Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoan Dupont, Laurent Romary, Eric Villemonte de La Clergerie, Benoît Sagot, Djamé Seddah

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Les modèles de langue contextuels CamemBERT pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement

Louis Martin*[1,2,3] Benjamin Muller*[2,3] Pedro Javier Ortiz Suárez*[2,3]
Yoann Dupont[3] Laurent Romary[2] Éric Villemonte de la Clergerie[2]
Benoît Sagot[2] Djamé Seddah[2]
[1]Facebook AI Research, Paris, France
[2]Inria, Paris, France
[3]Sorbonne Université, Paris, France
louismartin@fb.com,yoa.dupont@gmail.com,prenom.nom@inria.fr.

Résumé

Les modèles de langue neuronaux contextuels sont désormais omniprésents en traitement automatique des langues. Jusqu'à récemment, la plupart des modèles disponibles ont été entraînés soit sur des données en anglais, soit sur la concaténation de données dans plusieurs langues. L'utilisation pratique de ces modèles — dans toutes les langues sauf l'anglais — était donc limitée. La sortie récente de plusieurs modèles monolingues fondés sur BERT (Devlin *et al.*, 2019), notamment pour le français, a démontré l'intérêt de ces modèles en améliorant l'état de l'art pour toutes les tâches évaluées. Dans cet article, à partir d'expériences menées sur CamemBERT (Martin *et al.*, 2019), nous montrons que l'utilisation de données à haute variabilité est préférable à des données plus uniformes. De façon plus surprenante, nous montrons que l'utilisation d'un ensemble relativement petit de données issues du web (4Go) donne des résultats aussi bons que ceux obtenus à partir d'ensembles de données plus grands de deux ordres de grandeurs (138Go).

Abstract

CamemBERT Contextual Language Models for French: Impact of Training Data Size and Heterogeneity

Contextual word embeddings have become ubiquitous in Natural Language Processing. Until recently, most available models were trained on English data or on the concatenation of corpora in multiple languages. This made the practical use of models in all languages except English very limited. The recent release of monolingual versions of BERT (Devlin *et al.*, 2019) for French established a new state-of-the-art for all evaluated tasks. In this paper, based on experiments on CamemBERT (Martin *et al.*, 2019), we show that pretraining such models on highly variable datasets leads to better downstream performance compared to models trained on more uniform data. Moreover, we show that a relatively small amount of web crawled data (4GB) leads to downstream performances as good as a model pretrained on a corpus two orders of magnitude larger (138GB).

Mots-clés : Modèles de langue contextuels, BERT, CamemBERT, impact jeu de données.

Keywords: Contextual language models, BERT, CamemBERT, dataset impact.

∗. Les trois premiers auteurs ont contribué à parts égales à ce travail

Volume 2 : Traitement Automatique des Langues Naturelles, pages 54–65.
HAL : hal-02784755.

1 Introduction

En préface à son *Introduction to Deep Learning*, Charniak (2019) évoque son scepticisme initial face à la révolution apportée par l'apprentissage profond de réseaux neuronaux au traitement automatique des langues :

> « (..) I can rationalize this since this is the third time neural networks have threatened a revolution but only the first time they have delivered. *(Charniak, 2019, page* XI*)* »

En effet, la surprise apportée par l'avènement des plongements lexicaux et le gain de performance qu'ils ont permis en peu de temps (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Mikolov *et al.*, 2018) n'a eu pour équivalent que le saut qualitatif apporté par la prise en compte du contexte dans les représentations vectorielles, permettant *de facto* une prise en charge effective de la polysémie et donc l'obtention de modèles plus efficaces et plus fins (Peters *et al.*, 2018; Akbik *et al.*, 2018). Ces avancées ont ouvert la voie à des modèles contextuels plus larges, entraînés sur des objectifs de modèles de langue (Dai & Le, 2015). Ces approches, qui reposaient au départ sur des architectures LSTM (Howard & Ruder, 2018), ont évolué vers des architecturs de type *Transformer*, avec notamment GPT2 (Radford *et al.*, 2019), BERT (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019) et plus récemment ALBERT (Lan *et al.*, 2019) et T5 (Raffel *et al.*, 2019).
Bien que plusieurs modèles développés pour d'autres langues aient été publiés (modèles ELMo [1] pour le japonais, le portugais, l'allemand et le basque ; modèles BERT pour le chinois simplifié et classique (Devlin *et al.*, 2018) ou pour l'allemand (Chan *et al.*, 2019)), le différentiel quant à la taille de leurs données de pré-entraînement n'a pas permis l'émergence de travaux les comparant au modèle original. Cependant, des modèles multilingues reposant sur la concaténation de larges jeux de données (principalement basés sur Wikipedia) sont apparus (Devlin *et al.*, 2018; Conneau *et al.*, 2019) et ont permis des avancées notables via l'apprentissage par transfert (Pires *et al.*, 2019). Ce n'est toutefois que très récemment que des modèle monolingues à grande échelle ont été développés (Martin *et al.*, 2019; Le *et al.*, 2019; Virtanen *et al.*, 2019; Delobelle *et al.*, 2020) et ont permis de confirmer l'intérêt des modèles monolingues sur d'autres langues.

En ce qui concerne le français, Le *et al.* (2019) ont montré sur diverses tâches que leur modèle, FlauBERT, offrait un panel de performances équivalentes à celles de CamemBERT (Martin *et al.*, 2019), soulignant qui plus est la complémentarité des deux modèles sur des tâches d'analyse syntaxique. Sachant que ces modèles ont été entraînes sur des données *in fine* différentes bien que d'origine similaire (avec un filtrage plus intense et l'utilisation d'un équivalent francophone du *Bookcorpus* dans un cas, un filtrage principalement sur le bruit et l'identification de la langue cible dans l'autre), il est pertinent de s'interroger sur l'impact qu'ont les données de pré-entraînement, tant en termes de taille que de type de données, sur les performances des modèles de langue neuronaux contextuels. D'autres paramètres sont d'importance, en particulier la stratégie de *masking* utilisée (*subword* ou *whole-word* ?) et le nombre de couches et de têtes d'attention (modèle *Base* ou *Large* ?).

Nous présentons ici une série d'expériences construites autour de CamemBERT visant à répondre à ces questions. Nos résultats montrent que, contrairement à l'idée qui prévalait, il est possible d'obtenir des résultats étonnement bons, au niveau de l'état de l'art pour toutes les tâches ou presque, avec des modèles entraînés sur seulement 4Go de données. Le point essentiel est qu'il semble préférable d'utiliser des données à haute variabilité, éventuellement bruitées, plutôt que des données proprement éditées et stylistiquement homogènes telles qu'on peut en trouver dans des jeux de données tirés de Wikipedia. Ce résultat permet d'envisager l'entraînement de ce type de modèles pour

1. https://allennlp.org/elmo

des langues relativement peu dotées voire pour des domaines spécialisés, dans les cas où une stratégie de *fine-tuning* ne serait pas efficace.

2 Protocole expérimental

Nous reprenons le même environnement expérimental (paramètres, outils, métriques, etc.) que celui utilisé par Martin *et al.* (2019).

2.1 Modèles et architectures

BERT, RoBERTa et CamemBERT CAMEMBERT est basée sur ROBERTA (Liu *et al.*, 2019), une évolution de BERT (Devlin *et al.*, 2019) sur plusieurs plans, notamment par l'utilisation du *masked language model* comme seul objectif de pré-entrainement. Outre le modèle CAMEMBERT$_{\text{BASE}}$ originel entraîné avec 12 couches, 768 dimensions cachées et 12 têtes d'attention, soit 110M de paramètres, nous utilisons un CAMEMBERT$_{\text{LARGE}}$ entrainé avec 24 couches, 1024 dimensions cachées et 16 têtes d'attention, soit 340M paramètres.
Selon les expériences, nous évaluons nos modèles en fonction de plusieurs hyper-paramètres : (i) la stratégie de *masking* (*subword* ou *whole word*), (ii) l'architecture du modèle (*BASE* ou *LARGE*), (iii) le nombre d'étapes d'entraînement (steps) et (iv) l'utilisation du modèle via *fine-tuning* ou via l'extraction de plongements lexicaux.

Données d'entraînement Pour étudier l'impact des données d'entraînement sur les performances de CAMEMBERT, nous utilisons alternativement le sous-corpus français du corpus multilingue OSCAR extrait de Common Crawl (Ortiz Suárez *et al.*, 2019), un autre corpus extrait de Common Crawl nommé CCNET (Wenzek *et al.*, 2019) et un snapshot récent de la Wikipedia française.

— **OSCAR** (Ortiz Suárez *et al.*, 2019) est un ensemble de corpus monolingues extraits de Common Crawl (*dump* de novembre 2018). Les corpus ont été sélectionnés par un modèle de classification par langues en suivant l'approche de (Grave *et al.*, 2018) s'appuyant sur le classifieur linéaire FASTTEXT (Grave *et al.*, 2017; Joulin *et al.*, 2016) pré-entraîné sur les corpus Wikipedia, Tatoeba et SETimes, et couvrant 176 langues.
— **CCNet** (Wenzek *et al.*, 2019), un jeu de données extrait lui aussi de Common Crawl mais avec un filtrage différent de celui d'OSCAR. Il a été construit avec un modèle de langue entraîné sur Wikipedia, lui permettant ainsi de filtrer le bruit (code, tables, etc.). CCNET contient ainsi des documents plus longs en moyenne qu'OSCAR. Ce filtrage a pour effet de biaiser les données en leur donnant un aspect « Wikipedia » et nous permet de considérer CCNET comme se positionnant entre OSCAR, peu filtré voire bruité, et WIKIPEDIA, totalement édité.
— **Wikipedia**, un corpus homogène en termes de genre et de style. Nous utilisons le *dump* français officiel de Wikipedia (avril 2019). Le corpus est prétraité à l'aide de *WikiExtractor*[2].

Afin de pouvoir comparer équitablement l'impact du type de données de pré-entraînement, nous créons des échantillons aléatoires à partir de OSCAR et CCNET, et ce au niveau du document, de la même taille que celle de notre WIKIPEDIA, soit 4Go de texte brut non compressé. Ceci nous permet d'étudier également les effets de la taille des données d'entraînement sur les performances des modèles.

2. https://github.com/attardi/wikiextractor

Jeux de données et tâches d'évaluation Nous évaluons nos différents modèles en étiquetage morphosyntaxique, en analyse sytaxique, en reconnaissance d'entités nommées (NER) et en reconnaissance d'implication textuelle (*Natural Language Inference*, NLI), qui consiste à prédire la relation entre une phrase hypothèse et phrase prémisse (implication, contradiction, neutralité). Pour les évaluations en étiquetage morphosyntaxique (POS tagging) et analyse en dépendances (parsing), nous utilisons dans leurs versions *Universal Dependencies 2.2* (Nivre *et al.*, 2018) les corpus Sequoia (Candito & Seddah, 2012), UD French GSD, UD French Spoken et UD French ParTut. L'évaluation de la NER se fait sur l'instance du *French treebank* (Abeillé *et al.*, 2003) annotée en entités nommées par Sagot *et al.* (2012). Pour la tâche de NLI, nous utilisons la partie française du jeu de données XNLI (Conneau *et al.*, 2018) qui étend le corpus *Multi-Genre NLI)* (Williams *et al.*, 2018) [3].

Toutes nos expériences suivent les *splits* usuels et utilisent les métriques classiques associées à ces tâches (UPOS, LAS, F1 et exactitude) . La Table 1 présente des statistiques sur ces jeux de données.

Corpus	Taille (texte brut non compr.)	#tokens	#docs	tokens/doc quantiles : 5%	50%	95%
Wikipedia	4Go	990M	1.4M	102	363	2530
CCNet	135Go	31.9B	33.1M	128	414	2869
OSCAR	138Go	32.7B	59.4M	28	201	1946

TABLE 1 – Statistiques sur les jeux de données de pré-entrainement.

Corpus	#tokens	#phrases	Genres
GSD	389,363	16,342	Blogs, News Reviews, Wiki
Sequoia	68,615	3,099	Medical, News Non-fiction, Wiki
Spoken	34,972	2,786	Spoken
ParTUT	27,658	1,020	Legal, News, Wikis
FTB	350,930	27,658	News

TABLE 2 – Statistiques des corpus arborés utilisés en étiquetage morphosyntaxique, analyse en dépendance et NER.

2.2 Utilisation de CAMEMBERT pour des tâches en aval

Nous utilisons CAMEMBERT de deux façons. Dans la première, *fine-tuning*, nous affinons le modèle sur une tâche spécifique de bout en bout. Dans la seconde, nous extrayons de CAMEMBERT des plongements lexicaux contextuels figés. Les performances de ces deux approches complémentaires illustrent la qualité des représentations cachées que capture CAMEMBERT.

Fine-tuning Pour chaque tâche, nous ajoutons la couche prédictive pertinente au-dessus du modèle de CAMEMBERT. Suite au travail effectué sur BERT (Devlin *et al.*, 2019) en étiquetage de séquence, nous ajoutons une couche linéaire qui prend respectivement en entrée la dernière représentation cachée du token spécial <s> et la dernière représentation cachée du premier token de sous-mot de chaque mot. Pour l'analyse de dépendance, nous branchons une tête de prédiction de graphes *bi-affine* inspirée de Dozat & Manning (2017). Nous renvoyons le lecteur à cet article pour plus de détails sur ce module. Nous affinons CAMEMBERT sur XNLI en ajoutant une tête de classification composée d'une couche cachée avec une non-linéarité et une couche de projection linéaire, avec un dropout d'entrée pour chaque couche.

Nous affinons CAMEMBERT indépendamment pour chaque tâche et chaque ensemble de données. Nous optimisons le modèle en utilisant l'optimiseur Adam (Kingma & Ba, 2014) avec un taux d'apprentissage fixe. Nous effectuons une *grid-search* sur une combinaison de taux d'apprentissage

3. Seules les parties de validation et de test ont été manuellement traduites de l'anglais, la partie d'entraînement l'a été automatiquement (122k exemples d'entraînement, 2490 de développement et 5010 de test).

et de tailles de lots. Nous sélectionnons le meilleur modèle sur l'ensemble de validation parmi les 30 premières *epoch*. Pour la tache de NLI, nous utilisons les hyper-paramètres par défaut fournis par les auteurs de RoBERTa sur la tâche MNLI. [4] Bien que cela aurait pu encore accroître les performances, nous n'appliquons aucune technique de régularisation telle que le *weight decay*, *learning rate warm-up* ou un affinage discriminant, sauf dans le cas de NLI. En effet, les expériences de Martin *et al.* (2019) ont montré que ce n'était pas nécessaire étant donné qu'un affinage simple de CamemBERT a contribué à établir l'état de l'art sur toute les taches et surpasse les modèles BERT multilingues. [5] Les expériences d'étiquetage morpho-syntaxique, d'analyse syntaxique en dépendance et de reconnaissance d'entités nommées sont exécutées à l'aide de la bibliothèque Transformer d'HuggingFace étendue pour prendre en charge CamemBERT et l'analyse de dépendance (Wolf *et al.*, 2019). Les expériences NLI utilisent la bibliothèque FairSeq reposant sur l'implémentation de RoBERTa.

Plongements lexicaux Suivant en celà Straková *et al.* (2019) et Straka *et al.* (2019) pour mBERT et le BERT originel, nous utilisons aussi CamemBERT dans scénario d'extraction de plongements lexicaux. Afin d'obtenir une représentation pour un token donné, nous calculons d'abord la moyenne des représentations de chaque sous-mot dans les quatre dernières couches du Transformer, puis faisons la moyenne des vecteurs des sous-mot résultants.

Nous évaluons CamemBERT dans cette utilisation sous forme de plongements lexicaux dans des taches d'étiquetage morpho-syntaxique, d'analyse de dépendance et en NER, avec les implémentation open-source de Straková *et al.* (2019) et Straka *et al.* (2019) entraînés sur les jeux de données décrit auparavant. [6]

3 Facteurs influençant les performances des modèles

Dans cette section, nous étudions l'influence de plusieurs facteurs sur les performances des tâches aval. Dans ce but, nous produisons plusieurs versions de CamemBERT en faisant varier les données de pré-entraînement. Sauf indication contraire, nous utilisons l'architecture BASE et fixons le nombre d'étapes de pré-entraînement à 100k et permettons alors au nombre d'*epochs* de varier en conséquence (plus d'*epochs* pour des tailles de jeu de données plus petites).

3.1 Common Crawl vs. Wikipedia

Les résultats présentés à la Table 3 montrent que les modèles entraînés sur les versions réduites (4Go) d'OSCAR et de CCNet (issus tous deux de Common Crawl) obtiennent des performances constamment supérieures à celles du modèle entraîné sur Wikipedia, que l'on utilise les modèles en configuration *fine-tuning* ou comme sources de plongements lexicaux. Sans surprise, l'écart est plus grand sur les tâches impliquant des textes dont le genre et le style sont plus éloignés de Wikipédia, notamment pour l'étiquetage et l'analyse syntaxique du corpus French Spoken (transcriptions de

4. Voir https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md pour plus détails.
5. Résultat confirmé ensuite dans plusieurs travaux décrivant des modèles BERT monolingues, eg. (Le *et al.*, 2019).
6. UDPipe Future est disponible sur https://github.com/CoNLL-UD-2018/UDPipe-Future, et le code pour le *nested NER* est disponible sur https://github.com/ufal/acl2019_nested_ner.

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		**Average**		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55
Plongements lexicaux (avec UDPipe Future (tagging, parsing) ou LSTM+CRF (NER))													
Wiki	4GB	98.09	92.31	98.74	93.55	96.24	78.91	95.78	89.79	97.21	88.64	91.23	-
CCNet	4GB	**98.22**	**92.93**	99.12	94.65	97.17	**82.61**	**96.74**	89.95	97.81	90.04	**92.30**	-
OSCAR	4GB	98.21	92.77	99.12	**94.92**	97.20	82.47	**96.74**	**90.05**	**97.82**	**90.05**	91.90	-
OSCAR	138GB	98.18	92.77	**99.14**	94.24	**97.26**	82.44	96.52	89.89	97.77	89.84	91.83	-

TABLE 3 – Résultats sur quatre tâches aval de modèles de langues entraînés avec des jeux de données d'homogénéité et de taille variable. Nous rapportons les scores sur les ensemble de validation de chaque tâche (moyenne de 4 expériences de *fine-tuning* en POS tagging, en parsing et en NER, moyenne de 10 expériences de fine-tuning en NLI).

l'oral, sans ponctuation). L'écart de performance est également important en NLI, probablement en raison de la plus grande diversité thématique et en genre dans les corpus issus de Common Crawl, que l'on retrouve probablement dans les données XNLI, lui même divers thématiquement et en genre, et combinant données orales et écrites.

3.2 De combien de données avons-nous besoin ?

Un résultat inattendu de nos expériences est que le modèle CamemBERT standard, entraîné sur l'ensemble des 138Go de texte d'OSCAR, ne surpasse pas massivement le modèle entraîné « uniquement » sur l'échantillon de 4Go. Dans les configurations où le modèle de langue est utilisé comme plongements, le modèle entraîné sur 4Go conduit plus souvent à de meilleurs résultats que le CamemBERT standard entraîné sur 138Go, bien que les différences de scores soient rarement frappantes. Dans les configurations *fine-tuning*, le CamemBERT standard fonctionne généralement mieux que celui entraîné sur 4Go, mais là encore les différences sont toujours faibles.

En d'autres termes, lorsque les modèles sont entraînés sur des corpus tels que OSCAR et CCNet, hétérogènes en termes de genre et de style, 4Go de texte non compressé constitue un corpus de pré-entraînement suffisamment volumineux pour atteindre l'état de l'art avec l'architecture *BASE*, et notamment supérieurs dans tout les cas à ceux obtenus avec mBERT (pré-entraîné sur 60 Go de texte dans une centaine de langues). Cela remet en question la nécessité d'utiliser la totalité de très larges corpus tel qu'OSCAR ou CCNet lors du pré-entrainement de modèles tels que CamemBERT, sauf peut-être lorsque l'on utilise une architecture LARGE.

Cela signifie que des modèles de type CamemBERT peuvent être entraînés pour toutes les langues pour lesquelles un corpus varié d'au moins 4 Go peut être construit. OSCAR est disponible en 176 langues et fournit un tel corpus pour 38 langues. De plus, il est possible que des corpus légèrement plus petits (par exemple jusqu'à 1 Go) soient également suffisants pour entraîner des modèles de langue très performants.
Cependant, même avec une architecture BASE et 4 Go de données d'entraînement, la *validation loss* continue de diminuer au-delà de 100 000 *steps* (et 400 *epochs*). Cela suggère que nous sous-entraînons toujours sur le jeu de données de pré-entraînement de 4 Go, et qu'un entraînement plus long pourrait conduire à de meilleures performances. Quoiqu'il en soit, nos résultats ont été obtenus sur des modèles *BASE*, des recherches supplémentaires sont donc nécessaires pour confirmer la validité de nos résultats sur des architectures plus grandes et sur d'autres tâches plus complexes de

compréhension de la langue.

CORPUS	MASKING	ARCH.	#PARAM.	#STEPS	UPOS	LAS	NER	XNLI
	Stratégie de masking							
CCNET	*subword*	BASE	110M	100K	97.78	89.80	**91.55**	81.04
CCNET	*whole word*	BASE	110M	100K	**97.79**	**89.88**	91.44	**81.55**
	Taille du modèle							
CCNet	*whole word*	BASE	110M	100K	97.67	89.46	90.13	82.22
CCNet	*whole word*	LARGE	335M	100k	**97.74**	**89.82**	**92.47**	**85.73**
	Données d'entraînement							
CCNET	*whole word*	BASE	110M	100K	97.67	89.46	90.13	**82.22**
OSCAR	*whole word*	BASE	110M	100K	**97.79**	**89.88**	**91.44**	81.55
	Nombre de steps							
CCNet	*whole word*	BASE	110M	100k	**98.04**	89.85	90.13	82.20
CCNet	*whole word*	BASE	110M	500k	97.95	**90.12**	91.30	**83.04**

TABLE 4 – Comparaison des scores sur les ensemble de **Validation** des différents choix de conception. Les scores d'étiquetage morphosyntaxique et d'analyse syntaxique sont moyennés sur les 4 jeux de données.

3.3 Impact de la stratégie de *masking*

Dans le tableau 4, nous comparons les modèles entraînés avec une stratégie de *subword masking* à ceux en *whole word masking*. Le *whole word masking* a un impact positif sur les performances en NLI (mais seulement de 0,5 point de précision). À notre grande surprise et contrairement à l'anglais, cette stratégie de *masking* ne profite pas à des tâches de plus bas niveau (NER, étiquetage morphosyntaxique et analyse syntaxique).

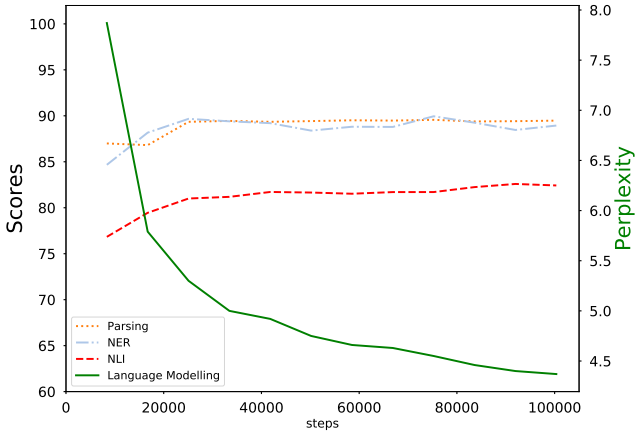
3.4 Impact de la taille du modèle

Le tableau 4 compare les modèles entraînés avec les architectures *BASE* et *LARGE*. Pour des raisons pratiques, ces modèles ont été entraînés avec le corpus CCNET (135 Go). Les résultats confirment l'impact positif de modèles plus grands sur les tâches NLI et NER. L'architecture *LARGE* conduit à une réduction d'erreur respectivement de 19,7% et 23,7% sur ces tâches. Étonnamment, sur les tâches d'étiquetage morphosyntaxique et d'analyse en dépendances, le fait d'avoir trois fois plus de paramètres ne conduit pas à des résultats significativement meilleurs qu'avec le modèle *BASE*.

Tenney *et al.* (2019) et Jawahar *et al.* (2019) ont montré que les informations morphosyntaxiques et syntaxiques sont apprises dans les couches inférieures de BERT tandis que les représentations sémantiques plus profondes se retrouvent dans les couches supérieures. Les couches inférieures de l'architecture *BASE* suffisent probablement à capturer ce qui est nécessaire aux tâches d'étiquetage morphosyntaxique et d'analyse syntaxique.

3.5 Impact du nombre de *steps*

La Figure ci-contre indique la perplexité du modèle de langue CAMEMBERT original ainsi que de ses performances sur nos tâches d'évaluation en fonction du nombre d'*epochs*, et ce à chaque *epoch* (8360 *steps*). Les résultats ci-contre suggèrent que plus la tâche est complexe, plus le nombre de *steps* a d'impact. Ainsi, alors qu'on peut observer un plateau pour les tâches bas-niveaux autour de 22000 *steps*, il semble que les performances continuent marginalement d'augmenter pour le NLI.



La comparaison entre deux modèles CCNET entraînés sur 100k et 500k *steps* respectivement (cf. Table 4) montre une légère augmentation des scores en NLI (+0,84) alors que ceux-ci stagnent en étiquetage et en analyse syntaxique. Ces résultats suggèrent que les représentations syntaxiques de bas niveau sont capturées bien plus tôt au cours de l'apprentissage que ne sont extraites les informations sémantiques complexes nécessaires au NLI.

4 Conclusion

Nous avons étudié l'impact de la taille et du niveau d'hétérogénéité des données de pré-entraînement sur la performance des modèles de langue neuronaux contextuels CAMEMBERT du français, ainsi qu'entre autres, l'impact de la taille du modèle et du nombre de *steps* de pré-entraînement. Nos résultats montrent que la taille des données d'entraînement n'a finalement que peu d'impact sur les performances globales et ouvrent donc la voie à des modèles de langages neuronaux contextuels spécialisés, liés à des domaines précis ou à des langues très peu dotées. La question de leur éventuelle complémentarité avec des modèles *fine-tuné* sur des modèles de langage générique est reste évidemment à explorer.

Entrainés sur des corpus *open-source* et disponibles sous une licence MIT, tous les modèles discutés dans cet article sont accessibles librement sur `https://camembert-model.fr`.

Remerciements

Nous tenons à remercier Clémentine Fourrier pour ses relectures et ses commentaires précieux, ainsi qu'Alix Chagué pour son fantastique logo. Ce travail a été en partie financé par trois projets de l'Agence Nationale de la Recherche accordés à Inria, les projets PARSITI (ANR-16-CE33-0021), SoSweet (ANR-15-CE38-0011) et BASNUM (ANR-18-CE38-0003), ainsi que par la chaire du dernier auteur dans l'Institut Prairie financée par l'ANR via le programme "Investissements d avenir" (ANR-19-P3IA-0001).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks*, p. 165–187. Kluwer : Dordrecht.

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, p. 1638–1649 : Association for Computational Linguistics.

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, France, June 4-8, 2012*, p. 321–334.

CHAN B., MÖLLER T., PIETSCH M., SONI T. & YEUNG C. M. (2019). German bert. https://deepset.ai/german-bert.

CHARNIAK E. (2019). *Introduction to deep learning*. The MIT Press.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint : 1911.02116.

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S. R., SCHWENK H. & STOYANOV V. (2018). XNLI : evaluating cross-lingual sentence representations. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, p. 2475–2485 : Association for Computational Linguistics.

DAI A. M. & LE Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 3079–3087.

DELOBELLE P., WINTERS T. & BERENDT B. (2020). RobBERT : a Dutch RoBERTa-based Language Model. arXiv preprint : 2001.06286.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md.

DEVLIN J., CHANG M., 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186 : Association for Computational Linguistics.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* : OpenReview.net.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. : European Language Resources Association (ELRA).

GRAVE E., MIKOLOV T., JOULIN A. & BOJANOWSKI P. (2017). Bag of tricks for efficient text classification. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2 : Short Papers*, p. 427–431 : Association for Computational Linguistics.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1 : Long Papers*, p. 328–339 : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

JAWAHAR G., SAGOT B., SEDDAH D., UNICOMB S., IÑIGUEZ G., KARSAI M., LÉO Y., KARSAI M., SARRAUTE C., FLEURY É. *et al.* (2019). What does bert learn about the structure of language ? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). Fasttext.zip : Compressing text classification models. arXiv preprint : 1612.03651.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint : 1412.6980.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2019). ALBERT : A lite BERT for self-supervised learning of language representations. arXiv preprint : 1909.11942.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. arXiv preprint : 1912.05372.

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. arXiv preprint : 1907.11692.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. arXiv preprint : 1911.03894.

MIKOLOV T., GRAVE E., BOJANOWSKI P., PUHRSCH C. & JOULIN A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, p. 3111–3119.

NIVRE J., ABRAMS M., AGIĆ Ž., AHRENBERG L., ANTONSEN L., ARANZABE M. J., ARUTIE G., ASAHARA M., ATEYAH L., ATTIA M., ATUTXA A., AUGUSTINUS L., BADMAEVA E., BALLESTEROS M., BANERJEE E., BANK S., BARBU MITITELU V., BAUER J., BELLATO S., BENGOETXEA K., BHAT R. A., BIAGETTI E., BICK E., BLOKLAND R., BOBICEV V., BÖRSTELL C., BOSCO C., BOUMA G., BOWMAN S., BOYD A., BURCHARDT A., CANDITO M., CARON B., CARON G., CEBIROĞLU ERYIĞIT G., CELANO G. G. A., CETIN S., CHALUB F., CHOI J., CHO Y., CHUN J., CINKOVÁ S., COLLOMB A., ÇÖLTEKIN Ç., CONNOR M., COURTIN M., DAVIDSON

E., DE MARNEFFE M.-C., DE PAIVA V., DIAZ DE ILARRAZA A., DICKERSON C., DIRIX P., DOBROVOLJC K., DOZAT T., DROGANOVA K., DWIVEDI P., ELI M., ELKAHKY A., EPHREM B., ERJAVEC T., ETIENNE A., FARKAS R., FERNANDEZ ALCALDE H., FOSTER J., FREITAS C., GAJDOŠOVÁ K., GALBRAITH D., GARCIA M., GÄRDENFORS M., GERDES K., GINTER F., GOENAGA I., GOJENOLA K., GÖKIRMAK M., GOLDBERG Y., GÓMEZ GUINOVART X., GONZÁLES SAAVEDRA B., GRIONI M., GRŪZĪTIS N., GUILLAUME B., GUILLOT-BARBANCE C., HABASH N., HAJIČ J., HAJIČ JR. J., HÀ MỸ L., HAN N.-R., HARRIS K., HAUG D., HLADKÁ B., HLAVÁČOVÁ J., HOCIUNG F., HOHLE P., HWANG J., ION R., IRIMIA E., JELÍNEK T., JOHANNSEN A., JØRGENSEN F., KAŞIKARA H., KAHANE S., KANAYAMA H., KANERVA J., KAYADELEN T., KETTNEROVÁ V., KIRCHNER J., KOTSYBA N., KREK S., KWAK S., LAIPPALA V., LAMBERTINO L., LANDO T., LARASATI S. D., LAVRENTIEV A., LEE J., LÊ HỒNG P., LENCI A., LERTPRADIT S., LEUNG H., LI C. Y., LI J., LI K., LIM K., LJUBEŠIĆ N., LOGINOVA O., LYASHEVSKAYA O., LYNN T., MACKETANZ V., MAKAZHANOV A., MANDL M., MANNING C., MANURUNG R., MĂRĂNDUC C., MAREČEK D., MARHEINECKE K., MARTÍNEZ ALONSO H., MARTINS A., MAŠEK J., MATSUMOTO Y., MCDONALD R., MENDONÇA G., MIEKKA N., MISSILÄ A., MITITELU C., MIYAO Y., MONTEMAGNI S., MORE A., MORENO ROMERO L., MORI S., MORTENSEN B., MOSKALEVSKYI B., MUISCHNEK K., MURAWAKI Y., MÜÜRISEP K., NAINWANI P., NAVARRO HORÑIACEK J. I., NEDOLUZHKO A., NEŠPORE-BĒRZKALNE G., NGUYỄN THỊ L., NGUYỄN THỊ MINH H., NIKOLAEV V., NITISAROJ R., NURMI H., OJALA S., OLÚÒKUN A., OMURA M., OSENOVA P., ÖSTLING R., ØVRELID L., PARTANEN N., PASCUAL E., PASSAROTTI M., PATEJUK A., PENG S., PEREZ C.-A., PERRIER G., PETROV S., PIITULAINEN J., PITLER E., PLANK B., POIBEAU T., POPEL M., PRETKALNIŅA L., PRÉVOST S., PROKOPIDIS P., PRZEPIÓRKOWSKI A., PUOLAKAINEN T., PYYSALO S., RÄÄBIS A., RADEMAKER A., RAMASAMY L., RAMA T., RAMISCH C., RAVISHANKAR V., REAL L., REDDY S., REHM G., RIESSLER M., RINALDI L., RITUMA L., ROCHA L., ROMANENKO M., ROSA R., ROVATI D., ROȘCA V., RUDINA O., SADDE S., SALEH S., SAMARDŽIĆ T., SAMSON S., SANGUINETTI M., SAULĪTE B., SAWANAKUNANON Y., SCHNEIDER N., SCHUSTER S., SEDDAH D., SEEKER W., SERAJI M., SHEN M., SHIMADA A., SHOHIBUSSIRRI M., SICHINAVA D., SILVEIRA N., SIMI M., SIMIONESCU R., SIMKÓ K., ŠIMKOVÁ M., SIMOV K., SMITH A., SOARES-BASTOS I., STELLA A., STRAKA M., STRNADOVÁ J., SUHR A., SULUBACAK U., SZÁNTÓ Z., TAJI D., TAKAHASHI Y., TANAKA T., TELLIER I., TROSTERUD T., TRUKHINA A., TSARFATY R., TYERS F., UEMATSU S., UREŠOVÁ Z., URIA L., USZKOREIT H., VAJJALA S., VAN NIEKERK D., VAN NOORD G., VARGA V., VINCZE V., WALLIN L., WASHINGTON J. N., WILLIAMS S., WIRÉN M., WOLDEMARIAM T., WONG T.-S., YAN C., YAVRUMYAN M. M., YU Z., ŽABOKRTSKÝ Z., ZELDES A., ZEMAN D., ZHANG M. & ZHU H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN & C. ILIADI, Éds., *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom : Leibniz-Institut für Deutsche Sprache. HAL : hal-02148693.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In A. MOSCHITTI, B. PANG & W. DAELEMANS, Éds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1532–1543 : ACL.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. A. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 2227–2237 : Association for Computational Linguistics.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual bert ? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*. DOI : 10.18653/v1/P19-1493.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners. preprint, https://paperswithcode.com/paper/language-models-are-unsupervised-multitask.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint : 1910.10683.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the paris 7 french treebank) [in french]. In G. ANTONIADIS, H. BLANCHON & G. SÉRASSET, Éds., *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, France, June 4-8, 2012*, p. 535–542 : ATALA/AFCP.

STRAKA M., STRAKOVÁ J. & HAJIC J. (2019). Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. arXiv preprint : 1908.07448.

STRAKOVÁ J., STRAKA M. & HAJIC J. (2019). Neural architectures for nested NER through linearization. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1 : Long Papers*, p. 5326–5331 : Association for Computational Linguistics.

TENNEY I., DAS D. & PAVLICK E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4593–4601, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1452.

VIRTANEN A., KANERVA J., ILO R., LUOMA J., LUOTOLAHTI J., SALAKOSKI T., GINTER F. & PYYSALO S. (2019). Multilingual is not enough : Bert for finnish. arXiv preprint : 1912.07076.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2019). CCNet : Extracting High Quality Monolingual Datasets from Web Crawl Data. arXiv preprint : 1911.00359.

WILLIAMS A., NANGIA N. & BOWMAN S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 1112–1122.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. arXiv preprint : 1910.03771.