Adaptation d'un modèle de langue encodeur-décodeur pour l'extraction de relations dans des rapports de renseignement

Adrien Guille

ERIC Lyon 2, EA 3083, Université de Lyon
https://github.com/adrienguille/textmine2025

1 Introduction

Cette proposition au Défi TextMine 2025 (Prieur et al., 2025) s'inscrit dans le cadre moderne du traitement automatique de la langue, en définissant la tâche d'extraction de relations comme une tâche de génération de texte.

Formulation de la tâche On s'inspire de la démarche proposée par Zhang et al. (2023), où l'extraction de relations est formulée comme un questionnaire à choix multiples, un modèle de langue se chargeant de choisir les options susceptibles de correspondre à des relations présentes dans un texte. Plutôt que d'inclure toutes les relations candidates en un seul prompt par le biais d'un QCM, on choisit de présenter chaque relation candidate dans un prompt indépendant, sous la forme d'une question fermée. Les relations candidates sont identifiées suivant les motifs (*type d'entité source*, *type de relation*, *type d'entité cible*) observés dans les données d'entraînement. Ci-après, un exemple de prompt et la réponse attendue :

Entrée reçue et sortie attendue

Does the relation (head_entity : [Constance Dupuis], relation_type : is_in_contact_with}, tail_entity : [Airîle, compagnie aérienne]), exists in the following text : "L'avion NY8 de la compagnie Airîle a lancé sa dernière position via le signal radio avant de se crasher dans une forêt en Malaisie le 19 février 2003. La compagnie aérienne a alerté les secours pour évacuer les passagers. Les hélicoptères d'urgence ont retrouvé l'appareil en feu. Les autorités malaisiennes ont recensé 15 morts au total. Cet incident n'a fait que peu de survivants, dont Constance Dupuis, présidente de l'association « des médicaments pour tous » en Grèce. D'après son témoignage, le NY8 a connu une défaillance technique que les pilotes n'ont pas pu contrôler. Les corps ont été transportés par brancard à la morgue." ?

---

no

Choix du type de modèle de langue Dans la lignée de nos propres travaux (Charpentier et al., 2024) et de travaux connexes tels que ceux de Qorib et al. (2024), qui mettent en avant les capacités sémantiques supérieures des encodeurs par rapport aux décodeurs, nous écartons les modèles de langue basés uniquement sur un décodeur (*e.g.* Llama, Mistral) au profit de modèles de langue basés sur une architecture encodeur-décodeur.

2 Modèles de langues

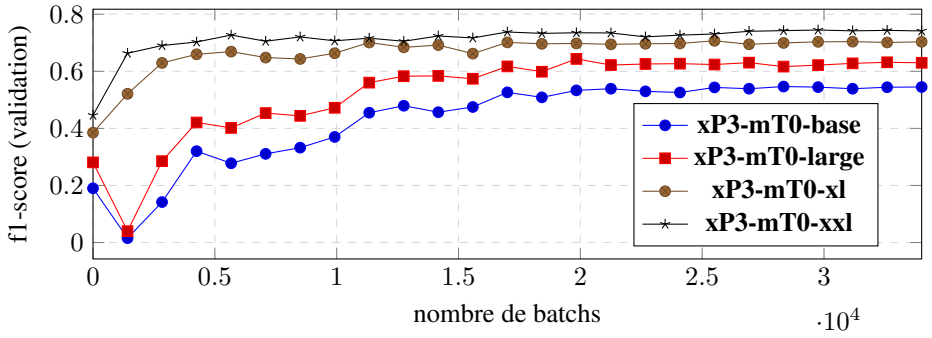
Modèles pré-entraînés On se limite à l'architecture encodeur-décodeur basée sur le bloc Transformer, telle que décrite par Raffel et al. (2020). Plus précisément, on considère le modèle pré-entraîné T5 et des modèles qui en découlent, brièvement décrits ci-après :

— **T5** (Raffel et al., 2020) : Modèle pré-entraîné sur le corpus anglophone C4 avec un objectif de débruitage de passages de texte ;
— **FLAN-T5** Chung et al. (2022) : Modèle **T5** spécialisé pour répondre à des instructions en poursuivant l'entraînement sur le corpus FLAN (où les instructions sont rédigées exclusivement en anglais et les réponses multilingues) ;
— **mT5** (Xue et al., 2021) : Même architecture que **T5** avec un vocabulaire de 250K tokens (au lieu de 32K pour T5), pré-entraîné sur le corpus multilingue mC4 ;
— **xP3-mT0** (Muennighoff et al., 2023) : Modèle **mT5** spécialisé pour répondre à des instructions en poursuivant l'entraînement sur le corpus xP3 (où les instructions sont rédigées exclusivement en anglais et les réponses multilingues).

Adaptation des modèles Pour adapter efficacement les modèles pré-entraînés à la tâche d'extraction de relations dans des rapports de renseignement, on procède à une adaptation de rang faible selon la méthode LoRA (Hu et al., 2022). Pour une matrice $W \in \mathbb{R}^{d_{\text{entrée}} \times d_{\text{sortie}}}$ paramétrant le modèle, cela consiste à approcher les ajustements à lui apporter, $\Delta W \in \mathbb{R}^{d_{\text{entrée}} \times d_{\text{sortie}}}$, par un produit de matrices de rang faible, $A \in \mathbb{R}^{d_{\text{entrée}} \times \text{rang}}$ et $B \in \mathbb{R}^{\text{rang} \times d_{\text{sortie}}}$: $\Delta W \approx A \times B$.

3 Résultats

On répartit les 800 documents annotés en 700 documents utilisés pour l'adaptation (matrices des projections en requêtes et en valeurs uniquement, rang 64) et 100 documents utilisés pour la validation. Par manque de place, on ne présente qu'une expérience sur l'effet de la taille du modèle **xP3-mT0**, sa variante **xxl** se classant première du défi. On observe que la performance est corrélée à la taille du modèle, que ce soit avant ou après l'adaptation. On remarque dans la figure ci-après que les grands modèles nécessitent peu de données pour atteindre leur meilleur score, tandis que l'adaptation des petits modèles est instable. On note par ailleurs que les variantes de **FLAN-T5** atteignent des scores similaires en validation, alors que les scores en test sont nettement inférieurs, ce qui suggère que l'entraînement multilingue de **xP3-mT0** lui confère une meilleure capacité de généralisation en français.



Références

Prieur, M., G. Gadek, A. Guille, H. Rawsthorne, P. Cuxac, et C. Lopez (2025). Défi TextMine'25 – Extraction de relations pour analyser des rapports de renseignement. In *Atelier TextMine (TextMine @ EGC 2025)*.

Charpentier, F., J. Cugliari, et A. Guille (2024). Exploring semantics in pretrained language model attention. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM @ ACL 2024)*.

Qorib, M., G. Moon, et H. T. Ng (2024). Are decoder-only language models better than encoder-only language models in understanding word meaning ? In *Findings of the Association for Computational Linguistics (ACL 2024)*.

Muennighoff, N., T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, et C. Raffel (2023). Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.

Zhang, K., B. Jimenez Gutierrez, et Y. Su (2023). Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics (ACL 2023)*.

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. hsin Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, et J. Wei (2022). Scaling instruction-finetuned language models. *Journal of Machine Learning Research 25*.

Hu, E. J., yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, et W. Chen (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR 2022)*.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, et C. Raffel (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL 2021)*.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, et P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21*.