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

Optimisation de la Recherche d'Information Juridiques à travers l'Agrégation des Signaux Contextuels Multi-niveaux des Modèles de Langue Préentraînés

Eya HAMMAMI Mohand BOUGHANEM Taoufiq DKAKI
IRIT, Université Toulouse, Toulouse, France
{Eya.Hammami, Mohand.Boughanem, Taoufiq.Dkaki}@irit.fr

RÉSUMÉ

L'accès croissant aux documents juridiques sous format numérique crée à la fois des opportunités et des défis pour les professionnels du droit et les chercheurs en intelligence artificielle. Cependant, bien que les Modèles de Langue Préentraînés (PLMs) excellent dans diverses tâches de TAL, leur efficacité dans le domaine juridique demeure limitée, en raison de la longueur et de la complexité des textes. Pour répondre à cette problématique, nous proposons une approche exploitant les couches intermédiaires des modèles du Transformer afin d'améliorer la représentation des documents juridiques. En particulier, cette méthode permet de capturer des relations syntaxiques et sémantiques plus riches, tout en maintenant les interactions contextuelles au sein du texte. Afin d'évaluer notre approche, nous avons mené des expérimentations sur des ensembles de données juridiques publiques, dont les résultats obtenus démontrent son efficacité pour diverses tâches, notamment la recherche et la classification de documents.

ABSTRACT

Optimization of Information Retrieval and Legal Case Classification through the Aggregation of Multi-Level Contextual Signals from Pretrained Language Models.

The increasing availability of legal documents in digital format creates opportunities and challenges for legal professionals and artificial intelligence researchers. However, although Pretrained Language Models (PLMs) excel in various NLP tasks, their effectiveness in the legal domain remains limited due to the length and complexity of legal texts. To address this issue, we propose an approach that leverages the intermediate layers of transformer-based models to enhance the representation of legal documents. In particular, this method captures richer syntactic and semantic relationships while preserving contextual interactions within the text. To evaluate our approach, we conducted experiments on publicly available legal datasets. The obtained results demonstrate its effectiveness across various tasks, including information retrieval and document classification..

MOTS-CLÉS : Recherche d'Information, Modèles de Langue Préentraînés, Domaine Juridique, Traitement Automatique des Langues, Apprentissage Automatique.

KEYWORDS: Information Retrieval, Pre-trained Language Models, Legal Domain, Natural Language Processing, Machine Learning.

Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA), pages 135–150.

1 Introduction

La Recherche d'Informations (IR) juridiques consiste à retrouver, au sein de vastes collections de textes juridiques, les documents les plus pertinents en réponse à une requête donnée. Cette tâche vise à identifier des affaires similaires, des lois applicables ou des précédents pertinents à partir de sources textuelles souvent non structurées. Ces systèmes jouent un rôle crucial dans l'aide à la prise de décision juridique et la préparation des affaires. Cependant, la performance des systèmes IR dans le domaine juridique dépend fortement de la manière dont les documents sont représentés. Compte tenu de la complexité, du caractère non structuré et de la terminologie spécifique aux documents juridiques, le choix de méthodes de représentation appropriées est essentiel pour capturer les subtilités du langage juridique. Dans ce contexte, différentes approches de recherche d'informations ont été développées pour améliorer l'accès aux textes juridiques, reposant principalement sur des modèles lexicaux et sémantiques. Les modèles lexicaux, tels que BM25, exploitent la correspondance exacte des termes pour évaluer la pertinence des documents. En parallèle, les modèles sémantiques, notamment ceux basés sur les embeddings de mots ou les modèles de langue préentraînés, tentent d'améliorer la compréhension du contexte en capturant des relations plus complexes entre les termes. Plus récemment, les architectures de type Transformers ont permis d'améliorer considérablement la qualité de la recherche en exploitant des représentations contextuelles riches. Toutefois, l'un des défis majeurs dans l'application de ces méthodes au domaine juridique réside dans la taille volumineuse des documents. Par conséquent, ce travail se concentre sur les récentes techniques de représentation d'embeddings contextuels afin d'améliorer la capacité des systèmes IR à fournir des résultats plus précis et adaptés au contexte juridique. Plus précisément, cette étude se concentre sur deux défis majeurs dans le traitement des textes juridiques. D'une part, la longueur et la complexité des documents juridiques. D'autre part, la spécificité du langage juridique, marquée par une terminologie spécialisée et des expressions nuancées, limite souvent l'efficacité des modèles de langue préentraînés, impliquant ainsi des adaptations ciblées pour mieux capturer ces subtilités. Pour résoudre ces problèmes, nous proposons plusieurs contributions essentielles dans le domaine du traitement des textes juridiques :

— Une approche consistant à agréger les sorties d'embeddings de différentes couches cachées intermédiaires de chaque modèle de langue préentraîné. Cette méthode multi-niveaux permet d'obtenir une compréhension plus riche et nuancée des textes juridiques, essentielle pour diverses tâches de traitement des textes.

— L'évaluation de ces méthodologies sur plusieurs tâches de traitement des textes juridiques, y compris le classement des précédents juridiques et la classification des affaires juridiques, démontrant leur potentiel à améliorer l'analyse des documents juridiques dans divers contextes.

2 Travaux connexes

Les modèles de langue préentraînés basés sur l'architecture Transformer (Vaswani *et al.*, 2017) ont récemment obtenu des performances de pointe en TALN, y compris en recherche d'information juridique (LIR). Plusieurs travaux, tels que ceux de (Mokanov *et al.*, 2019; Nguyen *et al.*, 2020; Shaghaghian *et al.*, 2020; Vold & Conrad, 2021; Licari & Comandè, 2022; Claveau, 2021), ont exploité des modèles comme BERT, RoBERTa et ALBERT en les ajustant pour améliorer la recherche et le classement des documents juridiques. Par exemple, (Shao *et al.*, 2020) ont développé BERT-PLI, un modèle optimisé sur un jeu de données de raisonnement juridique pour évaluer la pertinence des relations entre les affaires. Cependant, (Chalkidis *et al.*, 2020) ont montré que les modèles de langue

généralistes ne sont pas toujours adaptés aux spécificités du domaine juridique, ce qui a motivé la création de LEGAL-BERT, un modèle préentraîné spécifiquement sur des textes légaux. (Nguyen *et al.*, 2022) ont également proposé une approche combinant BM25 pour la correspondance lexicale et BERT pour l'analyse sémantique dans la recherche des affaires juridiques.

Toutefois, la question majeur dans l'application des Transformers aux documents juridiques réside dans leur longueur, souvent incompatible avec les contraintes des modèles traditionnels, comme la limite de 512 tokens imposée par BERT. Pour répondre à cette problématique, plusieurs solutions ont été proposées. Par exemple, (Xiao *et al.*, 2021) ont développé Lawformer, basé sur Longformer, pour mieux gérer les longues séquences grâce à l'attention globale et à une fenêtre glissante. Ce modèle, préentraîné sur un vaste corpus de documents légaux chinois, a démontré une amélioration significative dans des tâches telles que la prédiction des jugements et la sélection d'affaires similaires. De son côté, (Limsopatham, 2021) a exploré l'impact du préentraînement de BERT sur des textes juridiques spécialisés, mettant en évidence l'intérêt de modèles adaptés comme BigBird et Longformer. Ces recherches ont montré que le préentraînement sur des données spécifiques, associé à des stratégies comme la troncature intelligente et le pooling sur segments de texte, permet d'améliorer considérablement la performance des modèles sur des documents longs.

D'autres approches ont cherché à optimiser la recherche d'information juridique en combinant des techniques lexicales et sémantiques. (Bui *et al.*, 2022) ont ainsi développé un modèle intégrant LEGAL-BERT pour l'analyse sémantique et BM25 pour la sélection lexicale, améliorant la précision des résultats. (Askari *et al.*, 2022) ont proposé une reformulation des requêtes via KeyBERT, suivie d'un résumé automatique avec Longformer-Encoder-Decoder (LED) afin d'adapter les requêtes à la longueur des documents. Une fois ces requêtes reformulées, elles sont utilisées pour extraire un ensemble initial d'affaires via BM25, puis affinées avec Sentence-BERT pour mieux capturer les similarités sémantiques. Dans cette optique, (Nigam *et al.*, 2022) ont appliqué une approche combinant BM25 et Sentence-BERT pour sélectionner les affaires les plus pertinents, en exploitant une mesure de similarité cosinus avec max-pooling pour comparer les représentations textuelles. Par ailleurs, (Rabelo *et al.*, 2022) ont mis en œuvre une stratégie combinant deux modèles de sentence-transformers, all-mpnet-base-v2 et mpnet-base, afin de créer des représentations vectorielles multidimensionnelles des documents. Ces vecteurs sont ensuite utilisés pour entraîner un classificateur Gradient Boosting, permettant d'identifier les affaires les plus pertinentes. Plus récemment, (Li *et al.*, 2023) ont développé SAILER, un modèle sensible à la structure des textes juridiques, conçu pour capturer la logique d'écriture et les éléments juridiques clés, améliorant ainsi la recherche de documents légaux longs.

Malgré ces avancées, la majorité des travaux en LIR se limitent à l'exploitation de la dernière couche des modèles de langue préentraînés, négligeant les informations riches capturées par les couches intermédiaires. Pourtant, plusieurs études (Yang & Zhao, 2019; van Aken *et al.*, 2019; Huang *et al.*, 2021) ont démontré que les représentations les plus transférables proviennent souvent des couches intermédiaires, tandis que la dernière couche est davantage spécialisée pour la modélisation du langage. (Udagawa *et al.*, 2024) ont également montré que l'agrégation de plusieurs couches intermédiaires améliore la robustesse des modèles, notamment dans la reconnaissance automatique de la parole, renforçant ainsi leur rôle clé dans l'extraction d'informations contextuelles. Cependant, à notre connaissance, aucune étude en LIR n'a encore exploré de manière approfondie l'exploitation des couches intermédiaires des modèles de langue préentraîné. C'est précisément l'objet de notre approche à ce problème. Elle vise à exploiter les représentations intermédiaires issues de différentes couches des modèles de langue préentraînés, tout en prenant en compte l'intégralité du document juridique. Cette méthode permet de préserver les relations contextuelles entre les phrases adjacentes et d'intégrer davantage d'informations pour les tâches de recherche d'information juridique .

3 Représentation basée sur plusieurs couches cachées intermédiaires

Dans cette section, nous présentons notre approche proposée, qui exploite plusieurs couches cachées intermédiaires des PLMs. Nous discutons également des motivations théoriques qui rendent cette approche particulièrement intéressante.

3.1 Rôle des couches cachées dans les PLMs : de la compréhension syntaxique à la sémantique de haut niveau

Les travaux antérieurs de (Yang & Zhao, 2019; van Aken *et al.*, 2019; Huang *et al.*, 2021), qui ont analysé le rôle des différentes couches cachées des PLMs, ont démontré que les premières couches capturent principalement des propriétés syntaxiques élémentaires du texte. Cela inclut le tagging morphosyntaxique (part-of-speech tagging) et les dépendances grammaticales de base entre les mots proches. Ces couches sont sensibles au contexte local et aux caractéristiques lexicales, ce qui les rend essentielles pour comprendre la structure des phrases. Ensuite, les couches intermédiaires intègrent progressivement des structures syntaxiques plus complexes avec des informations sémantiques. À ce niveau, le modèle commence à construire une compréhension plus approfondie du sens des phrases, notamment en établissant des relations entre les entités présentes dans le texte. Ces couches jouent également un rôle important dans la résolution de coréférence (c'est-à-dire l'identification des noms faisant référence à la même entité) et la capture de certaines relations sémantiques. Enfin, les couches supérieures sont hautement spécialisées et adaptées aux tâches spécifiques sur lesquelles le modèle a été préentraîné ou affiné. Dans le cas de modèles comme BERT, qui sont entraînés sur une variété de tâches, ces couches combinent les informations des couches inférieures pour former des représentations sémantiques de haut niveau. Ces couches sont cruciales pour les tâches qui nécessitant une compréhension profonde du contexte, ainsi que pour l'analyse de la structure argumentative ou narrative d'un document. Dans le cadre de cet article, nous concentrons sur l'expérimentation et l'évaluation de la contribution des différentes couches cachées à la représentation des documents juridiques.

3.2 Phase de représentation des documents

Notre principal objectif dans ce travail est d'obtenir des représentations d'embeddings utiles et significatives pour l'ensemble du document juridique. Pour cela, nous avons choisi d'exploiter les représentations des couches cachées intermédiaires, ainsi que celles des couches supérieures des modèles de langue préentraînés. Dans un premier temps, nous avons utilisé la technique de fenêtre glissante permettant de contourner la contrainte de longueur des séquences imposée par les modèles Transformers. Elle permet de prendre en compte des documents de toute longueur, quel que soit le modèle préentraîné utilisé. Plus précisément, elle divise le texte d'entrée en segments plus petits, selon des paramètres de taille de fenêtre WS et de longueur de pas S. Le nombre total de segments générés est donné par :

$$N = \left\lfloor \frac{|d| - WS}{S} \right\rfloor + 1 \tag{1}$$

où $|d|$ représente la longueur totale du document. Chaque segment d_i est défini comme :

$$d_i = d[i : i + WS], \quad \text{pour } i \in \{0, S, 2S, \ldots, (N-1)S\} \tag{2}$$

Ensuite, un embedding est générés pour chaque segment, afin d'obtenir une représentation plus fine du texte. Ces embeddings sont produits à l'aide d'un PLM, qui extrait un vecteur de représentation pour chaque segment :

$$e_i = PLM(d_i) \tag{3}$$

L'ensemble des embeddings obtenus est donc :

$$E = \{e_1, e_2, \ldots, e_N\} \tag{4}$$

Par la suite, les embeddings des segments sont agrégés en un unique vecteur d'embedding, permettant de représenter le document globalement. À cet effet, plusieurs méthodes d'agrégation peuvent être utilisées, notamment la concaténation, la moyenne des embeddings et la moyenne pondérée des embeddings, chacune présentant des avantages spécifiques selon la tâche considérée. Dans ce travail, nous avons opté pour le moyenne des embeddings, qui garantit que tous les embeddings de tous les segments sont traités de manière équitable, assurant ainsi que l'ensemble du document est efficacement pris en compte dans sa représentation.

$$E_d = \frac{1}{N} \sum_{i=1}^{N} e_i \tag{5}$$

En parallèle avec cette méthode de fenêtre glissante, nous générons également l'embedding à partir des couches cachées intermédiaires du modèle de langue préentraîné, en agrégeant la sortie de n couches. Plus précisément, le texte d'entrée est découpé en tokens puis passé à travers le modèle, produisant une séquence de représentations des couches cachées pour chaque position de token. Ces représentations capturent des caractéristiques linguistiques de plus en plus abstraites au fur et à mesure qu'elles traversent les couches (van Aken *et al.*, 2019). Pour générer l'embedding d'une séquence donnée, les sorties des n couches cachées sont agrégées ensemble. Ce processus d'agrégation intègre les informations issues de plusieurs couches, offrant ainsi une représentation plus riche du texte. À ce niveau, nous avons choisi d'évaluer la fonction de moyenne qui considère les embeddings des couches cachées de manière équitable, en attribuant le même poids à chaque couche. Ainsi, chaque embedding d'un segment e_i est calculé comme suit :

$$e_i = \sum_{j=1}^{n} \alpha \cdot h_j(d_i) \tag{6}$$

où $h_j(d_i)$ représente la sortie de la j-ième couche cachée pour le segment d_i et chaque couche reçoit le même poids :

$$\alpha = \frac{1}{n} \tag{7}$$

L'agrégation des embeddings sur l'ensemble des segments permet d'obtenir la représentation finale du document :

$$Final_{E_d} = \frac{1}{N} \sum_{i=1}^{N} e_i \tag{8}$$

où chaque e_i est lui-même le résultat de l'agrégation des couches cachées.

Enfin, les représentations obtenues après agrégation servent d'embeddings qui encodent de manière complète l'information contextuelle du texte d'entrée. Ces embeddings peuvent ensuite être utilisés comme vecteurs de caractéristiques pour les tâches en aval.

4 Évaluation expérimentale et analyse des résultats

4.1 Paramètres expérimentaux

4.1.1 Jeux de données :

Nous avons utilisé trois jeux de données juridiques publics, représentant des documents d'affaires juridiques rédigés en anglais : COLIEE2022 (Kim *et al.*, 2022), AILA2019 (Bhattacharya *et al.*, 2019) et IRLeD2017 (Mandal *et al.*, 2017).

Le Tableau 1 présente des statistiques détaillées sur ces trois jeux de données.

TABLE 1 – Statistiques des collections de jeux de données juridiques.

JEUX DE DONNÉES	COLIEE 2022	IRLeD 2017	AILA 2019
Affaires de requêtes d'entraînement	900	1000	50
Affaires candidats d'entraînement par requête	3515	2000	2914
Affaires de requêtes de test	300	–	–
Affaires candidats de test par requête	1263	–	–
Longueur moyenne des affaires de requêtes d'entraînement	2724.42	304.94	251.04
Longueur moyenne des affaires candidats d'entraînement	27636.45	3524.97	1658.12
Longueur moyenne des affaires de requêtes de test	3513.18	–	–
Longueur moyenne des affaires candidats de test par requête	27759.92	–	–

4.1.2 Métriques d'évaluation :

Dans ce travail, nous avons évalué notre approche sur deux tâches distinctes : la classification et la recherche d'information appliquées aux documents juridiques. Afin d'assurer une évaluation comparative, nous avons suivi les mêmes métriques d'évaluation que celles utilisées dans les compétitions. Pour COLIEE 2022, nous rapportons la Précision, le Rappel et le F1-score. Plus précisément, nous indiquons F1-score@5 et F1-score@10 pour la tâche de recherche d'informations. Concernant IRLeD 2017 et AILA 2019, nous rapportons la Mean Average Precision (MAP) ainsi que la Précision@10.

4.1.3 Configuration des hyperparamètres :

Tous les PLMs utilisés dans cet article ont été directement importés depuis Hugging Face [1], sans aucun ajustement par fine-tuning. Le Tableau 2 présente leur principales caractéristiques. Là où nous avons utilisé cinq modèles Transformers de phrases préentraînés, basés sur des encodeurs [2], sélectionnés pour leurs performances moyennes optimales sur 14 tâches variées issues de différents domaines. De plus, nous avons intégré un PLM spécifique au domaine juridique. En association avec ces PLMs, nous avons exploré quatre configurations distinctes des paramètres WS et S afin de mettre en œuvre la méthode de représentation des documents : $(WS, S) = \{(128, 64), (256, 128), (512, 256)\}$ Pour la tâche de classification, nous avons exclusivement utilisé le jeu de données COLIEE2022, car il contient à la fois un ensemble d'entraînement et un ensemble de test. Toutefois, ce jeu de données étant

1. https ://huggingface.co/
2. https ://www.sbert.net/

fortement déséquilibré, nous avons dû ajuster la distribution des classes afin d'obtenir un ensemble d'entraînement plus représentatif, en générant environ 1000 échantillons négatifs par requête. Pour cela, nous avons appliqué une réduction d'échantillonnage en limitant le nombre d'échantillons négatifs à 1000, en nous appuyant sur OkapiBM25 pour extraire que 1000 échantillons négatifs. Parallèlement, la classe positive étant largement sous-représentée (moins de 5 échantillons positifs par requête en moyenne), nous avons procédé à un suréchantillonnage de ces exemples via une réplication simple (Rabelo *et al.*, 2022). Enfin, nous avons segmenté notre ensemble d'entraînement en dix sous-ensembles, afin d'assurer une répartition équilibrée de la variable cible et des groupes de requêtes définis. Enfin, nous avons utilisé l'algorithme Gradient Boosting de Scikit-Learn [3] avec ses paramètres par défaut comme modèle de classification. En complément, nous avons testé un réseau de neurones simple composé de trois couches denses. La première couche dense comprend 128 neurones, suivie d'une seconde couche de 64 neurones. Les deux premières couches utilisent la fonction d'activation Rectified Linear Unit (ReLU) afin de favoriser l'apprentissage des représentations non linéaires. Enfin, une troisième couche dense, constituée d'un seul neurone, applique une activation sigmoïde, adaptée à la nature binaire de la tâche de classification. L'ensemble des expérimentations a été réalisé à l'aide de PyTorch [4] et Keras [5].

TABLE 2 – PLMs entraînés et ajustés sur un large ensemble de données diversifié.

PLMs	Modèle de base	Dimensions	Taille	Longueur max.
all-roberta-large-v1	roberta-large	1024	1360 MB	512
all-mpnet-base-v1	mpnet-base	768	420 MB	512
gtr-t5-large	t5-large	768	640 MB	512
gtr-t5-xl	t5-3b	768	2370 MB	512
sentence-t5-xl	t5-3b	768	2370 MB	512
bert-base	bert-base	768	440 MB	512
legal-bert-base	bert-base	768	440 MB	512

4.2 Résultats expérimentaux :

4.2.1 Impact de la fenêtre glissante :

Le Tableau 3 présente les résultats obtenus en testant différentes combinaisons de la taille de la fenêtre glissante (WS) et du pas (S) sur plusieurs PLMs, dont RoBERTa, MPNet, BERT et Legal-BERT, pour la sélection des affaires juridiques sur COLIEE 2022, AILA 2019 et IRLeD 2017. L'évaluation vise à mesurer l'impact de la fenêtre glissante sur la sélection des affaires juridiques afin d'identifier les paramètres les plus efficaces. Trois configurations ont été comparées : ($WS = 128$, $S = 64$), ($WS = 256$, $S = 128$), et ($WS = 512$, $S = 256$). En plus de ces configurations, une condition sans fenêtre glissante a été testée. Dans ce cas, une seule fenêtre de la taille maximale autorisée par le modèle est utilisée, et le reste du texte est tronqué, ce qui permet d'évaluer l'apport réel de la technique de fenêtre glissante par rapport à une approche sans segmentation du texte.

Les résultats montrent que les modèles réagissent différemment selon la taille de la fenêtre et le pas. Sur COLIEE 2022, legal-bert-base affiche de meilleures performances avec des fenêtres plus grandes,

3. https ://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
4. https ://pytorch.org/
5. https ://keras.io/

suggérant qu'un contexte étendu améliore la sélection des affaires pertinents. De manière générale, les modèles performent mieux avec de grandes fenêtres, tandis que les résultats sans fenêtre glissante sont nettement inférieurs, confirmant son rôle clé dans la capture des détails essentiels. Une tendance similaire est observée sur AILA 2019, où une fenêtre plus large améliore les performances, soulignant la nécessité d'un contexte étendu pour associer correctement les requêtes aux affaires pertinents. Les résultats sans fenêtre glissante y sont aussi significativement plus faibles, renforçant l'importance de cette technique. Son impact est encore plus marqué sur IRLeD 2017, où l'augmentation de la taille de la fenêtre améliore substantiellement les scores MAP, confirmant son rôle dans l'optimisation de la recherche d'information juridique. En plus, des modèles comme gtr-t5-large et legal-bert-base affichent des gains notables avec des fenêtres plus grandes, exploitant mieux le contexte pour une recherche plus efficace.

Globalement, élargir la fenêtre glissante améliore les performances sur tous les jeux de données, montrant qu'un contexte plus large favorise le classement des affaires juridiques. Bien que l'impact du pas soit moindre, une valeur plus élevée permet de couvrir plus de contenu tout en réduisant les parties non pertinentes. Enfin, les résultats sans fenêtre glissante restent systématiquement inférieurs, prouvant que cette méthode est essentielle pour traiter efficacement les longs documents juridiques. De plus, Legal-BERT-Base se distingue comme le modèle le plus performant sur l'ensemble des jeux de données, confirmant son efficacité pour la recherche d'informations juridiques.

4.2.2 Impact de l'agrégation des couches cachées :

Les tableaux 4 et 5 comparent les performances de différents PLMs pour la recherche et la classification d'affaires juridiques sur COLIEE 2022, AILA 2019 et IRLeD 2017. Pour la classification, seul COLIEE 2022 a été utilisé, car il a déjà servi dans des tâches de classification binaire. Les modèles ont été évalués en combinant différentes couches cachées, de la dernière seule à l'agrégation des deux, trois ou quatre dernières.

Pour la recherche d'affaires (Tableau 4), utiliser uniquement la dernière couche donne des résultats inférieurs à l'agrégation de plusieurs couches. La performance s'améliore avec l'ajout de couches, là où Legal-BERT affichant des gains notables lorsqu'il intègre les quatre dernières couches, confirmant ainsi qu'une représentation plus riche améliore la sélection des affaires pertinentes. Cette tendance se vérifie sur tous les jeux de données.

En classification (Tableau 5), Gradient Boosting et Réseaux de Neurones ont été testés, avec la précision, le rappel et le F1-score comme métriques. Gradient Boosting surpasse généralement les réseaux de neurones, notamment avec l'agrégation de plusieurs couches cachées. Plus précisément l'agrégation des quatre dernières couches améliore systématiquement les performances et Legal-BERT affichant les meilleurs scores. Ces résultats confirment son efficacité pour l'analyse des textes juridiques lorsqu'une information contextuelle plus profonde est exploitée.

Globalement, combiner plusieurs couches cachées améliore les performances sur les deux tâches, prouvant l'importance d'une meilleure représentation contextuelle. Ainsi, Legal-BERT se distingue comme le modèle le plus performant pour les textes juridiques. Dans l'expérimentation suivante, comparant le pooling CLS et le pooling moyen de tous les tokens, nous avons conservé l'agrégation des quatre dernières couches cachées ainsi que les paramètres de la fenêtre glissante (WS = 512, S = 256).

TABLE 3 – Résultats obtenus avec différentes combinaisons des paramètres Taille de Fenêtre et Pas pour la sélection des affaires juridiques

Fenêtre glissante (WS = 128, S = 64)						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,176	0,163	0,104	0,026	0,193	0,122
all-mpnet-base-v1	0,204	0,177	0,103	0,026	0,202	0,130
gtr-t5-large	0,219	0,188	0,113	0,026	0,245	0,158
gtr-t5-xl	0,224	0,204	0,114	0,026	0,252	0,163
sentence-t5-xl	0,206	0,199	0,104	0,026	0,236	0,146
bert-base	0,207	0,230	0,107	0,026	0,243	0,143
legal-bert-base	0,245	0,247	0,116	0,026	0,256	0,168
Fenêtre glissante (WS = 256, S = 128)						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,223	0,198	0,113	0,038	0,216	0,132
all-mpnet-base-v1	0,252	0,215	0,112	0,034	0,224	0,139
gtr-t5-large	0,268	0,223	0,115	0,028	0,266	0,168
gtr-t5-xl	0,267	0,241	0,118	0,038	0,273	0,172
sentence-t5-xl	0,249	0,234	0,111	0,032	0,259	0,156
bert-base	0,252	0,226	0,117	0,036	0,269	0,159
legal-bert-base	0,277	0,252	0,120	0,041	0,278	0,176
Fenêtre glissante (WS = 512, S = 256)						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,252	0,227	0,116	0,040	0,239	0,142
all-mpnet-base-v1	0,283	0,245	0,112	0,034	0,245	0,149
gtr-t5-large	0,291	0,247	0,139	0,042	0,284	0,176
gtr-t5-xl	0,313	0,250	0,124	0,046	0,291	0,180
sentence-t5-xl	0,286	0,258	0,117	0,036	0,280	0,166
bert-base	0,301	0,241	0,123	0,039	0,279	0,167
legal-bert-base	**0,342**	**0,269**	**0,142**	**0,051**	**0,295**	**0,185**
Sans Fenêtre Glissante						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,140	0,103	0,074	0,018	0,170	0,111
all-mpnet-base-v1	0,176	0,159	0,072	0,018	0,180	0,119
gtr-t5-large	0,145	0,148	0,082	0,018	0,222	0,147
gtr-t5-xl	0,166	0,147	0,084	0,018	0,229	0,152
sentence-t5-xl	0,160	0,145	0,074	0,018	0,214	0,135
bert-base	0,133	0,086	0,071	0,011	0,201	0,121
legal-bert-base	0,176	0,173	0,089	0,018	0,233	0,156

TABLE 4 – Résultats obtenus avec différentes combinaisons de niveaux de couches cachées pour la sélection des affaires juridiques

Dernière couche cachée						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,252	0,227	0,116	0,040	0,239	0,142
all-mpnet-base-v1	0,283	0,245	0,112	0,034	0,245	0,149
gtr-t5-large	0,291	0,247	0,139	0,042	0,284	0,176
gtr-t5-xl	0,313	0,250	0,124	0,046	0,291	0,180
sentence-t5-xl	0,286	0,258	0,117	0,036	0,280	0,166
bert-base	0,303	0,261	0,131	0,038	0,279	0,156
legal-bert-base	0,318	0,265	0,142	0,065	0,294	0,186
Moyenne des 2 dernières couches cachées						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,276	0,247	0,153	0,064	0,262	0,154
all-mpnet-base-v1	0,285	0,240	0,150	0,058	0,269	0,161
gtr-t5-large	0,295	0,247	0,177	0,066	0,307	0,187
gtr-t5-xl	0,313	0,274	0,162	0,070	0,314	0,190
sentence-t5-xl	0,329	0,275	0,155	0,060	0,303	0,176
bert-base	0,317	0,268	0,159	0,062	0,281	0,159
legal-bert-base	0,333	0,280	0,182	0,071	0,319	0,197
Moyenne des 3 dernières couches cachées						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,345	0,282	0,194	0,076	0,284	0,164
all-mpnet-base-v1	0,347	0,276	0,190	0,070	0,292	0,171
gtr-t5-large	0,357	0,286	0,217	0,078	0,329	0,197
gtr-t5-xl	0,375	0,310	0,202	0,082	0,335	0,200
sentence-t5-xl	0,390	0,308	0,196	0,072	0,324	0,186
bert-base	0,382	0,300	0,192	0,072	0,319	0,198
legal-bert-base	0,396	0,315	0,223	0,089	0,341	0,211
Moyenne des 4 dernières couches cachées						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,394	0,315	0,220	0,080	0,307	0,176
all-mpnet-base-v1	0,396	0,308	0,216	0,074	0,313	0,182
gtr-t5-large	0,398	0,316	0,223	0,080	0,350	0,206
gtr-t5-xl	0,417	0,339	0,225	0,084	0,356	0,210
sentence-t5-xl	0,428	0,338	0,219	0,074	0,346	0,196
bert-base	0,396	0,318	0,222	0,079	0,341	0,206
legal-bert-base	**0,433**	**0,343**	**0,231**	**0,089**	**0,362**	**0,216**

TABLE 5 – Résultats obtenus avec différentes combinaisons de niveaux de couches cachées pour la classification des affaires juridiques en utilisant les classificateurs Gradient Boosting et Réseau de Neurones sur le jeu de données COLIEE 2022

Dernière couche cachée						
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,254	0,520	0,341	0,251	0,471	0,327
all-mpnet-base-v1	0,272	0,538	0,361	0,253	0,468	0,328
gtr-t5-large	0,278	0,521	0,363	0,248	0,476	0,326
gtr-t5-xl	0,277	0,522	0,362	0,254	0,478	0,332
sentence-t5-xl	0,279	0,524	0,364	0,257	0,481	0,335
bert-base	0,271	0,535	0,360	0,249	0,477	0,327
legal-bert-base	0,280	0,539	0,368	0,259	0,482	0,337
Moyenne des 2 dernières couches cachées						
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,264	0,513	0,349	0,258	0,476	0,335
all-mpnet-base-v1	0,272	0,538	0,362	0,261	0,472	0,336
gtr-t5-large	0,277	0,532	0,364	0,256	0,481	0,334
gtr-t5-xl	0,274	0,535	0,363	0,261	0,483	0,339
sentence-t5-xl	0,278	0,542	0,367	0,264	0,485	0,342
bert-base	0,276	0,538	0,365	0,252	0,481	0,331
legal-bert-base	0,282	0,540	0,370	0,262	0,483	0,340
Moyenne des 3 dernières couches cachées						
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,270	0,533	0,358	0,261	0,481	0,338
all-mpnet-base-v1	0,272	0,555	0,365	0,262	0,478	0,338
gtr-t5-large	0,277	0,554	0,370	0,258	0,484	0,337
gtr-t5-xl	0,278	0,556	0,371	0,265	0,486	0,343
sentence-t5-xl	0,282	0,554	0,373	0,268	0,489	0,346
bert-base	0,283	0,542	0,372	0,259	0,483	0,337
legal-bert-base	0,285	0,556	0,377	0,267	0,488	0,345
Moyenne des 4 dernières couches cachées						
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,296	0,534	0,381	0,280	0,461	0,349
all-mpnet-base-v1	0,303	0,547	0,390	0,283	0,460	0,350
gtr-t5-large	0,306	0,551	0,393	0,284	0,466	0,353
gtr-t5-xl	0,305	0,552	0,393	0,291	0,466	0,358
sentence-t5-xl	0,311	0,563	0,401	0,294	0,469	0,361
bert-base	0,307	0,553	0,395	0,263	**0,494**	0,343
legal-bert-base	**0,312**	**0,575**	**0,405**	**0,296**	0,478	**0,366**

4.2.3 Stratégie de pooling par moyenne de tous les tokens VS stratégie de pooling basée sur le token CLS :

Les tableaux 6 et 8 comparent les performances des PLMs précédemment évalués sur les mêmes jeux de données juridiques, en utilisant deux stratégies de pooling : Le Pooling par moyenne sur l'ensemble des tokens et le Pooling basé sur le token CLS, pour le classement et la classification des affaires juridiques.

Dans la recherche d'affaires (Tableau 6), le pooling moyenne sur l'ensemble des tokens obtient systématiquement de meilleurs F1 scores que le pooling basé que sur CLS token, notamment sur COLIEE 2022 où Legal-BERT atteint un F1score@5 de 0,433 et un F1score@10 de 0,343 contre 0,417 et 0,333 avec le pooling CLS. Cette tendance se retrouve sur AILA 2019 et IRLeD 2017, suggérant que le pooling moyen sur l'ensemble des tokens capture mieux les caractéristiques essentielles à la sélection des affaires. Toutefois, un test de significativité sur legal-bert-base ne révèle pas de différence statistiquement significative entre les deux méthodes.

Le tableau 7 compare nos résultats à l'état de l'art. Pour COLIEE 2022, nos performances surpassent celles de (Askari *et al.*, 2022), confirmant l'impact du pooling moyenne sur tous les tokens combiné à l'agrégation des quatre dernières couches cachées, notamment en F1score@5. Sur AILA 2019, nos résultats dépassent ceux de (Zhao *et al.*, 2019), qui rapportent un MAP score légèrement inférieur de 0,149 et un P@10 de 0,070. En revanche, sur IRLeD 2017, nos performances restent inférieures à celles rapportées par (Padigi *et al.*, 2019), qui obtiennent un MAP nettement plus élevé de 0,471 et un P@10 de 0,260.

TABLE 6 – Résultats obtenus avec différentes stratégies de pooling pour la sélection des affaires juridiques. Les tests t appariés significatifs comparant les stratégies de Pooling CLS et de Moyenne de tous les tokens pour la sélection des affaires juridiques sont effectués sur les modèles legal-bert-base et bert-base, où $p \prec 0.05$ et $p \succ 0.05$ sont indiqués respectivement avec ‡ et †

Moyenne des 4 dernières couches cachées basée sur le Token CLS						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,254	0,230	0,182	0,070	0,293	0,168
all-mpnet-base-v1	0,263	0,221	0,178	0,064	0,299	0,175
gtr-t5-large	0,276	0,230	0,205	0,072	0,336	0,199
gtr-t5-xl	0,244	0,295	0,191	0,078	0,336	0,199
sentence-t5-xl	0,316	0,257	0,191	0,076	0,332	0,189
bert-base	0,384‡	0,308‡	0,219†	0,073†	0,312‡	0,182‡
legal-bert-base	0,417†	0,333†	0,225†	**0,092**†	0,356†	**0,219**†
Moyenne des 4 dernières couches cachées basée sur la Moyenne de tous les tokens						
	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
all-roberta-large-v1	0,394	0,315	0,220	0,080	0,307	0,176
all-mpnet-base-v1	0,396	0,308	0,216	0,074	0,314	0,182
gtr-t5-large	0,398	0,316	0,223	0,080	0,350	0,206
gtr-t5-xl	0,417	0,339	0,225	0,084	0,356	0,210
sentence-t5-xl	0,428	0,338	0,219	0,074	0,346	0,196
bert-base	0,396‡	0,318‡	0,222†	0,079†	0,341‡	0,206‡
legal-bert-base	**0,433**†	**0,343**†	**0,231**†	0,089†	**0,362**†	0,216†

Dans la classification des affaires juridiques (Tableau 8), le pooling moyenne sur l'ensemble des tokens dépasse légèrement le pooling basé que sur CLS token en F1 score, aussi bien avec Gradient Boosting qu'avec les Réseaux de Neurones. Par exemple, avec un Réseau de Neurones, Legal-BERT

TABLE 7 – Comparaison avec les résultats de l'état de l'art pour la sélection des affaires juridiques

	COLIEE 2022		AILA 2019		IRLeD 2017	
PLMs	F1-score@5	F1-score@10	MAP	P@10	MAP	P@10
legal-bert-base$_{cls-Token}$	0,417	0,333	0,225	**0,092**	0,356	0,219
legal-bert-base$_{all-Tokens}$	**0,433**	**0,343**	**0,231**	0,089	0,362	0,216
(Askari *et al.*, 2022)	0,326	-	-	-	-	-
(Zhao *et al.*, 2019)	-	-	0,149	0,070	-	-
(Padigi *et al.*, 2019)	-	-	-	-	0,471	0,260

TABLE 8 – Résultats obtenus avec différentes stratégies de pooling pour la classification des affaires juridiques en utilisant les classificateurs Gradient Boosting et Réseau de Neurones sur le jeu de données COLIEE 2022. Les tests t appariés significatifs comparant les stratégies de Pooling CLS et de Moyenne de tous les tokens pour la classification des affaires juridiques sont effectués sur les modèles legal-bert-base et bert-base, où $p \prec 0.05$ et $p \succ 0.05$ sont indiqués respectivement avec ‡ et †

	Moyenne des 4 dernières couches cachées basée sur le Pooling CLS					
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,277	0,556	0,370	0,267	0,483	0,344
all-mpnet-base-v1	0,288	0,565	0,382	0,268	0,481	0,344
gtr-t5-large	0,289	0,567	0,383	0,263	0,492	0,343
gtr-t5-xl	0,286	0,566	0,380	0,275	0,493	0,353
sentence-t5-xl	0,292	0,571	0,386	0,279	**0,495**	0,357
bert-base	0,289†	0,569†	0,383†	0,262†	0,489†	0,341†
legal-bert-base	0,299†	**0,586**†	0,396†	0,289†	0,494†	0,364†
	Moyenne des 4 dernières couches cachées basée sur la Moyenne de tous les tokens					
	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
all-roberta-large-v1	0,296	0,534	0,381	0,280	0,461	0,349
all-mpnet-base-v1	0,303	0,547	0,390	0,283	0,460	0,350
gtr-t5-large	0,306	0,551	0,393	0,284	0,466	0,353
gtr-t5-xl	0,305	0,552	0,393	0,291	0,466	0,358
sentence-t5-xl	0,311	0,563	0,401	0,294	0,469	0,361
bert-base	0,307†	0,553†	0,395†	0,263†	0,494†	0,343†
legal-bert-base	**0,312**†	0,575†	**0,405**†	**0.296**†	0,478†	**0,366**†

TABLE 9 – Comparaison avec les résultats SOTA pour la tâche de classification des affaires juridiques sur le jeu de données COLIEE 2022

	Gradient Boosting			Réseau de Neurones		
PLMs	Précision	Rappel	Score F1	Précision	Rappel	Score F1
legal-bert-base$_{cls-Token}$	0,299	**0,586**	0,396	0,289	0,494	0,364
legal-bert-base$_{all-Tokens}$	0,312	0,575	**0,405**	0,296	0,478	0,366
(Rabelo *et al.*, 2022)	0,411	0,339	0,372	-	-	-

atteint un F1 de 0,366 en pooling de tous les tokens contre 0,364 en pooling que du CLS. Toutefois, le test de significativité sur legal-bert-base dans COLIEE 2022 ne montre aucune différence statistique notable entre les deux méthodes.

Bien que les performances des deux classificateurs soient similaires sous les deux stratégies de pooling, le pooling de tous les tokens affiche systématiquement une légère supériorité, suggérant son efficacité pour la classification des affaires juridiques sur différents PLMs. Toutefois, bien que legal-bert-base présente de solides résultats sur toutes les métriques, il ne dépasse pas l'état de l'art rapporté par (Rabelo *et al.*, 2022) (Tableau 9) en précision, mais obtient des scores de rappel et F1

plus élevés avec Gradient Boosting. Cela indique des pistes d'amélioration, notamment en explorant d'autres techniques d'optimisation.

Globalement, le pooling de tokens surpasse légèrement le pooling de CLS pour la recherche et la classification des affaires juridiques, probablement grâce à une meilleure capture du contexte. Ainsi, Legal-BERT se démarque particulièrement sous cette stratégie, obtenant systématiquement les meilleures performances.

4.3 Récapitulatif des résultats

L'analyse expérimentale porte sur trois aspects clés : l'impact de la fenêtre glissante sur la recherche d'affaires juridiques, l'effet de l'agrégation des couches cachées des PLMs sur la recherche et la classification, ainsi que la comparaison entre l'agrégation par moyenne sur l'ensemble des tokens et le l'agrégation par moyenne sur l'ensemble des CLS token. Cette étude offre des perspectives intéressantes sur l'optimisation des PLMs pour ces tâches.

D'abord, la fenêtre glissante, testée avec différentes tailles (WS) et longueurs de pas (S), améliore la sélection des affaires juridiques en offrant un contexte plus large. Les résultats montrent que des fenêtres plus grandes améliorent systématiquement les performances. Par exemple, legal-bert-base atteint son meilleur score sur COLIEE 2022 avec WS = 512 et S = 256, une tendance observée sur tous les jeux de données. Bien que la longueur du pas influence aussi les résultats, son impact reste moindre par rapport à la taille de la fenêtre. Cette configuration a donc été retenue pour les expérimentations suivantes.

Ensuite, l'agrégation des couches cachées améliore les performances par rapport à l'utilisation de la dernière couche seule. Cette tendance se confirme sur toutes les tâches, là où Legal-BERT obtenant le plus grand gain lorsqu'il combine les quatre dernières couches cachées, suggérant que les couches profondes enrichissent la compréhension des documents juridiques.

Enfin, la comparaison entre le pooling par moyenne sur l'ensemble des tokens et celui basé uniquement sur le token CLS montre que la première approche surpasse systématiquement la seconde. Elle offre une représentation plus riche du texte et se révèle mieux adaptée aux tâches juridiques.

Pour résumer, ces résultats soulignent l'importance de choisir des paramètres et stratégies optimaux pour maximiser l'efficacité des PLMs dans la recherche et la classification des documents juridiques.

5 Conclusion

Cet article a présenté une stratégie visant à améliorer la représentation des embeddings des documents juridiques. À travers des expérimentations menées sur trois collections de jeux de données juridiques publiques, nous avons démontré l'efficacité de notre approche sur deux tâches différentes. En exploitant l'intégralité du texte grâce à la technique de fenêtre glissante et en combinant les sorties de différentes couches cachées intermédiaires des PLMs, nous avons significativement amélioré la qualité des embeddings des documents juridiques. Nos résultats mettent en évidence le potentiel de notre méthode pour renforcer les performances des tâches LIR. En particulier, notre approche a montré des résultats prometteurs sur les collections COLIEE 2022 et AILA 2019, offrant ainsi une solution robuste pour exploiter des ressources limitées dans le traitement des textes juridiques.

Dans la continuité de ce travail, nous envisageons d'intégrer des méthodes basées sur la génération augmentée par récupération d'informations afin d'exploiter simultanément les avantages des modèles génératifs et des systèmes de recherche d'information. Plus précisément, nous prévoyons d'utiliser les Grands Modèles de Langue génératifs pour améliorer la reformulation des requêtes, générer des résumés explicatifs des documents sélectionnés et affiner la phase de reclassement en tenant compte du contexte juridique global.

Références

ASKARI A., PEIKOS G., PASI G. & VERBERNE S. (2022). Leibi@ coliee 2022 : aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. *arXiv :2205.13351*.

BHATTACHARYA P., GHOSH K., GHOSH S., PAL A., MEHTA P., BHATTACHARYA A. & MAJUMDER P. (2019). Fire 2019 aila track : Artificial intelligence for legal assistance. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, p. 4–6.

BUI Q. M., NGUYEN C., DO D.-T., LE N.-K., NGUYEN D.-H., NGUYEN T.-T.-T., NGUYEN M.-P. & NGUYEN M. L. (2022). Jnlp team : Deep learning approaches for tackling long and ambiguous legal documents in coliee 2022. In *JSAI International Symposium on Artificial Intelligence*, p. 68–83 : Springer.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). Legal-bert : The muppets straight out of law school. *arXiv preprint arXiv :2010.02559*.

CLAVEAU V. (2021). Neural text generation for query expansion in information retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, p. 202–209.

HUANG J., TANG D., ZHONG W., LU S., SHOU L., GONG M., JIANG D. & DUAN N. (2021). Whiteningbert : An easy unsupervised sentence embedding approach. *arXiv preprint arXiv :2104.01767*.

KIM M.-Y., RABELO J., GOEBEL R., YOSHIOKA M., KANO Y. & SATOH K. (2022). Coliee 2022 summary : Methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, p. 51–67 : Springer.

LI H., AI Q., CHEN J., DONG Q., WU Y., LIU Y., CHEN C. & TIAN Q. (2023). Sailer : structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1035–1044.

LICARI D. & COMANDÈ G. (2022). Italian-legal-bert : A pre-trained transformer language model for italian law. In *EKAW (Companion)*.

LIMSOPATHAM N. (2021). Effectively leveraging bert for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, p. 210–216.

MANDAL A., GHOSH K., BHATTACHARYA A., PAL A. & GHOSH S. (2017). Overview of the fire 2017 irled track : Information retrieval from legal documents. In *FIRE (Working Notes)*, p. 63–68.

MOKANOV I., SHANE D. & CERAT B. (2019). Facts2law : using deep learning to provide a legal qualification to a set of facts. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, p. 268–269.

NGUYEN H.-T., NGUYEN M.-P., VUONG T.-H.-Y., BUI M.-Q., NGUYEN M.-C., DANG T.-B., TRAN V., NGUYEN L.-M. & SATOH K. (2022). Transformer-based approaches for legal text processing : Jnlp team-coliee 2021. *The Review of Socionetwork Strategies*, **16**(1), 135–155.

NGUYEN H.-T., VUONG H.-Y. T., NGUYEN P. M., DANG B. T., BUI Q. M., VU S. T., NGUYEN C. M., TRAN V., SATOH K. & NGUYEN M. L. (2020). Jnlp team : Deep learning for legal processing in coliee 2020. *arXiv preprint arXiv :2011.08071*.

NIGAM S. K., GOEL N. & BHATTACHARYA A. (2022). nigam@ coliee-22 : Legal case retrieval and entailment using cascading of lexical and semantic-based models. In *JSAI International Symposium on Artificial Intelligence*, p. 96–108 : Springer.

PADIGI S. V., MAYANK M. & NATARAJAN S. (2019). Precedent case retrieval using wordnet and deep recurrent neural networks. In *CS & IT Conference Proceedings*.

RABELO J., KIM M.-Y. & GOEBEL R. (2022). Semantic-based classification of relevant case law. In *JSAI International Symposium on Artificial Intelligence*, p. 84–95 : Springer.

SHAGHAGHIAN S., FENG L. Y., JAFARPOUR B. & POGREBNYAKOV N. (2020). Customizing contextualized language models for legal document reviews. In *2020 IEEE international conference on big data (Big Data)*, p. 2139–2148 : IEEE.

SHAO Y., MAO J., LIU Y., MA W., SATOH K., ZHANG M. & MA S. (2020). Bert-pli : Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, p. 3501–3507.

UDAGAWA T., SUZUKI M., KURATA G., MURAOKA M. & SAON G. (2024). Multiple representation transfer from large language models to end-to-end asr systems. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 10176–10180 : IEEE.

VAN AKEN B., WINTER B., LÖSER A. & GERS F. A. (2019). How does bert answer questions ? a layer-wise analysis of transformer representations. CIKM '19, p. 1823–1832, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3357384.3358028.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VOLD A. & CONRAD J. G. (2021). Using transformers to improve answer retrieval for legal questions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, p. 245–249.

XIAO C., HU X., LIU Z., TU C. & SUN M. (2021). Lawformer : A pre-trained language model for chinese legal long documents. *AI Open*, **2**, 79–84.

YANG J. & ZHAO H. (2019). Deepening hidden representations from pre-trained language models. *arXiv :1911.01940*.

ZHAO Z., NING H., LIU L., HUANG C., LEI KONG L., HAN Y. & YUAN HAN Z. (2019). Fire2019@aila : Legal information retrieval using improved bm25. In *FIRE*.