Topic: Adaptation and Distillation of Large Language Models

In the last decade, Language Models (LMs) have become increasingly more important in Natural Language Processing (NLP). Nowadays, large Language Models [6, 7] constitute the basis of any NLP application. Since the introduction of Transformer architectures, the size of LMs has increased exponentially, with models reaching billions of learnable parameters [1]. Training and deploying models of such dimensions is extremely resource-demanding, and the costs are often not affordable in real-world applications. Moreover, these models often need to be adapted (i.e. fine-tuned) for specific domains.

Knowledge Distillation [4, 5] aim at reducing the computational costs while retaining (almost) the performances. Distillation consists of producing a lighter model through a student-teacher learning process, where the student network learns to emulate the teacher's output distribution. Therefore, token embeddings are usually kept the same, despite constituting a significant portion of the model's parameters. Limited literature investigates distillation and domain adaptation together. In [3] different distillation, adaptation stages are explored, but tokenization aspects are not investigated.

The goal of this thesis is to design, implement and evaluate approaches for distilling large language models in domain adaptation setting. The candidate will work on strategies that aim to further reduce the model's sizes, with particular focus on investigating solutions for distilling models having different (domain-dependent) tokenizers. As a case of study we will consider the financial domain. Data is already available.
The proposed work consists of the following parts:

- Review of related works in Knowledge Distillation and Domain Adaptation.
- Preparation of data for pre-training and adaptation.
- Analysis of domain-specific data: impacts of domain-shift in tokenization.
- Definition and preparation of metrics and baselines to evaluate performances, taking also into account the demand of resources (e.g. GPU, inference time with and without GPUs etc..).
- Implementation of Distillation Technique(s) as baseline models.
- Research on Distillation techniques in domain Adaptation scenarios, aiming at either reducing model's sizes or improving model's performances using domain-specific tokenization.
- Evaluate performances on NER task in financial domain corpus.
- Compare and discuss the results, highlighting advantages, limitations and promising future directions.

The thesis must contain a detailed description of all developed and used algorithms as well as a profound result evaluation and discussion. The implemented code in Python has to be documented and released. Extended research on literature, existing patents, and related works in the corresponding areas must be performed.

References

[1] Brown et al.: *Language models are few-shot learners.* arXiv preprint arXiv:2005.14165, 2020.
[2] Shen, Sheng et al.: *Q-bert: Hessian based ultra low precision quantization of bert.* Proceedings of the AAAI Conference on Artificial Intelligence, 2020.
[3] Gordon, Mitchell A and Duh, Kevin: *Distill, adapt, distill: Training small, in-domain models for neural machine translation.* arXiv preprint arXiv:2003.02877, 2020.
[4] Sanh, Victor et al: *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv preprint arXiv:1910.01108, 2019.
[5] Jiao, Xiaoqi et al. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351, 2019
[6] Devlin, Jacob, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
[7] Liu, Yinhan, et al. *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.