ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables

Laurent Vanni[1], Damon Mayaffre[2], Dominique Longree[3]

[1] UMR 7320 : Bases, Corpus, Langage - laurent.vanni@unice.fr

[2] UMR 7320 : Bases, Corpus, Langage - damon.mayaffre@unice.fr

[3] L.A.S.L.A. - dominique.longree@uliege.be

Abstract

This contribution confronts ADT and Machine learning. The extraction of statistical key-passages is undertaken following several calculations implemented using the Hyperbase software. An evaluation of these calculations according to the filters applied (taking into account only positive specificities, only substantives, etc.) is given. The extraction of key passages obtained by deep learning - passages that have the best recognition rate at the time of a prediction - is then proposed. The hypothesis is that deep learning is of course sensitive to the linguistic units on which the computation of the key statistical sentences are based, but also sensitive to phenomena other than frequency and other complex linguistic observables that the ADT has more difficulty taking into account - as would be the case with underlying patterns (Mellet et Longrée, 2009). If this hypothesis is confirmed, it would on the one hand permit better understanding of the *black box* of deep learning algorithms and on the other hand to offer the ADT community a new point of view.

Résumé

Cette contribution confronte ADT et Deep learning. L'extraction de passages-clefs statistiques est d'abord proposée selon plusieurs calculs implémentés dans le logiciel Hyperbase. Une évaluation de ces calculs en fonction des filtres appliqués (prise en compte des spécificités positives seulement, prise en compte de substantifs seulement, etc) est donnée. L'extraction de passages-clefs obtenus par deep learning - c'est-à-dire des passages qui ont le meilleur taux de reconnaissance au moment d'une prédiction - est ensuite proposée. L'hypothèse est que le deep learning est bien sûr sensible aux unités linguistes sur lesquelles le calcul des phrases-clefs statistiques se fondent, mais sensible également à d'autres phénomènes que fréquentiels et d'autres observables linguistiques complexes que l'ADT a plus de mal à prendre en compte - comme le seraient des motifs sous-jacents (Mellet et Longrée, 2009). Si cette hypothèse se confirmait, elle permettrait d'une part de mieux appréhender la *boîte noire* des algorithmes de deep learning et d'autre part d'offrir à la communauté ADT de nouveaux points de vue.

Mots-clés: ADT, deep learning, phrase-clef, motif, spécificités, nouveaux observables

1. Introduction

Pour des raisons techniques avant tout, l'ADT s'est constituée à partir des années 1960 autour du token, c'est-à-dire du mot graphico-informatique. Depuis lors, la discipline n'a cessé de varier et d'élargir ses observables, convaincue que le token seul rendait difficilement compte du texte dans sa complexité linguistique. Ainsi la tokenisation en particules graphiques élémentaires reste l'acte informatique premier des traitements textométriques, et le calcul des *spécificités*

lexicales reste l'entrée statistique privilégiée de nos parcours interprétatifs. Cependant, la recherche d'unités phraséologiques élargies et complexes, caractérisantes et structurantes des textes, est devenue le programme d'une discipline désormais adulte. Historiquement, dès 1987, le calcul des *segments répétés* (Salem, 1987) ou les n-grams a représenté une avancée puisque les segments significatifs du texte, de taille indéterminée, étaient automatiquement repérés ; et aujourd'hui la détection automatique, non supervisée, de *motifs* (Mellet et Longrée, 2009; Quiniou et al., 2012; Mellet et Longrée, 2012; Longrée et Mellet, 2013) - objets linguistiques complexes à empans variables et discontinus - apparait un enjeu décisif. C'est dans cette perspective que cette contribution travaille et met à l'épreuve l'idée de *passages-clefs* du texte, tels qu'ils sont implémentés dans les deux versions d'Hyperbase (locale développée par Etienne Brunet et web développée par Laurent Vanni) que l'UMR Bases, Corpus, Langage produit en collaboration avec le LASLA. La démonstration se fait en deux temps. D'abord, nous proposons une extraction statistique de *passages-clefs*, avec évaluation de leur pertinence interprétative sur un corpus français et un corpus latin. Ensuite une confrontation méthodologique avec le deep learning est mise en œuvre puisque le traitement deep learning attribue, après apprentissage, les passages de texte à leur auteur avec un taux de réussite éprouvé : par déconvolution nous repérons alors au sein de ces passages les *zones d'activation*, en soupçonnant qu'il s'agit, d'un point de vue linguistique, de motifs remarquables.

2. Les passages-clefs en ADT

2.1. Terminologie

Si nous préférons le terme de *passage-clef* à celui de *phrase-clef* c'est que les traitements ici présentés n'ont pas de modèle syntaxique, et que la ponctuation forte qui délimite habituellement la phrase est un jalon utile mais non-nécessaire à nos traitements. La notion de passage a été fortement théorisée par (Rastier, 2007) dans un article éponyme et désigne une « grandeur » du texte dont la valeur textuelle c'est-à-dire interprétative est patente. Un passage est donc un morceau de texte jugé suffisamment parlant, notamment par sa taille qui gagne à dépasser le mot, le segment voire la phrase, pour prétendre rendre compte d'un texte. Le passage-clef, quant à lui, s'appuie sur la définition rastirienne mais est une unité de surcroit textométrique ; c'est-à-dire une unité dont la pertinence est calculable et l'extraction automatique.

2.2. Implémentations

Les logociels ADT comme Hyperbase, Dtm-Vic, Iramuteq implémentent des calculs et l'extraction de passages-clefs. Dans tous les cas, les calculs proposés reposent sur l'examen des mots spécifiques (Lafon, 1984) : grosso modo, plus un passage concentre de spécificités, plus ce passage est jugé remarquable. Nous présentons ici deux types d'approche sur des passages arbitrairement constitués de 50 mots : un calcul naïf et sans filtre dans lequel tous les mots du passage sont considérés et un calcul filtré par nos connaissances linguistiques (sélection a priori des mots à considérer). Une évaluation de ces deux types d'approche est ensuite donnée.

2.2.1. Calcul sans filtre

Dans le cadre des études contrastives habituelles en ADT, l'indice de spécificité de chaque mot (Lafon, 1984) est sommé, qu'il soit positif ou négatif en postulant que si les mots positifs (les mots sur-utilisés par un auteur par exemple) doivent promouvoir le passage, il est légitime que

les mots négatifs (les mots sous-utilisés par un auteur) doivent l'handicaper. Chaque passage du corpus se trouve ainsi doté d'un super-indice de spécificité et Hyperbase fait remonter en bon ordre les passages les plus caractéristiques des textes comparés. Ainsi pour le français, sur le corpus de la présidentielle française 2017, le passage-clef le plus fortement indicé d'E. Macron (versus les autres candidats) est le suivant :

> *[...] nous croyons dans l'innovation, dans la transformation écologique et environnementale, parce que nous voulons réconcilier cette perspective et l'ambition de nos agriculteurs, parce que nous croyons dans la transformation digitale, parce que nous sommes pour une société de l'innovation, parce que nous voulons [...]*

Quoique naïf, le calcul apparait performant puisque l'interprétabilité socio-linguistique de ce passage est évidente : de fait Macron s'est fait élire sur un discours dynamique (*voulons* , *innovation* (deux fois), *transformation* (deux fois), *digitale*) et un discours rassembleur susceptible de transcender le clivage gauche/droite (*nous* (5 fois), *réconcilier*).

2.2.2. Calcul filtré

Par connaissances linguistiques et statistiques, le calcul peut être raffiné. Par exemple, seules les spécificités positives – et parmi elles, les spécificités les plus fortes – peuvent être considérées au motif qu'un objet s'identifie mieux par ses qualités que par ses défauts. Ensuite, les mots outils (conjonctions, déterminants) peuvent être écartés : ils présentent le double inconvénient d'avoir de très hautes fréquences (potentiellement déterminante pour le calcul des spécificités) et d'être peu parlants d'un point de vue sémantico-thématique. Et encore, la catégorie grammaticale peut être choisie : par exemple seuls les noms propres et communs, parfois plus chargés de sens, sont pris en compte. Ainsi pour le latin un passage-clef de Jules César, contrasté à de nombreux auteurs contenus dans la base du LASLA, est le suivant :

> *[...] partes Galliae uenire audere quas Caesar possideret neque exercitum sine magno commeatu atque molimento in unum locum contrahereposse sibi autem mirum uideri quid in sua Gallia quam bello uicisset aut Caesari aut omnino populo Romano negotii esset his responsis ad Caesarem relatis iterum ad eum Caesar [...]*

De fait, ce passage de la Guerre des Gaules peut être effectivement considéré comme très représentatif de l'œuvre de César. On relève des noms propres connus (*Galliae*, *Caesar*, *Gallia*) ou des noms communs correspondant à la réalité militaire du moment (*bello*, *commeatu*). Toutefois la méthode ne permet pas de repérer des structures caractéristiques de la langue et du style de César, comme par exemple une proposition participiale marquant la transition entre épisodes dans une négociation : *His responsis ad Caesarem relatis*, « Ces réponses ayant été rapportées à César ».

2.2.3. Evaluation

Calcul naïf ou calcul élaboré : nous récapitulons quelques performances. Dans un corpus contrastif, nous calculons le score de super-spécificité de chaque passage en fonction des différents auteurs comparés (Figure 1). Par exemple pour le français, sans aucun filtre 58% des passages du corpus de la présidentielle sont attribués justement à leur auteur ; et en ne considérant que les spécificités positives, le score descend à 52%. A l'opposé, en imposant le double filtre de la

	Passages-clefs - ADT				Deeplearning
	sans filtre	$z > 0$	substantifs	substantifs et $z > 0$	
français	58%	52%	88%	89%	90%
latin	69%	62%	84%	82%	85%

FIGURE 1 – Taux d'attribution ADT et taux de prédiction deep learning

catégorique grammaticale (seulement les substantifs) et de l'indice de spécificité (seulement les spécificités positives) nous élevons le taux de bonne attribution à 89% pour le français et 82% pour le corpus latin du LASLA.

3. Deep learning : à la recherche de nouveaux marqueurs linguistiques

3.1. Convolution et déconvolution, les principes

Le découpage du texte en segments de taille fixe est une méthode qui peut aussi être utilisée pour entraîner un réseau de neurones. Chaque segment devient alors une image d'un texte que le réseau va utiliser pour apprendre (Ducoffe et al., 2016) et faire ensuite de la prédiction. Sur nos deux corpus de référence (français et latin), les taux de précision convergent rapidement et atteignent le même niveau que ceux obtenus avec l'ADT (Figure 1). Si nous connaissons les paramètres à faire varier pour optimiser la détection des passages-clefs ADT, ceux issus du deep learning sont complètement non supervisés et découverts automatiquement par le réseau. L'idée des réseaux à convolution est de proposer un modèle capable de faire automatiquement une abstraction performante des données [1]. La convolution utilise pour cela un mécanisme de filtres qui va lire le texte avec une fenêtre coulissante pour extraire à chaque fois une partie de la matière linguistique présente dans la fenêtre (figure 2). Avec des centaines de filtres de tailles différentes, le texte est lu en utilisant tous les empans linguistiques possibles et le mécanisme de back-propagation [2] finit par accorder un certain poids à certains éléments du texte qui le pousse à prendre la bonne décision. Le deep learning est souvent considéré comme une boîte noire faute de pouvoir mettre en évidence précisément ces éléments. Nous avons donc ici concentré nos efforts sur la déconvolution. Ce mécanisme utilisé notamment en analyse d'images permet de démêler le réseau et de lui redonner une forme interprétable par l'humain. Notre modèle est composé d'une couche de pré-apprentissage (Mikolov et al., 2013) pour la représentation des mots en vecteurs, d'une couche de convolution (Kim, 2014), un maxpooling pour compresser l'information et enfin un réseau classique de perceptron à une couche cachée pour la classification (figure 2). La déconvoltution est en fait une simple copie partielle de ce réseau (jusqu'à la convolution) à laquelle on ajoute à la fin une transposée de la convolution. On copie bien sûr le poids de chaque neurone après l'entraînement dans cette copie de réseau et on obtient un nouveau réseau dont la couche de sortie correspond au résultat de chaque filtre de la convolution. Une simple somme de ces filtres pour chaque mot nous donne un indice d'activation du mot dans son contexte. Au final nous observons ici des zones de texte s'activer plus ou moins suivant l'importance que leurs a accordée le réseau.

1. L'abstraction des données peut être considérée comme les saillances lexicales d'un texte qui lui donnent une identité propre

2. Correction de l'erreur à chaque phase d'apprentissage.

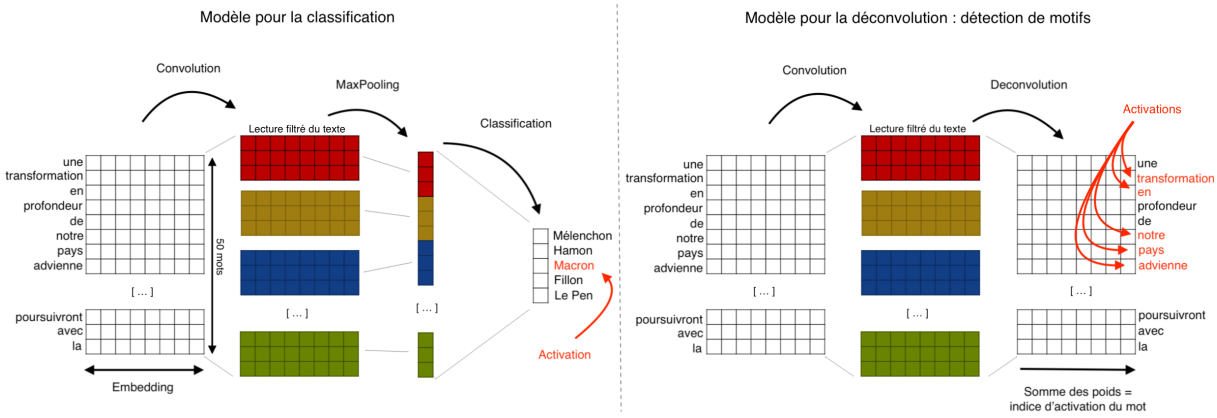

FIGURE 2 – Convolution et déconvolution d'un passage du discours d'E. Macron

3.2. Résultats et perspectives

A la lecture des résultats, nous voyons que le modèle identifie, sans surprise, des mots que le traitement statistique avait calculés comme spécifiques. Mais pas seulement. Certaines zones éclairées par le réseau semblent relever d'une nouvelle forme de lecture du texte. Nous pouvons illustrer ce constat avec un extrait des vœux d'E. Macron le 31 décembre 2017.

> *[...] une transformation en profondeur de notre pays advienne : à l' école pour nos enfants , au travail pour l' ensemble de nos concitoyens , pour le climat , pour le quotidien de chacune et chacun d' entre vous . Ces transformations profondes ont commencé et se poursuivront avec la [...]*

Dans ce passage, les mots « *transformation* » et « *notre* », fortement spécifique de Macron, sont activés : ici nous n'avons pas de plus-value heuristique par rapport à l'ADT. De la même manière, le segment répété « *chacune et chacun* », très spécifique, est repéré par le réseau. Mais il y a aussi les mots « *pays* » et « *advienne* » qui ne sont pas statistiquement spécifique de Macron et qui ont pourtant fortement contribué à la reconnaissance du passage. Si l'on regarde maintenant les activations autour de ces mots ciblés, on voit que c'est une expression formée de plusieurs mots, pas forcément contigus, qui est repérée par le réseau. Il semble donc que le deep learning ait identifié des structures phraséologiques ou motifs linguistiques sensibles aux occurrences et à leur organisation syntagmatique. Plus loin, la visualisation du passage dans son ensemble met au jour une topologie textuelle ou un rythme auxquels le deep larning a été sensible (Figure 3).

4. Conclusion

L'ADT et le deep learning ne sont peut-être pas des continents étrangers l'un à l'autre (Lebart, 1997). Cette contribution en croisant approche statistique et réseau de neurones nous a permis d'identifier des passage-clefs et peut-être des motifs susceptibles de nourrir nos traitements textuels. Si les observables qui ont présidé à la détection de passages-clefs par l'ADT (les spécificités lexicales) sont connus et éprouvés, les zones d'activation du deep learning semblent

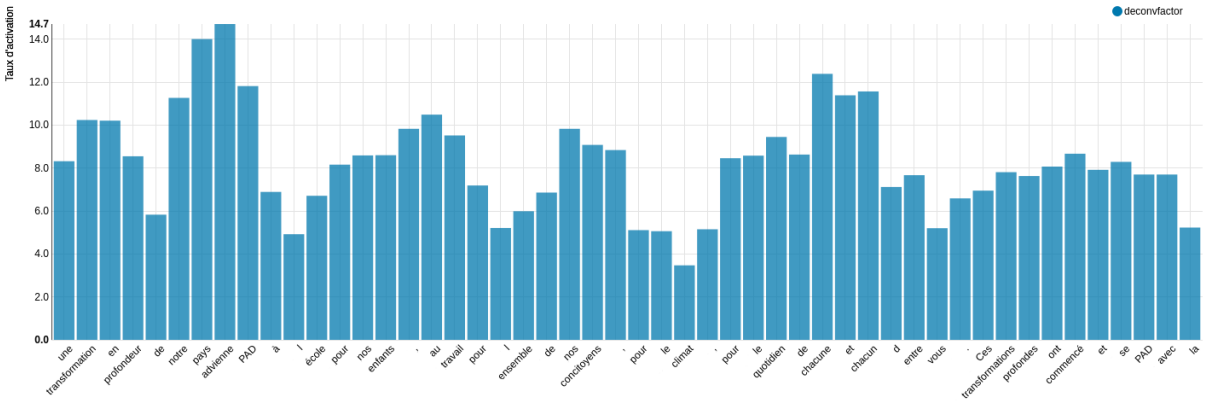

FIGURE 3 – Déconvolution : observation de la topologie d'un passage

relever de nouveaux observables linguistiques. Rappelons que la matière linguistique et la topologie des passages ne sauraient renvoyer au hasard : les zones d'activations permettent d'obtenir des taux de reconnaissance de plus de 90% sur le discours politique français et de 85% sur le corpus du LASLA ; soit des taux équivalents ou supérieurs aux taux obtenus par le calcul statistique des passages-clefs. Reste désormais à améliorer le modèle et à en comprendre tous les aboutissants mathématiques comme linguistiques. La première amélioration que l'on se propose désormais d'implémenter est l'injection d'informations morphosyntaxiques dans le réseau afin de mettre à l'épreuve des motifs linguistiques toujours plus complexes.

Références

Ducoffe, M., Precioso, F., Arthur, A., Mayaffre, D., Lavigne, F., et Vanni, L. (2016). Machine learning under the light of phraseology expertise : use case of presidential speeches, de gaulle - hollande (1958-2016). In *Actes de JADT 2016*, pages 155–168.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Lafon, P. (1984). Dépouillements et statistiques en lexicométrie. In *Genève-Paris, Slatkine-Champion*.

Lebart, L. (1997). Réseaux de neurones et analyse des correspondances. In *Modulad, (INRIA Paris), 18*, pages 21–37.

Longrée, D. et Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages 189*, pages 65–79.

Mellet, S. et Longrée, D. (2009). Syntactical motifs and textual structures. In *Belgian Journal of Linguistics 23*, pages 161–173.

Mellet, S. et Longrée, D. (2012). Légitimité d'une unité textométrique : le motif. In *Actes de JADT 2012*, pages 715–728.

Mikolov, T., Chen, K., Corrado, G., et Dean, J. (2013). Efficient estimation of word representations in vector space. In *arXiv :1301.3781*.

Quiniou, S., Cellier, P., Charnois, T., et Legallois, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In *Actes de JADT 2012*.

Rastier, F. (2007). Passages. *Corpus 6*, pages 25–54.

Salem, A. (1987). Pratique des segments répétés. essai de statistique textuelle. *Paris : Klincksieck*.