Adaptation des modèles de langue à des domaines de spécialité par un masquage sélectif fondé sur le genre et les caractéristiques thématiques

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour

Adaptation des modèles de langue à des domaines de spécialité par un masquage sélectif fondé sur le genre et les caractéristiques thématiques

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

RÉSUMÉ

Les modèles de langue pré-entraînés ont permis de réaliser des avancées significatives dans diverses tâches de Traitement Automatique du Langage Naturel (TALN). Une des caractéristiques des modèles reposant sur une architecture Transformeur concerne la stratégie de masquage utilisée pour capturer les relations syntaxiques et sémantiques inhérentes à une langue. Dans les architectures de type encodeur, comme BERT, les mots à masquer sont choisis aléatoirement. Cette stratégie ne tient néanmoins pas compte des caractéristiques linguistiques spécifiques à un domaine. Dans ce travail, nous proposons de réaliser un masquage sélectif des mots en fonction de leur saillance thématique dans les documents dans lesquels ils se produisent et de leur spécificité au genre de document. Les performances des modèles résultant d'un pré-entraînement continu dans le domaine juridique soulignent l'efficacité de notre approche sur la majorité des tâches de *benchmark* LexGLUE en langue anglaise.

ABSTRACT

Language Model Adaptation to Specialized Domains through Selective Masking based on Genre and Topical Characteristics

Pre-trained language models have made significant advances in a variety of natural language processing (NLP) tasks. One of the key components of these models using the transformers architecture is their training based on the masking task, where encoder models, such as BERT, randomly select the tokens to be masked. This masking approach does not take into account domain-specific linguistic features. Thus, we propose a new approach to selective word masking with the aim of adapting language models to specialty domains. Our approach weights words according to their thematic salience or document genre specificity, then uses this weight to select words for masking. The performance of the models resulting from continuous pre-training in the legal domain underlines the effectiveness of our strategy on the majority of tasks of the English-language LexGLUE benchmark.

MOTS-CLÉS : Modèle de langue ; stratégie de masquage ; BERT ; méta-discours ; tfidf.

KEYWORDS: Language model ; masking strategy ; BERT ; metadiscourse ; tfidf.

1 Introduction

Les modèles de langue pré-entraînés à grande échelle sont devenus indispensables pour modéliser le langage humain, améliorant considérablement les performances dans diverses tâches de Traitement Automatique du Langage Naturel (TALN) (Bao *et al.*, 2020; Guu *et al.*, 2020; Zhang *et al.*, 2022). Toutefois, le pré-entraînement de tels modèles pour des domaines de spécialité n'est pas toujours possible car il requiert une importante quantité de données qui n'est pas toujours disponible, mais aussi en raison du coût de calcul élevé qu'il représente. Une solution réside ainsi dans le pré-entraînement continu (*Continual Pre-Training*, CPT) pour spécialiser des modèles de fondation (Chalkidis *et al.*, 2020; Wu *et al.*, 2021; Ke *et al.*, 2022; Labrak *et al.*, 2023).

Classiquement, les modèles ainsi construits réutilisent l'algorithme d'apprentissage utilisé lors du pré-entraînement initial. Dans les modèles qui reposent sur une architecture de type encodeur à la BERT (Devlin *et al.*, 2019), l'algorithme utilisé est le Modèle de Langage Masqué (*Masked Language Modeling*, MLM). Dans cet algorithme, le modèle apprend à prédire un mot masqué aléatoirement dans une séquence. Les modèles résultants de cet apprentissage ont montré leur capacité à capturer les relations sémantiques complexes et les structures syntaxiques inhérentes au langage naturel. Certains travaux ont étendu le MLM, notamment pour affiner la capacité des modèles à capturer des expressions multimots (Sun *et al.*, 2019; Joshi *et al.*, 2020; Levine *et al.*, 2020; Li & Zhao, 2021). Cependant, aucune étude n'a envisagé d'étendre le MLM pour prendre en compte la spécificité du genre de document étudié.

Dans cet article, nous présentons une nouvelle approche de masquage sélectif pour l'adaptation de modèles de langue à des domaines de spécialité qui exploite les informations relatives à la spécificité au genre de document et à la topicalité. Notre approche pondère les mots en fonction de leur spécificité au genre de documents (on parle aussi de "caractère méta-discursif") et de leur saillance thématique. Nous utilisons un score TF×IDF pour mesurer la saillance et proposons une formule pour estimer la spécificité. En utilisant ces scores pour choisir les mots à masquer, nous forçons le modèle à s'adapter à la compréhension et à la prédiction des mots spécifiques à un domaine cible. Pour illustrer l'efficacité de notre approche, nous menons des expériences sur le pré-entraînement continu de modèles type BERT dans le domaine juridique, en comparant différentes stratégies de masquage des mots.

Nos contributions sont les suivantes :

— Nous proposons une nouvelle approche de masquage fondée sur la sélection de mots (méta-discours et TF×IDF) pour l'entraînement de modèles de langue ;
— Nous effectuons une analyse comparative de plusieurs manières de sélectionner les mots pendant le processus d'entraînement ;
— Nous partageons nos modèles et notre code source pour rendre plus facile le pré-entraînement de modèles de langue pour des domaines de spécialité, selon notre stratégie d'apprentissage [1] ;

2 Travaux connexes

L'approche de masquage classique utilisé dans BERT (Devlin *et al.*, 2019) consiste à sélectionner 15 % de tokens dans une séquence. Chacun de ces 15 % de tokens est ensuite substitué par le token spécial `[MASK]` (80 % de probabilité), remplacé par un mot tiré aléatoirement (10 %) ou bien laissé

1. github.com/ygorg/legal-masking

intact (10 %). L'objectif du modèle est de prédire les mots masqués originaux.

Dans le but d'enrichir les capacités de représentation des MLMs, ERNIE (Sun *et al.*, 2019) et SpanBERT (Joshi *et al.*, 2020) affinent la stratégie de masquage classique (aléatoire) utilisée par BERT. Ces modèles masquent respectivement des mots entiers et des parties contiguës de texte de manière aléatoire. Ces approches améliorent les performances des tâches de réponse aux questions et de résolution de coréférence. Yang *et al.* (2023) observent qu'à partir d'un moment de l'entraînement, les modèles cessent d'apprendre à partir de types de mots spécifiques, identifiés par des étiquettes morphosyntaxiques (*POS tags*), en fonction de la stabilité de la fonction de coût de modèle. Ainsi, ils introduisent une approche de masquage qui varie dans le temps, cette approche diffère des méthodes statiques qui maintiennent un contenu inchangeable tout au long de l'apprentissage. Cette stratégie améliore les performances sur le *benchmark* GLUE (Wang *et al.*, 2018).

D'autres méthodes s'intéressent à adapter les modèles à des domaines spécifiques et choisissent donc les mots à masquer en tenant compte des spécificités du domaine des documents. Dans le domaine des brevets, Althammer *et al.* (2021) masque les mots qui apparaissent fréquemment dans des groupes nominaux. Cette approche montre son efficacité dans les tâches de classification. Dans le domaine clinique, EntityBERT (Lin *et al.*, 2021) masque des tokens identifiés comme "entité" (symptôme, médicament, date) par un modèle pré-entraîné de détection d'entités. L'application de cette stratégie repose néanmoins sur la disponibilité de modèles de détection d'entités du domaine cible.

Moins générique, certaines approches adaptent les modèles à une tâche spécifique avant d'effectuer un affinage. Par exemple, pour des tâches de classification de document, l'approche proposée par Golchin *et al.* (2023) utilise KeyBERT (Grootendorst, 2020) pour masquer uniquement les mots identifiés comme mots-clés. L'approche *Difference-Masking* (Wilf *et al.*, 2023) quant à elle, masque les mots les plus similaires (en termes de plongement de mots) à des mots "ancres". Les ancres sont des mots ayant une fréquence plus élevée dans le corpus que dans un corpus général. Bien que ces techniques soient similaires à celles faisant de l'adaptation au domaine, les résultats de ces techniques d'adaptation à la tâche ne sont pas directement comparables car moins généralisables.

Le choix du taux de masquage influence de manière significative la tâche de Modèle de Langage Masqué et donc le pré-entraînement. L'étude menée par Wettig *et al.* (2023) indique que le taux de masquage optimal dépend de la taille du modèle à entraîner, 40% étant optimal pour les modèles larges et 20% pour les modèles de base, comme l'ont montré les évaluations des *benchmarks* GLUE et SQuAD.

L'approche présentée ici s'intéresse à sélectionner les mots à masquer en fonction de leur saillance thématique et de leur spécificité au genre de document. Elle s'inscrit dans les travaux d'adaptation au domaine et ainsi est indépendante des tâches en aval.

3 Stratégie de masquage fondée sur le genre et la topicalité

Contrairement à l'approche originale de BERT, qui sélectionne aléatoirement les tokens à masquer (Devlin *et al.*, 2019), notre approche se concentre sur le masquage au niveau des mots, les choisissant en fonction de leur spécificité par rapport au genre du texte ou bien à leur saillance thématique dans un document. Notre approche fonctionne en deux étapes. Tout d'abord, nous attribuons un *score de spécificité au genre* et un *score de saillance thématique* à chaque mot à partir de notre corpus spécifique à un domaine (Section 3.1). Ensuite, nous utilisons ces scores pour hiérarchiser et

sélectionner les mots à masquer (Section 3.2).

3.1 Pondération des mots

Nous proposons deux manières de pondérer les mots à partir d'un corpus de documents de spécialité. La première approche, le *score de topicalité* (TF×IDF), quantifie la saillance thématique d'un mot dans un document donné. Pour cela, nous utilisons la mesure classique TF×IDF (Jones, 1972), qui pondère un mot en fonction de son nombre d'occurrences dans un document particulier et du nombre de documents dans lequel il apparaît dans le corpus.

La seconde approche, le *score de spécificité au genre de texte* (`MetaDis`), évalue dans quelle mesure un mot est caractéristique d'un genre de documents. Un genre de documents est caractérisé par une structure commune (Biber & Conrad, 2019; Hyland, 1998). Par exemple, les jurisprudences (domaine juridique) présentent successivement des faits, puis des arguments et enfin un raisonnement pour parvenir à une décision. Chacune de ces parties utilise un lexique particulier à ce genre, nous désignons ce lexique par le terme de *méta-discours*. Bien que Hernandez & Grau (2003) ait utilisé le score de fréquence inverse de document pour évaluer la spécificité, leur mesure ne tient pas compte de la distribution des occurrences dans les documents. Nous supposons qu'un indicateur de méta-discours est présent dans une proportion constante dans les documents du même genre. Pour capturer de telles propriétés et calculer un score de méta-discours, nous proposons la formule décrite dans l'équation 1 :

$$s_t = \frac{df_t}{tf_t} * \left(1 - \frac{std(tf_{d,t})}{max(tf_{d,t})}\right) * \frac{df_t}{N} \quad (1)$$

Ici, df_t représente le nombre de documents dans lequel le mot t apparaît, tf_t le nombre d'occurrences de t dans le corpus, $tf_{d,t}$ le nombre d'occurrences de t dans le document d et N le nombre de documents dans le corpus. L'intuition derrière le premier terme est de donner un haut score aux mots apparaissant dans de nombreux documents. Le second mesure la constance d'apparition du mot et donne un haut score à ceux qui ont la même fréquence dans tous les documents (faible $std(tf_{d,t})$), le dénominateur normalise par la fréquence maximale du mot. Enfin, le troisième terme donne un haut score aux mots apparaissant dans de nombreux documents.

3.2 Stratégie de sélection des mots à masquer

Nous proposons deux manières de sélectionner les mots à masquer qui utilisent les scores TF×IDF ou `Metadis`. Notre première méthode, `TopN`, sélectionne les mots ayant les scores les plus élevés. Cette méthode est déterministe, pour un document les mots masqués seront toujours les mêmes.

La seconde méthode, `Samp` (échantillonnage, *sample* en anglais), vise à améliorer la robustesse du modèle en évitant le masquage systématique des mêmes mots. Cette méthode s'inspire du masquage dynamique utilisé dans RoBERTa (Liu *et al.*, 2019) et introduit un niveau d'aléa pondéré qui change à chaque itération de modèle. En pratique, nous échantillonnons aléatoirement des mots (sans remise) sur la base de la distribution des scores calculés.

2. Pour la stratégie `TopN` la fonction Max est utilisée.

Algorithm 1 Masquage sélectif

function MASK($tokens$)
  $\mathcal{M} \leftarrow \{\}$
  $W \leftarrow MotsEntiers(tokens)$
  $S \leftarrow CalculeScore(W)$
  while $|\mathcal{M}| < 0.15 * |tokens|$ **do**
    $i \leftarrow$ Echantillone(S) [2]
    Supprime $W[i]$ and $S[i]$
    if $|\mathcal{M}| + |W[i]| \leq 0.15 * |tokens|$ **then**
      $\mathcal{M} \leftarrow \mathcal{M} + w$
    end if
  end while
  return $\mathcal{M}$
end function

Une fois les mots sélectionnés, nous choisissons les tokens à masquer suivant l'Algorithme 1 pour masquer effectivement 15 % (Devlin *et al.*, 2019) des tokens de la séquence. Dans l'algorithme, M dénote les mots à masquer, W les mots segmentés et S les scores associés à chaque mot.

4 Paramètres expérimentaux

Nous utilisons comme base les modèles pré-entraînés BERT (Devlin *et al.*, 2019) et LegalBERT (Chalkidis *et al.*, 2020). L'efficacité de notre approche de masquage est évaluée par un pré-entraînement continu sur ces modèles, en se concentrant sur l'adaptation au domaine juridique. Dans cette section, nous présentons d'abord les données utilisées pour le pré-entraînement continu (Section 4.1). Puis, les tâches d'évaluation (Section 4.2) et enfin les détails expérimentaux (Section 4.3).

4.1 Corpus de pré-entraînement

Pour le pré-entraînement continu et la sélection du masquage de mots, nous avons choisi de nous concentrer sur le domaine juridique en utilisant un sous-ensemble du corpus LexFiles (Chalkidis *et al.*, 2023) représentatif du *benchmark* LexGLUE (Chalkidis *et al.*, 2022). Les documents ont été sélectionnés pour offrir une collection équilibrée et diversifiée, englobant les nuances linguistiques (voir la Table 1).

Sous-Corpus	**# Doc.**	**# Tokens**
EU Case Law	29,8K	178,5M (29%)
ECtHR Case Law	12,5K	78,5M (13%)
U.S. Case Law	104,7K	235,5M (39%)
Indian Case Law	34,8K	111,6M (19%)
Total	**181,8K**	**604,1M**

TABLE 1 – Détails de l'ensemble de données utilisé pour le pré-entraînement continu.

4.2 Tâches d'évaluation

Nous évaluons la performance de nos modèles à l'aide de LexGLUE (Chalkidis *et al.*, 2022) un *benchmark* conçu spécifiquement pour le domaine juridique. LexGLUE englobe une gamme variée de tâches provenant des systèmes juridiques de l'Europe, des États-Unis et du Canada. Une telle configuration permet de tester rigoureusement la capacité de notre stratégie dans un spectre de tâches complexes. Voici un aperçu de chacune de ces tâches : [3]

— **ECtHR A & B** consistent à déterminer quels articles de lois sont enfreints (ECtHR A), ou prétendument enfreints (ECtHR B), par une liste de faits. Les articles de loi sont 11 articles de la Cour Européenne des Droits de l'Homme.
— **SCOTUS** consiste à classer les opinions de la Cour suprême des États-Unis (SCOTUS) parmi 14 thèmes, tels que la procédure pénale, les droits civils, l'activité économique…
— **EUR-LEX** concerne la législation de l'Union européenne publiée sur le portail EUR-Lex. L'objectif est d'assigner aux textes de loi les concepts du thésaurus EUROVOC pertinents (parmi les 100 concepts les plus fréquents).
— **LEDGAR**, Labeled EDGAR, consiste à trouver le thème principal (parmi 100) de chaque paragraphe des contrats de la base de données EDGAR.
— **UNFAIR-ToS** cherche à détecter les clauses de conditions de services qui enfreignent potentiellement le droit des consommateurs européens. Chaque phrase est classée parmi 9 types de clauses contractuelles abusives.

4.3 Détails expérimentaux

Pour le pré-entraînement continu, nous avons mené des sessions d'entraînement totalisant plus de 20 heures en utilisant 16 GPU V100 sur le supercalculateur Jean Zay. Suivant la méthode de Labrak *et al.* (2023), nous avons adopté une taille de batch de 16 et configuré les étapes d'accumulation de gradient à 16, ce qui donne une taille de batch effective de 4 096. Pour déterminer les scores de performance de la tâche, nous avons calculé la moyenne des scores de trois exécutions indépendantes. Les métriques utilisées sont les mesures : micro (μF_1) et macro (m-F1) F-mesure.

5 Résultats et discussions

Les résultats sont détaillés dans la Table 2, chaque colonne correspondant à une tâche du *benchmark* LexGLUE.

Effets du pré-entraînement continu Nos résultats montrent l'efficacité du pré-entraînement continu dans toutes les tâches pour les modèles BERT et LegalBERT. Plus précisément, la macro F1 (m-F1) de la configuration de référence BERT+CPT augmente sensiblement de 60,39 à 64,69 dans la tâche ECtHR (B). De manière similaire, LegalBERT+CPT montre des améliorations substantielles dans la tâche EUR-LEX au niveau de la m-F1. Ces améliorations, même sans modification de la stratégie de masquage, suggèrent que le corpus utilisé contient de nouvelles caractéristiques spécifiques au domaine permettant d'enrichir les connaissances des modèles de langue.

3. Un problème dans le code de la tâche `case_hold` empêche son exécution, nous ne rapportons donc pas les résultats.

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS	
	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1	μF_1	m-F1
BERT	62,12	52,66	69,59	60,39	69,61	58,65	71,70	54,87	87,85	**82,30**	**95,66**	80,97
+ CPT (référence)	63,12	54,13	71,06	**64,69**	70,57	**60,38**	71,86	56,18	**87,90**	82,02	95,56	81,46
+ MetaDis - Samp	62,55	54,88	70,45	63,10	70,26	59,12	71,66	56,00	87,68	82,25	95,38	79,18
+ MetaDis - TopN	62,17	53,35	70,29	62,29	69,92	60,08	71,67	56,95	87,78	82,11	95,57	**81,51**
+ TF×IDF - Samp	**63,36**	**56,60**	71,32	64,58	69,69	59,10	**71,93**	55,82	87,69	82,11	95,50	78,63
+ TF×IDF - TopN	62,66	56,46	**71,50**	63,58	**70,71**	60,06	71,73	**57,73**	87,67	81,89	95,49	79,45
LegalBERT	63,41	53,19	72,10	63,68	73,61	61,50	71,93	55,47	87,91	81,67	95,81	81,27
+ CPT (référence)	63,64	**58,73**	72,60	64,95	**74,64**	**63,13**	72,01	55,12	**88,41**	**82,92**	**95,82**	79,70
+ MetaDis - Samp	63,39	56,39	73,08	65,76	74,21	62,97	**72,03**	54,76	88,38	82,58	95,20	80,26
+ MetaDis - TopN	**64,07**	58,56	72,53	**66,83**	73,88	62,57	71,96	55,01	88,32	82,16	94,80	73,67
+ TF×IDF - Samp	63,38	56,78	72,21	65,67	73,71	62,85	71,78	**55,82**	88,19	82,36	95,80	**82,12**
+ TF×IDF - TopN	62,89	53,58	**73,26**	65,86	74,38	63,10	71,90	55,08	88,27	82,65	95,59	81,20

TABLE 2 – Performances des modèles BERT et LegalBERT de base, après pré-entraînement continu avec masquage aléatoire (+CPT) ou masquage sélectif (+MetaDis, +TF×IDF) sur le *benchmark* LexGLUE. La meilleure valeur de chaque colonne est indiquée en gras, la deuxième meilleure est soulignée. Les fonds Verts et Oranges représentent respectivement les scores MetaDis et TF×IDF, les fonds foncés montrent les améliorations par rapport à la référence.

Masquage sélectif vs classique Pour évaluer l'efficacité de nos masquages (MetaDis, TF×IDF), nous les comparons au masquage aléatoire classique de BERT (+CPT (référence)) après pré-entraînement continu. Les résultats indiquent que, quelle que soit la stratégie de masquage utilisée, au moins un de nos modèles apporte des améliorations par rapport à la référence. Avec BERT, des améliorations notables sont observées dans les tâches ECtHR(A) et LEDGAR, obtenant respectivement des scores de 54,88 et 82,25 de m-F1 pour MetaDis - Samp. Cela peut être attribué à la capacité de nos modèles à exploiter les informations relatives au genre (MetaDis) et à la thématique (TF×IDF) propres au domaine juridique. Pour les modèles LegalBERT, des améliorations ont été observées dans les tâches ECtHR(B) et UNFAIR-ToS avec les deux pondérations. Ces résultats soulignent les avantages d'un masquage sélectif des mots, en particulier avec des modèles déjà adaptés. Par rapport aux résultats rapportés dans Chalkidis *et al.* (2022), notre stratégie obtient, sur la tâche SCOTUS, des résultats supérieurs aux modèles hiérarchiques, tout en requérant moins de paramètres et une architecture moins complexe. Cela souligne l'intérêt de développer des techniques de pré-entraînement qui se concentrent sur des caractéristiques linguistiques spécifiques à un domaine. Cela permet ainsi de se passer de modèles complexes requérant des paramètres supplémentaires ou bien des temps d'entraînement plus longs.

Méta-discours vs. TF×IDF En comparant les stratégies de masquage pondérant le genre (+MetaDis) et la topicalité (+TF×IDF) sur les modèles BERT et LegalBERT, nous avons observé des tendances différentes. Pour les modèles BERT, le méta-discours a démontré son efficacité dans les tâches où les caractéristiques linguistiques spécifiques au genre jouent un rôle important, comme dans les tâches ECtHR (A), LEDGAR et UNFAIR-ToS. Au contraire, TF×IDF montre des améliorations pour les tâches qui concernent la pertinence thématique, comme les tâches ECtHR (B), EURLEX et SCOTUS. Par exemple, la tâche EURLEX se concentre sur les textes législatifs de l'Union européenne, marqués par une diversité de concepts issus d'EuroVoc [4], le thésaurus multilingue, soulignant

4. eur-lex.europa.eu/browse/eurovoc.html

ainsi leur riche contenu thématique.

Pour le modèle LegalBERT, les deux stratégies présentent des performances similaires pour les tâches ECtHR (B) et UNFAIR-ToS. En revanche, pour la tâche ECtHR (A) la pondération basée sur le meta-discours obtient de meilleures performances, pour EURLEX c'est la pondération TF×IDF. Ces résultats suggèrent que la pertinence thématique est un facteur déterminant pour la tâche EURLEX, quel que soit le modèle de base. Cependant, pour la tâche ECtHR (A) les aspects liés au genre (MetaDis) sont particulièrement pertinents, soulignant l'importance de la structure des documents pour cette tâche.

Samp vs. TopN La comparaison des configurations BERT sur les deux pondérations indique que la stratégie Samp (échantillonnage) permet d'obtenir de meilleures performances dans les tâches ECtHR (A) et LEDGAR. La stratégie TopN, quant à elle, apporte une amélioration pour la tâche EURLEX. Avec la pondération TF×IDF, la stratégie TopN obtient de meilleures performances pour les tâches ECtHR (B) et SCOTUS. Ainsi, TopN semble plus efficace lorsque la topicalité est un facteur déterminant.

En ce qui concerne les modèles LegalBERT utilisant la pondération MetaDis, la stratégie Samp permet d'améliorer les performances dans les tâches ECtHR (B), EURLEX et UNFAIR-ToS. La stratégie TopN, quant à elle, permet de réaliser des progrès notables dans les tâches ECtHR (A) et (B), ce qui souligne son intérêt pour les tâches nécessitant une compréhension thématique des documents. Toujours en partant de LegalBERT et en utilisant la pondération TF×IDF, des améliorations sont observées dans les tâches ECtHR (B) et UNFAIR-ToS avec les deux stratégies TopN et Samp. Notons que la stratégie Samp obtient les meilleurs résultats sur la tâche EURLEX.

6 Conclusion et perspectives

Nos expériences montrent que nos approches de masquage sélectif, qui intègrent les caractéristiques du genre et du thème du document, jouent un rôle crucial dans l'adaptation des modèles à un domaine de spécialité. Nous observons des améliorations dans chacune les tâches du *benchmark* LexGLUE axé sur le domaine juridique. En particulier, des améliorations importantes ont été obtenues sur les tâches ECtHR et EUR-LEX pour les modèles BERT et LegalBERT. Néanmoins, les améliorations obtenues ne sont pas consistantes pour chaque modèle et chaque tâche et indiquent que le choix de la pondération semble dépendant de la tâche. Ainsi, plusieurs axes de recherche émergent : tout d'abord, mesurer la capacité de notre approche à généraliser à d'autres domaines, tels que le domaine clinique ou scientifique. Ensuite, étudier l'impact de notre approche dans le cadre de l'adaptation à la tâche.

Considérations éthiques

Concernant les risques et les biais potentiels inhérents aux modèles de langue entraînés sur des ensembles de données juridiques, les corpus peuvent comprendre des textes de qualité et de représentativité variables. L'utilisation de modèles entraînés sur des textes juridiques, tels que BERT, pourrait introduire des biais liés à la justice, à l'utilisation d'un langage genré, à la représentation de groupes de minorités et à la nature dynamique des normes juridiques au fil du temps. Il est impératif

que ces biais soient évalués et atténués de manière approfondie afin de garantir des performances équitables entre les différents groupes démographiques et de rester en phase avec l'évolution des normes juridiques.

Limitations

Notre travail propose une nouvelle façon d'aborder le pré-entraînement continu des modèles de langue dans le domaine juridique avec un masquage sélectif, mais présente néanmoins certaines limitations. En particulier, l'étude actuelle se concentre uniquement sur l'architecture BERT, limitant notre capacité à étudier une gamme plus large de modèles de langue. Ces derniers pourraient en effet présenter des comportements distincts et différentes sensibilités à notre stratégie de pré-entraînement continu et de masquage. Les études futures devraient explorer d'autres modèles, tels que DrBERT (Labrak *et al.*, 2023) et RoBERTa (Liu *et al.*, 2019), afin de fournir une compréhension plus complète des effets de notre stratégie. En outre, notre étude manque d'une comparaison directe avec un modèle entraîné à partir de zéro (*from scratch*) utilisant le masquage sélectif. Une telle comparaison constituerait un point de référence précieux permettant de déterminer d'autres avantages à notre méthode. Enfin, un réglage plus poussé des hyper-paramètres pourrait conduire à une amélioration des performances du modèle.

Remerciements

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

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

Références

ALTHAMMER S., BUCKLEY M., HOFSTÄTTER S. & HANBURY A. (2021). Linguistically informed masking for representation learning in the patent domain. In *2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2021)*, New York, NY, USA : Association for Computing Machinery.

BAO H., DONG L., WEI F., WANG W., YANG N., LIU X., WANG Y., PIAO S., GAO J., ZHOU M. & HON H.-W. (2020). Unilmv2 : Pseudo-masked language models for unified language model pre-training.

BIBER D. & CONRAD S. (2019). *Register, Genre, and Style*. Cambridge University Press. Google-Books-ID : x7OQDwAAQBAJ.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : The muppets straight out of law school. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.261.

CHALKIDIS I., GARNEAU N., GOANTA C., KATZ D. & SØGAARD A. (2023). LeXFiles and LegalLAMA : Facilitating English Multinational Legal Language Model Development. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15513–15535, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.865.

CHALKIDIS I., JANA A., HARTUNG D., BOMMARITO M., ANDROUTSOPOULOS I., KATZ D. & ALETRAS N. (2022). LexGLUE : A Benchmark Dataset for Legal Language Understanding in English. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4310–4330, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.297.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GOLCHIN S., SURDEANU M., TAVABI N. & KIAPOUR A. (2023). Do not mask randomly : Effective domain-adaptive pre-training by masking in-domain keywords. In B. CAN, M. MOZES, S. CAHYAWIJAYA, N. SAPHRA, N. KASSNER, S. RAVFOGEL, A. RAVICHANDER, C. ZHAO, I. AUGENSTEIN, A. ROGERS, K. CHO, E. GREFENSTETTE & L. VOITA, Éds., *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, p. 13–21, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.repl4nlp-1.2.

GROOTENDORST M. (2020). Keybert : Minimal keyword extraction with bert. DOI : 10.5281/zenodo.4461265.

GUU K., LEE K., TUNG Z., PASUPAT P. & CHANG M.-W. (2020). Realm : Retrieval-augmented language model pre-training.

HERNANDEZ N. & GRAU B. (2003). Automatic extraction of meta-descriptors for text description. In *International Conference on Recent Advances In Natural Language Processing (RANLP)*, Borovets, Bulgaria.

HYLAND K. (1998). Persuasion and context : The pragmatics of academic metadiscourse. *Journal of pragmatics*, **30**(4), 437–455.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**, 11–21. DOI : 10.1108/eb026526.

JOSHI M., CHEN D., LIU Y., WELD D. S., ZETTLEMOYER L. & LEVY O. (2020). SpanBERT : Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, **8**, 64–77. DOI : 10.1162/tacl_a_00300.

KE Z., SHAO Y., LIN H., KONISHI T., KIM G. & LIU B. (2022). Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LEVINE Y., LENZ B., LIEBER O., ABEND O., LEYTON-BROWN K., TENNENHOLTZ M. & SHOHAM Y. (2020). Pmi-masking : Principled masking of correlated spans.

LI Y. & ZHAO H. (2021). Pre-training universal language representation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 5122–5133, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.398.

LIN C., MILLER T., DLIGACH D., BETHARD S. & SAVOVA G. (2021). EntityBERT : Entity-centric masking strategy for model pretraining for the clinical domain. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 191–201, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.21.

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

SUN Y., WANG S., LI Y., FENG S., CHEN X., ZHANG H., TIAN X., ZHU D., TIAN H. & WU H. (2019). ERNIE : Enhanced Representation through Knowledge Integration. arXiv :1904.09223 [cs].

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In T. LINZEN, G. CHRUPAŁA & A. ALISHAHI, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WETTIG A., GAO T., ZHONG Z. & CHEN D. (2023). Should you mask 15% in masked language modeling ? In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2985–3000, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.217.

WILF A., AKTER S., MATHUR L., LIANG P., MATHEW S., SHOU M., NYBERG E. & MORENCY L.-P. (2023). Difference-masking : Choosing what to mask in continued pretraining. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 13222–13234, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.881.

WU T., CACCIA M., LI Z., LI Y.-F., QI G. & HAFFARI G. (2021). Pretrained language model in continual learning : A comparative study. In *International Conference on Learning Representations*.

YANG D., ZHANG Z. & ZHAO H. (2023). Learning Better Masking for Better Language Model Pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7255–7267, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.400.

ZHANG S., ROLLER S., GOYAL N., ARTETXE M., CHEN M., CHEN S., DEWAN C., DIAB M., LI X., LIN X. V., MIHAYLOV T., OTT M., SHLEIFER S., SHUSTER K., SIMIG D., KOURA P. S., SRIDHAR A., WANG T. & ZETTLEMOYER L. (2022). Opt : Open pre-trained transformer language models.

A Paramètres de pré-entraînement continu

Avant l'entraînement, les exemples ont été mélangés aléatoirement à trois reprises en utilisant la même graine. Nous entraînons chaque modèle en au moyen de la bibliothèque python `transformers`.

L'entraînement est réalisé sur 10 époques, représentant 4453 étapes pour BERT et 4396 étapes pour LegalBERT, les méthodes de segmentation en sous-mots de chaque modèle étant différentes.

Au total, nous estimons le temps de calcul total à $\simeq$ 4,100 h, à savoir 3,200 h d'entraînement, 380h pour l'évaluation des modèles et 520h pour le développement.

B Analyse des mots les plus masqués

Le score de TF×IDF a été calculé à l'aide du package python `scikit-learn`.

Pour mieux comprendre les différences de mots sélectionnés en fonction des deux scores d'importance, nous présentons dans la Table 3 les 50 mots les plus masqués pour 10 % du corpus d'entraînement.

TF×IDF
applicant, court, 2007, extradition, prosecutor, meshchanskiy, russian, dzhurayev, moscow, uzbekistan, tashkent, district, custody, government, convention, article, office, decision, §, detention, russia, ccp, 4, preventive, v, minsk, federation, 2, application, uzbek, proceedings, 1, 5, criminal, procedure, january, 38124, case, 29, merits, may, dismissed, law, rakhimovskiy, 466, request, decided, sobir, arrest, provisions
MetaDis
general, application, decision, january, september, decided, august, 4, 28, 9, 3, issued, request, rules, dismissed, 23, 29, indicated, basis, ordered, european, apply, be, 24, 17, date, 5, 30, held, final, december, 26, 6, 11, mentioned, applied, specified, 12, february, placed, 2, whether, remain, first, to, deliberated, represented, constitute, case, article

TABLE 3 – 50 mots les plus masqués avec les pondérations TF×IDF et `Metadis` ordonnés par fréquence