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
sylvain.parasie@sciencespo.fr, beatrice.mazoyer@sciencespo.fr
***INA, 18 Avenue des frères Lumière, 94366 Bry-sur-Marne, France
nherve@ina.fr

Résumé. L'identification automatique de séquences audiovisuelles pertinentes au sein de flux de journaux télévisés ou radiophoniques constitue un enjeu majeur pour l'analyse de contenu médiatique à grande échelle. Dans cet article, nous présentons une méthode d'extraction de segments dans des journaux télévisés et radiophoniques, développée à partir d'un cas d'étude consistant à retrouver toutes les séquences médiatiques traitant de la mort de Nahel Merzouk en 2023. La tâche est modélisée comme un calcul de similarité sémantique entre une requête textuelle et une fenêtre glissante appliquée à la transcription des émissions. Nos expériences s'appuient sur un corpus de transcriptions de chaînes de radio et de télévision dont une partie est annotée pour la segmentation du contenu. Les résultats montrent que des articles de presse quotidiens constituent des requêtes plus efficaces que des résumés pour identifier les segments pertinents.

1 Introduction

Alors que pendant près de deux décennies, l'intérêt des chercheurs en sciences sociales pour la télévision et la radio a connu un recul important, au bénéfice du numérique (Bourdon, 2018), on observe aujourd'hui un intérêt renouvelé pour l'étude de ces médias traditionnels par les sciences sociales. La principale raison étant que la télévision et la radio demeurent les premières sources d'information des Français, selon un rapport [1] de l'Autorité publique française de régulation de la communication audiovisuelle et numérique (ARCOM). Le premier mode d'accès à l'information est d'abord la télévision (à 80 %) et la radio en seconde place (37 %). De ce fait, le traitement de l'information par ces médias est une problématique récurrente dans l'analyse de contenu médiatique, et un sujet d'étude privilégié pour la recherche sociologique. Cependant, les émissions d'information produisent des volumes de données considérables, qui sont de mieux en mieux retranscrites mais qui sont fastidieuses à traiter manuellement.

1. Les Français et l'information, rapport de l'ARCOM, mars 2024 : https ://www.arcom.fr/se-documenter/etudes-et-donnees/etudes-bilans-et-rapports-de-larcom/les-francais-et-linformation

Dans cette contribution, nous nous focalisons sur un problème en particulier, qui concerne l'identification des séquences audiovisuelles liées à un événement d'actualité. Comment peut-on extraire de façon automatique un corpus pertinent à partir d'un ensemble très volumineux de retranscriptions du flux télévisuel ou radiophonique ? Dans la mesure où les transcriptions d'émissions ne sont pas segmentées en paragraphes, il n'est pas possible avec une simple recherche par mots-clés de délimiter les séquences associées à un événement d'actualité. Nous décrivons ici une méthode spécifique[2], qui a été élaborée et mise à l'épreuve dans le cadre d'un projet visant à observer les dynamiques d'influence entre les agendas parlementaires, médiatiques et citoyens. L'un des cas d'études du projet concerne la couverture médiatique, politique et publique de la mort de Nahel Merzouk, le 27 juin 2023, et des révoltes qu'elle a déclenchées. Dans ce cadre, nous avons développé une méthode permettant d'extraire toutes les séquences se référant à ces événements dans les transcriptions des journaux télévisés diffusés pendant cette période.

2 Cas d'étude : la mort de Nahel Merzouk

La mort de Nahel Merzouk est un événement notable dans la sphère médiatique d'une part en raison de la forte polarisation sociale qu'elle a entraînée et, d'autre part, en raison de sa forme : un enchaînement d'événements qui amplifient la portée sociale de l'événement initial.

Le 27 juin 2023 au matin, Nahel Merzouk est tué par un tir policier lors d'un refus d'obtempérer. La version de la légitime défense est rapidement contestée par les témoignages des passagers et par la diffusion, sur Twitter, d'une vidéo montrant le tir. Le soir même, des actes de révolte éclatent en Île-de-France et se prolongent jusqu'au 2 juillet, en s'étendant à d'autres régions. Le 29 juin, le policier auteur du tir est placé en détention provisoire. Le débat public autour de l'affaire se polarise progressivement entre les soutiens du policier, qui ouvrent une cagnotte à son intention, et ceux de Nahel et de sa famille, qui organisent une marche blanche le 30 juin. Le 1er juillet, la cagnotte dépasse le million d'euros tandis que la mère de Nahel organise un rassemblement d'appel au calme.

Lorsque les journaux télévisés ou radio rapportent un de ces sous-événements, ils ne mentionnent pas toujours les mots "Nahel" ou "émeutes". Cette fragmentation médiatique nuit particulièrement à la recherche par mots-clés, ce qui a motivé la mise en place d'une méthodologie s'appuyant sur la similarité sémantique entre plongements lexicaux. Le résultat de cette extraction est ensuite destiné à une analyse qualitative de l'événement.

3 État de l'art

La tâche qui nous intéresse ici peut être décomposée en deux tâches importantes en traitement automatique des langues : une tâche de segmentation thématique (Topic Segmentation), qui consiste à diviser un flux (textuel ou audiovisuel) en unités homogènes du point de vue du sujet qu'elles abordent, et une tâche de recherche d'information (Information Retrieval), dont l'objectif est d'optimiser la récupération des documents pertinents au sein d'un corpus à partir d'une requête utilisateur.

2. Le code est librement accessible sur github : `https://anonymous.4open.science/r/slide_extract-FE6C`

La segmentation de texte est une tâche déjà bien connue et décrite (Ghinassi et al., 2024). Toutefois, en ce qui concerne les flux audiovisuels, l'automatisation est encore difficile. Jusqu'à une période récente, la segmentation de flux audio reposait sur des éléments prosodiques plutôt que sur des informations sémantiques (Berlage et al., 2020), mais les progrès accomplis dans le domaine de la reconnaissance automatique de la parole (transcription) permettent désormais de bien segmenter en unités thématiques, au point que les méthodes reposant sur des données multimodales semblent pour le moment faire moins bien que celles s'appuyant sur le texte seul (Ghinassi et al., 2023; Shukla et al., 2024).

Notre étude se donne un objectif plus simple que celui de la segmentation complète des données, puisqu'il s'agit d'identifier uniquement les séquences où le sujet d'intérêt est abordé. Pour autant, il semble que peu de travaux aient abordé ce domaine, à l'exception d'articles s'intéressant à la catégorisation des séquences en thèmes prédéfinis (Leopold et Kindermann, 2006; Pelloin et al., 2022, 2024) tels que l'économie, la santé, l'environnement, le sport, etc. Ces méthodes fondées sur l'entraînement de classificateurs restent toutefois assez éloignées de notre tâche, puisque nous cherchons à extraire les séquences relatives à n'importe quel événement médiatique, sans avoir à ré-entraîner un classificateur pour chaque nouveau cas.

4 Méthodologie

Les données utilisées 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[3]. 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. Ce logiciel propriétaire utilise un processus de diarisation des locuteurs visant à diviser le flux audio en segments homogènes selon l'identité du locuteur. Cette segmentation est complétée par un découpage selon les pauses des locuteurs, correspondant à un tour de parole par ligne. Comme tout logiciel de transcription automatique, ce dernier est susceptible de commettre des erreurs, notamment sur l'orthographe des noms propres. Nous en avons tenu compte pour la sélection de nos mots-clés. Pour sélectionner les extraits effectivement liés à notre thématique, nous avons comparé plusieurs méthodes, toutes fondées sur la correspondance entre une requête, liée à notre sujet, et un segment du texte des transcriptions.

4.1 Première méthode : extraction par mots-clés

La première méthode est une simple recherche par mots-clés[4] : on sélectionne alors les tours de parole qui contiennent l'un de ces mots, sans pouvoir réellement définir la séquence temporelle de l'intervention. Cette méthode ne permet pas de relever les commentaires autour des mentions de l'événement bien qu'ils soient cruciaux pour une analyse sociologique du contenu médiatique. Pour remédier à ce problème, les méthodes suivantes mobilisent deux ressources : la prise en compte du flux temporel de la transcription et la représentation vectorielle du texte par des plongements lexicaux.

3. 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

4. "jeune conducteur", "refus d'obtempérer", "émeutes", "émeutiers", "cagnotte", "l'haÿ-les-roses", "lhaylesroses", "violences", "nanterre", "clichy-sous-bois", "naël", "nahel"

4.2 Plongements lexicaux et choix du découpage temporel

Plongements lexicaux Les deuxièmes et troisièmes méthodes exploitent des représentations par plongements lexicaux contextuels qui représentent les mots selon leur environnement syntaxique et sémantique. Le modèle retenu dans notre étude est un plongement de type Sentence-BERT (SBERT). Il est issu du modèle BERT (Devlin et al., 2019) qui repose sur une architecture Transformer bidirectionnelle mais qui ne se prête pas originellement aux mesures de similarité entre phrases. Sentence-BERT (Reimers et Gurevych, 2019) reformule BERT en ajoutant une couche de pooling pour obtenir un vecteur par phrase, pré-entraîné grâce à un réseau siamois. Nous utilisons un modèle pré-entraîné issu de CamemBERT (Martin et al., 2020) et adapté pour Sentence-BERT[5]. Les plongements de phrases ainsi obtenus facilitent le calcul d'une similarité entre phrases pour le français.

Découpage temporel La segmentation automatique des flux de parole par le logiciel Vocapia génère des durées de tour de parole très variables. Ainsi, un tour de parole de notre corpus correspond en moyenne à une durée de 19 secondes, mais cette durée varie drastiquement selon la prosodie des locuteurs (tableau 1). Pour atténuer ces variations, une fenêtre glissante est créée à partir d'un ensemble de tours de parole de la transcription, concaténées pour atteindre une fenêtre d'une minute de temps de parole. Cette fenêtre est décalée progressivement sur chaque média et sera comparée à chaque itération à une requête par le calcul d'un score de similarité cosinus.

Durée moyenne	Variance	Durée minimale	Durée maximale
19 sc	540 sc	0.18 sc	1582 sc

TAB. 1 – *Variation de durées en secondes des tours de parole.*

4.3 Deuxième méthode : requête par résumé

Nous appelons deuxième méthode le choix d'une requête unique : il s'agit d'un texte résumant les circonstances et les répercussions de l'affaire Nahel généré par un LLM (GPT-4o)[6]. Cette deuxième méthode se définit comme suit : **(1)** un plongement lexical est calculé pour le texte de résumé, **(2)** la distance de similarité entre les embeddings du résumé et les embeddings de la minute glissante est calculée et **(3)** toutes les minutes de texte de transcription suffisamment similaires à la requête, c'est à dire dont la distance de similarité dépasse un seuil (défini en Figure 1) sont sélectionnées. Cette méthode a pour limite de n'utiliser qu'une requête générale sur toute la période, alors que les éléments d'un événement médiatique apparaissent et évoluent au cours du temps.

5. disponible sur https ://huggingface.co/Lajavaness/sentence-camembert-large

6. Les requêtes utilisées sont les suivantes : "fais moi un résumé de l'affaire Nahel qui a eu lieu en france en 2023.","peux tu développer au sujet des indignations sur les violences policières ?","développe le sujet des violences urbaines avec un langage oral."

4.4 Troisième méthode : requête par article de presse

Notre dernière méthode se fonde sur la presse écrite comme ancrage quotidien, permettant de faire évoluer la requête. Nos données couvrent 433 titres de presse et sont issues du projet OTMedia (Viaud et al., 2018). La troisième méthode se définit comme suit : **(1)** les articles de la presse du jour sont sélectionnés s'ils contiennent des mots-clés prédéfinis [7] **(2)** les embeddings des titres et débuts de chaque article sont calculés à partir du modèle SBERT, **(3)** la distance de similarité entre les embeddings de presse et les embeddings de chaque fenêtre télévisuelle est calculée et **(4)** la minute de texte est sélectionnée si la distance dépasse un seuil (défini en Figure 1). Nos requêtes de presse peuvent contenir entre 300 et 3000 articles selon l'évolution du traitement médiatique. Le changement de requête permet de tenir compte de la nature changeante de l'événement en le caractérisant différemment chaque jour. Cette méthode permet également d'avoir une définition homogène de l'événement entre presse écrite et audiovisuelle.

5 Résultats

Le travail de segmentation manuel réalisé par les documentalistes [8] de l'INA sur certaines chaînes (TF1 et France 2) permet d'évaluer les trois méthodes avec une vérité de terrain. Ce corpus a été étendu avec une annotation de Europe1 et CNews par segments d'une heure par jour. En utilisant la presse écrite comme texte de référence, nous obtenons des scores de préci-

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. 2 – *Performances des extractions selon le texte de requête.*

sion globale (accuracy) et de F-mesure (F1) plus élevés pour les articles que pour les résumés, ces mesures évaluant respectivement la proportion de prédictions correctes et l'équilibre entre précision et rappel (tableau 2). Après avoir observé que les seuils de similarité optimaux étaient différents pour les deux types de requête (presse ou résumé), les scores reportés ici sont basés sur le seuil optimal pour chaque type de texte de requête (voir Figure 1).

Nos résultats montrent que la presse écrite peut jouer un rôle d'ancrage évolutif pour améliorer sensiblement l'extraction de segments dans des flux audiovisuels. Cette approche apporte une meilleure couverture de la diversité des faits abordés et des types d'interventions, ainsi qu'une détection plus précoce (dès le premier jour, alors que le nom de "Nahel" n'est pas encore utilisé à la télévision).

L'évaluation de la durée des séquences dédiées à l'affaire Nahel dans les journaux télévisés (tableau 3) permet également de prendre la mesure de la couverture médiatique de l'événement, exceptionnelle par son ampleur (en pourcentage du temps de parole total) et par sa durée (le sujet est encore très présent à l'antenne après une semaine). Le temps d'antenne dédié à Nahel

7. mots-clés "nahel", "nanterre", "émeutes"
8. `https://catalogue.ina.fr/`

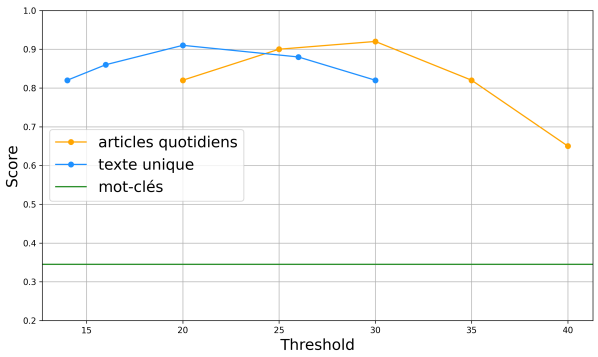

FIG. 1 – *Les scores F1 des extractions varient selon les seuils de similarité, le seuil optimal de chaque méthode est celui utilisé pour l'évaluation finale.*

Jour	**Durée totale**	**Durée dédiée à Nahel**	**Proportion (%)**	**Nombre d'articles utilisés en requête**
2023-06-27	137h 1m	12h 18m	8.99 %	332
2023-06-28	137h 26m	41h 42m	30.35 %	1258
2023-06-29	137h 54m	51h 7m	37.07 %	2183
2023-06-30	135h 21m	46h 17m	34.2 %	3324
2023-07-01	121h 54m	35d 35m	29.2 %	1933
2023-07-02	121h 23m	34h 34m	28.49 %	1526
2023-07-03	134h 21m	24h 37m	18.33 %	2888

TAB. 3 – *Temps d'antenne dédié à Nahel par jour : comparaison entre le corpus complet et l'extraction obtenue par une requête de presse quotidienne, pour la totalité des chaînes.*

selon l'extraction par requête de presse montre des durées impressionnantes, mais ceci reflète la proportion importante des chaînes d'information en continu dans le corpus. Les chaînes d'actualité en continu représentent 52 % du texte de notre corpus et, de part leur objectif rédactionnel, elles choisissent des formats qui privilégient la répétition des événements les plus récents (Bucy et al., 2007). En figure 2, l'analyse du temps pour chaque chaîne révèle la part importante des chaînes d'information en continu dans le traitement de l'événement.

6 Perspectives et Limitations

Évaluation La méthodologie présentée dans cet article est transférable, mais n'a pu être évaluée sur d'autres événements, ce qui limite notre estimation des performances à grande échelle. Nos premières observations (sur les variations de valeurs de similarité en fonction du temps entre plusieurs événements) semblent valider l'hypothèse d'une application pour tous types d'événements, mais nécessitent une étude plus poussée.

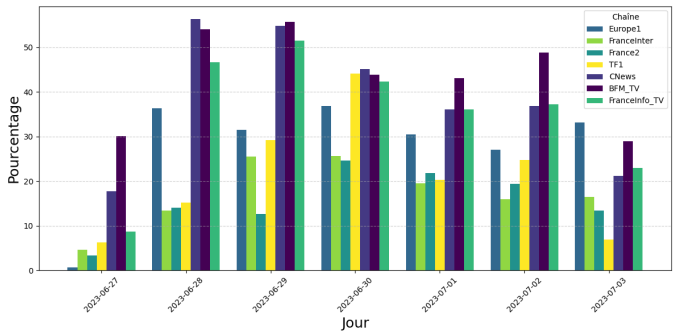

FIG. 2 – *Pourcentage des interventions dédiées à Nahel et aux émeutes sur le temps d'antenne de chaque chaîne.*

Perspectives à propos du requêtage par LLMs L'écart de performance entre les articles de presse quotidiens et le résumé généré par GPT-4o est de deux points seulement. Ceci est concordant avec la qualité des requêtes "zero-shot" produites par les LLMs, qui reflète la connaissance du monde obtenue par leur entraînement, surtout lorsqu'elle est complétée par une recherche internet (comme pour chatGPT). Néanmoins, la faiblesse des LLMs vient de leur absence de mise à jour des informations récentes qui les rend moins fiables dans le cas d'une extraction en temps réel à partir de flux d'informations. Les textes de presse, utilisés à la volée, sont par nature à jour des informations les plus récentes. Une analyse approfondie des types d'erreurs produites par les deux méthodes pourrait apporter un éclairage sur la qualité des LLMs dans ce domaine.

7 Conclusion

Cette étude introduit une méthode reliant presse écrite et transcriptions audiovisuelles pour extraire automatiquement les séquences liées à un événement. À représentation égale (plongements lexicaux), l'usage de requêtes issues d'articles quotidiens surpasse un résumé global unique, car il suit l'évolution des cadrages et des sous-événements et améliore la détection des segments pertinents. Cette approche offre un cadre opérationnel pour constituer des corpus ciblés à grande échelle et outiller des analyses médiatiques sensibles au tempo de l'actualité. Nos résultats restent néanmoins limités à un seul cas et à des paramètres fixés ; des validations sur d'autres événements renforceraient la robustesse et la portée de la méthode. En résumé, on peut dire que la presse fournit une source simple et efficace pour rapprocher textes journalistiques et flux audiovisuels dans une perspective d'extraction et d'analyse dynamique.

Références

Berlage, O., K.-M. Lux, et D. Graus (2020). Improving automated segmentation of radio shows with audio embeddings. In *ICASSP 2020-2020 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pp. 751–755. IEEE.

Bourdon, J. (2018). Is the end of television coming to an end ? *VIEW Journal of European Television History and Culture 7*, 80, doi: 10.18146/2213-0969.2018.jethc144.

Bucy, E., W. Gantz, et Z. Wang (2007). Media technology and the 24 hour news cycle.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). Bert : Pre-training of deep bidirectional transformers for language understanding.

Ghinassi, I., L. Wang, C. Newell, et M. Purver (2023). Multimodal topic segmentation of podcast shows with pre-trained neural encoders. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pp. 602–606.

Ghinassi, I., L. Wang, C. Newell, et M. Purver (2024). Recent trends in linear text segmentation : A survey. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, pp. 3084–3095.

Leopold, E. et J. Kindermann (2006). Content classification of multimedia documents using partitions of low-level features.

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.

Pelloin, V., F. Dary, N. Hervé, B. Favre, N. Camelin, A. Laurent, et L. Besacier (2022). Asr-generated text for language model pre-training applied to speech tasks. In *Proc. Interspeech 2022*, pp. 3453–3457, doi: 10.21437/Interspeech.2022-352.

Pelloin, V., L. Dodson, É. Chapuis, N. Hervé, et D. Doukhan (2024). Automatic classification of news subjects in broadcast news : Application to a gender bias representation analysis. In *Interspeech 2024*, pp. 3055–3059. ISCA.

Reimers, N. et I. Gurevych (2019). Sentence-bert : Sentence embeddings using siamese bert-networks.

Shukla, S. D., P. Denisov, et M. A. T. Turan (2024). Advancing topic segmentation of broadcasted speech with multilingual semantic embeddings. In *European Conference on Artificial Intelligence 2024*.

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.

Summary

In this paper, we present a method for extracting segments from television and radio news programs, developed from a case study consisting of finding all media sequences dealing with the killing of French 17-year-old Nahel Merzouk in 2023. The task is modeled as a semantic similarity calculation between a text query and a sliding window applied to the transcription of the programs. Our experiments are based on an ASR corpus from radio and television channels, part of which is annotated for content segmentation. The results show that daily news articles constitute more effective queries than summaries for identifying relevant segments.