Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Le rôle du contexte dans la classification séquentielle de phrases pour les documents longs

Anas Belfathi, Nicolas Hernandez, Laura Monceaux, Richard Dufour
LS2N, UMR CNRS 6004, Nantes Université
{firstname.lastname}@univ-nantes.fr

Résumé

La classification séquentielle de phrases étend la classification traditionnelle en intégrant un contexte plus large. Cependant, les approches de pointe rencontrent deux défis majeurs dans le traitement automatique des documents longs : les modèles de langue préentraînés sont limités par des contraintes de longueur d'entrée, tandis que les modèles hiérarchiques proposés introduisent souvent du contenu non pertinent. Pour surmonter ces limitations, nous proposons une approche de recherche d'information au niveau du document visant à extraire uniquement le contexte le plus pertinent. Plus précisément, nous introduisons deux types d'heuristiques : **Séquentiel**, qui capture l'information locale, et **Sélectif**, qui sélectionne les phrases les plus sémantiquement similaires. Nos expériences sur trois corpus juridiques en anglais montrent que ces heuristiques améliorent les performances. Les heuristiques séquentielles surpassent les modèles hiérarchiques sur deux des trois jeux de données, démontrant l'apport du contexte ciblé.

Abstract

The Role of Context in Sequential Sentence Classification for Long Documents

Sequential sentence classification extends traditional classification by incorporating broader context. However, state-of-the-art approaches face two major challenges in long documents : pretrained language models struggle with input-length constraints, while proposed hierarchical models often introduce irrelevant content. To address these limitations, we propose a document-level retrieval approach that extracts only the most relevant context. Specifically, we introduce two heuristic strategies : **Sequential**, which captures local information, and **Selective**, which retrieves the most semantically similar sentences. Experiments on legal domain datasets written in English show that both heuristics improve performance. Sequential heuristics outperform hierarchical models on two out of three datasets, demonstrating the benefits of targeted context.

Mots-clés : Extraction d'information, classification séquentielle de phrases, documents longs, modèles de langue préentraînés.

Keywords: Information Extraction, Sequential Sentence Classification, Long Documents, Pretrained Language Models.

Article : **Soumis à** la conférence ACL.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 488–501.

1 Introduction

La classification séquentielle de phrases (*Sequential Sentence Classification*, SSC) consiste à catégoriser des phrases en fonction de leur rôle discursif au sein d'un document. Le sens d'une phrase étant souvent influencé par le contexte qui l'entoure, la SSC est particulièrement utile pour l'analyse de textes structurés, tels que les décisions juridiques. L'identification des principaux rôles rhétoriques (par exemple, le préambule, l'énoncé du problème ou l'analyse ; voir Figure 1) facilite plusieurs tâches en aval, notamment la recherche d'information (Neves *et al.*, 2019; Safder & Hassan, 2019) et le résumé automatique de documents (Kalamkar *et al.*, 2022; Muhammed *et al.*, 2024).

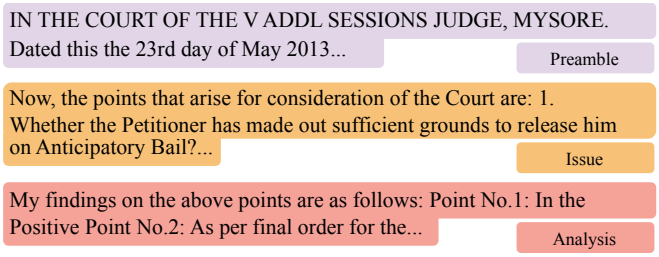


FIGURE 1 – Extrait d'un document juridique avec des phrases étiquetées en fonction de leur rôle.

Les modèles hiérarchiques dans l'état de l'art ont démontré des performances remarquables en SSC en traitant l'ensemble des séquences d'un document simultanément, ce qui permet de capturer un contexte plus large (Jin & Szolovits, 2018; Brack *et al.*, 2021; Kalamkar *et al.*, 2022). Cependant, nous supposons que prendre en compte toutes les phrases n'est pas toujours nécessaire, car cela peut introduire du bruit dû à des contenus non pertinents (Shi *et al.*, 2023). Par ailleurs, les modèles de langue préentraînés (*Pretrained Language Models*, PLMs) restent limités par leur longueur d'entrée (Warner *et al.*, 2024), malgré les avancées récentes avec les grands modèles de langue (Large Language Models, LLMs BehnamGhader *et al.*, 2024).

Des études récentes ont exploré différentes stratégies pour sélectionner des informations pertinentes au niveau du document (Amalvy *et al.*, 2023b; Lan *et al.*, 2024). Cependant, à notre connaissance, aucun travail existant n'a explicitement étudié comment extraire le contexte le plus pertinent afin d'optimiser les performances des PLMs en SSC.

Dans cet article, nous apportons deux contributions principales : (1) Une analyse du rôle du contexte en SSC à travers l'introduction de deux types heuristiques d'extraction d'information—*Séquentiel*, qui exploite l'information locale autour de chaque phrase, et *Sélectif*, qui identifie les phrases les plus proches sémantiquement à l'échelle du document. (2) démontrer comment ces stratégies améliorent les PLMs en leur fournissant un contexte plus pertinent.

Pour l'évaluation, nous utilisons des corpus de documents juridiques, qui constituent la principale référence pour la tâche SSC. Afin de garantir la transparence et la reproductibilité, nous mettons notre code à disposition sous une licence open source [1].

1. https://anonymous.4open.science/r/ACL-2025-4BE2

2 Travaux connexes

2.1 Contraintes de longueur des séquences d'entrée dans les PLMs

Les modèles encodeurs comme BERT (Devlin *et al.*, 2019) offrent un bon compromis entre taille et performance, ce qui en fait une alternative intéressante aux architectures basées sur les décodeurs pour les tâches de classification. Cependant, la complexité quadratique du mécanisme d'auto-attention dans les Transformeurs classiques limite leur capacité à traiter de longues séquences en entrée, posant ainsi des défis pour le traitement des documents longs. Pour atténuer ces limitations, des mécanismes d'attention parcimonieux (*Sparse attention*) ont été introduits afin de réduire les coûts computationnels (Zaheer *et al.*, 2020; Wang *et al.*, 2020; Beltagy *et al.*, 2020; Choromanski *et al.*, 2021). Bien que ces méthodes étendent la portée du contexte, elles ne parviennent toujours pas à résoudre complètement les contraintes liées au traitement des textes longs (Warner *et al.*, 2024; Nussbaum *et al.*, 2025).

2.2 SSC pour les documents longs

Les premiers travaux sur la SSC se sont concentrés sur des modèles hiérarchiques afin d'intégrer un contexte plus large dans les représentations de phrases. Le réseau de labellisation séquentielle hiérarchique (*Hierarchical Sequential Labeling Network*, HSLN) a été l'un des premiers modèles à traiter des séquences de documents entiers pour produire des représentations contextualisées (Jin & Szolovits, 2018; Shang *et al.*, 2021; Brack *et al.*, 2021; Kalamkar *et al.*, 2022). Des études plus récentes ont exploré des stratégies d'apprentissage avancées : T.y.s.s. *et al.* (2024) ont appliqué l'apprentissage contrastif et prototypique pour améliorer les représentations des phrases en exploitant les similitudes sémantiques, tandis que T.y.s.s *et al.* (2024) ont proposé un cadre d'apprentissage curriculaire hiérarchique permettant d'améliorer progressivement la capacité du modèle à distinguer les étiquettes rhétoriques à différents niveaux de granularité.

Bien que ces travaux aient principalement porté sur l'amélioration de HSLN, notre étude adresse un défi différent : surmonter les contraintes de longueur d'entrée dans les PLMs en récupérant uniquement le contexte le plus pertinent, réduisant ainsi le bruit et améliorant l'efficacité de la SSC.

3 Extraction du Contexte

Pour analyser l'impact du contexte sur les performances des PLMs, nous introduisons deux types d'heuristiques : **Séquentiel**, qui exploite le contexte local, et **Sélectif**, qui identifie les phrases les plus pertinentes. Celles-ci s'inspirent de recherches antérieures sur l'enrichissement contextuel aves les LLMs (Amalvy *et al.*, 2023a; Wang *et al.*, 2024; Nussbaum *et al.*, 2025).

Heuristiques Séquentielles Elles permettent d'extraire le contexte des phrases voisines à la phrase cible au sein du même document. Nous définissons trois stratégies :

— **Précédente** : sélectionne les k phrases précédant la phrase cible.

— **Suivante** : sélectionne les k phrases suivant la phrase cible.
— **Environnante** : sélectionne $\frac{k}{2}$ phrases avant et après la phrase cible.

Heuristiques Sélectives Contrairement aux stratégies séquentielles, les heuristiques sélectives récupèrent des phrases dans n'importe quelle partie du document, indépendamment de leur position par rapport à la phrase cible. Nous explorons trois techniques d'extraction :

— **Aléatoire** : sélectionne k phrases de manière aléatoire dans l'ensemble du document.
— **BM25** : extrait les k phrases les plus pertinentes à l'aide de BM25 (Trotman *et al.*, 2014), une fonction de classement qui attribue un score aux phrases en fonction d'un schéma de pondération basé sur la fréquence des termes et l'inverse de la répartition du terme dans les documents (TF-IDF). BM25 est largement utilisé en recherche d'information pour le classement lexical par pertinence.
— **Sentence-BERT** : sélectionne les k phrases les plus proches sémantiquement de la phrase cible en utilisant les représentations de Sentence-BERT (Reimers & Gurevych, 2019), qui capturent la similarité entre phrases grâce à un réseau BERT siamois pré-entraîné *(pre-trained Siamese BERT)*.

Compte tenu des contraintes computationnelles, nous limitons notre analyse à $k = 6$. Notons que les heuristiques sélectives peuvent récupérer des phrases déjà incluses dans le contexte séquentiel, car elles opèrent sur l'ensemble du document. Des exemples illustratifs sont fournis dans le Tableau 3 en annexe.

Ordonnancement des phrases Nous examinons l'influence de l'ordre des phrases récupérées sur la performance de la SSC. Cette analyse s'inspire de NAREOR (Gangal *et al.*, 2022), qui explore la réorganisation des phrases afin d'évaluer la cohérence narrative. Dans ce cadre, nous évaluons l'impact de la modification de l'ordre des phrases d'un document lorsque ($k = N$) sur les performances du modèle, en appliquant nos heuristiques.

Dans l'approche Séquentielle, l'ordre d'apparition des phrases dans le document est conservé afin de préserver la structure logique et discursive du texte. Dans l'approche Sélective, les phrases sont réorganisées en fonction de leur pertinence par rapport à la phrase cible, tout en garantissant l'inclusion de l'ensemble des phrases extraites afin de maintenir une comparaison équitable entre les deux types d'heuristiques.

4 Protocole Expérimental

4.1 Jeux de données d'évaluation

Notre étude se concentre sur le domaine juridique, le seul domaine disposant de jeux de données annotés à l'échelle du document. Nous utilisons trois ensembles de données [2] :

1. **DeepRhole** (Bhattacharya *et al.*, 2023) : contient 50 jugements de la Cour suprême de l'Inde, couvrant cinq domaines juridiques. Il regroupe un total de 9 380 phrases,

2. Les jeux de données sont divisés en 80% pour l'entraînement, 10% pour la validation et 10% pour le test.

avec une moyenne de 188 phrases par document. Chaque phrase est annotée selon sept rôles rhétoriques.

2. **Legal-Eval** (Kalamkar *et al.*, 2022) : est composé de 214 jugements de la Cour suprême indienne. L'ensemble de données comprend 31 865 phrases, soit en moyenne 115 phrases par document. L'annotation suit un schéma de 13 rôles rhétoriques.
3. **SCOTUS** (Lavissière & Bonnard, 2024) : regroupe 180 jugements de la Cour suprême des États-Unis. Il inclut 22 600 phrases, avec une moyenne de 130 phrases par document. Chaque phrase est annotée selon 13 rôles rhétoriques.

Par ailleurs, d'autres corpus (Dernoncourt *et al.*, 2017; Gonçalves *et al.*, 2020; Lan *et al.*, 2024) sont principalement dédiés à l'analyse de résumés scientifiques et biomédicaux, avec une moyenne de 10 phrases par échantillon. L'absence d'annotations à l'échelle du document les rend inadaptés à notre étude.

L'évaluation des modèles est réalisée à l'aide du F1-score pondéré.

4.2 Modèle SSC pour l'Analyse du Contexte

Afin d'agréger l'ensemble des informations issues du contexte extrait, notre analyse repose sur le modèle hiérarchique HSLN (Brack *et al.*, 2021), avec deux modifications mineures : (1) Motivés par des études d'ablation (Jin & Szolovits, 2018; Chen *et al.*, 2023), qui ont identifié la couche d'enrichissement contextuel des phrases comme étant le principal facteur d'efficacité de HSLN, nous avons supprimé la couche Conditional Random Field (CRF). (2) Nous optimisons uniquement la prédiction de la phrase cible, enrichie avec le contexte sélectionné par nos heuristiques.

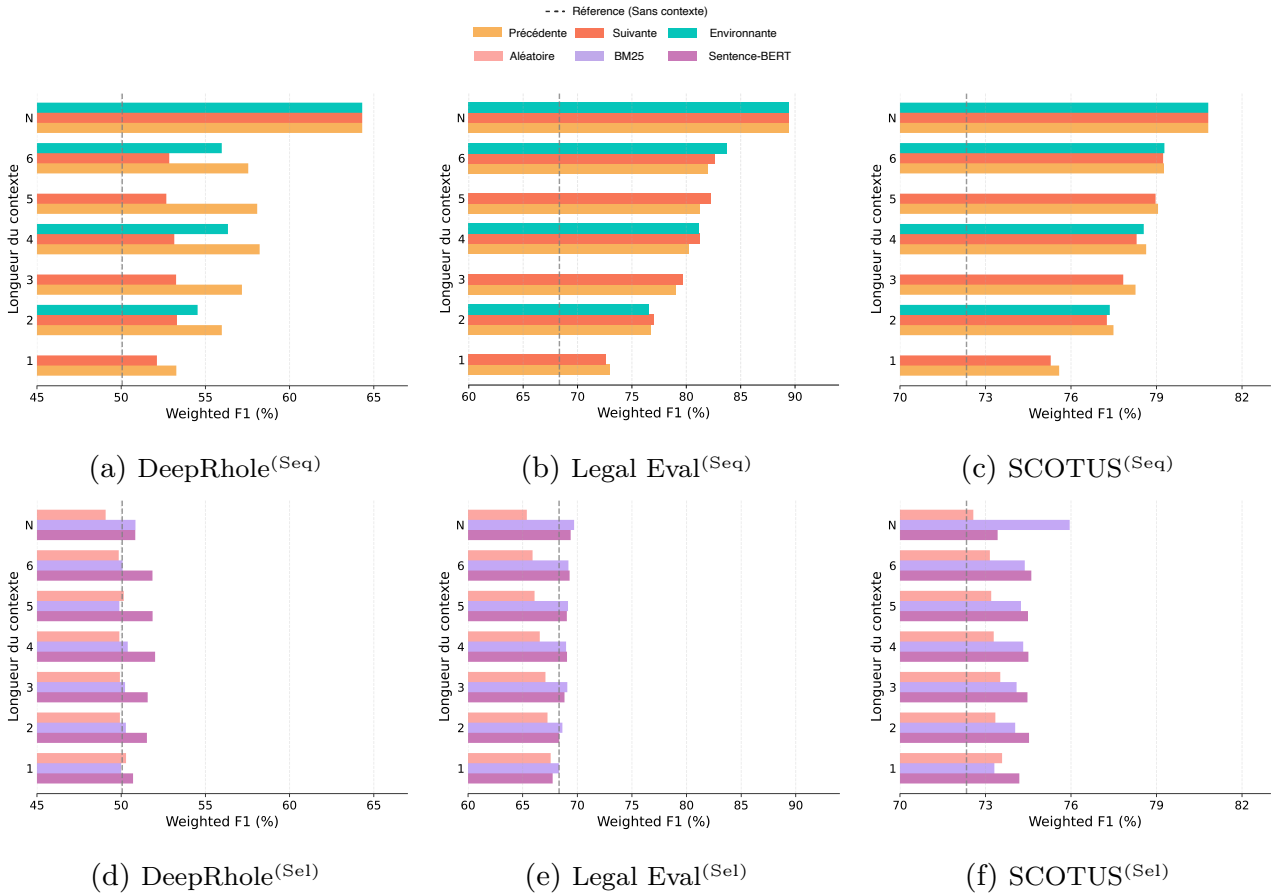
Aperçu du modèle avec nos améliorations :

— **Encodage des mots :** La phrase cible et son contexte récupéré sont encodés séparément à l'aide de BERT (Devlin *et al.*, 2019), générant ainsi des embeddings au niveau des mots.
— **Encodage des phrases :** Un Bi-LSTM (Hochreiter, 1997) traite ces représentations de mots, suivi d'un pooling attentionnel permettant d'obtenir des représentations de phrases.
— **Enrichissement contextuel :** Cette couche modélise les relations inter-phrases afin d'affiner les représentations contextualisées.
— **Couche de sortie :** Une transformation linéaire mappe la représentation de la phrase cible vers des logits, avec des étiquettes prédites via softmax.

5 Résultats

5.1 Analyse du Contexte

La Figure 2 montre que l'intégration du contexte améliore de manière constante la performance de classification sur l'ensemble des jeux de données, indépendamment de l'heuristique appliquée. Ce résultat souligne l'importance d'une sélection efficace du contexte dans SSC.



(a) DeepRhole$^{(\text{Seq})}$ (b) Legal Eval$^{(\text{Seq})}$ (c) SCOTUS$^{(\text{Seq})}$

(d) DeepRhole$^{(\text{Sel})}$ (e) Legal Eval$^{(\text{Sel})}$ (f) SCOTUS$^{(\text{Sel})}$

FIGURE 2 – Scores F1 pondérés en fonction de la longueur du contexte k sur les trois jeux de données. La première ligne (a, b, c) correspond aux résultats obtenus avec les heuristiques séquentielles$^{(\text{Seq})}$, tandis que la seconde ligne (d, e, f) présente les performances des heuristiques sélective$^{(\text{Sel})}$. La condition $k = N$ indique que l'ensemble du document est utilisé afin d'évaluer l'impact de l'ordonnancement des phrases sur la classification. Afin d'assurer une comparabilité entre les différentes heuristiques, la valeur de k est imposée comme un nombre pair pour l'heuristique Environnante.

Heuristiques Séquentielles Elles renforcent progressivement la classification à mesure que davantage de phrases sont incluses. Dans Legal-Eval et SCOTUS, l'heuristique *Environnante* atteint le meilleur score F1 (83.6% et 79.2% pour $k = 6$, respectivement). Cependant, dans DeepRhole, l'heuristique *Précédente* est la plus performante, atteignant 58.2%.
Une analyse plus approfondie indique que 71% des prédictions correctes sont partagées entre les différentes heuristiques séquentielles, suggérant une convergence des performances, quel que soit le choix heuristique adopté.

En revanche, les **Heuristiques Sélectives** offrent des gains marginaux : BM25 étant la plus efficace, atteignant $\approx 74\%$ de score F1 dans SCOTUS lorsque $k \leq 6$.
L'efficacité limitée de ces heuristiques pourrait s'expliquer par deux facteurs : (1) Lorsque les documents manquent de phrases sémantiquement similaires, les heuristiques récupèrent des phrases non pertinentes, ajoutant du bruit (comme observé dans DeepRhole), et (2) les heuristiques sont plus efficaces lorsque les phrases récupérées partagent la même étiquette de la phrase cible (Figure 3 dans l'Annexe).

Modèle	Seq	DeepRhole	Legal Eval	SCOTUS
BERT (référence)	512	52.23	69.74	75.58
+ Précédente		**67.18**†	<u>78.41</u>†	<u>79.74</u>†
+ Suivante		56.72†	**79.74**†	**81.34**†
+ Environnante		<u>62.87</u>†	77.27†	75.47
+ Aléatoire		46.86	67.05	74.70
+ BM25		51.59	69.43	75.96
+ Sentence-BERT		52.23	68.98	76.24
Nomic-BERT (référence)	2048	50.32	68.90	75.50
+ Précédente		**67.89**†	<u>80.54</u>†	<u>81.12</u>†
+ Suivante		57.75†	**81.11**†	**81.32**†
+ Environnante		<u>65.51</u>†	78.20†	80.81†
+ Aléatoire		51.61	68.43	75.73
+ BM25		53.90	70.82‡	77.06†
+ Sentence-BERT		54.02‡	70.76‡	77.17‡
BERT-HSLN (SOTA)	$512 \times N$	54.45	93.06	79.66

TABLE 1 – Performances des PLMs avec la meilleure configuration identifiée dans l'analyse du contexte pour $k \leq 6$ selon chaque heuristique. Les valeurs en gras indiquent l'amélioration significative par rapport à la référence (sans contexte), tandis que les valeurs soulignées correspondent à la deuxième meilleure amélioration. BERT-HSLN représente l'état de l'art pour la tâche de SSC. Les marqueurs † et ‡ signalent une différence statistiquement significative par rapport à la réference aux seuils $p = 0.05$ et $p = 0.01$, respectivement.

À $k = N$, l'expérience sur l'**Ordonnancement des phrases** confirme que SSC est sensible à la structure du contexte—les meilleurs scores étant obtenus lorsque le flux logique du document est préservé. À l'inverse, la réorganisation des phrases à l'aide des Heuristiques Sélectives suggère que prendre l'intégralité du document n'est pas forcément nécessaire ; au contraire, prioriser uniquement les phrases les plus pertinentes permet d'obtenir une performance compétitive.

5.2 Enrichissement du Contexte pour les PLMs

Pour examiner comment les PLMs bénéficient de l'enrichissement contextuel [3], nous menons des expériences avec BERT (Devlin *et al.*, 2019) et le récent Nomic-BERT (Nussbaum *et al.*, 2025), comme indiqué dans la Table 1.

Nos résultats indiquent que les Heuristiques Séquentielles offrent généralement les améliorations les plus importantes, surpassant significativement la réference sans contexte. Notamment, elles surpassent l'état de l'art BERT-HSLN [4], qui traite tous les séquences du document simultanément pour DeepRhole et SCOTUS.
Nous attribuons cette amélioration substantielle, en particulier pour DeepRhole, à deux facteurs : (1) Cet ensemble de données contient moins de labels rhétoriques par rapport aux autres, et (2) D'un point de vue statistique, en moyenne, un nouveau label rhétorique persiste pendant environ 8.56 phrases avant de passer à un autre label. Par conséquent, les modèles

3. Les phrases de contexte ont été intégrées avec la phrase cible dans l'entrée du PLM tout en conservant l'ordre naturel des phrases pour les heuristiques séquentielles.

4. Pour une comparaison équitable, nous nous comparons au modèle original, qui n'inclut pas nos modifications introduites dans l'analyse du contexte.

entièrement hiérarchiques comme BERT-HSLN, qui traitent de plus larges segments du document, peuvent avoir des difficultés avec ces transitions, entraînant une perte d'informations importantes [5] [6].

Cependant, Legal-Eval reste un défi, car ces PLMs n'ont pas encore atteint la performance de l'état de l'art. Une explication plausible est la complexité plus élevée des labels, rendant difficile pour de petits modèles comme BERT d'obtenir une forte discrimination, comme mentionné dans le guide d'annotation de SCOTUS (Lavissière & Bonnard, 2024).

Des résultats supplémentaires obtenus avec RoBERTa (Liu *et al.*, 2019), LegalBERT (Chalkidis *et al.*, 2020) et Longformer (Beltagy *et al.*, 2020) sont présentés dans le Tableau 2 en annexe.

6 Conclusion et Travaux Futurs

Dans cette étude, nous avons examiné comment le rôle du contexte influence la tâche de SSC dans les longs documents juridiques. Nos résultats montrent que les Heuristiques Séquentielles, qui préservent le flux du texte, conduisent systématiquement à des gains de performance plus importants que les Heuristiques Sélectives. De plus, l'enrichissement des PLMs tels que BERT avec un contexte pertinent a permis des améliorations significatives par rapport aux modèles hiérarchiques qui traitent l'ensemble du document. Les travaux futurs devraient donner la priorité à (1) l'extension de l'étude à l'échelle du corpus, afin d'examiner l'impact du contexte issu de plusieurs documents et (2) l'amélioration des Heuristiques Sélectives afin d'extraire un contexte de bonne qualité sans ajouter de bruit.

7 Limitations

Bien que cette étude démontre les avantages de l'information contextuelle pour SSC, certaines limitations doivent être prises en compte :

— Nous avons délibérément maintenu les heuristiques simples, car notre objectif n'était pas d'optimiser la performance maximale. Néanmoins, des approches plus sophistiquées pourraient obtenir des scores plus élevés que ceux que nous présentons.

— Nos expériences se sont concentrées sur un seul document. En pratique, l'intégration du contexte de plusieurs documents pourrait potentiellement offrir une information plus riche pour les Heuristiques Sélectives.

— Nous ne pouvons pas exclure l'hypothèse que nos conclusions sur l'utilité du contexte ne soient pas universellement généralisables à d'autres tâches. Notre analyse s'est centrée sur des ensembles de données juridiques et, par conséquent, des recherches supplémentaires sont nécessaires pour déterminer si des gains similaires apparaîtraient dans d'autres contextes.

5. Un segment fait référence à des unités d'annotation consécutives (phrases) partageant le même label au sein d'un document.

6. Les statistiques ont été calculées sur la base de notre analyse du corpus.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014882 attribuée par GENCI.

Ce travail a été financé, en totalité ou en partie, par l'Agence Nationale de la Recherche (ANR), projet ANR-22-CE38-0004.

Références

Amalvy A., Labatut V. & Dufour R. (2023a). Learning to rank context for named entity recognition using a synthetic dataset. In H. Bouamor, J. Pino & K. Bali, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 10372–10382, Singapore : Association for Computational Linguistics. doi : 10.18653/v1/2023.emnlp-main.642.

Amalvy A., Labatut V. & Dufour R. (2023b). The role of global and local context in named entity recognition. In A. Rogers, J. Boyd-Graber & N. Okazaki, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 714–722, Toronto, Canada : Association for Computational Linguistics. doi : 10.18653/v1/2023.acl-short.62.

BehnamGhader P., Adlakha V., Mosbach M., Bahdanau D., Chapados N. & Reddy S. (2024). LLM2Vec : Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Beltagy I., Peters M. E. & Cohan A. (2020). Longformer : The long-document transformer. *arXiv preprint arXiv :2004.05150*.

Bhattacharya P., Paul S., Ghosh K., Ghosh S. & Wyner A. (2023). Deeprhole : deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, p. 1–38.

Brack A., Hoppe A., Buschermöhle P. & Ewerth R. (2021). Sequential sentence classification in research papers using cross-domain multi-task learning. corr. *arXiv preprint arXiv :2102.06008*.

Chalkidis I., Fergadiotis M., Malakasiotis P., Aletras N. & Androutsopoulos I. (2020). LEGAL-BERT : The muppets straight out of law school. In T. Cohn, Y. He & Y. Liu, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. doi : 10.18653/v1/2020.findings-emnlp.261.

Chen Y., Zhang Y., Wang J. & Zhang X. (2023). YNU-HPCC at SemEval-2023 task 6 : LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar & E. Sartori, Éds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, p. 2075–2081, Toronto, Canada : Association for Computational Linguistics. doi : 10.18653/v1/2023.semeval-1.286.

Choromanski K. M., Likhosherstov V., Dohan D. *et al.* (2021). Rethinking attention with performers. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, Virtual Event, Austria : OpenReview.net. Disponible à l'adresse https://openreview.net/forum?id=Ua6zuk0WRH.

Dernoncourt F., Lee J. Y. & Szolovits P. (2017). Neural networks for joint sentence classification in medical paper abstracts. In M. Lapata, P. Blunsom & A. Koller, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 694–700, Valencia, Spain : Association for Computational Linguistics.

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio, Éds., *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, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

Gangal V., Feng S. Y., Alikhani M., Mitamura T. & Hovy E. (2022). Nareor : The narrative reordering problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10), 10645–10653. DOI : 10.1609/aaai.v36i10.21309.

Gonçalves S., Cortez P. & Moro S. (2020). A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, **32**(11), 6793–6807.

Hochreiter S. (1997). Long short-term memory. *Neural Computation MIT-Press*.

Jin D. & Szolovits P. (2018). Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In E. Riloff, D. Chiang, J. Hockenmaier & J. Tsujii, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3100–3109, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1349.

Kalamkar P., Tiwari A., Agarwal A., Karn S., Gupta S., Raghavan V. & Modi A. (2022). Corpus for automatic structuring of legal documents. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4420–4429, Marseille, France : European Language Resources Association.

Lan M., Zheng L., Ming S. & Kilicoglu H. (2024). Multi-label sequential sentence classification via large language model. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 16086–16104, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.944.

Lavissière M. C. & Bonnard W. (2024). Who's really got the right moves ? analyzing recommendations for writing american judicial opinions. *Languages*, **9**(4). DOI : 10.3390/languages9040119.

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.

Muhammed A., Muslihuddeen H., Sankar S. & Kumar M. A. (2024). Impact of rhetorical roles in abstractive legal document summarization. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, p. 1–6 : IEEE.

Neves M., Butzke D. & Grune B. (2019). Evaluation of scientific elements for text similarity in biomedical publications. In B. Stein & H. Wachsmuth, Éds., *Proceedings of the 6th Workshop on Argument Mining*, p. 124–135, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4515.

Nussbaum Z., Morris J. X., Duderstadt B. & Mulyar A. (2025). Nomic embed : Training a reproducible long context text embedder.

Reimers N. & Gurevych I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng & X. Wan, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

Safder I. & Hassan S.-U. (2019). Bibliometric-enhanced information retrieval : a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, **119**, 257–277.

Shang X., Ma Q., Lin Z., Yan J. & Chen Z. (2021). A span-based dynamic local attention model for sequential sentence classification. In C. Zong, F. Xia, W. Li & R. Navigli, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 198–203, Online : Association for Computational Linguistics. doi : 10.18653/v1/2021.acl-short.26.

Shi F., Chen X., Misra K., Scales N., Dohan D., Chi E. H., Schärli N. & Zhou D. (2023). Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato & J. Scarlett, Éds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 de *Proceedings of Machine Learning Research*, p. 31210–31227 : PMLR.

Trotman A., Puurula A. & Burgess B. (2014). Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, p. 58–65, New York, NY, USA : Association for Computing Machinery. doi : 10.1145/2682862.2682863.

T.y.s.s S., Isaia A., Hong S. & Grabmair M. (2024). HiCuLR : Hierarchical curriculum learning for rhetorical role labeling of legal documents. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 7357–7364, Miami, Florida, USA : Association for Computational Linguistics. doi : 10.18653/v1/2024.findings-emnlp.433.

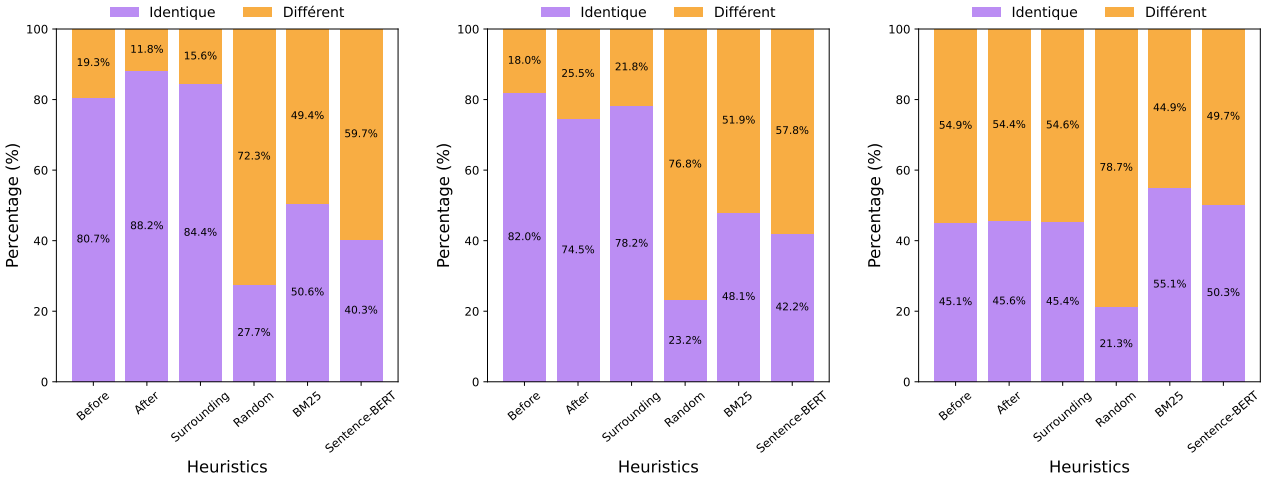
T.y.s.s. S., Sarwat H., Abdou A. M. A. & Grabmair M. (2024). Mind your neighbours : Leveraging analogous instances for rhetorical role labeling for legal documents. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 11296–11306, Torino, Italia : ELRA and ICCL.

Wang L., Yang N. & Wei F. (2024). Learning to retrieve in-context examples for large language models. In Y. Graham & M. Purver, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1752–1767, St. Julian's, Malta : Association for Computational Linguistics.

Wang S., Li B. Z., Khabsa M., Fang H. & Ma H. (2020). Linformer : Self-attention with linear complexity. *CoRR*, **abs/2006.04768**.

Warner B., Chaffin A., Clavié B., Weller O., Hallström O., Taghadouini S., Gallagher A., Biswas R., Ladhak F., Aarsen T., Cooper N., Adams G., Howard J. & Poli I. (2024). Smarter, better, faster, longer : A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Zaheer M., Guruganesh G., Dubey K. A., Ainslie J., Alberti C., Ontanon S., Pham P., Ravula A., Wang Q., Yang L. & Ahmed A. (2020). Big bird : Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 17283–17297 : Curran Associates, Inc.



(a) DeepRhole (b) Legal Eval (c) SCOTUS

FIGURE 3 – Analyse des phrases extraites pour chaque heuristique afin de mesurer la proportion de phrases issues du contexte partageant la même étiquette que la phrase cible.

Modèle	**Seq**	**DeepRhole**	**Legal Eval**	**SCOTUS**
Roberta-base (référence)	512	52.63	72.43	76.28
+ Précédente		**68.29**†	78.3†	**81.75**†
+ Suivante		60.3†	**80.12**†	81.43†
+ Environnante		63.86†	78.40†	80.10†
+ Aléatoire		50.04	72.35	75.79
+ BM25		53.54	72.79	77.78‡
+ Sentence-BERT		53.33	73.25‡	77.84‡
Legal-BERT (référence)	512	54.06	69.43	76.85
+ Précédente		**69.10**†	79.65†	81.40†
+ Suivante		63.19†	**80.99**†	**82.81**†
+ Environnante		67.15†	78.55†	78.72
+ Aléatoire		50.32	68.55	76.56
+ BM25		54.59	70.77‡	77.06
+ Sentence-BERT		56.30	70.55	77.47
Longformer (référence)	4096	53.83	72.57	76.26
+ Précédente		**67.62**†	79.89†	**81.58**†
+ Suivante		61.16†	**80.09**†	81.09†
+ Environnante		64.83†	73.09†	81.35†
+ Aléatoire		52.55	72.54	75.78
+ BM25		54.82	73.22	77.44†
+ Sentence-BERT		54.3	77.95‡	77.47‡

TABLE 2 – Performances des PLMs avec la meilleure configuration identifiée dans l'analyse du contexte pour $k \leq 6$ selon chaque heuristique. Les valeurs en gras indiquent l'amélioration significative par rapport à la réference (sans contexte), tandis que les valeurs soulignées correspondent à la deuxième meilleure amélioration. Les marqueurs † et ‡ signalent une différence statistiquement significative par rapport à la réference aux seuils $p = 0.05$ et $p = 0.01$, respectivement.

Target Sentence : *"This case focuses upon the requirement of 'fair presentation.'"*

Heuristic	Extracted Sentence
Précédente	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)."*
Suivante	*"Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Environnante	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)."* *"Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Aléatoire	*"In such instances, the nature of the issue may matter more than does the legal validity of the lower court decision."*
BM25	*"For another thing, the opinion-reading requirement would impose a serious burden upon judges of state appellate courts, particularly those with discretionary review powers."*
Sentence-BERT	*"The petition provides no citation of any case that might have alerted the court to the alleged federal nature of the claim."*

TABLE 3 – Exemples de phrases extraites à l'aide de différentes heuristiques depuis le jeu de données SCOTUS.

Jeu de données	Source	Sous-domaine	Instances entraînement/dév/test	Cibles
DeepRhole	(Bhattacharya *et al.*, 2023)	Droit indien	7 591 / 857 / 932	7 classes
Legal Eval	(Kalamkar *et al.*, 2022)	Droit indien	28 986 / 2 879 / 4 158	13 classes
SCOTUS	(Lavissière & Bonnard, 2024)	Droit américain	21 396 / 2 450 / 2 481	13 classes

TABLE 4 – Statistiques des jeux de données utilisés pour l'évaluation.