Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle

Othman Zennaki, Nasredine Semmar, Laurent Besacier

Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle

Othman Zennaki[1,2] Nasredine Semmar [1] Laurent Besacier [2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie de Contenus, F-91191, Gif-sur-Yvette, France
(2) Laboratoire d'Informatique de Grenoble, Univ. Grenoble-Alpes, Grenoble, France
othman.zennaki, nasredine.semmar@cea.fr, laurent.besacier@imag.fr

Résumé. La construction d'outils d'analyse linguistique pour les langues faiblement dotées est limitée, entre autres, par le manque de corpus annotés. Dans cet article, nous proposons une méthode pour construire automatiquement des outils d'analyse via une projection interlingue d'annotations linguistiques en utilisant des corpus parallèles. Notre approche n'utilise pas d'autres sources d'information, ce qui la rend applicable à un large éventail de langues peu dotées. Nous proposons d'utiliser les réseaux de neurones récurrents pour projeter les annotations d'une langue à une autre (sans utiliser d'information d'alignement des mots). Dans un premier temps, nous explorons la tâche d'annotation morpho-syntaxique. Notre méthode combinée avec une méthode de projection d'annotation basique (utilisant l'alignement mot à mot), donne des résultats comparables à ceux de l'état de l'art sur une tâche similaire.

Abstract.

Use of Recurrent Neural Network for Part-Of-Speech tags projection from a parallel corpus.
In this paper, we propose a method to automatically induce linguistic analysis tools for languages that have no labeled training data. This method is based on cross-language projection of linguistic annotations from parallel corpora. Our method does not assume any knowledge about foreign languages, making it applicable to a wide range of resource-poor languages. No word alignment information is needed in our approach. We use Recurrent Neural Networks (RNNs) as cross-lingual analysis tool. To illustrate the potential of our approach, we firstly investigate Part-Of-Speech (POS) tagging. Combined with a simple projection method (using word alignment information), it achieves performance comparable to the one of recently published approaches for cross-lingual projection.

Mots-clés : Multilinguisme, transfert crosslingue, étiquetage morpho-syntaxique, réseaux de neurones récurrents.

Keywords: Multilingualism, cross-Lingual transfer, part-of-speech tagging, recurrent neural network.

1 Introduction

L'annotation linguistique de ressources consiste à ajouter des informations de nature interprétative aux données brutes originales (Garside *et al.*, 1997). Ces informations peuvent être d'ordre terminologique, lexical, morphologique, syntaxique ou sémantique et les ressources linguistiques peuvent être des lexiques, des dictionnaires, des transcriptions de dialogue ou des corpus de textes (Véronis, 2000). Ces ressources linguistiques sont annotées par des outils d'analyse linguistique et utilisées dans de nombreuses applications : recherche d'information translingue, fouille de textes, extraction d'informations, traduction automatique, etc.

Dans la littérature, il a été montré que les outils d'analyse linguistique les plus performants sont ceux construits pour les quelques langues (richement dotées) disposant des ressources linguistiques manuellement annotées nécessaires aux algorithmes d'apprentissage supervisé. Cependant, la plus grande majorité des langues (faiblement dotées) ne disposent pas de telles ressources annotées.

La construction manuelle de ces ressources est lente et coûteuse, rendant ainsi l'utilisation des approches supervisées difficile voire impossible. Dans cet article, nous nous intéressons à l'induction de ressources linguistiques adéquates à moindre coût pour les langues faiblement dotées, et aussi à la construction automatique d'outils d'analyse linguistique pour ces langues. Pour cela, nous proposons d'utiliser des approches fondées sur *la projection interlingue d'annotations*.

Celles-ci s'articulent autour de l'exploitation des corpus parallèles multilingues entre une langue source richement dotée (disposant d'outils d'analyse linguistique) et une langue cible faiblement dotée. En partant d'un corpus parallèle dont les

textes en langue *source* sont déjà annotés, les textes en langue *cible* sont annotés par projection des annotations à l'aide de techniques d'alignement automatique au niveau des mots.

Bien que prometteuses, ces approches non supervisées ont des performances assez éloignées de celles des méthodes supervisées. Par exemple, pour une tâche d'analyse morpho-syntaxique supervisée, (Petrov *et al.*, 2012) obtient une précision moyenne de 95.2% pour 22 langues richement dotées, tandis que les analyseurs morpho-syntaxiques non supervisés construits par (Das & Petrov, 2011; Duong *et al.*, 2013) donnent une précision moyenne de 83.4% pour 8 langues Européennes.

Dans cet article, nous explorons la possibilité d'employer les réseaux de neurones récurrents (RNN) pour induire des outils multilingues d'analyse linguistique. Dans un premier temps, nous abordons la possibilité de les utiliser comme analyseurs morpho-syntaxiques. Pour cela, nous utilisons un corpus parallèle entre une langue bien dotée et une autre langue moins bien dotée, pour assigner aux mots du corpus parallèle (appartenant aux vocabulaires des langues source et cible) une représentation commune, obtenue à partir d'un alignement au niveau des phrases. Cette représentation commune permet d'apprendre — à partir d'une seule langue étiquetée parmi N — un seul analyseur multilingue capable de traiter N langues.

Après un bref état de l'art présenté dans la section 2, notre modèle est décrit dans la partie 3 et son évaluation est présentée dans la partie 4, la partie 5 conclut notre étude et présente nos travaux futurs.

2 État de l'art

La projection interlingue d'annotations a été introduite par (Yarowsky *et al.*, 2001), en utilisant un corpus parallèle pour adapter des outils monolingues (analyseurs morpho-syntaxiques, analyseurs syntaxiques de surface et analyseurs morphologiques) à de nouvelles langues. Le transfert entre les langues a été rendu effectif en utilisant les alignements au niveau des mots entre les phrases d'un corpus parallèle. Plusieurs outils permettent d'obtenir automatiquement de tels alignements, dont GIZA++ (Och & Ney, 2000). Cet outil implémente divers modèles de traduction (IBM 1, 2, 3, 4, 5 et HMM). Ces modèles utilisent l'algorithme EM (Dempster *et al.*, 1977) pour l'apprentissage à partir de corpus bilingues. L'alignement des mots est réalisé à l'aide d'un algorithme de recherche de type Viterbi. GIZA++ est un outil efficace pour aligner les mots simples, mais il est moins performant, d'une part, lorsque les langues source et cible ont des morphologies et des structures syntaxiques différentes, et d'autre part, pour aligner les expressions multi-mots (Allauzen & Wisniewski, 2009; Abdulhay, 2012).

Cette méthode a été ensuite utilisée avec succès dans plusieurs autres travaux. Ainsi, (Das & Petrov, 2011; Duong *et al.*, 2013) ont montré qu'il était possible d'apprendre des analyseurs morpho-syntaxiques de bonne qualité de cette manière. Dans cette lignée, (Wisniewski *et al.*, 2014; Täckström *et al.*, 2013) ont obtenu de meilleures performances encore, en combinant les informations obtenues par projection avec les informations extraites d'un dictionnaire qui associe à chaque mot (de la langue cible) l'ensemble des étiquettes morpho-syntaxiques autorisées, puis en utilisant des méthodes d'apprentissage faiblement supervisées.

La projection interlingue a été aussi adaptée avec succès pour transférer d'autres types d'annotations. Par exemple, la projection d'annotations en sens réalisée par (Bentivogli *et al.*, 2004; Van der Plas & Apidianaki, 2014), l'annotation en rôles sémantiques sur l'allemand par projection interlingue à partir de la paire de langues anglais-allemand (Padó & Lapata, 2005, 2006), dont la généricité a plus spécifiquement été évaluée dans (Pado & Pitel, 2007). De plus cette méthode permet la portabilité multilingue des applications utilisant les annotations linguistiques, (Jabaian *et al.*, 2013) l'ont utilisé pour la portabilité d'un système de compréhension de la parole pour des langues ou domaines différents.

Dans ces approches, les annotations du côté source sont projetées vers le côté cible, à travers les alignements automatiques du corpus parallèle obtenus au niveau des mots. Cette annotation partielle et bruitée des textes cibles est ensuite utilisée par des méthodes d'apprentissage robustes. Cependant, les performances des algorithmes d'alignement au niveau des mots ne sont pas toujours satisfaisantes (du point de vue de la qualité des alignements prédits) et l'étape d'alignement au niveau des mots (un alignement n'est pas toujours 1-1, il peut être 1-N, N-N, etc.) constitue aujourd'hui un facteur limitant la projection d'annotations linguistiques (Fraser & Marcu, 2007). Pour cette raison, notre approche utilise un corpus parallèle aligné au niveau des phrases seulement et n'applique aucun pré-traitement du type *alignement automatique en mots* qui est source d'erreurs et de bruit.

3 Méthode proposée

Pour faire face aux limitations relatives à l'étape d'alignement mot à mot des phrases du corpus parallèle, nous proposons de ne pas prendre en compte les informations bruitées issues de cet alignement, mais de représenter ces informations

de façon intrinsèque dans l'architecture du réseau de neurones. Dans ce travail initial, nous implémentons un analyseur morpho-syntaxique multilingue basé sur les réseaux de neurones récurrents, et nous montrons que ses performances sont proches de l'état de l'art des autres analyseurs morpho-syntaxiques non supervisés.

Avant de décrire notre étiqueteur morpho-syntaxique multilingue basé sur les réseaux de neurones récurrents (RNN), nous décrivons tout d'abord l'approche par projection simple à laquelle nous allons nous comparer (et qui sera aussi combinée — au cours des expériences qui vont suivre — avec la méthode que nous proposons).

3.1 Annotateur morpho-syntaxique non supervisé par projection simple

L'approche pour construire notre étiqueteur morpho-syntaxique non supervisé par projection simple (décrit par l'algorithme 1) est très proche de celle introduite par (Yarowsky *et al.*, 2001). Cette approche, qui a été réutilisée plus récemment par (Duong *et al.*, 2013), correspond à l'état de l'art des annotateurs morpho-syntaxiques non supervisés. Ces auteurs utilisent l'alignement automatique en mots (obtenu à partir d'un corpus parallèle) pour projeter les annotations de la langue source vers la langue cible, en vue de construire des annotateurs morpho-syntaxiques pour la langue cible.

L'algorithme 1 est décrit dans l'encadré ci-dessous :

Algorithme 1 : Méthode de référence par projection d'annotations selon un alignement automatique en mots

1 : Annoter le côté source du corpus parallèle.
2 : Aligner automatiquement le corpus parallèle en utilisant GIZA++ ou un autre outil d'alignement en mots.
3 : Projeter les annotations directement pour les alignements 1-1.
4 : Pour les correspondances N-1, projeter l'annotation du mot se trouvant à la position N/2 arrondi à l'entier supérieur.
5 : Annoter les mots non-alignés avec l'étiquette la plus fréquente qui leur est associée dans le corpus.
6 : Apprendre un analyseur morpho-syntaxique à partir de la partie cible du corpus désormais annotée (par exemple, dans notre cas, nous utilisons TNT tagger (Brants, 2000).

3.2 Annotateur morpho-syntaxique non supervisé fondé sur les réseaux de neurones récurrents

Les réseaux de neurones sont généralement classés dans deux grandes catégories : les réseaux de neurones *Feed-forward* (Bengio *et al.*, 2006) et les réseaux de neurones Récurrents (on utilise l'acronyme RNN en anglais - pour *Recurrent Neural Networks*) (Mikolov *et al.*, 2010). (Sundermeyer *et al.*, 2013) ont montré que les modèles de langue statistiques basés sur une architecture récurrente présentent de meilleurs performances que les modèles basés sur une architecture *Feed-forward*. Cela vient du fait que les réseaux de neurones récurrents utilisent un contexte de taille non limitée, contrairement aux réseaux *Feed-forward* dont la topologie limite la taille du contexte pris en compte. Cette propriété a motivé notre choix d'utiliser, dans nos expériences, un réseau de neurones de type récurrent (Elman, 1990).

Dans cette section, nous décrivons en détail l'approche proposée pour la construction d'un étiqueteur morpho-syntaxique multilingue basé sur les RNNs. L'approche, qui ne nécessite aucune ressource externe, requiert simplement un corpus parallèle et un annotateur morpho-syntaxique pré-existant dans la langue source.

3.2.1 Description du modèle

Un RNN est au minimum composé d'une succession de trois couches de neurones : une couche d'entrée au temps *t* notée *x(t)*, une couche cachée *h(t)* (aussi appelée couche de contexte), et une couche de sortie *y(t)*. Chaque neurone de la couche d'entrée est relié à tous les neurones de la couche cachée par les matrices des poids U et W. La matrice des poids V connecte tout neurone de la couche cachée à chaque neurone de la couche de sortie, cf. (Figure 1).

Dans notre modèle, la couche d'entrée est formée par la concaténation de la représentation vectorielle *w(t)* du mot courant, et de la couche cachée au temps précédent *h(t-1)* (information de la première des couches cachées, dans le cas où on utilise plusieurs). La première étape de notre modèle est donc d'associer à chaque mot *w* (appartenant aux vocabulaires des langues source et cible) une représentation vectorielle spécifique.

Notre idée est la suivante : si on arrive à construire un espace de représentation commun, où un mot source et sa traduction cible possèdent des représentations vectorielles proches, nous pourrons — à partir de cette représentation commune —

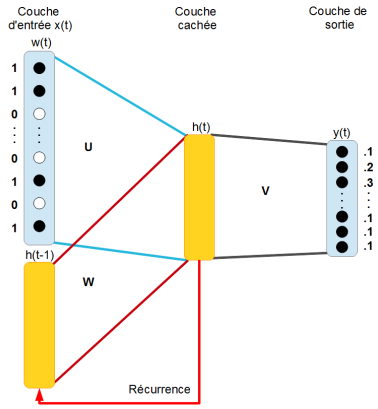

FIGURE 1 – Exemple de réseau de neurones récurrent

utiliser l'annotateur morpho-syntaxique de type RNN (appris initialement sur le coté source) pour annoter un texte en langue cible.

En général, un mot source et sa traduction cible apparaissent le plus souvent ensemble dans les mêmes bi-phrases, et donc leurs empreintes distributionnelles sont proches. Nous faisons le choix de construire notre espace de représentation commun, en associant à chaque mot (source, cible) son empreinte distributionnelle *Vw* de dimension *N* (nombre de bi-phrases dans le corpus parallèle) indiquant si le mot apparaît ou pas dans chaque bi-phrase *Phi* $\{i = 1, ..., N\}$ du corpus parallèle :

$$V_w = \begin{cases} V_{wi} = 1 & si \quad w \in Ph_i \\ V_{wi} = 0 & sinon \end{cases} \tag{1}$$

Par ailleurs, nous utilisons deux couches cachées (des expériences préliminaires ont montré que ceci permet d'obtenir de meilleures performances), avec des tailles variables (de 80 à 1024 neurones). Pour la fonction d'activation, nous utilisons la fonction *sigmoïde*. Nous pensons que ces couches cachées devraient permettre de capturer intrinsèquement des informations d'alignement au niveau des mots.

De plus, pour pouvoir transférer les annotations morpho-syntaxiques d'une langue à une autre, il est nécessaire que ces annotations soient décrites de la même manière dans les deux langues. (Petrov *et al.*, 2012) définissent un ensemble de 12 étiquettes morpho-syntaxiques universelles à gros grain, communes au plus grand nombre de langues (*Universal Tagset*). Ces étiquettes universelles sont les suivantes : NOUN (noms), VERB (verbes), ADJ (adjectifs), ADV (adverbes), PRON (pronoms), DET (déterminants et articles), ADP (prépositions et postpositions), NUM (numéraux), CONJ (conjonctions), PRT (particules), « . » (symboles de ponctuations) et X (pour tout ce qui échappe aux autres catégories). Dans nos travaux, nous adoptons ces étiquettes morpho-syntaxiques universelles. Par conséquent, la couche de sortie de notre modèle comporte 12 neurones, chaque neurone correspondant à une étiquette morpho-syntaxique universelle. On utilise la fonction d'activation *softmax* sur la couche de sortie afin d'obtenir des scores assimilables à des probabilités, le mot *w* en entrée du réseau est annoté par l'étiquette la plus probable en sortie du réseau.

3.2.2 Construction du modèle

Tout d'abord, avant l'apprentissage du modèle, quelques étapes de pré-traitement sont nécessaires. Celles-ci sont appliquées sur notre corpus d'apprentissage (corpus parallèle source / cible) et / ou sur notre corpus de validation en langue source :

— Étiqueter le côté source du corpus parallèle et le corpus de validation (avec l'étiqueteur supervisé disponible) :

— Construire les représentations vectorielles communes (empreintes distributionnelles) des mots source et cible, à partir du corpus parallèle initial [1].

Ensuite, le réseau de neurones est entraîné sur plusieurs itérations (époques). L'algorithme 2 présenté ci-dessous décrit une époque d'entraînement du réseau.

1. Il est important de noter que si ce corpus parallèle change - par exemple si de nouvelles données sont disponibles - les représentations vectorielles pourront être soient conservées à l'identique soit mises à jour (en augmentant la taille du vecteur) avant le ré-apprentissage du RNN.

Algorithme 2 : Apprentissage d'un analyseur morpho-syntaxique multilingue basé sur un RNN

1 : Initialiser les matrices des poids du réseau avec une distribution normale.
2 : Initialiser le compteur du temps t=0, et initialiser l'état des neurones de la couche cachée h(t) à 1.
3 : Incrémenter le compteur du temps t de 1.
4 : Présenter le vecteur représentant le mot w(t) dans la couche d'entrée.
5 : Recopier l'état de la couche cachée h(t-1) dans la couche d'entrée.
6 : Calculer la valeur de la couche cachée h(t) et de la couche de sortie y(t).
7 : Calculer l'erreur de prédiction $e_0(t) = d(t) - y(t)$ (différence entre la sortie prédite et la sortie attendue).
8 : Mettre à jour les matrices des poids V et U avec l'algorithme de rétropropagation (RP) du gradient de l'erreur (Rumelhart *et al.*, 1985).
9 : Mettre à jour la matrice des poids de récurrence W avec l'algorithme de la rétro-propagation du gradient de l'erreur à travers le temps (RPTT) (Rumelhart *et al.*, 1985).
10 : Si le corpus d'apprentissage comporte encore des exemples, alors revenir à 3.

Les matrices des poids du réseau sont mises à jour en utilisant l'erreur de prédiction pondérée par un pas d'apprentissage α, initialement fixé à 0.1.

Après chaque époque, le corpus de validation est annoté en utilisant le réseau de neurones appris jusque-là. Les sorties sont comparées aux sorties de l'annotateur supervisé, pour calculer le taux d'erreur du réseau. Si le taux d'erreur diminue d'une époque à une autre, le pas d'apprentissage reste inchangé et l'apprentissage continue durant une nouvelle époque. Sinon, le pas d'apprentissage est diminué de moitié au début de la nouvelle époque. Pour éviter un sur-apprentissage des poids du réseau, l'algorithme d'apprentissage est arrêté si le taux d'erreur ne diminue plus durant deux époques successives. Généralement le réseau converge en 5 à 10 époques.

La deuxième étape de notre approche consiste simplement à utiliser le modèle entraîné sur le côté source comme annotateur morpho-syntaxique pour la langue cible, via l'utilisation de la représentation vectorielle commune. Il est important de noter que si l'on dispose d'un corpus parallèle en N langues (au lieu de 2), un même réseau RNN pourra étiqueter toutes ces langues sans être re-entraîné. On dispose donc d'un véritable étiqueteur multilingue.

4 Expérimentations et Résultats

4.1 Corpus et outils

Initialement, nous avons expérimenté notre approche sur le couple de langues anglais-français, où le français est considéré comme langue cible. Le français n'est certainement pas une langue faiblement dotée, mais le fait qu'il dispose d'un annotateur morpho-syntaxique supervisé (TreeTagger (Schmid, 1995)), nous a permis de construire une *pseudo vérité terrain* (sur le corpus de test) pour évaluer notre approche. Nous avons utilisé un corpus d'apprentissage de 10000 bi-phrases, extrait du corpus parallèle (anglais-français) ARCADEII (Véronis *et al.*, 2008), dont le côté source a été annoté par l'outil *TreeTagger* (Schmid, 1995) pour l'anglais. Notre corpus de validation (en anglais - pour le règlage du RNN) contient 1000 phrases (non présentes dans le corpus d'apprentissage), et est aussi extrait du corpus ARCADEII puis annoté par le toolkit *TreeTagger* pour l'anglais. Nous avons construit notre corpus test (français) à partir de 1000 phrases extraites du corpus ARCADEII, et annoté par le toolkit *TreeTagger* pour le français, puis corrigées manuellement.

Ayant obtenu des résultats intéressants sur le couple de langues anglais-français, nous nous sommes ensuite intéressés à la généralisation de notre approche sur d'autres langues : l'allemand, le grec et l'espagnol. Afin de pouvoir rendre nos résultats comparables avec ceux de (Das & Petrov, 2011) et (Duong *et al.*, 2013), nous suivons leur protocole : nous partons de l'anglais comme langue source et utilisons un corpus parallèle et un corpus de validation (anglais) extraits d'Europarl (Koehn, 2005). Nous évaluons les résultats de nos approches sur les mêmes corpus de test, qui sont ceux des campagnes d'évaluation d'analyse en dépendances CoNLL (Buchholz & Marsi, 2006). Ces corpus ont été annotés manuellement par des experts linguistes. Nous utilisons aussi la même métrique d'évaluation (le taux d'erreur d'étiquetage) et le même jeu d'étiquettes (*Universal Tagset* (Petrov *et al.*, 2012)).

Afin de construire nos modèles par projection simple (Algorithme 1), la partie cible des corpus d'apprentissage est étiquetée par projection des annotations du côté source (annoté par le toolkit TreeTagger pour l'anglais) en utilisant les alignements obtenus par GIZA++. Par souci d'uniformité, nous avons aussi transformé les étiquettes morpho-syntaxiques

fines (de TreeTagger et de CoNLL) en leurs équivalents dans le jeu étiquettes universelles via les règles de (Petrov *et al.*, 2012).

Pour implémenter notre approche (décrite dan l'Algorithme 2), nous avons adapté l'outil *Recurrent Neural Network Language Modeling Toolkit* (RNNLM) fourni par (Mikolov *et al.*, 2011), pour apprendre et tester notre annotateur morpho-syntaxique neuronal [2].

Pour tirer parti des avantages de chacun de ces deux modèles *M1* (Projection Simple) et *M2* (RNN), il est intéressant d'étudier un moyen de les combiner. Le mot *w* est annoté avec l'étiquette t_w la plus probable, en utilisant la fonction *f* donnée par l'équation ci-dessous :

$$f(w) = \arg\max_t(\mu P_{M1}(t|w, C_{M1}) + (1-\mu)P_{M2}(t|w, C_{M2})) \quad (2)$$

Où, C_{M1} et C_{M2} sont, respectivement les contextes de *w* considérés par *M1* et *M2*. Le paramètre d'interpolation μ (importance de chaque modèle) est ajusté par validation croisée sur le corpus de test.

4.2 Résultats et discussion

Les résultats obtenus par notre approche sont résumés dans le tableau 1. Les scores obtenus par (Das & Petrov, 2011) et (Duong et al., 2013) sont également inclus lorsque ceux-ci sont disponibles sur le même corpus de test

Modèle	français		allemand		grec		espagnol	
	Tous mots	OOV	Tous mots	OOV	Tous mots	OOV	Tous mots	OOV
Projection Simple	80.3 %	77.1 %	78.9%	73%	77.5%	72.8%	80%	79.7%
RNN-640-160	78.5 %	70 %	76.1%	76.4%	75.7%	70.7%	78.8%	72.6%
Projection+RNN	**84.5%**	**78.8%**	81.5 %	**77%**	78.3%	**74.6%**	83.6%	**81.2%**
(Das, 2011)	na	na	82.8%	na	**82.5%**	na	**84.2%**	na
(Duong, 2013)	na	na	**85.4%**	na	80.4%	na	83.3%	na

TABLE 1 – Performances en taux d'erreur d'étiquetage (Projection Simple, RNN et Projection+RNN) - et comparaison avec Das & Petrov (2011) et Duong et al (2013).

Nous avons évalué plusieurs topologies de réseaux de neurones récurrents, avec une ou deux couches cachées, et avec différentes tailles. Les meilleures performances obtenues sont celles des annotateurs basés sur des réseaux de neurones à deux couches cachées, contenant respectivement 640 et 160 neurones (RNN-640-160). Ces performances sont proches des annotateurs par projection simple, les différences proviennent de la gestion des mots inconnus (OOV) qui est pour l'instant quasi-inexistante dans notre approche. En effet, les représentations vectorielles des OOV sont nulles, et pour les annoter, le réseau de neurones n'utilise que l'information de récurrence (étiquette précédente prédite) ce qui est une information insuffisante pour une bonne annotation. Une perspective à très court terme consistera à traiter le cas des mots inconnus dans notre RNN.

En attendant, nous avons également combiné l'approche classique avec notre méthode par réseaux récurrents. Dans l'ensemble, les résultats expérimentaux de notre combinaison (Projection+RNN) sont proches de ceux de l'état de l'art des annotateurs morpho-syntaxiques non supervisés (Das & Petrov, 2011; Duong *et al.*, 2013) et montre une bonne complémentarité entre projection simple et RNN. Toutefois, nos résultats sont légèrement inférieurs aux résultats de référence, et une meilleure gestion des OOV, ainsi que l'utilisation d'une faible quantité de données cible annotée pour *adapter* le réseau, semblent des perspectives intéressantes à court terme.

5 Conclusion

Dans cet article, nous avons présenté une approche utilisant les réseaux de neurones récurrents comme annotateurs morpho-syntaxiques multilingues (non supervisés pour les langues cibles). Cette approche n'a besoin que d'un corpus parallèle et d'un annotateur morpho-syntaxique pré-existant en langue source. Bien que nos résultats initiaux soient positifs, ils doivent être améliorés. Dans nos futurs travaux, nous envisageons donc d'utiliser une meilleure représentation pour les OOV. Par ailleurs, nous envisageons d'utiliser une technique similaire pour des tâches plus complexes du TALN (par exemple annotation en sens, en entités nommées et en rôles sémantiques).

2. L'adaptation du RNNLM est disponible à l'url `https://github.com/othman-zennaki/RNN_POS_Tagger.git`

Références

ABDULHAY A. (2012). *Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné*. PhD thesis, Université de Grenoble.

ALLAUZEN A. & WISNIEWSKI G. (2009). Modèles discriminants pour l'alignement mot à mot. *Traitement Automatique des Langues*, **50**(3), 173–203.

BENGIO Y., SCHWENK H., SENÉCAL J.-S., MORIN F. & GAUVAIN J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, p. 137–186. Springer.

BENTIVOGLI L., FORNER P. & PIANTA E. (2004). Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 364 : Association for Computational Linguistics.

BRANTS T. (2000). Tnt : a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, p. 224–231 : Association for Computational Linguistics.

BUCHHOLZ S. & MARSI E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, p. 149–164 : Association for Computational Linguistics.

DAS D. & PETROV S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, p. 600–609 : Association for Computational Linguistics.

DEMPSTER A. P., LAIRD N. M. & RUBIN D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, p. 1–38.

DUONG L., COOK P., BIRD S. & PECINA P. (2013). Simpler unsupervised pos tagging with bilingual projections. In *ACL (2)*, p. 634–639.

ELMAN J. L. (1990). Finding structure in time. *Cognitive science*, **14**(2), 179–211.

FRASER A. & MARCU D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, **33**(3), 293–303.

GARSIDE R., LEECH G. N. & MCENERY T. (1997). *Corpus annotation : linguistic information from computer text corpora*. Taylor & Francis.

JABAIAN B., BESACIER L. & LEFEVRE F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transactions on Audio, Speech & Language Processing*, **21**(3), 636–648. (Impact-F 1.67 estim. in 2012).

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, p. 79–86.

MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKỲ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, p. 1045–1048.

MIKOLOV T., KOMBRINK S., DEORAS A., BURGET L. & CERNOCKY J. (2011). Rnnlm-recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, p. 196–201.

OCH F. J. & NEY H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 440–447 : Association for Computational Linguistics.

PADÓ S. & LAPATA M. (2005). Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 859–866 : Association for Computational Linguistics.

PADÓ S. & LAPATA M. (2006). Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 1161–1168 : Association for Computational Linguistics.

PADO S. & PITEL G. (2007). Annotation précise du français en sémantique de rôles par projection cross-linguistique. *Proceedings of TALN-07, Toulouse, France*.

PETROV S., DAS D. & MCDONALD R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

RUMELHART D. E., HINTON G. E. & WILLIAMS R. J. (1985). *Learning internal representations by error propagation*. Rapport interne, DTIC Document.

SCHMID H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop* : Citeseer.

SUNDERMEYER M., OPARIN I., GAUVAIN J.-L., FREIBERG B., SCHLUTER R. & NEY H. (2013). Comparison of feedforward and recurrent neural network language models. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 8430–8434 : IEEE.

TÄCKSTRÖM O., DAS D., PETROV S., MCDONALD R. & NIVRE J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, **1**, 1–12.

VAN DER PLAS L. & APIDIANAKI M. (2014). Cross-lingual word sense disambiguation for predicate labelling of french. *Proceedings of TALN-14, Marseille, France*, p.46̃.

VÉRONIS J. (2000). *Chapitre 4*, In *Annotation automatique de corpus : panorama et état de la technique*. Editions Hermés.

VÉRONIS J., HAMON O., AYACHE C., BELMOUHOUB R., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F. & WAJDI Z. (2008). *Chapitre 2*, In *ArcadeII Action de recherche concertée sur l'alignement de documents et son évaluation*. Editions Hermés.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014). Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue. *Proceedings of TALN-14, Marseille, France*, p. 173.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.