Abstract of the dissertation: Natural Language Generation: Systems and Evaluation

In recent years, the Natural Language Generation (NLG) field has changed drastically. This shift, which can be partially attributed to the notable advance in hardware, led to recent efforts in NLG to be focused on data-driven methods leveraging large pretrained Neural Networks (NNs). However, this progress gave rise to new challenges related to computational requirements, accessibility, and evaluation strategies, to name a few. In this dissertation, we are primarily concerned with contributing to the efforts to mitigate these challenges.

To address the lack of monolingual generative models for some languages, we start by introducing ***BARThez*** and ***AraBART***, the first large-scale pretrained seq2seq models for French and Arabic, respectively. Being based on BART, these models are particularly well-suited for generative tasks. We evaluate BARThez on five discriminative tasks from the FLUE benchmark and two generative tasks from a novel summarization dataset, OrangeSum, that we created for this research. We show BARThez to be very competitive with state-of-the-art BERT-based French language models such as CamemBERT and FlauBERT. We also continue the pretraining of a multilingual BART on BARThez' corpus, and show our resulting model, mBARThez, to significantly boost BARThez' generative performance. On the other hand, We show that AraBART achieves the best performance on multiple abstractive summarization datasets, outperforming strong baselines.

Next, we study new ***adaptation strategies*** to reduce the computation time required to fully pretrain large language models. We show that our proposed strategies show on-par performance with a full pretraining from scratch, while providing a significant speed-up making it possible for the NLP community to pretrain their own models with relatively affordable resources.

Finally, we focus on the NLG system evaluation by proposing ***DATScore*** and ***FrugalScore***. DATScore uses data augmentation techniques to improve the evaluation of machine translation and other NLG tasks. Our main finding is that introducing data augmented translations of the source and reference texts is greatly helpful in evaluating the quality of the generated translation. We also propose two novel score averaging and term weighting strategies to improve the original score computing process of BARTScore. Experimental results on WMT show that DATScore correlates better with human meta-evaluations than the other recent state-of-the-art metrics, especially for low-resource languages. On the other hand, FrugalScore is an approach to learn a fixed, low-cost version of any expensive NLG metric while retaining most of its original performance. Experiments with BERTScore and MoverScore on summarization and translation show that FrugalScore is on par with the original metrics (and sometimes better), while having several orders of magnitude fewer parameters and running several times faster. On average overall learned metrics, tasks, and variants, FrugalScore retains 96.8% of the performance, runs 24 times faster, and has 35 times fewer parameters than the original metrics.

Moussa KAMAL EDDINE