Few-shot learning with Large Language Models for Natural Language Inference tasks for French

Maximos Skandalis (LIRMM) Michael Sioutis (LIRMM)

Résumé

En traitement automatique du langage naturel (TALN), la tâche de l'inférence textuelle (*natural language inference* ou *recognising textual entailment*, en anglais) est une tâche de classification de paires de phrases (*sentence-pair classification task*) avec trois étiquettes/classes (*entailment*, *neutral*, *contradiction*).

Presque tous les modèles d'apprentissage profond pour cette tâche en français sont entraînés sur le sous-ensemble d'entraînement de XNLI, qui n'est qu'une traduction par machine au français du sous-ensemble d'entraînement de la version anglaise originale de XNLI.

Le projet de TER proposé consiste à étudier les possibilités du *few-shot learning* avec différents grands modèles de langage (GPT-5, Gemini 3, Mistral 3, CamemBERT) sur les nouveaux jeux de données que nous avons introduits pour le français (DACCORD, RTE3-FR, GQNLI-FR, SICK-FR, LingNLI-FR). L'objectif serait de tester à quelle mesure le *few-shot learning*, avec différentes combinaisons de modèles de langage et de jeux de données, pourrait influencer la performance des modèles sur tous les jeux de données disponibles pour le français. Il s'agirait aussi de voir si le paramétrage sur un (ou plusieurs) de nos jeux de données des modèles initialement entraînés sur XNLI est efficace pour améliorer les performances des modèles sur des exemples issus d'autres jeux de données distincts.

Les missions du TER comprennent :

1. paramétrer ou fine-tuner plusieurs modèles récents d'apprentissage profond sur nos jeux de données (par exemple, avec des méthodes PEFT déjà disponibles) ;
2. évaluer les performances de ces modèles sur tous les jeux de données mentionnés ;
3. tester des techniques de *few-shot learning* (par exemple, *Few-Shot Prompting*, *Chain-of-Thought (CoT) Prompting*) sur différents modèles et avec différents jeux de données parmi ceux qui sont indiqués ci-dessus.

1 Contexte

En traitement automatique du langage naturel (TALN), la tâche de l'inférence textuelle (connue en anglais sous les noms de *natural language inference* ou *recognising textual entailment*) est une tâche de classification de paires de phrases (*sentence-pair classification task*) avec, de préférence, trois étiquettes/classes (*entailment*, *neutral*, *contradiction*), alors que la tâche de détection automatique de contradictions est une tâche similaire de classification binaire (contradiction, pas de contradiction) de paires de phrases. Les métriques utilisées pour ces deux tâches sont traditionnellement celles d'accuracy et de score F1.

Presque tous les modèles d'apprentissage profond pour ces deux tâches en français sont entraînés sur le sous-ensemble d'entraînement de XNLI, qui est composé de 392702 paires de phrases mais qui n'est qu'une traduction par machine de l'anglais au français du sous-ensemble d'entraînement de la version initiale de XNLI en anglais. De ce fait, la généralisation des modèles vers d'autres jeux de données de nature différente que celui de l'entraînement n'est pas satisfaisante. Beaucoup de ces modèles sont disponibles sur huggingface.

Récemment, nous [8] avons construit DACCORD, un nouveau jeu de données avec 1034 paires de phrases pour la tâche de détection automatique de contradictions entre phrases en français. Nous [8] avons aussi produit des traductions pour plusieurs jeux de données de l'anglais en français pour la tâche d'inférence textuelle, à savoir RTE-3 (1600 paires de phrases), GQNLI (300 paires de phrases), SICK (9840 paires de phrases), et LingNLI.

2 Projet

Le projet de TER proposé consiste à étudier les possibilités du *few-shot learning* [2, 4, 5, 7, 9, 10] avec différents grands modèles de langage sur ces jeux de données récemment introduits. Une évaluation de certains LLMs (avant la sortie de Gemini 3 et Mistral 3) a déjà été effectuée par [1], mais sans *few-shot learning*. L'objectif ici serait de tester à quelle mesure le *few-shot learning*, avec différentes combinaisons de modèles de langage et de jeux de données, pourrait influencer la performance des modèles sur tous les jeux de données disponibles pour le français (XNLI, FraCaS, DACCORD [1], RTE3-FR [2], GQNLI-FR [3], SICK-FR [4], LingNLI-FR [5]). Il s'agirait aussi de voir si le paramétrage [6] sur un (ou plusieurs) de nos jeux de données des modèles initialement entraînés sur XNLI est efficace pour améliorer les performances des modèles sur des exemples issus d'autres jeux de données ayant leurs spécificités.

Les articles cités ici ont dans leur plupart déjà leur code disponible sur github (voir le lien github en note de bas de page dans chaque article).

3 Tâches

Les missions du TER comprennent :

1. paramétrer ou fine-tuner [6] [3] plusieurs modèles récents d'apprentissage profond (GPT-5, Gemini 3, FlauBERT, CamemBERT, mDeBERTa, XLM-R, DistilCamemBERT [7], Mistral 3) sur nos (un ou plusieurs chaque fois) jeux de données ;
2. évaluer les performances de ces modèles sur tous les 6 ou 7 jeux de données mentionnés ;
3. tester des techniques de *few-shot learning* [8] sur différents modèles et avec différents jeux de données parmi ceux qui sont indiqués ci-dessus.

1. https://huggingface.co/datasets/maximoss/daccord-contradictions
2. https://huggingface.co/datasets/maximoss/rte3-french
3. https://huggingface.co/datasets/maximoss/gqnli-fr
4. https://huggingface.co/datasets/maximoss/sick-fr-mt
5. https://huggingface.co/datasets/maximoss/lingnli-multi-mt
6. https://huggingface.co/blog/peft
7. https://huggingface.co/cmarkea/distilcamembert-base-nli.
8. https://skllm.beastbyte.ai/docs/few-shot-text-classification

4. Une étude de cas : Les 74 premiers exemples de FraCaS portent sur l'usage de quantificateurs généralisés. Ce serait pertinent d'essayer du *few shot learning* sur ces exemples, puis d'évaluer le modèle resultant sur GQNLI, un jeu de données plus compliqué que FraCas mais qui est lui aussi dédié aux quantificateurs généralisés. L'inverse (c.-à-d. *few shot learning* sur GQNLI, test sur ces 74 exemples de FraCas) est aussi à envisager comme expérience.

Co-Supervisors: Maximos Skandalis (LIRMM), and Michael Sioutis (LIRMM)

Ce sujet de TER est lié aux travaux d'une thèse co-financée par l'Agence de l'innovation de défense (AID) de la Direction générale de l'armement (DGA) et par l'Institut Cybersécurité Occitanie (ICO).

Références

[1] David Beauchemin, Yan Tremblay, Mohamed Amine Youssef, and Richard Khoury. Cole : a comprehensive benchmark for french language understanding evaluation, 2025.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora : Efficient finetuning of quantized llms, 2023.

[4] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics.

[5] Jiaxin Ge, Hongyin Luo, Yoon Kim, and James Glass. Entailment as robust self-learner. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 13803–13817, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc., 2022.

[7] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict : A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), 2023.

[8] Maximos Skandalis, Richard Moot, Christian Retoré, and Simon Robillard. New datasets for automatic detection of textual entailment and of contradictions between sentences in French. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12173–12186, Torino, Italia, May 2024. ELRA and ICCL.

[9] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv preprint arXiv :2104.14690*, 2021.

[10] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. More samples or more prompts ? exploring effective few-shot in-context learning for LLMs with in-context sampling. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics : NAACL 2024*, pages 1772–1790, Mexico City, Mexico, June 2024. Association for Computational Linguistics.