Segmentation thématique d'interviews politiques

Stage de fin d'études d'Ingénieur ou de Master 2 – Année académique 2023-2024

Mots clés : deep learning, large language models, spoken language understanding, natural language processing, machine learning, digital humanities

Contexte

L'Institut National de l'Audiovisuel (INA) est un établissement public à caractère industriel et commercial (EPIC), dont la mission principale consiste à sauvegarder et promouvoir le patrimoine audiovisuel français à travers la vente d'archives et la gestion du dépôt légal. Ce dernier permet aux chercheurs en sciences sociales, notamment dans le cadre de projets de recherche pluri-disciplinaires, de travailler sur les collections de l'INA. À ce titre, l'Institut capte en continu 180 chaînes de télévision et radio et stocke plus de 25 millions d'heures de contenu audiovisuel. L'INA assure également des missions de formation, de production et de recherche scientifique.

Le service de la recherche de l'INA mène depuis plus de 20 ans des travaux de recherche dans le domaine de l'indexation et de la description automatique de ces fonds selon l'ensemble des modalités : textes, sons et images. Le service participe à de nombreux projets collaboratifs de recherche que ce soit dans un cadre national et européen et accueille des stages de Master ainsi que des doctorants en co-encadrement avec des laboratoires nationaux d'excellence.

Ce stage est proposé au sein de l'équipe de recherche (https://recherche.ina.fr).

D'autres sujets de stage sont également proposés dans l'équipe :
https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/stages

Objectifs du stage

Dans le cadre d'un projet de recherche pluri-disciplinaire, nous cherchons à constituer un corpus conséquent d'interviews politiques annotées provenant d'émissions radiophoniques et télévisées. Ce corpus se distingue par une annotation fine de chaque émission selon deux axes principaux : un axe thématique et un axe reflétant la nature des tours de paroles selon un vocabulaire spécifique d'interruptions. Ce corpus annoté doit in fine servir de support à des études quantitatives portant notamment sur l'interaction entre intervieweurs et invité.e.s. en croisant ces divers axes et les métadonnées recueillies sur les participants.

L'objectif du stage consiste à mettre au point des méthodes de segmentation et de classification multi-labels des segments selon des catégories thématiques, en utilisant des méthodes d'apprentissage automatique. Les catégories ciblées appartiennent à un thésaurus spécifique du domaine politique français riche d'une centaine de termes et élaboré par nos partenaires chercheurs en sciences politiques.

Cette tâche s'appuiera sur les transcriptions timecodées générées par des outils d'ASR état de l'art ainsi que sur la diarisation de ces interviews.

Une partie du corpus, soient plusieurs dizaines d'interviews, a été annotée manuellement par ces mêmes partenaires selon les deux axes et constitue à la fois un ensemble d'apprentissage et d'évaluation de la tâche de segmentation/classification.

La segmentation et la classification de textes constituent des tâches fondamentales de l'IA pour lesquelles il existe une grande diversité d'approches reposant initialement sur des approches lexicales statistiques (TF-IDF, Bag of words, Latent Dirichlet Allocation) et évoluant progressivement vers l'utilisation de représentations sémantiques.

Ces dernières prennent la forme d'espaces vectoriels dits de plongement ("embedding"), tout d'abord du mot seul (word2vec) puis du mot dans le contexte de la phrase. Des architectures neuronales comme les RNN puis, plus récemment, les encodeurs bi-directionnels possédant des mécanismes d'attention sont utilisés pour générer ces représentations sémantiques prenant en compte le contexte.

Appliquées aux problèmes de la segmentation et de la classification de textes, ces approches ont donné lieu à des solutions telles que textTiling[1], TopicTiling[2] et plus récemment [3][4] et [5].

Les tâches de similarité et de classification de texte sont toutes deux fortement dépendantes de l'adaptation au domaine étudié de l'espace de plongement utilisé pour encoder les parties de textes.

Certains travaux récents proposent d'adapter les encodeurs de type BERT existants en utilisant une ou plusieurs stratégies comme le réentraînement non supervisé sur des textes du domaine [9], l'utilisation d'apprentissage contrastif utilisant des réseaux siamois (SBERT, SimCSE) [6][7] ou encore, pour le cas de la classification, d'approches semi-supervisées utilisant une définition des classes sous forme de mots-clés dans la construction de l'espace de plongement [8].

La difficulté de la tâche visée réside à la fois dans, 1) la nature du texte à annoter thématiquement : traiter automatiquement une transcription d'interviews représente une tâche bien plus difficile en comparaison avec de la parole préparée (i.e un sujet de JT), 2) le (relativement) peu de données annotées disponibles au regard de la taille du vocabulaire cible, ainsi que d'un déséquilibre significatif des classes et 3) l'annotation multi-labels des annotations, tâche plus complexe que le cas multi-classes.

Pour remédier à ces difficultés une tâche de collecte de données destinée à disposer de davantage de données d'apprentissage sera mise en place. Plusieurs sources seront explorées : 1) Des retranscriptions de sujets de JT annotés selon le thésaurus thématique de l'INA. Il sera nécessaire de trouver un mapping entre les catégories cherchées du thésaurus cible et les catégories et descripteurs INA servant à la documentation du fonds ; 2) le scrapping de données textuelles du Web illustrant ces catégories.

Les principales tâches envisagées sont les suivantes :

- bibliographie et état de l'art sur la segmentation et classification thématique de textes ;

- recueil de données selon les pistes évoquées dans le paragraphe précédent ;
- élaboration de plusieurs architectures de réseaux de type transformers et/ou LSTM utilisant notamment des modèles français de type BERT. Les pistes d'optimisation de l'espace de plongement de ces architectures à la tâche et au domaine seront étudiées, notamment en s'appuyant sur les travaux décrits précédemment ;
- évaluation des résultats sur une partie du corpus annotée

Valorisation du stage

Différentes stratégies de valorisation des travaux seront envisagées, en fonction de leur degré de maturité et des orientations envisagées pour la suite des travaux :

- Diffusion des outils d'analyse réalisés sous licence open-source via le dépôt GitHub de l'INA : https://github.com/ina-foss
- Rédaction de publications scientifiques

Conditions du stage

Le stage se déroulera sur une période de 4 à 6 mois, au sein du service de la Recherche de l'Ina. Il aura lieu sur le site Bry 2, situé au 28 Avenue des frères Lumière, 94360 Bry-sur-Marne. La·le stagiaire sera encadré·e par Steffen LALANDE. Un ordinateur sera fourni ainsi qu'un accès au cluster de calcul GPU de l'Institut.

Gratification : 760 € brut / mois + 50 % pass navigo

Candidature

Envoyez par email et en PDF aux adresses slalande@ina.fr et abeloued@ina.fr, les documents suivants : CV et relevés de notes + liste des enseignements suivis en M2 et en M1 ou en école d'ingénieurs. Précisez le sujet du stage pour lequel vous candidatez.

Profil recherché

- Étudiant·e en dernière année d'un bac +5 dans le domaine de l'informatique et de l'IA
- Forte appétence pour la recherche académique
- Intérêt pour les sciences sociales computationnelles
- Maîtrise de Python et expérience dans l'utilisation de bibliothèques de ML/NLP
- Capacité à effectuer des recherches bibliographiques
- Rigueur, synthèse, autonomie, capacité à travailler en équipe

Bibliographie

[1] Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23, 1: 33–64.

[2] Martin Riedl and Chris Biemann. 2012. TopicTiling: A Text Segmentation Algorithm based on LDA. In Proceedings of ACL 2012 Student Research Workshop, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

[3] Glavas, Goran and Swapna Somasundaran. "Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation." ArXiv abs/2001.00891 (2020): n. pag.

[4] Michael Lukasik, Boris Dadachev, Gonc¸alo Simoes and Kishore Papineni. 2020a. Text segmentation by cross segment attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4707–4716.

[5] Iacopo Ghinassi. 2021. Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. In 2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021) at the ACM 417 International Conference on Interactive Media Experiences (IMX 2021) (DataTV-2021).

[6] Reimers, Nils and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” Conference on Empirical Methods in Natural Language Processing (2019).

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[8] Tim Schopf, Daniel Braun, and Florian Matthes. 2023. Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches. In Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR '22). Association for Computing Machinery, New York, NY, USA, 6–15.

[9] Guillaume Lefebvre, Haytham Elghazel, Théodore Guillet, Alexandre Aussem, & Matthieu Sonnati (2023). BERTEPro : Une nouvelle approche de représentation sémantique dans le domaine de l'éducation et de la formation professionnelle. Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances, RNTI-E-39, 211-222.