Appariement de trames narratives : dépasser le chevauchement lexical en recherche d'information

Ahmed Hamdi[1] Emanuela Boros[2] José G. Moreno[1] Antoine Doucet[2, 3]
(1) IRIT, Université de Toulouse, France
(2) L3i, La Rochelle université, France
(3) Université de Ljubljana, Slovénie
ahmed.hamdi@irit.fr, emanuela.boros@univ-lr.fr, jose.moreno@irit.fr,
antoine.doucet@univ-lr.fr

RÉSUMÉ

Les modèles de similarité sémantique actuels, bien que performants sur les benchmarks standards, peinent à reconnaître l'équivalence narrative entre textes relatant une même histoire. Pour pallier cette limitation en recherche d'information orientée narration, nous présentons une nouvelle ressource multilingue constituée de résumés de films appariés, extraite automatiquement de Wikipedia. Ce jeu de données permet un apprentissage supervisé à grande échelle de la similarité narrative au-delà du simple chevauchement lexical. Nous l'exploitons pour fine-tuner des modèles d'embeddings via un apprentissage contrastif et évaluons leur performance dans un système de recherche d'information à deux étages (premier appel BM25 suivi d'un re-ranking). Nous comparons les modèles en configuration zero-shot avec leurs versions fine-tunées sur notre ressource, démontrant l'apport de l'appariement narratif cross-lingue pour cette tâche.

ABSTRACT

Narrative Frame Matching for Information Retrieval Beyond Lexical Overlap

Current semantic similarity models, while performing well on standard benchmarks, often struggle to recognize narrative equivalence between texts that tell the same story, limiting their effectiveness in narrative-oriented information retrieval. To address this limitation, we introduce a new multilingual resource consisting of paired movie plot summaries, automatically extracted from Wikipedia. This dataset enables large-scale supervised learning of narrative similarity beyond simple lexical overlap. We leverage this resource to fine-tune embedding models using contrastive learning objectives and evaluate their performance within a two-stage information retrieval system (initial BM25 retrieval followed by re-ranking). We compare zero-shot models against their fine-tuned counterparts on our dataset, demonstrating the value of cross-lingual narrative alignment for this task.

MOTS-CLÉS : recherche d'information, narrative, appariement, résumé de films, ressource.

KEYWORDS: information retrieval, narrative, alignment, film plots, resource.

1 Introduction

Les modèles de plongements lexicaux de phrases et de documents, tels que SBERT et ses variantes, sont devenus la référence pour l'estimation de la similarité sémantique et la recherche d'information (Devlin *et al.*, 2019; Liu *et al.*, 2019; Reimers & Gurevych, 2019a; Thakur *et al.*, 2021). Leur succès

repose sur leur capacité à capturer des relations sémantiques au-delà du simple chevauchement lexical, permettant ainsi d'identifier des textes traitant de sujets similaires même lorsqu'ils emploient une terminologie différente. Cependant, ces modèles présentent une limitation fondamentale : ils restent sensibles aux variations de surface comme les noms de personnages, les lieux ou les époques et échouent souvent à reconnaître l'équivalence narrative entre deux textes qui racontent pourtant la même histoire (Elson, 2012; McCoy *et al.*, 2019; Hill *et al.*, 2016). Cette faiblesse découle de leur conception : ils sont optimisés pour la similarité thématique ou stylistique, et non pour la correspondance des structures événementielles qui constituent la trame d'un récit.

L'impact de cette lacune est particulièrement problématique dans les domaines où la dimension narrative est centrale : archives culturelles, bases de données médiatiques, bibliothèques éducatives ou encore analyse comparative des récits. Dans ce contexte, les utilisateurs cherchent des documents liés par leur trame narrative comme l'identification des variantes d'un conte à travers différentes cultures ou l'adaptation d'une histoire d'un média à un autre. Or, les modèles standards, qu'ils soient lexicaux (comme BM25) ou sémantiques (embeddings contextuels), privilégient la proximité thématique et négligent les documents qui partagent des structures narratives similaires mais divergent dans leur formulation, leurs personnages, ou leur langue d'expression.

Pour répondre à ce défi, nous proposons dans cet article une approche qui reformule la similarité narrative comme une tâche de recherche d'information cross-lingue. Notre hypothèse centrale est que l'alignement de résumés de films racontant la même histoire à travers différentes langues, fournit un signal d'apprentissage naturel et puissant pour capturer l'équivalence narrative. Ces paires partagent la même séquence d'événements fondamentaux tout en variant dans leur expression linguistique et leurs détails de surface, forçant ainsi le modèle à abstraire la structure narrative sous-jacente.

À partir de cette intuition, nous construisons une ressource multilingue à grande échelle dédiée à l'apprentissage de la similarité narrative. Nous extrayons des résumés de films alignés dans plusieurs langues (anglais, français, espagnol, portugais) et les traduisons automatiquement vers l'anglais pour créer un corpus de paires narratives comparables. Cette approche nous permet de générer une supervision à grande échelle sans annotation manuelle, chaque paire de résumés décrivant le même film constituant un exemple positif d'équivalence narrative [1]

La suite de cet article est organisée comme suit : la section 2 examine les travaux connexes en narratologie computationnelle et similarité sémantique. La section 3 décrit la construction de notre ressource à partir de résumés de films multilingues. La section 4 présente notre cadre expérimental, les modèles comparés et les résultats obtenus pour la recherche d'information narrative. Enfin, la section 5 conclut et discute les perspectives de recherche.

2 Travaux connexes

Notre travail s'inscrit à l'intersection de la narratologie computationnelle et de l'apprentissage de représentations sémantiques. Des travaux fondateurs ont formalisé les récits comme des séquences d'événements (McIntyre & Lapata, 2010; Goyal *et al.*, 2010) ou des scripts (Chambers & Jurafsky, 2008, 2009; Pichotta & Mooney, 2016; Keith Norambuena *et al.*, 2023). Plus récemment, des approches neuronales ont exploré la cohérence narrative ou la génération d'histoires, mais les jeux de

1. L'ensemble du code source ainsi que les ressources utilisées dans ce travail sont disponibles pour la communauté : https://github.com/ahHamdi/narrative-IR-for-coria-2026.

données associés (ex. ROCStories, StoryCloze) sont conçus pour évaluer la compréhension de récits courts, non pour mesurer la similarité entre textes narratifs complets.

Parallèlement, le domaine de la similarité sémantique a produit de nombreux modèles et benchmarks pour les phrases (Agirre *et al.*, 2012, 2016; Cer *et al.*, 2017). Leur extension aux textes longs se concentre souvent sur la similarité thématique ou documentaire (Conneau *et al.*, 2017), inadéquate pour capturer la correspondance narrative.

En recherche d'information, BM25 reste une baseline lexicale robuste, mais elle est fondamentalement aveugle à la structure narrative sous-jacente. Enfin, les ressources existantes pour l'étude de la similarité narrative sont rares et de taille limitée (Chaturvedi *et al.*, 2018; Piskorski *et al.*, 2025; Johnson *et al.*, 2025), ne permettant pas un apprentissage contrastif à grande échelle. La ressource que nous introduisons comble ce manque en fournissant des milliers de paires narratives alignées, ouvrant ainsi la voie à des modèles spécifiquement adaptés à la recherche d'information narrative.

3 Modélisation de la similarité narrative pour la recherche d'information

Cette section présente l'ensemble de notre démarche méthodologique. Nous décrivons d'abord la construction d'une ressource pour la similarité narrative à partir de résumés de films multilingues. Nous présentons ensuite la recherche d'information narrative comme tâche d'évaluation, en présentant le benchmark que nous concevons pour mesurer la capacité des modèles à retrouver des récits équivalents. Enfin, nous exposons notre approche d'apprentissage contrastif et l'intégration des modèles dans un système de recherche à deux étages, combinant un premier appel lexical BM25 pour le filtrage initial et un re-ranking basé sur des embeddings.

3.1 Mise en place de la ressource pour la similarité narrative

Pour entraîner des modèles capables de reconnaître l'équivalence narrative, nous construisons une ressource à grande échelle à partir de résumés de films extraits de Wikipédia en quatre langues : anglais (EN), français (FR), espagnol (ES) et portugais (PT). La liste des films candidats est dérivée des pages d'index regroupées sous Wikipédia[2]. Pour chaque film sélectionné, nous récupérons le contenu de l'article Wikipédia correspondant via le jeu de données Wikimedia Wikipedia de Hugging Face[3], qui fournit des dumps structurés des pages adaptés à l'interrogation et au traitement à grande échelle.

À partir des pages, nous collectons les descriptions des résumés dans les quatre langues. Les sections narratives incluent respectivement le contenu étiqueté `plot`, `synopsis`, `sinopsis`, `synops` lorsqu'elles sont disponibles. Les résumés sont appariés entre les langues en utilisant l'alignement des titres et les liens interlangues. Pour augmenter la comparabilité et permettre un appariement narratif contrôlé, les résumés non anglais sont automatiquement traduits en anglais via l'API Google Translate[4]. Ce choix de projection dans un espace linguistique unique facilite l'apprentissage contrastif

2. https://en.wikipedia.org/wiki/Lists_of_films
3. https://huggingface.co/datasets/wikimedia/wikipedia
4. https://cloud.google.com/translate

en créant des paires de textes narrativement équivalents : chaque résumé anglais original est associé à ses versions traduites depuis les autres langues.

Partition	#Résumés	Tokens/Résumé (EN)	Tokens/Résumé (Traduction)
EN-PT	7 107	565,83	149,31
EN-ES	3 644	554,44	211,54
EN-FR	17 630	515,92	171,21
EN-ENS	28 381	533,36	170,90

TABLE 1 – Statistiques de la ressource par partition linguistique. ENS est l'ensemble des synopsis de toutes les langues.

Les résumés traduits apparaissent significativement plus courts que les résumés anglais originaux car les versions disponibles dans les autres langues sont souvent plus condensées et moins détaillées dans les sources Wikipédia correspondantes. La traduction automatique préserve globalement cette densité informationnelle réduite plutôt que d'enrichir le contenu, ce qui explique l'écart observé en nombre moyen de tokens.

Ce processus produit des paires qui partagent la même histoire sous-jacente tout en présentant des variations lexicales et syntaxiques contrôlées, issues des différences linguistiques et des artefacts de traduction. Le jeu de données résultant consiste donc en variantes narratives multilingues projetées dans un espace de représentation anglais partagé. Le tableau 1 fournit une description statistique du jeu de données, rapportant le nombre de résumés anglais et leurs variantes traduites correspondantes pour chaque paire de langues.

3.2 Recherche d'information narrative

Pour évaluer l'utilité pratique de nos plongements dans un contexte de recherche réaliste, nous construisons un benchmark dédié à partir du corpus de résumés de films collectés. L'objectif de cette évaluation est de mesurer si des représentations apprises sur la ressource améliorent la recherche de documents pertinents au niveau des narrations, au-delà de la similarité lexicale de surface.

Notre ensemble de test est composé d'environ 1 200 résumés de films EN qui sont appariés avec leurs trois traductions en français, portugais et espagnol. Pour chaque résumé EN, nous générons une requête en supprimant la dernière phrase du résumé et en utilisant cette phrase retenue comme requête de recherche. Cette requête simule un scénario réaliste dans lequel un utilisateur fait référence à un tournant narratif saillant proche de la fin et cherche à retrouver l'histoire complète ou des détails antérieurs à partir d'une collection. Chaque requête est donc associée à quatre documents pertinents : le résumé anglais original (dont la dernière phrase a été supprimée) et ses trois résumés traduits. Ces documents décrivent le même contenu narratif à travers les langues et servent de cibles positives dans la tâche de recherche. Tous les autres résumés de la collection sont traités comme des documents non pertinents pour cette requête.

L'ensemble de résumés non inclus dans ce sous-ensemble de benchmark sont réservés exclusivement pour l'entraînement et l'apprentissage des représentations, garantissant une séparation stricte entre les données d'entraînement et les données de test. Le benchmark résultant supporte l'évaluation de la recherche multilingue et cross-lingue, où une requête dérivée d'une fin narrative anglaise doit retrouver des documents narrativement équivalents dans plusieurs langues.

Ce protocole évalue si les plongements appris sont de qualité suffisante pour capturer la structure narrative et la sémantique des événements et pour supporter une recherche d'information adaptée à la narration, plutôt que le chevauchement d'entités ou à la similarité lexicale locale.

3.3 Modèles et apprentissage contrastif

À partir de cette ressource, nous formulons l'apprentissage de la similarité narrative comme un problème d'optimisation contrastive. L'objectif est d'apprendre une fonction de plongement f qui projette un texte x dans un espace vectoriel où la distance reflète l'équivalence narrative : des résumés racontant la même histoire doivent être rapprochés, tandis que des résumés d'histoires différentes doivent être éloignés.

Pour chaque résumé anglais x_i et sa version traduite x_i^+ (depuis une autre langue), nous constituons des paires positives. Les paires négatives sont échantillonnées aléatoirement parmi les autres résumés du corpus. Nous utilisons une fonction de perte contrastive :

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(f(x_i), f(x_i^+))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f(x_i), f(x_j^-))/\tau)} \quad (1)$$

où sim est la similarité cosinus, τ un paramètre de température, et N le nombre d'exemples négatifs par positif. Nous utilisons la similarité cosinus qui est mieux adaptée à notre cadre multilingue, où les différences de longueur et de densité informationnelle entre résumés originaux et traductions peuvent induire des variations importantes de norme vectorielle. La normalisation implicite opérée par la similarité cosinus permet de réduire l'influence de ces variations de longueur et de focaliser l'apprentissage contrastif sur la proximité narrative et sémantique des contenus.

Nous expérimentons avec l'architectures `all-MiniLM-L6-v2`, un modèle léger et efficace qui offre un bon équilibre entre performance et rapidité d'exécution. Ce choix est motivé par la nécessité de disposer d'un modèle rapide pour le re-ranking dans notre pipeline de recherche d'information à deux étages, tout en maintenant une qualité de représentation suffisante pour capturer la similarité narrative. Pour cette architecture, nous comparons deux versions : le modèle pré-entraîné disponible publiquement, et sa version finetunée sur notre ressource.

Le finetuning est effectué en utilisant la librairie Sentence-Transformers avec un objectif d'apprentissage contrastif exploitant les négatifs intra-batch. La tâche d'entraînement est formulée comme une tâche de paires dans Sentence-Transformers, où l'objectif est de maximiser la similarité entre le résumé anglais original et sa traduction tout en minimisant la similarité avec les autres résumés présents dans le batch[5].

3.4 Système de recherche à deux étages

Pour évaluer ces modèles dans des conditions réalistes, nous les intégrons dans un système de recherche d'information à deux étages. La première étape utilise le modèle lexical BM25 pour

5. Nous utilisons une taille de batch de 8 paires et entraînons pendant 3 époques. La longueur maximale des séquences est fixée à 128 tokens afin de se concentrer sur le contenu narratif principal. L'optimisation est réalisée avec l'optimiseur AdamW et un taux d'apprentissage de 3×10^{-5}.

sélectionner un ensemble restreint de candidats pertinents parmi l'ensemble de la collection. Cette étape de pré-filtrage réduit l'espace de recherche et permet d'appliquer dans un second temps des modèles plus coûteux computationnellement.

La seconde étape consiste en un re-ranking des candidats sélectionnés par BM25 à l'aide de nos modèles d'embeddings. Pour chaque paire (requête, document candidat), nous calculons la similarité entre leurs représentations vectorielles et réordonnons les documents selon ce score.

Nous comparons systématiquement deux configurations pour chaque modèle d'embedding :

— **Zero-shot** : le modèle est utilisé tel quel, sans finetuning préalable sur notre ressource, afin d'évaluer sa capacité intrinsèque à capturer la similarité narrative ;
— **Finetuné** : le modèle est d'abord entraîné sur notre ressource via l'apprentissage contrastif décrit ci-dessus, puis évalué sur la tâche de recherche.

Cette comparaison nous permet d'isoler l'apport spécifique de l'apprentissage supervisé sur paires narratives et de mesurer dans quelle mesure notre ressource améliore les performances des modèles pour la recherche d'information narrative.

4 Expériences et résultats

Cette section présente le cadre expérimental mis en place pour évaluer l'apport de notre ressource pour la recherche d'information narrative. Nous décrivons d'abord les modèles comparés et les métriques d'évaluation, puis nous présentons et analysons les résultats obtenus.

4.1 Protocole expérimental

Modèles évalués. Nous évaluons trois configurations dans le cadre de notre système de recherche à deux étages :

— **BM25 seul** : modèle lexical servant de baseline, utilisé pour la première étape de recherche et comme référence de performance.
— **BM25 + all-MiniLM-L6-v2 (zero-shot)** : le modèle de base `all-MiniLM-L6-v2` de la librairie Sentence-Transformers (Reimers & Gurevych, 2019b), utilisé sans finetuning préalable pour le re-ranking des candidats BM25.
— **BM25 + MiniLM-L6 finetuné** : le même modèle architecture, mais finetuné sur notre ressource de paires narratives via l'apprentissage contrastif décrit en section 3.

Pour toutes les configurations, la première étape de recherche utilise BM25 pour sélectionner les 100 documents les plus pertinents. Cela est maintenu pour toutes les expériences, garantissant que les différences observées dans les métriques de re-ranking sont attribuables uniquement aux modèles d'embeddings.

Métriques d'évaluation. Nous évaluons la performance des modèles à l'aide des métriques standards en recherche d'information :

— **Précision@k (P@k)** : proportion de documents pertinents parmi les k premiers documents retournés.
— **Mean Reciprocal Rank (MRR)** : moyenne de l'inverse du rang du premier document pertinent.

— **nDCG@k** : *Normalized Discounted Cumulative Gain*, qui pénalise les documents pertinents apparaissant plus bas dans le classement.

Nous rapportons ces métriques pour k=10, ainsi que le MRR, qui sont particulièrement adaptés pour évaluer la qualité du classement.

4.2 Résultats et discussion

Le tableau 2 présente les résultats comparatifs sur l'ensemble des 1160 requêtes de test. La première étape BM25 atteint un MRR de 0,7521 et un nDCG@10 de 0,7466, confirmant que BM25 reste une baseline robuste grâce au chevauchement lexical résiduel entre requêtes et documents.

Modèle	**Entraînement**	**P@10**	**MRR**	**nDCG@10**
BM25	–	0,1650	**0,7521**	**0,7466**
Re-ranking avec embeddings				
MiniLM-V6	zero-shot	0,1690	0,6254	0,6393
	PT	0,1695	0,6174	0,6337
	FR	0,1724	0,6281	0,6422
	ES	0,1729	0,6403	0,6556
	All	**0,1735**	0,6209	0,6446

TABLE 2 – Performance globale des modèles (1 160 requêtes)

Les résultats montrent que les modèles fine-tunés améliorent la précision à 10 par rapport à BM25, passant de 0,1650 à 0,1735 pour la version entraînée sur toutes les langues, soit un gain de +5,2%. Cependant, nous observons une baisse apparente du MRR et du nDCG pour tous les modèles d'embeddings par rapport à BM25. Ce phénomène s'explique par la nature même de notre protocole expérimental : la requête est construite à partir de la dernière phrase du résumé anglais, qui partage inévitablement un certain vocabulaire avec le début du même résumé. BM25 exploite ce chevauchement lexical pour positionner très haut le document anglais pertinent, ce qui lui confère un avantage sur le MRR et le nDCG.

Pour isoler la capacité des modèles à capturer la similarité narrative au-delà du chevauchement lexical, nous avons analysé la distribution des overlaps entre requêtes et documents. La figure 1 présente la distribution cumulative des chevauchements lexicaux moyens, définis par la proportion moyenne de tokens partagés entre une requête et un document.

Cette distribution révèle que seulement 31 requêtes (soit 2,7% du total) présentent un chevauchement moyen inférieur à 0,2, tandis que la grande majorité des requêtes bénéficient d'un chevauchement lexical non négligeable avec les documents pertinents. Ces requêtes à très faible overlap constituent un véritable test de la capacité des modèles à généraliser au-delà de la similarité lexicale de surface. Le tableau 3 présente les résultats sur les 31 requêtes les plus difficiles, où le chevauchement lexical moyen est inférieur à 0,2.

Sur les requêtes difficiles où le chevauchement lexical est minimal, la précision à 10 de BM25 chute à 0,0903 (contre 0,1650 sur l'ensemble), confirmant que son apparente bonne performance globale était largement due à l'exploitation du chevauchement lexical. Privé de ce signal, BM25 peine à identifier les documents pertinents. Par ailleurs, tous les modèles d'embeddings surclassent très nettement BM25. Le modèle zero-shot double déjà la précision (0,1806 vs 0,0903) et améliore le MRR de

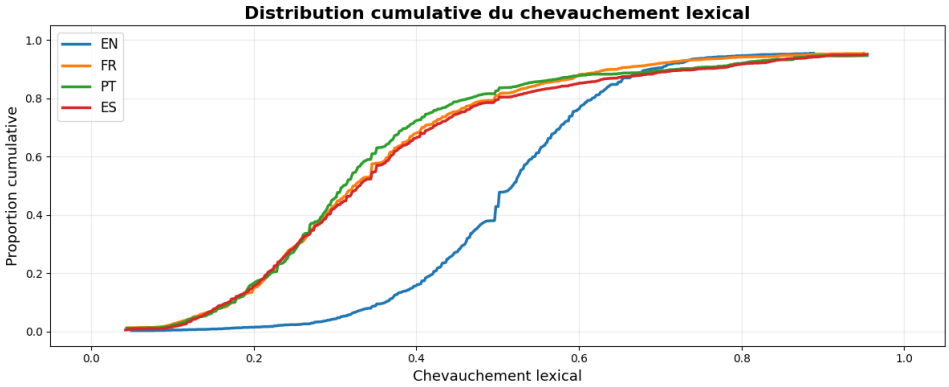

FIGURE 1 – Distribution cumulative du chevauchement lexical moyen requêtes-documents

Modèle	Entraînement	P@10	MRR	nDCG@10
BM25	–	0,0903	0,5804	0,5775
Re-ranking avec embeddings				
MiniLM-V6	zero-shot	0,1806	0,7270	0,7084
	PT	0,2065	0,6398	0,6806
	FR	**0,2097**	**0,7765**	**0,7234**
	Es	0,1968	0,6858	0,7049
	All	0,1935	0,7243	0,7094

TABLE 3 – Performance sur les requêtes à faible chevauchement lexical (31 requêtes, chevauchement < 0,2)

0,5804 à 0,7270 (+25%). Par rapport au modèle zero-shot, les versions fine-tunées améliorent encore les performances. Le meilleur modèle (fine-tuné sur le français) atteint une précision à 10 de 0,2097 (soit +132% par rapport à BM25 et +16% par rapport au zero-shot) et un MRR de 0,7765 (+34% par rapport à BM25, +7% par rapport au zero-shot).

Ces résultats démontrent que sur les requêtes où l'information lexicale est insuffisante, les embeddings (même zero-shot) réussissent à identifier les documents pertinents, cela prouve qu'ils encodent une forme de similarité plus profonde. Les modèles fine-tunés surpassent systématiquement la version zero-shot, validant la pertinence de notre ressource pour apprendre la similarité narrative. Plus la tâche est difficile (faible chevauchement), plus l'apport de notre approche est marqué, cela est traduit sur les requêtes les plus exigeantes par un gain en précision qui atteint +132% par rapport à BM25.

La baisse apparente du MRR sur l'ensemble des requêtes pour les modèles fine-tunés s'explique par un comportement de classement différent. Plutôt que de privilégier systématiquement le document anglais qui partage quelques termes avec la requête, ces modèles cherchent à identifier une véritable correspondance narrative. Ainsi, lorsque des documents dans d'autres langues présentent une correspondance narrative plus pertinente, le document anglais peut être relégué à un rang inférieur. Ce phénomène peut entraîner une diminution du MRR, tout en traduisant un comportement plus conforme à l'objectif d'une recherche d'information centrée sur la similarité narrative.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une nouvelle ressource pour l'apprentissage de la similarité narrative, construite à partir de résumés de films multilingues alignés. Nous avons formulé la tâche comme un problème de recherche d'information en utilisant un apprentissage contrastif. Nos expériences sur les requêtes considérées difficiles montrent que BM25 s'effondre (P@10 = 0,0903), tandis que nos modèles fine-tunés atteignent une précision à 10 de 0,2097, soit un gain de +132% par rapport à BM25 et de +16% par rapport à la version zero-shot. Ces résultats confirment que notre ressource permet de capturer l'équivalence narrative au-delà du simple chevauchement lexical.

Nos travaux futurs exploreront l'extension de cette approche à d'autres domaines comme les archives historiques et la littérature ainsi que l'utilisation de modèles plus puissants que MiniLM et l'intégration de représentations explicites de la structure événementielle.

Remerciements

Ce travail a été cofinancé par l'Agence Nationale de la Recherche (ANR) à travers du projet ANR-25-CE38-6695 (MILL-EHNAS) ainsi que par l'Union européenne via la subvention HORIZON-WIDERA-2023-TALENTS-01-01 n°101186647 - AI4DH.

Références

AGIRRE E., BANEA C., CER D., DIAB M., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, p. 497–511.

AGIRRE E., CER D., DIAB M. & GONZALEZ-AGIRRE A. (2012). Semeval-2012 task 6 : A pilot on semantic textual similarity. in* sem 2012 : The first joint conference on lexical and computational semantics–volume 1 : Proceedings of the main conference and the shared task, and volume 2 : Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montréal, QC, Canada*, p. 7–8.

CER D., DIAB M., AGIRRE E., LOPEZ-GAZPIO I. & SPECIA L. (2017). Semeval-2017 task 1 : Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv :1708.00055*.

CHAMBERS N. & JURAFSKY D. (2008). Unsupervised learning of narrative event chains. In J. D. MOORE, S. TEUFEL, J. ALLAN & S. FURUI, Éds., *Proceedings of ACL-08 : HLT*, p. 789–797, Columbus, Ohio : Association for Computational Linguistics.

CHAMBERS N. & JURAFSKY D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 602–610.

CHATURVEDI S., SRIVASTAVA S. & ROTH D. (2018). Where have i heard this story before ? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers), p. 673–678.

CONNEAU A., KIELA D., SCHWENK H., BARRAULT L. & BORDES A. (2017). Supervised learning of universal sentence representations from natural language inference data. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 670–680, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1070.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies, volume 1 (long and short papers)*, p. 4171–4186.

ELSON D. K. (2012). *Modeling narrative discourse*. Columbia University.

GOYAL A., RILOFF E. & DAUMÉ III H. (2010). Automatically producing plot unit representations for narrative text. In H. LI & L. MÀRQUEZ, Éds., *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 77–86, Cambridge, MA : Association for Computational Linguistics.

HILL F., CHO K. & KORHONEN A. (2016). Learning distributed representations of sentences from unlabelled data. In K. KNIGHT, A. NENKOVA & O. RAMBOW, Éds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1367–1377, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1162.

JOHNSON N., BERTSCH A., DEAL M.-E. & STRUBELL E. (2025). Ficsim : A dataset for multi-faceted semantic similarity in long-form fiction. In *Findings of the Association for Computational Linguistics : EMNLP 2025*, p. 25228–25246.

KEITH NORAMBUENA B. F., MITRA T. & NORTH C. (2023). A survey on event-based news narrative extraction. *ACM Computing Surveys*, **55**(14s), 1–39.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MCCOY R. T., PAVLICK E. & LINZEN T. (2019). Right for the wrong reasons : Diagnosing syntactic heuristics in natural language inference. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3428–3448, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1334.

MCINTYRE N. & LAPATA M. (2010). Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1562–1572 : Association for Computational Linguistics.

PICHOTTA K. & MOONEY R. (2016). Learning statistical scripts with lstm recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

PISKORSKI J., MAHMOUD T., NIKOLAIDIS N., CAMPOS R., JORGE A. M., DIMITROV D., SILVANO P., YANGARBER R., SHARMA S., CHAKRABORTY T. *et al.* (2025). Semeval 2025 task 10 : Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, p. 2610–2643.

REIMERS N. & GUREVYCH I. (2019a). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

REIMERS N. & GUREVYCH I. (2019b). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

THAKUR N., REIMERS N., DAXENBERGER J. & GUREVYCH I. (2021). Augmented sbert : Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics : Human language technologies*, p. 296–310.