Adaptation au domaine pour l'analyse morpho-syntaxique

Éléonor Bartenlian Margot Lacour Matthieu Labeau Alexandre Allauzen
Guillaume Wisniewski François Yvon
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
prénom.nom@limsi.fr

RÉSUMÉ

Ce travail cherche à comprendre pourquoi les performances d'un analyseur morpho-syntaxiques chutent fortement lorsque celui-ci est utilisé sur des données hors domaine. Nous montrons à l'aide d'une expérience jouet que ce comportement peut être dû à un phénomène de masquage des caractéristiques lexicalisées par les caractéristiques non lexicalisées. Nous proposons plusieurs modèles essayant de réduire cet effet.

ABSTRACT

Domain Adaptation for PoS tagging

This work aims at understanding the performance drop observed when a PoS-tagger is applyed on out-domain data. We design a toy experiment showing that this observation may result from the fact that lexicalized features are masking non lexicalized features. We propose several improvements of history-based model aiming at reducing this phenomenon.

MOTS-CLÉS : Analyse morpho-syntaxique, adaptation au domaine, modèles statistiques, UGC.

KEYWORDS: PoS tagging, Domain Adaptation, UGC.

1 Introduction

L'étiquetage morpho-syntaxique *(PoS tagging)* est une tâche qui est généralement considérée comme résolue : à partir du moment où suffisamment de données du domaine considéré sont disponibles, les performances des méthodes statistiques sont telles que les seules erreurs de prédiction restantes résultent d'erreurs d'annotation ou de cas particulièrement ambigus (Manning, 2011). Pourtant dès que l'on change de domaine ou de type de documents, les performances chutent de manière spectaculaire (Martínez Alonso *et al.*, 2016). Les raisons généralement avancées pour expliquer cette chute sont l'augmentation du nombre de mots hors vocabulaire (HV) qui n'ont jamais été observés sur le corpus d'apprentissage et l'incapacité des modèles statistiques à généraliser les connaissances extraites du corpus d'apprentissage à ceux-ci (Foster, 2010; Seddah *et al.*, 2012).

L'objectif de ce travail est de fournir une première explication de cette observation : nous montrons par plusieurs expériences, qu'une raison possible peut être liée à un phénomène de *masquage* des caractéristiques non lexicalisées (robustes au changement de domaine mais avec un faible pouvoir prédictif) par les caractéristiques lexicalisées (permettant d'obtenir de très bonnes performances sur des données du même domaine, mais avec un faible pouvoir de généralisation). Nous proposons et évaluons également plusieurs méthodes cherchant à limiter cet effet.

caractéristiques lexicalisées	caractéristiques non lexicalisées
◇ mot à la i^{e} position	◇ trois dernières lettres des i^{e}, $(i \pm 1)^{\text{e}}$ mots
◇ mots aux positions $i-2$, $i-1$, $i+1$, $i+2$	◇ dernière lettre des i^{e} et $(i-1)^{\text{e}}$ mot
◇ conjonction du i^{e} mot et de la $(i-1)^{\text{e}}$ étiquette	◇ étiquettes des $(i-1)^{\text{e}}$ et $(i-2)^{\text{e}}$ mot
	◇ conjonction des deux étiquettes précédentes

TABLE 1 – Patron des caractéristiques utilisées dans notre analyseur morpho-syntaxique. Nous distinguons 6 patrons lexicalisés (reposant sur la connaissance du token) de 8 patrons non lexicalisées.

Le reste de cet article est organisé de la manière suivante : nous commencerons par mettre en évidence l'impact du changement de domaine sur les performances d'un analyser morpho-syntaxique (§2) ; nous chercherons ensuite à expliquer pourquoi les analyseurs morpho-syntaxiques ne sont pas robustes à un changement de domaine (§3) avant de proposer et d'évaluer trois approches visant à résoudre ces problèmes (§4).

2 Évaluation hors domaine d'un analyseur morpho-syntaxique

Nos expériences utilisent un analyseur morpho-syntaxique à base d'historique (Black *et al.*, 1992; Tsuruoka *et al.*, 2011). Dans ces modèles, la prédiction d'une séquence d'étiquettes morpho-syntaxiques est modélisée sous la forme d'une suite de problèmes de décision, consistant chacun à prédire l'étiquette d'une observation. Chaque décision est prise par un classifieur multi-classe utilisant comme descripteurs des informations extraites de la structure d'entrée, ainsi que les décisions prises antérieurement. Nous utilisons, dans toutes nos expériences, un perceptron moyenné comme classifieur multi-classe (Collins, 2002). Les caractéristiques utilisées sont décrites dans le Tableau 1 [1]. Ce sont des caractéristiques simples et indépendantes de la langue qui sont, à notre connaissance, utilisées dans tous les étiqueteurs morpho-syntaxiques. Une description détaillée de ce modèle est présentée dans (Wisniewski *et al.*, 2014b,a).

Pour mesurer l'impact du changement de domaine sur les performances de l'analyse morpho-syntaxique, nous avons mené une série d'expériences sur le français. Les paramètres de l'analyseur sont estimés sur le corpus d'apprentissage du projet UD [2] qui a pour objectif de fournir, pour de nombreuses langues, des corpus annotés en PoS, morphologie et dépendances dans un schéma commun. Ses performances sont évaluées sur un corpus du domaine (la partie test du corpus UD) ainsi que sur quatre corpus hors domaine :

— deux corpus de contenu généré par des utilisateurs (*User Generated Content*, UGC) collectés sur les forums de discussion du site MARMITON [3] et du jeu MINECRAFT [4]. La collecte et l'analyse de ces corpus sont présentées dans (Martínez Alonso *et al.*, 2016).

— deux corpus de tweets, un parlant de football, FOOT et l'autre de catastrophe naturelle, NATDIS. Ces deux corpus sont décrits plus précisément dans (Adda *et al.*, 2017).

Ces corpus diffèrent du corpus d'apprentissage sur plusieurs points : ils sont composés de phrases

1. Les entrées sont également transformées : tous les nombres, les URL et les mentions sont remplacées par un même token
2. http://universaldependencies.org/
3. http://www.marmitton.org
4. http://www.minecraft.com

MARMITON	① Que faire de facile avec un reste de boeuf pour une fondue ? ? ? ② Qu' avez -vous dans le vôtre ?
MINECRAFT	③ tu fais kit artisan ④ Apr'es tu veux alli sale noob de Azarid ...
FOOT	⑤ Ça va faire un petit Juventus - Barça , excellent aha 🙂 ⑥ Ça c' est de l' attaque ! La BBC 😉 #HalaMadrid #RealMadrid http ://t.co/iCknoGm9mF
NATDIS	⑦ A Lire #Liban Une vague de chaleur exceptionnelle frappe le Liban , 40 degrés à midi http ://t.co/moDWsrreaT ⑧ Au @Radiojournal_RC 7h Un nouveau tremblement de terre sème la terreur au Népal http ://t.co/UPtZpytQ0q

TABLE 2 – Extrait des différents corpus hors domaine considérés.

corpus	#phrases	# mots	% mots HV	% erreur	% erreur HV	% erreur ambigu
UD	298	6 829	5,8%	4,4%	18,0%	5,6%
FOOT	743	13 985	26,2%	16,0%	37,2%	10,4%
NATDIS	622	1 233	23,4%	8,8%	18,3%	6,9%
MARMITON	285	2 058	18,5%	13,2%	35,7%	11,1%
MINECRAFT	234	900	38,7%	40,2%	69,6%	23,9%

TABLE 3 – Caractéristiques des corpus considérés pour l'évaluation et taux d'erreurs obtenus selon le type de mots.

plus courtes, comportant de nombreuses idiosyncrasies et fautes d'orthographe et diffèrent par leur style, leur genre, leur registre et les thématiques qu'ils abordent. La Figure 2 montre des extraits de ces corpus. Les étiquettes de ces corpus ont été converties manuellement vers le schéma de l'UD [5].

Les performances obtenues par notre modèle sur un corpus du domaine (cf. Tableau 3) sont proches de celles de l'état de l'art ((Straka *et al.*, 2016) obtient, par exemple, un taux d'erreur de 2,92%), bien que le modèle d'apprentissage soit extrêmement simple et qu'aucun effort particulier n'ait porté sur le choix et la conception des caractéristiques [6].

Comme on pouvait s'y attendre, changer de domaine a un fort impact négatif sur la qualité de la prédiction : l'analyseur réalise entre 2 et 4 fois plus d'erreurs sur les corpus hors domaine que sur le test du domaine. Pour mieux comprendre l'impact du changement de domaine, nous proposons d'évaluer, les performances de l'analyseur sur différents types de mots :

— les mots hors vocabulaire (HV), c'est-à-dire ceux qui n'apparaissent jamais dans le corpus d'apprentissage ;
— les mots ambigus, c'est-à-dire ceux qui *i)* apparaissent plus d'une fois dans le corpus d'apprentissage et *ii)* n'apparaissent pas toujours avec la même étiquette.

Les résultats de cette évaluation sont donnés dans le Tableau 3. Sans surprise, le nombre de mots hors

5. Ces corpus ont été étiquetés en suivant les conventions du FTB qui définit un jeu d'étiquettes plus riche que celui de l'UD. La conversion est donc déterministe et ne pose pas de problèmes particuliers.

6. Comme souligné dans de nombreux travaux (Manning, 2011), cette évaluation est toutefois optimiste puisqu'elle prend en compte des occurrences dont l'étiquette peut être prédite de manière déterministe (par exemple les symboles de ponctuation).

corpus	#exemples train/test/HV	Δ erreur train	Δ erreur test	Δ erreur HV	# mises à jour lex/delex
UD	5 917/638/88	-18,84%	-13.26%	1,36%	10 476 / 55 364
SPMRL	10 440/1 769/145	-20,06%	-17,00%	0,14%	6 674 / 81 554

TABLE 4 – Résultats de notre expérience jouet : nous reportons, pour les différents corpus, la différence entre la précision du modèle lexicalisé et celle du modèle non lexicalisé.

vocabulaire augmente fortement dès que l'on change de domaine. On observe également, de manière plus surprenante, que globalement les étiquettes des mots HV et des mots ambigus sont nettement moins bien prédites sur le corpus hors domaine que sur le corpus du domaine. Cette observation, similaire à celle faite par (Barrett *et al.*, 2007), pourrait indiquer que la structure des phrases (qui conditionne les mots et étiquettes au voisinage d'un mot HV) change d'un domaine à l'autre et ne permet plus de désambiguïser les mots HV.

3 L'effet de masquage des caractéristiques lexicalisées

Il est surprenant que les performances sur les mots HV soient beaucoup plus faibles que celles sur les mots observés en apprentissage. Il est bien connu que, pour l'analyse morpho-syntaxique, la caractéristique la plus utile est la connaissance du mot, puisque la plupart des mots ne sont pas ou peu ambigus. Ainsi, sur le corpus UD, un analyseur prédisant l'étiquette la plus fréquente associée à chaque mot ne fait que 11.0% d'erreur, les mots HV étant considérés comme des erreurs. Toutefois, les modèles de l'état de l'art comportent tous des caractéristiques décrivant la structure (étiquettes voisines) et une approximation plus ou moins grossière de la « morphologie » des mots (à minima les suffixes et préfixes), afin d'être capable de *généraliser* les observations faites sur le corpus d'apprentissage aux mots HV. Par exemple, un mot inconnu se trouvant entre un déterminant *(une)* et un nom *(voiture)* sera très probablement un adjectif [7] ; les caractéristiques de forme aident à prédire qu'un mot se terminant en *-ment* sera très certainement un adverbe.

Les résultats présentés dans le Tableau 3 montrent que ces caractéristiques n'arrivent pas à jouer leur rôle. Nous faisons l'hypothèse que cela est la conséquence du très fort pouvoir prédictif des caractéristiques *lexicalisées* : comme expliqué dans le paragraphe précédent, connaître le mot dont on cherche à prédire l'étiquette permet de déterminer celle-ci de manière quasi déterministe et, l'on peut supposer qu'il suffit d'observer une fois un mot pour « retenir » l'étiquette qui lui est associée. Mais, une fois que l'étiquette d'un mot est correctement prédite (ce qui est, à priori, possible à partir d'une unique caractéristique lexicale), celui-ci ne participe plus à l'apprentissage puisque dans les modèles à base de correction d'erreur, comme le perceptron, les paramètres ne sont mis-à-jour lorsqu'une prédiction est erronée et dans les modèles maximisant la vraisemblance (log-loss), comme les réseaux neuronaux ou les CRF, la norme du gradient est proportionnelle à l'erreur. L'estimation du poids des autres caractéristiques sera alors moins fiable, puisqu'étant moins souvent mis à jour, leur variance sera plus grande. Tout se passe comme si les caractéristiques lexicales *masquaient* les autres. Une observation similaire a été faite par (Klein & Manning, 2002) qui en faisant le lien avec le phénomène *explaining-away* (Pearl, 1988), montre comment certaines caractéristiques peuvent être ignorées lors de l'apprentissage. Les expériences menées par (Pécheux *et al.*, 2015) mettent également en évidence

7. Cela pourrait également être un nom comme dans « un chien voiture »

ce phénomène en étudiants l'impact de contraintes appliquées lors de l'apprentissage d'un analyseur morpho-syntaxique.

Pour confirmer cette hypothèse, nous proposons l'expérience jouet suivante : est-il possible d'identifier si un mot se terminant en `-ent` est un verbe (conjugué à la troisième personne du pluriel) ou un adverbe, à partir de la connaissance du mot et de l'étiquette du mot précédent ? En réduisant le nombre de caractéristiques et d'étiquettes, nous espérons limiter les interactions entre caractéristiques et obtenir des modèles plus facilement interprétables.

Plus précisément, nous extrayons des corpus UD et SPMRL (Seddah *et al.*, 2013) français tous les mots se terminant en `-ent` ainsi que l'étiquette (oracle) du mot précédent. Ces mots sont des verbes, des noms, des adjectifs ou des adverbes. Nous considérons ensuite deux perceptrons moyennés essayant d'identifier ces classes à partir de deux jeux de caractéristiques différents : le premier considère des caractéristiques *lexicalisées* (l'étiquette du mot précédent et le mot courant), le second uniquement des caractéristiques non lexicalisées (l'étiquette du mot précédent).

Les résultats, présentés Table 4, confirment notre hypothèse. Les caractéristiques lexicalisées ont un fort pouvoir prédictif (le taux d'erreur global du modèle lexicalisé est nettement plus faible que celui du modèle non lexicalisé), mais elles ne permettent pas de généraliser les connaissances apprises aux mots qui n'ont pas été vus en apprentissage : le modèle non lexicalisé obtient de meilleures performances que le modèle lexicalisé si l'on restreint l'évaluation aux mots qui n'ont pas été vus en apprentissage. La comparaison du nombre de mises à jour effectuées pour les deux modèles semble indiquer que cette observation est bien le résultat du phénomène de *masquage* décrit précédemment.

4 Méthode d'adaptation au domaine

Nous proposons dans cette section deux améliorations de notre analyseur morpho-syntaxique visant à améliorer les performances de celui-ci sur des données hors domaine.

Méthodes La première méthode, SPEC, repose sur une spécialisation du perceptron utilisé dans notre analyseur morpho-syntaxique. Elle consiste simplement à utiliser deux perceptrons différents pour prendre les décisions successives lors de l'inférence : un premier classifieur, utilisant l'ensemble des caractéristiques de la Table 1 sera utilisé pour prédire les étiquettes des mots observés en apprentissage ; un second, n'utilisant que les caractéristiques non lexicalisées prédira les étiquettes des mots HV. Ces deux classifieurs sont appris successivement sur l'ensemble des données d'apprentissage. Nous espérons par cette spécialisation des modèles mieux estimer les paramètres liées aux caractéristiques de structure et de morphologie que nous supposons plus robuste au changement de domaine que les caractéristiques reposant sur les mots, tout en continuant à exploiter les caractéristiques lexicalisées pour les mots vus en apprentissage.

La seconde méthode, DROPOUT-T, est directement inspirée de la méthode proposée par (Kouw *et al.*, 2016) : partant du constat que les différences entre domaines se traduisent souvent par des changements dans la probabilité d'observation des caractéristiques (typiquement une caractéristique très pertinente n'apparaît qu'en apprentissage et pas en test), les auteurs proposent d'apprendre un modèle plus robuste en *bruitant* les données d'apprentissage. C'est une approche classique en apprentissage statistique (Van Der Maaten *et al.*, 2013; Wang *et al.*, 2013) permettant de réduire le sur-apprentissage.

	UD	FOOT	NATDIS	MARMITON	MINECRAFT
BASELINE	4,4% 18,0%/5,5%	16,0% 37,2%/10,4%	8,8% 18,3%/6,9%	13,2% 35,7%/11,1%	40,2% 69,6%/23,9%
SPEC	4,3% 21,0%/5,3%	16,3% 37,9%/10,9%	9,5% 22,8%/6,4%	13,9% 37,5%/11,5%	40,1% 72,4%/22,6%
DROPOUT-T	4,9% 20,5%/6,1%	16,2% 37,5%/11,0%	8,7% 18,3%/7,0%	12,8% 35,7%/10,6%	38,3% 68,7%/26,0%
DROPOUT-A	4,3% 17,7%/5,6%	16,0% 37,1%/10,6%	8,6% 19,2%/6,6%	12,4% 32,0%/10,9%	37,6% 68,7%/23,4%
NEURONALE	4,2%	15,0%	8,9%	12,3%	41,8%

TABLE 5 – Précision obtenue par les différentes méthodes d'adaptation au domaine proposées. Les taux d'erreur sur les mots HV et les mots ambigus sont détaillés.

Plus précisément, (Kouw *et al.*, 2016) propose de supprimer aléatoirement certaines caractéristiques lors de l'apprentissage afin d'encourager le modèle à utiliser les autres caractéristiques et à estimer « correctement » leur poids. En pratique, chaque caractéristique f d'une observation est supprimée avec une probabilité :

$$\max\left\{0, 1 - \frac{p_{test}}{p_{train}}\right\} \tag{1}$$

où p_{test} est la probabilité d'observer la caractéristique f sur les données de test et p_{train} la probabilité d'observer f sur les données d'apprentissage. Intuitivement, plus une caractéristique est typique des observations du domaine d'apprentissage ($p_{train} \gg p_{test}$), plus elle aura des chances d'être souvent supprimée ; à la limite, une caractéristique n'apparaissant jamais dans les données de test devrait être supprimée avec probabilité 1. Pour mesurer l'intérêt de cette approche, nous considérons également la méthode DROP-A dans laquelle la probabilité de supprimer une caractéristique est arbitrairement fixée à 0,3.

La troisième méthode, NEURONAL, repose sur un étiqueteur neuronal. Cette approche permet de s'affranchir des problèmes liés aux mots HV en modélisant directement la séquence de caractères composant le mot. Elle construit la représentation vectorielle continue d'un mot à partir de la séquence de ses caractères, à l'aide d'une couche de convolution. La prédiction de l'étiquette est réalisée avec une couche cachée classique, qui prend en entrée les représentations vectorielle du mot considéré et de ses voisins. Une description plus détaillée de cette approche est faite dans (Labeau *et al.*, 2015).

Évaluation expérimentale Les résultats de ces deux approches sont décrits dans le tableau 5. Aucun des méthodes proposées ne parvient à réduire significativement l'impact du changement de domaine. De manière plus surprenante, sur plusieurs corpus, les résultats obtenus par les approches proposées sont très proches de ceux obtenus par un simple perceptron. Nous pensons que cette observations peut-être liées à deux raisons : *i)* des différences systématiques dans la préparation (p. ex. la segmentation en mots) et l'annotation des données, une hypothèse également étudiée par (Rabary *et al.*, 2015) *ii)* une mauvaise compréhension ou modélisation des caractéristiques robustes au changement de domaine (p. ex. est-ce que la structure des phrases est conservée lors du changement de domaine). Nous menons actuellement des expériences pour mettre en évidence l'impact de ces deux raisons.

Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (projet PARSITI, ANR-16-CE33-0021). Nous remercions les relecteurs pour leurs commentaires et suggestions.

Références

ADDA G., CREGO J.-M., LACROIX O., LI G., MOSTEFA D., VLADIMIR P., WISNIEWSKI G. & FRANÇOIS Y. (2017). *A multilingual corpus of tweets for domain adaptation*. Research report, LIMSI-CNRS ; ELDA ; Systran.

BARRETT L., GREENBERG D. F. & SCHWARTZ M. (2007). A syntactic feature counting method for selecting machine translation training corpora. *Language and Computers : Studies in Practical Linguistics*, **60**, 1–19.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

COLLINS M. (2002). Discriminative training methods for hidden markov models : Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, p. 1–8 : Association for Computational Linguistics.

FOSTER J. (2010). "cba to check the spelling" : Investigating parser performance on discussion forum posts. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 381–384, Los Angeles, California : Association for Computational Linguistics.

KLEIN D. & MANNING C. D. (2002). Conditional structure versus conditional estimation in nlp models. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, p. 9–16, Stroudsburg, PA, USA : Association for Computational Linguistics.

KOUW W. M., VAN DER MAATEN L. J., KRIJTHE J. H. & LOOG M. (2016). Feature-level domain adaptation. *Journal of Machine Learning Research*, **17**(171), 1–32.

LABEAU M., LÖSER K. & ALLAUZEN A. (2015). Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 232–237, Lisbon, Portugal : Association for Computational Linguistics.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In *Proceedings of the Computational Linguistics and Intelligent Text Processing, 12th International Conference*, CICLing 2011, p. 171–189 : Springer.

MARTÍNEZ ALONSO H., SEDDAH D. & SAGOT B. (2016). From noisy questions to minecraft texts : Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 13–23, Osaka, Japan : The COLING 2016 Organizing Committee.

PEARL J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

PÉCHEUX N., ALLAUZEN A., LAVERGNE T., WISNIEWSKI G. & YVON F. (2015). Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles crf. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 37–48, Caen, France : Association pour le Traitement Automatique des Langues.

RABARY C., LAVERGNE T. & NÉVÉOL A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 508–514, Caen, France : Association pour le Traitement Automatique des Langues.

SEDDAH D., SAGOT B., CANDITO M., MOUILLERON V. & COMBET V. (2012). The French Social Media Bank : a treebank of noisy user generated content. In *Proceedings of COLING 2012*, p. 2441–2458, Mumbai, India : The COLING 2012 Organizing Committee.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & DE LA CLERGERIE E. V. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

STRAKA M., HAJIČ J. & STRAKOVÁ J. (2016). UDPipe : trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France : European Language Resources Association (ELRA).

TSURUOKA Y., MIYAO Y. & KAZAMA J. (2011). Learning with lookahead : Can history-based models rival globally optimized models ? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL'11, p. 238–246, Portland, Oregon, USA : Association for Computational Linguistics.

VAN DER MAATEN L., CHEN M., TYREE S. & WEINBERGER K. Q. (2013). Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, p. I–410–I–418 : JMLR.org.

WANG S., WANG M., WAGER S., LIANG P. & MANNING C. D. (2013). Feature noising for log-linear structured prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1170–1179, Seattle, Washington, USA : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014a). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014b). Cross-lingual pos tagging through ambiguous learning : First experiments (apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue) [in french]. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 173–183, Marseille, France : Association pour le Traitement Automatique des Langues.