

# Exploration de l'apprentissage par transfert pour l'analyse de textes des réseaux sociaux

Sara Meftah<sup>1</sup> Nasredine Semmar<sup>1</sup> Youssef Tamaazousti<sup>3</sup>

Hassane Essafi<sup>1</sup> Fatiha Sadat<sup>2</sup>

(1) CEA LIST, LASTI, Gif-sur-Yvette, France

(2) UQAM, Montréal, Canada

(3) MIT, CSAIL, Massachusetts, États-Unis

{prenom.nom}@cea.fr, ytamaaz@mit.edu, sadat.fatiha@uqam.ca

## RÉSUMÉ

---

L'apprentissage par transfert représente la capacité qu'un modèle neuronal entraîné sur une tâche à généraliser suffisamment et correctement pour produire des résultats pertinents sur une autre tâche proche mais différente. Nous présentons dans cet article une approche fondée sur l'apprentissage par transfert pour construire automatiquement des outils d'analyse de textes des réseaux sociaux en exploitant les similarités entre les textes d'une langue bien dotée (forme standard d'une langue) et les textes d'une langue peu dotée (langue utilisée en réseaux sociaux). Nous avons expérimenté notre approche sur plusieurs langues ainsi que sur trois tâches d'annotation linguistique (étiquetage morpho-syntaxique, annotation en parties du discours et reconnaissance d'entités nommées). Les résultats obtenus sont très satisfaisants et montrent l'intérêt de l'apprentissage par transfert pour tirer profit des modèles neuronaux profonds sans la contrainte d'avoir à disposition une quantité de données importante nécessaire pour avoir une performance acceptable.

## ABSTRACT

---

### Exploring neural transfer learning for social media text analysis

Transfer learning consists in learning a parent neural network on a source problem with enough data, then transferring a part of its weights to represent data of a target problem with few training examples. We present in this paper an approach based on transfer learning to automatically build tools to analyze social media texts by exploiting similarities between texts of a resource-rich language (standard language) and texts of a low-resource language (social media). We conducted experiments on various languages and three Natural Language Processing tasks : Morpho-Syntactic tagging, Part-Of-Speech tagging and Named entity recognition. The obtained results are very satisfactory and show the interest of transfer learning to take advantage of deep neural models without the constraint of having a large amount of data required to obtain an acceptable performance.

**MOTS-CLÉS :** Apprentissage par transfert, Contenus des réseaux sociaux, Langues peu dotées, Adaptation au domaine, Étiquetage morpho-syntaxique, Reconnaissance d'entités nommées.

**KEYWORDS:** Transfer learning, Social media content, Low-resource languages, Domain adaptation, Part-Of-Speech tagging, Named entity recognition.

---

# 1 Introduction

Les méthodes d'apprentissage automatique qui reposent sur les Réseaux de Neurones (RNs) obtiennent des performances qui s'approchent de plus en plus de la performance humaine dans plusieurs applications du Traitement Automatique de la Langue (TAL) qui bénéficient de la capacité des différentes architectures des RNs à généraliser à partir des régularités apprises à partir d'exemples d'apprentissage. En particulier, la structure ordonnée et les dépendances temporelles des données textuelles nécessitent un traitement spécifique. En effet, le contexte joue un rôle important pour identifier le sens d'un mot ou comprendre une phrase dans un document. Pour ce type de tâches, les RNs récurrents (RNNs pour *Reccurent Neural Networks*) et ses variantes ; dont les deux principales : le modèle *Long Short-Term Memory* (LSTM) et sa version simplifiée *Gated Recurrent Units* (GRUs) sont les plus adaptés grâce à leur capacité à conserver en mémoire les informations pertinentes en analysant les mots (ou les caractères) dans un ordre précis. En outre, les RNs à convolutions (CNNs pour *Convolutional Neural Networks*) ont aussi montré leur efficacité pour l'encodage des caractères. Plusieurs études (Jozefowicz *et al.*, 2016) ont montré que les CNNs représentaient l'architecture idéale pour l'extraction et l'encodage des informations morphologiques (racine, préfixe, suffixe, etc.), en particulier pour les langues avec une morphologie riche (Chiu & Nichols, 2015; Ma & Hovy, 2016) comme l'arabe, les langues slovaques, le hindi, etc.

Toutefois, pour être performants, ces modèles neuronaux ont besoin de corpus annotés de taille importante. Par conséquent, uniquement les langues et les domaines bien dotés peuvent bénéficier directement de l'avancée apportée par les RNs, comme par exemple les formes standard de l'anglais, du français, de l'arabe, entre autres. Cependant, la majorité des langues ne sont pas dotées en données d'apprentissage ou ayant des données annotées de très petite taille (Baumann & Pierrehumbert, 2014). Dès lors, les systèmes les plus performants de l'état de l'art pour les langues peu dotées sont ceux fondés sur des règles, construites manuellement dans la majorité des cas. La principale limitation de ces systèmes réside dans leur incapacité à s'adapter à de nouvelles langues et de nouveaux domaines.

Pour pallier au problème de la dépendance des RNs aux données annotées, de nombreux travaux récents s'intéressent à la construction de modèles statistiques et particulièrement des modèles neuronaux pour les domaines peu dotés en exploitant les données annotées des domaines bien dotés (Duong, 2017; Meftah *et al.*, 2018a). Particulièrement, la succession de couches dans les RNs leur permet d'apprendre les connaissances d'une façon hiérarchique, en d'autres termes, des connaissances générales aux couches inférieures et des connaissances spécifiques au problème d'apprentissage aux couches supérieures. Par exemple en traitement d'images les premières couches ont tendance à apprendre les bordures (Yosinski *et al.*, 2014) et les connaissances morphologiques en TAL (Peters *et al.*, 2018), d'où l'intérêt de l'apprentissage par transfert neuronal (TL pour *Transfer Learning*) qui consiste à transférer les connaissances acquises lors de l'apprentissage des problèmes sources aux problèmes cibles.

Nous étudions plus particulièrement dans cet article l'apport du TL par *fine-tuning* pour l'adaptation au domaine, pour contourner le problème de manque de données d'apprentissage dans le domaine des réseaux sociaux (RS). En effet, les performances des outils TAL utilisant des modèles neuronaux appris sur des corpus volumineux annotés manuellement tels que le corpus *Wall Street Journal* (WSJ) de *Penn TreeBank* (PTB) (Marcus *et al.*, 1993) et évalués sur des données du même domaine se rapprochent des performances humaines (97.96% de précision par Bohnet *et al.* (2018)). En revanche, les performances de ces outils chutent fortement lorsque ceux-ci sont appliqués sur des données hors domaine telles que les textes générés par les internautes sur les RS notamment les textes de nature conversationnelle (Twitter, SMS, etc.). Cela est dû principalement aux erreurs linguistiques,

aux incohérences orthographiques, aux abréviations informelles et au style idiosyncratique. De plus, Twitter pose un problème supplémentaire en imposant une limite de 280 caractères pour chaque Tweet. Les expériences menées sur trois tâches : l'étiquetage des séquences des textes des RS : l'annotation en parties de discours (*PoS tagging*), l'annotation morpho-syntaxique (*MS tagging*) et la reconnaissance des entités nommées (*NE recognition*) montrent l'efficacité du *fine-tuning* pour l'annotation des textes des RS.

Le reste de cet article est organisé comme suit : nous commençons dans la section 2 par formaliser l'approche que nous utilisons pour le transfert de connaissances apprises à partir de la forme standard des langues pour l'amélioration de l'annotation des textes des RS. Ensuite, nous décrivons dans la section 3 l'architecture neuronale, les tâches et les corpus sur lesquels nous avons expérimenté notre approche. Nous discutons, par la suite, les résultats que nous avons obtenus dans la section 4. Et finalement, la section 5 conclut notre étude et présente nos travaux futurs.

## 2 Notre approche pour l'annotation des textes de RS

Considérons que nous avons un problème source  $\mathcal{P}_s = (\mathcal{D}_s, \mathcal{T}_s)$  et un problème cible  $\mathcal{P}_c = (\mathcal{D}_c, \mathcal{T}_c)$ , où  $\mathcal{D}_s$  et  $\mathcal{D}_c$  représentent les domaines source et cible, respectivement, et  $\mathcal{T}_s$  et  $\mathcal{T}_c$  représentent les tâches source et cible, respectivement. Nous visons à améliorer l'apprentissage de la fonction de prédiction cible à partir des connaissances apprises de  $\mathcal{D}_s$  et  $\mathcal{T}_s$ . Pour cela, nous transférons les connaissances du problème source  $\mathcal{P}_s$  au problème cible  $\mathcal{P}_c$ , tels que  $\mathcal{T}_s = \mathcal{T}_c$ ,  $\mathcal{D}_s \neq \mathcal{D}_c$  et  $n_s > n_c$  (le domaine source est plus riche en données annotées que le domaine cible). Pour notre cas, le domaine source  $\mathcal{D}_s$  représente les textes de la forme standard de la langue et le domaine cible  $\mathcal{D}_c$  représente les textes des RS de la même langue, la tâche source et cible  $\mathcal{T} = \mathcal{T}_s = \mathcal{T}_c$  est le *PoS tagging*, le *MS tagging* ou *NE recognition*.

Nous utilisons le schéma du *fine-tuning* standard, comme illustré sur la figure 1, nous avons un RNs source  $\mathcal{N}_s$  avec un ensemble de paramètres  $\theta_s$  répartis en deux ensembles  $\theta_s = (\theta_s^1, \theta_s^2)$ , et un RNs cible  $\mathcal{N}_c$  avec un ensemble de paramètres  $\theta_c$  répartis en deux ensembles :  $\theta_c = (\theta_c^1, \theta_c^2)$ . Le transfert est effectué en trois étapes : (1)  $\mathcal{N}_s$  est entraîné sur le problème source avec les données annotées du domaine source  $\mathcal{D}_s$ . (2) Les poids du premier ensemble de paramètres du  $\mathcal{N}_s$  sont transférés au réseau cible  $\mathcal{N}_c$  :  $\theta_c^1 = \theta_s^1$  (les poids des paramètres  $\theta_c^1$  sont initialisés avec les poids appris par le RNs source et ceux des paramètres  $\theta_c^2$  sont initialisés aléatoirement) (3)  $\mathcal{N}_c$  est affiné sur le problème cible en lançant l'entraînement sur le petit corpus du domaine cible  $\mathcal{D}_c$ . Notant que le choix du nombre de paramètres (nombre de couches) à transférer dépend de la similarité entre les problèmes source et cible et la disponibilité des données annotées pour le problème cible (Mou *et al.*, 2016).

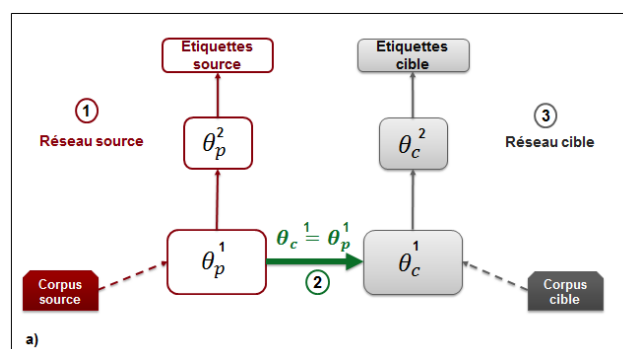


FIGURE 1 – Le schéma de l'adaptation aux domaines par *fine-tuning*.

# 3 Expérimentations

## 3.1 Description des tâches et des jeux de données

Nous expérimentons l'approche discutée sur trois tâches : *MS tagging*, *PoS tagging* et *NE recognition*.

1. Le *PoS tagging* consiste à étiqueter chaque mot dans la phrase avec sa partie de discours (Nom, Adjective, etc.). Pour le domaine source, nous utilisons la partie **WSJ** du PTB (Marcus *et al.*, 1993), annotée avec l'ensemble d'étiquettes du PTB. Et nous évaluons le performances du *fine-tuning* sur trois corpus des Tweets : **TPoS** (Ritter *et al.*, 2011), annoté également avec l'ensemble d'étiquettes du PTB ; **ARK** (Owoputi *et al.*, 2013) annoté avec un ensemble de 25 classes conçues pour les textes des RS ; et **TweeBank**, récemment proposé par Liu *et al.* (2018) et annoté avec l'ensemble d'étiquettes universelles.
2. Le *MS tagging* consiste à étiqueter chaque mot de la phrase avec une étiquette morphosyntaxique, où la partie de discours est enrichie avec les informations morphologiques du mot. Comme illustré sur la figure 2, chaque lettre de l'étiquette représente une catégorie. La première position représente la partie de discours et les autres positions représentent des catégories morphosyntaxiques, comme le nombre, le genre, etc. Nous utilisons les données fournies par la compagnie d'évaluation *MTT* du **Vardial18** (Zampieri *et al.*, 2018), contenant un corpus de la forme standard et un autre des Tweets pour chacune des langues : Slovène, Croate et Serbe, annotés en *MS tagging*.
3. Pour le *NE recognition*, nous utilisons comme corpus source l'ensemble de données de l'anglais standard de la compagnie d'évaluation **CONLL-03** (Sang & De Meulder, 2003) contenant 4 types d'entités nommées : Personne, Organisation, Lieu et *Misc*, la classe des autres entités n'appartenant à aucune des classes précédentes. Concernant le domaine cible, nous utilisons **WNUT-17** de la compagnie d'évaluation *Emerging Entity Detection* (Derczynski *et al.*, 2017), contenant 6 types, dont trois communs avec le corpus source : Personne, Organisation et Lieu et trois différents : Produit, Groupe et Travail Créatif.

POS = pronoun Type = demonstrative Gender = neuter Number = singular Case = nominative  MS tag = Pd-nsn <b>To</b>	POS = verb Type = auxiliary Vform = present Person = third Number = singular Negative = yes  Va-r3s-y <b>ni</b>	POS = pronoun Type = negative Gender = feminine Number = singular Case = nominative  Pz-fsn <b>nobena</b>	POS = noun Type = common Gender = feminine Number = singular Case = nominative  Ncfsn <b>novost</b>	POS = Punctuation          <b>.</b>
--	---	--	--	---

FIGURE 2 – Exemple d'une phrase Slovène annotée en classes morphosyntaxiques (To ni nobena novost - "Ce n'est pas une nouveauté.") .

Les statistiques des jeux de données sur lesquels nous avons évalué nos modèles sont résumées dans le tableau 1.

## 3.2 Architecture neuronale

Étant donnée une phrase  $S = [w_1, \dots, w_n]$  de  $n$  mots successifs  $w_i$ , l'objectif d'un étiqueteur est de prédire la classe  $c_i \in \mathcal{C}$  pour chaque  $w_i$ , où  $\mathcal{C} \in \mathbb{R}^C$  est l'ensemble de classes. Nous utilisons une architecture neuronale communément utilisée pour l'étiquetage des séquences (Ma & Hovy, 2016; Meftah *et al.*, 2018b). Tout d'abord, chaque mot  $w_i$  de la phrase est représenté par  $x_i$  : une concaténation de deux plongements de mots hybrides. Le premier est le résultat de l'encodage de tous les caractères du mot avec un LSTMs bi-directionnels (bi-LSTMs) permettant un meilleur traitement

Tâche	Langue	Domaine	Corpus	# classes	# phrases	# mots
MS	Slovène	Standard RS	Vardial18	1,304	27,829	586,248
				1102	6,670	64,108
	Croate	Standard RS	Vardial18	772	24,611	506,460
654				6,763	75,907	
Serbe	Standard RS	Vardial18	557	3,891	86,765	
			589	5,884	78,616	
POS	Anglais	Standard RS	WSJ	36	67,786	1,173,766
			T-POS	40	787	15,000
			Ark	25	2,374	34,301
			TweeBank	17	3,550	55,607
NER	Anglais	Standard RS	CONLL-03	4	20,744	301,418
			WNUT-17	6	5,690	101,857

TABLE 1 – Statistiques des jeux de données utilisés. Pour chaque langue, le corpus de la langue standard est utilisé pour le pré-apprentissage et le corpus du domaine des RS est utilisé pour l’affinement.

des mots hors vocabulaire en capturant leurs caractéristiques morphologiques. Le deuxième est une représentation contextuelle du mot qui permet de capturer sa sémantique. Ensuite, les plongements  $[x_1, \dots, x_n]$  sont fournis dans l’ordre chronologique des mots de la phrase à un extracteur de représentations à base de bi-LSTMs, dont les sorties  $[h_1, \dots, h_n]$  sont acheminées via une couche finale linéaire avec une activation Softmax afin de générer une distribution de probabilité pour chaque classe.

### 3.3 Détails d’implémentation

Les hyper-paramètres utilisés pour notre modèle sont les suivants : nous avons fixé la dimension de la couche de l’embedding des caractères à 50, la dimension de l’encodeur des caractères à base de bi-LSTMs à 100 pour le *NE recognition* et le *PoS tagging* et 200 pour le *MS tagging*. Les plongements de mots sont de dimension 300, ces derniers sont initialisés avec des vecteurs pré-appris et disponibles publiquement. Plus particulièrement, nous avons utilisé les vecteurs Glove (Pennington *et al.*, 2014) pré-appris sur 42 milliards de mots du Web, pour le *PoS tagging* et le *NE recognition*, et les vecteurs pré-appris avec FastText sur des données du web<sup>1</sup> pour le *MS tagging*. Pour la couche bi-LSTMs de l’extracteur de représentations, nous avons fixé sa dimension à 200. Finalement, pour toutes nos expérimentations, l’apprentissage est effectué en utilisant l’algorithme SGD avec momentum et des mini-batches de 8 phrases.

## 4 Résultats expérimentaux et discussion

Nous rapportons dans la première partie du tableau 2 les meilleurs résultats de l’état de l’art pour chaque corpus sur lequel nous avons expérimenté notre approche. L’étiqueteur **ARK** (Owoputi *et al.*, 2013) est basé sur le modèle de Markov d’entropie maximale, utilisant les *brown clusters* et des règles soigneusement conçues à la main. **TPANN** (Gui *et al.*, 2017) est un modèle neuronal, dont l’approche se base sur l’apprentissage antagoniste (Ganin *et al.*, 2016) permettant d’exploiter les Tweets non annotés (1,17M mots) et WSJ. **UH&CU** est le système du *MS tagging* proposé par Silfverberg & Drobac (2018) pour la campagne d’évaluation Vardial 2018. Leur approche est fondée

1. <https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md>

sur un modèle neuronal qui au lieu de considérer une étiquette morphosyntaxique comme une entité, génère les différentes catégories morphosyntaxiques comme une séquence de caractères en rajoutant un décodeur à base de bi-LSTMs à la sortie du modèle.

Dans la deuxième partie du tableau 2, nous fournissons les résultats expérimentaux de l'utilisation de l'apprentissage par transfert pour l'adaptation des connaissances apprises lors de l'entraînement sur la forme standard d'une langue, au domaine des RS. **Modèle source** consiste à évaluer sur le corpus cible, les performances du modèle appris uniquement sur le corpus source sans adaptation<sup>2</sup>. **Modèle cible** présente les résultats lorsque nous n'entraînons le RNs que sur le corpus des RS (corpus cible), c'est-à-dire, l'initialisation aléatoire de ses paramètres. **Apprentissage par Transfert** montre les résultats de l'approche proposé dans cet article, avec le pré-apprentissage sur le corpus annoté de la langue standard et le *fine-tuning* sur le corpus cible des Tweets.

Langue	Anglais				Sl	Cr	Ser
Tâche	PoS tagging			NER	MS tagging		
Corpus	T-Pos	ARK	Tweeb	WNUT	Vardial		
ARK	90.40	<u>93.2</u>	<u>94.6</u>	n/a	n/a	n/a	n/a
TPANN	<u>90.92</u>	92.8	n/a	n/a	n/a	n/a	n/a
(Aguilar <i>et al.</i> , 2018)	n/a	n/a	n/a	<u>45.55</u>	n/a	n/a	n/a
UH&CU	n/a	n/a	n/a	n/a	<u>88.4</u>	<u>88.7</u>	<u>90</u>
<b>Modèle source (sans adaptation)</b>	76.03	n/a	n/a	n/a	70.65	78.95	73.39
<b>Modèle cible (sans préapprentissage)</b>	87.76	90.96	91.64	41.00	82.87	84.25	85.38
<b>Apprentissage par Transfert</b>	<b>90.70</b>	<b>91.81</b>	<b>93.00</b>	<b>43.57</b>	<b>87.83</b>	<b>88.30</b>	<b>87.06</b>

TABLE 2 – Les performances des étiqueteurs de séquence (la précision (%) du *PoS tagging* et du *MS tagging* et la mesure F1 (%) du *NE recognition*) des textes des réseaux sociaux.

En comparant les résultats sans et avec pré-apprentissage, nous constatons que le transfert améliore considérablement les performances pour tous les corpus. En revanche, nous constatons que l'apport est plus important pour *T-PoS* (+3%) et les langues slovaques (+5% pour le slovène). En effet ces corpus partagent les mêmes classes que celles du corpus source utilisé pour le pré-apprentissage, mais cela n'est pas le cas pour *ArK*, *Tweebank* et *WNUT*. Par conséquent, la dernière couche du RNs est pré-apprise pour les expérimentations sur *T-PoS* et les langues slovaques et initialisée aléatoirement pour *ArK*, *Tweebank* et *WNUT*.

Nous traçons dans la figure 3 la courbe des performances de notre modèle sans TL et avec TL sur l'ensemble du développement de *Tweebank* avec différentes tailles de l'ensemble d'apprentissage des Tweets. Nous pouvons remarqué que le gain apporté par le TL est plus important dans les scénarios où les exemples d'apprentissage du domaine cible sont rares.

Pour mieux comprendre l'impact du pré-apprentissage, la figure 4 présente le pourcentage des prédictions corrigées et celles falsifiées en introduisant le pré-apprentissage du RNs sur le domaine source par rapport à l'initialisation aléatoire. Nous constatons que le pré-apprentissage améliore considérablement les prédictions. Cependant, les falsifications causées par le transfert négatif des régularités spécifiques aux corpus sources (Meftah *et al.*, 2019) réduisent l'apport final du pré-apprentissage. Parmi les erreurs causées par le transfert négatif que nous avons constaté :

1. Les mots avec une première lettre en majuscule ; en effet, dans la forme standard des langues,

2. Notant que cette expérimentation n'est pas possible pour tous les corpus, notamment, les cas où les corpus cibles et sources ne partagent pas le même ensemble de classes (ARK, TweepBank et WNUT)

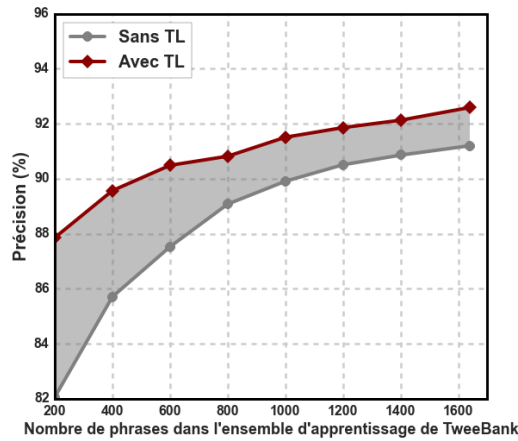


FIGURE 3 – Les performances (sur l’ensemble du développement du TweeBank) avec différentes tailles de l’ensemble d’apprentissage des Tweets.

la première lettre des noms propres est mise en majuscule et le modèle pré-entraîné ne parvient pas à oublier cette régularité qui n’est souvent pas respectée dans les textes des RS.

2. Les mots contenant des ponctuations sont souvent prédits comme *PUNCT*, par exemple : *I’M*.
3. Les verbes (*Did, has, is*) sont prédits comme des auxiliaires dans les cas où ils se trouvent avant un deuxième verbe. Par exemple : *stay informed*.

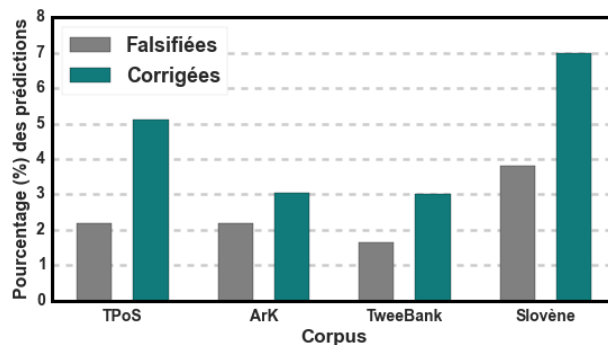


FIGURE 4 – Pourcentage des prédictions corrigées et celles falsifiées en introduisant le pré-apprentissage sur le corpus de la langue standard par rapport à l’initialisation aléatoire.

## 5 Conclusion

Dans cet article, nous avons proposé et implémenté une approche fondée sur l’apprentissage par transfert pour construire automatiquement des outils d’analyse de textes des réseaux sociaux. Cette approche s’appuie sur les similarités entre les textes de la langue standard et les textes des réseaux sociaux relatifs à cette langue. Nous avons montré la validité de notre approche et sa généralité en l’expérimentant sur plusieurs langues et trois tâches d’annotation linguistique. En outre, nous avons constaté que malgré son impact positif, l’apprentissage par transfert est accompagné par un effet de falsification de quelques prédictions des modèles appris sans cet apprentissage.

Nos travaux futurs s’orientent, d’une part, vers l’expérimentation de cette approche sur d’autres langues morphologiquement riches comme l’arabe et ses dialectes, et d’autre part, vers la modélisation des similarités entre la langue standard et ses langues proches en vue d’intégrer cette connaissance externe dans le modèle neuronal pour le forcer à prendre en compte cette connaissance lors de la phase de prédiction.

# Références

- AGUILAR G., MONROY A. P. L., GONZÁLEZ F. & SOLORIO T. (2018). Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1401–1412.
- BAUMANN P. & PIERREHUMBERT J. B. (2014). Using resource-rich languages to improve morphological analysis of under-resourced languages. In *LREC*, p. 3355–3359.
- BOHNET B., McDONALD R., SIMÕES G., ANDOR D., PITLER E. & MAYNEZ J. (2018). Morpho-syntactic tagging with a meta-bilstm model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2642–2652.
- CHIU J. P. & NICHOLS E. (2015). Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv :1511.08308*.
- DERCZYNSKI L., NICHOLS E., VAN ERP M. & LIMSOPATHAM N. (2017). Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 140–147.
- DUONG L. (2017). *Natural language processing for resource-poor languages*. PhD thesis.
- GANIN Y., USTINOVA E., AJAKAN H., GERMAIN P., LAROCHELLE H., LAVIOLETTE F., MARCHAND M. & LEMPITSKY V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, **17**(1), 2096–2030.
- GUI T., ZHANG Q., HUANG H., PENG M. & HUANG X. (2017). Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2411–2420.
- JOZEFOWICZ R., VINYALS O., SCHUSTER M., SHAZEER N. & WU Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv :1602.02410*.
- LIU Y., ZHU Y., CHE W., QIN B., SCHNEIDER N. & SMITH N. A. (2018). Parsing tweets into universal dependencies. In *NAACL*, volume 1, p. 965–975.
- MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv :1603.01354*.
- MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, **19**(2), 313–330.
- MEFTAH S., SEMMAR N. & SADAT F. (2018a). A neural network model for part-of-speech tagging of social media texts. In *LREC*.
- MEFTAH S., SEMMAR N., SADAT F. & RAAIJMAKERS S. (2018b). Using neural transfer learning for morpho-syntactic tagging of south-slavic languages tweets. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, p. 235–243.
- MEFTAH S., TAMAAZOUSTI Y., SEMMAR N., ESSAFI H. & SADAT F. (2019). Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. In *HLT-NAACL*.
- MOU L., MENG Z., YAN R., LI G., XU Y., ZHANG L. & JIN Z. (2016). How transferable are neural networks in nlp applications? *arXiv preprint arXiv :1603.06111*.
- OWOPUTI O., O’CONNOR B., DYER C., GIMPEL K., SCHNEIDER N. & SMITH N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*, p. 380–390.



- PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.
- PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*.
- RITTER A., CLARK S., ETZIONI O. *et al.* (2011). Named entity recognition in tweets : an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1524–1534 : Association for Computational Linguistics.
- SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- SILFVERBERG M. & DROBAC S. (2018). Sub-label dependencies for neural morphological tagging—the joint submission of university of colorado and university of helsinki for vardial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, p. 37–45.
- YOSINSKI J., CLUNE J., BENGIO Y. & LIPSON H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, p. 3320–3328.
- ZAMPIERI M., MALMASI S., NAKOV P., ALI A., SHON S., GLASS J., SCHERRER Y., SAMARDŽIĆ T., LJUBEŠIĆ N., TIEDEMANN J., VAN DER LEE C., GRONDELAERS S., OOSTDIJK N., VAN DEN BOSCH A., KUMAR R., LAHIRI B. & JAIN M. (2018). Language Identification and Morphosyntactic Tagging : The Second VarDial Evaluation Campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

