Exploiter la presse écrite pour l'extraction des séquences audiovisuelles liées à un événement d'actualité

Marjolaine Ray*, Thierry Poibeau*
Sylvain Parasie**, Nicolas Hervé***
Béatrice Mazoyer**

*Lattice, ENS-PSL, 1 Rue Maurice Arnoux, 92120 Montrouge, France
marjolaineray.me@gmail.com, thierry.poibeau@ens.psl.eu
**médialab, SciencesPo, 84 Rue de Grenelle, 75007 Paris, France
***INA, 18 Avenue des frères Lumière, 94366 Bry-sur-Marne, France

Dans cet article, nous présentons une méthode d'extraction de segments dans des journaux télévisés et radiophoniques en utilisant la presse écrite comme ancrage dynamique. Cette méthode [1] a été élaborée dans le cadre du projet ANR Médialex, qui vise à renouveler la compréhension des dynamiques d'influence entre les agendas parlementaires, médiatiques et citoyens. L'un des cas d'étude du projet concerne la couverture médiatique de la mort de Nahel Merzouk et des révoltes qui ont suivi. La tâche est modélisée comme un calcul de similarité entre une requête textuelle et une fenêtre glissante parcourant la transcription des émissions.

Données Les données de journaux radio et télévisuels (JTs) proviennent des transcriptions automatiques de journaux d'information diffusés entre le 27 juin et le 3 juillet 2023 sur 27 chaînes de radio et télévision [2]. L'ensemble totalise 925 heures d'émissions. Ces transcriptions ont été réalisées avec le logiciel de reconnaissance et de transcription automatique de la parole Vocapia et mises à disposition par l'Institut National de l'Audiovisuel (INA). Nos données de presse sont issues du projet OTMedia (Viaud et al., 2018) et couvrent 433 titres de presse.

Méthodologie Pour sélectionner les extraits effectivement liés à notre thématique, nous avons testé plusieurs méthodes, toutes fondées sur la correspondance entre une *requête*, liée à notre sujet, et un segment du texte des transcriptions. La première méthode est une simple recherche par mots-clés (tels que "Nahel", "Nael", "refus d'obtempérer", "émeutes", etc.). La deuxième méthode repose sur un plongement de type Sentence-BERT (SBERT) (Reimers et Gurevych, 2019) issu du modèle CamemBERT (Martin et al., 2020). La requête devient alors la représentation vectorielle d'un résumé (généré par GPT-4o) décrivant les circonstances et les répercussions de l'affaire Nahel. Notre troisième méthode se fonde sur la presse écrite pour obtenir une nouvelle requête chaque jour. Nous utilisons les plongements SBERT des articles de la presse du jour, sélectionnés car ils contiennent un des mots-clés évoqués ci-dessus. Nous tenons ainsi compte de la nature changeante de l'événement en le caractérisant différemment

1. https://github.com/Hortatori/slide_extract
2. Arte, BFM TV et radio, C8, CNews, Europe 1, France 2, France 3, France Inter, France 5, FranceBleu Régions, France Culture, FranceInfo TV et Radio, LCI, LCP, M6, Radio Classique, RFI, RMC, RTL, TF1, TMC

chaque jour. Chaque requête composée de plongements lexicaux (pour les méthodes 2 et 3) est ensuite comparée à une minute de JT, qui est sélectionnée si elle est suffisamment similaire.

Résultats et Discussion Les méthodes sont évaluées à partir de segments annotés manuellement par l'INA [3] et nous-mêmes (TF1, France 2, Europe 1 et CNews). Les scores de précision

Texte de requête	Accuracy	Score F1	Precision	Recall
mots-clés	0.72	0.35	0.98	0.21
unique résumé	0.92	0.90	0.85	0.95
articles quotidiens	**0.94**	**0.92**	0.90	0.95

TAB. 1 – *Performances des extractions selon le texte de requête (le seuil optimal de chaque méthode est celui utilisé pour l'évaluation).*

globale (accuracy) et de F-mesure (F1) sont plus élevés pour les articles que pour les résumés (tableau 1). Nos résultats montrent que la presse écrite apporte une meilleure couverture de la diversité des faits abordés, ainsi qu'une détection plus précoce (dès le premier jour, alors que le nom de "Nahel" n'est pas encore utilisé à la télévision).

Conclusion et Perspectives À représentation égale (des plongements lexicaux identiques), l'usage de requêtes issues d'articles quotidiens surpasse un résumé global unique généré par un LLM. La méthodologie présentée dans cet article est transférable, mais n'a pu être évaluée sur d'autres événements. Une telle évaluation exige de définir une procédure à grande échelle, qui n'a pu être mise en place ici. D'autre part, nous avons étudié un événement passé (en 2023), pour lequel il est facile d'obtenir un résumé via un LLM en ligne. En temps réel, de tels LLMs souffrent de leur absence de mise à jour des informations récentes, au contraire des textes de presse, utilisés à la volée, qui sont par nature à jour des informations les plus récentes.

Références

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. Sagot (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, doi: 10.18653/v1/2020.acl-main.645.

Reimers, N. et I. Gurevych (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992.

Viaud, M.-L., A. Saulnier, N. Hervé, B. Renoust, et J. Thièvre (2018). OTMedia : Outils de fouille multimodales transmedia de l'actualité. In *En Quête d'archives : Bricolages Méthodologiques En Terrains Médiatiques*. Ina Editions.

3. `https://catalogue.ina.fr/`