Traduction automatique de documents: le cas des documents scientifiques

François Yvon - Rachel Bawden

Avril 2023

Projet Scientifique

La plupart des systèmes de traduction automatique (TA) dits “neuronaux” modélisent le processus de génération d'un document cible y à partir d'un document source x en décomposant ce processus phrase par phrase, chaque phrase étant traduite independamment des phrases voisines. Le modèle probabiliste sous-jacent est alors de la forme $P(y|x;\theta)$, où θ représente l'ensemble des paramètres du modèle (par exemple un modèle Transformer [Vas17]). Une fois θ appris, la génération d'une traduction repose sur la recherche de la traduction la plus probable réalisant : $\arg\max_y P(y|x;\theta)$.

Cette modélisation est naïve et ignore les multiples dépendances qui existent entre les phrases au sein d'un même document. Pour pallier cette déficience, de nombreuses architectures alternatives ont été proposées pour intégrer un contexte discursif c au modèle, conduisant à des modèles de la forme $P(y|x,c;\theta)$. Selon les implantations, le contexte c représente quelques phrases précédent x, ou bien tout le document source, ou bien également le début de la traduction (les phrases précédant y). Plusieurs manières d'encoder c (avec un encodeur dédié, ou bien en utilisant le même encodeur que pour x) ont été proposées dans la litérature. Les principales architectures de ce type, dédiées à la TA de documents (TAD), sont décrites dans [Mar21].

Deux obstacles principaux rendent cette extension difficile à implanter : (a) les ressources computationnelles (mémoire et calcul) nécesssaires à l'encodage d'un contexte étendu croissent de manière quadratique avec la longueur du contexte (pour les architectures Transformer) ; (b) l'apprentissage des dépendances entre y et c est rendu difficile par la relative rareté des mots pour lesquels le contexte étendu c est utile. La plupart des études dans le cadre de la TAD s'intéressent au problème (a) et considèrent soit des approximations du calcul de l'attention (voir [Tay21] pour un état des lieux récent), soit des architectures alternatives au modèle Transformer (par exemple [Gu21]) pour encoder des séquences.

Le travail de thèse proposé s'intéresse au problème de la traduction de docu-

ments complets, en se focalisant sur un type de documents particulier, à savoir les écrits académiques (articles, communications, rapports de recherche). Il s'agit de documents relativement longs, qui sont régis par des principes d'organisation et de présentation rigides et propres à ce genre textuel - organisation en sections en sous-sections -, ainsi que par des stratégies argumentatives propres à ce genre de textes : introduction de concepts et de définititions, raisonnements explicites devant venir à l'appui de démonstrations et de conclusions précises, etc.

L'objectif principal de la thèse est de parvenir à assurer que les documents générés par traduction automatique (a) reproduisent correctement la structure générale du texte d'entrée ; (b) manifestent le même niveau de cohésion et de cohérence, en particulier dans le choix des termes, que le texte source ; (c) reproduisent fidèlement les stratégies argumentatives (prémisses, déductions, conclusions) qui sont présentes dans le texte source, et (d) énoncent dans la langue cible les mêmes conclusions générales que dans la langue source. Parmi les autres difficultés de la tâche, qui pourront faire l'objet d'une attention particulière, mentionnons : la présence de nombreux extra-lexicaux (chiffres, symboles mathématiques, noms propres) et de parties non-textuelles (formules, équations, tableaux, graphiques). Notons enfin que les méthodes considérées devront être adaptées à une situation où les données monolingues sont abondantes, mais les données parallèles sont extrêmement rares : ce contexte est propice à l'utilisation de grands modèles de langue pré-entrainés.

Pour parvenir à ces fins, on s'intéressera par exemple aux architectures, déjà évoquées ci-dessus, qui exploitent un contexte discursif étendu, que ce soit pour la traduction automatique ou pour le résumé de longs documents [Koh22], ou encore aux méthodes de planification également utilisées pour la génération automatique de textes [Pup22]. L'enjeu principal de cet axe sera de rechercher les meilleurs compromis entre la complexité algorithmique du traitement de grands contextes (à l'apprentissage et à l'inférence) et le bénéfice tangible de ces efforts mesuré par de l'accomplissement des objectifs (a-d).

Un second axe de travail s'intéressera à la modélisation des stratégies discursives et des objectifs de communication associés à chacune des parties du document : une manière simple d'approximer ces objectifs s'appuie sur la structure interne des documents, mais des approches plus sophistiquées, utilisant des modèles à données latentes, devront également être considérées.

Contexte La thèse se déroule dans le cadre du projet ANR-MaTOS. Le candidat sera co-encadré par François Yvon (DR CNRS) au sein de l'équipe MLIA de l'Institut des Systèmes Intelligents et de Robotique et Rachel Bawden, CR Inria, dans l'équipe Almanach du Centre Inria Paris et inscrit dans l'école doctorale informatique, télécommunication et électronique (EDITE) de Paris.

Le ou la candidate recruté-e bénéficiera d'un contrat doctoral de Sorbonne Université, et aura, si il ou elle le souhaite, l'opportunité de réaliser des missions

d'enseignement au sein de l'Université.

Qualifications désirées

- Master en Informatique ou Mathématiques Appliquées avec une spécialisation en Intelligence Artificielle, Apprentisage Automatique, Traitement des Langues, ou diplôme équivalent).
- Solides compétences en programmation (PyTorch),
- Communication en anglais écrit et parlé
- Créativité et capacité à formuler et à résoudre des problèmes de manière autonome

Pour candidater Les candidatures seront reçues par email auprès de François Yvon (prenom.nom@cnrs.fr) et Rachel Bawden (prenom.nom@inria.fr) avec pour unique sujet "[Phd-DLMT] Candidature de Prenom Nom" avant le 22 mai 2023.

Merci de joindre à votre dossier :

- CV détaillé,
- Lettre de motivation,
- Relevés de notes (M1 et M2),
- Lettre(s) de recommandation

Références

- [Gu21] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
- [Koh22] H. Y. Koh, J. Ju, M. Liu, and S. Pan. An empirical survey on long document summarization : Datasets, models, and metrics. ACM Comput. Surv., 55(8), dec 2022.
- [Mar21] S. Maruf, F. Saleh, and G. Haffari. A survey on document-level neural machine translation : Methods and evaluation. ACM Comput. Surv., 54(2), Mar. 2021.
- [Pup22] R. Puduppully, Y. Fu, and M. Lapata. Data-to-text generation with variational sequential planning. Transactions of the Association for Computational Linguistics, 10 :697–715, 2022.
- [Tay21] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers : A survey. ACM Comput. Surv., apr 2022.
- [Vas17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017.