ALIBERT: A PRETRAINED LANGUAGE MODEL FOR FRENCH BIOMEDICAL TEXT A PREPRINT

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, Jean-Daniel Zucker

ALIBERT: A PRETRAINED LANGUAGE MODEL FOR FRENCH BIOMEDICAL TEXT

A PREPRINT

Aman Berhe*
Quinten
IRD, Sorbonne University, UMMISCO,
91, bvd Hopital, F-75013, Paris, France
amanzaid.berhe@ird.fr

Guillaume Draznieks*
Quinten
8 Rue Vernier, 75017 Paris
gdraznieks@student.ethz.ch

Vincent Martenot
Quinten
8 Rue Vernier, 75017 Paris
v.martenot@quinten-france.com

Valentin Masdeu
Quinten
8 Rue Vernier, 75017 Paris
v.masdeu@quinten-france.com

Lucas Davy
Quinten
8 Rue Vernier, 75017 Paris
l.davy@quinten-france.com

Jean-Daniel Zucker
IRD, Sorbonne University, UMMISCO,
INSERM, Sorbonne University, NUTRIOMICS,
91, bvd Hopital, F-75013, Paris, France
jean-daniel.zucker@ird.fr

ABSTRACT

Over the past few years, domain specific pretrained language models have been investigated and have shown remarkable achievements in different downstream tasks, especially in biomedical domain. These achievements stem on the well known BERT architecture which uses an attention based self-supervision for context learning of textual documents. However, these domain specific biomedical pretrained language models mainly use English corpora. Therefore, non-English, domain-specific pretrained models remain quite rare, both of these requirements being hard to achieve. In this work, we proposed AliBERT, a biomedical pretrained language model for French and investigated different learning strategies. AliBERT is trained using regularized Unigram based tokenizer trained for this purpose. AliBERT has achieved state of the art F1 and accuracy scores in different down-stream biomedical tasks. Our pretrained model manages to outperform some French non domain-specific models such as CamemBERT and FlauBERT on diverse down-stream tasks, with less pretraining and training time and with much less corpora.

1 Introduction

Recent contextual language models have achieved a tremendous results in almost all domains using textual information. Transformers [Vaswani et al., 2017] based pretrained language models (T-PLM) have contributed and continue to contribute to the success of natural language processing (NLP) in multiple domains of expertise. Furthermore, very large transformer based models which require hundreds of billions of parameters have shown extra-ordinary achievements and became more accessible. However, these huge pretrained language models (PMLs) such as Generative pretrained Transformer (GPT-3) and No Language Left Behind (NLLB) [NLLB Team et al., 2022] have addressed different languages in the general domain of NLP tasks but not domain specific NLP.

*The first two authors have equal contribution.

Although, there are a few that are multi-lingual [Devlin et al., 2018, NLLB Team et al., 2022], multi-domain pretrained language models (PLMs) [Maronikolakis and Schütze, 2021], are very scarce. Yet, PLMs can be trained in domain specific and language specific for better performance. The biomedical field is one of the most important domains and its associated textual corpora is one of the first rapidly growing sources of information in multiple languages. Hence, researchers have been taking advantage of PLMs to represent biomedical knowledge from different sources, following their success in the general domain. There are quite fascinating biomedical pretrained language models (B-PLMs) that have achieved interesting findings and that help on the decision making of biomedical domain, such as, BioBERT [Lee et al., 2020], PubMedBERT [Gu et al., 2022], BioELECTRA [raj Kanakarajan et al., 2021], etc.

PLMs are trained using different training mechanisms. The most common are masked language modeling (MLM) [Devlin et al., 2019], replaced token detection (RTD) [Clark et al., 2020] or Next Sentence Prediction (NSP) [Devlin et al., 2019]. Training a biomedical language model using different strategies will benefit the different down stream tasks. Furthermore, B-PLMs apply various pretraining methods since they borrow some characteristics from already existing PLMs. The commonly used pretraining methods are continual pretraining (CPT), mixed domain pretraining, domain specific pretraining (DSPT), etc. In this work, DSPT was used for training, our proposed model, from scratch using domain specific French corpora. Moreover, these B-PLMs use tokens as their input. Tokenization is the basic step in the training of the language models, since it can directly be used as discrete input to the pretraining of the models. There are different ways to tokenize a text input. The most common tokenization techniques are Byte Pair Encoding (BPE) (such as; SentencePiece, WordPiece, etc) and Unigram sub-word based tokenization. Considering and implementing different tokenization technique is equally important for better performance of the B-PLMs, specially when the model is language specific. Language-specific PLMs can use common tokenization techniques like BPE, moreover they can adapt the tokenization process and train a tokenizer that can fit to a specific language and domain under consideration. In similar way the biomedical text differs from general domain texts, hence using tailored tokenization helps to better represent most words of biomedical vocabulary.

Another key stage to consider while training huge PLMs is hyperparameters optimization. Optimization of the hyperparameters of the T-PLMs have an impact on the performance and training time of the models. One of the hyperparameters in the deep architecture of transformers is its optimizer; that help to minimize the output of the loss function. Recently, an optimizer known as LAMB —which stands for "Layer-wise Adaptive Moments optimizer for Batch training"— was shown to greatly reduce the training time of BERT [Devlin et al., 2018] from 3 days to 76 minutes [You et al., 2019]. Hence, choosing the right optimizer can speed up the training of PLMs. In our work we did investigate the effectiveness of the LAMB optimizer compared to other commonly used optimizers.

However, biomedical languages models in other languages (other than English), a domain and language specific PLMs, are quite scarce. In the language specific NLP models there are some works that focus on French language, such as CamemBERT [Martin et al., 2020] and FlauBERT [Le et al., 2019]. French is a very rich language and French-based PLMs [Martin et al., 2020, Le et al., 2019] has shown the importance of such a model for different purposes. However, French biomedical textual information have not been represented using transformers based PLM. There are some word embedding of a French language in different domains. [Dynomant et al., 2019] compared different word embedding techniques (word2vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014]) for a French health related documents. Considering the drawbacks of word embedding for word representation it is necessary to build B-PLMs for better representation. In this work, we propose AliBERT (named after the renowned french dermatologist), a BERT-based language-specific and domain-specific Biomedical language model. AliBERT uses a masked language model (MLM) pretraining mechanism which randomly masks some of the tokens from the input biomedical text and predicts the masked tokens based on the context of the input. Thereby learning the context of each word according to the biomedical text input. A Unigram based tokenizer with a novel regularization algorithm is trained for the pretraining purpose of AliBERT. In addition to the MLM, we have also trained ELECTRA-based [Clark et al., 2020] models called AliBERT-ELECTRA. AliBERT-ELECTRA is trained using the replaced token detection mechanism using the same vocabularies and tokenization steps as AliBERT. In addition, the LAMB optimizer is studied to analyze its computational speed gain during model pretraining. Here are the main contributions of our work:

- A French biomedical language model, a language-specific and domain-specific PLM, which can be used to represent French biomedical text for different downstream tasks.
- A normalization of a Unigram sub-word tokenization of French biomedical textual input which improves our vocabulary and overall performance of the models trained.
- AliBERT outperforms other French PLMs in different downstream tasks. AliBERT models and code will be available at https://gitlab.par.quinten.io/qlab/dagobert/-/tree/main

This paper is organized in the following manner: first the related work is briefly discussed in section 2, different language-specific and domain specific PLMs and their pretraining objectives and strategies are discussed. Second,

section 3 presents our B-PLM AliBERT with details on architecture, tokenization and optimization. Then, section 4 discusses the fine-tuning and evaluation of our models in downstream tasks. Next, section 5 explain the experiments and results on the down-stream tasks. Then, section 6 discusses the results found and the drawbacks we encountered in detail. Finally, section 7 concludes the findings of this paper and points out our future directions concerning the domain-specific and language-specific PLMs.

2 Related work

In the recent years, there have been a fast growing number of transformer based language models and their performance have been remarkable in many domains. The corner stone of these models is the attention architecture known as "attention is all you need" [Vaswani et al., 2017] which is composed of an encoder, decoder and an attention mechanism. The pioneers of transformers based PLMs (T-PLMs) are BERT [Devlin et al., 2018] and GPT [Radford et al., 2018] which are transformer encoder and transformer decoder based models, respectively. Consequently, the T-PLMs can be mainly divided as transformer encoder based models such as ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], ELCTRA [Clark et al., 2020], and transformer decoder based model such as BART [Lewis et al., 2019], PEGASUS [Zhang et al., 2019], and T5 [Raffel et al., 2020]. [Devlin et al., 2018] played an important role for the increase of T-PLMs and fine-tuning many down-steam tasks. Devlin et al. also paved the way for other languages (other than English), such as [Martin et al., 2020, Le et al., 2019, Delobelle et al., 2020, Cañete et al., 2020], to develop language specific (monolingual) language models.

There are very few French language models [Martin et al., 2020, Le et al., 2019, Copara et al., 2020, Douka et al., 2021, Cattan et al., 2022]. CamemBERT [Martin et al., 2020] and FlauBERT [Le et al., 2019] are trained on general knowledge French corpora. CamemBERT used OSCAR[2] dataset which is composed of 130 Gigabytes (GB) of raw French text with 32.7 Billion tokens where as FlauBERT utilized 71 GB of raw text with 12.7 Billion of token. BERTweetFR [Guo et al., 2021] is another French PLM trained on French tweets. BERTweetFR is a general domain which is initialized using CamemBERT utilizing the largest French tweets corpora which is composed of 16 GB of 226 Million tweets. They took tweets with an average length of 30 tokens. Kamal Eddine et al. developed a BART based french language model names as BARThez which a generative language model based on BART[3] [Lewis et al., 2019]. BARThez used 66 GB (110 GB after tokenization) raw text for pretraining. Cattan et al. investigated the usability of transformer based models for French question answering task and provided a model known as FrALBERT which is based on a famous compact language models (parameter efficient BERT) known as ALBERT [Lan et al., 2019]. FrALBERT is a compact language model pretrained on the French version of the Wikipedia encyclopedia of 04/05/2021. Their dataset is composed of 4 GB of text and 17 million sentences. There are two French domain specific PLMs. JuriBERT [Douka et al., 2021] is a French legal language model (language and domain specific) which is trained on 6.3 GB of raw legal text[4]. CamemBioBERT [Copara et al., 2020] is a fine-tuned CamemBERT [Martin et al., 2020] using biomedical text from a French language challenge known as DEFT ("Défi Fouille de Textes")[5].

Moving on towards domain specific-language models, Lee et al.2020 built the first BERT based language model in English in the biomedical domain, known as BioBERT. BioBERT [Lee et al., 2020] is built on top of the BERT [Devlin et al., 2018] model using the abstract of articles from the PubMed [6] as well as PMC [7] full articles. BioBERT provided different sizes of pretrained models and they claimed to have achieved state of the art in multiple down-stream task including named entity recognition (NER), biomedical text classification, relation extraction, etc. Following the publication of BioBERT, there has been tremendous increase in biomedical language models. Thought the B-PLMs are variants of the pioneer transformers [Vaswani et al., 2017] architecture, they have key differences. Recently, a survey [Kalyan et al., 2021] has studied many publicly available language models in the biomedical domain and provided a survey of systematic literature review, known as AMMU. AMMU included 121 articles of biomedical language models that exist until their paper was published in 2021.

AMMU, Kalyan et al., have investigated the core B-PLMs concepts, such as pretraining methods, pretraining tasks, fine-tuning methods and embeddings. Furthermore, Kalyan et al. disclosed different types of corpora along with the language models that used the corpus. The main corpora included were electronic health record (EHR), radiology reports, social

[2]OSCAR is a set of monolingual corpora

[3]BART: De-noising Sequence-to-Sequence pretraining for Natural Language Generation, Translation, and Comprehension

[4]Number of token used in JuriBERT [Douka et al., 2021] not mentioned in the paper

[5]DEFT is a scientific evaluation campaign on Francophone text mining.

[6]PubMed comprises more than 34 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites

[7]PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM))

media texts and scientific literature. During pretraining, different B-PLMs took different learning objectives. Kalyan et al. have listed out the most common ways and these are Masked Language Modeling (MLM), Replaced Token Detection (RTD), Next Sentence Prediction (NSP), Sentence Order Prediction (SOP) and Span Boundary Objective (SBO). MLM was introduced by BERT and many B-PLMs [Lee et al., 2020, López-García et al., 2021] used this to train their model. MLM considers context from both directions (left and right) of an input tokens to predict the missing/masked token. There are different masking techniques of MLM, for example, *Dynamic* masking [Liu et al., 2019], *whole word* masking [Gu et al., 2022], *whole entity* masking [Lin et al., 2021] and *whole span* masking [Zhang et al., 2020]. RTD learns by checking whether each token is replaced or not. It has two sub-models, a generator which predicts words for a masked words and a discriminator which detects if each prediction was the original word or not. RTD was first introduced by ELECTRA [Clark et al., 2020] then used by BioElECTRA [raj Kanakarajan et al., 2021] as a domain specific training. NSP was also introduced by BERT, it is training a model on sentence-level to predict if two sentences come one after the other or not (binary classification). SOP [Lan et al., 2019] is a recent sentence-level prediction strategy to learn the coherence between sentences. In the biomedical domain BioALBERT [Naseem et al., 2022] has used this technique. SBO is a pretraining task that learns on predicting the whole masked span of a context. SpanBERT used this technique for conceptual representation of biomedical text [Joshi et al., 2020]. More details is left for reader to enjoy the AMMU, survey paper [Kalyan et al., 2021]. We will focus on the non-English biomedical language models.

There are few non-English in-domain (biomedical) transformer based B-PLMs [Terumi Rubel Schneider et al., 2020, Bressem et al., 2020, López-García et al., 2021]. Most of the models are pretrained using the continual pretraining (CPT) approach which means they used already pretrained language specific general knowledge PTM or multilingual BERT (mBERT) [Devlin et al., 2018] as a starting point and continue the training using biomedical and clinical textual documents. Terumi Rubel Schneider et al. developed a B-PLM for Brazilian Portuguese using biomedical scientific paper abstracts and clinical notes. They initialized their models using mBERT provided by Devlin et al. and achieved state of the art Portuguese biomedical named entities recognition (NER). They have also experimented the combination of clinical notes and abstracts of biomedical scientific articles and they claimed the model trained with both datasets outperformed the other two models trained separately (using only one of the datasets). Cañete et al. developed a clinical coding model for Spanish medical documents using transformer based PLM. They developed different transformer based models by continuing the training from existing multilingual and language specific models such as mBERT [Devlin et al., 2018], BETO[8] [Canete et al., 2020] and XLM-RoBERTa [Conneau et al., 2019] using a private corpus of deidentified real-world oncology clinical texts written in Spanish. The corpora are composed of 30.9 thousand documents, 64.4 million words. Then, they further fine-tuned the models on clinical coding tasks using publicly available clinical Spanish corpora. López-García et al. showed that language and domain specific PLMs perform better than the mixed domain and multilingual language models. Bressem et al. developed German biomedical transformer based PLMs trained from already existing German language models such as mBERT, GermanBERT [Chan et al., 2020], etc. and continued training for biomedical text documents. They have also trained the BERT model from scratch and named it FS-BERT based on a text corpora compromised of 3.8 million unstructured radiology reports which are composed of around 4.16 billion words. Bressem et al. created a custom WordPiece vocabulary because the existing German language models did not have the vocabularies that come from radiology texts. There are more language specific B-PLMs that are note covered here, we invite readers to read AMMU [Kalyan et al., 2021], a survey in B-PLMs.

To the best of our knowledge, there is yet no any French biomedical transformer based PLM trained from scratch. From the literature we can clearly see that there is a gap in pretrained language models for French biomedical text mining. Hence filling this gap is our primary purpose besides improving the tokenization process for French biomedical texts. Furthermore, tokenization process has been regularized for French biomedical text rather than just using general tokenization methods.

3 AliBERT: a pretrained language model for French biomedical text

This sections focuses on how the proposed pretrained language model, `AliBERT`, was built. It describes the pretraining strategy and architecture, pretraining corpora, tokenization and optimization of our models.

3.1 Pretraining strategies

There are different kinds of pretraining strategies [Kalyan et al., 2021] to train a transformers based models, such as pretraining from scratch (PTS), continual pretraining (CPT), simulated pretrained (SPT), etc. as discussed in section 2. pretraining from scratch (PTS) is utilized for training AliBERT and its variants from scratch using biomedical corpora

[8]BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique.

for better representation of biomedical context of words. Training our models from scratch helps to represent vocabulary that only exist in biomedical text which will be discussed in subsection 3.3.

The models selected are based on the transformers [Vaswani et al., 2017] architecture and the famous BERT [Devlin et al., 2018] model is used as masked language model (MLM), transformers and BERT architecture will not be discussed here because they have been discussed enough in many research works. Therefore, the AliBERT is trained in the coarse of self-supervised learning by masking the 15% of the words from the input text (sequence of words). All necessary steps and configurations are discussed in the following sub-sections.

3.2 Pretraining data

The pretraining corpus is gathered from different sub-corpora of French biomedical textual documents. The sources used are Drug Database, RCT, biomedical articles from ScienceDirect[9], Thesis manuscripts in French and clean articles from Cochrane [10] database. It can be inferred from the names of the corpora, they cover various topics in the biomedical domain and they have different writing styles. Table 1 summarises the different corpora collected and used for pretraining AliBERT models.

Name	Type	Quantity	Size
Drug database	Description	23 K	550 Mb
RCP	Description	35 K	2200 Mb
Articles	Scientific articles	500 K	4300 Mb
Thesis	Thesis summaries	300 K	300 Mb
Cochrain	Articles pages	7.6 K	27 Mb

Table 1: Corpora used to pretrain AliBERT

The corpora were collected from different sources. Scientific articles are collected from the ScienceDirect using an API provided on subscription and French articles in biomedical domain are selected. The summaries of thesis manuscripts are collected form "Système universitaire de documentation (SuDoc)" which is a catalog of university documentation system. Short texts and some complete sentences were collected from the public drug database which lists the characteristics of tens of thousands of drugs. Furthermore, a similar drug database known as "Résumé des Caractéristiques du Produit (RCP)"[11] is also used to represent description of medications that are intended to be utilized by biomedicine professionals. Pages of biomedical articles from Cochrane are also collected. Hence, our corpus for pretraining is composed of around 7 gigabyte (GB) textual documents.

When comparing with the corpora of already existing French T-PLMs, our corpus is big enough to represent a biomedical text. Table 2 compares the different corpora used for pretraining french language models.

Model	Domain	Size	Source
CamemBERT [Martin et al., 2020]	general	138 GB	OSCAR
FlauBERT [Le et al., 2019]	general	71 GB	WMT19, OPUS, Wikmedia
BERTweetFR [Guo et al., 2021]	general	16 GB	French tweets
JuriBERT [Douka et al., 2021]	legal	6.3 GB	LégalFrance & Court of Causation
FrAlbert [Cattan et al., 2022]	general	4.0 GB	Wikipedia
AliBERT (Proposed)	biomedical	7.0 GB	ScineceDirect, SuDoc, Drug datbases and Cochrane

Table 2: Comparing the corpus with already existing corpora for French PLMs

3.3 Tokenization

In the case of PLMs, for a neural network, the tokenization is a key process of splitting a text into smaller input types, known as tokens. Most BERT based PLMs use sub-word tokenization scheme such as Bite Pair Encoding (BPE),

[9]ScienceDirect is a website which provides access to a large bibliographic database of scientific and medical publications of the Dutch publisher Elsevier.

[10]Cochrane is a British international charitable organisation formed to organise medical research findings to facilitate evidence-based choices about health interventions involving health professionals, patients and policy makers.

[11]The "Résumé des Caractéristiques du Produit" (RCP) database aims at providing more accurate information than the medication note given for medicines.

WordPiece and SentencePiece. However, the tokenization process can be regularized or trained to fit to the specific purpose and to represent a vocabulary in specific domain. Hence, it consists a series of rules, that include a learned vocabulary, to segment a text into tokens which are members of the learned vocabulary. Hence we decided to use our own tokenizer so that our vocabulary represent the necessary biomedical tokens.

A normalization step was used to enhance our vocabulary. In this step we added a space after and before every punctuation mark. These normalise the representation of the text, facilitate the tokenization and learning by the neural network. Hence, there is a significant reduction of duplicates, such as, ("MOT", "_MOT"),("_siècle","_siècles") which were introduced due to punctuation marks, like "(", " :", "-", etc. in the text.

We have trained different tokenizers, such as Unigram, WordPiece with different parameters (vocabulary size, regularization). Unlike BPE, Unigram starts from a big vocabulary and removes tokens until it reaches the desired vocabulary size. During training, at every step, Unigram computes a loss over the corpus given the current vocabulary. Then, for each symbol it calculates how much the overall loss would increase if the symbol was removed, and looks for the symbols that would decrease it the most. Figure 1 depicts the steps taken during tokenization with an example and compares Unigram tokenizers trained from scratch and the tokenizer from CamemBERT[Martin et al., 2020].

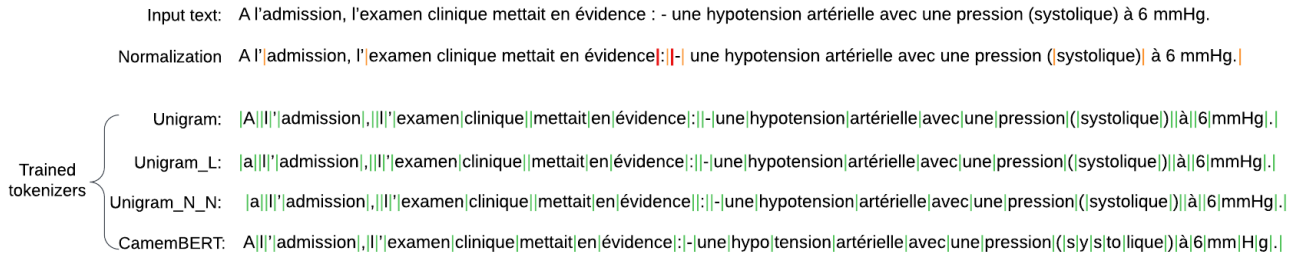

Figure 1: Normalization and tokenization example. During normalization step the input text is normalized by adding a space after the punctuation (shown by the orange lines) and removing a space before it (shown by the red lines) and then used to train the tokenizer (Unigram). The Unigram tokenizers are trained from scratch while developing AliBERT, Unigram uses text input as it is (does not change the cases), Unigram_L lower cased the input text and Unigram_N_N is the not-normalized version of Unigram and CamemBERT is the tokenizer used by CamemBERT [Martin et al., 2020], a French PLM.

3.3.1 Training configurations

During training a large language model, it is necessary to consider different configuration that are necessary for building a well performing model. Hence, the model architectures, training strategy, optimization and computation are key parameters to consider.

Model architectures and training: We have mainly developed two architectures of our French B-PLM namely AliBERT, a BERT [Lan et al., 2019] based and AliBERT-ELECTRA, an ELECTRA [Clark et al., 2020] based, models. BERT and ELECTRA differ only in their learning strategy. The former uses masked language modeling (MLM) and the later uses replaced token detection (RTD). AliBERT$_{base}$ have the same architecture with BERT$_{base}$which has a length (L) of 12, height (H) of 512 and a self-attention head (A) of 12.

Figure 2 illustrates how MLM works. A sequence of words is given as input and 15 % of the words are hidden. The input goes through the tokenization stage and the words are tokenized. The tokens are padded or truncated to have a maximum length of 512 tokens. Hence, special tokens "[CLS]","[PAD]" are added if the sequence length is less than 512 tokens. Then the embeddings of the tokens are passed to the transformer layers to learn the context of the input and the relationship of the tokens. Finally the output of the transformers is passed to a feed forward neural network to compute the probability distribution of the token to predict the masked toked words. For more detail on this training method see the original work BERT [Devlin et al., 2019].

Another strategy different from MLM is RTD, illustrated on the Figure 3. In RTD the objective is to predict which tokens have been replaced and which have not. A very simple pretrained model is used as generator to predict a masked word from the input text. Then, the predicted words are used to replace the masked inputs and the unmasked sentence is used as input text in the discriminator model. Finally the discriminator model is used to identify the original words of the original input text. For more details of the architecture, we invite our readers to refer to the original work of ELECTRA [Ozyurt, 2020].

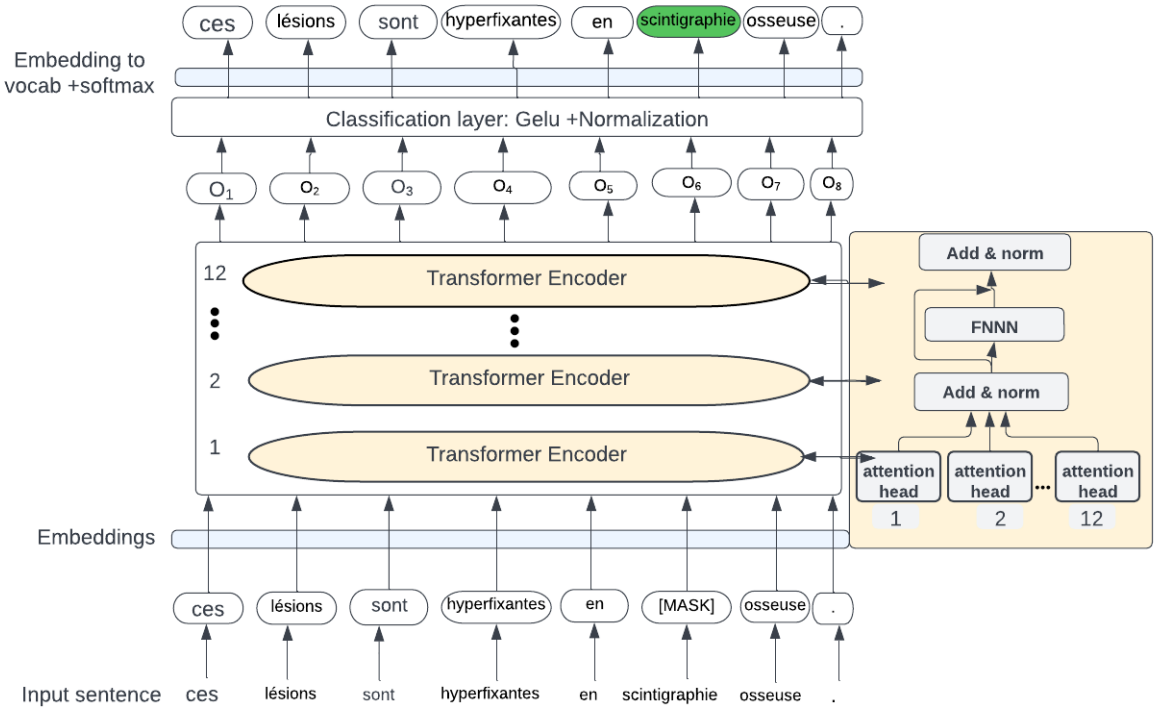

Figure 2: Masked Language Modeling (MLM) strategy. In this example, a sentence related to the French medical domain is given as input where some of the words are hidden. In the output, the hidden words are predicted.

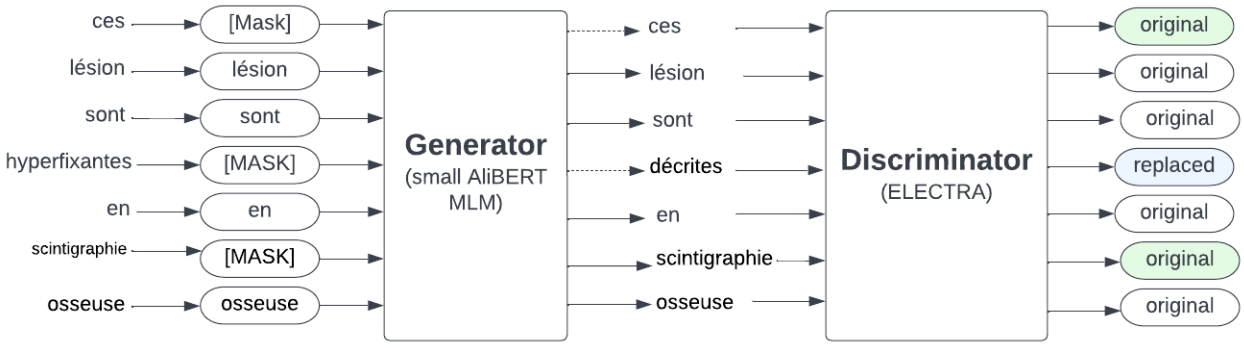

Figure 3: Replaced Token Detection (RTD) strategy. In this example a sentence in French medical language is given as input, some words are hidden and then replaced by the generator. The discriminator then predicts whether the words are those of the original sentence or have been replaced.

Optimization: AliBERT was trained using the ADAM [12] optimizer for faster and better training as used in BERT. Meanwhile, a recent work by You et al.2019 introduced an optimizer known as LAMB that minimizes the training time of BERT to 76 minutes from 3 days (4320 minutes). Therefore, AliBERT was also trained using LAMB optimizer.

The models trained using LAMB optimizer trained much faster than the counter part (using ADAM). However the performance of the models trained with LAMB was not as good as the models trained with ADAM. Figure 4 shows the comparison of time taken to train using LAMB and ADAM atomizers on our models. The loss of the model quickly reduces when LAMB optimizer is used during training.

[12]Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments.

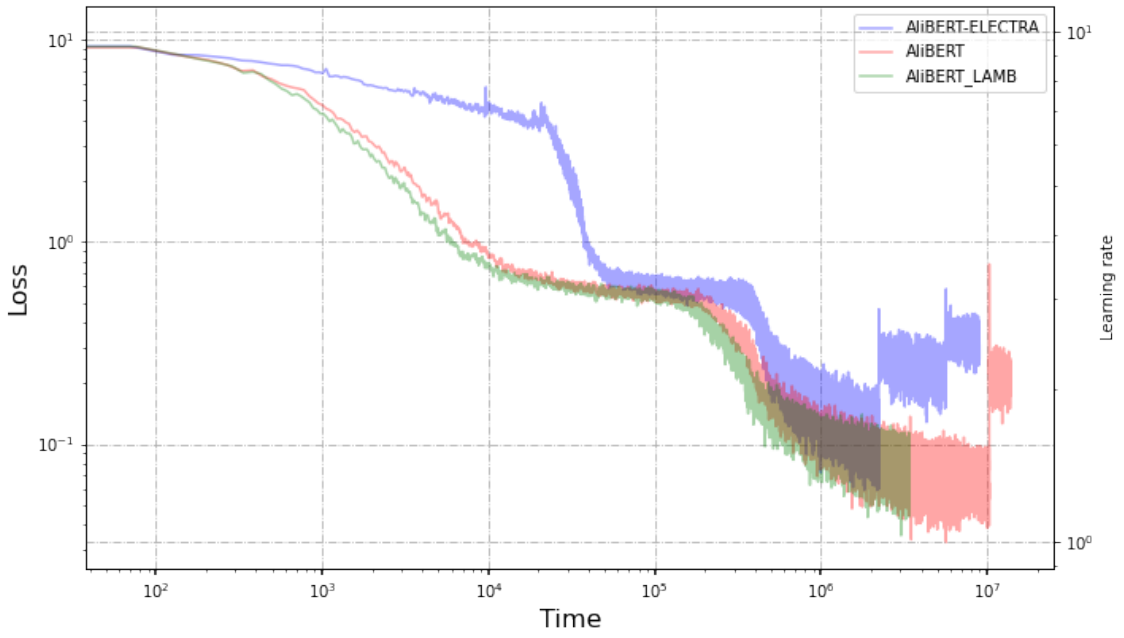

Figure 4: Training time comparison between models using ADAM optimizer and LAMB. The latter allows for faster training but does not lead to better performance.

4 Fine-tuning and model evaluation

In order to evaluate AliBERT's level of understanding French biomedical tasks, we have fine-tuned AliBERT to standard task of evaluating pretrained language models such as biomedical named entity recognition (NER), biomedical text classification, etc. Below, we discussed how the tasks are trained.

4.1 Biomedical Named Entity Recognition (NER)

For the NER task we have used HuggingFace[13] token classification pipeline using our AliBERT models. The dataset used is from the work of Groun et al. 2021 which is used in different challenges of French biomedical text challenge known as, "DEFT (Défis Fouille de Texte)". It is composed of clinical French texts which focuses on specific specialities of medical domain such as cardiology, urology, oncology, obstetrics, pneumatic, etc. The annotation in this dataset include plenty of biomedical entities where some of them have not adequate annotation. Hence, we have kept only 5 types of annotation and these are anatomy, pathology, symptom, substance and value. Table 3 descries the annotated dataset used in NER task for fine-tuning and evaluation purposes.

5 Experiments and results

AliBERT$_{base}$ was trained on 48 GPUs Nvidia A100 ((12 nodes each with 4 GPUs) for 20 hours with 512 input tokens and a batch size of 960 (20 batch size for each GPU). We have used a vocabulary of 40K sub-word units which are built using Unigram tokenization algorithm.

Our models have been evaluated using the above mentioned fine-tuning models and on the masked token prediction. The results found using our models have been compared to the CamemBERT [Martin et al., 2020] French PLM which is the state of the art in French language. Unfortunately, we were not able to compare our models with biomedical PLMs due to lack of French PLM in biomedical domain.

Here are the down stream task that our models has been evaluated on

[13]HuggingFace: the AI community building the future. https://huggingface.co/

Annotation	Occurrences	Description
Substance	2009	Refers to the pharmacological substances used by the patient (drugs, commercial names and generics)
Symptom	5240	Entities that are used to make a diagnosis that reveals the pathology of the patient.
Anatomy	4780	Refers to all anatomical parts (arms, cells, cytoplasm, etc.)
Value	1743	Refers to values and units, grades, etc. corresponding to examination results, or descriptions of Symptoms
Pathology	764	Concerns diseases and all that is pathological (adenocarcinoma, carcinoma, fistula, etc.)

Table 3: NER corpus used for evaluation

Biomedical Named Entity Recognition (NER) A token classification models was fine-tuned from the pretrained models mainly in 5 biomedical entity types, these are symptoms, anatomy, substance, value and pathology. Our models have outperformed CamemBERT in most of the entities and in their macro and micro average of precision (P), recall(R) and F1 score (F1).

	Models								
	CamemBERT			AliBERT			AliBERT-ELECTRA		
Entities	P	R	F1	P	R	F1	P	R	F1
Substance	**0.96**	0.87	0.91	**0.96**	**0.91**	**0.93**	0.95	0.91	0.93
Symptom	0.89	0.91	0.90	**0.96**	**0.98**	**0.97**	0.94	**0.98**	0.96
Anatomy	0.94	0.91	0.88	**0.97**	**0.97**	**0.98**	0.96	**0.97**	0.96
Value	0.88	0.46	0.60	**0.98**	**0.99**	**0.98**	0.93	0.93	0.93
Pathology	0.79	**0.70**	**0.74**	**0.81**	0.39	0.52	0.85	0.57	0.68

Table 4: Biomedical named entity recognition (NER) results

The results found in Table 4 are trained upon a batch size of 80. learning rate (lr) of 2e-5 and weight decay of 0.01 and the dataset used for each of the entities is discussed on Table 3.Table 4 illustrates that AliBERT and AliBERT-ELECTRA outperformed CamemBERT considering the precision of the models to detect the entities. CamemBERT achieved higher F1 score higher than our models for the "Pathology" entity. This is due to the reason that pathology words also exist in the general french language words.However, our models outperformed CamemBERT by huge margin in F1 score for the other entities.

Masking language modeling We have also compared the ability of the models to predict masked tokens, in the same way our proposed models have outperformed CamemBERT. For this experiment of unmasking evaluation a sunset of 3000 text of clean texts (1000 articles of ScienceDirect, 1000 articles from Cochrane, 1000 thesis abstracts from SuDuc) is used. Table 5 illustrates the performance of different models for the prediction of the masked word until the top 5 predictions.

Model	Acc	Top 3 Acc	Top 5 Acc
CamemBERT	0.49	0.57	0.62
AliBERT	**0.72**	**0.83**	**0.87**
AliBERT-ELECTRA	0.71	**0.83**	**0.87**

Table 5: Results predicting the masked tokens (MLM)

AliBERT has outperformed CamemBERT on predicting a masked word prediction. It can be seen in Table 5 AliBERT has an increase of 23% in accuracy when compared with CamemBERT. Hence, it clearly shows that in-domain pretrained language models are really important while dealing with a domain specific texts and hence domain specific down stream tasks.

Figure 5 presents few biomedical text examples for the prediction of masked words. Predicted words colored in green are the correct predictions. Blue colors shows the prediction is correct in the top 2 predictions, purple color depicts that the correct prediction is the top 3 and the red colors show the correct word has not been predicted. As can be seen, Figure 5 AliBERT and AliBERT-ELECTRA outperformed the two French PLMs. This confirms that the need for training domain specific language models, specifically B-PLMs.

Sentence	AliBERT word score	CamemBERT word score	FlauBERT word score	AliBERT-ELECTRA word score
La prise de greffe a été systématiquement réalisée au niveau de la face interne de la [MASK] afin de limiter la plaie cicatricielle.	cuisse 0.913 jambe 0.051 main 0.022 joue 0.004 face 0.002	peau 0.129 jambe 0.117 cuisse 0.094 plaie 0.073 main 0.057	plaie 0.216 lésion 0.067 cellule 0.061 peau 0.053 feuille 0.047	cuisse 0.805 jambe 0.066 main 0.065 joue 0.017 fesse 0.006
Ces lésions sont hyperfixantes en [MASK] osseuse.	scintigraphie 0.987 surface 0.003 pathologie 0.001 phase 0.001 périphérie 0.001	moelle 0.218 densité 0.139 masse 0.088 croissance 0.050 structure 0.034	densité 0.307 masse 0.254 valeur 0.04 quantité 0.039 matière 0.027	scintigraphie 0.791 pathologie 0.074 moelle 0.025 ils 0.018 imagerie 0.01
A l'admission, l'examen clinique mettait en évidence : - une hypotension artérielle avec une pression [MASK] à 6 mmHg.	artérielle 0.434 systolique 0.349 diastolique 0.185 moyenne 0.008 intracrânienne 0.003	inférieure 0.521 supérieure 0.407 supérieur 0.012 inférieur 0.008 artérielle 0.006	supérieure 0.664 inférieure 0.265 égale 0.018 supérieur 0.011 estimée 0.005	artérielle 0.686 diastolique 0.095 systolique 0.093 capillaire 0.050 cardiaque 0.015
En mars 2001, le malade fut opéré, mais vu le caractère hémorragique de la tumeur, une simple biopsie surrénalienne a été réalisée ayant montré l'aspect de [MASK] malin non Hodgkinien de haut grade de malignité.	lymphome 0.992 sarcome 0.001 processus 0.001 lymphomes 0.001 thymome 0.001	cancer 0.402 tumeur 0.189 virus 0.071 maladie 0.067 diabète 0.034	tumeur 0.240 cancer 0.199 tissu 0.161 syndrome 0.057 type 0.034	lymphome 0.940 mélanome 0.007 thymome 0.004 gliome 0.004 lymphomes 0.004
La cytologie urinaire n'a mis en évidence que des cellules [MASK] normales et l'examen cyto-bactériologique des urines était stérile.	épithéliales 0.710 rénales 0.111 souches 0.034 sanguines 0.023 interstitielles 0.017	souches 0.682 musculaires 0.019 rouges 0.017 parfaitement 0.017 humaines 0.015	blanches 0.208 grises 0.103 souches 0.045 jaunes 0.039 noires 0.032	épithéliales 0.208 rénales 0.199 urinaires 0.130 tumorales 0.104 sanguines 0.042

Figure 5: MLM prediction examples and comparison between different Language Model for French Text. For each sentence where a word has been masked, the list of the first five most probable words according to the model are given. The colors show the position of the correct prediction, i.e. green is 1^{st}, blue is 2^{nd}, purple is 3^{rd} and red indicates the correct word is not within the list.

6 Discussion

Our pretrained language models trained on in-domain (biomedical) textual documents tend to outperform models that are trained on general domain textual documents which is also seen on the literature review of pretrained language models for English language such as BioBERT [Lee et al., 2020], PubMedBERT [Gu et al., 2022], etc. Training the PLMs using the masking language model (MLM) objective shows a bit better results, thought the difference is not significant in comparison with replaced token prediction (MLM) objective. Moreover, choosing the right optimizer like LAMB have an effect on the training speed of the pretrained models but not on the performance of the models. During the training of our models different types of tokenizers, such as, Unigram, WordPiece, SentencePiece, BPE, etc. are trained and compared with each other. Unigram tokenizer along with our normalization (see section 3) step tend to outperform other tokenizers. Unigram also was trained into two ways, cased and uncased respectively. Lower casing the input text achieved better results than letting upper cases as it is. Biomedical text tend to have lots of words that are written in capital letters. But, we have seen they are not enough to be used for training our models as upper cases. Biomedical named entity recognition (B-NER) and biomedical text classification (private data, hence results not reported) where used to fine-tune our models to a specific task. Our models tend to generalize faster than the counter part French general PLMs. For AliBERT or AliBERT-ELECTRA less examples of B-NER text inputs where required to start learning and generalize quickly and accurately. Whereas, Camembert took a while to generalize and with less accuracy for biomedical entities which is understandable because it was not trained using in-domain text. In the same manner, this behaviour was reflected during biomedical text classification task. This can also be seen as a comparison to the vocabularies used by CamemBERT and our models. Our tokenizer's (Unigram) vocabulary and CamemBERT tokenizer's (SentencePiece) have huge difference in content and size. The Unigram tokenizers used to train our models have a vocabulary size of 40008 while CamemBERT has 32005. CamemBERT's vocabulary does not include most biomedical words. The two tokenizers have about 10,000 tokens in common in their vocabularies. Although the performance of our models are remarkable, more and various corpora might improve the capability of the models. For example, medical notes, which often come in electronic health records (EHR) data might help to represent the knowledge and experience of medical practitioners.

Furthermore, to improve the models, continual training over general purpose pretrained language, like CamemBERT, could be implemented. Since our tokenizers were a bit different and our goal is to study a purely biomedical PLM, we have not investigated it yet.

7 Conclusion

In this work, French biomedical pretrained language models were proposed which are trained on different corpora of French biomedical textual documents. The proposed models have used different pretraining strategies, AliBERT a BERT [Devlin et al., 2018] based pretrained model used masking language models (MLM) pretraining strategy and AliBERT-ELECTRA (an ELECTRA [Clark et al., 2020] based model) used a replaced token prediction (RTP) learning strategy. Furthermore, a tokenization adaptation strategy was introduced as a building block for pretraining the proposed models. LAMB optimizer has been introduced to AliBERT to speed up training. The proposed pretraining models have been tested on different downstream tasks and achieved state of the art results on multiple tasks. Biomedical entity recognition (NER) and biomedical text classification downstream tasks are fine-tuned using different biomedical textual documents. Hence, AliBERT is expected to be used by different organization and practitioners that work with biomedical text for better understanding and to help make informed decisions concerning biomedical situations. Though our models performed well in all downstream tasks we believe that there is room for improvements. Hence, we plan to work on integrating clinical documents e.g. EHR data, specifically physician notes, to make the model more robust to any kind of biomedical documents. The models can also be enlarged by using continual learning strategy from well known French pretrained language models. CamemBERT [Martin et al., 2020] can be used as a base model and the training can be continued using our biomedical corpus, like BioBERT [Lee et al., 2020] and others did.

We are working on a new version of AliBERT with more data and a greater diversity of corpora. We plan to include text from EHR and medical notes in our corpora. Finally, we also plan to train AliBERT to generate biomedical texts for different purposes.

Acknowledgements

The authors acknowledge a support from the French Institute for Sustainable Science (IRD) within the framework of the France Reliance plan to support research and development (R&D) employment and collaborative research. This work was partly performed using HPC resources from a GENCI-IDRIS grant.

References

K. K. Bressem, L. C. Adams, R. A. Gaudin, D. Tröltzsch, B. Hamm, M. R. Makowski, C.-Y. Schüle, J. L. Vahldiek, and S. M. Niehues. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36:5255–5261, 2020.

J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10, 2020.

O. Cattan, C. Servan, and S. Rosset. On the usability of transformers-based models for a french question-answering task. *arXiv preprint arXiv:2207.09150*, 2022.

J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

B. Chan, S. Schweter, and T. Möller. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, 2020. doi:10.18653/v1/2020.coling-main.598.

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*, pages 1–18, 2020.

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. *arXiv preprint arXiv:1911.02116*, 2019.

J. Copara, J. Knafou, N. Naderi, C. Moro, P. Ruch, and D. Teodoro. Contextualized French Language Models for Biomedical Named Entity Recognition. In *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). Atelier DÉfi Fouille de Textes*, pages 36–48, 2020.

P. Delobelle, T. Winters, and B. Berendt. Robbert: a dutch roberta-based language model. *CoRR*, abs/2001.06286, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171—4186, 2019.

S. Douka, H. Abdine, M. Vazirgiannis, R. E. Hamdani, and D. R. Amariles. Juribert: A masked-language model adaptation for french legal text. *CoRR*, abs/2110.01485, 2021.

E. Dynomant, R. Lelong, B. Dahamna, C. Massonnaud, G. Kerdelhué, J. Grosjean, S. Canu, and S. J. Darmoni. Word embedding for the french natural language in health care: Comparative study. *JMIR Med Inform*, 7(3):e12310, 2019.

C. Groun, N. Grabar, and G. Illouz. Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne deft 2021. In *Actes DEFT 2021*, 2021. URL `https://deft.lisn.upsaclay.fr/2021/`.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, pages 1–23, 2022.

Y. Guo, V. Rennard, C. Xypolopoulos, and M. Vazirgiannis. Bertweetfr: Domain adaptation of pre-trained language models for french tweets. *arXiv preprint arXiv:2109.10234*, 2021.

M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

K. S. Kalyan, A. Rajasekharan, and S. Sangeetha. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, page 103982, 2021.

M. Kamal Eddine, A. Tixier, and M. Vazirgiannis. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*, 2019.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pages 1234–1240, 2020.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, and V. S. andLuke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova. Entitybert: Entity-centric masking strategy for model pretraining for the clinical domain. In *Association for Computational Linguistics*, 2021.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas. Transformers for clinical coding in spanish. *IEEE Access*, 9:72387–72397, 2021.

A. Maronikolakis and H. Schütze. Multidomain pretrained language models for green NLP. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 1–8, 2021.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

U. Naseem, A. G. Dunn, M. Khushi, and J. Kim. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15, 2022.

NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

I. B. Ozyurt. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 104–112, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.sdp-1.12. URL `https://aclanthology.org/2020.sdp-1.12`.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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*, 21(140):1–67, 2020.

K. raj Kanakarajan, B. Kundumani, and M. Sankarasubbu. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, 2021.

E. Terumi Rubel Schneider, J. V. Andrioli de Souza, J. D. M. Knafou, L. E. Silva e Oliveira, J. L. Copara Zea, Y. Bonescki Gumiel, L. Ferro Antunes de Oliveira, E. Cabrera Paraiso, D. Teodoro, and C. M. Cabral Moro Barra. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72. Association for Computational Linguistics, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.

N. Zhang, Q. Jia, K. Yin, L. Dong, F. Gao, and N. Hua. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*, 2020.