Introduction au numéro spécial — Robustesse et limites des modèles de traitement automatique des langues

Caio Corro[*] — **Gaël Lejeune**[**] — **Vlad Niculae**[***]

[*] *INSA Rennes, IRISA, Inria, CNRS, Université de Rennes*
[**] *STIH/CERES, Sorbonne Université Paris, France*
[***] *Language Technology Lab, IVI, FNWI, University of Amsterdam*

RÉSUMÉ. *Les chercheurs en traitement automatique des langues (TAL) sont amenés à traiter des tâches et des données de plus en plus variées. Ce numéro de la revue TAL s'intéresse à la capacité des systèmes de TAL à s'adapter à la variation des données, à leur robustesse. Les articles présentés ici s'intéressent à deux types de données qui questionnent la robustesse : les données générées par des utilisateurs et les données issues de la reconnaissance optique de caractères.*

MOTS-CLÉS : *robustesse, normalisation lexicale, contenus produits par les utilisateurs, correction automatique d'OCR, reconnaissance optique de caractères, reconnaissance d'entités nommées.*

TITLE. *Introduction to the special issue on robustness and limits of NLP systems*

ABSTRACT. *Researchers in natural language processing (NLP) are required to address a variety of tasks and data. This issue of the TAL journal focuses on the ability of NLP systems to adapt to data variability, to their robustness. The articles presented here explore two types of data that challenges robustness: user-generated content and data derived from optical character recognition.*

KEYWORDS: *robustness, lexical normalisation, user generated content, automatic OCR correction, optical character recognition, named entity recognition.*

1. Introduction

Les méthodes modernes d'apprentissage automatique ont permis d'atteindre des résultats spectaculaires en traitement automatique des langues, notamment à partir de la construction de modèles préentraînés comme les réseaux de plongements contextuels, c'est-à-dire BERT (Devlin *et al.*, 2019) et ses dérivés, et les (grands/giga) modèles de langue (Radford *et al.*, 2018). D'un point de vue scientifique, les avancées se sont faites sans changement concret de paradigme depuis une trentaine d'année. En effet, la production scientifique en traitement automatique des langues et en apprentissage automatique peut être caractérisée par son caractère routinier, par exemple :

– le web ou *world wide web* met à disposition un très grand nombre de données qui peuvent être exploitées par les modèles de TAL (Resnik, 1999 ; Grefenstette, 1999 ; Keller *et al.*, 2002 ; Ortiz Suárez *et al.*, 2019). Ces données sont souvent complétées par des données de meilleure qualité mais en quantité moindre, comme des productions journalistiques (Charniak, Eugene *et al.*, 2000) ;

– les réseaux de neurones sont utilisés depuis longtemps comme fondement pour les modèles de langue (Bengio *et al.*, 2000 ; Schwenk et Gauvain, 2002) ;

– l'apprentissage des paramètres d'un réseau de neurones nécessite de développer des méthodes d'optimisation pouvant s'appliquer sur des fonctions non convexes avec un très grand nombre de paramètres, par exemple en tirant bénéfice de la structure géométrique des données, ce qui est le cas de la descente de gradient naturelle (Amari, 1998), mais également de Adam (Kingma et Ba, 2015) qui sert de fondement à la plupart des approches d'optimisation actuelles ;

– l'utilisation de très grands jeux de données pour l'apprentissage de modèles génératifs nécessite de mettre en place des architectures neuronales qui peuvent pleinement profiter de la parallélisation des calculs, c'est par exemple le cas des architectures attentionnelles (Vaswani *et al.*, 2017) qui sont aujourd'hui hégémoniques, mais la masse des données était déjà un enjeu pour les *self-organizing maps* (Seiffert et Michaelis, 2001), pour lesquels la question de la parallélisation matérielle des calculs pour l'entraînement était importante : « *In general there are two main reasons to implement artificial neural networks on parallel hardware. [...] the second motive becomes evident when dealing with demanding real-world applications, when training times are increasing up to and above the pain threshold* » (Seiffert, 2004).

Il est d'ailleurs parfois difficile de savoir à partir de deux titres d'article qui ont presque 20 ans de différence lequel est le plus ancien :

– dans les actes d'HLT/EMNLP 2005 : « *Training Neural Network Language Models On Very Large Corpora* » (Schwenk et Gauvain, 2005) ;

– dans les actes d'un atelier d'EMNLP 2024 : « *Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations* » (Csordás *et al.*, 2024).

Ceci étant, ces dernières années ont quand même été marquées par un tournant sur plusieurs plans : passage à l'échelle des méthodes de TAL (autant en termes de

données utilisées qu'en taille des modèles), amélioration des performances pour de nombreuses tâches cibles, en particulier la tâche de génération de textes qui est aujourd'hui utilisée pour réaliser toutes sortes d'autres tâches, et enfin démocratisation de l'accès à ces modèles (où les interfaces conversationnelles sont devenues incontournables pour les utilisateurs).

La communauté a développé de nombreux jeux de données (ou *benchmarks*) sur lesquels les résultats évoluent très rapidement, donnant l'impression que de nombreux problèmes liés au traitement automatique des langues sont « résolus » ou en passe de l'être. La production scientifique du domaine s'apparente beaucoup à la création d'une collection de timbres[1], où chacun y va de son nouveau modèle, de son nouveau jeu de données, de sa nouvelle tâche, etc.

Pourtant, la question de la capacité des méthodes de TAL à être efficaces, voire simplement utilisables, sur différents types de données et cas d'usage reste ouverte. Par exemple, des applications comme la pharmacovigilance ou encore l'épidémiologie digitale nécessitent d'analyser en continu de larges volumes de données produits sur les réseaux sociaux, souvent écrits dans une langue non standard. Cela requiert d'une part d'avoir des méthodes d'analyse robustes, rapides et de préférence à faible consommation énergétique, mais d'autre part de penser finement les métriques d'évaluation qui peuvent être difficiles à agréger en un seul nombre : quelques faux positifs peuvent avoir un impact négligeable, alors qu'un faux négatif peut être catastrophique ; détecter un nouveau cluster de cas de rhumes semble moins primordial qu'un nouveau cluster de cas de virus Ebola. En traduction automatique, des métriques standards comme le score BLEU (Papineni *et al.*, 2002), mais également les métriques automatisées plus modernes comme COMET (Rei *et al.*, 2020), peuvent passer à côté d'éléments importants, comme une terminologie spécifique mais cruciale (même si elle ne concerne que peu de mots) ou la consistance de traduction des termes au sein d'un document (Semenov *et al.*, 2023). Il nous semble donc essentiel que la communauté s'intéresse à la robustesse concrète des méthodes de TAL au-delà de la collection de modèles et la surenchère de *benchmarks*.

2. Contexte de l'appel et relecture

Une journée d'études organisée par l'ATALA en 2022[2] avait été l'occasion pour la communauté française de présenter un éventail de travaux sur la robustesse et sur les limites des modèles actuels de TAL (Corro et Lejeune, 2022), c'est-à-dire sur la capacité de ces modèles d'offrir des performances comparables sur des données et des cas d'usage variés (Yu *et al.*, 2022). Cette journée d'études était focalisée sur les données dites « non standards », qui avaient été définies de manière large comme des données présentant des variations vis-à-vis d'un état de langue attendu : variation de la langue

1. Pour reprendre les termes d'une citation attribuée de façon incertaine à Ernest Rutherford : « *all science is either physics or stamp collecting* ».

2. `https://www.atala.org/content/robustesse-des-systemes-de-tal`

en diachronie, variations régionales, variation dans l'ordre des mots, *code-switching*, *user generated content*, orthographe inconsistante, données accidentellement bruitées suite à un prétraitement, données incomplètes ou encore présence d'un vocabulaire de domaine spécialisé. Plusieurs problématiques liées à la robustesse avaient été abordées telles que la reproductibilité des résultats, la portabilité des algorithmes, ou encore l'influence de la qualité des données. La variété des tâches abordées (reconnaissance d'entités nommées, traduction automatique, reconnaissance automatique de la parole, *web scraping* ou encore similarité sémantique) a naturellement amené à proposer de faire de cette thématique un numéro spécial de la revue TAL.

À la suite de cette journée, nous avons donc lancé un appel pour un numéro thématique de la revue TAL, visant à questionner la robustesse et les limites des modèles de TAL, en particulier en ce qui concerne les trois points suivants :

– données « non standards » : utilisation de modèles sur des données présentant des variations vis-à-vis d'un certain attendu en termes d'état de langue ;

– données hors domaine : utilisation de modèles sur des données d'un domaine différent par rapport aux données d'entraînement ;

– généralisation à des structures linguistiques non observées à l'entraînement : généralisation compositionelle (Kim et Linzen, 2020), généralisation structurelle (Yao et Koller, 2022) ou encore généralisation du genre (Stanovsky *et al.*, 2019), entres autres.

En terme de thématique, nous avons ouvert l'appel à un large éventail de contributions :

– identification et évaluation des phénomènes linguistiques problématiques pour les modèles neuronaux et les autres systèmes de TAL ;

– analyse et correction de la propagation des erreurs dans les systèmes fondés sur une analyse en cascade ;

– retours d'expérience sur l'utilisation de systèmes de TAL qui se sont révélés non fonctionnels sur des types de données particuliers ;

– critique de jeux de données utilisés pour l'apprentissage ou pour l'évaluation ;

– construction de jeux de données permettant d'évaluer la robustesse aux variations linguistiques ;

– augmentation artificielle de données pour améliorer la robustesse des modèles ;

– adaptation hors domaine ou apprentissage avec des domaines peu représentés dans les données ;

– architectures neuronales et méthodes d'entraînement améliorant la robustesse des modèles.

De plus, toutes les tâches standards du traitement automatique des langues pouvaient être considérées et les travaux portant sur d'autres langues que le français étaient les bienvenus, car nous avions pour objectif d'identifier les cas d'usage intéressants pour la recherche sur la robustesse.

3. Articles acceptés

Ce numéro comporte deux articles en français sélectionnés à l'issue du processus de relecture, portant tous les deux sur le prétraitement des entrées dans une chaîne d'analyse textuelle afin d'en normaliser le contenu : la normalisation lexicale pour le premier article et la correction des erreurs d'OCR (*optical character recognition*) pour le second.

Le premier d'entre eux, « *Étude sur la normalisation lexicale de contenus produits par les utilisateurs* » écrit par Lydia Nishimwe, Benoît Sagot et Rachel Bawden, propose un état de l'art sur la normalisation lexicale des contenus produits par des utilisateurs (ou UGC pour *User Generated Content*). La normalisation est ici définie comme la transformation des formes non standards par leur variantes standards, comme « jvien » en « je viens ». L'article propose de classifier les travaux de la littérature en deux catégories : les méthodes de correction de mot et les méthodes de traduction de phrase. De plus, l'article détaille les métriques et les corpus utilisés pour évaluer la qualité de la normalisation. Il montre que la normalisation n'améliore pas systématiquement les résultats, puisqu'elle introduit elle-même du bruit et que son efficacité dépend de la tâche et de la langue. Les autrices et l'auteur expliquent que les approches fondées sur des modèles comme BERT ont tendance à être moins sensibles aux données non standards mais que la normalisation se justifie sans doute encore sur des contextes avec peu de ressources (langues ou domaines peu dotés).

Le second article, « *Analyse multilingue de l'impact de la correction automatique de l'OCR sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires* » écrit par Caroline Koudoro-Parfait, Ljudmila Petkovic et Glenn Roe s'intéresse quant à lui à la normalisation des données produites par OCR. Il propose une analyse multilingue (français, anglais et portugais) de la robustesse de systèmes de reconnaissance d'entités nommées (REN) utilisés sur des données bruitées obtenues par OCR, en utilisant des outils existants pour les différentes étapes de traitement et de normalisation. Les autrices et l'auteur partent du constat que l'OCR génère des erreurs qui posent des problèmes à des systèmes de REN entraînés sur des données non bruitées. L'objectif de l'article est double : (1) évaluer si la correction automatique des transcriptions (ou post-correction OCR), censée améliorer la qualité de l'entrée, améliore la qualité de la REN et (2) proposer une critique des métriques d'évaluation strictes qui pourraient contribuer à sous-évaluer la robustesse des systèmes de REN. L'article montre que les outils ont une forte tendance à la surcorrection, avec des modifications erronées qui affectent la qualité (et le nombre) d'entités correctes extraites. Une typologie de l'impact de la correction sur la REN est également proposée.

4. Remerciements

Nous remercions le comité éditorial et scientifique de la revue TAL, ainsi que le comité scientifique invité, en particulier les relecteurs et relectrices, qui ont contribué par leur temps et leurs efforts à la qualité de ce numéro.

5. Bibliographie

Amari S.-i., « Natural Gradient Works Efficiently in Learning », *Neural Computation*, vol. 10, n° 2, p. 251-276, 1998.

Bengio Y., Ducharme R., Vincent P., « A Neural Probabilistic Language Model », *in* T. Leen, T. Dietterich, V. Tresp (eds), *Advances in Neural Information Processing Systems*, vol. 13, MIT Press, 2000.

Charniak, Eugene, Blaheta, Don, Ge, Niyu, Hall, Keith, Hale, John, Johnson, Mark, « BLLIP 1987-89 WSJ Corpus Release 1 », 2000.

Corro C., Lejeune G. (eds), *Actes de la journée d'étude sur la robustesse des systemes de TAL*, 2022.

Csordás R., Potts C., Manning C. D., Geiger A., « Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations », *in* Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, H. Chen (eds), *Proceedings of the 7th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Miami, Florida, US, p. 248-262, November, 2024.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *in* J. Burstein, C. Doran, T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Grefenstette G., « The World Wide Web as a Resource for Example-Based Machine Translation Tasks », *Proceedings of Translating and the Computer 21*, Aslib, London, UK, November 10-11, 1999.

Keller F., Lapata M., Ourioupina O., « Using the Web to Overcome Data Sparseness », *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, p. 230-237, July, 2002.

Kim N., Linzen T., « COGS : A Compositional Generalization Challenge Based on Semantic Interpretation », *in* B. Webber, T. Cohn, Y. He, Y. Liu (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 9087-9105, November, 2020.

Kingma D. P., Ba J., « Adam : A Method for Stochastic Optimization », *Proceedings of the International Conference on Learning Representations*, 2015.

Ortiz Suárez P. J., Sagot B., Romary L., « Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures », Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, p. 9 - 16, 2019.

Papineni K., Roukos S., Ward T., Zhu W.-J., « Bleu : a Method for Automatic Evaluation of Machine Translation », *in* P. Isabelle, E. Charniak, D. Lin (eds), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311-318, July, 2002.

Radford A., Narasimhan K., Salimans T., Sutskever I. *et al.*, « Improving language understanding by generative pre-training », 2018.

Rei R., Stewart C., Farinha A. C., Lavie A., « COMET : A Neural Framework for MT Evaluation », *in* B. Webber, T. Cohn, Y. He, Y. Liu (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 2685-2702, November, 2020.

Resnik P., « Mining the Web for Bilingual Text », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, College Park, Maryland, USA, p. 527-534, June, 1999.

Schwenk H., Gauvain J.-L., « Connectionist language modeling for large vocabulary continuous speech recognition », *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. I-765-I-768, 2002.

Schwenk H., Gauvain J.-L., « Training Neural Network Language Models on Very Large Corpora », *in* R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (eds), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, British Columbia, Canada, p. 201-208, October, 2005.

Seiffert U., « Artificial neural networks on massively parallel computer hardware », *Neurocomputing*, vol. 57, p. 135-150, 2004. New Aspects in Neurocomputing : 10th European Symposium on Artificial Neural Networks 2002.

Seiffert U., Michaelis B., « Multi-Dimensional Self-Organizing Maps on Massively Parallel Hardware », *Advances in Self-Organising Maps*, Springer London, London, p. 160-166, 2001.

Semenov K., Zouhar V., Kocmi T., Zhang D., Zhou W., Jiang Y. E., « Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies », *in* P. Koehn, B. Haddow, T. Kocmi, C. Monz (eds), *Proceedings of the Eighth Conference on Machine Translation*, Association for Computational Linguistics, Singapore, p. 663-671, December, 2023.

Stanovsky G., Smith N. A., Zettlemoyer L., « Evaluating Gender Bias in Machine Translation », *in* A. Korhonen, D. Traum, L. Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 1679-1684, July, 2019.

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*, vol. 30, Curran Associates, Inc., 2017.

Yao Y., Koller A., « Structural generalization is hard for sequence-to-sequence models », *in* Y. Goldberg, Z. Kozareva, Y. Zhang (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, p. 5048-5062, December, 2022.

Yu Y., Khan A. R., Xu J., « Measuring robustness for NLP », *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3908-3916, 2022.