Apprentissage profond pour le traitement automatique des langues

Alexandre Allauzen* — **Hinrich Schütze****

* *LIMSI-CNRS, Université Paris-Sud et Université Paris-Saclay*
** *Center for Information and Language Processing, Ludwig-Maximilians-Universität (LMU), München*

RÉSUMÉ. *Ces dernières décennies, les réseaux de neurones artificiels et plus généralement leur apprentissage dit profond ont renouvelé les perspectives de recherche en traitement automatique des langues (TAL). Par leur capacité en termes d'apprentissage de représentations, les réseaux neuronaux ont permis des avancées importantes pour le TAL et ce pour de nombreuses tâches par exemple l'analyse syntaxique, la classification de documents, la reconnaissance automatique de la parole et la traduction automatique. Néanmoins, ces évolutions récentes posent de nombreuses questions scientifiques. Si les performances obtenues par les modèles neuronaux impressionnent souvent, les architectures déployées sont complexes à concevoir et à optimiser. La vision de ces modèles comme une boîte noire est problématique tant l'interprétation des résultats et la compréhension de ce qui est appris restent obscures. Ce numéro spécial propose donc d'explorer les apports de l'apprentissage profond pour le TAL, d'en montrer à la fois les promesses, les limites et les singularités.*

ABSTRACT. *During the last decades, artificial neural networks and deep learning approaches have strongly renewed the research perspectives in Natural Language Processing (NLP). Neural networks provide an efficient way for representation learning, yielding important improvement in several tasks. These tasks include document classification, syntactic parsing, automatic speech recognition and machine translation. Whereas the performances achieved by neural networks are impressive, their conception and optimization are still challenging. Moreover, these architectures are merely understood as efficient black boxes and their results remain difficult to interpret and explain. This special issue explores contributions of deep learning to NLP, their promises along with their limits and peculiarities.*

MOTS-CLÉS : *réseaux de neurones, apprentissage profond.*

KEYWORDS: *neural network, deep-learning.*

1. Introduction

Au cours des dernières décennies, les réseaux de neurones artificiels et plus généralement l'apprentissage profond de ces modèles (*deep-learning*) ont renouvelé les perspectives de recherche en traitement automatique des langues (TAL). Comme dans d'autres domaines, le potentiel de ces approches à modéliser des données et des tâches complexes explique leur impact. Ainsi, de nombreuses applications du TAL, avec pourtant des objectifs différents, sont concernées. Par exemple, la classification de textes implique de représenter un texte de longueur variable afin d'en extraire les caractéristiques nécessaires à la prédiction d'une classe. En traduction automatique, l'enjeu peut se schématiser par la représentation d'une phrase dans un but de génération d'une séquence de mots dans une autre langue. Cet enjeu se retrouve de manière similaire dans d'autres applications comme en reconnaissance automatique de la parole, où la séquence d'entrée est un signal acoustique, et en analyse syntaxique, où la structure à engendrer est un arbre.

Face à cette diversité des structures manipulées, de nombreux travaux en TAL ont exploré de manière souvent disjointe, d'une part, la représentation des données d'entrée (mots, phrases et documents), et d'autre part, les modèles de prédiction, par exemple effectuant la classification d'un texte ou l'inférence d'une traduction. Ainsi de nombreux travaux ont été pionniers dans l'apprentissage de représentations vectorielles pour les mots et les documents, grâce à l'analyse sémantique latente et ses variantes (Benzécri, 1981 ; Deerwester *et al.*, 1990 ; Schütze, 1992) ou des modèles de thèmes probabilistes (Hofmann, 2001 ; Blei *et al.*, 2003). Par ailleurs, l'inférence de structures a donné lieu à des modèles spécifiques comme le perceptron structuré (Collins, 2002) et les champs aléatoires conditionnels (Lafferty *et al.*, 2001) où la représentation des données d'entrée doit être spécifiée par l'utilisateur.

Une des difficultés majeures, en termes d'apprentissage automatique, est que les données langagières se caractérisent par des distributions particulières suivant la loi de Zipf (Zipf, 1935). Ces distributions sont souvent qualifiées de parcimonieuses ou de creuses et elles portent sur des unités discrètes dont l'inventaire est potentiellement grand. Ainsi la plupart des modèles utilisés par le passé ont été confrontés à des difficultés de généralisation, et les enjeux scientifiques se sont alors concentrés sur, d'une part, le développement d'estimateurs statistiques plus robustes, illustré par les nombreux travaux sur les modèles de langues et leur estimation (H. Ney, 1994 ; Chen et Goodman, 1996 ; Teh, 2006), et, d'autre part, la conception de modèles pouvant tenir compte d'une description riche des données manipulées dont les champs conditionnels aléatoires sont les représentants (Lavergne *et al.*, 2010).

2. Réseaux de neurones pour le TAL

Dès les premiers travaux, l'application des réseaux de neurones au TAL poursuit le même objectif lié à la représentation des unités linguistiques qui sont discrètes, par exemple des mots, des caractères, des catégories morphosyntaxiques ou séman-

tiques. L'objectif est de substituer à ces unités discrètes une représentation numérique continue sous forme de vecteur afin de les « plonger » dans un espace où il est possible de définir des notions de similarité qui sont donc plus propices à la généralisation. Cette notion de plongement apparaît d'abord dans le contexte des réseaux sémantiques (Hinton, 1981 ; Hinton, 1986). Puis dans (Sejnowski et Rosenberg, 1986 ; Sejnowski et Rosenberg, 1988), les auteurs introduisent les premiers plongements (ou *embedding*) de caractères pour faire de la conversion graphème-phonème et dans (Nakamura et Shikano, 1988 ; Nakamura *et al.*, 1990) les unités considérées sont des classes morphosyntaxiques.

Cette notion de plongement sera formalisée plus avant dans (Bengio *et al.*, 2003) avec cette fois-ci comme application la définition d'un modèle de langue neuronal n-grammes. L'unité considérée dans ces travaux est le mot et on parle alors de plongements lexicaux ou *word embeddings*. L'apport des réseaux de neurones réside dans leur capacité à représenter dans un espace continu les unités discrètes manipulées. Le modèle peut ainsi mieux généraliser ses connaissances extraites des données d'apprentissage en exploitant la similarité entre les unités manipulées, dans l'espace continu de représentations. La différence fondamentale avec d'autres types de représentations vectorielles comme l'analyse sémantique latente (Deerwester *et al.*, 1990) dans un contexte applicatif similaire (Bellegarda, 2000 ; Afify *et al.*, 2007) est que les représentations sont apprises conjointement avec la fonction de similarité nécessaire à l'application. Cette idée a permis des avancées rapides et notables en reconnaissance automatique de la parole (Schwenk et Gauvain, 2002) puis en traduction automatique (Schwenk *et al.*, 2006 ; Le *et al.*, 2012).

3. De l'apprentissage de représentations aux systèmes de bout en bout

Au-delà de ces premières applications, les plongements lexicaux ont par la suite été utilisés dans de nombreuses tâches du TAL. Dans (Collobert et Weston, 2008 ; Collobert *et al.*, 2011), les auteurs proposent une architecture unifiée permettant d'exploiter les plongements lexicaux pour différentes tâches d'étiquetage de séquences. Partant du constat que, dans beaucoup de langues, les textes sous format électronique sont des ressources facilement accessibles et exploitables, il est possible de pré-apprendre les plongements lexicaux sur ces données textuelles non annotées disponibles en grande quantité puis de raffiner ces représentations pour une tâche précise en utilisant cette fois les données annotées qui sont, quant à elles, disponibles en faible quantité (Collobert *et al.*, 2011 ; Mikolov *et al.*, 2013). Ce type d'approche a connu récemment un regain d'intérêt très important avec le déploiement d'architectures neuronales plus complexes et des capacités d'apprentissage bien plus importantes, tant au niveau des ressources de calcul que des données disponibles (Peters *et al.*, 2018 ; Howard et Ruder, 2018 ; Devlin *et al.*, 2018).

Les réseaux de neurones se distinguent également par leur capacité à représenter une phrase au-delà d'un simple sac de mots. Cette capacité s'appuie principalement sur deux types d'architectures qui se distinguent par les mécanismes de représenta-

tions d'une séquence, c'est-à-dire sur la façon de la décomposer ainsi que de tenir compte des dépendances et des interactions entre les mots qui la constituent. Les réseaux convolutifs sont le premier type. Inspirés par l'opérateur de convolution en traitement du signal, ces réseaux peuvent être perçus comme une généralisation des modèles n-grammes. Ils s'appuient sur des fenêtres glissantes de différentes tailles, permettant l'extraction de caractéristiques locales. Ces caractéristiques sont ensuite combinées afin de représenter la phrase dans son ensemble (Waibel *et al.*, 1990 ; Collobert et Weston, 2008 ; Kim, 2014). L'autre type d'architecture utilise les réseaux récurrents (Elman, 1990) et leurs évolutions récentes comme les réseaux LSTM, pour *Long Short Term Memory* (Hochreiter et Schmidhuber, 1997 ; Graves, 2008). Un réseau récurrent parcourt la phrase, par exemple de gauche à droite, un mot après l'autre, mettant à jour la mémoire interne du réseau à chaque pas, accumulant ainsi une vision globale de la séquence. Afin de renforcer la modélisation des dépendances à longue distance, les réseaux récurrents peuvent être bidirectionnels, parcourant la phrase de gauche à droite et de droite à gauche (Schuster et Paliwal, 1997).

L'enjeu pour ces architectures est double. D'une part, elles permettent d'apprendre à représenter les unités qui composent une séquence grâce aux plongements lexicaux, et, d'autre part, elles modélisent le mécanisme combinant ces plongements afin de représenter la séquence dans son ensemble. Ainsi, les modèles neuronaux ont dépassé le cadre de l'apprentissage de représentations pour évoluer vers des architectures de plus en plus profondes, permettant de modéliser de bout en bout des tâches d'inférence de plus en plus complexes, comme la génération d'une phrase ou d'un arbre syntaxique. Désormais, pour de nombreuses applications, les approches considérées comme état de l'art ont rapidement évolué ces dernières années, que ce soit en traduction automatique (Bahdanau *et al.*, 2014 ; Vaswani *et al.*, 2017) en reconnaissance automatique de la parole (Chan *et al.*, 2016 ; Chiu *et al.*, 2018) en synthèse vocale (van den Oord *et al.*, 2016 ; Li *et al.*, 2018), ou pour la génération de légendes d'images (Xu *et al.*, 2015). Ces systèmes, qui s'appuyaient auparavant sur différents types de modèles, sont désormais constitués d'un seul réseau de neurones.

4. Contenu du numéro

Ce numéro spécial de la revue TAL comprend les trois articles suivants :

– « *Classifying Semantic Clause Types with Recurrent Neural Networks : Analysis of Attention, Context and Genre Characteristics* », de Maria Becker, Michael Staniek, Vivi Nastase, Alexis Palmer, et Anette Frank ;

– « Prédiction de performances des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs », de Zied Elloumi, Laurent Besacier, Olivier Galibert, et Benjamin Lecouteux ;

– « *Adversarial networks for machine reading* », de Quentin Grail, Julien Perez, et Tomi Silander.

Le premier article s'intéresse à la classification sémantique des propositions. Les auteurs déploient une architecture combinant des réseaux récurrents avec un mécanisme d'attention. Il explore en particulier la capacité de ce type de modèle à apprendre une représentation pertinente pour les propositions, et l'usage du mécanisme d'attention permet de mieux caractériser la nature des propriétés linguistiques apprises par le réseau. Dans le deuxième article, bien que le choix de l'architecture diffère avec la construction d'un réseau convolutif, les auteurs explorent également les propriétés apprises par le réseau afin de mieux expliquer les performances du modèle. L'application est ici de prédire les performances d'un système de reconnaissance automatique de la parole lorsqu'il est confronté à des conditions d'utilisation nouvelles. Le troisième article explore une architecture complexe qui couple deux réseaux « adversaires » afin d'apprendre un système de réponses à des questions. Ce dispositif d'apprentissage permet d'accroître la robustesse du système au bruit pouvant être présent dans les données, et de relâcher certaines contraintes de l'apprentissage supervisé.

Ces trois articles couvrent des applications différentes, avec des enjeux scientifiques bien distincts, allant de l'apprentissage de représentations aux stratégies d'apprentissage. Une thématique récurrente dans ces travaux est de montrer la capacité des réseaux neuronaux, pour des tâches complexes, d'apprendre automatiquement des caractéristiques à la fois pertinentes, en partie explicables, et robustes. Ces trois articles illustrent donc bien le potentiel et les promesses de ce type d'approche. Ils montrent aussi la complexité des solutions qui sont explorées. Ainsi la conception d'architectures neuronales pour le TAL reste un enjeu scientifique et technique qui, loin d'être une solution facile et toute prête, porte néanmoins des promesses importantes de progrès.

Remerciements

Nous tenons à remercier Sophie Rosset pour le suivi de ce numéro, les membres du comité permanent de la revue TAL, ainsi que les membres du comité spécifique à ce numéro :

– Marianna Apidianaki, LIMSI-CNRS
– Loic Barrault, Université du Maine, LIUM
– Fethi Bougares, LIUM, Université du Maine
– Marie Candito, LLF, Université Paris Diderot
– Marta R. Costa-jussà, Universitat Politecnica de Catalunya
– Benoit Crabbé, LLF, Université Paris Diderot
– Richard Dufour, LIA, Université d'Avignon
– Benoit Favre, LIS, Aix-Marseille Université
– Joseph Leroux, LIPN, Université Paris-Nord
– Matthieu Labeau, LIMSI, Université Paris-Sud

– Gwénolé Lecorvé, IRISA, Université de Rennes I, ENSSAT
– Fabrice Lefevre, LIA, Université d'Avignon
– Thomas Pellegrini, IRIT, Université de Toulouse III - Paul Sabatier
– Christian Raymond, IRISA, INSA de Rennes
– Lynda Tamine-Lechani, IRIT, Université de Toulouse III - Paul Sabatier
– Tim Van de Cruys, IRIT, CNRS

5. Bibliographie

Afify M., Siohan O., Sarikaya R., « Gaussian Mixture Language Models for Speech Recognition », *Proceedings of the IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*, vol. 4, p. 29-32, April, 2007.

Bahdanau D., Cho K., Bengio Y., « Neural Machine Translation by Jointly Learning to Align and Translate », *CoRR*, 2014.

Bellegarda J. R., « Exploiting latent semantic information in statistical language modeling », *Proc. of the IEEE, Special Issue on Speech Recognition and Understanding*, vol. 88, n° 8, p. 1279-1296, 2000.

Bengio Y., Ducharme R., Vincent P., Janvin C., « A neural probabilistic language model », *Journal of Machine Learning Research*, vol. 3, p. 1137-1155, 2003.

Benzécri J.-P., *Pratique de l'analyse des données. Linguistique et lexicologie*, Dunod, 1981.

Blei D. M., Ng A. Y., Jordan M. I., « Latent dirichlet allocation », *Journal of Machine Learning Research*, vol. 3, p. 993-1022, 2003.

Chan W., Jaitly N., Le Q., Vinyals O., « Listen, attend and spell : A neural network for large vocabulary conversational speech recognition », *Proceedings of the IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*, p. 4960-4964, March, 2016.

Chen S. F., Goodman J., « An Empirical Study of Smoothing Techniques for Language Modeling », *in* A. Joshi, M. Palmer (eds), *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, Morgan Kaufmann Publishers, San Francisco, p. 310-318, 1996.

Chiu C.-C., Sainath T., Wu Y., Prabhavalkar R., Nguyen P., Chen Z., Kannan A., Weiss R. J., Rao K., Gonina K., Jaitly N., Li B., Chorowski J., Bacchiani M., « State-of-the-art Speech Recognition With Sequence-to-Sequence Models », *Proceedings of the IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*, 2018.

Collins M., « Discriminative training methods for hidden Markov models : theory and experiments with perceptron algorithms », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1-8, 2002.

Collobert R., Weston J., « A unified architecture for natural language processing : deep neural networks with multitask learning », *Proceedings of the International Conference of Machine Learning (ICML)*, ACM, New York, NY, USA, p. 160-167, 2008.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P., « Natural Language Processing (Almost) from Scratch », *Journal of Machine Learning Research*, vol. 12, p. 2493-2537, 2011.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391-407, 1990.

Devlin J., Chang M., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *CoRR*, 2018.

Elman J. L., « Finding structure in time », *Cognitive Science*, vol. 14, n° 2, p. 179-211, 1990.

Graves A., Supervised sequence labelling with recurrent neural networks, PhD thesis, Technical University Munich, 2008.

H. Ney U. Essen R. K., « On Structuring Probabilistic Dependences in Stochastic Language Modelling », *Computer Speech and Language*, vol. 8, n° 1, p. 1-38, 1994.

Hinton G. E., « Implementing Semantic Networks in Parallel Hardware », *in* G. E. Hinton, J. A. Anderson (eds), *Parallel Models of Associative Memory*, Erlbaum, p. 161-187, 1981.

Hinton G. E., « Learning Distributed Representations of Concepts », *Annual Conference of the Cognitive Science Society*, 1986.

Hochreiter S., Schmidhuber J., « Long Short-Term Memory », *Neural Comput.*, vol. 9, n° 8, p. 1735-1780, November, 1997.

Hofmann T., « Unsupervised Learning by Probabilistic Latent Semantic Analysis », *Machine Learning*, vol. 42, n° 1, p. 177-196, 2001.

Howard J., Ruder S., « Universal Language Model Fine-tuning for Text Classification », *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, p. 328-339, 2018.

Kim Y., « Convolutional Neural Networks for Sentence Classification », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, p. 1746-1751, October, 2014.

Lafferty J., McCallum A., Pereira F., « Conditional random fields : probabilistic models for segmenting and labeling sequence data », *icml*, p. 282-289, 2001.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, Uppsala, Sweden, p. 504-513, July, 2010.

Le H.-S., Allauzen A., Yvon F., « Continuous Space Translation Models with Neural Networks », *Proceedings of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, Association for Computational Linguistics, Montréal, Canada, p. 39-48, June, 2012.

Li N., Liu S., Liu Y., Zhao S., Liu M., Zhou M., « Close to Human Quality TTS with Transformer », *CoRR*, 2018.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient estimation of word representations in vector space », *International Conference on Learning Representations (ICLR)*, 2013.

Nakamura M., Maruyama K., Kawabata T., Kiyohiro S., « Neural network approach to word category prediction for english texts », *Proceedings of the International Conference on Computational Linguistics (COLING)*, vol. 3, p. 213-218, 1990.

Nakamura M., Shikano K., « A study of English word category prediction based on neural networks », *The Journal of the Acoustical Society of America*, 1988.

Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L., « Deep Contextualized Word Representations », *Proceedings of the North American Chapter of*

the Association for Computational Linguistics (NAACL), Association for Computational Linguistics, p. 2227-2237, 2018.

Schuster M., Paliwal K., « Bidirectional Recurrent Neural Networks », *IEEE Transaction on Signal Processing*, vol. 45, nº 11, p. 2673-2681, November, 1997.

Schütze H., « Word Space », *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, p. 895-902, 1992.

Schwenk H., Dchelotte D., Gauvain J.-L., « Continuous space language models for statistical machine translation », *Proceedings of the International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 723-730, 2006.

Schwenk H., Gauvain J.-L., « Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition », *Proceedings of the IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*, Orlando, p. 765-768, May, 2002.

Sejnowski T. J., Rosenberg C. R., NETtalk : A parallel network that learns to read aloud, Technical Report nº 86/01, Johns Hopkins University Department of Electrical Engineering and Computer Science Technical, 1986.

Sejnowski T. J., Rosenberg C. R., *Neurocomputing : Foundations of Research*, MIT Press, Cambridge, MA, USA, chapter NETtalk : A Parallel Network That Learns to Read Aloud, p. 661-672, 1988.

Teh Y. W., « A Hierarchical Bayesian Language Model based on Pitman-Yor Processes », *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 985-992, 2006.

van den Oord A., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A., Kavukcuoglu K., « WaveNet : A Generative Model for Raw Audio », *Arxiv*, 2016.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., p. 6000-6010, 2017.

Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K. J., *Readings in Speech Recognition*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, chapter Phoneme Recognition Using Time-delay Neural Networks, p. 393-404, 1990.

Xu K., Ba J., Kiros R., Cho K., Courville A. C., Salakhutdinov R., Zemel R. S., Bengio Y., « Show, Attend and Tell : Neural Image Caption Generation with Visual Attention », *Proceedings of the International Conference of Machine Learning (ICML)*, 2015.

Zipf G., *The Psychobiology of Language : An Introduction to Dynamic Philology*, M.I.T. Press, Cambridge, Mass., 1935.