

Benchmarking Transformers-based models on French Spoken Language Understanding tasks

Oralie Cattan, Sahar Ghannay, Christophe Servan, Sophie Rosset

▶ **To cite this version:**

Oralie Cattan, Sahar Ghannay, Christophe Servan, Sophie Rosset. Benchmarking Transformers-based models on French Spoken Language Understanding tasks. INTERSPEECH 2022, Sep 2022, Incheon, South Korea. hal-03715340v2

HAL Id: hal-03715340
https://hal.science/hal-03715340v2

Submitted on 19 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Benchmarking Transformers-based models on French Spoken Language Understanding tasks

Oralie Cattan[1,2]*, Sahar Ghannay*[1]*, Christophe Servan*[1,2]*, Sophie Rosset*[1]

[1]Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
[2] QWANT, 10 boulevard Haussmann, 75009 Paris, France

[1]firstname.lastname@lisn.upsaclay.fr
[2]inital.lastname@qwant.com

Abstract

In the last five years, the rise of the self-attentional Transformer-based architectures led to state-of-the-art performances over many natural language tasks. Although these approaches are increasingly popular, they require large amounts of data and computational resources. There is still a substantial need for benchmarking methodologies ever upwards on under-resourced languages in data-scarce application conditions. Most pre-trained language models were massively studied using the English language and only a few of them were evaluated on French. In this paper, we propose a unified benchmark, focused on evaluating models quality and their ecological impact on two well-known French spoken language understanding tasks. Especially we benchmark thirteen well-established Transformer-based models on the two available spoken language understanding tasks for French: MEDIA and ATIS-FR. Within this framework, we show that compact models can reach comparable results to bigger ones while their ecological impact is considerably lower. However, this assumption is nuanced and depends on the considered compression method.

Index Terms: French Language, Language Models, Spoken Language Understanding, Benchmarking, Model costs

1. Introduction

The development of pre-trained language models based on Transformer architectures [1], such as BERT [2] has recently led to significant progress in the field of natural language processing (NLP). However, the trend of training large pre-trained language models on ever larger corpora, with an ever-increasing amount of parameters has raised questions related to the usability of these approaches [3]. Because these models require considerable computational resources, major efforts have been made to develop compact models in order to reduce the cost of using them. Compact models offer alternatives to the energy-intensive models with comparable performances while reducing their computational complexity and size. These models also allow to solve some industrial problems related to online speech processing. In particular, some applications (speech recognition, speech to text, etc.) have some known problems associated to network latency, transmission path difficulties, or privacy concerns.

Spoken Language Understanding (SLU) in language dialogue systems refers to the task of producing a semantic analysis and a formalization of the user's utterance. SLU traditionally encompasses the processes of determining a broad range of information conveyed in dialogue such as identifying the domain, the intent and the concepts of the conversation.

Benchmarking models have been extensively used in the performance evaluation of NLP-based systems [4, 5, 6]. With respect to our task of interest, recent research has focused on evaluating word embeddings representations. Word embeddings have proven to be effective in capturing semantic relationships between words. They are also an essential element of deep learning-based architectures.

In [7], contextual (ELMo [8]) and flat (Word2Vec [9], GloVe [10] and Fast-Text [11]) representations have been evaluated in order to investigate different input representations and their influence on the results obtained in SLU tasks. They highlighted the competitiveness of Word2Vec and ELMO on the French SLU corpus: MEDIA. More recently, [12] investigated the transferability of two French pre-trained BERT models [2] and their integration in BiLSTM and BiLSTM+CNN-based architectures. They obtained state-of-the-art results on MEDIA using CamemBERT [13] model. Lately, [14] gave a comparison of various cross-lingual transfer approaches including the use of the multilingual BERT encoder, evaluated on MultiATIS++, a multilingual corpus for the task of language understanding. They show that mBERT brings substantial improvements on multilingual training and cross-lingual transfer tasks, yielding up to 1.4% of relative improvements on the French subcorpus.

While there are now many English corpora and models for various tasks, there are still very few resources in French. Direct comparisons between SLU models are difficult because of the lack of a unified evaluation framework and size diverse Transformer-based models in French.

Contributions[1]: In this study we are investigating the use of transformer-based architectures for French language in solving an SLU problem, a concept detection task [15]. We also assess the impact related to model compacteness on SLU performance and their ecological impact (e.g.: their CO_2 cost). We focus on the benchmarking of the existing French and multilingual Transformer-based models on two French SLU corpora: MEDIA and ATIS-FR. Models considered for this benchmark are detailed in section 2, the experiments, the data, the model optimization and the experimental protocol are described in section 3. Section 4 presents the results, and section 5 proposes an analysis of some chosen results.

2. Transformer models considered

In order to compare the models performances on French SLU tasks, we highlight in this section and in Table 1 some characteristics of the considered models.

Among the language models that have been proposed in recent years we consider the most commonly used multilingual benchmarking models that have "reasonable" resource

[1]This work has been made possible thanks to the Saclay-IA computing platform

Model	Objectives	Data	Vocabulary size	Tokenization	# parameters	Model size
FlauBERT$_{\text{base}}$	MLM	24 French subcorpora (71 Gb of text)	68,729	BPE	138 M	553 Mb
CamemBERT$_{\text{base}}$	MLM	French OSCAR (138 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	French CCNet (135 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	French OSCAR (4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	French CCNet (4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	French Wikipedia (4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{large}}$	MLM	French CCNet (135 Gb of text)	32,005	SentencePiece	335 M	1.35 Gb
FrALBERT$_{\text{base}}$	MLM and SOP	French Wikipedia (4 Gb of text)	32,005	SentencePiece	12 M	50 Mb
XLM-R$_{\text{base}}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	278 M	1.12 Gb
XLM-R$_{\text{large}}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	559 M	1.24 Gb
mBERT$_{\text{base}}$	MLM and NSP	Wiki-100	119,547	WordPiece	177 M	714 Mb
small-mBERT$_{\text{base}}$ FR	MLM and NSP	Wiki-100	24,495	WordPiece	104 M	420 Mb
distil-mBERT$_{\text{base}}$	MLM and NSP	Wiki-100	119,547	WordPiece	134 M	542 Mb

Table 1: *Model characteristics, in terms of pre-training settings, sources of data and models size.*

consumption: XLM-R [16] and mBERT [2], as well as the French models: CamemBERT [13] and FlauBERT [5] [2] since these are the only two available large models for French at the time of writing. We also evaluate compact models that are scalable Transformer-based models. We exploit distil-mBERT [18], a distilled version of mBERT, small-mBERT [19] a mBERT model whose original vocabulary has been reduced to the French language and FrALBERT [3], a recently released model for French.

As mentioned in [20], the performance of language models is determined by multiple factors such as the pre-training objective, the layers specification or the dataset size.

Indeed, pre-training objectives vary by model and may depend on the downstream task being solved [21]. BERT-like models adopt the masked language modeling (MLM) and next sentence prediction (NSP) objectives. MLM is a fill-in-the-blank task consisting of predicting tokens of the input sequence that have been masked whereas NSP is a binary classification task to predict whether two segments are adjacent in the original text. XLM-R differs from mBERT solely in the pre-training procedure, eliminating the NSP task to address its ineffectiveness. This is also the case for CamemBERT and FlauBERT models. FrALBERT pre-training objective consists of MLM and Sentence Order Prediction (SOP) objective. SOP effectively models inter-sentence coherence by predicting whether a sentence order in a given sentence pair is swapped or not.

For monolingual and multilingual language modeling, the quality and representativeness of the large-scale datasets used are important for efficient modeling and for producing good generalization. Most models are pre-trained on the content of either Wikipedia, the web pages gathered by Common Crawl (CC) or a combination of diverse corpora. This represents a few gigabytes to several hundreds or even several terabytes of text. The other noteworthy difference is the vocabulary size with several tens of thousands of tokens for monolingual models to several hundred for multilingual models.

Models like Text-To-Text Transfer Transformer (T5) [22] or GPT-3 [23], have been excluded from our study, since they pose several usability problems (emphasized in the introduction). Although the GPT-3 and T-5 models are models for English, the data used for their pre-training contains many other languages, however, these languages represent only a small portion of the dataset compared to English (93% in word count for GPT-3). Furthermore, in terms of size, GPT-3 and T5 are huge compared to the size of established general NLP models such as BERT, and these earlier models are already quite expensive to run on GPUs. For example, GPT-3 is trained on billions of parameters, 470 times larger than the BERT model. We highlight the question of their usability in the introduction section.

In terms of availability finally, BERT is an open-source tool and easily available for users to access and fine-tune according to their needs and solve various downstream tasks. GPT-3 on the other hand is not open-sourced. It has limited access to users.

3. Experiments

Experiments are conducted on two well known French SLU task: MEDIA and ATIS-FR.

3.1. Datasets

The French MEDIA[3] corpus, composed of 1258 transcribed dialogues, which is about hotel reservation and information [15]. The corpus was manually annotated with semantic concepts characterized by a label and its value. There are in total 76 semantic labels. The corpus is split into three parts: a training corpus composed of 13k sentences, a development corpus composed of 1.3k sentences, and a test corpus composed of 3.5k sentences. The MEDIA task is also known as one of the most challenging slot-filling tasks, according to [24].

The French version of the Air Travel Information System (ATIS) corpus from the recent MultiATIS++ extension [14], named ATIS-FR, concerns flight information [25]. The ATIS-FR corpus corresponds to the manual translation of the original ATIS sentences (in English). It is composed of 84 semantic labels, and is split into three parts: a training corpus composed of 4.5k sentences, a development corpus composed of 490 sentences, and a test corpus composed of 893 sentences.

3.2. Experimental protocol

Spoken language understanding (SLU) is considered here as a sequence labeling task that assigns a concept label (from a predefined set) to each token of a sentence. We follow BIO-tagging scheme, where each concept is associated with two labels, B-label (for Beginning) and I-label (for Intermediate). Finally, the O-label (for Other) identifies the non-concept tokens. SLU performance is evaluated in terms of F-measure or F1 and Concept Error Rate (CER). The CER score is the official metric used in

[2] We did not include the small CamemBERT models [17] and the large version of FlauBERT, the former not being available and the latter not converging in our experiments despite our efforts.

[3] MEDIA is available for academic use: *https://catalogue.elra.info/ELRA-S0272*.

the MEDIA campaign [15] which is estimated in the same way as the classical word error rate but applied to semantic concepts instead of words (the lower the better). The significant results are marked with a star and measured using the 95% confidence interval.

We rely in this study on the standard approach introduced in [2], that corresponds to a model-based transfer learning method, used to facilitate the modeling of the target task with the knowledge learned from the language modeling task. Specifically, it consists of adding on top of the pre-trained model a token-level classifier with two hidden linear layers (with ReLU activations and dropout) that takes as input the last hidden state of the input sequence and outputs probabilities over concepts. In our experiments, we use Adam optimizer and conduct an automated hyperparameter optimization on a development dataset.

This optimization is based on the population-based learning algorithm [26], with proven efficiency, and in which a population of models and their hyperparameters are jointly optimised. Among the hyperparameters considered are the number of training epochs from 5 to 100, the batch size in the interval of 8 and 32 or the learning rate in the range between 1 and 5.

4. Results

This section presents the evaluation results of the different Transformer-based pre-trained models described in section 2 on both French datasets MEDIA and ATIS-FR.

4.1. MEDIA results

Performances on the MEDIA test dataset are presented in the two first columns of the Table 2. Monolingual models scores are very close and vary from 89 to 90 of F1 while CER varies between 7.5 and 8.6. FlauBERT$_{\text{base}}$ gets the worst F1 score (89.0) while CamemBERT$_{\text{base, Wiki 4 Gb}}$, pre-trained only on 4 Gb of Wikipedia, gets the best F1 score (90.0) with a CER score of 8.4. Considering the CER, as the official metric of the MEDIA task, the best model is CamemBERT$_{\text{base, CCNet 135 Gb}}$, a base CamemBERT model pre-trained with CCNet on 135 Gb of text. It obtains the lowest CER, at 7.5 for an F1 at 89.9.

The multilingual models (mBERT$_{\text{base}}$, distill-mBERT$_{\text{base}}$, small-mBERT$_{\text{base-fr}}$, XLM-R$_{\text{base}}$ and XLM-R$_{\text{large}}$) obtained CER scores ranging from 10.1 to 8.0, for the worst one, distill-mBERT$_{\text{base}}$, and the best one XLM-R$_{\text{large}}$, respectively. Regarding the F1 scores, they are comparable to the French monolingual models, for XLM-R models with 89.5 for the *base* one and 89.9 for the *large* one. On the other side, multilingual BERT and its *distilled* and *small* versions obtained performances comparable to the FlauBERT$_{\text{base}}$, with 88.9 F1 at best.

4.2. ATIS-FR results

The performance of the monolingual models on the ATIS-FR task in terms of F1 varies between 92.5 and 94.1 and between 3.3 and 5.3 of CER (Table 2). FlauBERT gets the worst F1 and CER scores while the large version of the CamemBERT model gets the best F1 and CER scores. FrALBERT and CamemBERT$_{\text{base, Wiki 4 Gb}}$ obtain similar performance with 0.3 points of F1 and 0.1 point of CER difference. The best model is CamemBERT$_{\text{large, CCNet 135 Gb}}$, which obtains the lowest CER, at 3.3 for an F1 at 94.1. The F1 scores vary from 88.1 (CER at 6.0) for the *distilled* version of mBERT to 93.6 (CER at 5.0) for mBERT$_{\text{base}}$. The two versions of XLM-R obtained comparable results in terms of CER and F1 scores to CamemBERT$_{\text{base, Wiki 4 Gb}}$ and FrALBERT$_{\text{base, Wiki 4 Gb}}$.

	MEDIA		ATIS-FR	
Model	**F1**	**CER**	**F1**	**CER**
FlauBERT$_{\text{base}}$	89.0	8.1	92.5	*5.3
CamemBERT$_{\text{large, CCNet 135 Gb}}$	89.2	7.8	**94.1**	***3.3**
CamemBERT$_{\text{base, CCNet 135 Gb}}$	89.9	***7.5**	94.0	3.7
CamemBERT$_{\text{base, OSCAR 138 Gb}}$	89.3	7.9	93.9	3.7
CamemBERT$_{\text{base, OSCAR 4 Gb}}$	89.7	8.3	93.6	3.7
CamemBERT$_{\text{base, CCNet 4 Gb}}$	89.7	8.3	93.8	3.8
CamemBERT$_{\text{base, Wiki 4 Gb}}$	***90.0**	8.4	92.5	4.2
FrALBERT$_{\text{base, Wiki 4 Gb}}$	89.8	8.6	92.8	4.3
XLM-R$_{\text{base}}$	89.5	8.5	92.5	4.3
XLM-R$_{\text{large}}$	89.9	8.0	92.7	4.4
mBERT$_{\text{base}}$	88.9	8.7	93.6	5.0
distill-mBERT$_{\text{base}}$	*87.5	*10.1	*88.1	*6.0
small-mBERT$_{\text{base-fr}}$	*88.8	*8.1	93.3	*5.3

Table 2: *SLU performances on the MEDIA and ATIS-FR test dataset. Results are given in terms of F-measure (F1) and Concept Error Rate (CER). Significant results are marked with a star.*

5. Analysis

5.1. In deph analysis

Benchmarking Transformer-based models on the two French SLU tasks (MEDIA & ATIS-FR) allows to observe some trends. In both tasks, the CamemBERT models perform the best, in terms of Concept Error Rate (CER). FrALBERT obtains comparable results to CamemBERT$_{\text{base, Wiki 4 Gb}}$, this may come from that the two models are trained on the same kind and amount of data (Wikipedia 4G). We also observe that FrALBERT has comparable performances to XLM-R$_{\text{base}}$ in both tasks even if the training data and structure are both different. In addition, we notice the underachievement of distill-mBERT$_{\text{base}}$ compared to other BERT models and especially to FrALBERT. Finally, FlauBERT$_{\text{base}}$ has comparable results to CamemBERT models in the MEDIA task, but the CER score of FlauBERT in the ATIS-FR task, is significantly worse than the CamemBERT models, which led us to go deeper in the analysis.

We propose to focus the analysis on the most representative labels of each task. For MEDIA we focus on COMMAND-TACHE, TEMPS-DATE, NOMBRE-CHAMBRE, LOCALISATION-VILLE, NOM-HOTEL, OBJET, labels. Their F1 varies between 70.15 and 95.76. In the ATIS task, we focus on CITY_NAME, AIRLINE_NAME, DEPART_TIME.PERIOD_OF_DAY,DEPART_DATE.DAY_NAME, TOLOC.CITY_NAME, FROMLOC.CITY_NAME labels and their F1 varies between 59.41 and 100.

The first trend we observed is an underperformance of the distill-mBERT$_{\text{base}}$ on named entity tags on both tasks. For instance, STATE NAME, CITY_NAME in ATIS or NOM-HOTEL, LOCALISATION-VILLE for MEDIA have up to 2 F1 points less than the other models. Note that these tags are ones of the most frequent in both tasks. On the other side, we could not observe such a big trend by comparing larger models and compact models, or multilingual models versus French monolingual models. Differences between models are very small, for instance the CamemBERT$_{\text{base, Wiki 4 Gb}}$ model will be better than the FrALBERT$_{\text{base, Wiki 4G}}$ on the MEDIA labels NOMBRE-CHAMBRE (93,14 versus 91,85 of F1 respectively), NOM-HOTEL (82,76 versus 79,60 of F1 respectively).

When diving deeper in the MEDIA task, we can detect some small trends when we look at the whole results. For in-

Steps	Fine-tuning *(1 epoch)*						Inference					
Tasks	MEDIA			ATIS-FR			MEDIA			ATIS-FR		
Models	**Time (s)**	**Energy (kWh)**	**CO_2 (g)**	**Time (s)**	**Energy (kWh)**	**CO_2 (g)**	**Time (s)**	**Energy (kWh)**	**CO_2 (g)**	**Time (s)**	**Energy (kWh)**	**CO_2 (g)**
$\text{FlauBERT}_{\text{base}}$	121.89	765.24	554.04	52.08	231.13	167.34	9.44	26.52	19.20	1.44	4.01	2.90
$\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$	144.31	659.34	477.36	56.65	345.33	250.02	23.67	66.58	48.20	3.44	9.64	6.98
$\text{CamemBERT}_{\text{base, OSCAR 138 Gb}}$	130.06	789.69	571.74	53.39	234.84	170.02	9.46	26.58	19.24	1.53	4.23	3.06
$\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$	118.58	671.42	486.11	51.67	230.01	166.53	7.28	20.44	14.80	1.19	3.22	2.33
$\text{CamemBERT}_{\text{base, OSCAR 4 Gb}}$	116.59	623.44	451.37	51.66	229.96	166.49	7.43	20.87	15.11	1.18	3.27	2.37
$\text{CamemBERT}_{\text{base, CCNet 4 Gb}}$	115.54	662.78	479.86	50.74	227.35	164.60	7.31	20.53	14.86	1.21	3.31	2.40
$\text{CamemBERT}_{\text{base, Wiki 4 Gb}}$	109.57	645.96	467.68	50.40	226.39	163.91	7.19	20.19	14.62	1.17	3.24	2.35
$\text{FrALBERT}_{\text{base, Wiki 4 Gb}}$	72.65	474.69	343.67	28.55	97.26	70.41	4.26	11.95	8.65	0.65	1.78	1.29
$\text{XLM-R}_{\text{base}}$	125.26	549.48	397.82	56.30	243.02	175.94	8.04	22.59	16.35	1.21	3.34	2.42
$\text{XLM-R}_{\text{large}}$	196.74	1 155.15	836.33	64.08	433.70	314.00	26.03	73.24	53.02	3.75	9.88	7.61
$\text{mBERT}_{\text{base}}$	119.36	673.56	487.66	52.61	232.65	168.44	8.39	23.56	17.06	1.16	3.21	2.33
$\text{distill-mBERT}_{\text{base}}$	80.10	545.46	394.91	49.48	240.96	166.53	7.04	19.75	14.30	1.10	3.02	2.18
$\text{small-mBERT}_{\text{base-fr}}$	112.08	589.84	427.04	50.67	227.17	164.47	7.69	21.59	15.63	1.15	3.19	2.31

Table 3: *Estimation of fine-tuning and inference costs on MEDIA and ATIS-FR corpora.*

stance, even if the OBJET label is one of the most frequent, it seems that all models have difficulties to correctly detect it. In the same way, the label NOM is highly difficult, even for one of the best model ($\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$). It seems that the biggest difference between models occurs in their ability to transfer their knowledge according to the amount of data used for pretraining. In this way, the models trained with the biggest corpora are enable to handle the best, the most infrequent labels.

What we observe in the MEDIA task can also be observed in the ATIS-FR task, even if the overall performance of the models is higher. This leads us to suggest that a better sample of the training data could bring interesting results. Moreover, recent works in few-shot learning for slot-filling approaches [27] enable models to perform at high levels with a strong subsampling of examples, which in this case could be a plus.

5.2. Ecological and computational costs

We conducted an impact study of Transformer models we considered by measuring the ecological impact of these models [28]. The table 3 presents the time to process one fine-tuning epoch, and one inference step, associated to the energy consumed and the CO_2 produced, for each task. Each score is the mean of five runs for each step.

$\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$ and $\text{XLM-R}_{\text{large}}$ produce the most over all tasks and steps. $\text{FlauBERT}_{\text{base}}$, $\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$ and $\text{mBERT}_{\text{base}}$ produce nearly the same amount of CO_2 during fine-tuning and inference steps and for the both tasks (MEDIA and ATIS-FR), which lead us to correlate the amount of parameters and their impact in terms of energy consumed and CO_2 produced.

$\text{FrALBERT}_{\text{base, Wiki 4 Gb}}$ is the model which uses the less energy (in both fine-tuning and inference steps), and produces the less CO_2 of all models. Especially, in the inference step, the FrALBERT model consumes nearly the half of other compact models: $\text{distill-mBERT}_{\text{base}}$ and $\text{small-mBERT}_{\text{base-fr}}$.

This study shows compact models like FrALBERT, which obtained comparable results to bigger models (especially with large BERT models in Table 2), have also an important lower ecological impact than big Transformer models. For instance, the impact of FrALBERT is more than 5 time lower than the impact of large models such as $\text{XML-R}_{\text{large}}$ and $\text{CamemBERT}_{\text{large}}$ in the inference step for both tasks and in the fine-tuning step, the FrALBERT model is more than 3 times lower than the $\text{XML-R}_{\text{large}}$ model.

Surprisingly, the compression approaches of compact models may not have the same impact. We can observe that $\text{distill-mBERT}_{\text{base}}$ is a little bit different from $\text{small-mBERT}_{\text{base}}$ with a resource requirement and thus an impact a little bit lower but still quite comparable to the $\text{mBERT}_{\text{base}}$ model, but with significantly lower performances, compared to mBERT.

6. Conclusion

Transformer-based architectures are currently the state-of-the-art model for many NLP tasks. In this study we have proposed to benchmark the existing French and multilingual Transformer-based pre-trained models, for the purpose of comparing their performance on two French SLU corpora: MEDIA and ATIS-FR. We have also assessed the ecological impact related to model compacteness on SLU performances and conducted an extensive side-by-side comparison of thirteen recently proposed efficient Transformer models.

The experimental results show that these tasks are very challenging even for large models. Experimental results show that for both tasks, the CamemBERT models perform the best, in term of Concept Error Rate (CER). The best CER results on MEDIA and ATIS-FR are respectively achieved by $\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$ and $\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$. In both tasks, the French compact model FrALBERT obtains comparable results to the large model CamemBERT (*base* and *Wiki 4 Gb* configuration). Moreover, this model achieves comparable performances to multilingual models in both tasks and outperforms the distilled version.

Then, our detailed analysis of F1 scores provides interesting model insights in each task. From those analyses we observe that the biggest difference between models occurs in their ability to transfer their knowledge according to the amount of pre-trained data.

Finally, the ecological study conducted on these models shows that compact models have significantly less ecological impact compared to big Transformer models in both fine-tuning and inference steps. This less CO_2 is a plus in a context of reducing our ecological impact but it also means less energy consumed, which can reach more than 5 time less between the FrALBERT model and large BERT models.

We plan to open source our code and benchmarks to facilitate future benchmarking, research and model development.

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NACL*, 2019.

[3] O. Cattan, C. Servan, and S. Rosset, "On the Usability of Transformers-based models for a French Question-Answering task," in *RANLP*.

[4] J. Guo, Q. Liu, J. G. Lou, Z. Li, X. Liu, T. Xie, and T. Liu, "Benchmarking meaning representations in neural semantic parsing," in *EMNLP*, 2020.

[5] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, B. C. A. Allauzen, L. Besacier, and D. Schwab, "FlauBERT: Unsupervised language model pre-training for French," in *LREC*, 2020.

[6] I. A. Farha and W. Magdy, "Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection," in *WANLP*, 2021.

[7] S. Ghannay, A. Neuraz, and S. Rosset, "What is best for spoken language understanding: small but task-dependant embeddings or huge but out-of-domain embeddings?" in *ICASSP*, 2020.

[8] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL*, 2018.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[10] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *EMNLP*, 2014.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, 2017.

[12] S. Ghannay, C. Servan, and S. Rosset, "Neural networks approaches focused on French spoken language understanding: application to the MEDIA evaluation task"," in *COLING*, 2020.

[13] L. Martin, B. Muller, P. J. Suárez, Y. Dupont, L. Romary, E. V. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty french language model," in *ACL*, 2020.

[14] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," in *EMNLP*, 2020.

[15] H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefevre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, and J. Villaneau, "Results of the French Evalda-Media evaluation campaign for literal understanding," in *LREC*, 2006.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *ACL*, 2020.

[17] V. Micheli, M. d'Hoffschmidt, and F. Fleuret, "On the importance of pre-training data volume for compact language models," in *EMNLP*, 2020.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *NIPS*, 2019.

[19] A. Abdaoui, C. Pradel, and G.Sigel, "Load what you need: Smaller versions of mutililingual BERT," in *SustaiNLP*, 2020.

[20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *CoRR*, 2020.

[21] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving Pre-training by Representing and Predicting Spans," *TACL*, 2020.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[23] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[24] F. Béchet and C. Raymond, "Benchmarking benchmarks: introducing new automatic indicators for benchmarking spoken language understanding corpora," in *InterSpeech*, 2019.

[25] P. J. Price, "Evaluation of spoken language systems: The atis domain," in *HLT*, 1990.

[26] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu, "Population based training of neural networks," *arXiv*, 2017.

[27] O. Cattan, S. Rosset, and C. Servan, "On the cross-lingual transferability of multilingual prototypical models across NLU tasks," in *META-NLP*, 2021.

[28] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1–43, 2020.