Étude comparative de modèles Transformers en compréhension de la parole en français

Oralie Cattan, Sahar Ghannay, Christophe Servan, Sophie Rosset

Étude comparative de modèles Transformers en compréhension de la parole en français

Oralie Cattan[1, 2] Sahar Ghannay[1] Christophe Servan[2] Sophie Rosset[1]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
lastname@lisn.upsaclay.fr
(2) QWANT, 10 Boulevard Haussmann, 75009 Paris, France
inital.lastname@qwant.com

RÉSUMÉ

Au cours des cinq dernières années, les approches par transfert utilisant les modèles de type Transformers ont établit l'État de l'Art sur de nombreuses tâches du Traitement Automatique des Langues. Ces approches deviennent de plus en plus populaires et nécessitent une grande quantité de données et de ressources computationnelles. La plupart des modèles de langue pré-entraînés ont fait l'objet de nombreuses études en anglais et seulement quelques-uns d'entre eux ont été évalués sur une tâche de compréhension de la parole en français. Dans cet article, nous proposons un tour d'horizon de référence, axé sur l'évaluation de la qualité de treize modèles bien établis basés sur les modèles Transformers. Les deux tâches de compréhension de la parole considérées sont MEDIA et ATIS-FR.

ABSTRACT

Benchmarking Transformer-based models on French Spoken Language Understanding tasks

In the past five years, model-based transfer learning, with the rise of the Transformer architecture led to state-of-the-art performances over many natural language tasks. These approaches are becoming increasingly popular and require large amounts of data and computational resources. Most pre-trained language models were massively studied using the English language and only a few of them were evaluated on a spoken language understanding task, in French. In this paper, we propose a unified benchmark, focused on evaluating models quality of thirteen well-established Transformer-based models. The two available spoken language understanding tasks considered are MEDIA and ATIS-FR.

MOTS-CLÉS : Modèles Transformers, Compréhension de la parole, Comparatif, Modèles de langue, Étude comparative.

KEYWORDS: Transformers Models, French Language, Language Models, Spoken Language Understanding, Benchmarking.

1 Introduction

Le développement de l'apprentissage par transfert et la disponibilité de modèles de langue pré-entraînés basés sur des architectures Transformers (Vaswani *et al.*, 2017), comme BERT (Devlin *et al.*, 2019), ont récemment permis de réaliser d'importants progrès dans le domaine du Traitement Automatique des Langues (TAL). Toutefois, la tendance consistant à créer de grands modèles pré-entraînés sur des quantités de données de plus en plus grandes, avec une quantité croissante de

paramètres, a soulevé des questions liées à l'utilisabilité de ces approches. Étant donné que ces modèles nécessitent des ressources informatiques considérables, d'importants efforts ont été déployés pour élaborer des modèles compacts afin de réduire le coût de leur utilisation (notamment énergétique). Une alternative possible à ces modèles à forte consommation énergétique sont les modèles compacts qui ont des performances comparables aux modèles plus gros, tout en réduisant la complexité de calcul et leur taille. Ces modèles permettent également de résoudre certains besoins industriels liés au traitement automatique des langues et de la parole en temps réel.

La compréhension de la parole (spoken language understanding – SLU) dans le cadre d'un système de dialogue fait référence à la tâche de produire une analyse sémantique et une formalisation du discours de l'utilisateur. La SLU englobe traditionnellement les processus de détermination d'un large éventail d'informations transmises dans le dialogue comme l'identification du domaine, l'intention et les concepts de la conversation.

Les études comparative ont été largement employées pour l'évaluation de modèles en TAL (Guo *et al.*, 2020; Le *et al.*, 2020; Farha & Magdy, 2021). En ce qui concerne notre tâche, les recherches récentes ont porté sur l'évaluation des représentations continue de mots (ou plongements de mots). Ces représentations s'avèrent essentielles dans la capture des relations sémantiques entre les mots. Elles sont également un élément essentiel des architectures issues de l'apprentissage profond.

Dans Ghannay *et al.* (2020a), les représentations contextuelles (ELMo (Peters *et al.*, 2018)) à plat (Word2Vec (Mikolov *et al.*, 2013), celle de GloVe (Pennington *et al.*, 2014) et de Fast-Text (Bojanowski *et al.*, 2017)) ont été évaluées afin d'étudier les différentes entrées de ces représentations et leur influence sur les résultats obtenus dans les tâches de compréhension. Ils ont souligné la compétitivité de Word2Vec et ELMO sur la tâche de compréhension en Français : MEDIA.

Plus récemment, Ghannay *et al.* (2020b) ont étudié la transférabilité de deux modèles de BERT pré-préentraînés (Devlin *et al.*, 2019) et leur intégration dans les architectures de BiLSTM et de BiLSTM+CNN. Ils ont également établit un nouvel État-de-l'Art sur MEDIA en utilisant le modèle CamemBERT (Martin *et al.*, 2020).

Dernièrement, Xu *et al.* (2020) ont proposé une comparaison de diverses approches de transfert interlingue, avec notamment le modèle BERT multilingue (mBERT), évalué sur MultiATIS++, un corpus multilingue pour la tâche de compréhension de la parole. Ils montrent que mBERT apporte des améliorations substantielles lors d'un apprentissage multilingue et aux tâches de transfert interlingue, ce qui donne jusqu'à 1,4% d'amélioration sur le sous-corpus français de MultiATIS++ par rapport à l'État-de-l'Art.

Bien qu'il existe maintenant de nombreux corpus et modèles anglais pour diverses tâches, il y a encore peu de ressources en français dans le domaine. Les comparaisons directes entre les modèles SLU sont difficiles en raison de l'absence d'un cadre d'évaluation unifié et de la taille de différents modèles Transformers en français.

Contributions : Dans cette étude, nous étudions l'utilisation d'architectures Transformers en français pour résoudre le problème de compréhension de la parole, une tâche de détection de concept (Bonneau-Maynard *et al.*, 2006). Nous évaluons également l'impact lié à la compacité des modèles sur les performances d'analyse. Nous nous concentrons sur l'analyse comparative des modèles français et multilingues existants de Transformers sur deux tâches françaises de compréhension de la parole : MEDIA et ATIS-FR. Les modèles pris en compte pour ce point de référence sont détaillés dans la section 2, les expériences, les données, l'optimisation du modèle et le protocole expérimental sont décrits dans la section 3. La section 4 présente les résultats, et section 5 propose une analyse.

Modèles	Objectifs	Données	Taille Vocabulaire	Tokenisation	# paramètres	Taille Modèle
FlauBERT$_{base}$	MLM	24 sous-corpus français (71 Go de texte)	68 729	BPE	138 M	553 Mb
CamemBERT$_{base}$	MLM	OSCAR français(138 Go de texte)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet français(135 Go de texte)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	OSCAR français (4 Go de texte)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet français (4 Go de texte)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	Wikipedia français (4 Go de texte)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{large}$	MLM	CCNet français (135 Go de texte)	32 005	SentencePiece	335 M	1,35 Gb
FrALBERT$_{base}$	MLM et SOP	Wikipedia français (4 Go de texte)	32 005	SentencePiece	12 M	50 Mb
XLM-R$_{base}$	MLM	CC-100 (2,5 To de texte)	250 002	BPE	278 M	1,12 Gb
XLM-R$_{large}$	MLM	CC-100 (2,5 To de texte)	250 002	BPE	559 M	1,24 Gb
mBERT$_{base}$	MLM et NSP	Wiki-100	119 547	WordPiece	177 M	714 Mb
small-mBERT$_{base}$ FR	MLM et NSP	Wiki-100	24 495	WordPiece	104 M	420 Mb
distil-mBERT$_{base}$	MLM et NSP	Wiki-100	119 547	WordPiece	134 M	542 Mb

TABLE 1 – Caractéristiques des modèles, avec les paramètres de pré-entraînement, les sources de données et les tailles des modèles.

Modèles	Objectifs	Données	Taille Vocabulaire	Tokenisation	# para--mètres	Taille Modèle
FlauBERT$_{base}$	MLM	24 sous-corpus (71 Go)	68 729	BPE	138 M	553 Mb
CamemBERT$_{base}$	MLM	OSCAR (138 Go)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet (135 Go)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	OSCAR (4 Go)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet (4 Go)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	Wikipedia (4 Go)	32 005	SentencePiece	110 M	445 Mb
CamemBERT$_{large}$	MLM	CCNet (135 Go)	32 005	SentencePiece	335 M	1,35 Gb
FrALBERT$_{base}$	MLM et SOP	Wikipedia (4 Go)	32 005	SentencePiece	12 M	50 Mb
XLM-R$_{base}$	MLM	CC-100 (2,5 To)	250 002	BPE	278 M	1,12 Gb
XLM-R$_{large}$	MLM	CC-100 (2,5 To)	250 002	BPE	559 M	1,24 Gb
mBERT$_{base}$	MLM et NSP	Wiki-100	119 547	WordPiece	177 M	714 Mb
small-mBERT$_{base}$ FR	MLM et NSP	Wiki-100	24 495	WordPiece	104 M	420 Mb
distil-mBERT$_{base}$	MLM et NSP	Wiki-100	119 547	WordPiece	134 M	542 Mb

TABLE 2 – Caractéristiques des modèles, avec les paramètres de pré-entraînement, les sources de données et les tailles des modèles.

2 Modèles Transformers étudiés

Afin de comparer les performances des modèles sur les tâches de compréhension en français, nous soulignons dans cette section et dans le tableau 2 quelques caractéristiques des modèles considérés.

Parmi les modèles de langue proposés ces dernières années, nous considérons les modèles multilingues : XLM-R (Conneau *et al.*, 2020) et mBERT (Devlin *et al.*, 2019), ainsi que les modèles français : Camembert (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020) [1]. A ce jour et à notre connaissance, ce sont les deux seuls grands modèles disponibles librement pour le français. Nous évaluons également des modèles compacts qui sont des modèles évolutifs de Transformers. Nous exploitons le distil-mBERT (Sanh *et al.*, 2019), une version distillée de mBERT, small-mBERT un modèle mBERT dont le vocabulaire original a été réduit à la langue française et FrALBERT, un modèle récemment publié pour le français (Cattan *et al.*, 2021b).

Tel que mentionnés dans Kaplan *et al.* (2020), les performances des modèles de langue sont déterminées par plusieurs facteurs tels que les objectifs de pré-entraînement, les caractéristiques des couches ou encore la taille des ensembles de données d'apprentissage.

En effet, les objectifs de pré-formation varient selon les modèles et peuvent dépendre de la tâche en aval résolue (Joshi *et al.*, 2020). Les modèles de type BERT adoptent les objectifs de modélisation du langage à masque (masked language modeling – MLM) et de prévision de la prochaine phrase (next sentence prediction – NSP). Le MLM est une tâche de remplissage consistant à prévoir les tokens de la séquence d'entrée masqués alors que le NSP est une tâche de classification binaire pour prédire si deux segments sont adjacents dans le texte original. XLM-R diffère de mBERT uniquement dans la procédure de pré-entraînement, éliminant la tâche de NSP. C'est également le cas pour les modèles CamemBERT et FlauBERT. Les objectifs de pré-entraînement de FrALBERT sont le MLM et la prédiction de l'ordre des phrases (Sentence Order Prediction – SOP). Le SOP modélise efficacement la cohérence entre les phrases en prédisant si l'ordre de la phrase dans une paire de phrases est interchangeable ou non.

Pour la modélisation monolingue et multilingue, la qualité et la représentativité des ensembles de données à grande échelle utilisés sont importantes pour une modélisation efficace et pour une bonne généralisation. La plupart des modèles sont pré-entraînés sur le contenu de Wikipedia et de pages web recueillies par Common Crawl (CC) ou une combinaison de diverses sociétés. Cela représente quelques gigaoctets à plusieurs centaines ou même plusieurs teraoctets de texte. L'autre différence notable est la taille du vocabulaire avec plusieurs dizaines de milliers de tokens pour les modèles monolingues à plusieurs centaines pour les modèles multilingues.

3 Expériences

Les expériences sont réalisées sur deux tâches et corpus en français MEDIA et ATIS-FR dédiées à la tâche de compréhension de la parole.

1. Nous n'avons pas inclus les petits modèles CamemBERT et la grande version de FlauBERT, le premier n'étant pas disponible et le second ne convergeant pas dans nos expériences malgré nos efforts.

3.1 Données

Le corpus *MEDIA* [2] est composé de 1258 dialogues transcrits, sur la réservation d'hôtel et d'information touristiques (Bonneau-Maynard *et al.*, 2006). Le corpus a été annoté manuellement avec des concepts sémantiques caractérisés par une étiquette et sa valeur. Il y a au total 76 étiquettes sémantiques. Le corpus est divisé en trois parties : un corpus d'entraînement composé de 13k phrases, un corpus de développement composé de 1,3k phrases et un corpus d'évaluation composé de 3,5k phrases. La tâche MEDIA est également connue comme l'une des tâches les plus difficiles, selon (Béchet & Raymond, 2019).

La version française du corpus *Air Travel Information System (ATIS)* issu de la récente extension MultiATIS++ (Xu *et al.*, 2020), nommé ATIS-FR, sur le domaine de l'information de vol (Price, 1990). Le corpus ATIS-FR correspond à la traduction manuelle des phrases originales de ATIS (depuis l'anglais). Il est composé de 84 étiquettes de concepts et est divisé également en trois parties : un corpus de d'entraînement composé de 4,5k phrases, un corpus de développement composé de 490 phrases et d'un corpus d'évaluation composé de 893 phrases.

3.2 Protocole expérimental

La compréhension de la parole (SLU) est considérée ici comme une tâche d'étiquetage qui attribue une étiquette de concept (d'un ensemble prédéfini) à chaque token d'une phrase. Nous suivons le schéma de marquage BIO, où chaque concept est associé à deux étiquettes, B (pour le début – Begin) et I (pour l'intermédiaire – Intermediate). Enfin, l'étiquette O (pour les autres – Others) identifie les tokens associés à aucun concept. La performance de la compréhension est évaluée en termes de F-mesure (ou F1) et de taux d'erreur de concepts (Concept Error Rate – CER). Le score CER est la mesure utilisée dans la campagne MEDIA (Bonneau-Maynard *et al.*, 2006) qui est estimée de la même manière que le taux d'erreur de mots classiques mais appliquée aux concepts sémantiques au lieu des mots (le plus bas le mieux).

Comme présenté par Devlin *et al.* (2019), nous ajoutons à la suite du modèle pré-entraîné un classificateur (au niveau du token) avec deux couches linéaires cachées (avec des activations ReLU et un *dropout*) qui prend comme entrée le dernier état caché de la séquence d'entrée et produit des probabilités sur les concepts.

Dans nos expériences, nous utilisons l'optimiseur Adam et effectuons une optimisation d'hyperparamètres automatisée sur les ensembles de données de développement. Cette optimisation d'hyperparamètres est fondée sur l'algorithme d'apprentissage bioinspiré sur les populations (Jaderberg *et al.*, 2017) et dans lequel une population de modèles et leurs hyperparamètres sont optimisés conjointement. Parmi les hyperparamètres pris en considération sont le nombre d'époques d'apprentissage de 5 à 100, la taille du *batch* dans un intervalle de 8 à 32 et avec un taux d'apprentissage compris entre 1e-05 et 5e-05. Nous sélectionnons les meilleurs modèles obtenus selon le meilleur score CER.

2. MEDIA est disponible librement à des fins de recherche : *https://catalogue.elra.info/ELRA-S0272*.

4 Resultats

Cette section présente les résultats de l'évaluation des modèles Transformers décrits dans la section 2 sur les corpus MEDIA et ATIS-FR.

4.1 Résultats sur MEDIA

Les performances sur MEDIA sont présentées dans les deux premières colonnes du tableau 3. Les résultats des modèles monolingues sont très proches et varient de 89 à 90 de F1, alors que les scores de CER varient entre 7,5 et 8,6. FlauBERT$_{\text{base}}$ obtient les moins bons scores de F1 et CER (respectivement 89,0 et 8,1) tandis que CamemBERT$_{\text{base,Wiki 4 Go}}$, le modèle CamemBERT pré-entraîné sur seulement 4 Gb de Wikipedia, obtient le meilleur score F1 (90,0) avec un score CER de 8,4. Selon le CER, la mesure historique de la tâche MEDIA, le meilleur modèle est CamemBERT$_{\text{base, CCNet 135 Go}}$, un modèle de base CamemBERT pré-entraîné avec CCNet sur 135 Gb de texte. Il obtient le CER le plus bas, à 7,5 pour un F1 à 89,9.

Les modèles multilingues (mBERT$_{\text{base}}$, distill-mBERT$_{\text{base}}$, small-mBERT$_{\text{base-fr}}$, XLM-R$_{\text{base}}$ and XLM-R$_{\text{large}}$) ont obtenu les meilleurs scores de CER allant de 10,1 à 8,0. Le modèle le moins performan est distill-mBERT$_{\text{base}}$, et le meilleur XLM-R$_{\text{large}}$. En ce qui concerne les scores F1, ils sont comparables aux modèles monolingues français, pour les modèles XLM-R avec 89,5 pour le *base* et 89,9 pour la version *large*. D'autre part, BERT multilingue et ses versions *distill* et *small* ont obtenu des performances comparables au FlauBERT$_{\text{base}}$, avec 88,9 F1.

4.2 Résultats sur ATIS-FR

Les performances des modèles monolingues sur la tâche ATIS-FR en F1 varient de 92,5 à 94,1 et de 3,3 à 5,3 du CER (tableau 3, colonnes de droite). FlauBERT obtient les meilleurs résultats F1 et CER tandis que la version *large* du modèle CamemBERT obtient les meilleurs résultats de F1 et de CER. FrALBERT et Camembert$_{\text{base, Wiki 4 Go}}$ obtiennent des performances comparables avec 0,3 point de F1 et 0,1 point de différence CER.

Le meilleur modèle est CamemBERT$_{\text{large, CCNet 135 Go}}$ obtenant un CER de 3,3 pour une F1 à 94,1. Les scores F1 varient de 88.1 (CER à 6,0), pour la version *distill* de mBERT), à 93.6 (CER à 5.0) pour mBERT$_{\text{base}}$. Les deux versions de XLM-R ont obtenus des résultats comparables en termes de CER et F1 à CamemBERT$_{\text{base, Wiki 4 Go}}$ et FrALBERT$_{\text{base, Wiki 4 Go}}$.

4.3 Impact écologique des modèles

Nous avons conduit une étude d'impact des modèles en mesurant la quantité de CO_2 produit à travers le temps passé et l'énergie consommée. Nous utilisons l'approche proposée par Henderson *et al.* (2020) dont les résultats sont présenté dans le tableau 4. Ce dernier présente trois informations de temps passé d'apprentissage, énergie consommée et donc de production de CO_2 pour les modèles considérés pour la tâche ATIS-FR avec trois époques. Le nombre d'époque fut choisi dans un soucis de préservation de l'énergie.

Model	MEDIA F1	MEDIA CER	ATIS-FR F1	ATIS-FR CER
$\text{FlauBERT}_{\text{base}}$	89,0	8,1	92,5	5,3
$\text{CamemBERT}_{\text{base, OSCAR 138 Go}}$	89,3	7,9	93,9	3,7
$\text{CamemBERT}_{\text{base, CCNet 135 Go}}$	89,9	**7,5**	94,0	3,7
$\text{CamemBERT}_{\text{base, OSCAR 4 Go}}$	89,7	8,3	93,6	3,7
$\text{CamemBERT}_{\text{base, CCNet 4 Go}}$	89,7	8,3	93,8	3,8
$\text{CamemBERT}_{\text{base, Wiki 4 Go}}$	**90,0**	8,4	92,5	4,2
$\text{CamemBERT}_{\text{large, CCNet 135 Go}}$	89,2	7,8	**94,1**	**3,3**
$\text{FrALBERT}_{\text{base, Wiki 4 Go}}$	89,8	8,6	92,8	4,3
$\text{XLM-R}_{\text{base}}$	89,5	8,5	92,5	4,3
$\text{XLM-R}_{\text{large}}$	89,9	8,0	92,7	4,4
$\text{mBERT}_{\text{base}}$	88,9	8,7	93,6	5,0
$\text{distill-mBERT}_{\text{base}}$	87,5	10,1	88,1	6,0
$\text{small-mBERT}_{\text{base-fr}}$	88,8	8,1	93,3	5,3

TABLE 3 – Performances mesurées sur les données de test des corpus MEDIA et ATIS-FR. Les résultats sont donnés en termes de F-mesure (F1) et de taux d' erreur de concepts (CER).

Les modèles ayant le plus grand nombre de paramètres sont ceux qui consomment le plus ($\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$ et $\text{XLM-R}_{\text{large}}$). On constate que les modèles $\text{FlauBERT}_{\text{base}}$, $\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$ et $\text{mBERT}_{\text{base}}$ à nombre de paramètres équivalent, produisent presque autant de CO_2 (respectivement 7,98 g, 8,10 g et 8,80 g). $\text{FrALBERT}_{\text{base, Wiki 4 Gb}}$ est le modèle qui consomme le moins (0,88 kWh), et qui produit le moins d e CO_2 (1,15g) devant les modèles compacts $\text{distill-mBERT}_{\text{base}}$ (1,50 kWh & 4,35 g) et $\text{small-mBERT}_{\text{base-fr}}$ (1,45 kWh & 6,29 g).

5 Analyse des résultats

L'évaluation des modèles Transformers en compréhension de la parole en français sur les corpus MEDIA & ATIS-FR permet d'observer certaines tendances. Dans les deux tâches, les modèles CamemBERT sont les meilleurs, en considérant le taux d'erreur de concepts (CER). FrALBERT obtient des résultats comparables à $\text{Camembert}_{\text{base, Wiki 4 Go}}$, ce qui peut provenir du fait que les deux modèles sont entraînés sur le même type et la même quantité de données (4Go de Wikipedia). Nous constatons également que FrALBERT a des performances comparables à $\text{XLM-R}_{\text{base}}$ dans les deux tâches, même si les données d'entraînement et la structure sont toutes deux différentes. En outre, nous constatons une sous-performance du modèle $\text{distill-mBERT}_{\text{base}}$ par rapport à d'autres modèles BERT et en particulier à FrALBERT. Enfin , $\text{FlauBERT}_{\text{base}}$ a des résultats comparables aux modèles CamemBERT dans la tâche MEDIA, cependant le score CER de FlauBERT dans la tâche ATIS-FR est moins bon que les modèles CamemBERT. Ce dernier point nous a conduit à aller plus loin dans l'analyse.

Nous examinons en détail les résultats pour chaque étiquette ou groupe d'étiquettes du même type. La première tendance que nous avons observée est une sous-performance du modèle $\text{distill-mBERT}_{\text{base}}$ sur les étiquettes d'entité nommées sur les deux tâches. Par exemple, STATE NAME, CITY NAME ou MEAN DESCRIPTION dans ATIS ou HOTEL-NAME, LOC-CITY pour MEDIA ont jusqu'à 2 points

Model	Temps (s)	Energie (kWh)	CO_2 (g)
$\text{FlauBERT}_{\text{base}}$	78.58	1.98	7.98
$\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$	76.85	1.68	8.10
$\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$	172.43	4.22	18.39
$\text{FrALBERT}_{\text{base, Wiki 4 Gb}}$	**46.13**	**0.88**	**1.15**
$\text{XLM-R}_{\text{base}}$	97.6	2.46	10.25
$\text{XLM-R}_{\text{large}}$	197.00	5.08	22.90
$\text{mBERT}_{\text{base}}$	85.93	2.07	8.80
$\text{distill-mBERT}_{\text{base}}$	79.13	1.50	4.35
$\text{small-mBERT}_{\text{base-fr}}$	54.86	1.45	6.29

TABLE 4 – Temps passé, énergie consommée et quantité CO_2 produite par les modèles pour un entraînement en trois époques.

F1 de moins que les autres modèles. Précisons que ces étiquettes sont parmis des plus fréquentes dans les deux tâches. D'autre part, nous n'avons pas pu observer une telle tendance en comparant des modèles plus grands et des modèles compacts, ou des modèles multilingues par rapport aux modèles monolingues français. Les différences entre les modèles sont très faibles, par exemple le modèle $\text{FrALBERT}_{\text{base , Wiki 4Go}}$ sera meilleur que $\text{CamemBERT}_{\text{base , Wiki 4 Go}}$ sur le label MEDIA LINKREF-COREF (91,47 contre 90,62 de F1, respectivement). Au contraire, ce dernier modèle est meilleur que le premier sur l'étiquette COMMAND-TASK avec respectivement 83,38 et 84,90 de F1, dans la tâche MEDIA .

Lorsque nous allons plus en profondeur sur la tâche MEDIA, nous pouvons détecter quelques tendances lorsque nous examinons l'ensemble des résultats. Par exemple, même si le label OBJECT est l'un des plus fréquents, il semble que tous les modèles ont des difficultés à le détecter correctement. De la même manière, le label NAME est très difficile, même pour un des meilleurs modèles ($\text{CamemBERT}_{\text{base, CCNet 135 Go}}$). Il semble que la plus grande différence entre les modèles réside dans leur capacité à transférer leurs connaissances en fonction de la quantité de données utilisées pour le pré-entraînement. De cette façon, les modèles entraînés avec la plus grande quantité de corpus sont en mesure de mieux gérer les étiquettes les plus rares. Ce que nous observons dans la tâche MEDIA peut également être observé dans la tâche ATIS-FR, même si le rendement global des modèles est plus élevé. Cela nous amène à suggérer qu'un meilleur choix des données de pré-entraînement pourrait apporter des résultats intéressants. De plus, les travaux récents dans l'apprentissage de type *few-shot* pour les approches de labellisation (Cattan *et al.*, 2021a) permettent d'obtenir de hautes performances avec des sous-échantillonnages élevés, qui dans notre cas pourrait être un plus.

6 Conclusion

Les modèles fondés sur les modèles Transformers sont actuellement les modèles à l'État-de-l'Art pour de nombreuses tâches de TAL. Dans cette étude, nous avons proposé de comparer les modèles Transformers pré-entraînés français et multilingues existants, afin de comparer leurs performances sur deux corpus de compréhension de la parole en français : MEDIA et ATIS-FR. Nous avons également évalué l'impact lié à la compacité des modèles sur les performances de compréhension et nous avons

procédé à une comparaison de treize modèles de Transformers.

Les résultats expérimentaux montrent que ces tâches sont très difficiles même pour les grands modèles. Pour les deux tâches, les modèles CamemBERT sont les meilleurs, en termes de taux d'erreur de concepts (CER). Les meilleurs résultats de CER obtenus sur MEDIA et ATIS-FR sont associées au modèles $\text{CamemBERT}_{\text{base, CCNet 135 Go}}$ et $\text{CamemBERT}_{\text{large, CCNet 135 Go}}$. Dans les deux tâches, le modèle compact français FrALBERT obtient des résultats comparables au grand modèle $\text{CamemBERT}_{\text{base, Wiki 4 Go}}$. En outre, ce modèle compact atteint des performances comparables à celles des modèles multilingues dans les deux tâches et il dépasse la version distillée de BERT (distill-mBERT).

Enfin, notre analyse détaillée des résultats de F1 fournit une vue d'ensemble intéressante de chaque tâche. À partir de ces analyses, nous constatons que la plus grande différence entre les modèles se produit dans leur capacité à transférer leurs connaissances en fonction de la quantité de données de pré-entraînement. Nous prévoyons de mettre en ligne notre code afin de faciliter de futures études et analyses comparatives, la recherche et l'élaboration de modèles à l'avenir.

7 Remerciements

Ce travail a été financé en partie à travers le projet ANR TextToKids (AAPG2019). Ce travail a été rendu possible grâce à la plateforme de calcul Saclay-IA

Références

BÉCHET F. & RAYMOND C. (2019). Benchmarking benchmarks : introducing new automatic indicators for benchmarking spoken language understanding corpora. In *InterSpeech*.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *TACL*.

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. & VILLANEAU J. (2006). Results of the French Evalda-Media evaluation campaign for literal understanding. In *LREC*.

CATTAN O., ROSSET S. & SERVAN C. (2021a). On the cross-lingual transferability of multilingual prototypical models across NLU tasks. In *META-NLP*.

CATTAN O., SERVAN C. & ROSSET S. (2021b). On the Usability of Transformers-based models for a French Question-Answering task. In *RANLP*.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

DEVLIN J., CHANG M. W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NACL*.

FARHA I. A. & MAGDY W. (2021). Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *WANLP*.

GHANNAY S., NEURAZ A. & ROSSET S. (2020a). What is best for spoken language understanding : small but task-dependant embeddings or huge but out-of-domain embeddings ? In *ICASSP*.

GHANNAY S., SERVAN C. & ROSSET S. (2020b). Neural networks approaches focused on French spoken language understanding : application to the MEDIA evaluation task". In *COLING*.

GUO J., LIU Q., LOU J. G., LI Z., LIU X., XIE T. & LIU T. (2020). Benchmarking meaning representations in neural semantic parsing. In *EMNLP*.

HENDERSON P., HU J., ROMOFF J., BRUNSKILL E., JURAFSKY D. & PINEAU J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, **21**(248), 1–43.

JADERBERG M., DALIBARD V., OSINDERO S., CZARNECKI W. M., DONAHUE J., RAZAVI A., VINYALS O., GREEN T., DUNNING I., SIMONYAN K., FERNANDO C. & KAVUKCUOGLU K. (2017). Population based training of neural networks. *arXiv*.

JOSHI M., CHEN D., LIU Y., WELD D. S., ZETTLEMOYER L. & LEVY O. (2020). SpanBERT : Improving Pre-training by Representing and Predicting Spans. *TACL*.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). Scaling laws for neural language models. *CoRR*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., A. ALLAUZEN B. C., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. In *LREC*.

MARTIN L., MULLER B., SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE E. V., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty french language model. In *ACL*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *ICLR*.

PENNINGTON J., SOCHER R. & MANNING C. (2014). GloVe : Global vectors for word representation. In *EMNLP*.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *NAACL*.

PRICE P. J. (1990). Evaluation of spoken language systems : The atis domain. In *HLT*.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter. In *NIPS*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *NIPS*.

XU W., HAIDER B. & MANSOUR S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In *EMNLP*.