données brutes présentées dans la sous-section 4.1. Les données non étiquetées ont été annotées automatiquement et nous avons gardé les meilleures : 1295 phrases, soit l'équivalent de 90% des données annotées manuellement, selon la mesure de confiance.

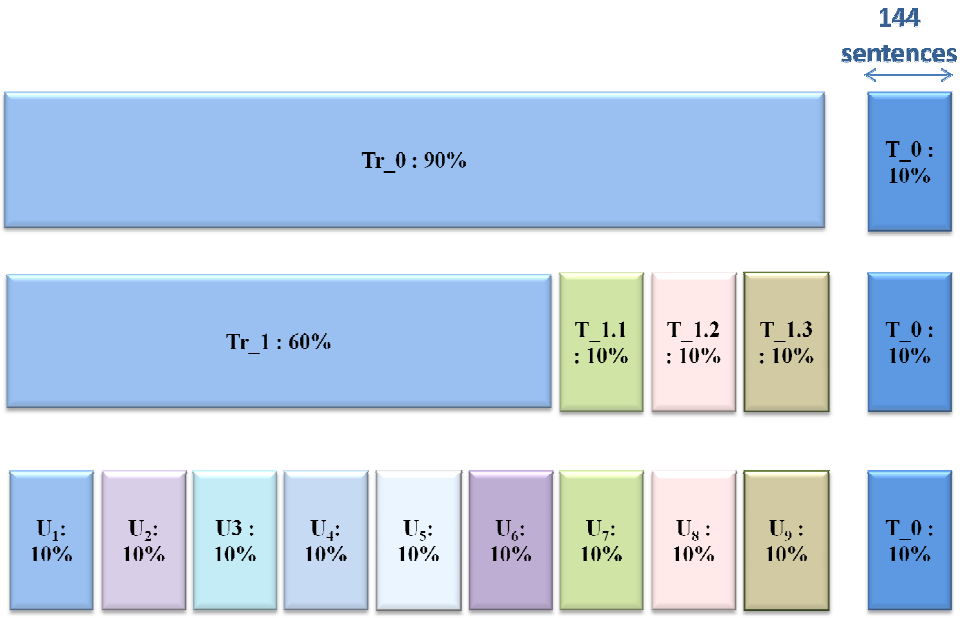

Figure 2 : Subdivision des données pour les expérimentations sur l'apprentissage semi-supervisé

Dans cette expérimentation, le critère de sélection est basé sur la mesure de la confiance donnée par le système. Après, ce corpus a été subdivisé en 9 parties U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, et U_9, où chacune des parties U_i contient 144 phrases avec i variant de 1 à 9 (l'équivalent de 10% du nombre total des phrases du corpus annoté manuellement). La subdivision du corpus est présentée dans la figure 2.

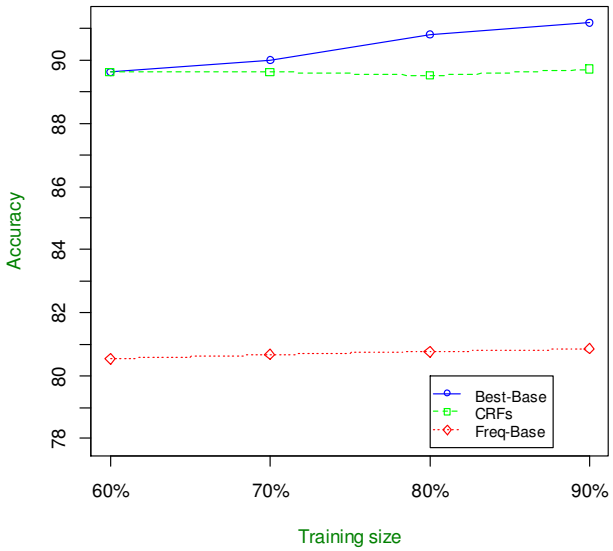

Figure 3 : Apprentissage du modèle en utilisant la mesure de confiance du mot comme moyen de sélection

Nous avons remarqué qu'au fur et à mesure que la performance augmente, elle devient difficile à améliorer. Néanmoins, la différence d'amélioration ne diminue pas de façon régulière, elle fluctue légèrement. Par exemple, le taux d'amélioration entre 70 et 80% (0.81) est supérieur au taux d'amélioration entre 60 et 70% (0.66) lorsqu'on fait l'entraînement des modèles à l'aide des données annotées manuellement. A l'analyse des fichiers en sortie de l'étiqueteur, il s'avère que les mots hors vocabulaire constituent un facteur important dans l'amélioration de la précision de l'étiqueteur. Aussi, la performance des modèles basés sur les CACs est-elle nettement supérieure à celle du modèle à base des fréquences.

Pour ce qui est des résultats du modèle utilisant et les données du corpus de référence et les données brutes, l'amélioration est légère. Les résultats de l'expérimentation montrent qu'il y a une réduction du taux d'erreur de 1,3% (voir figure 2).

4.4. Expérimentation : choix aléatoire des données pour l'apprentissage

Pour étudier l'effet d'ignorer la confiance et voir si ce critère est important ou non, nous avons conduit une expérimentation où nous commenceons par M_{init} et à chaque itération de l'opération d'apprentissage nous ajoutons 144 phrases de U annotées automatiquement par M_{init} et choisies aléatoirement.

Tels que montrés dans la figure 4, les résultats de l'apprentissage à partir des données choisies aléatoirement sont moins précis que ceux qui se basent sur la sélection des données en utilisant la mesure de la confiance. Ceci confirme l'utilité

de cette mesure dans la sélection des phrases dans l'auto-apprentissage de notre étiqueteur morphosyntaxique.

Dans la figure 4, CRFs-R représente le modèle généré à partir des données sélectionnées aléatoirement et CRF-BS le modèle généré en utilisant la mesure de confiance du mot comme moyen de sélection des données pour l'apprentissage.

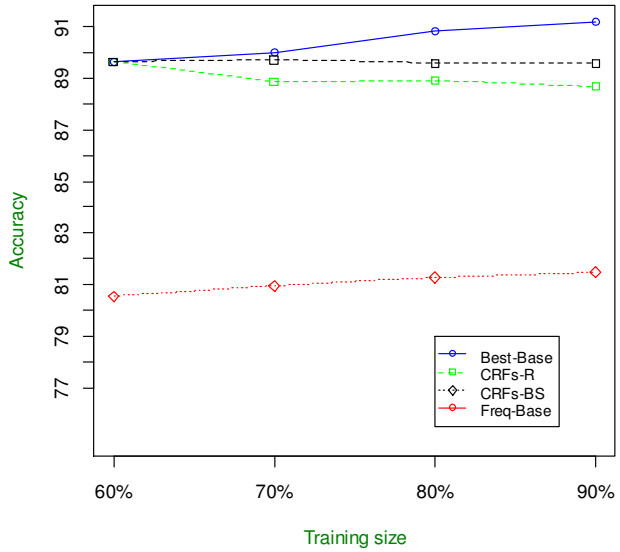

Figure 4 : Apprentissage à partir de données sélectionnées aléatoirement

Afin de vérifier l'hypothèse que le bruit de l'auto-apprentissage n'empêche pas la réduction du taux d'erreur lors de l'entraînement de notre modèle, nous avons conduit l'expérimentation suivante :

- génération de M_{init} à partir des parties U_1, U_2,...U_6 constituant 60% de la taille du corpus de référence ;

- ajout à chaque itération de l'apprentissage de 144 phrases au corpus d'apprentissage jusqu'à ce que le corpus d'apprentissage atteigne l'équivalent de 90% du corpus de référence.

Les résultats de l'expérimentation montrent qu'il y a une réduction du taux d'erreur de 5,9% entre M_{init} et le modèle appris en utilisant U_1, U_2,…U_9. Ce qui montre que, même si le bruit existe, le système continue d'apprendre.

5. Conclusions

La langue amazighe, comme la plupart des langues de moindre diffusion, souffre encore de la pénurie d'outils et des ressources pour son traitement automatique, en particulier les corpus annotés. Dans ce papier, nous avons présenté les expérimentations préliminaires d'utilisation de ressources externes, consistant en un corpus de textes bruts de 225.240 morphèmes et d'un corpus manuellement

annoté d'environ 20k morphèmes et leur impact sur la performance de la tâche d'étiquetage morphosyntaxique de la langue amazighe.

Les résultats des expérimentations montrent une réduction du taux d'erreur de 1,3%. Aussi la réduction du taux d'erreur est-elle de 5,9%, lorsque le modèle est complètement entraîné par les phrases du corpus brut annotées automatiquement.

Dans le futur proche, nous étudierons l'impact de l'utilisation du caractère informatif des MHVs et la mesure de confiance lors de l'utilisation des méthodes d'apprentissage semi-supervisé sur l'amélioration de la performance de l'étiqueteur morphosyntaxique.

Références

Ameur, M., Bouhjar, A., Boukhris, F., Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi, E. et Souifi, H. (2004), *Initiation à la langue Amazighe*, Rabat, Publications de l'IRCAM.

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

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., and Rosso P. (2010). Arabic Named Entity Recognition: A Feature-Driven Study. *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.

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

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics.In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park.AAAI Press/MIT Press.

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

Cieri, C., and Liberman, M. (2008). 15 Years of Language Resource Creation and Sharing. A Progress Report on LDC Activities. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC'08, Marrakech.

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

Jurafsky, D., and Martin, J.H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd Ed. New Jersey: Prentice Hall.

Hammersley, J.M. et Clifford, P. (1971). Markov Fields on finite graphs and lattices. *Manuscrit non publié*.

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

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

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

Li W., and McCallum, A. (2003). Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction. In *ACM Transactions on Computational Logic*, pp 290--294.

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

McCallum, A., Freitag, D. and Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. *International Conference on Machine Learning*, pages 591—598.

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

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

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

Outahajala, M., Benajiba, Y., Rosso, P. et Zenkouar, L. (2012), « L'étiquetage grammatical de l'amazighe en utilisant les propriétés n-grammes et un prétraitement de segmentation », *e-TI - la revue électronique des technologies d'information*, N° 6.

Outahajala, M., Zenkouar, L, Benajiba, Y., and Rosso, P. (2013). The Development of a Fine Grained Class Set for Amazigh POS Tagging. In *proceedings of ACS/IEEE 10th conference*. AICCSA 2013. Fes, Morocco.

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

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

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

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