# Étiquetage morpho-syntaxique de tweets avec des CRF

Tian Tian, Marco Dinarelli, Isabelle Tellier, Pedro Cardoso

# Etiquetage morpho-syntaxique de tweets avec des CRF

Tian Tian[1, 2] Marco Dinarelli[1] Isabelle Tellier[1] Pedro Cardoso[2]
(1) Lattice, UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge
(2) Synthesio, 8-10 rue Villedo, 75001 Paris
ttian@synthesio.com, pedro@synthesio.com, isabelle.tellier@univ-paris3.fr, marco.dinarelli@ens.fr

**Résumé.** Nous nous intéressons dans cet article à l'apprentissage automatique d'un étiqueteur mopho-syntaxique pour les tweets en anglais. Nous proposons tout d'abord un jeu d'étiquettes réduit avec 17 étiquettes différentes, qui permet d'obtenir de meilleures performances en exactitude par rapport au jeu d'étiquettes traditionnel qui contient 45 étiquettes. Comme nous disposons de peu de tweets étiquetés, nous essayons ensuite de compenser ce handicap en ajoutant dans l'ensemble d'apprentissage des données issues de textes bien formés. Les modèles mixtes obtenus permettent d'améliorer les résultats par rapport aux modèles appris avec un seul corpus, qu'il soit issu de Twitter ou de textes journalistiques.

**Abstract.**
**Part-of-speech Tagging for Tweets with CRFs**
We are insterested in this paper in training a part-of-speach tagger for tweets in English. We first propose a reduced tagset with 17 different tags, which allows better results in accuracy than traditional tagsets which contain 45 tags. Since we have few annoted tweets, we then try and overcome this difficulty by adding data from other more standard texts into the training set. The obtained models reach better results compared to models trained with only one corpus, whether coming of Twitter or of journalistic texts.

**Mots-clés :** tweets, CRF, étiquettage morpho-syntaxique.

**Keywords:** tweets, CRFs, part-of-speech tagging.

## 1 Introduction

Les réseaux sociaux sont devenus la principale source de textes générés par des utilisateurs sur Internet. Ces textes constituent des données massives potentiellement porteuses de beaucoup d'information, mais aussi difficiles à traiter automatiquement du fait de leur grande variété en genres (textes de blogs, forums, tweets...), domaines et styles. Après la phase préliminaire de tokenisation, l'étiquetage morpho-syntaxique de ces textes apparaît comme une étape fondamentale de traitement, permettant d'éventuelles analyses syntaxiques ultérieures.

Dans cet article, nous nous intéressons à l'étiquetage morpho-syntaxique des tweets, qui présentent souvent le plus grand écart à la norme. Nous évoquons dans un premier temps la spécificité de ces données ainsi que les difficultés majeures qui en découlent pour la tâche d'étiquetage morpho-syntaxique. Puis nous introduisons les corpus (en anglais) utilisés dans nos expériences : l'un d'eux est constitué de tweets, l'autre de textes journalistiques plus respectueux de la norme linguistique standard. Pour les traiter, nous proposons d'abord un jeu d'étiquettes morpho-syntaxiques réduit par rapport à ceux utilisés dans les corpus annotés habituellement disponibles. Nous décrivons ensuite l'approche que nous avons adoptée pour apprendre un étiqueteur morpho-syntaxique de tweets anglais avec des CRF. Nous essayons en particulier d'apprendre des modèles à partir de mélanges de textes issus des deux corpus. Les paramètres que nous faisons varier dans nos expériences sont donc : les propriétés prises en compte dans les textes et les patrons qui définissent les fonctions caractéristiques des CRF, les paramètres de régularisation de l'implémentation et les proportions des différents textes sources dans l'ensemble d'apprentissage. Les résultats de nos expériences montrent que notre jeu d'étiquettes permet d'améliorer les performances de l'étiqueteur et qu'un modèle appris avec un mélange de tweets et de textes de journaux donne de meilleurs résultats que ceux appris sur un seul type de textes.

## 2 Tâche et état de l'art

Le cadre général de ce travail est celui de l'analyse d'opinion portant sur des noms de produits ou de marques cités dans des tweets. L'étiquetage morpho-syntaxique est un préalable nécessaire car nous souhaitons identifier les produits/marques évoqués (présents sous la forme de noms communs ou de noms propres) ainsi que les mots éventuellement porteurs de sentiments (principalement les adjectifs, les verbes et les adverbes). Le but de notre étiqueteur morpho-syntaxique est donc de différencier les grandes classes de catégories grammaticales, pas de vérifier des propriétés morpho-syntaxiques fines comme les accords en genre et en nombre ni de préparer une analyse syntaxico-sémantique profonde.

Malgré cette simplification, l'étiquetage automatique des tweets reste difficile, surtout à cause de leur caractère mal formé, qu'illustre par exemple la figure 1.

Today wasz Fun cusz anna Came juss for me <3( : hahaha

FIGURE 1 – Exemple 1 de tweet

Le tweet de la figure 1 est issu du corpus (Ritter *et al.*, 2011). La phrase "correcte" devrait être :

Today was fun because Anna came just for me <3( : hahaha

Dans cet exemple, les difficultés sont multiples :
– présence de fautes d'orthographe : wasz (was), cusz (because), juss (just)
– inversion majuscule/minuscule : Fun (fun), anna (Anna), Came (came)
– émoticon : <3( :
– interjection : hahaha
Dans un dictionnaire anglais, les mots comme "wasz", "cusz", "juss", "<3( :" et "hahaha" n'existent probablement pas. Les étiqueteurs à base de règles, fondées sur des listes de mots associés à leurs catégories (comme was : verbe) ne fonctionneront donc pas avec ce genre de tweets, à moins de mises à jour massives des ressources qu'elles exploitent, ou de pré-traitements permettant de "corriger" les textes initiaux. Plutôt que de chercher à réaliser ce genre de prétraitements, nous choisissons d'étiqueter ce type de textes comme s'ils étaient "bien écrits". Ceci revient à considérer que "wasz" est une variante possible du mot "was", etc. Pour cela, nous allons compter sur les capacités de généralisation de l'apprentissage automatique.

Eduardo Surita : your a freaking ... `http://tumblr.com/xmciuda0t`

FIGURE 2 – Exemple 2 de tweet

La figure 2 illustre une autre difficulté de l'étiquetage morpho-syntaxique des tweets. En anglais standard, il faudrait écrire "you're" au lieu de "your". Le mot "your" de ce tweet est ainsi porteur de deux catégories morpho-syntaxiques : pronom et verbe. Pour remédier à ce genre de problèmes, plusieurs solutions sont possibles :
– ajouter un pré-traitement de normalisation afin de substituer "you're" à "your" et traiter ensuite le tweet comme du texte écrit standard ;
– annoter le texte tel qu'il est, avec une seule étiquette syntaxique pour "your". Dans ce cas, on peut soit créer une étiquette nouvelle spéciale (pronom+verbe) comme dans (Gimpel *et al.*, 2011), soit choisir une étiquette "traditionnelle" (par exemple "pronom") comme dans (Ritter *et al.*, 2011). Cette dernière solution entrainera la possibilité de séquences d'étiquettes en principe interdites pour certaines phrases, comme "pronom déterminant adjectif", signe d'une construction sans verbe.

Ces deux exemples expliquent que les performances des étiqueteurs appris sur des corpus "bien formés" comme le Penn TreeBank (Marcus *et al.*, 1993) (aussi appelé PTB par la suite) ou le French TreeBank (Abeillé *et al.*, 2003) chutent quand ils sont confrontés à des tweets. Le Maximum Entropy POS Tagger [1] appris avec le Penn TreeBank dans (Toutanova & Manning, 2000) a obtenu 96.86% d'exactitude en validation croisée mais seulement 81.3% sur les tweets (Ritter *et al.*, 2011). Les expériences menées sur le français montrent des résultats similaires : l'étiqueteur appris sur le French TreeBank

1. Stanford Pos Tagger : http ://nlp.stanford.edu/software/tagger.shtml

dans (Constant *et al.*, 2011) atteint 97.3% d'exactitude en validation croisée, alors que celui utilisé dans (Nooralahzadeh *et al.*, 2014) a eu seulement un score de 91.7% sur le corpus French Social Media (Seddah *et al.*, 2012).

Beaucoup de travaux ont été consacrés à l'amélioration du traitement des tweets. (Foster *et al.*, 2011), par exemple, essaient d'appendre un analyseur syntaxique qui leur est dédié. (Gimpel *et al.*, 2011) propose d'utiliser un tokeniseur spécifique différent et un jeu d'étiquettes particulier (par exemple nom+verbe pour les tokens comme "I'll"). (Ritter *et al.*, 2011) cherchent aussi à construire un étiqueteur morho-syntaxique avec un modèle appris sur un mélange de plusieurs corpus : le Penn TreeBank (Marcus *et al.*, 1993), le NPS IRC Corpus (un corpus de *chatroom* introduit dans (Forsyth, 2007)) et son propre corpus twitter (T-POS). Notre travail se situe dans la même lignée, consistant à mélanger des données issues de corpus de textes bien formés et mal formés pour l'apprentissage. Nous créons ainsi un modèle mixte que nous testons sur les données de Twitter. Notre contribution dans cet article porte sur une proposition de jeu d'étiquettes réduit et sur l'étude des effets de diverses proportions de corpus "standard" et de corpus cible (Twitter) pour apprendre un modèle, associée à une optimisation des paramètres de régularisation du modèle d'apprentissage.

# 3 Les corpus T-POS, Penn TreeBank et le jeu d'étiquettes universel

(Ritter *et al.*, 2011) ont mis à disposition un corpus Twitter annoté en étiquettes morpho-syntaxiques et en entités nommées. Les sujets de discussions y sont très variés : vie quodidienne, équipes sportives, films, groupes musicaux, etc. Nous avons choisi ce corpus comme référence pour Twitter. La partie annotée est toutefois très limitée : elle ne contient que 787 tweets, soit 15972 tokens. La taille de ce corpus ne permet donc pas d'apprendre un modèle complet.

Ce corpus Twitter contient de nombreux exemples d'une autre spécificité des tweets : la présence de caractères spéciaux désignant des "hashtags" (qui servent à l'indexation des tweets par mots clés), des "retweets" (pour une rediffusion de tweets), des URL et des "usernames" (comptes d'utilisateurs de tweets), aux catégories morpho-syntaxiques souvent ambiguës. Prenons l'exemple des hashtags, repérables à leur symbole "#" : ils peuvent se substituer à un mot simple (un adjectif dans l'exemple 3), à un constituant syntaxique complet (exemple 4) ou être simplement un terme d'indexation sans rôle syntaxique précis (exemple 5).

| | |
|---|---|
| 3 | My #twitter age is 458 days 0 hours 3 minutes 49 seconds |
| 4 | On Thanksgiving after you done eating its #TimeToGetOut unless you wanna help with the dishes |
| 5 | New book blogger @GennaSarnak launches weekly feature , Poetry Sunday : http ://tinyurl.com/47vbdy5 #Books #Poetry |

TABLE 1 – Les hashtags dans les tweets

(Ritter *et al.*, 2011) ont choisi de ne pas traiter les hashtags et autres mots spéciaux utilisés dans Twitter comme des composants linguistiques comme les autres. Ils ont ajouté dans leur corpus de tweets quatre étiquettes spécifiques : USR pour "*at mention*", HT pour "*hashtag*", URL et RT pour "*retweet*" en plus des 45 étiquettes définies dans le Penn TreeBank. Ils n'ont donc pas pris en compte les éventuels rôles syntaxiques des hashtags, illustrés dans les exemples 3 et 4. Ce genre de tweets nécessiterait un étiqueteur morpho-syntaxique particulièrement flexible et robuste.

Notre étiqueteur morpho-syntaxique est construit dans un contexte d'analyse multi-langue. Un seul jeu d'étiquettes pour toutes les langues est dans ce cas nécessaire. Comme, de plus, l'objectif final de notre travail relève de la fouille d'opinion dans des données massives et devrait se passer d'une analyse syntaxique complète de ces données, nous n'avons pas besoin de certaines des distinctions du PennTreebank, comme verbes au passé / verbes au présent, nom commun singulier / pluriel, etc. Cela nous a amené à nous intéresser au jeu d'étiquettes proposé dans (Petrov *et al.*, 2012), qui se veut universel tout en laissant la possibilité de réaliser une analyse syntaxique rudimentaire. Il comporte 12 étiquettes différentes et des correspondances (*mapping*) avec les jeux d'étiquettes utilisés dans 25 TreeBanks de 25 langues différentes sont disponibles. Il a été testé sur le PTB et permet d'obtenir des résultats légèrement meilleurs en exactitude (96.8% contre 96.7% avec les étiquettes originales). Nous l'avons étendu en lui adjoignant les quatre étiquettes dédiées à Twitter utilisées dans (Ritter *et al.*, 2011). Enfin, nous avons rétabli la distinction entre les noms propres et les noms communs (qui sont assimilés dans (Petrov *et al.*, 2012)), afin de préparer le terrain à la tâche ultérieure d'extraction des entités nommées. Le nombre total d'étiquettes que nous considérons est donc finalement de 17 : NUM (nombres), PUNCT, NN (nom commun), NP (nom propre), VB (verbe), ADJ (adjectif), ADV (adverb), DET (déterminant), PRON (pronom), CC (conjonction de coordination), PREPCS (préposition et conjonction de subordination), PRT (particule), X (mot inconnu, interjection, émotion), RT, URL, USR, et HT.

Le tableau 2 montre les correspondances entre ce jeu et les étiquettes du PTB, ainsi qu'avec celles de T-POS. Les diffé-

rences d'étiquettes entre ces deux corpus sont marquées en gras.

| Tag universel | Tag Penn TreeBank | Tag T-POS | Remarques |
|---|---|---|---|
| NUM | CD | CD | nombres |
| PUNCT | " , **-LRB- -RRB-** . : **- ”** | " ( ) . , : **NONE O** LS | |
| NN | NN NNS | NN NNS | noms communs |
| NP | NNP NNPS | NNP NNPS | noms propres |
| VB | MD VB VBD VBG VBN VBP VBZ | MD VB VBD VBG VBN VBP VBZ **VPP** | verbes |
| ADJ | **AFX** JJ JJR JJS | JJ JJR JJS | AFX : *Yes* |
| ADV | RB RBR RBS WRB | RB RBR RBS WRB | WRB : *where*, *when* |
| DET | DET PDT **PRP$** | PDT DT WDT **EX TD** | PDT : *half*, PRP$ : *his* |
| PRON | PRP WP | PRP PRP$ WP WP$ | pronoms |
| CC | CC | CC | |
| PREPCS | IN | IN | préposition et CS |
| PRT | POS TO RP | POS TO RP | *particule* |
| X | **# $** FW **NIL** SYM INTJ | INTJ SYM FW | FW : mot inconnu |
| RT | | RT | *Retweet* |
| HT | | HT | *Hashtag* |
| URL | | URL | Adresse d'un site web |
| USR | | USR | Compte d'utilisateur |

TABLE 2 – Correspondance entre les étiquettes du PTB et celles de Ritter

# 4 Les CRF, les fonctions caractéristiques et les patrons

Les CRF (Conditional Random Fields), introduits dans (Lafferty *et al.*, 2001), font désormais partie des méthodes d'apprentissage automatique supervisé standard, ils sont particulièrement efficaces pour l'annotation de séquences. Différents choix sont possibles pour définir les "patrons" qui leur seront utiles pour la tâche d'annotation morpho-syntaxique. (Lavergne *et al.*, 2010) et (Suzuki & Isozaki, 2008) ont ainsi chacun proposé un ensemble de patrons destinés à l'étiquetage morpho-syntaxique de l'anglais, tandis que (Nooralahzadeh *et al.*, 2014) et (Constant *et al.*, 2011) ont construit des modèles pour le français. Nos patrons sont inspirés de ceux de (Constant *et al.*, 2011), définis pour des textes écrits. Etant donnée l'irrégularité des tweets, nous avons utilisé seulement des unigrammes d'étiquettes associés aux propriétés du tableau 3, ainsi que les bigrammes d'étiquettes seules, pour prendre en compte leurs transitions (comme dans un modèle de type *HMM*). Dans ce tableau, le chiffre 0 désigne le token courant, -1 le précédent, 1 le suivant, etc. Toutes nos expériences ont été menées avec le logiciel Wapiti [2].

| Type | Nom de propriété | Fenêtre |
|---|---|---|
| valeur de token | valeur de token | [-2, -1, 0, 1, 2] |
| valeur de token | valeur de token bigramme | [-1, 1], [-1, 0], [0, 1] |
| type de token en binaire | fstUpper, allUpper, hasDash, hasNumb | 0 |
| lowercase | lower | 0 |
| prefixe/suffixe | prefixe_n, suffixe_n (n = 1..5) | 0 |
| ressource externe en binaire | catétories dans PTB | [-2, -1, 0, 1, 2] |

TABLE 3 – Patron des CRF pour les expériences

# 5 Expériences

Notre but est tout d'abord de mesurer l'effet sur l'apprentissage de passer des jeux d'étiquettes initiaux du T-POS et du PTB (45 étiquettes) à notre jeu d'étiquettes universel spécifique des tweets (17 étiquettes). Nous nous attendons à de meilleures performances en annotation morpho-syntaxique avec moins d'étiquettes. Ensuite, nous essayons d'évaluer les performances d'un modèle mixte appris en mélangeant, dans diverses proportions, des données du PTB et de T-POS.

2. wapiti 1.5.0, https ://wapiti.limsi.fr

Pour ce faire, nous voulions construire notre propre baseline en apprenant un modèle sur le Penn TreeBank entier et en le testant sur le corpus T-POS. Mais notre serveur ne s'est pas avéré sufisamment puissant pour mener à bout ces expériences. Nous nous contentons donc de reproduire ici le résultat présenté dans (Ritter *et al.*, 2011), obtenu avec un modèle de Maximum d'Entropie : l'exactitude annoncée est de 81.3%.

Nous avons également procédé à une validation croisée en 10 blocs sur le corpus T-POS seul. L'évaluation en exactitude est dans ce cas une moyenne des 10 tests sur 1/10 de T-POS chacun.

Enfin, notre dernière série d'expériences teste des modèles appris avec un mélange de corpus PTB et les 9 blocs de T-POS, avec la même répartition aléatoire du corpus Twitter T-POS que dans la partie précédente.

L'algorithme d'optimisation utilisé pour tous les apprentissage de wapiti est rprop+.

## 5.1 Validation croisée avec T-POS uniquement

Dans cette section, nous avons testé nos patrons en validation croisée à 10 blocs, en calculant l'exactitude moyenne obtenue sur l'ensemble du jeu d'étiquettes Ritter (49 étiquettes). Pour optimiser la régularisation de la log-vraisemblance du CRF, pondérée par les paramètres L1 et L2, nous avons d'abord fixé L2 à 0.00001 (valeur par défaut dans wapiti) et nous avons testé toutes les valeurs possibles de L1 dans l'ensemble suivant : $\{0.01, 0.03, 0.1, 0.3, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$. Ensuite, nous avons gardé la valeur de L1 qui donne le meilleur résultat et nous avons fait varier la valeur de L2 dans l'ensemble $\{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1\}$ en gardant de nouveau celle qui donne le meilleur résultat.

| Jeu d'étiquettes | L1 | L2 | Moyenne d'exactitude |
|---|---|---|---|
| Ritter | 0.1 | 1 | 87.21% |
| Universal | 0.01 | 1 | 89.25% |

TABLE 4 – Validation croisée avec le corpus T-POS

D'après les résultats du tableau 4, nous remarquons que notre jeu d'étiquettes universel permet d'augmenter de 3% l'exactitude des étiquettes morpho-syntaxiques, bien que ce jeu entraîne de nouvelles séquences d'étiquettes auparavant impossibles. Ce résultat nous laisse espérer mieux analyser morpho-syntaxiquement les données de Twitter.

## 5.2 Modèles mixtes

Pour construire des modèles mixtes, trois différentes proportions de données issues du PTB ont été ajoutées à celles initialement disponibles en apprentissage (en validation croisée) dans T-POS : le même nombre de séquences, 4 fois plus et 9 fois plus. Les données de ces trois parties de PTB sont disjointes. Les ensembles d'apprentissage des corpus mixtes sont construits de la manière suivante : pour chaque itération de la validation croisée, les données provenant du corpus PTB restent les mêmes, seule la partie du corpus T-POS change. Les résultats figurent dans le tableau 5. L'évaluation des modèles est calculée par la moyenne des 10 blocs. Ensuite, l'optimisation des paramètres L1 et L2 se fait de manière identique à la partie précédente.

| Jeu d'étiquettes | Proportion | L1 | L2 | Moyenne d'exactitude |
|---|---|---|---|---|
| Ritter | 1 :1 | 1.5 | 0.0001 | 85.40% |
| Ritter | 4 :1 | 1.8 | 0.001 | 86.72% |
| Ritter | 9 :1 | 1.9 | 0.0001 | 87.18% |
| Universal | 1 :1 | 1.0 | 0.01 | 89.11% |
| Universal | 4 :1 | 1.0 | 0.5 | 89.27% |
| Universal | 9 :1 | 1.6 | 0.01 | 88.95% |

TABLE 5 –

Ce résultat montre qu'avec le jeu d'étiquettes Ritter, le fait d'ajouter des textes issus du PTB n'augmente pas vraiment l'exactitude par rapport aux validations croisées (qui était de 87.21%), C'est sans doute dû au fait que les deux corpus ne se ressemblent pas sufisamment : les nouvelles données introduisent de nouveaux tokens et de nouvelles séquences

d'étiquettes qui n'aident pas à mieux reconnaître celles de Twitter. L'effet semble légèrement moindre avec le jeu d'étiquettes universel, qui permet une très légère amélioration. Nous remarquons néanmoins qu'ajouter successivement plus de données du PTB dans l'ensemble d'apprentissage (mélangées avec celles du corpus T-POS) améliore les performances.

# 6 Conclusion et perspectives du travail

Dans cet article, nous proposons tout d'abord un jeu d'étiquettes réduit par rapport aux 45 étiquettes du PTB, qui sera facilement exploitable pour d'autres langues tout en préservant les spécificités de Twitter. Ce jeu d'étiquettes rend l'étiqueteur morpho-syntaxique plus fiable en regroupant entre elles les catégories similaires/proches. Mais sa limite est qu'il ne distingue pas bien certaines catégories. Une nouvelle version du jeu de Tags universel a d'ors et déjà été proposée dans `http://universaldependencies.github.io/docs/u/pos/index.html`. Cette version sépare les conjonctions de subordination des prépositions, les auxiliaires des verbes, les symboles et les interjections des X et les noms propres des noms communs. Nous avons déjà pris en compte cette dernière distinction. Comme la prochaine étape de notre travail est d'analyser les opinions dans les tweets, nous allons tester l'impact des deux jeux d'étiquettes pour cette analyse d'opinions.

D'autre part, nous avons essayé d'apprendre un étiqueteur morpho-syntaxique à partir de textes écrits et de tweets, pour étiqueter des tweets. Les résultats de ces expériences montrent qu'ajouter des données éloignées de la cible dans la phase d'apprentissage permet d'améliorer l'exactitude. Mais, pour apprendre un étiqueteur plus performant, rien ne vaut l'augmentation de la taille des données d'apprentissage proches de la cible. À défaut, il faudrait envisager l'utilisation de ressources linguistiques externes ou de données non étiquetées.

Une autre piste serait de trouver des patrons plus adaptés aux données de Twitter (nous avons vu que les bigrammes d'étiquettes avec les caractéristiques ne fonctionnent pas). Il peut être aussi intéressant d'ajouter d'autres caractéristiques pertinentes. L'exemple donné au début de l'article dans la figure 1 suggère l'utilisation de transcriptions phonétiques pour normaliser des tokens comme "wasz" ou "cusz" , comme proposé dans (Clark, 2003).

Des clusters de grandes quantités de tweets comme dans (Nooralahzadeh *et al.*, 2014), une optimisation des paramètres L1 et L2 à la façon de (Dinarelli & Rosset, 2012) sont aussi des pistes exploitables pour cette tâche. Enfin, un traitement spécifique des hashtags (s'ils jouent un vrai rôle syntaxique) pourrait aussi permettre un étiquetage plus fin et plus régulier des messages.

# Références

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

CLARK A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, p. 59–66, Stroudsburg, PA, USA : Association for Computational Linguistics.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *TALN*, volume 1, p. 321, Montpellier, France.

DINARELLI M. & ROSSET S. (2012). Tree-structured named entity recognition on ocr data : Analysis, processing and results.

FORSYTH E. N. (2007). Improving automated lexical and discourse analysis of online chat dialog.

FOSTER J., ÇETINOGLU Ö., WAGNER J., LE ROUX J., HOGAN S., NIVRE J., HOGAN D., VAN GENABITH J. *et al.* (2011). # hardtoparse : Pos tagging and parsing the twitterverse. In *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, p. 20–25.

GIMPEL K., SCHNEIDER N., O'CONNOR B., DAS D., MILLS D., EISENSTEIN J., HEILMAN M., YOGATAMA D., FLANIGAN J. & SMITH N. A. (2011). Part-of-speech tagging for twitter : Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, p. 42–47, Stroudsburg, PA, USA : Association for Computational Linguistics.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 504–513, Stroudsburg, PA, USA : Association for Computational Linguistics.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of english : The penn treebank. *COMPUTATIONAL LINGUISTICS*, **19**(2), 313–330.

NOORALAHZADEH F., BRUN C. & ROUX C. (2014). Part of speech tagging for french social media data. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, August 23-29, 2014, Dublin, Ireland*, p. 1764–1772.

PETROV S., DAS D. & MCDONALD R. (2012). A universal part-of-speech tagset. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

RITTER A., CLARK S., MAUSAM & ETZIONI O. (2011). Named entity recognition in tweets : An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1524–1534, Stroudsburg, PA, USA : Association for Computational Linguistics.

SEDDAH D., C B. S. M., MOUILLERON V. & COMBET V. (2012). The french social media bank : a treebank of noisy user generated content.

SUZUKI J. & ISOZAKI H. (2008). Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *In ACL*.

TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, p. 63–70, Stroudsburg, PA, USA : Association for Computational Linguistics.