Un retour d'expériences sur l'adaptation de modèles de langue à la presse régionale : connaissance n'est pas compétence !

Soumission anonyme
Institution anonyme
anonyme@entite.fr

RÉSUMÉ

Nous présentons une étude sur l'adaptation de modèles de langue (ML) généralistes à la presse régionale. En particulier, nous nous intéressons aux faits et relations spécifiques au journal appris lors de l'adaptation du ML et à l'impact sur des tâches classiques. Nous analysons les conséquences de l'adaptation du vocabulaire et de la poursuite du pré-entraînement pour deux modèles bidirectionnels récents et mettons en évidence que ces deux étapes permettent de mieux capturer les spécificités du journal et d'acquérir des connaissances. Nous évaluons ensuite l'impact sur différentes tâches canoniques montrent que l'adaptation des ML n'améliore pas les performances sur nos tâches, en dehors du cas où les données d'apprentissage sont fortement limitées.

ABSTRACT

A feedback on adapting language models to the regional press: Knowledge does not mean skills!

We present a study on the adaptation of general language models (LM) to regional newspapers. In particular, we focus on the facts and relationships specific to newspapers learned during the adaptation of the LM and the impact on classic tasks. We analyze the consequences of vocabulary adaptation and We analyze the consequences of vocabulary adaptation and continued pre-training for two recent bidirectional models and highlight that these two steps make it possible to better capture the specificities of the newspaper and acquire knowledge. We then evaluate the impact on various canonical tasks and show that ML adaptation does not improve performance on our tasks, except in cases where the training data is very limited.

MOTS-CLÉS : modèles de langue, adaptation de vocabulaire, pré-apprentissage continu, acquisition de connaissance, classification, entités nommées, recherche d'information.

KEYWORDS: language models, vocabulary adaptation, continued pre-training, knowledge acquisition, classification, named entities, information retrieval.

1 Introduction

Les grands modèles pré-entrainés comme modèles de langue, comme GPT ou BERT et leurs multiples descendants, sont des modèles généralistes, dont les paramètres sont appris sur de grandes quantités de texte ne relevant pas d'un domaine ou d'un sujet particulier. Il a été montré dans de nombreux travaux que poursuivre l'apprentissage du modèle de langue sur des données d'un domaine ciblé avant de le spécialiser sur des tâches précises permet d'améliorer les résultats (Gururangan *et al.*, 2020; Sun *et al.*, 2020; Wu *et al.*, 2022; Xie *et al.*, 2024) : par eemple, dans le domaine de la finance (Xie *et al.*,

2024), dans les domaines juridique (Chalkidis *et al.*, 2020; Guo *et al.*, 2025) ou encore médical (Wu *et al.*, 2024; Guo *et al.*, 2025). Il s'agit typiquement de domaines de spécialité qui font appel à des termes très spécifiques, en général peu connus des modèles généralistes, et à un style rédactionnel parfois significativement différent du langage modélisé par de tels modèles.

Dans cet article, nous nous intéressons à l'impact de la poursuite de l'apprentissage du modèle de langue dans le cadre de la presse régionale, en l'occurrence celle publiée par le journal Ouest France. Sans être, bien sûr, à proprement parler dans le cadre d'un domaine de spécialité, les caractéristiques des textes publiés dans le journal divergent des données généralistes utilisées dans les modèles génériques pour le français comme FlauBERT (Le *et al.*, 2020), CamemBERT 2.0 (Antoun *et al.*, 2024) ou ModernCamemBERT (Antoun *et al.*, 2025). Le genre journalistique est globalement assez bien capturé par ces modèles généralistes, et le style rédactionnel de Ouest France, raisonnablement proche du style journalistique général, également. En revanche, les entités mentionnées et les faits qui s'y rapportent sont spécifiques au journal, avec notamment de nombreux lieux et personnalités typiques de l'actualité dans l'ouest de la France. Nous parlons par la suite – certes quelque peu abusivement – d'adaptation au domaine du journal Ouest France ou à son style rédactionnel pour référer aux études menées pour chercher à mieux intégrer dans les modèles généralistes ces entités et faits particuliers utilisés dans les articles, et autres spécificités de rédaction du quotidien [1].

Plus précisément, nous étudions l'impact de l'adaptation d'un modèle de langue bidirectionnel généraliste, d'une part en augmentant son vocabulaire pour prendre en compte les entités spécifiques et, d'autre part, en poursuivant son apprentissage de manière à lui permettre de capturer plus d'informations et de faits contenus dans les archives du journal avant de le spécialiser sur des tâches d'intérêt. Notre étude empirique exploite 15 ans d'archives du journal Ouest France pour poursuivre l'entraînement de modèles de langue, en l'occurrence CamemBERT 2.0 et ModernCamemBERT.

Dans un premier temps, dans la section 2, nous nous intéressons à l'adaptation du vocabulaire et à la poursuite du pré-apprentissage. Nous mettons en évidence que ces deux étapes sont indispensables et permettent d'aboutir à des modèles qui correspondent mieux au style rédactionnel du journal et qui ont acquis une meilleure connaissance des faits qui y sont relatés. Nous observons aussi de manière classique que plus un fait est mentionné dans les données utilisées pour poursuivre l'apprentissage, mieux il sera intégré par le modèle. Dans un second temps, dans la section 3, nous nous intéressons à l'adaptation des modèles pour des tâches de classification, de détection des entités nommées, et de recherche d'information. Les résultats expérimentaux montrent que la poursuite du pré-apprentissage n'amène en général pas de gains significatifs sur ces tâches. Nous mettons cependant en évidence que l'adaptation des modèles généralistes facilite l'apprentissage avec très peu de données, ce qui est un scénario fréquent lorsque l'on cherche à développer des modèles de classification dans le contexte d'une rédaction.

2 Adaptation de modèles au domaine du journal Ouest France

Nous décrivons tout d'abord les données utilisées pour adapter au domaine du journal deux modèles de langues généralistes récents couramment utilisés pour le français, CamemBERT 2.0 et ModernCamemBERT que nous désignerons par la suite comme CamemBERT et ModernBERT. Nous détaillons ensuite la procédure d'adaptation au domaine avant d'évaluer l'impact de celle-ci sur les modèles de

1. Il serait plus approprié de parler d'adaptation au discours et au style particulier de Ouest France au sein du genre journalistique.

langue et sur les connaissances qu'ils intègrent.

2.1 Données

Notre étude s'appuie sur une partie des archives du quotidien Ouest France, unique par sa couverture géographique et temporelle. Le fonds documentaire utilisé dans cet article comprend l'intégralité des publications parues entre 2010 et 2024, soit 15 années d'actualité journalistique couvrant les 12 départements de l'ouest de la France, ainsi que, dans une moindre mesure, l'actualité nationale et internationale. La diversité éditoriale est caractéristique de ce corpus : faits divers, sports, vie culturelle et associative, politique locale, société, éditoriaux et tribunes constituent des domaines variés. Sur le plan quantitatif, le corpus totalise 16M d'articles, représentant environ 4,6 milliards de *tokens*. Cette volumétrie, bien qu'importante, demeure modeste comparée aux volumes de données d'apprentissage analysés par les modèles généralistes contemporains. Notre corpus représente en effet approximativement 16% (resp. 5%) du corpus d'entraînement de CamemBERT (resp. ModernBERT). Cette disproportion justifie pleinement notre choix méthodologique d'une adaptation de domaine plutôt qu'un entraînement de zéro, stratégie qui permet de préserver les capacités généralistes des modèles tout en spécialisant leurs représentations sur la variété linguistique de la presse régionale. Pour mesurer l'impact de la quantité de données pour l'adaptation au domaine, nous avons par ailleurs extrait de ces 15 ans les articles des années 2022 et 2024 pour constituer un second corpus plus restreint, comportant 1M de publications, soit 510M de *tokens*.

Le traitement des entités nommées constitue un enjeu central dans le domaine de la presse, les entités constituant l'ossature de l'information en répondant aux questions fondamentales des « 5W » (*who, what, where, when, why*) qui structurent traditionnellement le récit de presse (Hamborg *et al.*, 2018). La capacité d'un modèle de langue à représenter les acteurs et les lieux est d'autant plus critique dans notre contexte que l'ancrage référentiel des entités régionales (personnalités locales, toponymes infra-régionaux, etc.) est généralement sous-représenté dans les corpus généralistes utilisés pour le pré-entraînement. Pour considérer cet aspect, nous nous sommes appuyés sur le référentiel du journal qui recense les entités nommées apparaissant dans les archives et les articles correspondants afin de créer une liste des entités pertinentes à considérer dans nos modèles adaptés. À l'heure actuelle, le référentiel se compose de 219 467 entités nommées, catégorisées en une dizaine de types. On y trouve principalement des noms de personnes (118 544), de lieux (44 150), de sociétés (14 776), d'associations (2 906), d'institutions (1 850), ainsi que d'autres éléments moins fréquents.

2.2 Adaptation des modèles

Modèles généralistes utilisés. Nous rappelons tout d'abord les caractéristiques principales des deux modèles utilisés dans cette étude. Dérivé de l'architecture RoBERTa (Liu *et al.*, 2019), CamemBERT est entraîné comme modèle de langue masqué pour le français sur 275 milliards de *tokens*. Le modèle ModernBERT est quant à lui une version française de (Warner *et al.*, 2025), entraînée sur 1 000 milliards de *tokens*. Cette architecture est connue pour être plus efficace en temps de calcul que CamemBERT, tant à l'apprentissage qu'à l'inférence, et permet de traiter des contextes plus longs. Les deux modèles s'appuient sur un tokeniseur fondé sur l'algorithme WordPiece avec un vocabulaire de 32 768 *tokens*. Dans les deux cas, les textes utilisés pour l'apprentissage sont en grande partie issus du web et incluent des données journalistiques ou proches du genre journalistique dans une proportion qui n'est *a priori* pas majeure. Nous savons par exemple qu'environ 70 000 articles de Ouest France

figurent dans le corpus OSCAR, ne constituant qu'une très faible proportion des textes vus lors de l'entraînement des modèles généralistes. Quoiqu'il en soit, si des données proches de Ouest France peuvent être présentes dans les données d'apprentissage, il est raisonnable de considérer que ni ces modèles ni leurs tokeniseurs n'y sont spécifiquement adaptés. Les statistiques sur la décomposition des entités nommées en *tokens* présentées dans la section suivante tendent d'ailleurs à confirmer ce point.

Adaptation du vocabulaire. Les entités nommées au cœur de Ouest France, comme le nom des petites communes bretonnes ou encore des figures politiques locales, sont souvent absentes des corpus utilisés pour l'apprentissage des modèles de langue généralistes et des tokeniseurs associés. Cela conduit en général à une segmentation de ces entités en plusieurs *tokens*, ce qui peut nuire aux performances des modèles sur des tâches sensibles aux entités (reconnaissance d'entités nommées, extraction de connaissance). Parmi les 83 891 mots-formes uniques constituant les entités nommées du référentiel, seuls 4,7 % sont tokenisés en un seul *token* par le tokeniseur des modèles généralistes. Les autres sont en moyenne représentées par 2,6 *tokens*, avec un maximum de 8.

Dans des travaux préliminaires, nous avons expérimenté l'apprentissage d'un nouveau tokeniseur fondé sur l'algorithme WordPiece, mais la nécessité de ré-entraîner le modèle de langue généraliste en fait une solution peu pratique et efficace [2]. Nous nous sommes donc tournés vers l'ajout de *tokens* au vocabulaire du modèle initial, en adoptant une stratégie de sélection des nouveaux *tokens* tenant compte des occurrences des entités nommées de tout type dans notre corpus. Plus précisément, nous observons dans un premier temps les 80 000 entités les plus fréquentes dans le corpus, que nous segmentons ensuite par espaces (pour dissocier notamment noms et prénoms) et par tirets (pour dissocier les entités de lieux de type *Saint-...*). Cette procédure aboutit à un ensemble de 40 519 formes distinctes, dont seules 7,3 % sont présentes dans le vocabulaire du tokeniseur : les 37 552 restantes sont ajoutés au vocabulaire. On y retrouve majoritairement des formes correspondant aux entités de lieux (50 % des ajouts) et de personnes (37 %).

Adaptation des modèles. Après adaptation du vocabulaire, nous poursuivons le pré-apprentissage des modèles sur les données de Ouest France dans le cadre d'une tâche de modèle de langage masqué, avec 40 % des *tokens* masqués, taux utilisé dans les pré-entraînements les plus récents comme celui de ModernBERT. Pour ce dernier, nous fixons la longueur maximale des séquences à 2 000 *tokens* pour des raisons d'efficacité, bien que le modèle généraliste puisse en considérer jusqu'à 8 192 : la longueur moyenne des articles étant de 290 *tokens*, limiter à 2 000 n'aboutit à la troncature que de 0,2 % des articles, ce qui est raisonnable par rapport aux gains en terme de calcul.

Pour les modèles entraînés respectivement sur 2 ans et sur 15 ans d'archives, 2 000 et 8 013 articles issus des périodes correspondantes sont réservés à la validation. Concernant l'évaluation, les deux modèles sont testés sur un même ensemble de 1 000 articles extraits du corpus de 15 ans. Initialement, 8 013 articles avaient également été retenus pour le test, effectif que nous avons réduit par la suite afin de limiter le temps d'inférence, sans impact notable sur les résultats obtenus. L'entraînement est réalisé sous PyTorch Lightning en configuration DDP multi-GPU, en précision mixte bf16. L'optimisation repose sur AdamW. Nous mettons en œuvre une stratégie de *learning rate scheduling* linéaire : le taux d'apprentissage suit une phase de *warmup* couvrant 10 % des pas d'entraînement, durant laquelle il croît progressivement jusqu'à 10^{-5} avant de décroître linéairement jusqu'à zéro.

2. Malgré plusieurs approches pour initialiser les paramètres d'un modèle CamemBERT efficacement avec un nouveau tokeniseur en s'appuyant sur le modèle initial, la quantité de données reste insuffisante pour un apprentissage efficace.

2.3 Analyse de la qualité des modèles

Nous nous intéressons ici à la qualité intrinsèque des modèles adaptés aux spécificités des contenus Ouest France. Dans un premier temps, nous considérons la tâche classique de prédiction de mots masqués afin de mettre en évidence les capacités de ces modèles à représenter le style du journal et les entités qu'il utilise. Dans un second temps, nous cherchons à évaluer la connaissance acquise, c'est-à-dire les faits du domaine appris par les modèles.

Évaluation en modèle de langue masqué. Nous rapportons dans le tableau 1 les taux de prédiction correcte des mots masqués des différents modèles (*cf.* légende de la figure) en nous focalisant sur différents types de masques. Il est important de noter que le masquage est effectué sur les mots et non sur les *tokens* : un mot dont la forme n'apparait pas dans le vocabulaire du tokeniseur ne peut donc être prédit correctement puisque nous ne pouvons trouver de *token* correspondant dans les prédictions.

Les deux premières colonnes de résultats correspondent à une stratégie de masquage proche de celle utilisée pour l'apprentissage, en masquant aléatoirement 15 % des mots dans les énoncés avant tokenisation, tous mots confondus pour la première, uniquement les mots qui ne correspondent pas à des entités pour la seconde. Ces résultats montrent clairement que la poursuite de l'apprentissage et la mise à jour du vocabulaire permet aux modèles de s'adapter aux spécificités des contenus du journal de manière globale. La poursuite de l'apprentissage sans adaptation du vocabulaire (ligne 3) ou avec (ligne 4) montre peu de différence sur la prédiction des mots masqués en général : l'adaptation du vocabulaire, avec une augmentation significative de la taille de ce dernier, n'a que peu d'impact sur la prédiction de l'ensemble des mots ; la poursuite du pré-apprentissage permet de mieux capturer les caractéristiques rédactionnelles du journal. Les colonnes suivantes s'intéressent plus particulièrement aux entités nommées et à leur prédiction lorsque seules ces dernières sont masquées dans l'énoncé d'origine. Nous avons fait le choix de masquer 50 % des entités (ou des entités d'un type) d'un énoncé. Ce choix résulte d'un compromis entre rendre la tâche de prédiction suffisamment complexe – masquer 15 % des entités ne présente pas de grandes difficultés en raison des relations fortes existant entre elles – tout en assurant sa faisabilité – masquer l'ensemble des entités ne donne pas suffisamment de contexte pour rendre la tâche réalisable. Nous rapportons tout d'abord les prédictions en masquant aléatoirement des entités sans distinction de type, puis, dans les colonnes suivantes, en ne masquant qu'un type d'entité – resp. lieux, personnes, sociétés qui correspondent aux entités les plus fréquentes. Ces résultats mettent en évidence que l'adaptation du vocabulaire est cruciale pour faciliter la prédiction d'entités, *a priori* utile pour des tâches comme l'extraction d'information, de relations ou encore la détection des entités nommées. On note aussi que la quantité de données, qui a un impact très modéré sur la prédiction des mots en général, est un élément clé pour mieux prédire les entités, ces dernières apparaissant moins fréquemment que les mots en général, en particulier les mots grammaticaux. Enfin, on notera une légère amélioration des taux de prédiction correcte entre CamemBERT et ModernBERT.

Globalement, ces résultats montrent qu'augmenter le vocabulaire pour une meilleure gestion des entités nommées et poursuivre le pré-apprentissage du modèle de langue permet de mieux capturer le style rédactionnel des contenus ainsi que leurs spécificités en terme d'entités présentes.

Évaluation des connaissances relationnelles du modèle. L'amélioration des prédictions sur les *tokens* masqués correspondant à des entités, et l'importance de ces dernières dans le domaine journalistique, nous amènent à nous interroger sur les connaissances que les modèles ont pu acquérir

modèle		*tokens* (15 %)		entités (50 %)			
		tous	hors ent.	toutes	lieux	pers.	sociétés
CamemBERT		51,29	54,29	11,80	15,97	9,96	17,73
ModernBERT		52,68	55,39	11,70	14,96	8,64	16,87
CamemBERT	2 ans	54,02	56,43	14,56	18,35	10,84	22,03
CamemBERT	tok, 2 ans	54,02	56,04	16,74	22,09	15,73	24,27
CamemBERT	tok, 15 ans	55,68	57,41	21,55	27,59	19,48	30,21
ModernBERT	tok, 15 ans	56,90	58,87	21,90	28,38	19,48	29,41

TABLE 1 – Taux de prédiction correcte des *tokens* masqués pour les modèles généralistes (lignes 1-2), CamemBERT adapté sur 2 ans avec son tokeniseur d'origine (ligne 3), les modèles adaptés avec le vocabulaire augmenté des entités (lignes 4-6).

relation	#faits	#occ	*prompt*
(X, situé dans, Y)	462	1121	X est une commune située dans le département suivant : [MASK].
(X, a lieu à, Y)	105	127	Le festival X se déroule à [MASK].
(X, maire de, Y)	2 563	45	X est maire de [MASK].

TABLE 2 – Statistiques et *prompt* des trois relations testées.

avec la poursuite du pré-entrainement. En effet, les modèles de langue sont connus pour emmagasiner des connaissances sur des faits mentionnés dans les contenus utilisés pour l'apprentissage (Petroni *et al.*, 2019; Cao *et al.*, 2021; Heinzerling & Inui, 2021), permettant de les utiliser dans une certaine mesure comme base de connaissance. Dès lors, la question se pose de savoir dans quelle mesure nos modèles adaptés sont porteurs d'une meilleure connaissance du domaine que les modèles généralistes.

Nous nous intéressons ici aux connaissances factuelles, typiquement formalisées sous la forme d'un triplet (sujet, prédicat, objet) qui permet facilement de mobiliser des méthodes de *prompting* pour tester leur présence ou non dans un modèle de langue (Petroni *et al.*, 2019; AlKhamissi *et al.*, 2022; Zhong *et al.*, 2021). Nous avons testé nos modèles adaptés pour trois relations (prédicats) choisies en raison de la fréquence des mentions des faits (triplets) correspondants dans le corpus et rapportées dans le tableau 2, où #faits est le nombre de triplets pour un prédicat, et #occ le nombre médian de co-occurrences du sujet et de l'objet d'un fait au sein d'une même phrase dans le corpus. Par exemple, la relation "maire de" (ligne 3) s'instancie dans 2 563 triplets dont le nombre de mentions médian est 45. Si la mention de cette relation dans les articles du journal est fréquente, chaque triplet est très peu mentionné ce qui rend sa mémorisation par le modèle de langue difficile. À l'inverse, la relation "situé dans" (ligne 1) n'a que peu d'instances, chacune apparaissant fréquemment – en particulier pour les principales communes comme Rennes, Brest ou encore Caen. Enfin, la relation concernant le lieu dans lequel un festival récurrent se déroule présente un intermédiaire entre ces deux extrêmes.

Nous mesurons les connaissances acquises par un modèle à l'aide d'énoncés décrivant un triplet appartenant à une relation et dont l'objet est masqué. Les énoncés utilisés pour chacune de nos relations sont donnés dans la dernière colonne du tableau 2, où X est remplacé par le sujet pour un fait donné. Le tableau 3 donne le taux de prédiction correcte de l'objet, prédit comme le *token* le plus probable selon la distribution de probabilité sur le vocabulaire donnée par le modèle. L'amélioration dépend fortement du nombre d'occurrences moyen des faits dans les données d'adaptation, la relation

	modèles généralistes		modèles adaptés	
	CamemBERT	ModernBERT	CamemBERT	ModernBERT
(X, situé dans, Y)	17,7	17,8	66,9	70,5
(X, a lieu à, Y)	1,8	5,5	16,1	14,3
(X, maire de, Y)	0	0	0,1	0,07

TABLE 3 – Taux de prédiction correct (en %) de l'objet pour chacune des relations.

"maire de" étant un cas extrême dans lequel les modèles adaptés n'arrivent que très marginalement à une prédiction correcte de l'objet masqué. Pour les deux autres relations, l'adaptation aux spécificités de Ouest France permet une amélioration significative de la prédiction de l'objet pour les deux modèles avec, là encore sans surprise, un gain plus significatif pour la relation "situé dans" où chaque fait est mentionné en moyenne un nombre suffisant de fois. Il est à noter qu'une partie du gain est imputable à l'adaptation du vocabulaire : la tâche consistant à prédire un seul *token* pour le masque, les entités dont la forme n'apparaît pas dans le vocabulaire du tokeniseur ne peuvent être prédites correctement [3].

Nous avons également testé une approche fondée sur l'apprentissage de *prompt* (*soft prompting*), connue pour pallier la sensibilité à l'énoncé de la technique précédente et sa difficulté à faire ressortir des faits rares dans les données d'apprentissage. En s'appuyant sur l'algorithme OptiPrompt (Zhong *et al.*, 2021), nous avons une amélioration globale des performances pour la relation "situé dans", avec, pour ModernBERT, une précision de 41,7 % pour le modèle généraliste et de 93,2 % pour le modèle adapté. Pour la relation difficile "maire de", l'apprentissage de *prompts* n'apporte en revanche aucun gain, la tâche restant toujours impossible. Il paraît naturel que, quelle que soit la technique de *prompting* utilisée, il faille un seuil minimum d'occurrences pour que le modèle de langue « mémorise » un fait.

3 Adaptation à des tâches canoniques

Les résultats précédents montrent que l'adaptation au domaine de modèles généralistes capture efficacement les spécificités de Ouest France, notamment ses entités, tant en modélisation du langage qu'en extraction de relations, témoignant de l'acquisition de certains faits fréquents. Nous évaluons désormais l'impact de l'adaptation au domaine pour des tâches essentielles à l'exploitation et à l'enrichissement des archives de presse, à savoir la classification d'articles, la détection des entités nommées et la recherche d'information (du moins, la comparaison d'articles). Ces tâches présentent en outre l'intérêt de s'inscrire à des échelles différentes (article, mot, collection) et peuvent donc être impactées par l'adaptation des modèles de manières différentes.

3.1 Classification thématique en régime de données limitées

Pour la tâche de classification de textes, nous exploitons un corpus de 4 378 articles, manuellement annotés selon 337 catégories thématiques dérivées de la taxonomie IPTC, largement employée pour

3. Nous avons testé de nombreuses stratégies pour prédire plusieurs *tokens* pour un masque, avec peu de succès en raison de la difficulté à prédire le nombre de *tokens* à prédire.

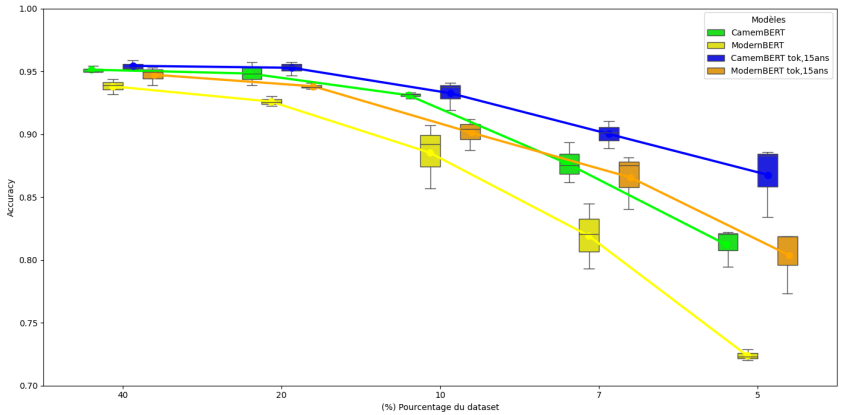

FIGURE 1 – Distribution des taux de classification correcte (3 apprentissages sur des sous-échantillons différents) en fonction de la quantité de données d'apprentissage pour la classification de texte.

caractériser la presse. Notre étude se focalise sur une sélection de 20 catégories fréquentes, avec un nombre d'articles par catégorie allant de 174 (centre de loisirs) à 417 (élections européennes). Des expériences préliminaires ayant mis en évidence un effet de saturation (performances *quasi* optimales) pour l'ensemble des modèles lorsque l'intégralité du corpus est exploitée, nous avons orienté notre évaluation vers un protocole évaluant l'efficacité des modèles en régime de données limité, un scénario très fréquent dans l'exploitation des archives à des fins variées. Nous avons ainsi artificiellement et progressivement réduit la taille du corpus d'apprentissage, en faisant varier la proportion des données disponibles de 100 % jusqu'à 5 % – soit entre 8 et 20 articles par classe dans ce dernier cas. Pour assurer la robustesse statistique malgré la rareté, trois apprentissages sont réalisés sur des sous-échantillons différents, le jeu de test restant constant.

Pour l'entraînement nous appliquons une tête de classification à la représentation du *token* CLS, composée de deux couches linéaires successives pour CamemBERT et d'une couche linéaire suivie d'une fonction d'activation et d'une normalisation de type *LayerNorm* pour ModernBERT. Le modèle est entraîné sur 25 *epochs* avec une taille de *batch* de 10, en utilisant l'optimiseur *AdamW*. Le *learning rate* initial est fixé à 10^{-5} et suit un *scheduler* linéaire avec 10 % de *warm-up*.

Les résultats, figure 1, montrent une amélioration des modèles adaptés au domaine lorsque la quantité de données d'apprentissage pour la tâche diminue, l'écart avec les modèles généralistes se creusant au fur et à mesure que la quantité de données d'entraînement baisse.

3.2 Reconnaissance d'entités nommées

Pour la détection des entités nommées, 54 000 articles issus des archives ont été annotés à l'aide du modèle *zero-shot* GliNER (Zaratiana *et al.*, 2024) selon 23 catégories [4], dont quelques unes correspondant aux entités considérées dans la procédure d'adaptation des modèles (personne, organisation, lieu). La tête de classification en étiquette IOB des *tokens* consiste en une une projection linéaire

4. personne, organisation, lieu, date, heure, nombre, argent, produit, événement, loi, maladie, substance chimique, livre, film, logiciel, véhicule, pays, ville, université, parti politique, compétitions, équipes et récompense

	modèles généralistes		modèles adaptés	
	CamemBERT	ModernBERT	CamemBERT	ModernBERT
reconnaissance des entités	84,2	82,1	83,8	82,8
RI, encodage conjoint	97,6	98,0	97,6	98,3
RI, encodage séparé	91,5	88,6	91,7	86,3

TABLE 4 – Résultats pour les tâches de reconnaissance des entités nommées (ligne 1, F-score), et de recherche d'information (lignes 2-3, taux de classification correct).

pour CamemBERT, et suit pour ModernBERT la même architecture que pour la classification. Les modèles sont entraînés pendant 10 *epochs* avec une taille de *batch* de 8, les autres paramètres restant inchangés par rapport à la section 3.1.

Le tableau 4 (haut) présente les scores F1 évalués au niveau des entités complètes, une entité étant considérée comme correcte si ses frontières recouvrent exactement les *tokens* correspondant de la référence. Outre le fait que ModernBERT est moins bon que CamemBERT, l'impact de l'adaptation du modèle est négligeable sur la prédiction des entités, avec une très légère baisse pour CamemBERT et une très légère amélioration pour ModernBERT. Ce résultat est en partie surprenant car on est en droit d'espérer de l'ajout d'entités au vocabulaire qu'il facilite la prédiction des étiquettes IOB correspondant à ces entités, ce qui n'est globalement pas le cas. Une explication possible de cela est que la prédiction d'étiquettes au niveau du *token* est relativement robuste quelle que soit la tokenisation adoptée pour les entités et ne dépend au final que peu des entités elles-mêmes. Ce constat amène à réfléchir sur la manière de formuler la tâche de détection des entités nommées pour qu'elle puisse bénéficier de la connaissance des entités par le modèle de langue, ainsi que des relations entre entités implicitement apprises.

3.3 Recherche d'information

Pour évaluer la performance des modèles sur une tâche proche de la recherche d'information (RI), nous exploitons le corpus initialement développé dans le cadre du challenge *EvalLLM2025* (Rousseau *et al.*, 2025). Dédié aux domaines de la politique étrangère et de la défense, ce corpus agrège 1 712 articles provenant de trois sources de données complémentaires et parus entre janvier 2022 et mai 2025. Afin de formuler la tâche sous forme d'apprentissage supervisé, des paires question-document ont été générées automatiquement – *cf*. (Rousseau *et al.*, 2025) pour les détails – et le corpus distingue les paires pertinentes des paires non pertinentes, obtenues par échantillonnage négatif. Le jeu de données final comprend un total de 28 543 paires positives et de 28 381 paires négatives.

Nous comparons deux architectures couramment utilisées en RI. La première se fonde sur un encodage conjoint de la question et de l'article, avec l'ajout d'un *token* de séparation entre les deux. La représentation de la paire question et article est obtenue par moyennage des vecteurs en sortie du modèle (*average pooling*) et sert d'entrée à un classifieur binaire. La seconde se fonde sur un encodage séparé de la question et d'un article, sans interaction entre les deux afin de permettre le pré-calcul des descriptions des articles. Les représentations de la question, q, et de l'article, a, sont obtenues en moyennant les vecteurs en sortie des encodeurs respectifs, puis combinées pour l'étape de classification binaire. La combinaison des deux représentations q et a est donnée par la concaténation de q, a, $|q - a|$ et $q \odot a$, où $|q - a|$ est la valeur absolue de la différence et $q \odot a$ le produit de Hadamard. Empiriquement, cette combinaison s'est révélée plus efficace que l'utilisation

d'une fonction de similarité ou la simple concaténation de q et a.

Le tableau 4 (bas) donne les taux de classification correcte des paires question/article pour chacun des modèles avec les deux stratégies d'encodage, conjoint ou disjoint. Comme pour les tâches précédentes, nous n'observons pas d'amélioration significative des performances avec les modèles adaptés, quelque soit l'encodage utilisé. Nous avons également testé ces modèles sur une tâche plus proche de la recherche d'information, visant à ordonner les articles en fonction de leur pertinence pour apporter une réponse à la question. La pertinence est obtenue avec les scores des modèles appris pour la classification binaire, interprétés ici comme (une estimation de) la probabilité que l'article contienne la réponse à la question. Sur cette dernière tâche, les modèles adaptés sont même légèrement moins performants que les modèles généralistes. Pour CamemBERT par exemple, le rang réciproque moyen des articles pertinents décroît de 0,69 à 0,66 avec l'adaptation, cette dernière classant moins d'articles pertinents en première position (55 % contre 58 %).

4 Discussion

L'ensemble des résultats présentés dans cet article montre clairement que l'adaptation de modèles de langue généralistes bidirectionnels à la presse régionale, qui sans être un domaine de spécialité présente des singularités fortes au niveau des entités et faits relatés, permet d'obtenir des modèles de langue plus proches des contenus. Cette adaptation, y compris celle du vocabulaire, s'avère donc importante dans notre cadre, tout comme elle l'est pour les domaines de spécialité, et requiert une quantité importante de données. On met également en évidence que si les modèles de langue adaptés sont meilleurs à prédire les entités et qu'ils ont pu acquérir un certain nombre de faits, cela n'apporte de manière générale pas de bénéfice significatif pour réaliser des tâches canoniques, sauf lorsque les données d'apprentissage pour la tâche ciblée sont en quantité très limitée ; ce dernier résultat, mis en évidence sur la classification de textes, demande cependant à être généralisé à d'autres tâches.

Généralement, si les modèles de langue adaptés ont acquis de la connaissance sur le domaine, ils n'ont pas acquis la compétence qui leur permet de mieux réaliser les tâches sur lesquels on souhaite les spécialiser. Ces constatations soulèvent plusieurs questionnements. Il est en règle générale admis, du moins dans les domaines de spécialité, qu'adapter les modèles de langue généralistes au domaine avant de les spécialiser sur des tâches est bénéfique. Si on peut arguer que la presse régionale n'est pas un domaine de spécialité, nos résultats soulèvent la question de savoir dans quels contextes l'adaptation de domaine des modèles généralistes est bénéfique pour les rendre compétents, sans nécessité de recourir à une approche empirique en adaptant le modèle pour voir les bénéfices éventuels. Notons pour alimenter ce débat que plusieurs travaux (*preprints*) font des constats similaires au nôtre dans des domaines de spécialité (Chen *et al.*, 2025; Geng *et al.*, 2021), le second montrant également que « *domain adaptive pre-training is only helpful with low-resource downstream tasks* ». Les résultats mis en avant dans ce retour d'expériences interrogent aussi sur la manière de formaliser des tâches comme la détection des entités et la recherche d'information, qui devraient facilement pouvoir bénéficier de la connaissance implicitement acquise lors de l'adaptation des modèles généralistes, en particulier sur les entités du domaine. Des modèles génératifs plus complexes devraient pouvoir apporter une réponse à cette question (Wu *et al.*, 2024), en mobilisant des approches de classification sans exemple (Wei *et al.*, 2022), au prix cependant d'un coût calculatoire nettement plus élevé et de la nécessité de disposer de plus de données d'adaptation. Certains scénarios ne peuvent cependant se permettre ce coût et appellent à des alternatives avec des modèles plus petits et tout aussi performants.

Références

ALKHAMISSI B., LI M., CELIKYILMAZ A., DIAB M. & GHAZVININEJAD M. (2022). A review on language models as knowledge bases. DOI : https ://doi.org/10.48550/arXiv.2204.06031.

ANTOUN W., KULUMBA F., TOUCHENT R., ÉRIC DE LA CLERGERIE, SAGOT B. & SEDDAH D. (2024). CamemBERT 2.0: A smarter French language model aged to perfection. DOI : https ://doi.org/10.48550/arXiv.2411.08868.

ANTOUN W., SAGOT B. & SEDDAH D. (2025). ModernBERT or DeBERTaV3? Examining architecture and data influence on transformer encoder models performance. In *Proc. Intl. Joint Conference on Natural Language Processing and the the Asia-Pacific Chapter of the ACL*, p. 3061–3074.

CAO B., LIN H., HAN X., SUN L., YAN L., LIAO M., XUE T. & XU J. (2021). Knowledgeable or educated guess ? Revisiting language models as knowledge bases. In *Proc. Annual Meeting of the Association for Computational Linguistics and Intl. Joint Conference on Natural Language Processing*, p. 1860–1874.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : Preparing the muppets for court. In *Proc. Conf. Empirical Methods in Natural Language Processing*, p. 2898–2904.

CHEN P.-E., LIAN D.-C., HSIEH S.-K., HUANG S.-C., SHAO H.-L., CHIU J.-W., LIN Y.-H., CHEN Z.-C., HUANG E. T., SEE S. *et al.* (2025). Continual pre-training is (not) what you need in domain adaption. DOI : https ://doi.org/10.48550/arXiv.2504.13603.

GENG S., LEBRET R. & ABERER K. (2021). Legal transformer models may not always help. DOI : https ://doi.org/10.48550/arXiv.2109.06862.

GUO Y., FU J., ZHANG H. & ZHAO D. (2025). Efficient domain continual pretraining by mitigating the stability gap. In *Proc. Annual Meeting of the Association for Computational Linguistics*, p. 32850–32870.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360.

HAMBORG F., LACHNIT S., SCHUBOTZ M., HEPP T. & GIPP B. (2018). Giveme5w: Main event retrieval from news articles by extraction of the five journalistic W questions. In G. CHOWDHURY, Éd., *Transforming Digital Worlds : 13th International Conference, iConference 2018*, volume 10766 de Lecture Notes in Computer Science, p. 356–366.

HEINZERLING B. & INUI K. (2021). Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proc. Conf. of the European Chapter of the Association for Computational Linguistics*, p. 1772–1791.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert: Unsupervised language model pre-training for French. In *Proc. Language Resources and Evaluation Conference*, p. 2479–2490.

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. DOI : https ://doi.org/10.48550/arXiv.1907.11692.

PETRONI F., ROCKTÄSCHEL T., RIEDEL S., LEWIS P., BAKHTIN A., WU Y. & MILLER A. (2019). Language models as knowledge bases ? In *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Intl. Joint Conference on Natural Language Processing*, p. 2463–2473.

ROUSSEAU I., PERROUX C., ADAM P., GIRAULT T., DELPHIN-POULAT L., VEYRET M., LECORVÉ G. & DAMNATI G. (2025). O_FT@EvalLLM2025 : étude comparative de choix de données et de stratégies d'apprentissage pour l'adaptation de modèles de langue à un domaine. In *Proc. atelier EvalLLM 2025*.

SUN Y., WANG S., LI Y., FENG S., TIAN H., WU H. & WANG H. (2020). Ernie 2.0: A continual pre-training framework for language understanding. In *Proc. AAAI conference on artificial intelligence*.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T. *et al.* (2025). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proc. Annual Meeting of the Association for Computational Linguistics*, p. 2526–2547.

WEI J., BOSMA M., ZHAO V., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2022). Finetuned language models are zero-shot learners. In *Proc. International Conference on Learning Representations*.

WU C., LIN W., ZHANG X., ZHANG Y., XIE W. & WANG Y. (2024). PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, **31**(9), 1833–1843.

WU T., CACCIA M., LI Z., LI Y.-F., QI G. & HAFFARI G. (2022). Pretrained language model in continual learning: A comparative study. In *Proc. International Conference on Learning Representations*.

XIE Y., AGGARWAL K. & AHMAD A. (2024). Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics*, p. 10184–10201.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2024). GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics*, p. 5364–5376.

ZHONG Z., FRIEDMAN D. & CHEN D. (2021). Factual probing is [MASK]: Learning vs. learning to recall. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 5017–5033.