PALM : Un modèle neuronal pour l'étiquetage morphosyntaxique des textes médiévaux

Aude Mairey [1], Mourad Aouini[2]

[1] University of Paris 1 La Sorbonne – aude.mairey@univ-paris1.fr

[2]CNRS – mourad.aouini@cnrs.fr

Abstract

PALM is an online system, which allows executing linguistic analysis on medieval texts in three languages, Middle French, Middle English and Medieval Latin. In this paper, we wish to show the implementation and possibilities offered by PALM, as well as to evoke future prospects in order to improve the system, especially the passage from N-gram sequences to a system based on neuronal network concerning Part-of-Speech tagging and lemmatization. In this context, we will present a recurrent neuronal network (RNN) GRU bidirectional for POS tagging of the three medieval languages treated by PALM. The first results are encouraging. This type of architecture brings up promising answers for the analysis of non-uniformed ancient languages.

Keywords: PALM, POS tagging, RNN, GRU

Résumé

La plateforme PALM est un système en ligne qui permet de procéder à des analyses linguistiques aux textes médiévaux en trois langues moyen français, moyen anglais et latin médiéval.
Cette communication se propose de présenter la mise en œuvre et les possibilités offertes par PALM, ainsi que d'évoquer les perspectives futures en cours de réflexion afin d'améliorer le système, en particulier le passage d'un modèle d'analyse des séquences N-gram à un système fondé sur les réseaux de neurones pour l'étiquetage morphosyntaxique et la lemmatisation. Dans ce contexte, nous présentons une architecture neuronale récurrente (RNN) GRU bidirectionnelles pour l'étiquetage morphosyntaxique de trois langues médiévales moyen français, moyen anglais et latin médiéval. Les premiers résultats sont très encourageants. Ce type d'architecture apporte des solutions prometteuses pour l'analyse des langues anciennes non uniformisées.

Mots clés : PALM, étiquetage morphosyntaxique, RNN, GRU

1. Introduction

La plateforme PALM[1] est un système en ligne qui permet de conduire des analyses linguistiques appliquées à des textes médiévaux en trois langues, le moyen français, le moyen anglais et le latin médiéval. Elle a été développée dans le cadre d'un programme ERC dirigé par Jean-Philippe Genet entre 2010 et 2014, intitulé *Signs and States* (Genet 2015). Ce programme, dont PALM n'est que l'une des multiples facettes, a eu pour objectif d'examiner les différentes manières par lesquelles le développement et l'application de nouvelles formes de gouvernance et de pouvoir étatique à la fin du Moyen Âge ont transformé les sociétés européennes, non

[1] http://palm.huma-num.fr/PALM/.

seulement sur les plans sociaux et économiques, mais aussi sur le plan culturel. Il s'est concentré sur le pouvoir symbolique.

PALM est à la fois constitué par une bibliothèque de textes numérisés et par un ensemble d'outils d'annotation semi-automatique de corpus. Les origines de la bibliothèque remontent à MEDITEXT, un projet collectif au long cours initié dans les années 1980 par Jean-Philippe Genet et Claude Gauvard et dont l'objet était de constituer une bibliothèque numérique de textes politiques, au sens le plus large du terme et en tenant compte du fait que l'usage de cette notion de politique est parfois contestée pour le Moyen Âge (Mairey 2009). Le corpus est pour l'heure de documents français et anglais, la palette des genres étant très vaste – traités, poèmes, sermons, etc., sur une période allant du XII^{e} au XVI^{e} siècle (pour une présentation détaillée, voir Genet *et al*. 2012). Il a été largement enrichie lors du programme ERC et l'est encore ; il comprend à l'heure actuelle des milliers de textes médiévaux accessibles dans les trois langues (la liste complète est disponible sur le site).

Notre plateforme permet d'accompagner les utilisateurs pour trouver, partager et (ré)utiliser les textes bruts et/ou annotés du MEDITEXT en essayant d'implémenter progressivement les principes FAIR sur la gestion des données de recherches.

Les outils d'annotation de PALM permettent d'enrichir les textes par des annotations grammaticales et sémantiques afin d'assister le chercheur dans des tâches allant de la simple recherche lexicale, portant sur les concordances par exemple, à l'usage d'outils sémantiques plus sophistiqués, par exemple le développement des grammaires ou des expressions rationnelles pour l'identification de collocations et d'entités nommées, jusqu'à à des méthodes d'exploration statistique multidimensionnelle comme l'analyse factorielle par correspondance.

2. Pourquoi PALM ?

La version initiale a été conçue pour traiter des textes politiques d'origine française et anglaise, en latin médiéval, moyen français ou moyen anglais, sur une période allant du XII^{e} au début du XVI^{e} siècle, tout en laissant ouverte la possibilité d'ajouter par la suite d'autres thèmes, d'autres langues et d'autres régions. Avant PALM, il était particulièrement difficile de préparer un grand corpus annoté de cette période à cause de l'absence d'une graphie normalisée ainsi que, surtout dans le cas du moyen anglais, de la nature toujours changeante de sa grammaire et de sa syntaxe. De manière plus spécifique, PALM est dotée d'interfaces graphiques intuitives pour constituer un corpus annoté dans l'une des trois langues, par des parties du discours, des lemmes, des traits sémantiques et des entités nommées tout en respectant le standard XML-TEI et le jeu de caractères universel Unicode. Elle offre également des facilités pour la correction manuelle des annotations produites automatiquement, inévitable pour des textes médiévaux, avec des outils qui permettent de modifier les résultats des analyses. L'annotation et la lemmatisation de corpus médiévaux sont de fait indispensables si l'on veut pouvoir effectuer des analyses textométriques sur des textes de cette période (Mairey 2011).

Les utilisateurs (qui doivent simplement ouvrir un compte pour avoir accès à la plateforme) ont donc généralement pour objectif de constituer des corpus annotés afin d'explorer les textes via les outils de textométrie comme Hyperbase, TXM ou

Lexico3. L'annotation offerte par PALM est importante pour lancer des opérations textométriques ou de simple recherche. Par exemple, il est possible de recenser toutes les flexions d'un verbe, de trouver des fonctions et des titres juridiques et militaires grâce à des traits sémantiques ou d'identifier les cooccurrences associées à un lieu ou un personnage.

L'annotation de PALM affiche des bonnes performances pour extraire les concordances et les cooccurrences des formes permettant ainsi d'étudier l'uniformisation et la variance de la graphie, pour retrouver les structures *répétées* qui commencent à se stabiliser et pour étudier les contextes d'apparition des formes, des parties de discours ou un ensemble de formes qui partagent des traits sémantiques en commun.

3. Jeux d'étiquettes

Un jeu d'étiquettes (*tagset*) constitue un formalisme qui permet de décrire les propriétés morphosyntaxiques d'une unité lexicale dans son contexte d'énonciation. La définition d'un jeu d'étiquettes adéquat est indispensable pour assurer une meilleure performance de l'étiqueteur morphosyntaxique.

La définition des jeux d'étiquettes des trois langues est le résultat d'une phase expérimentale sur notre corpus Meditext durant laquelle des annotations manuelles et des mesures de performances ont été effectuées.

Par ailleurs, l'utilisation de ressources linguistiques diverses pour une seule langue qui n'ont pas le même jeu d'étiquettes comme l'*Anglo-Normand Dictionary* (Bennett, 2007) et le *Dictionnaire du moyen français (DMF)* (Martin & Bazin-Tacchella, 2012), nous a conduit à mettre en place un jeu d'étiquettes simple qui supporte des formalismes différents. Nous avons dressé des jeux d'étiquettes composées d'environ 16 catégories morphosyntaxiques pour chaque langue dont 9 communes aux trois langues (Tableau 1).

Cette description des catégories, qui se veut consensuelle, a l'avantage d'améliorer considérablement la performance d'étiquetage. Mais, nous avons tenté autant que possible d'adapter les parties du discours à chaque langue, d'où la présence, par exemple pour l'anglais, des catégories comme « verbe+nom » et « nom verbal » (*giving*, *making*…).

Abréviation	Description
ADV	Adverbe
A	Adjectif
Npropre	nom propre
Ncommun	nom commun
PREP	préposition
PRO	Pronom
V	Verbe

INTJ	interjection
PUNC	Ponctuation

Tableau 1. Catégories morphosyntaxiques communes aux trois langues : moyen français, moyen anglais et latin médiéval.

Dans un avenir proche, nous souhaitons enrichir ces étiquettes par des informations morphologiques comme, pour le français par exemple, le genre et le nombre pour les noms communs et les adjectifs ainsi que la personne, le genre, le nombre, le mode et la voix pour les verbes.

4. L'étiquetage morphosyntaxique des textes médiévaux

L'étiquetage morphosyntaxique permet une catégorisation des unités lexicales, qui structurent le texte en parties du discours. Cette tâche implique la résolution de difficultés liées à l'ambiguïté posée par les unités lexicales. Les méthodes d'étiquetage morphosyntaxique affichent des performances élevées pour les langues standardisées, y compris pour des textes issus des réseaux sociaux (Meftah et al., 2018).

L'évaluation des performances d'un étiqueteur est une tâche complexe car elle dépend de plusieurs facteurs liés généralement à l'état de la langue, à la complexité du jeu d'étiquettes utilisé et au choix du corpus de test (Véronis et al., 1995). Pour les langues standardisées, les étiqueteurs sont généralement efficaces quelle que soit la technologie et la méthode d'évaluation adoptées. Pour ces langues, en effet, un bon nombre d'unités lexicales ne sont pas ambigües et une grande part de l'ambiguïté est détenue par un petit nombre de mots fréquents qui sont généralement des mots grammaticaux (Tzourkermann et al., 1996 ; Véronis, 2000). De ce fait, il est possible d'atteindre des bons résultats dépassant largement 90 % en utilisant des systèmes d'analyse de séquences qui permettent de traiter les unités lexicales fréquentes et ambiguës.

Pour les langues non uniformisées de la période médiévale, les résultats de ces méthodes sont moins spectaculaires (Aouini, 2018), à cause de l'instabilité des phénomènes linguistiques et la présence des traits régionaux. En effet, le moyen français et le moyen anglais, plus encore que le latin médiéval, sont, on l'a vu, des langues en pleine évolution dont la graphie, le système flexionnel et la syntaxe ne sont pas stables. Elles se singularisent principalement par une importante variation graphique. Comme l'ont souligné Souvay & Pierrel (2009) pour le moyen français, « on se rend bien compte qu'il est difficile voire impossible d'établir une liste exhaustive des formes possibles pour une entrée lexicale ».

Plusieurs facteurs peuvent expliquer cette forte variation graphique. Le contact avec le latin et les différentes langues régionales a largement influencé la graphie française et anglaise. Par exemple, la forme *chiel* est utilisée pour *ciel* dans les textes picards et la forme *bastoun* pour *baston* dans les textes anglo-normands. Le lexique est donc caractérisé par une variabilité tant géographique que chronologique.

Le système flexionnel reste complexe même si on constate une tendance à la stabilisation. Pour le moyen français, par exemple, la conjugaison des verbes du

premier groupe tend à s'aligner sur un seul paradigme flexionnel pour chaque temps avec un emploi régulier de certains suffixes. Mais les alternances de bases verbales, en particulier pour le présent de l'indicatif, et l'utilisation des formes qui datent de l'ancien français (XII^e^-XIII^e^ siècles) n'est pas rare. Autre exemple, l'emploi courant de S, X ou Z se généralise pour le pluriel des substantifs tels que *ciels*, *cielz* et *cielx* pour *ciel.* Mais, cette régularité de la formation des substantifs apparue au XIV^e^ siècle et qui a continué son évolution au XV^e^ siècle, nous a permis de distinguer la présence de plusieurs paradigmes irréguliers pour des substantifs (Aouini, 2018).

Comme le français et l'anglais sont des langues parlées, plusieurs variantes sont dues au système phonétique. En effet, les phonèmes peuvent changer de graphies selon le contexte sonore ou selon leur position dans le mot. Par exemple, ils sont marqués par une réduction des hiatus, tels OU au lieu d'AOU ou OI au lieu d'ËOI. Afin de s'approcher de la prononciation des locuteurs, nous constatons également l'introduction de la cédille pour distinguer la lettre C prononcée K de celle C prononcée S et l'introduction des accents tels « *à* », « *â* », « *ê* », « *ô* » pour distinguer d'autres prononciations.

Finalement, les langues vernaculaires médiévales sont marquées par une nette évolution de la syntaxe. En effet, l'ordre des mots commence à se stabiliser en sujet suivi d'un verbe, lui-même suivi d'un complément. Cependant, les séquences verbe-sujet continuent d'exister et l'emploi des prépositions et des conjonctions est en nette augmentation au fil du temps rendant les phrases de plus en plus longues et complexes.

Ces aspects que nous venons d'exposer conduisent à considérer l'analyse morphosyntaxique de ces langues, en particulier le moyen anglais et le moyen français, comme une tâche complexe et non triviale. La distribution des probabilités d'apparition des formes dans un corpus en moyen français ou en moyen anglais est plus complexe que celle d'une langue contemporaine qui contient un nombre fini de formes fréquentes et ambigües.

Ce constat est confirmé par les travaux d'étiquetage morphosyntaxique sur des corpus en moyen français et en moyen anglais. Un modèle d'analyse de séquence fondé sur des arbres de décisions Treetagger (Schmid, 1994) a été entraîné sur un ensemble de textes de la Base de Français Médiéval (BFM) (Guillot-Barbance *et al*., 2017) qui couvre la période entre le IX^e^ et la fin du XV^e^ siècle. En utilisant le jeu d'étiquettes « Cattex 2009 », ce modèle affiche des taux de performance inférieurs à 75% pour les étiquettes peu fréquentes (Guillot *et al*., 2015).

Nous avons entraîné en utilisant le corpus Meditext trois modèles Treetagger pour chacune des langues. Après plusieurs tests et ajustement de paramètres d'entraînement, Treetagger affiche un taux de performance de 87.41 % pour le moyen français et environ 84.78% pour le moyen anglais.

Les modèles Treetagger du latin médiéval affichent des performances proches des langues contemporaines entre 92% et 98%. Nous faisons référence au modèle entraîné dans le cadre de l'ANR Omnia (Outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins, voir Bon 2011) menée entre 2008 et 2012 et celui entraîné sur MEDITEXT (97.95%).

En utilisant le jeu d'étiquettes très fines du Laboratoire d'analyse statistique des langues anciennes (LASLA), l'évaluation des performances de trois étiqueteurs morphosyntaxiques MBT, TnT et TreeTagger sur des textes latins classiques n'a pas dépassé les 89% (Poudat et Longree, 2009). Un des objectifs de cet article est de présenter les travaux réalisés récemment pour améliorer les performances de notre système d'étiquetage morphosyntaxique. De ce fait, les deux derniers points de cette communication portent sur la mise en place et l'évaluation d'un système basé sur une architecture neuronale en utilisant les réseaux neuronaux récurrents RNN.

5. Un modèle Bi-GRU RNN pour l'étiquetage morphosyntaxique

L'étiquetage morphosyntaxique est une tâche d'analyse de séquences. Il consiste à prédire un ensemble d'étiquettes $Y=\{y_1, y_2,..., y_m\}$ à partir d'un ensemble d'éléments $X=\{x_1, x_2,..., x_n\}$ avec m = n. De ce fait, pour chaque élément de la séquence x_i, le modèle prédit une catégorie y_i. En étiquetage morphosyntaxique, le couple (x_i,y_i) correspond donc respectivement au mot et à la catégorie morphosyntaxique.

Les algorithmes d'apprentissage supervisé, qui apprennent à partir d'un corpus d'entraînement annoté manuellement à classer les mots selon un jeu d'étiquettes, ont été plus étudiés que les méthodes de modélisation statistiques ou les méthodes à base de règles pour résoudre l'étiquetage morphosyntaxique. Parmi ces algorithmes, citons les chaînes de Markov (Merialdo, 1994), les arbres de décisions (Schmid, 1994), l'entropie maximale (Ratnaparkhi, 1996) et les champs aléatoires conditionnels (CRFs) (Lafferty et al., 2001). Le principal inconvénient de ces méthodes dites de type n-gram est qu'elles supportent des séquences de mots X de petite taille n généralement inférieure à 5 ($2<n<5$).

Les architectures des réseaux neuronaux profonds (RNP) peuvent pallier cet inconvénient en analysant des séquences de taille importante. Ces architectures apportent une précision supérieure aux méthodes n-gram (Wang et al., 2015, Plank et al., 2016). En outre, elles peuvent s'entraîner sur des corpus annotés volumineux pour produire des modèles puissants sans risque de sur-apprentissage.

Les réseaux neuronaux récurrents (RNN) (Elman, 1990) sont les RNP les plus employés pour l'analyse des séquences en général et pour l'étiquetage morphosyntaxique en particulier. Ce type d'architecture neuronale permet une distribution de probabilité sur un jeu d'étiquettes prédéfini, de taille fixe, en traitant des séquences de mots de taille variable. En effet, les RNN sont composés d'un état caché pour chaque pas-de-temps i. Chaque état reçoit la sortie de l'état caché précédent h_{i-1} qui sera multiplié par une matrice W_{hh} et une représentation de mot x_i multiplie par une matrice W_{hx}. Finalement, la couche décisionnelle permet une distribution sur le jeu d'étiquettes afin de prédire la catégorie ayant la probabilité la plus élevée.

Bien que les RNN puissent théoriquement capturer des dépendances à long terme, leurs implémentations montrent une limite majeure : la perte d'information. En effet, le problème est qu'il est difficile pour un RNN d'apprendre à conserver les informations sur de nombreux pas-de-temps (Bengio et al., 1994; Pascanu et al., 2013a). En traitant des longues séquences de mots, les matrices de poids ne se propagent pas parfaitement d'un pas-de-temps à un autre et donc l'accès à des pas-de-

temps à partir d'une position i lointaine devient difficile. Ce problème est dû à la disparition du gradient lors de la phase de rétro-propagation. En effet, lors de l'apprentissage, les valeurs de gradient disparaissent progressivement au fur et à mesure qu'elles se propagent aux pas-de-temps antérieurs.

Pour remédier à ce problème, plusieurs variantes de RNN ont été proposées comme LSTM (Long short-term memory, en français réseau récurrent à mémoire court et long terme) (Hochreiter & Schmidhuber, 1997) et GRU (Gated recurrent units, en français réseau récurrent à portes) (Cho et al., 2014) qui sont les plus utilisés. Ces deux architectures neuronales reposent sur l'utilisation d'unités d'activation complexe qui sont conçues de manière à avoir une mémoire persistante. Cette sorte de mémoire séparée permet aux RNN de capturer plus facilement les dépendances à long terme.

Nous avons mis en place une architecture GRU bidirectionnel (figure 1) pour l'étiquetage morphosyntaxique. Cette architecture GRU assure la propagation d'informations contextuelles sur des pas-de-temps lointains. Elle est aussi bidirectionnelle : le réseau analyse des mots du contexte gauche et des mots du contexte droit afin de prédire la catégorie morphosyntaxique d'un mot central.

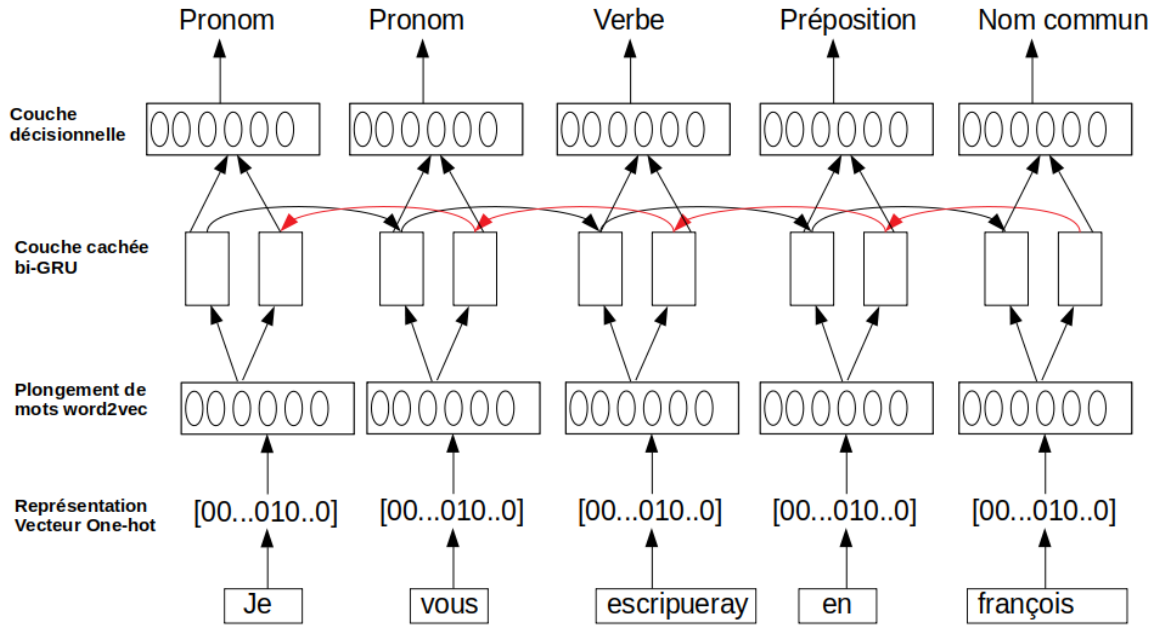

Figure 1. Architecture Bi-GRU RNN pour l'étiquetage morphosyntaxique

Notre modèle exploite donc deux couches GRU, une pour la propagation de gauche à droite et une autre pour la propagation de droite à gauche.

$$\overrightarrow{h_i} = f(\overrightarrow{W_{xh}}\, x_i + \overrightarrow{W_h}\, \overrightarrow{h_{i-1}} + \overrightarrow{b})$$

$$\overleftarrow{h_i} = f(\overleftarrow{W_{xh}}\, x_i + \overleftarrow{W_h}\, \overleftarrow{h_{i+1}} + \overleftarrow{b})$$

Le résultat final de la classification, $\hat{y}_i$, est généré en combinant les résultats des scores produits par les deux couches GRU cachées.

$$\hat{y}_i = g(U[\overrightarrow{h_i} + \overleftarrow{h_i}]+c)$$

Les entrées de notre modèle neuronal d'étiquetage morphosyntaxique sont les plongements des mots qui sont des représentations des mots par des vecteurs denses, permettant d'encoder des notions de similarité et de différence, dans un espace multidimensionnel, généralement entre 300 et 1000 dimensions. Nous n'avons pas initialisé la couche d'entrée par des plongements des mots pré-entraînées pour les trois langues moyen-français, moyen anglais et latin médiéval à cause de la non-disponibilité de corpus suffisamment volumineux pour ces langues.

6. Expérimentation et évaluation

Notre modèle Bi-GRU RNN a été implémenté avec Pytorch. Il est composé simplement de trois couches : une couche d'*embeddings*, une couche GRU bidirectionnelle et une couche de décision. Cette couche de décision sera appliquée à chaque position de la phrase pour prédire l'étiquette associée. Nous avons effectué plusieurs tests sur les trois langues de corpus en modifiant la taille de l'état caché et la longueur maximale d'une séquence de mots en apprentissage selon la performance du modèle. Ces deux mesures ont été fixées respectivement à 128 et à 16. La taille de la couche d'*embeddings* a été fixée à 300. Lors de la phase de rétro-propagation, le critère d'entropie croisée (*cross-entropy loss*) est utilisé comme la fonction *loss* qui permet de calculer la performance dans le but d'ajuster les poids du modèle. L'entraînement a été effectué avec des *batches* de taille 64 et un taux d'apprentissage (*Learning rate*) constant à 0.01. Rappelons que la couche d'*embeddings* n'a pas été initialisée par des plongements pré-entraînés.

Comme le montre le tableau 2, nous avons évalué notre modèle en utilisant un corpus de test pour chacune des trois langues. Il a été comparé au Treetagger, entraîné et évalué sur les mêmes corpus d'entraînement et de test et sur les mêmes jeux d'étiquettes. Notre modèle est plus performant que Treetagger pour des langues non uniformisées comme le moyen français et le moyen anglais et il affiche un taux de performance légèrement supérieure pour le latin médiéval.

Langues	Bi-GRU RNN	Treetagger
Moyen français	95.67 %	87.41 %
Moyen anglais	96.14 %	84.78 %
Latin médiéval	98.02 %	97.95%

Tableau 2. Tableau comparatif des F-mesure (F1-score) s entre Bi-GRU RNN et Treetagger

Il faut toutefois interpréter ces résultats avec précaution. En effet, pendant la phase d'apprentissage, nous parcourons l'ensemble du corpus plusieurs fois (80 à 100 epochs selon le corpus) afin d'ajuster les poids de notre modèle Bi-GRU RNN. Par

ailleurs, le corpus annoté en moyen anglais est plus volumineux que le corpus en moyen français ce qui explique le score élevé de notre modèle pour le moyen anglais. Treetagger de son côté analyse des séquences de caractères pour identifier les affixes et utilise des ressources externes, à savoir un dictionnaire contenant le lexique de la langue. Cela n'est pas le cas de notre modèle qui pourrait être encore plus compétitif en utilisant des représentations des mots et de caractères pré-entraînées et des ressources lexicales. De ce fait, nous considérons que les résultats de notre modèle sont plus qu'encourageants pour analyser des langues médiévales non normalisées comme le moyen français et le moyen anglais.

7. Conclusion

La proposition explicitée dans cette communication d'une architecture Bi-GRU RNN pour l'étiquetage morphosyntaxique des textes médiévaux en moyen français, en moyen anglais et en latin médiéval, a donné des premiers résultats très encourageants.

Mais notre modèle neuronal d'étiquetage morphosyntaxique est conçu pour évoluer. Plusieurs améliorations pourraient être apportées à notre architecture, tels l'utilisation des plongements de mots et de caractères et l'introduction des multicouches, des multi-classes (Plank et al., 2016) ou des lexiques (Sagot et Matinez Alonso, 2017). Pour conclure, les architectures RNN bidirectionnelles comme notre modèle Bi-GRU RNN apportent une solution efficace pour l'étiquetage morphosyntaxique des langues non uniformisées et méritent une exploration plus approfondie pour d'autres problématiques.

References

Aouini M. (2018). Approche multi-niveaux pour l'analyse des données textuelles non-standardisées: corpus de textes en moyen français. Thèse de doctorat. Franche-comté.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2): 157–166.

Bennet Ph. (2007). Anglo-Norman Dictionary. *The Modern Language Review*, 2007, 102(2): 500-503.

Bon B. (2011). OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3), BUCEMA, 11, https://doi.org/10.4000/cem.12015.

Cho K., Van Merrienboer B., Gulcehre C., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Empiricial Methods in Natural Language Processing* (EMNLP 2014).

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

Genet J.-P. *et al*. (2012). Une plateforme internet pour l'analyse linguistique de textes médiévaux (PALM) – Un corpus de textes politiques d'origine anglaise et française de la fin du Moyen Âge (Méditext), http://archive-2013-2016.lamop.fr/IMG/pdf/PALM-Meditext.pdf.

Genet J.-P. (2015). Pouvoir symbolique, légitimation et genèse de l'État moderne. in *ibid* editor, *La légitimité implicite*, vol. 1, Publications de la Sorbonne, pp. 9-47.

Guillot C., Heiden S., Lavrentiev A. & Pincemin B. (2015). L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique. In Jeppesen Kragh K. & Lindschouw J. editors, *Les variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du Colloque DIA II à Copenhague (19-21 nov. 2012)*, pp. 15-27.

Guillot-Barbance C., Heiden S. and Lavrentiev A. (2017). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques*, 7: 168-184.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Lafferty J. D., McCallum A. and Pereira F. C. N. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In International Conference on Machine Learning (*ICML*, pp. 282–289.

Leemans, H. (2018). *Traitement automatique du moyen français : analyse de données, étiquetage morphosyntaxique et désambiguïsation contextuelle.* Faculté de philosophie, arts et lettres, Université atholique de Louvain.

Mairey A. (2009). Les langages politiques – Introduction, *Médiévales* 57 : 5-14.

Mairey A. (2011), Quelles perspectives pour la textométrie des états des langues passées ?. In Genet J. P., Zorzi A. editors, *Les historiens et l'informatique : un métier à réinventer*, pp. 157-170

Marchello-Nizia C. (2005) *La langue française aux XIVe et XVe siècles*. Armand Colin.

Martin R. and Bazin-Tacchella S. (2012). *Dictionnaire du moyen français*. (DMF2012).

Meftah S. and Semmar N. (2018). A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018).

Merialdo B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20(2):155–171.

Mikolov T. Sutskever I., Chen K. *et al.* (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)* pp. 3111-3119.

Pascanu R., Mikolov, T., and Bengio, Y. (2013). *On the difficulty of training recurrent neural networks*. In International Conference on Machine Learning (ICML) .

Plank B., Søgaard A. and Goldberg Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proc. of the 54th Annual Meeting of the ACL. Berlin, Germany*.

Poudat, C. and Longrée D. (2009). Variations langagières et annotation morphosyntaxique du latin classique. *Traitement Automatique des Langues*, 50.2: 129–148.

Ratnaparkhi A. (1996). *A maximum entropy model for part-of-speech tagging*. In In Conference on Empirical Methods in Natural Language Processing (*EMNLP)*, 1: 133–142.

Sagot B. and Martínez Alonso H. (2017). Improving neural tagging with lexical information. In *Proceedings of the 15th International Conference on Parsing Technologies. Association for Computational Linguistics*, pp. 25–31. http://www.aclweb.org/anthology/W17-63 4.

Schmid H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In *Proceedings of International Conference on New Methods in Language Processing*.

Souvay G. and Pierrel J. M. (2009). LGeRM Lemmatisation des mots en moyen français. *Traitement Automatique des Langues*, 50(2), p. 149-172.

Tzoukermann E. and Radev D. R. (1996). Using word class for part-of-speech disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pp. 1-13.

Véronis J. (2000). Annotation automatique de corpus: panorama et état de la technique. *Ingénierie des langues*, 4.

Véronis J. and Khouri L. (1995). Étiquetage grammatical multilingue: le projet MULTEXT. *TAL. Traitement automatique des langues*, 36(1-2): 233-248.

Wang P., Qian Y., Soong F. K., He L. and Zhao H.. 2015. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. pre-print, abs/1510.06168.

.