TALN 2010, Montréal, 19–23 juillet 2010

Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français

Pascal Denis & Benoît Sagot
Alpage, INRIA Paris–Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{pascal.denis,benoit.sagot}@inria.fr

Résumé. Cet article présente MElt_{fr}, un étiqueteur morpho-syntaxique automatique du français. Il repose sur un modèle probabiliste séquentiel qui bénéficie d'informations issues d'un lexique exogène, à savoir le Le*fff*. Evalué sur le FTB, MElt_{fr} atteint un taux de précision de 97.75% (91.36% sur les mots inconnus) sur un jeu de 29 étiquettes. Ceci correspond à une diminution du taux d'erreur de 18% (36.1% sur les mots inconnus) par rapport au même modèle sans couplage avec le Le*fff*. Nous étudions plus en détail la contribution de cette ressource, au travers de deux séries d'expériences. Celles-ci font apparaître en particulier que la contribution des traits issus du Le*fff* est de permettre une meilleure couverture, ainsi qu'une modélisation plus fine du contexte droit des mots.

Abstract. This paper presents MElt_{fr}, an automatic POS tagger for French. This system relies on a sequential probabilistic model that exploits information extracted from an external lexicon, namely Le*fff*. When evaluated on the FTB corpus, MElt_{fr} achieves an accuracy of 97.75% (91.36% on unknow words) using a tagset of 29 categories. This corresponds to an error rate decrease of 18% (36.1% on unknow words) compared to the same model without Le*fff* information. We investigate in more detail the contribution of this resource through two sets of experiments. These reveal in particular that the Le*fff* features allow for an increased coverage and a finer-grained modeling of the context at the right of a word.

Mots-clés : Etiquetage morpho-syntaxique, modèles à maximisation d'entropie, français, lexique.

Keywords: POS tagging, maximum entropy models, French, lexicon.

1 Introduction

De nombreux systèmes pour l'étiquetage automatique en parties du discours ont été développés pour un large éventail de langues. Parmi les systèmes les plus performants, on trouve ceux qui s'appuient sur des techniques d'apprentissage automatique[1]. Pour certaines langues comme l'anglais ou d'autres langues très étudiées, ces systèmes ont atteint des niveaux de performance proches des niveaux humains. Il est intéressant de constater que la majorité de ces systèmes n'ont pas recours à une source externe d'informations lexicales, et se contentent d'un lexique extrait « à la volée » à partir du corpus d'apprentissage (cf. cependant (Hajič, 2000)). On est donc en droit de se demander s'il est possible d'améliorer encore

[1] Se reporter à (Manning & Schütze, 1999) pour un panorama complet.

les performances des étiqueteurs en exploitant ce type de ressources. Un avantage potentiel de l'utilisation d'un lexique externe consiste en un meilleur traitement des mots « inconnus », c'est-à-dire des mots absents du corpus d'apprentissage, dès lors qu'ils sont présents dans le lexique externe.

Dans (Denis & Sagot, 2009), nous avons montré qu'un modèle d'étiquetage peut bénéficier à la fois d'informations provenant d'un corpus d'entraînement et d'un lexique exogène à large couverture. Nous avons pour cela utilisé des modèles markoviens à maximisation d'entropie, à savoir une classe de modèles discriminants adaptés aux problèmes séquentiels et par ailleurs très rapides à entraîner. Les expériences menées en couplant ainsi le Corpus Arboré de Paris 7, dorénavant FTB pour *French TreeBank* (Abeillé *et al.*, 2003), et le lexique Le*fff* (Sagot, 2010) ont ainsi conduit à un étiqueteur de niveau état de l'art pour le français, nommé MElt_{fr} et distribué librement[2], qui a une précision de $97,7\%$ sur le jeu de test. (Denis & Sagot, 2009) montre par ailleurs que l'utilisation d'informations extraites du Le*fff* permet, pour des taux de précision similaires, de réduire le volume du corpus d'entraînement par un facteur de 2 à 3. Indépendamment, un travail comparable, mais qui intègre en plus dans le modèle les informations de lemmatisation, a également montré la pertinence de ce type d'approches (Chrupała *et al.*, 2008).

Toutefois, les causes précises de l'amélioration des performances lorsque les informations du Le*fff* sont exploitées n'avaient pas encore été explorées de façon systématique. On s'attend naturellement à ce que de telles informations supplémentaires améliorent l'étiquetage des mots inconnus. Mais les informations extraites du Le*fff* sont-elles les plus utiles à propos du mot courant ou du contexte gauche ou droit ? Sont-elles plus cruciales pour étiqueter les mots appartenant à des classes fermées ou ouvertes ?

Pour apporter des éléments de réponse à ces questions, nous avons conduit plusieurs séries d'expériences. Nous commençons par décrire les ressources utilisées (section 2). Nous détaillons ensuite le fonctionnement et les performances de l'étiqueteur MElt_{fr}, amélioré depuis sa présentation dans (Denis & Sagot, 2009), et nous le comparons avec d'autres étiqueteurs entraînés sur les mêmes données (section 3). Afin de mieux comprendre la façon dont le couplage avec le Le*fff* améliore les performances, nous présentons alors des expériences faisant varier la façon dont les informations extraites du Le*fff* sont exploitées (section 4), puis des expériences faisant varier les jeux d'étiquettes du corpus et du lexique (section 5).

2 Ressources utilisées

Le corpus annoté en parties du discours que nous avons utilisé est une variante du FTB. Il diffère du FTB originel en ceci que tous les composés qui ne correspondent pas à une séquence régulière de parties du discours sont fusionnés en un token unique, alors que les autres sont représentés par des séquences de plusieurs tokens (Candito, c.p.). Le corpus résultat contient $350\,931$ tokens pour $12\,351$ phrases. Dans le FTB originel, les mots sont répartis en 13 catégories principales, elles-mêmes réparties en 34 sous-catégories. La version du corpus que nous avons utilisée a été obtenue en convertissant ces sous-catégories en un jeu de 29 parties du discours, avec une granularité intermédiaire entre catégories et sous-catégories. Ces 29 étiquettes améliorent les catégories principales par des informations sur le mode des verbes, ainsi que par quelques traits lexicaux supplémentaires. Ce jeu d'étiquettes est celui qui conduit aux meilleurs résultats d'analyse syntaxique probabiliste pour le français (Crabbé & Candito, 2008)[3]. Comme dans (Crabbé & Candito, 2008), le FTB est divisé en 3 sections : entraînement (80%), développement (10%) et test (10%).

[2] `http://gforge.inria.fr/projects/lingwb/`.

[3] Ce jeu d'étiquettes est nommé TREEBANK+ dans (Crabbé & Candito, 2008).

Les tailles respectives de ces sections sont détaillées à la table 1, ainsi que les nombres et proportions de tokens inconnus.[4]

Section	**# de phrases**	**# de tokens**	**# de tokens inconnus**	**# de tokens inconnus et absents du Lefff**
FTB-TRAIN	9 881	278 083		
FTB-DEV	1 235	36 508	1 790 (4,9%)	604 (1,7%)
FTB-TEST	1 235	36 340	1 701 (4,7%)	588 (1,6%)

TAB. 1 – Jeux de données

La source d'informations lexicales que nous avons utilisée est le lexique Le*fff* (Sagot, 2010)[5]. Nous avons extrait du Le*fff* 502 223 entrées distinctes de la forme *(forme, étiquette)*, les étiquettes correspondant après conversion au jeu de 29 étiquettes de la variante du FTB décrite ci-dessus et utilisée pour l'apprentissage.

3 Étiquetage avec un modèle à maximisation d'entropie

Dans cette section, nous décrivons le modèle markovien à maximisation d'entropie (MMME) sur lequel repose l'étiqueteur MElt_{fr}. Nous présentons d'abord la variante $\text{MElt}_{\text{fr}}^{\text{nolex}}$, qui n'exploite pas les informations lexicales du Le*fff*. Il est comparable aux systèmes de Ratnaparkhi (1996) et Toutanova & Manning (2000), à la fois quant au modèle et quant aux traits utilisés. Aujourd'hui, les étiqueteurs reposant sur ce type de modèles sont parmi les meilleurs pour l'anglais.[6] Un avantage important de ces modèles (sur les modèles de Markov cachés, notamment) est de permettre de combiner ensemble des traits très divers, éventuellement redondants, sans qu'il soit nécessaire de faire une hypothèse d'indépendance entre eux. C'est ce qui nous a permis de construire MElt_{fr} en rajoutant à $\text{MElt}_{\text{fr}}^{\text{nolex}}$ des traits lexicaux extraits du Le*fff*. Enfin, ces modèles sont également attrayants du fait qu'ils sont très rapides à entraîner[7].

3.1 Le modèle de base

Étant donné un jeu d'étiquettes T et une chaîne de mots (tokens) w_1^n, on définit la tâche d'étiquetage comme le processus consistant à assigner à w_1^n la séquence d'étiquettes $\hat{t}_1^n \in T^n$ de vraisemblance maximale. Suivant (Ratnaparkhi, 1996), on peut alors approcher la probabilité conditionnelle $P(t_1^n|w_1^n)$ de sorte que :

$$\hat{t}_1^n = \arg\max_{t_1^n \in T^n} P(t_1^n|w_1^n) \approx \arg\max_{t_1^n \in T^n} \prod_{i=1}^{n} P(t_i|h_i), \tag{1}$$

où t_i est l'étiquette du mot w_i et h_i est le *contexte* de (w_i, t_i), qui comprend la séquence des étiquettes déjà assignées t_1^{i-1} et la séquence des mots w_1^n.

[4] Il s'agit des tokens inconnus à la fois sous leur forme d'origine *et* sous leur forme minusculisée.

[5] Le Le*fff* est librement distribué sous licence LGPL-LR à l'adresse `http://alexina.gforge.inria.fr/`.

[6] (Ratnaparkhi, 1996) et (Toutanova & Manning, 2000) obtiennent des précisions respectives de 96.43 et 96.86 sur les sections 23 et 24 du Penn Treebank.

[7] Les Champs Aléatoires Conditionnels (Conditional Random Fields, CRF) sont souvent considérés comme plus adaptés aux problèmes de prédiction séquentielle et structurée (Lafferty *et al.*, 2001), mais ils sont aussi nettement plus lents en temps d'entraînement.

Dans un modèle à maximisation d'entropie, on estime les paramètres d'un modèle exponentiel qui a la forme suivante :

$$P(t_i|h_i) = \frac{1}{Z(h)} \cdot \exp\left(\sum_{i=j}^{m} \lambda_j f_j(h_i, t_i)\right) \quad (2)$$

Les f_1^m sont des traits, fonctions définies sur l'ensemble des étiquettes t_i et des historiques h_i (avec $f(h_i, t_i) \in \{0, 1\}$), les λ_1^m sont les paramètres associés aux f_1^m, et $Z(h)$ est un facteur de normalisation sur les différentes étiquettes. Dans ce type de modèle, le choix des paramètres est assujetti à des contraintes garantissant que l'espérance de chaque trait soit égale à son espérance empirique telle que mesurée sur le corpus d'apprentissage (Berger *et al.*, 1996). Dans nos expériences, les paramètres ont été estimés en utilisant l'algorithme dit *Limited Memory Variable Metric Algorithm* (Malouf, 2002) implémenté au sein du système Megam[8] (Daumé III, 2004).

Les classes de traits que nous avons utilisées pour la conception du modèle de base $\text{MElt}_{\text{fr}}^{\text{nolex}}$ d'étiquetage du français, c'est-à-dire du modèle n'utilisant pas le lexique L*efff*, est un sur-ensemble des traits utilisé par (Ratnaparkhi, 1996) et (Toutanova & Manning, 2000) pour l'anglais (qui étaient largement indépendants de la langue). Ces traits peuvent être regroupés en deux sous-ensembles. Le premier rassemble des traits dits *internes* qui essaient de capturer les caractéristiques du mot à étiqueter. Il s'agit notamment du mot w_i lui-même, de ses préfixes et suffixes de longueur 1 à 4, ainsi que de traits booléens qui testent si w_i contient ou non certains caractères particuliers comme les chiffres, le tiret ou les majuscules. Le deuxième ensemble de traits, dits *externes*, modélise le contexte du mot à étiqueter. Il s'agit tout d'abord des mots qui sont dans les contextes gauche et droit de w_i (à une distance d'au plus 2). Ensuite, nous intégrons comme traits l'étiquette t_{i-1} assignée au mot précédent, ainsi que la concaténation des étiquettes t_{i-1} et t_{i-2} pour les deux mots précédents w_i. La liste détaillée des classes de traits utilisées dans $\text{MElt}_{\text{fr}}^{\text{nolex}}$ est indiquée à la table 2.

Traits internes	
$w_i = X$	& $t_i = T$
Préfixe de $w_i = P, \lvert P\rvert < 5$	& $t_i = T$
Suffixe de $w_i = S, \lvert S\rvert < 5$	& $t_i = T$
w_i contient un nombre	& $t_i = T$
w_i contient un tiret	& $t_i = T$
w_i contient une majuscule	& $t_i = T$
w_i contient uniquement des majuscules	& $t_i = T$
w_i contient une majuscule et n'est pas le premier mot d'une phrase	& $t_i = T$
Traits externes	
$t_{i-1} = X$	& $t_i = T$
$t_{i-2}t_{i-1} = XY$	& $t_i = T$
$w_{i+j} = X, j \in \{-2, -1, 1, 2\}$	& $t_i = T$

TAB. 2 – Traits de base utilisés par $\text{MElt}_{\text{fr}}^{\text{nolex}}$

Une différence importante avec le jeu de traits de (Ratnaparkhi, 1996) vient du fait que nous n'avons pas restreint l'application des traits de type préfixes et suffixes aux mots qui sont rares dans le corpus d'apprentissage. Dans notre modèle, ces traits sont toujours construits, même pour les mots fréquents. En effet, nous avons constaté lors du développement que la prise en compte systématique de ces traits conduit à de meilleurs résultats, notamment sur les mots inconnus. De plus, ces traits sont probablement

[8]Librement disponible à l'adresse `http://www.cs.utah.edu/~hal/megam/`.

plus discriminants sur le français que sur l'anglais, puisque le français est morphologiquement plus riche. Une autre différence entre notre modèle de base et les travaux antérieurs concerne le lissage. (Ratnaparkhi, 1996) et (Toutanova & Manning, 2000) seuillent leurs traits à un nombre d'occurrence de 10 pour éviter les données statistiquement non significatives. Nous n'avons pas seuillé nos traits mais avons utilisé à la place une régularisation gaussienne sur les poids, ce qui est une technique de lissage plus motivée statistiquement.

3.2 Intégration des informations lexicales dans l'étiqueteur

L'avantage du modèle sous-jacent à $\text{MElt}_{\text{fr}}^{\text{nolex}}$ est de permettre un ajout aisé de traits supplémentaires, y compris de traits dont les valeurs sont calculées à partir d'une ressource externe au corpus d'apprentissage, et notamment d'un lexique comme le Le*fff*.

Pour chaque mot w_i, nous générons une nouvelle série de traits internes basés sur la présence (ou non) de w_i dans le Le*fff* et, le cas échéant, les étiquettes associées à w_i par le Le*fff*. Si w_i est associé à une étiquette unique t_i, nous générons un trait qui encode l'association non ambiguë entre w_i et $t_{i,0}$. Lorsque w_i est associé à plusieurs étiquettes $t_{i,0}, \ldots, t_{i,m}$ par le Le*fff*, nous générons un trait interne pour chacune de ses étiquettes possibles $t_{i,j}$, ainsi qu'un trait interne qui représente la disjonction de m étiquettes. Enfin, si w_i n'est pas recensé dans le Le*fff*, nous créons un trait spécifique qui encode le statut d'inconnu du Le*fff*.[9]

De même, nous utilisons le Le*ffff* pour construire de nouveaux traits externes : nous construisons l'équivalent des traits internes pour les mots des contextes gauche et droit à une distance de moins de 2 du mot courant. Nous générons également des traits bigrammes correspondant à la concaténation des étiquettes du Le*fff* pour les 2 mots à gauche, les 2 mots à droite, et les deux mots qui entourent w_i. Lorsque ces mots sont ambigus pour le Le*fff*, seule leur disjonction contribue au bigramme, et si l'un ce ces mots est inconnu, la valeur *unk* tient lieu d'étiquette.[10]

La liste détaillée des classes de traits utilisées pour MElt_{fr}, en plus de ceux de la table 2, est indiquée à la table 3. Ce jeu de traits étend légèrement celui présenté par (Denis & Sagot, 2009).

Traits lexicaux internes	
$t_i =_{uni} X$, if $\mathsf{lefff}(w_i) = \{X\}$	& $t_i = T$
$t_i = X, \forall X \in \mathsf{lefff}(w_i)$ if $\lvert\mathsf{lefff}(w_i)\rvert > 1$	& $t_i = T$
$t_i = \bigvee \mathsf{lefff}(w_i)$ if $\lvert\mathsf{lefff}(w_i)\rvert > 1$	& $t_i = T$
$t_i = unk$, if $\mathsf{lefff}(w_i) = \emptyset$	& $t_i = T$
Traits lexicaux externes	
$t_{i+j} = \bigvee \mathsf{lefff}(w_{i+1}), j \in \{-2, -1, 1, 2\}$	& $t_i = T$
$t_{i+j}t_{i+k} = \bigvee \mathsf{lefff}(w_{i+j}) \bigvee \mathsf{lefff}(w_{i+k}), (j,k) \in \{(-2,-1),(+1,+2),(-1,+1)\}$	& $t_i = T$

TAB. 3 – Traits lexicaux ajoutés au modèle de base dans MElt_{fr}

Ces différents traits permettent d'avoir une information, ne serait-ce qu'ambiguë, sur les étiquettes dans le contexte droit du mot, ce que ne permettent pas les traits de base utilisés par $\text{MElt}_{\text{fr}}^{\text{nolex}}$. Ceux-ci n'incluent

[9] Pour les mots qui apparaissent en position initiale dans la phrase, nous vérifions au prélalable que la version décapitalisée du mot n'est pas présente non plus dans le Le*fff*.

[10] Différentes valeurs de « fenêtre » ont été essayées lors de la phase de développement : 1, 2 et 3. Bien que le passage de 1 à 2 mène à une amélioration significative, le passage de 2 à 3 mène à une légère dégradation.

que les étiquettes sur le contexte gauche, les seules à pouvoir être intégrées dans un décodage gauche-droite. Par ailleurs, cette manière d'intégrer les informations issues du lexique au modèle sous forme de traits supplémentaires a l'avantage de ne pas ajouter de contraintes *fortes*, et d'être ainsi robuste à d'éventuelles erreurs ou incomplétudes du lexique.

Une autre façon, plus directe, d'exploiter une ressource lexicale exogène consiste en effet à utiliser les informations lexicales comme *filtre*. A savoir, on contraint l'étiqueteur à choisir pour un mot w une étiquette correspondant soit à une occurrence de w dans le corpus, soit à une entrée du lexique pour w. C'est l'approche employée par exemple par (Hajič, 2000) pour des langues à morphologie très riche, et notamment pour le tchèque. Dans (Denis & Sagot, 2009), nous avons montré que cette stratégie ne permet d'améliorer que marginalement les performances de $\text{MElt}_{\text{fr}}^{\text{nolex}}$, et restent largement en-deçà de celles de MElt_{fr}.

3.3 Décodage

La procédure de décodage (c'est-à-dire l'étiquetage proprement dit une fois le modèle construit) repose sur un algorithme de type *beam search* pour trouver la séquence d'étiquettes la plus probable pour une phrase donnée. Autrement dit, chaque phrase est décodée de gauche à droite, et l'on conserve pour chaque mot w_i les n séquences d'étiquettes candidates les plus probables du début de la phrase jusqu'à la position i. Pour nos expériences, nous avons utilisé un *beam* de taille 3[11]. De plus, la procédure de test utilise un *dictionnaire d'étiquettes*) qui liste pour chaque mot les étiquettes qui lui sont associées dans le corpus d'apprentissage. Ceci réduit considérablement l'ensemble des étiquettes parmi lesquelles l'étiqueteur peut choisir pour étiqueter un mot donné, ce qui conduit, comme le montrent nos expériences, à de meilleures performances tant en termes de précision que d'efficacité en temps.

3.4 Comparaisons avec d'autres systèmes

Nous avons comparé les résultats de $\text{MElt}_{\text{fr}}^{\text{nolex}}$ et de MElt_{fr} à divers autres étiqueteurs, dont les deux premiers n'utilisent pas le Le*fff*, mais qui ont tous été (ré)entraînés d'une façon ou d'une autre sur le corpus d'apprentissage du FTB :

- UNIGRAM, un étiqueteur de base qui fonctionne comme suit : pour un mot présent dans le corpus d'entraînement, l'étiqueteur assigne l'étiquette la plus fréquemment trouvée dans le corpus ; pour les autres mots, il utilise l'étiquette la plus fréquente du corpus (ici, NC) ;
- TreeTagger, un étiqueteur statistique[12] qui repose sur les arbres de décision (Schmid, 1994) réentraîné sur notre corpus d'apprentissage.
- $\text{UNIGRAM}_{\text{Le}\mathit{fff}}$, comme UNIGRAM, est un modèle unigramme qui repose sur le corpus d'apprentissage, mais qui utilise le Le*fff* pour étiqueter les mots inconnus : parmi les étiquettes que le Le*fff* associe à un mot inconnu du corpus, l'étiquette la plus fréquente à l'échelle de tout le corpus est utilisée ; les mots qui sont inconnus et du corpus et du Le*fff* reçoivent l'étiquette la plus fréquente (ici, NC) ;
- $\text{TreeTagger}_{\text{Le}\mathit{fff}}$ est une variante de TreeTagger, le Le*fff* étant fourni comme lexique externe ;
- F-BKY, une instance de l'analyseur syntaxique de Berkeley tel qu'adapté au français par Crabbé & Candito (2008), et utilisée comme étiqueteur.

[11] Nous avons essayé d'autres valeurs (5, 10, 20) pendant le développement, mais ces valeurs n'ont pas conduit à des variations significatives de performance.

[12] Disponible sur `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

Les résultats de cette compraison sur le corpus de test du FTB font l'objet du tableau 4.

Étiqueteur	**Précision globale (%)**	**Précision sur les mots inconnus (%)**
UNIGRAM	91,90	24,50
TreeTagger	96,14	75,77
UNIGRAM$_{\text{Le}fff}$	93,40	55,00
TreeTagger$_{\text{Le}fff}$	96,55	82,14
F-BKY	97,25	82,90
MElt$_{\text{fr}}^{\text{nolex}}$	97,25	86,47
MElt$_{\text{fr}}$	**97,75**	**91,36**

TAB. 4 – Comparaison des performances de divers étiqueteurs en partie du discours pour le français

Parmi les étiqueteurs ne faisant pas usage du Le*fff*, on constate que MElt$_{\text{fr}}^{\text{nolex}}$ atteint déjà une précision de $97,25\%$, avec 86.47% sur les mots inconnus. Ceci est significativement meilleur que TreeTagger, avec un gain de plus de 10% sur les mots inconnus[13]. On peut avancer plusieurs hypothèses pour expliquer des écarts si importants sur l'étiquetage des mots inconnus. Tout d'abord, l'estimation des paramètres dans un modèle à maximisation d'entropie est moins sujette au problème du manque de données pour certains traits ou certaines valeurs de traits que d'autres approches comme les arbres de décision (utilisés par TreeTagger), notamment parce qu'aucune partition des données d'entraînement n'est effectuée. Par ailleurs, TreeTagger n'est pas en mesure de faire autant de généralisations que MElt$_{\text{fr}}^{\text{nolex}}$ sur les traits internes, puisqu'il ne prend en compte que les suffixes, et ce, uniquement sur les mots inconnus.

Parmi les étiqueteurs faisant usage du Le*fff*, le meilleur d'entre eux est MElt$_{\text{fr}}$, avec une exactitude de $97,75\%$ globalement et $91,36\%$ sur les mots inconnus. Ces deux résultats constituent des améliorations significatives de $0,5\%$ et $4,89\%$ par rapport au modèle sans Le*fff*. Ces scores sont meilleurs que ceux de tous les étiqueteurs que nous avons pu tester, y compris l'analyseur F-BKY qui exploite des résultats d'analyse syntaxique probabiliste, et ce, avec un écart significatif[14].

D'autres étiqueteurs ont été proposés pour le français, notamment lors de la campagne d'évaluation GRACE[15]. Bien qu'une comparaison directe soit difficile, étant donné la différence de corpus d'évaluation et de jeux d'étiquettes, notons que le meilleurs résultats reportés lors de cette campagne sont de 96% (Adda *et al.*, 1999) et qu'ils ont été obtenus par des analyseurs syntaxiques. Notons par ailleurs que (Nasr & Volanschi, 2004) reporte des scores de 97.82 sur le corpus de Paris 7, mais leur étiqueteur/chunker ne prend pas en compte les mots inconnus.

Une analyse détaillée des causes d'erreurs de MElt$_{\text{fr}}$ (Denis & Sagot, 2009) peut être résumée ainsi : $43,5\%$ des erreurs sont des erreurs classiques (dont 4% d'erreurs sur *de*, *du* et *des*, et $5,5\%$ de confusions entre adjectifs et participes passés), $15,5\%$ des erreurs concernent des nombres, $27,5\%$ des erreurs sont liées à des entités nommées, et $13,5\%$ des erreurs n'en sont pas vraiment, soit que le corpus de référence contient lui-même une erreur (9% des cas), soit que l'étiquette de référence et l'étiquette proposée par MElt$_{\text{fr}}$ semblent toutes deux correctes ($4,5\%$ des cas).

[13]Des tests de significativité statistique (tests du χ^2) ont été appliqués aux écarts de précision, avec un paramètre p à $0,01$.

[14]Une adaptation au français de Morfette (Chrupała *et al.*, 2008) utilisant le FTB et le Le*fff* a été réalisée par G. Chrupała et D. Seddah (c.p.). Leur précision est comparable à celle de MElt$_{\text{fr}}$ (sur les mêmes jeux de données, Henestroza et Candito (c.p.) ont obtenu $97,68\%$). Sur d'autres variantes du FTB (tokenisation d'origine), Chrupała et Seddah (c.p.) obtiennent $97,9\%$. Ces comparaisons sont à nuancer dans la mesure où des informations supplémentaires (les lemmes) sont extraites des corpus d'apprentissage et prises en compte dans ce modèle.

[15]`http://www.limsi.fr/TLP/grace/`

Nous avons cherché à comprendre au mieux la façon dont les informations extraites du Le*fff* permettent d'améliorer les résultats. Pour cela, nous avons mené un certain nombre d'expériences, notamment en faisant varier les jeux de traits et d'étiquettes.

4 Impact de différents jeux de traits extraits du Le*fff*

En vue de mieux comprendre l'impact des informations extraites du Le*fff* sur notre modèle, nous nous sommes livrés à plusieurs expériences d'ablation sur les traits décrits dans le tableau 3. Plus spécifiquement, nous avons évalué les 8 configurations possibles qui consistent à inclure (ou non) les traits lexicaux internes (INT), les traits lexicaux externes définis sur le contexte gauche (LEFT) et les traits lexicaux externes définis sur le contexte droit (RIGHT). Les résultats de ces différentes expériences menées sur le corpus de développement FTB-DEV sont repris dans le tableau 5. Notons que les configurations les plus extrêmes, $\emptyset$ et INT+LEFT+RIGHT, correspondent respectivement aux systèmes $\text{MElt}_{\text{fr}}^{\text{nolex}}$ et MElt_{fr}.

Traits Le*fff*	**Précision globale (%)**	**Précision sur les mots inconnus (%)**
$\emptyset$ ($\text{MElt}_{\text{fr}}^{\text{nolex}}$)	96,54	83,95
INT	97,04	91,4
LEFT	96,38	85,36
RIGHT	96,39	86,48
INT+LEFT	96,92	91,28
INT+RIGHT	97,30	92,01
LEFT+RIGHT	96,57	86,93
INT+LEFT+RIGHT (MElt_{fr})	97,41	92,35

TAB. 5 – Comparaison des performances de MElt_{fr} avec différents sous-ensembles de traits sur FTB-DEV

Ces résultats indiquent que c'est la combinaison des traits internes et des traits sur le contexte droit qui apporte le plus d'informations à l'étiqueteur. Le sous-ensemble INT+RIGHT donne en effet les meilleurs scores après MElt_{fr} lui-même, aussi bien sur l'ensemble des mots que sur les mots inconnus seuls. Ces deux sous-ensembles sont complémentaires : les traits INT permettent d'améliorer la couverture lexicale de l'étiqueteur (certains mots inconnus, c'est-à-dire absents du corpus d'entraînement, sont couverts par le lexique), alors que les traits RIGHT fournissent des informations importantes sur le contexte droit que les traits de $\text{MElt}_{\text{fr}}^{\text{nolex}}$ ne modélisent que frustement.

5 Impact de différents jeux d'étiquettes

Indépendamment des expériences présentées à la section précédente, nous avons entraîné différentes versions de MElt_{fr} en faisant varier à chaque fois le jeu d'étiquettes utilisé par le lexique et par le corpus d'apprentissage. Rappelons en effet que ces deux jeux d'étiquettes n'ont aucunement besoin d'être identiques, les traits lexicaux permettant d'intégrer les informations issues du lexique quels que soient les deux jeux d'étiquettes utilisés. La comparaison entre ces différentes variantes de MElt_{fr} est utile à deux points de vue au moins. Tout d'abord, elle permet de donner une idée des performances de MElt_{fr} avec différents jeux de paramètres. Certaines tâches de traitement automatique ou d'extraction d'informations n'ont peut-

être pas besoin de la granularité de notre jeu d'étiquettes d'origine. Par ailleurs, ces expériences permettent d'aborder par un autre angle l'analyse de l'impact des informations lexicales sur les performances.

Nous avons donc utilisé différentes variantes du jeu d'étiquettes, qui sont les suivantes :

`29` le jeu de départ (cf. section 2) ;

`15` les catégories principales du FTB (cf. section 2) ;

`open` le même que 15, où toutes les classes fermées (autres que `NC`, `NPP`, `ADJ`, `ADV` et `V`) sont regroupées en une seule classe *CLOSED* ;

`gram` le même que 15, où toutes les classes ouvertes sont regroupées en une seule classe *LEX*.

Jeu d'étiquettes du corpus	**Jeu d'étiquettes du lexique**				
	pas de Lefff	`gram`	`open`	`15`	`29`
`gram`	*96,74%*	98,55%	**98,82%**	98,81%	98,76%
`open`	*97,29%*	97,88%	98,12%	98,19%	**98,25%**
`15`	*96,86%*	97,15%	97,75%	97,87%	**97,87%**
`29`	*96,54%*	96,63%	97,04%	97,29%	**97,41%**

TAB. 6 – Comparaison des performances de MElt_{fr} sur FTB-DEV avec divers jeux d'étiquettes. L'étiquetage et donc l'évaluation se font sur le jeu d'étiquettes du corpus.

Ces résultats permettent de tirer quelques enseignements généraux :
– comme attendu, plus le jeu d'étiquettes sur lequel on s'évalue est riche, plus les résultats se dégradent ; une exception toutefois : dès lors que l'on utilise une des variantes du Le*fff*, étiqueter les 10 classes fermées est plus facile qu'étiqueter les 5 classes ouvertes ;
– en général, les informations du Le*fff* sont d'une aide d'autant plus grande que les étiquettes qui en sont extraites sont riches ;
– les informations lexicales semblent plus améliorer l'étiquetage des classes fermées que celui des classes ouvertes ; nous soupçonnons que ceci s'explique par l'ambiguïté des mots grammaticaux et le fait qu'ils soient difficiles à étiqueter sans aucune information lexicale spécifique.

6 Conclusions et perspectives

Nous avons présenté un étiqueteur morpho-syntaxique hybride du français, MElt_{fr}, qui a des performances état-de-l'art. Il a pour particularité de chercher à exploiter au mieux des informations extraites d'un lexique exogène non probabilisé, le Le*fff*, en plus de celles extraites d'un corpus d'apprentissage extrait du FTB. Nous avons essayé de comprendre au mieux de quelle façon ces informations lexicales contribuent au gain de performance observé entre MElt_{fr} et sa contrepartie $\text{MElt}_{\text{fr}}^{\text{nolex}}$ qui n'utilise pas le Le*fff*.

Les perspectives de ce travail sont nombreuses. Tout d'abord, des travaux préliminaires ont été menés qui montrent la pertinence du modèle sous-jacent à MElt_{fr} sur d'autres langues que le français, en particulier si une ressource lexicale comparable au Le*fff* est disponible. Ensuite, des informations supplémentaires pourraient être extraites du Le*fff*, qui sont susceptibles d'améliorer les performances. Ainsi, des informations de sous-catégorisation verbale, disponibles dans le Le*fff*, pourraient par exemple améliorer l'étiquetage d'un mot tel que *de*, ambigu entre d'une part, une préposition qui introduit parfois un argument de type objet indirect et d'autre part, un déterminant partitif qui débute parfois un argument de type objet direct.

Enfin, nous souhaitons permettre à MElt_{fr} de prendre en entrée pas seulement une séquence de mots mais plus généralement un graphe de formes, afin de permettre d'utiliser MElt_{fr} pour lever des ambiguïtés de segmentation, voire de correction orthographique, en plus de fournir une annotation en parties du discours.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Kluwer, Dordrecht.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M. & LECOMTE J. (1999). Métrique et premiers résultats de l'évaluation grace des étiqueteurs morphosyntaxiques pour le français. In *TALN*.

BERGER A., PIETRA S. D. & PIETRA V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1), 39–71.

CHRUPAŁA G., DINU G. & VAN GENABITH J. (2008). Learning morphology with morfette. In *Proceedings of the 6th Language Resource and Evaluation Conference*, Marrakech, Maroc.

CRABBÉ B. & CANDITO M. (2008). Expériences d'analyses syntaxique statistique du français. In *Proceedings of TALN'08*, Avignon, France.

DAUMÉ III H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong Kong.

HAJIČ J. (2000). Morphological Tagging : Data vs. Dictionaries. In *Proceedings of ANLP'00*, p. 94–101, Seattle, WA, USA.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*, p. 282–289.

MALOUF R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, p. 49–55, Taipei, Taiwan.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA : MIT Press.

NASR A. & VOLANSCHI A. (2004). Couplage d'un étiqueteur morpho-syntaxique et d'un analyseur partiel représentés sous la forme d'automates finis pondérés. In *TALN*.

RATNAPARKHI A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, p. 133–142.

SAGOT B. (2010). The Le*fff*, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of the 7th Language Resource and Evaluation Conference*, La Valette, Malte. à paraître.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of International Conference on New Methods in Language Processing*, p. 63–70, Hong Kong.