Détection automatique de citations dans les transcriptions TV et radio

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) et se place dans le cadre de deux projets collaboratifs financés par l'ANR : Medialex et Pantagruel.

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

Objectifs du stage

La détection de citations consiste à localiser des passages de paroles citées dans un texte. Il s'agit d'une tâche très utile notamment dans une perspective de « fact-checking » [11] et d'analyse sociologique des médias [12].

Selon [3], la citation n'est pas un phénomène linguistique simple et n'a pas été largement étudiée pour le français, bien qu'il y ait eu un regain d'intérêt pour son étude ces dernières années. Il y a donc très peu de corpus disponibles pour travailler sur le sujet. Cette tâche vise à détecter les portions de texte qui correspondent au contenu d'une citation dans un texte. D'apparence simple, elle l'est en fait beaucoup moins : une citation peut être annoncée par un indice et peut ou non être placée entre guillemets - parfois, elle contient également des guillemets trompeurs. Elle peut être très longue et discontinue, ou se superposer à un élément de repère. Les citations sont généralement divisées en trois types : directes (entre guillemets), indirectes (paraphrasant des mots ou des phrases) et non directes et mixtes ou partiellement indirecte (une combinaison des deux). Tous ces types de citations peuvent être trouvés indifféremment dans différents types de textes, dans des textes littéraires ou dans

des documents d'information, ce qui rend leur identification automatique d'autant plus difficile.

Nous nous situons dans le cadre de l'analyse des discours et des informations provenant d'émissions radiophoniques et télévisées. Une spécificité de ce contexte est que nous travaillons donc sur de la parole orale et non sur des textes écrits. Nous disposons pour cela de transcriptions timecodées générées par des outils d'ASR état de l'art. Toutefois, la présence d'indices tels que les guillemets n'existe pas dans ces textes transcrits et les codes syntaxiques visant à introduire une citation peuvent différer des pratiques utilisées typiquement pour la presse écrite. De plus, deux difficultés supplémentaires à la langue orale sont la détection de la fin de la citation et les aléas liés aux erreurs potentielles de transcription des logiciels d'ASR.

Dans ce context, nous nous proposons d'étudier des méthodes d'étiquetage de séquences : CRF [10] et modèle semi-markovien [9]. Nous accorderons également de l'importance aux approches basées sur les LLMs.

Ainsi les principales tâches envisagées sont les suivantes :

- bibliographie et état de l'art sur la détection de citations directes et indirectes
- expérimentation et évaluation de quelques approches sur un corpus de presse écrite
- adaptation de l'approche et expérimentation sur un corpus audiovisuel
- expérimentation avec et sans bibliothèque de citations pré-existante
- stratégie de création de corpus d'entraînement et/ou d'évaluation
- évaluation de l'impact du système de transcription sur les performances
- évaluation de l'adaptation du LLM à la langue orale sur les performances

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 Émile Chapuis. 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 echapuis@ina.fr et nherve@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
- Bonne compétences en programmation

- Maîtrise de Python et expérience dans l'utilisation de bibliothèques de ML/NLP (Sklearn, Pytorch)
- Capacité à effectuer des recherches bibliographiques
- Rigueur, synthèse, autonomie, capacité à travailler en équipe

Bibliographie

[1] V.-G. Soumah, P. Rao, P. Eibl, et M. Taboada, « ***Radar de Parité: An NLP system to measure gender representation in French news stories*** ». 2023.
[2] S. Song, H. Song, K. Park, J. Han, et M. Cha, « ***Detecting contextomized quotes in news headlines by contrastive learning*** ». 2023.
[3] A. Richard, L. Alonzo-Canul, et F. Portet, « ***FRACAS: A FRench annotated corpus of attribution relations in newS*** ». 2023.
[4] A. Kathirgamalingam, F. Lind, et H. G. Boomgaarden, « ***Automated detection of voice in news text – evaluating tools for reported speech and speaker recognition*** », Computational Communication Research, vol. 5, nº 1. Amsterdam University Press, p. 85, 2023.
[5] M. Janicki, A. Kanner, et E. Mäkelä, « ***Detection and attribution of quotes in Finnish news media: BERT vs. rule-based approach*** », in Proceedings of the 24th nordic conference on computational linguistics (NoDaLiDa)
[6] R. Gangi Reddy, H. Elfardy, H. P. Chan, K. Small, et H. Ji, « ***SumREN: Summarizing Reported Speech about Events in News*** », AAAI, vol. 37, nº 11, p. 12808-12817, juin 2023
[7] A. Richard, G. Bastin, et F. Portet, « ***GenderedNews: Une approche computationnelle des écarts de représentation des genres dans la presse française*** ». 2022.
[8] N. Hervé et al., « ***Using ASR-Generated text for spoken language modeling*** », in « Challenges & Perspectives in Creating Large Language Models », ACL 2022 workshop, mai 2022.
[9] Scheible et al., « **Model Architectures for Quotation Detection** », in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.
[10] Pareti et al., « **Automatically Detecting and Attributing Indirect Quotations** », Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.
[11] Govaert et al., « **Deceptive journalism: Characteristics of untrustworthy news items** », Journalism Practice, 2020.
[12] Jiyoung Han and Gunho Lee, « **A comparative study of the accuracy of quotation-embedded headlines in chosun ilbo and the new york times from 1989 to 2009** », Korea Journal, 2013