### Internship Topic

Recently, pre-trained sequence-to-sequence models such as mBART, IndicBART, mT5 etc have helped improve natural language generation quality, however large models are difficult to train and deploy, especially because computing requirements become high. To this end, in this internship, we plan to explore two efficient approaches for sequence-to-sequence pre-training and fine-tuning.

The first approach falls under the data selection paradigm, where we seek to select representative data. Instead of pre-training large models on hundreds of millions of sentences, we suppose that selecting the most relevant data should help us train smaller models, faster while sacrificing a small amount of performance. As an initial approach, we plan on obtaining sentence representations and use distributed clustering methods to identify sentence groups and then select the most representative ones. We hope that this can help reduce the amount of training data by more than 80% which will need smaller and hence more efficient models.

The second approach is more relevant to fine-tuning, where instead of updating all model parameters we modify the inputs using a prompt, a random or learned set of embeddings, which is tuned for a fraction of the cost it would take to update the entire model. Prompt-based approaches have been used successfully in models such as GPT2 but not as much for natural language generation, and we plan to show that such approaches can have a huge impact in the case of machine translation, summarization and relevant sequence generation tasks.
募集情報

応募条件
・学年：大学院以上
・専攻学科等：計算機科学
・その他：自然言語処理、機械翻訳に関する知見がある方を優先的に採択する。

実施場所
ユニバーサルコミュニケーション研究所（けいはんな）

実施期間
2022年4月1日～2023年3月31日の間で60日間

受入予定人数
1名

選考課題
なし

備考