Extraction terminologique juridique à faible supervision : une méthode hybride combinant LLM, règles syntaxiques et CamemBERT

Julien Breton[1,2], Mokhtar Boumedyen Billami[2], Max Chevalier[1], Cassia Trojahn[1]

[1] Institut de Recherche en Informatique de Toulouse, IRIT
[2] Berger-Levrault

julien.breton@irit.fr, mb.billami@berger-levrault.com,
max.chevalier@irit.fr, cassia.trojahn@irit.fr

Résumé

Le secteur juridique se caractérise par un nombre important de documents et par leur complexité. Les entreprises ont l'obligation d'appliquer ces dispositions juridiques. En raison de l'évolution constante de ces documents, un intérêt croissant se manifeste pour l'automatisation du traitement des textes juridiques afin de faciliter la conformité réglementaire. Une étape clé de ce processus réside dans l'extraction des termes juridiques. Les méthodes état de l'art, telles que les systèmes à base de règles, les réseaux Bi-LSTM et BERT, requièrent une quantité importante de données annotées pour atteindre des performances satisfaisantes, une tâche particulièrement chronophage pour les experts du domaine. Avec l'essor des grands modèles de langage (LLM), la recherche s'oriente de plus en plus vers l'exploitation de leurs capacités, notamment à travers des approches faiblement supervisées. Dans cet article, nous présentons un système hybride qui distille les connaissances de GPT-4 vers un modèle CamemBERT, tout en appliquant un filtrage syntaxique. Cette approche réduit non seulement le besoin d'intervention d'experts par rapport au système CamemBERT classique, mais elle surpasse également le système reposant uniquement sur GPT-4, en améliorant le score F1 de 7 à 24 points de pourcentage.

Mots-clés

Extraction terminologique juridique, Faible supervision, CamemBERT, Grands modèles de langage (LLM), Distillation des connaissances

1 Introduction

Le domaine juridique se caractérise par un volume considérable de documents en constante évolution, tels que les contrats, la législation, les décisions de justice ou encore les décrets. Ces documents sont denses, complexes et rédigés dans un langage hautement spécialisé, ce qui rend leur analyse et application à la fois chronophages et susceptibles aux erreurs humaines. Cependant, les entreprises ont l'obligation légale de se mettre en conformité avec ces dispositions juridiques, sous peine d'amendes. Comme le soulignent Sassier et al. [23], en France, « plus de 10 500 lois, 120 000 décrets, 7 400 traités, 17 000 textes communautaires, des dizaines de milliers de pages réparties dans 62 codes distincts » sont en vigueur. Certains de ces textes font d'ailleurs l'objet de modifications fréquentes : « 6 modifications par jour ouvrable pour le Code des impôts de 2006 ». C'est dans ce sens que la recherche vise à automatiser le traitement des documents juridiques. Elle souhaite non seulement accélérer leur analyse, tout en délestant les experts juridiques de cette tâche chronophage et à faible valeur ajoutée.

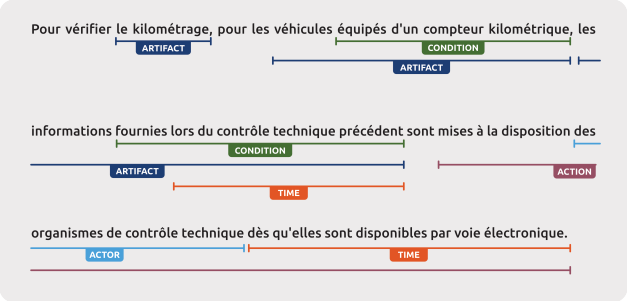

FIGURE 1 – Extraction des termes juridiques à partir de la phrase suivante : « Pour vérifier le kilométrage, pour les véhicules équipés d'un compteur kilométrique, les informations fournies lors du contrôle technique précédent sont mises à la disposition des organismes de contrôle technique dès qu'elles sont disponibles par voie électronique ».

La tâche fondamentale pour l'extraction des règles juridiques consiste à les formaliser de manière structurée. Deux tâches entrent alors en jeu : l'extraction des termes juridiques et l'extraction des relations entre ces termes. Le présent article se concentre sur l'extraction terminologique, comme l'illustre la Figure 1, qui provient d'un exemple issu du jeu de données utilisé dans cette étude.
Le jeu de données exploité dans le cadre de cette recherche a été introduit par Sleimi et al. [26], qui se sont appuyés sur des documents juridiques du Luxembourg. Dans leur étude, les auteurs ont développé un système fondé sur des

règles syntaxiques pour l'extraction de termes juridiques. Ils ont obtenu une précision notable de 0,874 et un rappel de 0,855. Cependant, atteindre de telles performances requiert un investissement conséquent en temps de la part d'experts pour annoter les données et concevoir les règles syntaxiques. D'autres méthodes, telles que les LSTM (Long Short-Term Memory) [24] ou BERT (Bidirectional Encoder Representations from Transformers) [17], rencontrent les mêmes contraintes liées aux données annotées. Néanmoins, l'émergence des grands modèles de langage (LLMs) [36] ouvre de nouvelles perspectives en réduisant l'implication humaine requise pour l'extraction. En tirant parti de leur connaissance fondationnelle, les LLMs sont par exemple capables de réaliser de l'extraction d'entités nommées dans des documents en biologie [29].

Cet article propose une approche hybride combinant LLM, méthodes fondées sur des règles syntaxiques et modèle de langage. La distillation des connaissances fournies par un LLM vers un modèle CamemBERT, avec un filtrage basé sur des règles, répond à plusieurs objectifs. Premièrement, cette approche limite l'implication des experts à la formulation des instructions pour le LLM et à la définition des règles de filtrage. Deuxièmement, elle diminue les besoins en ressources de calcul et améliore l'interprétabilité des résultats, en offrant une meilleure compréhension du processus d'extraction ainsi que du jeu de données utilisé pour l'entraînement.

Les travaux présentés dans cet article sont l'adaptation en version française de l'article [4] paru lors de la conférence EKAW 2024.

Le reste de l'article est organisé comme suit : la section 2 présente les principaux travaux connexes ; la Section 3 décrit les modèles de référence utilisés (CamemBERT et GPT-4) ; la Section 4 introduit le système hybride proposé ; la Section 5 détaille le jeu de données utilisé ainsi que les résultats obtenus selon les différentes stratégies ; enfin, la Section 6 conclut l'article et propose des perspectives de recherche futures.

2 Travaux connexes

Le traitement des connaissances constitue un domaine vaste ayant suscité de nombreuses contributions de la part de la communauté scientifique. Récemment, une tendance marquée s'est manifestée en faveur de l'utilisation des réseaux de neurones pour des tâches d'analyse [28] ou de classification [6, 22]. Par exemple, Biener et al. [2] ont développé un système d'anonymisation d'entités nommées (Named Entity Recognition, NER [18]) dans des documents financiers rédigés en allemand. Leur système identifie des entités telles que les noms et prénoms, les adresses postales et électroniques, ainsi que les localisations. L'étude a évalué diverses architectures, notamment les RNN, LSTM et Conditional Random Fields (CRF) [15]. Les résultats démontrent que les architectures combinant RNN et CRF obtiennent les meilleures performances, avec un rappel de plus de 97 % sans post-traitement et environ 99 % après post-traitement, tout en maintenant une précision supérieure à 90 %. Malgré l'efficacité de cette approche, elle présente certaines limites en raison de la forte implication des experts et du caractère chronophage des tâches. L'annotation manuelle d'un corpus de 407 documents financiers allemands publiés, représentant un total de 189 000 tokens, constitue un travail particulièrement lourd et coûteux en temps.

Le modèle BERT, proposé par Vaswani et al. [30], vise à atténuer ce problème grâce à un entraînement initial sur un large corpus de données, comprenant environ 3,3 millions de mots issus du Toronto BookCorpus et de Wikipédia en anglais. Cet entraînement préliminaire confère au modèle une base de connaissances génériques qui peut ensuite être affinée sur des jeux de données de taille plus réduite. La communauté juridique a adopté cette architecture pour la reconnaissance d'entités nommées, comme en témoignent les travaux menés sur des décisions judiciaires italiennes [21] ou des textes juridiques brésiliens [32]. Ces approches associent BERT à des architectures de type Bi-LSTM et CRF. Toutefois, malgré leurs performances, ces méthodes nécessitent toujours un volume conséquent de données annotées et un investissement important d'expertise humaine.

Avec l'essor récent des grands modèles de langage (LLMs), de nombreuses études ont exploré leur potentiel pour l'extraction d'informations, obtenant des résultats significatifs [9, 33, 7, 1]. Des avancées notables ont aussi été accomplies dans la distillation des performances de ces modèles vers des modèles de plus petite taille [12]. Yuxian et al. [11] ont démontré que la distillation constitue une méthode efficace pour transférer les connaissances d'un grand modèle génératif vers un modèle plus compact. Ce processus favorise l'obtention de réponses plus précises ainsi qu'une amélioration des performances. Cependant, des travaux constatent des limites dans l'utilisation des LLM pour l'extraction d'informations, que ce soit dans les capacités de calcul [13, 31] ou même dans les résultats obtenus [14]. [10] montre que les LLM ne parviennent pas à surpasser les modèles traditionnels (BERT) dans les tâches de NER, soulignant les limites des LLM dans l'extraction d'entités complexes spécifiques à un domaine. Xie et al. [34] examinent les performances des LLM dans les tâches de NER non supervisées, notant que bien que des améliorations puissent être réalisées, des défis subsistent pour atteindre un haut niveau de précision sans recourir à des stratégies personnalisées.

Dans le domaine juridique, comme le souligne l'étude de Solihin et al. [27], les travaux se sont majoritairement concentrés sur la NER, ce qui a donné lieu à une recherche relativement limitée et à peu de jeux de données spécifiquement consacrés à l'extraction de termes juridiques. Bien que ces deux tâches puissent sembler similaires a priori, elles s'en distinguent considérablement. Les entités nommées, telles que les personnes ou les organisations, sont généralement représentées par une unité courte et bien délimitée, comme *« Stephen Hawking »*. À l'inverse, les entités juridiques englobent des concepts plus larges, des entités nommées ou même des expressions complètes. Par exemple, le concept juridique d'« Acteur » peut corres-

pondre à une entité comme *« le conducteur »*, qui n'est pas une entité nommée. De même, la notion de « Condition » peut inclure un contenu tel que *« si l'ensemble couplé de véhicules se compose de deux véhicules automoteurs »*. Ces exemples illustrent que notre tâche dépasse largement le cadre NER traditionnel et peut s'apparenter à l'extraction de segments textuels. Il est donc légitime de remettre en question l'efficacité directe des approches conçues pour la NER [37, 35, 19, 20] dans l'extraction de termes juridiques.

Concernant l'extraction de termes juridiques, certaines études, telles que celle de Sleimi et al. [25], ont développé leurs propres jeux de données et mis en œuvre une approche fondée sur des règles syntaxiques. Des travaux plus récents, tels que ceux de Castano et al. [5], explorent une démarche similaire en se focalisant sur l'extraction conjointe de concepts et d'entités à partir de documents juridiques européens. Les éléments extraits sont ensuite intégrés dans un système de gestion des connaissances. D'autres contributions, comme celle de Dragoni et al. [8], montrent la pertinence et l'efficacité d'une approche hybride pour l'extraction de règles juridiques à partir de documents textuels. Ces résultats suggèrent que la combinaison de plusieurs techniques peut considérablement améliorer la précision et l'efficacité de l'extraction, tout en compensant leurs défauts mutuels.

Basé sur ces observations, notre étude a pour objectif de comparer notre système hybride avec les modèles de référence, à savoir CamemBERT et GPT-4, qui seront présentés dans la section suivante.

3 Systèmes de référence

Dans le but de comparer notre approche aux systèmes actuels de l'état de l'art, nous introduisons les modèles de langage pour la tâche d'extraction de termes juridiques. Les sections suivantes décrivent leur mise en œuvre, tout en mettant en lumière les avantages et les limites propres à chacune de ces architectures.

3.1 Extraction terminologique juridique à l'aide de CamemBERT

En raison de la langue française du jeu de données utilisé (décrit en Section 5.1), nous avons opté pour un modèle CamemBERT, au lieu du modèle BERT original. Le modèle Legal-CamemBERT-base [16], spécifiquement réentraîné sur plus de 22 000 articles juridiques du droit belge en français, se révèle plus adapté au traitement et à la compréhension des textes juridiques en langue française. Ce choix vise à mieux capter les subtilités et les spécificités linguistiques présentes dans notre corpus, améliorant ainsi la représentation des plongements lexicaux.

L'extraction de termes juridiques avec CamemBERT requiert un corpus annoté manuellement par des experts. Ce jeu de données doit être divisé en deux parties : l'une pour l'entraînement et la seconde pour l'évaluation, comme indiqué à la Figure 2. L'extraction terminologique au moyen de CamemBERT repose traditionnellement sur le schéma d'annotation Inside-Outside-Beginning (IOB), courramment utilisé en traitement automatique des langues pour l'étiquetage de séquences, notamment dans les tâches de NER. Toutefois, dans le cas de l'extraction de termes juridiques, des chevauchements d'annotations sont fréquemment observés, ce qui nécessite une adaptation de l'architecture interne du modèle. Par exemple, dans la Figure 1, nous constatons que le segment « équipés d'un compteur kilométrique » est à la fois catégorisé comme `Artifact` et comme `Condition`.

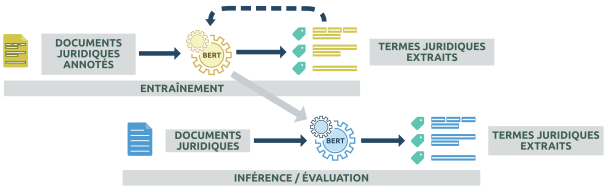

FIGURE 2 – Processus global basé sur CamemBERT : (i) réentraînement du modèle CamemBERT, (ii) évaluation de l'inférance du modèle réentraîné.

FIGURE 3 – Matrice de tokenisation basée sur la phrase courte : « les véhicules équipés d'un compteur kilométrique ». Les mots activent les concepts juridiques tels que `Condition` ou `Artifact` à l'aide de valeurs binaires (0 ou 1).

Une première modification porte donc sur la matrice d'entrée exploitée dans le modèle PyTorch, pour permettre une classification multi-label et multi-classe. La Figure 3 illustre cette modification, qui permet l'activation de plusieurs concepts juridiques pour un même token, comme pour le mot « compteur ». La seconde modification concerne la mesure utilisée lors de l'entraînement ; nous avons opté pour un macro score F1, c'est-à-dire une moyenne des scores F1 calculés pour chaque concept individuellement. Enfin, une dernière adaptation consiste à modifier le classifieur de sortie à l'aide de la bibliothèque Transformers de Huggingface. Cette modification substitue la fonction de perte CrossEntropyLoss [1] par la fonction BCEWithLogitsLoss [2], qui combine l'entropie croisée binaire à une fonction sigmoïde, convenant aux tâches multi-label. Cette architecture est schématisée par la Figure 4 et les expériences décrites dans cet article sont disponibles dans notre dépôt GitLab [3].

Il est important de mentionner que le réentraînement sur les 22 000 articles, réalisé pour créer le modèle Legal-CamemBERT-base [16], diffère des réentraînements réalisés dans notre étude. En effet, les articles du droit belge ont eu pour objectif d'améliorer la représentation sémantique,

1. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html
2. https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html
3. https://gitlab.irit.fr/ala/legal-concepts-extraction

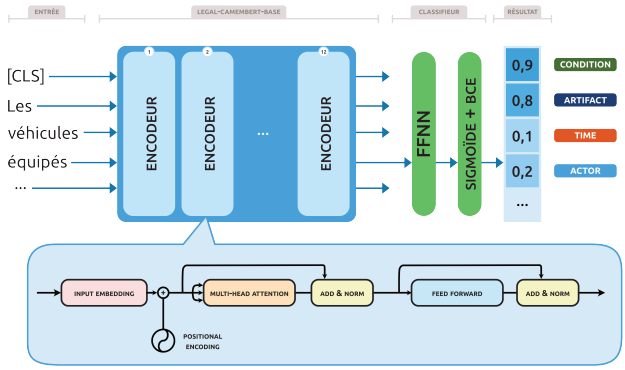

FIGURE 4 – Architecture du modèle CamemBERT utilisé dans le cadre de notre étude. Les modifications concernent le classifieur en sortie du modèle CamemBERT, incluant la fonction de perte.

tandis que nos réentraînements effectués via le corpus de Sleimi et al. [26] visent le classifieur de sortie, comme illustré dans la Figure 4.

L'entraînement du classifieur en sortie de CamemBERT nécessite un jeu de données conséquent dont l'élaboration représente une tâche fastidieuse pour les experts. L'apparition récente de modèles génératifs, tels que GPT-4, a ouvert de nouvelles voies de recherche autour de l'extraction de connaissances faiblement supervisée. Ces modèles sont en mesure d'effectuer des tâches d'extraction sans recourir à un grand volume de données annotées. Cette approche est détaillée dans la section suivante.

3.2 Extraction terminologique juridique fondée sur un LLM

La Figure 5 illustre le processus global d'utilisation d'un LLM pour l'extraction de termes à partir de textes juridiques. En employant l'ingénierie des requêtes (*prompt engineering*), nous utilisons le LLM pour une tâche d'extraction terminologique. L'un des principaux atouts de cette approche repose sur l'entraînement fondationnel du modèle, qui lui permet d'obtenir de bonnes performances à partir d'instructions minimales.

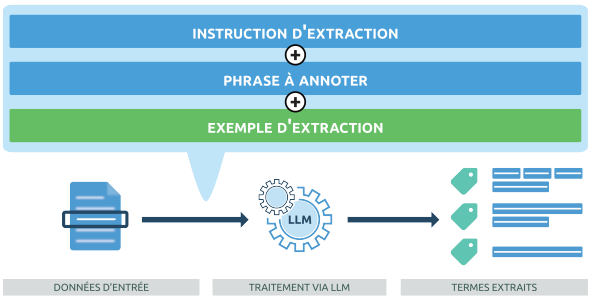

FIGURE 5 – Processus d'extraction terminologique employant un LLM via l'ingénierie des requêtes.

La structure de la requête utilisée a émergé d'un processus empirique et d'expérimentations itératives visant à identifier la formulation la plus robuste. Ce travail s'appuie sur les recommandations et bonnes pratiques proposées par OpenAI [4]. Des ajustements successifs ont été opérés sur les différentes composantes, incluant la définition du rôle du modèle, la description des tâches et le format de sortie attendu. Les requêtes, disponibles dans notre dépôt GitLab [5], sont organisées comme suit : en premier lieu, un rôle est assigné au modèle (par exemple « expert en TAL »), conjointement à la tâche visée (« extraire des termes à partir d'énoncés »). Ensuite, sont présentés les différents concepts juridiques accompagnés de leurs définitions. Ces définitions constituent la seule connaissance externe introduite, nécessitant une contribution minimale des experts. À cela s'ajoute un exemple d'extraction avec la sortie attendue (en JSON), fournissant ainsi un aperçu explicite de la tâche à effectuer. La Figure 5 illustre la structure des différentes composantes au sein de la requête fournie en entrée du LLM.

Après avoir présenté l'application des LLMs à l'extraction d'information par ingénierie de requête, nous décrivons à présent notre approche hybride combinant un LLM, un filtrage syntaxique et un CamemBERT.

4 Approche hybride

Comme introduit précédemment, les modèles de type BERT offrent une efficacité notable mais requièrent un jeu de données conséquent. À l'inverse, les LLMs, du fait de leur pré-entraînement, ne dépendent pas d'un corpus spécifique, mais peuvent souffrir de précision dans la délimitation des termes. Pour surmonter les limites de ces deux approches, nous avons développé une architecture hybride, illustrée par la Figure 6. Cette approche combine un LLM afin d'amorcer l'extraction, suivie de règles syntaxiques pour filtrer les hallucinations du LLM, et se termine par un modèle CamemBERT permettant d'apprendre et de compresser cette connaissance.

La première étape de notre chaîne de traitement consiste donc à amorcer une extraction à l'aide de GPT-4, identique au processus décrit en Section 3.2. Les experts élaborent une requête contenant les définitions des concepts juridiques, un exemple commenté, et l'énoncé cible à analyser. Le LLM génère ensuite la prédiction correspondante. Une fois ce corpus synthétique généré, les termes extraits par le LLM sont filtrés via des règles syntaxiques. L'objectif consiste à améliorer la précision des annotations, en s'appuyant sur 15 règles définies par des experts [25]. Par exemple, dans la phrase illustrée à la Figure 1, les concepts de type `Artifact` peuvent être associés à des groupes nominaux, tandis que `Action` relèvent de groupes verbaux. Cette étape permet de traiter la problématique des limites (*boundary issue*) et de garantir la qualité des termes annotés.

Après avoir filtré les termes juridiques, ce corpus est utilisé afin de réentraîner le classifieur CamemBERT, selon

4. https://platform.openai.com/docs/guides/prompt-engineering
5. https://gitlab.irit.fr/ala/legal-entity-extraction/-/raw/main/modules/llm/utils.py

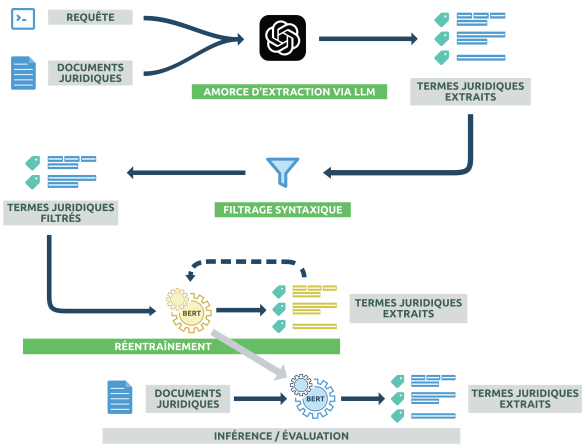

FIGURE 6 – Processus global de l'approche hybride : (i) amorçage de l'extraction terminologique via GPT-4, (ii) filtrage des résultats à l'aide de règles syntaxiques, (iii) réentraînement d'un CamemBERT.

les modalités présentées en Section 3.1. Le modèle est modifié par l'ajout d'une couche linéaire finale et l'ajustement de la fonction de perte adaptée au contexte multi-label. À l'issue de l'entraînement, le modèle est prêt à effectuer des extractions sur de nouveaux documents.

En distillant les connaissances d'un LLM dans un modèle CamemBERT, cette approche permet d'exploiter les atouts complémentaires des deux paradigmes, aboutissant à un système robuste et performant pour l'extraction de termes juridiques. La section suivante détaille les résultats obtenus et les compare avec les systèmes de référence.

5 Expérimentations et résultats

Après avoir présenté l'architecture hybride, nous nous intéressons à présent à l'évaluation de son efficacité et la comparons avec des systèmes de référence. Nous commençons par décrire le jeu de données utilisé dans l'ensemble de nos expérimentations, puis nous évaluons les performances du modèle CamemBERT seul, suivi du LLM avec GPT-4, et terminons avec les résultats du modèle hybride.

5.1 Jeu de données

Le jeu de données utilisé dans notre étude comprend 200 énoncés en français extraits du Code de la route luxembourgeois, annotés manuellement par des experts, identifiant 1339 segments au total [26]. L'annotation porte sur 14 concepts juridiques, et l'ensemble du corpus est disponible en ligne [6]. Dans le cadre de notre étude, nous nous concentrons sur un sous-ensemble de 8 concepts : `Action`, `Actor`, `Artifact`, `Condition`, `Location`, `Modality`, `Reference` et `Time`. Cette sélection correspond à des travaux antérieurs ayant conduit à la création du modèle sémantique SEMLEG [3], dédié à la formalisation de règles juridiques. Ce choix repose aussi sur le constat que certains concepts étaient sous-représentés dans les données d'origine. Afin de favoriser la réutilisabilité du corpus dans d'autres domaines applicatifs, nous avons choisi de nous limiter à ces huit concepts, définis dans le Tableau 1.

6. https://sites.google.com/view/metax-re2018/

Concept	Définition
Action	Le processus de faire quelque chose.
Actor	Entité qui a la capacité d'agir.
Artifact	Objet matériel ou immatériel impliqué dans une action.
Condition	Une contrainte précisant les propriétés qui doivent être respectées.
Location	Un lieu où une action peut être réalisée.
Modality	Représente la contrainte d'application d'une règle.
Reference	Mention textuelle d'une autre source juridique.
Time	Moment, durée ou occurrence d'une action.

TABLE 1 – Définitions des huit concepts de [26] que nous utilisons dans notre étude.

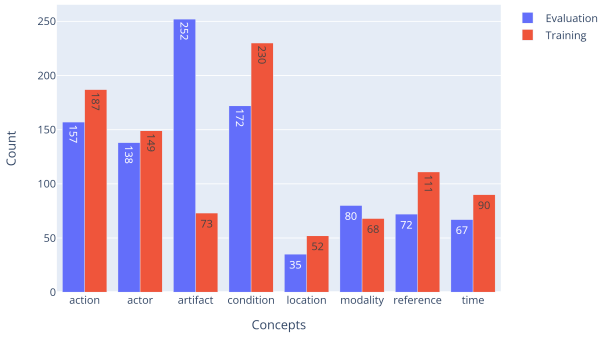

FIGURE 7 – Distribution des concepts juridiques dans les données d'entraînement et d'évaluation [26]. Huit concepts sont conservés : `Action`, `Actor`, `Artifact`, `Condition`, `Location`, `Modality`, `Reference` et `Time`.

La Figure 7 détaille la distribution des concepts retenus. Comme le montre cette distribution, le jeu de données issu de [26] présente un déséquilibre significatif entre les différentes classes. Un cas particulièrement notable concerne le concept `Artifact`, qui ne compte que 73 instances dans l'ensemble d'entraînement contre 252 dans l'ensemble d'évaluation. Ce déséquilibre peut influencer les performances des modèles d'apprentissage, en limitant leur capacité à généraliser sur des classes sous-représentées. Afin de garantir une comparaison avec l'approche des auteurs, nous avons choisi de conserver la distribution originale des annotations sans procéder à un rééquilibrage. Cette décision nous permet non seulement d'évaluer les performances de nos méthodes dans un contexte comparable, mais également d'analyser la robustesse de nos méthodes face à des situations dans lesquelles certaines catégories sont sous-représentées. Ainsi, cette configuration constitue une opportunité d'étudier dans quelle mesure l'approche hybride peut s'adapter aux contraintes inhérentes aux données déséquilibrées.

5.2 Résultats du modèle CamemBERT réentraîné

Les performances du modèle d'extraction terminologique juridique fondé sur Legal-CamemBERT-base [16] sont présentées dans le Tableau 2. Celui-ci fournit les scores de précision, rappel et F1 pour chacun des huit concepts étudiés.

	Concepts juridiques							
	Action	Actor	Artifact	Condition	Location	Modality	Reference	Time
Précision	0.93	0.84	0.69	0.80	0.65	0.91	0.88	0.85
Rappel	0.41	0.60	0.16	0.83	0.53	0.48	0.67	0.66
F1	0.57	0.70	0.26	0.82	0.59	0.63	0.76	0.74

TABLE 2 – Résultats de précision, rappel et F1 obtenus avec Legal-CamemBERT-base [16] pour l'extraction de termes juridiques.

L'approche basée sur CamemBERT obtient un score F1 moyen supérieur à 0.63. La précision moyenne atteint 0.81, ce qui montre que le modèle extrait les termes avec une très bonne précision. Les concepts `Action`, `Actor`, `Condition`, `Modality`, `Reference` et `Time` obtiennent tous une précision égale ou supérieure à 0.80. Toutefois, le rappel moyen est plus faible, autour de 0.54, ce qui révèle une difficulté à détecter l'intégralité des termes présents dans les documents.
Cela est notamment dû au concept `Artifact`, qui affiche des performances significativement inférieures. Bien que la précision soit modérée (0.69), le rappel chute drastiquement à 0.16, entraînant un score F1 très faible de 0.26. Nos investigations montrent que ce résultat s'explique majoritairement par un fort déséquilibre de la distribution des instances dans le jeu de données d'entraînement, comme cela apparaît dans la Figure 7. Alors que la plupart des concepts sont répartis selon un ratio équilibré entre apprentissage et évaluation, le concept `Artifact` présente un déséquilibre marqué, avec seulement 74 occurrences dans l'ensemble d'entraînement contre 252 dans l'ensemble d'évaluation. Ce déséquilibre compromet la capacité du modèle à généraliser efficacement pour ce concept, ce qui explique les difficultés rencontrées. Pour remédier à cette situation, un réajustement des répartitions entre données d'apprentissage et d'évaluation serait nécessaire.
En résumé, le réentraînement du modèle CamemBERT a permis d'atteindre de bonnes performances, avec un score F1 moyen de 0.69 si l'on exclut le concept `Artifact`. Toutefois, les résultats obtenus soulignent clairement la dépendance du modèle à une quantité significative de données annotées, en particulier pour garantir une couverture adéquate. La constitution de tels jeux de données demeure une tâche exigeante en termes de temps et de compétences. Nous évaluons, dans la section suivante, la capacité des LLM à s'affranchir du corpus d'entraînement.

5.3 Résultats avec un LLM

Cette section présente les résultats obtenus à l'aide de l'ingénierie des requêtes et du LLM. Dans notre cas, nous utilisons GPT-4 pour l'extraction des termes juridiques. Les performances en précision, rappel et F1 sont présentées dans le Tableau 3.

	Concepts							
	Action	Actor	Artifact	Condition	Location	Modality	Reference	Time
Précision	0.65	0.67	0.58	0.76	0.53	0.54	0.71	0.81
Rappel	0.27	0.58	0.33	0.53	0.36	0.54	0.53	0.34
F1	0.38	0.63	0.42	0.63	0.42	0.54	0.61	0.48

TABLE 3 – Résultats de précision, rappel et F1 avec GPT-4 pour l'extraction de termes juridiques.

Les résultats obtenus révèlent une forte hétérogénéité selon les concepts. Le meilleur score de précision concerne le concept `Time` (0.81), ce qui illustre la capacité du modèle à identifier correctement les expressions temporelles. Des scores de précision élevés sont également observés pour les concepts `Condition` (0.76) et `Reference` (0.71). Sur le plan du rappel, le concept `Actor` atteint un score notable de 0.58, indiquant une bonne couverture des termes relevant de cette catégorie. En revanche, le concept `Action` affiche le rappel le plus faible (0.27), traduisant des difficultés à identifier de manière exhaustive les actions décrites dans les textes juridiques.
Ces résultats montrent une efficacité en termes de précision, mais soulignent aussi les limitations des LLM en matière de rappel. Le modèle GPT-4 sait identifier les termes de manière exacte lorsqu'il les reconnaît, mais il en omet un nombre significatif. Comme évoqué dans l'état de l'art, cette limitation a déjà été identifiée par des travaux précédents, révélant que les LLM ont tendance à annoter qu'une partie des expressions attendues.
Néanmoins, cette approche sans réentraînement offre des avantages notables, en particulier en termes de réduction du coût lié à l'annotation manuelle. En exploitant ses capacités fondationnelles, GPT-4 permet d'automatiser partiellement le processus d'extraction sans avoir recours à un corpus annoté dédié, ce qui facilite l'extension à d'autres domaines. Malgré des performances limitées sur certains concepts, cette approche réduit l'intervention humaine. Nous verrons dans la section suivante comment l'approche hybride permet d'en améliorer les résultats.

5.4 Résultats du système hybride

Les résultats de notre approche hybride, qui combine GPT-4, CamemBERT et un filtrage syntaxique, montrent une amélioration significative des performances globales pour l'extraction de termes juridiques. Le Tableau 4 présente les scores en précision, rappel et F1 obtenus pour chacun des concepts étudiés.

	Concepts							
	Action	Actor	Artifact	Condition	Location	Modality	Reference	Time
Précision	0.36	0.80	0.58	0.68	0.35	0.73	0.66	0.64
Rappel	0.78	0.52	0.54	0.73	0.47	0.57	0.75	0.81
F1	0.50	0.63	0.56	0.70	0.41	0.64	0.70	0.72

TABLE 4 – Résultats de précision, rappel et F1 avec notre approche hybride pour l'extraction de termes juridiques.

Le Tableau 5.4 met en évidence l'amélioration significative apportée par l'approche hybride par rapport à GPT-4. En combinant les capacités d'extraction initiale de GPT-4, le filtrage syntaxique et la distillation dans un CamemBERT, nous obtenons un gain de 7 à 24 points de pourcentage sur

Concept	CamemBERT	GPT-4	Hybride
Action	0.57	0.38	**0.50 (+12 %)**
Actor	0.70	0.63	0.63 (0 %)
Artifact	0.26	0.42	**0.56 (+14 %)**
Condition	0.82	0.63	**0.70 (+7 %)**
Location	0.59	**0.42 (+1 %)**	0.41
Modality	0.63	0.54	**0.64 (+10 %)**
Reference	0.76	0.61	**0.70 (+9 %)**
Time	0.74	0.48	**0.72 (+24 %)**

TABLE 5 – Comparaison des performances en F1 pour les trois approches : CamemBERT, GPT-4 et système hybride. En gras, les meilleures performances entre GPT-4 et Hybride.

le score F1 selon les concepts. Ces améliorations sont particulièrement notables pour les concepts les plus complexes ou mal représentés tels que `Artifact`, `Reference` et `Time`.

L'utilisation de règles syntaxiques permet d'atténuer les erreurs d'extraction apparaissant avec les LLMs. Ce filtrage améliore sensiblement la qualité des annotations automatiques générées, qui sont ensuite utilisées pour réentraîner CamemBERT. Le modèle résultant bénéficie à la fois de la capacité générative des LLMs et de l'efficacité prédictive d'un classifieur BERT, tout en limitant l'implication des experts.

Il convient toutefois de noter que le système hybride n'égale pas les performances du modèle CamemBERT entraîné de manière supervisée (sur un jeu de données annoté manuellement par des experts). Ce constat rejoint les observations présentées dans des travaux similaires, tels que celui de Tang et al. [29], soulignant qu'un entraînement basé sur des annotations expertes produit de meilleurs résultats.

Malgré cela, notre analyse démontre que l'approche hybride constitue une solution efficace pour l'extraction de termes juridiques tout en réduisant la dépendance à une annotation experte. Elle ouvre ainsi la voie à des systèmes plus autonomes, adaptés au traitement de corpus juridiques de grande taille, tout en maintenant un bon compromis entre qualité, coût, et effort humain.

6 Conclusion

Notre article a proposé une approche visant à améliorer l'extraction de termes juridiques, tout en réduisant l'implication des experts. Nous avons évalué trois stratégies distinctes pour l'extraction de termes juridiques : le réentraînement d'un modèle CamemBERT, l'exploitation d'un grand modèle de langage (LLM) via l'ingénierie des requêtes, ainsi qu'une approche hybride combinant LLMs, méthodes à base de règles syntaxiques, et un modèle CamemBERT. L'évaluation de ces stratégies a été menée sur un même jeu de données issu de la législation luxembourgeoise.

Nous avons montré que le réentraînement du modèle CamemBERT permet d'obtenir des performances élevées, démontrant sa capacité à extraire efficacement des termes juridiques à partir de textes en français. Par ailleurs, nous avons exploré le potentiel de GPT-4 pour réduire la nécessité d'annotation experte, en soulignant le rôle central de l'ingénierie des requêtes dans la production de sorties structurées. Enfin, l'approche hybride, fondée sur la distillation des connaissances d'un LLM vers CamemBERT via un filtrage par règles, s'est montrée particulièrement prometteuse. Celle-ci améliore significativement les performances de GPT-4 seul, tout en limitant l'intervention des experts à la définition des concepts, de leurs définitions, et à l'élaboration des règles syntaxiques de filtrage.

L'approche hybride présente des atouts majeurs par rapport aux méthodes traditionnelles de l'état de l'art. En tirant parti de GPT-4 pour générer des données annotées (« amorçage » ou bootstrapping), la dépendance au travail d'annotation devient minimale, à condition que le domaine étudié soit bien couvert par les connaissances du LLM.

Cependant, il existe des limites à cette approche hybride. Premièrement, le filtrage syntaxique suppose un accès à un analyseur syntaxique fiable, qui peut varier selon la langue. Ainsi, les règles développées pour le français ne sont pas directement transposables à d'autres langues, comme l'anglais, et requièrent un nouveau travail d'expertise. Par ailleurs, le modèle CamemBERT est spécifiquement entraîné pour le français ; son utilisation sur des documents rédigés dans d'autres langues nécessiterait l'adoption de modèles adaptés, tels que BERT pour l'anglais.

Après l'extraction terminologique, les perspectives de ce travail visent à extraire les relations entre termes juridiques. Des améliorations de l'approche hybride sont également envisagées, en particulier sur le plan du filtrage syntaxique. Au lieu de s'en remettre exclusivement aux experts pour la formalisation des règles, des techniques d'apprentissage non supervisé pourraient être envisagées pour assister, voire automatiser, leur génération, réduisant ainsi davantage l'effort manuel requis.

Remerciements

Ce travail a bénéficié d'un accès aux ressources de calcul intensif de l'IDRIS dans le cadre de l'allocation 2024-AD011014922 accordée par GENCI.

Références

[1] Patrizio Bellan, Mauro Dragoni, and Chiara Ghidini. Extracting business process entities and relations from text using pre-trained language models and in-context learning. In *International Conference on Enterprise Design, Operations, and Computing*, pages 182–199. Springer, 2022.

[2] David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Patrick Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Anonymization of german financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, pages 151–161, 2022.

[3] Julien Breton, Mokhtar Boumedyen Billami, Max Chevalier, and Trojahn Cassia. Leveraging seman-

tic model and llm for bootstrapping a legal entity extraction : An industrial use case. In *20th International Conference on Semantic Systems (SEMANTICS 2024)*, 2024. to appear.

[4] Julien **Breton**, Mokhtar Boumedyen Billami, Max Chevalier, and Cassia Trojahn. Empowering CamemBERT Legal Entity Extraction With LLM Boostrapping. In *Knowledge Engineering and Knowledge Management*, volume 15370, pages 86–101. Springer Nature Switzerland, Cham, 2025.

[5] Silvana Castano, Alfio Ferrara, Emanuela Furiosi, Stefano Montanelli, Sergio Picascia, Davide Riva, and Carolina Stefanetti. Enforcing legal information extraction through context-aware techniques : The ASKE approach. *Computer Law & Security Review*, 52 :105903, 2024.

[6] Yun Chen, Bo Xiao, Zhiqing Lin, Cheng Dai, Zuochao Li, and Liping Yan. Multi-label text classification with deep neural networks. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 409–413. IEEE, 2018.

[7] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1) :1418, 2024.

[8] Mauro Dragoni, Serena Villata, Williams Rizzi, and Guido Governatori. Combining NLP approaches for rule extraction from legal documents. In *1st Workshop on MIning and REasoning with Legal texts (MIREL 2016)*, 2016.

[9] Alex Dunn, John Dagdelen, N. Walker, Sanghoon Lee, Andrew S. Rosen, G. Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *ArXiv*, 2022.

[10] Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. Mining experimental data from materials science literature with large language models : an evaluation study. *Science and Technology of Advanced Materials : Methods*, 4, 2024.

[11] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm : Knowledge distillation of large language models, 2024.

[12] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step ! outperforming larger language models with less training data and smaller model sizes. *ArXiv*, 2023.

[13] Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, et al. Information extraction from clinical notes : Are we ready to switch to large language models ? *arXiv preprint arXiv :2411.10020*, 2024.

[14] Tomoki Ito and Shun Nakagawa. Tender document analyzer with the combination of supervised learning and llm-based improver. In *Companion Proceedings of the ACM Web Conference 2024*, pages 995–998, 2024.

[15] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Icml*, page 3. Williamstown, MA, 2001.

[16] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Finding the law : Enhancing statutory article retrieval via graph neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 2753–2768, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[17] Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67) :2, 2001.

[18] Behrang Mohit. Named entity recognition. In *Natural language processing of semitic languages*, pages 221–245. Springer, 2014.

[19] Vitor Oliveira, Gabriel Nogueira, Thiago Faleiros, and Ricardo Marcacini. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. *Artificial Intelligence and Law*, 2024.

[20] Kalyani Pakhale. Comprehensive overview of named entity recognition : Models, domain-specific applications and challenges. *arXiv preprint arXiv :2309.14084*, 2023.

[21] Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. Named entity recognition and linking for entity extraction from italian civil judgements. In Roberto Basili, Domenico Lembo, Carla Limongelli, and Andrea Orlandini, editors, *AIxIA 2023 – Advances in Artificial Intelligence*, pages 187–201. Springer Nature Switzerland, 2023.

[22] P Lakshmi Prasanna and D Rajeswara Rao. Text classification using artificial neural networks. *International Journal of Engineering & Technology*, 7(1.1) :603–606, 2018.

[23] Philippe Sassier and Dominique Lansoy. *Ubu loi*. Arthème Fayard, France, 2008.

[24] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *ArXiv*, 2018.

[25] Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, Marcello Ceci, and John Dann. An automated framework for the extraction of semantic legal metadata from legal texts. *Empirical Software Engineering*, 26 :1–50, 2021.

[26] Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, and John Dann. Automated extraction of semantic legal metadata using natural language processing. *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 124–135, October 2018.

[27] Firdaus Solihin, Indra Budi, Rizal Fathoni Aji, and Edmon Makarim. Advancement of information extraction use in legal documents. *International Review of Law, Computers & Technology*, 35(3) :322–351, 2021.

[28] Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2) :268–287, 2022.

[29] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining ? arxiv 2023. *arXiv preprint arXiv :2303.04360*, 2023.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Karthik S Vedula, Annika Gupta, Akshay Swaminathan, Ivan Lopez, Suhana Bedi, and Nigam H Shah. Distilling large language models for efficient clinical information extraction. *arXiv preprint arXiv :2501.00031*, 2024.

[32] Zhili Wang, Yufan Wu, Pengbin Lei, and Cheng Peng. Named entity recognition method of brazilian legal text based on pre-training model. *Journal of Physics : Conference Series*, 1550 :032149, 2020.

[33] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv :2302.10205*, 2023.

[34] Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. Empirical study of zero-shot ner with chatgpt. *arXiv preprint arXiv :2310.10035*, 2023.

[35] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER : Generalist model for named entity recognition using bidirectional transformer.

[36] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv :2303.18223*, 2023.

[37] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. UniversalNER : Targeted distillation from large language models for open named entity recognition.