Zero/few shot learning pour le NLU en français

ALBAR Boris, b.albar@catie.fr, CATIE, Talence (Orateur)

BOURDOIS Loïck, l.bourdois@catie.fr, CATIE, Talence

BEDU Pierre, p.bedu@catie.fr, CATIE, Talence

Thématique : **Thème 6 : Intelligence Artificielle**

Résumé : *Le domaine du NLU a vu des avancées significatives avec le développement de larges modèles de langue (BERT, GPT, ...). Néanmoins adapter ces modèles à des tâches spécifiques requiert souvent un nombre suffisant de données labellisées. Dans cette présentation, nous étudierons un ensemble de techniques pour le zero ou few-shot learning sur des problématiques de NLU dans le cadre de la langue française.*

Mots clés : *NLU, zero/few shot learning, modèles de langue*

1. Introduction

Le domaine du NLU a vu des avancées significatives avec le développement de larges modèles de langue entraînés de manière autosupervisée (BERT, GPT, ...). Néanmoins, la quantité de données nécessaire et le coût important en calcul rend difficile l'application directe de ces modèles à des problématiques spécifiques telles que celles trouvées dans un contexte industriel. A titre d'exemple, le modèle français FlauBERT dans sa version "Large" a été entrainé sur 138Gb de données non-labellisés pendant 390h sur 128 GPUs [2]. Les techniques classiques basées sur du « fine-tuning » nécessitent généralement plusieurs centaines d'exemples labellisés et nécessitent de réentraîner les modèles pour chaque nouvelle tâche.

De nombreuses techniques ont été développés pour pouvoir aborder des problèmes où aucune donnée n'est disponible (zero-shot learning) ou lorsqu'uniquement un ensemble limité d'exemples est disponible pour la tâche (few-shot learning). Ces approches reposent notamment sur le tuning d'un petit ensemble des paramètres du modèle (parameter-efficient fine-tuning) ou encore sur le développement de « prompts ».

Néanmoins, ces avancées ont été principalement testées dans le cadre de la langue anglaise et les travaux dans d'autres langues sont parcellaires ou incomplets. Au vu des multiples cas d'usage de ces travaux, il paraît important de développer et d'évaluer ces approches dans le cadre de la langue française.

2. Méthodologie

Nous proposons une analyse comparative d'un ensemble de méthodes pour l'apprentissage en zero/few shot learning adaptées au cas de la langue française. Nous considèrerons notamment les approches à base de prompts [5], les approches à base d'apprentissage multitâches [7,8] ainsi que les approches dites « parameter-efficient fine-tuning » [6].

Pour les approches nécessitant un modèle de langue pré-entrainé (prompt et parameter-efficient fine-tuning), nous utiliserons les modèles publiquement disponibles pour le français (BARThez [1], FlauBERT [2], CamemBERT [3]).

Les modèles multitâches seront quant à eux utilisés dans une version multilingue [7] ainsi que dans une version entraîné directement à partir de données anglaises traduites en français de manière automatique.

Nous considérerons les performances en zero ou few shot learning sur la base d'un ensemble de tâches, dont la classification de texte, la détection de paraphrases, le « question-answering » et le résumé de texte. Pour cela, nous utiliserons les protocoles d'évaluation issus de la littérature et disponibles publiquement pour la langue française : FLUE [2], FQuad [4] pour la tâche de « question-answering » et Orange-sum [1] pour la tâche de résumé de texte.

Enfin, les résultats seront comparés par rapport à des approches dites de « fine-tuning » (soit via les résultats de la littérature si des modèles existent pour la tâche donnée, soit en réentraînant les modèles à partir des données directement). Nous prendrons aussi en compte le coût en termes de temps de calcul dans la comparaison des différentes méthodes à la fois en terme d'entraînement et d'inférence.

3. Originalité / perspective

Ces travaux proposent une étude des approches de zero/few shot learning pour le NLU en français. Bien que ces techniques aient été largement étudiées dans le cadre de la langue anglaise, les travaux dédiés à la langue française restent limités à l'heure actuelle. De tels travaux sont pourtant nécessaires pour aborder les nombreux cas d'usage industriels pour lesquels l'usage de modèles en français est une nécessité. Ces travaux permettent donc d'évaluer la pertinence de ces techniques et d'établir de premières « baselines » pour un large ensemble de tâches classiques dans le domaine du NLU en français.

Un ensemble d'éléments techniques (code source) et de modèles pré-entrainés développés dans le cadre de cette étude seront mis librement à disposition de la communauté au travers de la plateforme Vaniila [9] développée par le CATIE.

Enfin, une partie de ces méthodes repose sur l'utilisation de modèles de langue restant lourd en termes de nombre de paramètres. Une perspective possible serait d'envisager l'utilisation de modèles pré-entraînés plus légers et de déterminer si l'usage de tels modèles plus petits impacte les performances des méthodes présentées.

Références

[1] Eddine, M. K., Tixier, A. J. P., & Vazirgiannis, M. (2020). Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

[2] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., ... & Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

[3] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2019). CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

[4] d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., & Vidal, M. (2020). FQuAD: French question answering dataset. *arXiv preprint arXiv:2002.06071*.

[5] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

[6] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.

[7] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., ... & Raffel, C. (2022). Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

[8] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

[9] Vaniila, Plateforme Visant l'Accompagnement de Nouveaux Intervenants dans l'Intelligence Artificille, https://www.vaniila.ai/