FlauBERT: Unsupervised Language Model Pre-training for French

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbe, Laurent Besacier, Didier Schwab

FlauBERT: Unsupervised Language Model Pre-training for French

Hang Le[1] **Loïc Vial**[1] **Jibril Frej**[1] **Vincent Segonne**[2] **Maximin Coavoux**[1]
Benjamin Lecouteux[1] **Alexandre Allauzen**[3] **Benoît Crabbé**[2] **Laurent Besacier**[1] **Didier Schwab**[1]

[1]Univ. Grenoble Alpes, CNRS, LIG [2]Université Paris Diderot [3]E.S.P.C.I, CNRS LAMSADE, PSL Research University
{thi-phuong-hang.le, loic.vial, jibril.frej}@univ-grenoble-alpes.fr
{maximin.coavoux, didier.schwab, benjamin.lecouteux, laurent.besacier}@univ-grenoble-alpes.fr
{vincent.segonne@etu, bcrabbe@linguist}.univ-paris-diderot.fr, alexandre.allauzen@espci.fr

Abstract

Language models have become a key step to achieve state-of-the art results in many different Natural Language Processing (NLP) tasks. Leveraging the huge amount of unlabeled texts nowadays available, they provide an efficient way to pre-train continuous word representations that can be fine-tuned for a downstream task, along with their contextualization at the sentence level. This has been widely demonstrated for English using contextualized representations (Dai and Le, 2015; Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019b). In this paper, we introduce and share FlauBERT, a model learned on a very large and heterogeneous French corpus. Models of different sizes are trained using the new CNRS (French National Centre for Scientific Research) *Jean Zay* supercomputer. We apply our French language models to diverse NLP tasks (text classification, paraphrasing, natural language inference, parsing, word sense disambiguation) and show that most of the time they outperform other pre-training approaches. Different versions of FlauBERT as well as a unified evaluation protocol for the downstream tasks, called FLUE (French Language Understanding Evaluation), are shared to the research community for further reproducible experiments in French NLP.

Keywords: FlauBERT, FLUE, BERT, Transformer, French, language model, pre-training, NLP benchmark, text classification, parsing, word sense disambiguation, natural language inference, paraphrase.

1. Introduction

A recent game-changing contribution in Natural Language Processing (NLP) was the introduction of deep *unsupervised* language representations pre-trained using only plain text corpora. Previous word embedding pre-training approaches, such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), learn a single vector for each wordform. By contrast, these new models are trained to produce *contextual embeddings*: the output representation depends on the entire input sequence (*e.g.* each token instance has a vector representation that depends on its left and right context). Initially based on recurrent neural networks (Dai and Le, 2015; Ramachandran et al., 2017; Howard and Ruder, 2018; Peters et al., 2018), these models quickly converged towards the use of the Transformer (Vaswani et al., 2017), such as GPT (Radford et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019b), RoBERTa (Liu et al., 2019). Using these pre-trained models in a transfer learning fashion has shown to yield striking improvements across a wide range of NLP tasks. One can easily build state-of-the-art NLP systems thanks to the publicly available pre-trained weights, saving time, energy, and resources. As a consequence, unsupervised language model pre-training has become a *de facto* standard in NLP. This has been, however, mostly demonstrated for English even though multi-lingual or cross-lingual variants are also available, taking into account more than a hundred languages in a single model: mBERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), XLM-R (Conneau et al., 2019).

In this paper, we describe our methodology to build FlauBERT – **F**rench **La**nguage **U**nderstanding via **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, a French BERT[1] model that outperforms multi-lingual/cross-lingual models in several downstream NLP tasks, under similar configurations. FlauBERT relies on freely available datasets and is made publicly available in different versions.[2] For further reproducible experiments, we also provide the complete processing and training pipeline as well as a general benchmark for evaluating French NLP systems. This evaluation setup is similar to the popular GLUE benchmark (Wang et al., 2018), and is named FLUE (French Language Understanding Evaluation).

2. Related Work

2.1. Pre-trained Language Models

Self-supervised[3] pre-training on unlabeled text data was first proposed in the task of neural language modeling (Bengio et al., 2003; Collobert and Weston, 2008), where it was shown that a neural network trained to predict next word from prior words can learn useful embedding representations, called *word embeddings* (each word is represented by a fixed vector). These representations were shown to play an important role in NLP, yielding state-of-the-art performance on multiple tasks (Collobert et al., 2011), especially

[1]We learned of a similar project that resulted in a publication on arXiv (Martin et al., 2019). However, we believe that these two works on French language models are complementary since the NLP tasks we addressed are different, as are the training corpora and preprocessing pipelines. We also point out that our models were trained using the CNRS (French National Centre for Scientific Research) public research computational infrastructure and did not receive any assistance from a private stakeholder.

[2]https://github.com/getalp/Flaubert

[3]*Self-supervised learning* is a special case of *unsupervised learning* where unlabeled data is used as a supervision signal.

after the introduction of word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), efficient and effective algorithms for learning word embeddings.

A major limitation of word embeddings is that a word can only have a single representation, even if it can have multiple meanings (*e.g.* depending on the context). Therefore, recent works have introduced a paradigm shift from context-free word embeddings to *contextual embeddings*: the output representation is a function of the entire input sequence, which allows encoding complex, high-level syntactic and semantic characteristics of words or sentences.

This line of research was started by Dai and Le (2015) who proposed pre-training representations via either an encoder-decoder language model or a sequence autoencoder. Ramachandran et al. (2017)[4] showed that this approach can be applied to pre-training sequence-to-sequence models (Sutskever et al., 2014). These models, however, require a significant amount of in-domain data for the pre-training tasks. Peters et al. (2018, ELMo) and Howard and Ruder (2018, ULMFiT) were the first to demonstrate that leveraging huge general-domain text corpora in pre-training can lead to substantial improvements on downstream tasks. Both methods employ LSTM (Hochreiter and Schmidhuber, 1997) language models, but ULMFiT utilizes a regular multi-layer architecture, while ELMo adopts a *bidirectional* LSTM to build the final embedding for each input token from the concatenation of the left-to-right and right-to-left representations. Another fundamental difference lies in how each model can be tuned to different downstream tasks: ELMo delivers different word vectors that can be interpolated, whereas ULMFiT enables robust fine-tuning of the whole network *w.r.t.* the downstream tasks. The ability of fine-tuning was shown to significantly boost the performance, and thus this approach has been further developed in the recent works such as MultiFiT (Eisenschlos et al., 2019) or most prominently Transformer-based (Vaswani et al., 2017) architectures: GPT (Radford et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019b), XLM (Lample and Conneau, 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), T5 (Raffel et al., 2019). These methods have one after the other established new state-of-the-art results on various NLP benchmarks, such as GLUE (Wang et al., 2018) or SQuAD (Rajpurkar et al., 2018), surpassing previous methods by a large margin.

2.2. Pre-trained Language Models Beyond English

Given the impact of pre-trained language models on NLP downstream tasks in English, several works have recently released pre-trained models for other languages. For instance, ELMo exists for Portuguese, Japanese, German and Basque,[5] while BERT and variants were specifically trained for simplified and traditional Chinese[8] and German.[6] A Portuguese version of MultiFiT is also available.[7] Recently, more monolingual BERT-based models have been released, such as for Arabic (Antoun et al., 2020), Dutch (de Vries et al., 2019; Delobelle et al., 2020), Finnish (Virtanen et al., 2019), Italian (Polignano et al., 2019), Portuguese (Souza et al., 2019), Russian (Kuratov and Arkhipov, 2019), Spanish (Cañete et al., 2020), and Vietnamese (Nguyen and Nguyen, 2020). For French, besides pre-trained language models using ULMFiT and MultiFiT configurations,[7] CamemBERT (Martin et al., 2019) is a French BERT model concurrent to our work.

Another trend considers one model estimated for several languages with a shared vocabulary. The release of multilingual BERT for 104 languages pioneered this approach.[8] A recent extension of this work leverages parallel data to build a cross-lingual pre-trained version of LASER (Artetxe and Schwenk, 2019) for 93 languages, XLM (Lample and Conneau, 2019) and XLM-R (Conneau et al., 2019) for 100 languages.

2.3. Evaluation Protocol for French NLP Tasks

The existence of a multi-task evaluation benchmark such as GLUE (Wang et al., 2018) for English is highly beneficial to facilitate research in the language of interest. The GLUE benchmark has become a prominent framework to evaluate the performance of NLP models in English. The recent contributions based on pre-trained language models have led to remarkable performance across a wide range of Natural Language Understanding (NLU) tasks. The authors of GLUE have therefore introduced SuperGLUE (Wang et al., 2019a): a new benchmark built on the principles of GLUE, including more challenging and diverse set of tasks. A Chinese version of GLUE[9] is also developed to evaluate model performance in Chinese NLP tasks. As of now, we have not learned of any such benchmark for French.

3. Building FlauBERT

In this section, we describe the training corpus, the text preprocessing pipeline, the model architecture and training configurations to build FlauBERT$_{\text{BASE}}$ and FlauBERT$_{\text{LARGE}}$.

3.1. Training Data

Data collection Our French text corpus consists of 24 sub-corpora gathered from different sources, covering diverse topics and writing styles, ranging from formal and well-written text (*e.g.* Wikipedia and books)[10] to random text crawled from the Internet (*e.g.* Common Crawl).[11] The data were collected from three main sources: (1) monolingual data for French provided in WMT19 shared tasks (Li et al., 2019, 4 sub-corpora); (2) French text corpora offered in the OPUS collection (Tiedemann, 2012, 8 sub-corpora); and (3) datasets available in the Wikimedia projects (Meta, 2019, 8 sub-corpora).

We used the WikiExtractor tool[12] to extract the text from Wikipedia. For the other sub-corpora, we either used our

[4] It should be noted that learning contextual embeddings was also proposed in (McCann et al., 2017), but in a *supervised* fashion as they used annotated machine translation data.

[5] https://allennlp.org/elmo

[6] https://deepset.ai/german-bert

[7] https://github.com/piegu/language-models

[8] https://github.com/google-research/bert

[9] https://github.com/chineseGLUE/chineseGLUE

[10] http://www.gutenberg.org

[11] http://data.statmt.org/ngrams/deduped2017

[12] https://github.com/attardi/wikiextractor

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	$\text{FlauBERT}_{\text{BASE}}$/$\text{FlauBERT}_{\text{LARGE}}$
Language	English	English	French	French
Training data	13 GB	160 GB	138 GB[†]	71 GB[‡]
Pre-training objectives	NSP and MLM	MLM	MLM	MLM
Total parameters	110 M	125 M	110 M	138 M/ 373 M
Tokenizer	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masking strategy	Static + Sub-word masking	Dynamic + Sub-word masking	Dynamic + Whole-word masking	Dynamic + Sub-word masking

[†], [‡]: 282 GB, 270 GB before filtering/cleaning.

Table 1: Comparison between FlauBERT and previous work.

own tool to extract the text or download them directly from their websites. The total size of the uncompressed text before preprocessing is 270 GB. More details can be found in Appendix A.1.

Data preprocessing For all sub-corpora, we filtered out very short sentences as well as repetitive and non-meaningful content such as telephone/fax numbers, email addresses, *etc*. For Common Crawl, which is our largest sub-corpus with 215 GB of raw text, we applied aggressive cleaning to reduce its size to 43.4 GB. All the data were Unicode-normalized in a consistent way before being tokenized using Moses tokenizer (Koehn et al., 2007). The resulting training corpus is 71 GB in size.
Our code for downloading and preprocessing data is made publicly available.[13]

3.2. Models and Training Configurations

Model architecture FlauBERT has the same model architecture as BERT (Devlin et al., 2019), which consists of a multi-layer bidirectional Transformer (Vaswani et al., 2017). Following Devlin et al. (2019), we propose two model sizes:

- $\text{FlauBERT}_{\text{BASE}}$: $L = 12, H = 768, A = 12$,
- $\text{FlauBERT}_{\text{LARGE}}$: $L = 24, H = 1024, A = 16$,

where L, H and A respectively denote the number of Transformer blocks, the hidden size, and the number of self-attention heads. As Transformer has become quite standard, we refer to Vaswani et al. (2017) for further details.

Training objective and optimization Pre-training of the original BERT (Devlin et al., 2019) consists of two supervised tasks: (1) a *masked language model* (MLM) that learns to predict randomly masked tokens; and (2) a *next sentence prediction* (NSP) task in which the model learns to predict whether `B` is the actual next sentence that follows `A`, given a pair of input sentences `A`,`B`.
Devlin et al. (2019) observed that removing NSP significantly hurts performance on some downstream tasks. However, the opposite was shown in later studies, including Yang et al. (2019b, XLNet), Lample and Conneau (2019, XLM), and Liu et al. (2019, RoBERTa).[14] Therefore, we only employed the MLM objective in FlauBERT.
To optimize this objective function, we followed Liu et al. (2019) and used the Adam optimizer (Kingma and Ba, 2014) with the following parameters:

- $\text{FlauBERT}_{\text{BASE}}$: warmup steps of 24k, peak learning rate of 6e−4, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\text{e}{-6}$ and weight decay of 0.01.
- $\text{FlauBERT}_{\text{LARGE}}$: warmup steps of 30k, peak learning rate of 3e−4, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\text{e}{-6}$ and weight decay of 0.01.

Training $\text{FlauBERT}_{\text{LARGE}}$ Training very deep Transformers is known to be susceptible to instability (Wang et al., 2019b; Nguyen and Salazar, 2019; Xu et al., 2019; Fan et al., 2019). Not surprisingly, we also observed this difficulty when training $\text{FlauBERT}_{\text{LARGE}}$ using the same configurations as $\text{BERT}_{\text{LARGE}}$ and $\text{RoBERTa}_{\text{LARGE}}$, where divergence happened at an early stage.
Several methods have been proposed to tackle this issue. For example, in an updated implementation of the Transformer (Vaswani et al., 2018), layer normalization is applied *before* each attention layer by default, rather than after each residual block as in the original implementation (Vaswani et al., 2017). These configurations are called *pre-norm* and *post-norm*, respectively. It was observed by Vaswani et al. (2018), and again confirmed by later works *e.g*. (Wang et al., 2019b; Xu et al., 2019; Nguyen and Salazar, 2019), that pre-norm helps stabilize training. Recently, a regularization technique called *stochastic depths* (Huang et al., 2016) has been demonstrated to be very effective for training deep Transformers, by *e.g*. Pham et al. (2019) and Fan et al. (2019) who successfully trained architectures of more than 40 layers. The idea is to randomly drop a number of (attention) layers at each training step. Other techniques are also available such as progressive training (Gong et al., 2019), or improving initialization (Zhang et al., 2019a; Xu et al., 2019) and normalization (Nguyen and Salazar, 2019).
For training $\text{FlauBERT}_{\text{LARGE}}$, we employed pre-norm attention and stochastic depths for their simplicity. We found that these two techniques were sufficient for successful training. We set the rate of layer dropping to 0.2 in all the experiments.

Other training details A vocabulary of 50K sub-word units is built using the Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2016). The only difference between our work and RoBERTa is that the training data are preprocessed and tokenized using a basic tokenizer for French (Koehn et al., 2007, Moses), as in XLM (Lample and Conneau, 2019), before the application of BPE. We use *fastBPE*,[15] a very efficient implementation to extract the BPE units and encode the corpora.

[13] `https://github.com/getalp/Flaubert`

[14] Liu et al. (2019) hypothesized that the original BERT implementation may only have removed the loss term while still retaining a bad input format, resulting in performance degradation.

[15] `https://github.com/glample/fastBPE`

FlauBERT$_{\text{BASE}}$ is trained on 32 GPUs Nvidia V100 in 410 hours and FlauBERT$_{\text{LARGE}}$ is trained on 128 GPUs in 390 hours, both with the effective batch size of 8192 sequences. Finally, we summarize the differences between FlauBERT and BERT, RoBERTa, CamemBERT in Table 1.

4. FLUE

In this section, we compile a set of existing French language tasks to form an evaluation benchmark for French NLP that we called FLUE (French Language Understanding Evaluation). We select the datasets from different domains, level of difficulty, degree of formality, and amount of training samples. Three out of six tasks (Text Classification, Paraphrase, Natural Language Inference) are from cross-lingual datasets since we also aim to provide results from a monolingual pre-trained model to facilitate future studies of cross-lingual models, which have been drawing much of research interest recently.
Table 2 gives an overview of the datasets, including their domains and training/development/test splits. The details are presented in the next subsections.

Dataset		Domain	Train	Dev	Test
CLS-FR	Books	Product reviews	2 000	-	2 000
	DVD		1 999	-	2 000
	Music		1 998	-	2 000
PAWS-X-FR		General domain	49 401	1 992	1 985
XNLI-FR		Diverse genres	392 702	2 490	5 010
French Treebank		Daily newspaper	14 759	1 235	2 541
FrenchSemEval		Diverse genres	55 206	-	3 199
Noun Sense Disambiguation		Diverse genres	818 262	-	1 445

Table 2: Descriptions of the datasets included in our FLUE benchmark.

4.1. Text Classification

CLS The Cross Lingual Sentiment CLS (Prettenhofer and Stein, 2010) dataset consists of Amazon reviews for three product categories: books, DVD, and music in four languages: English, French, German, and Japanese. Each sample contains a review text and the associated rating from 1 to 5 stars. Following Blitzer et al. (2006) and Prettenhofer and Stein (2010), ratings with 3 stars are removed. Positive reviews have ratings higher than 3 and negative reviews are those rated lower than 3. There is one train and test set for each product category. The train and test sets are balanced, including around 1 000 positive and 1 000 negative reviews for a total of 2 000 reviews in each dataset. We take the French portion to create the binary text classification task in FLUE and report the accuracy on the test set.

4.2. Paraphrasing

PAWS-X The Cross-lingual Adversarial Dataset for Paraphrase Identification PAWS-X (Yang et al., 2019a) is the extension of the Paraphrase Adversaries from Word Scrambling PAWS (Zhang et al., 2019b) for English to six other languages: French, Spanish, German, Chinese, Japanese and Korean. PAWS composes English paraphrase identification pairs from Wikipedia and Quora in which two sentences in a pair have high lexical overlap ratio, generated by LM-based word scrambling and back translation followed by human judgement. The paraphrasing task is to identify whether the sentences in these pairs are semantically equivalent or not. Similar to previous approaches to create multilingual corpora, Yang et al. (2019a) used machine translation to create the training set for each target language in PAWS-X from the English training set in PAWS. The development and test sets for each language are translated by human translators. We take the related datasets for French to perform the paraphrasing task and report the accuracy on the test set.

4.3. Natural Language Inference

XNLI The Cross-lingual NLI (XNLI) corpus (Conneau et al., 2018) extends the development and test sets of the Multi-Genre Natural Language Inference corpus (Williams et al., 2018, MultiNLI) to 15 languages. The development and test sets for each language consist of 7 500 human-annotated examples, making up a total of 112 500 sentence pairs annotated with the labels *entailment*, *contradiction*, or *neutral*. Each sentence pair includes a premise (p) and a hypothesis (h). The Natural Language Inference (NLI) task, also known as recognizing textual entailment (RTE), is to determine whether p entails, contradicts or neither entails nor contradicts h. We take the French part of the XNLI corpus to form the development and test sets for the NLI task in FLUE. The train set is obtained from the machine translated version to French provided in XNLI. Following Conneau et al. (2018), we report the test accuracy.

4.4. Parsing and Part-of-Speech Tagging

Syntactic parsing consists in assigning a tree structure to a sentence in natural language. We perform parsing on the French Treebank (Abeillé et al., 2003), a collection of sentences extracted from French daily newspaper Le Monde, and manually annotated with both constituency and dependency syntactic trees and part-of-speech tags. Specifically, we use the version of the corpus instantiated for the SPMRL 2013 shared task and described by Seddah et al. (2013). This version is provided with a standard split representing 14 759 sentences for the training corpus, and respectively 1 235 and 2 541 sentences for the development and evaluation sets.

4.5. Word Sense Disambiguation Tasks

Word Sense Disambiguation (WSD) is a classification task which aims to predict the sense of words in a given context according to a specific sense inventory. We used two French WSD tasks: the FrenchSemEval task (Segonne et al., 2019), which targets verbs only, and a modified version of the French part of the Multilingual WSD task of SemEval 2013 (Navigli et al., 2013), which targets nouns.

Verb Sense Disambiguation We made experiments of sense disambiguation focused on French verbs using FrenchSemEval (Segonne et al., 2019, FSE), an evaluation dataset in which verb occurrences were manually sense annotated with the sense inventory of Wiktionary, a collaboratively edited open-source dictionary. FSE includes both the evaluation data and the sense inventory. The evaluation data consists of 3 199 manual annotations among a selection of

66 verbs which makes roughly 50 sense annotated occurrences per verb. The sense inventory provided in FSE is a Wiktionary dump (04-20-2018) openly available via Dbnary (Sérasset, 2012). For a given sense of a target key, the sense inventory offers a definition along with one or more examples. For this task, we considered the examples of the sense inventory as training examples and tested our model on the evaluation dataset.

Noun Sense Disambiguation We propose a new challenging task for the WSD of French, based on the French part of the Multilingual WSD task of SemEval 2013 (Navigli et al., 2013), which targets nouns only. We adapted the task to use the WordNet 3.0 sense inventory (Miller, 1995) instead of BabelNet (Navigli and Ponzetto, 2010), by converting the sense keys to WordNet 3.0 if a mapping exists in BabelNet, and removing them otherwise.
The result of the conversion process is an evaluation corpus composed of 306 sentences and 1 445 French nouns annotated with WordNet sense keys, and manually verified.
For the training data, we followed the method proposed by Hadj Salah (2018), and translated the SemCor (Miller et al., 1993) and the WordNet Gloss Corpus[16] into French, using the best English-French Machine Translation system of the *fairseq* toolkit[17] (Ott et al., 2019). Finally, we aligned the WordNet sense annotation from the source English words to the the translated French words, using the alignment provided by the MT system.
We rely on WordNet sense keys instead of the original BabelNet annotations for the following two reasons. First, WordNet is a resource that is entirely manually verified, and widely used in WSD research (Navigli, 2009). Second, there is already a large quantity of sense annotated data based on the sense inventory of WordNet (Vial et al., 2018) that we can use for the training of our system.
We publicly release[18] both our training data and the evaluation data in the UFSAC format (Vial et al., 2018).

5. Experiments and Results

In this section, we present FlauBERT fine-tuning results on the FLUE benchmark. We compare the performance of FlauBERT with Multilingual BERT (Devlin et al., 2019, mBERT) and CamemBERT (Martin et al., 2019) on all tasks. In addition, for each task we also include the best non-BERT model for comparison. We made use of the open source libraries (Lample and Conneau, 2019, XLM) and (Wolf et al., 2019, Transformers) in some of the experiments.

5.1. Text Classification

Model description We followed the standard fine-tuning process of BERT (Devlin et al., 2019). The input is a degenerate text-$\varnothing$ pair. The classification head is composed of the following layers, in order: dropout, linear, $\tanh$ activation, dropout, and linear. The output dimensions of the linear layers are respectively equal to the hidden size of the Transformer and the number of classes (which is 2 in this case as the task is binary classification). The dropout rate was set to 0.1.
We trained for 30 epochs using a batch size of 16 while performing a grid search over 4 different learning rates: $1e{-}5$, $5e{-}5$, $1e{-}6$, and $5e{-}6$. A random split of 20% of the training data was used as validation set, and the best performing model on this set was then chosen for evaluation on the test set.

Model	Books	DVD	Music
MultiFiT†	91.25	89.55	93.40
mBERT†	86.15	86.90	86.65
CamemBERT	92.30	93.00	94.85
$\text{FlauBERT}_{\text{BASE}}$	93.10	92.45	94.10
$\text{FlauBERT}_{\text{LARGE}}$	**95.00**	**94.10**	**95.85**

† Results reported in (Eisenschlos et al., 2019).

Table 3: Accuracy on the CLS dataset for French.

Results Table 3 presents the final accuracy on the test set for each model. The results highlight the importance of a monolingual French model for text classification: both CamemBERT and FlauBERT outperform mBERT by a large margin. $\text{FlauBERT}_{\text{BASE}}$ performs moderately better than CamemBERT in the books dataset, while its results on the two remaining datasets of DVD and music are lower than those of CamemBERT. $\text{FlauBERT}_{\text{LARGE}}$ achieves the best results in all categories.

5.2. Paraphrasing

Model description The setup for this task is almost identical to the previous one, except that: (1) the input sequence is now a pair of sentences `A,B`; and (2) the hyper-parameter search is performed on the development data set (*i.e.* no validation split is needed).

Results The final accuracy for each model is reported in Table 4. One can observe that the monolingual French models perform only slightly better than the multilingual model mBERT, which could be attributed to the characteristics of the PAWS-X dataset. Containing samples with high lexical overlap ratio, this dataset has been proved to be an effective measure of model sensitivity to word order and syntactic structure (Yang et al., 2019a). A multilingual model such as mBERT, therefore, could capture these features as well as a monolingual model.

Model	Accuracy
ESIM† (Chen et al., 2017)	66.20
mBERT†	89.30
CamemBERT	**90.14**
$\text{FlauBERT}_{\text{BASE}}$	89.49
$\text{FlauBERT}_{\text{LARGE}}$	89.34

† Results reported in (Yang et al., 2019a).

Table 4: Results on the French PAWS-X dataset.

[16] The set of WordNet glosses semi-automatically sense annotated which is released as part of WordNet since version 3.0.

[17] `https://github.com/pytorch/fairseq`

[18] `https://zenodo.org/record/3549806`

5.3. Natural Language Inference

Model description As this task was also considered in (Martin et al., 2019, CamemBERT), for a fair comparison, here we replicate the same experimental setup. Similar to paraphrasing, the model input of this task is also a pair of sentences. The classification head, however, consists of only one dropout layer followed by one linear layer.

Results We report the final accuracy for each model in Table 5. The results confirm the superiority of the French models compared to the multilingual model mBERT on this task. FlauBERT$_{\text{LARGE}}$ performs moderately better than CamemBERT. Both of them clearly outperform XLM-R$_{\text{BASE}}$, while cannot surpass XLM-R$_{\text{LARGE}}$.

Model	Accuracy
XLM-R$_{\text{LARGE}}$ †	**85.2**
XLM-R$_{\text{BASE}}$ †	80.1
mBERT‡	76.9
CamemBERT ‡	81.2
FlauBERT$_{\text{BASE}}$	80.6
FlauBERT$_{\text{LARGE}}$	83.4

† Results reported in (Conneau et al., 2019).
‡ Results reported in (Martin et al., 2019).

Table 5: Results on the French XNLI dataset.

5.4. Constituency Parsing and POS Tagging

Model description We use the parser described by Kitaev and Klein (2018) and Kitaev et al. (2019). It is an openly available[19] chart parser based on a self-attentive encoder. We compare (i) a model without any pre-trained parameters, (ii) a model that additionally uses and fine-tunes fastText[20] pre-trained embeddings, (iii) models based on pre-trained language models: mBERT, CamemBERT, and FlauBERT. We use the default hyperparameters from Kitaev and Klein (2018) for the first two settings and the hyperparameters from Kitaev et al. (2019) when using pre-trained language models, except for FlauBERT$_{\text{LARGE}}$. For this last model, we use a different learning rate (0.00001), batch size (8) and ignore training sentences longer than 100 tokens, due to memory limitation. We jointly perform part-of-speech (POS) tagging based on the same input as the parser, in a multitask setting. For each setting we perform training 3 times with different random seeds and select best model according to development F-score.

For final evaluation, we use the evaluation tool provided by the SPMRL shared task organizers[21] and report labelled F-score, the standard metric for constituency parsing evaluation, as well as POS tagging accuracy.

Results We report constituency parsing results in Table 6. Without pre-training, we replicate the result from Kitaev and Klein (2018). FastText pre-trained embeddings do not bring improvement over this already strong model.

[19] https://github.com/nikitakit/self-attentive-parser

[20] https://fasttext.cc/

[21] http://pauillac.inria.fr/~seddah/evalb_spmrl2013.tar.gz

Model	Dev		Test	
	F_1	POS	F_1	POS
Best published (Kitaev et al., 2019)			87.42	
No pre-training	84.31	97.6	83.85	97.5
fastText pre-trained embeddings	84.09	97.6	83.64	97.7
mBERT	87.25	98.1	87.52	98.1
CamemBERT (Martin et al., 2019)	88.53	98.1	88.39	**98.2**
FlauBERT$_{\text{BASE}}$	88.95	**98.2**	**89.05**	98.1
FlauBERT$_{\text{LARGE}}$	**89.08**	**98.2**	88.63	**98.2**
Ensemble: FlauBERT$_{\text{BASE}}$ + CamemBERT	**89.32**		**89.28**	

Table 6: Constituency parsing and POS tagging results.

When using pre-trained language models, we observe that CamemBERT, with its language-specific training improves over mBERT by 0.9 absolute F_1. FlauBERT$_{\text{BASE}}$ outperforms CamemBERT by 0.7 absolute F_1 on the test set and obtains the best published results on the task for a single model. Regarding POS tagging, all large-scale pre-trained language models obtain similar results (98.1-98.2), and outperform models without pre-training or with fastText embeddings (97.5-97.7). FlauBERT$_{\text{LARGE}}$ provides a marginal improvement on the development set, and fails to reach FlauBERT$_{\text{BASE}}$ results on the test set.

In order to assess whether FlauBERT and CamemBERT are complementary for this task, we evaluate an ensemble of both models (last line in Table 6). The ensemble model improves by 0.4 absolute F_1 over FlauBERT on the development set and 0.2 on the test set, obtaining the highest result for the task. This result suggests that both pre-trained language models are complementary and have their own strengths and weaknesses.

5.5. Dependency parsing

Model We use our own reimplementation of the parsing model of Dozat and Manning (2016) with maximum spanning tree decoding adapted to handle several input sources such as BERT representations. The model does not perform part of speech tagging but uses the predicted tags provided by the SPMRL shared task organizers.

Our word representations are a concatenation of word embeddings and tag embeddings learned together with the model parameters on the French Treebank data itself, and at most one of (fastText, CamemBERT, FlauBERT$_{\text{BASE}}$, FlauBERT$_{\text{BASE}}$, mBERT) word vector. As Dozat and Manning (2016), we use word and tag dropout ($d = 0.5$) on word and tag embeddings but without dropout on BERT representations. We performed a fairly comprehensive grid search on hyperparameters for each model tested.

Results The results are reported in Table 7. The best published results in this shared task (Constant et al., 2013) were involving an ensemble of parsers with additional resources for modelling multi word expressions (MWE), typical of the French treebank annotations. The monolingual French BERT models (CamemBERT, FlauBERT) perform better and set the new state of the art on this dataset with a single parser and without specific modelling for MWEs. One can observe that both FlauBERT models perform marginally better than CamemBERT, while all of them outperform mBERT by a large margin.

Model	UAS	LAS
Best published (Constant et al., 2013)	89.19	85.86
No pre-training	88.92	85.11
fastText pre-training	86.32	82.04
mBERT	89.50	85.86
CamemBERT	91.37	88.13
FlauBERT$_{\text{BASE}}$	91.56	88.35
FlauBERT$_{\text{LARGE}}$	**91.61**	**88.47**

Table 7: Dependency parsing results.

5.6. Word Sense Disambiguation

Verb Sense Disambiguation Disambiguation was performed with the same WSD supervised method used by Segonne et al. (2019). First we compute sense vector representations from examples found in the Wiktionary sense inventory: given a sense s and its corresponding examples, we compute the vector representation of s by averaging the vector representations of its examples. Then, we tag each test instance with the sense whose representation is the closest based on cosine similarity. We used the contextual embeddings output by FlauBERT as vector representations for any given instance (from the sense inventory or the test data) of a target word. We proceeded the same way with mBERT and CamemBERT for comparison. We also compared our model with a simpler context vector representation called averaged word embeddings (AWE) which consists in representing context of target word by averaging its surrounding words in a given window size. We experimented AWE using fastText word embeddings with a window of size 5. We report results in Table 8. BERT-based models set the new state of the art on this task, with the best results achieved by CamemBERT and FlauBERT$_{\text{LARGE}}$.

Model	F_1
fastText	34.90
mBERT	49.83
CamemBERT	50.02
FlauBERT$_{\text{BASE}}$	43.92
FlauBERT$_{\text{LARGE}}$	**50.48**

Table 8: F_1 scores (%) on the Verb Disambiguation Task.

Noun Sense Disambiguation We implemented a neural classifier similar to the classifier presented by Vial et al. (2019). This classifier forwards the output of a pre-trained language model to a stack of 6 trained Transformer encoder layers and predicts the synset of every input words through softmax. The only difference between our model and Vial et al. (2019) is that we chose the same hyperparameter as FlauBERT$_{\text{BASE}}$ for the d_{ff} and the number of attention heads of the Transformer layers (more precisely, $d_{ff} = 3072$ and $A = 12$).

At prediction time, we take the synset ID which has the maximum value along the softmax layer (no filter on the lemma of the target is performed). We trained 8 models for every experiment, and we report the mean results, and the standard deviation of the individual models, and also the result of an ensemble of models, which averages the output of the softmax layer. Finally, we compared FlauBERT with CamemBERT, mBERT, fastText and with no input embeddings. We report the results in Table 9. On this task and with these settings, we first observe an advantage for mBERT over both CamemBERT and FlauBERT$_{\text{BASE}}$. We think that it might be due to the fact that the training corpora we used are machine translated from English to French, so the multilingual nature of mBERT makes it probably more fitted for the task. Comparing CamemBERT to FlauBERT$_{\text{BASE}}$, we see a small improvement in the former model, and we think that this might be due to the difference in the sizes of pre-training corpora. Finally, with our FlauBERT$_{\text{LARGE}}$ model, we obtain the best scores on the task, achieving more than 1 point above mBERT.

Model	Single		Ensemble
	Mean	Std	
No pre-training	45.73	±1.91	50.03
fastText	44.90	±1.24	49.41
mBERT	53.03	±1.22	56.47
CamemBERT	52.06	±1.25	56.06
FlauBERT$_{\text{BASE}}$	51.24	±1.33	54.74
FlauBERT$_{\text{LARGE}}$	53.53	±1.36	**57.85**

Table 9: F_1 scores (%) on the Noun Disambiguation Task.

6. Conclusion

We present and release FlauBERT, a pre-trained language model for French. FlauBERT was trained on a multiple-source corpus and achieved state-of-the-art results on a number of French NLP tasks, surpassing multilingual/cross-lingual models. FlauBERT is competitive with CamemBERT (Martin et al., 2019) – another pre-trained language model for French – despite being trained on almost twice as fewer text data. In order to make the pipeline entirely reproducible, we not only release preprocessing and training scripts, together with FlauBERT, but also provide a general benchmark for evaluating French NLP systems (FLUE). FlauBERT is also now supported by Hugging Face's `transformers` library.[22]

7. Acknowledgements

This work benefited from the *'Grand Challenge Jean Zay'* program and was also partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

We thank Guillaume Lample and Alexis Conneau for their active technical support on using the XLM code.

A Appendix

A.1 Details on our French text corpus

Table 10 presents the statistics of all sub-corpora in our training corpus. We give the description of each sub-corpus below.

[22] `https://huggingface.co/transformers/`

Dataset	Post-processed text size	Number of Tokens (Moses)	Number of Sentences
CommonCrawl (Buck et al., 2014)	43.4 GB	7.85 B	293.37 M
NewsCrawl (Li et al., 2019)	9.2 GB	1.69 B	63.05 M
Wikipedia (Meta, 2019)	4.2 GB	750.76 M	31.00 M
Wikisource (Meta, 2019)	2.4 GB	458.85 M	27.05 M
EU Bookshop (Skadins et al., 2014)	2.3 GB	389.40 M	13.18 M
MultiUN (Eisele and Chen, 2010)	2.3 GB	384.42 M	10.66 M
GIGA (Tiedemann, 2012)	2.0 GB	353.33 M	10.65 M
PCT	1.2 GB	197.48 M	7.13 M
Project Gutenberg	1.1 GB	219.73 M	8.23 M
OpenSubtitles (Lison and Tiedemann, 2016)	1.1 GB	218.85 M	13.98 M
Le Monde	664 MB	122.97 M	4.79 M
DGT (Tiedemann, 2012)	311 MB	53.31 M	1.73 M
EuroParl (Koehn, 2005)	292 MB	50.44 M	1.64 M
EnronSent (Styler, 2011)	73 MB	13.72 M	662.31 K
NewsCommentary (Li et al., 2019)	61 MB	13.40 M	341.29 K
Wiktionary (Meta, 2019)	52 MB	9.68 M	474.08 K
Global Voices (Tiedemann, 2012)	44 MB	7.88 M	297.38 K
Wikinews (Meta, 2019)	21 MB	3.93 M	174.88 K
TED Talks (Tiedemann, 2012)	15 MB	2.92 M	129.31 K
Wikiversity (Meta, 2019)	10 MB	1.70 M	64.60 K
Wikibooks (Meta, 2019)	9 MB	1.67 M	65.19 K
Wikiquote (Meta, 2019)	5 MB	866.22 K	42.27 K
Wikivoyage (Meta, 2019)	3 MB	500.64 K	23.36 K
EUconst (Tiedemann, 2012)	889 KB	148.47 K	4.70 K
Total	**71 GB**	**12.79 B**	**488.78 M**

Table 10: Statistics of sub-corpora after cleaning and pre-processing an initial corpus of 270 GB, ranked in the decreasing order of post-processed text size.

Datasets from WMT19 shared tasks We used four corpora provided in the WMT19 shared task (Li et al., 2019).[23]

- *Common Crawl* includes text crawled from billions of pages in the internet.
- *News Crawl* contains crawled news collected from 2007 to 2018.
- *EuroParl* composes text extracted from the proceedings of the European Parliament.
- *News Commentary* consists of text from news-commentary crawl.

Datasets from OPUS OPUS[24] is a growing resource of freely accessible monolingual and parallel corpora (Tiedemann, 2012). We collected the following French monolingual datasets from OPUS.

- *OpenSubtitles* comprises translated movies and TV subtitles.
- *EU Bookshop* includes publications from the European institutions.
- *MultiUN* composes documents from the United Nations.
- *GIGA* consists of newswire text and is made available in WMT10 shared task.[25]
- *DGT* contains translation memories provided by the Joint Research Center.
- *Global Voices* encompasses news stories from the website Global Voices.
- *TED Talks* includes subtitles from TED talks videos.[26]
- *Euconst* consists of text from the European constitution.

[23] http://www.statmt.org/wmt19/translation-task.html

[24] http://opus.nlpl.eu

[25] https://www.statmt.org/wmt10/

[26] https://www.ted.com

Wikimedia database This includes Wikipedia, Wiktionary, Wikiversity, *etc*. The content is built collaboratively by volunteers around the world.[27]

- *Wikipedia* is a free online encyclopedia including high-quality text covering a wide range of topics.
- *Wikisource* includes source texts in the public domain.
- *Wikinews* contains free-content news.
- *Wiktionary* is an open-source dictionary of words, phrases *etc*.
- *Wikiversity* composes learning resources and learning projects or research.
- *Wikibooks* includes open-content books.
- *Wikiquote* consists of sourced quotations from notable people and creative works.
- *Wikivoyage* includes information about travelling.

Project Gutenberg This popular dataset contains free ebooks of different genres which are mostly the world's older classic works of literature for which copyright has expired.

EnronSent This dataset is provided by (Styler, 2011) and is a part of the Enron Email Dataset,[28] a massive dataset containing 500K messages from senior management executives at the Enron Corporation.

PCT This sub-corpus contains patent documents collected and maintained internally by the GETALP[29] team.

Le Monde This is also collected and maintained internally by the GETALP team, consisting of articles from Le Monde[30] collected from 1987 to 2003.

Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F., (2003). *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Buck, C., Heafield, K., and van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May.

Cañete, J., Chaperon, G., Fuentes, R., and PÃ©rez, J. (2020). Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Constant, M., Candito, M., and Seddah, D. (2013). The ligm-alpage architecture for the spmrl 2013 shared task: Multiword expression analysis and dependency parsing. In *Proceedings of the EMNLP Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Delobelle, P., Winters, T., and Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. In *ICLR*.

Eisele, A. and Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

[27] https://dumps.wikimedia.org/other/cirrussearch/current/

[28] https://www.cs.cmu.edu/~enron/

[29] http://lig-getalp.imag.fr/en/home/

[30] https://www.lemonde.fr

Eisenschlos, J., Ruder, S., Czapla, P., Kardas, M., Gugger, S., and Howard, J. (2019). Multifit: Efficient multilingual language model fine-tuning. In *Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Fan, A., Grave, E., and Joulin, A. (2019). Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*.

Gong, L., He, D., Li, Z., Qin, T., Wang, L., and Liu, T. (2019). Efficient training of bert by progressively stacking. In *International Conference on Machine Learning*, pages 2337–2346.

Hadj Salah, M. (2018). *Arabic word sense disambiguation for and by machine translation*. Theses, Université Grenoble Alpes ; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, December.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.

Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Machine Translation Summit, 2005*, pages 79–86.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. In *Advances in neural information processing systems*.

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

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). Findings of the first shared task on machine translation robustness. *WMT 2019*, page 91.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2015: Extracting large parallel corpora from movie and tv subtitles. In *International Conference on Language Resources and Evaluation*.

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

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894, Nov.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Meta. (2019). Data dumps — meta, discussion about wikimedia projects.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Navigli, R., Jurgens, D., and Vannella, D. (2013). SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, feb.

Nguyen, D. Q. and Nguyen, A. T. (2020). Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Nguyen, T. Q. and Salazar, J. (2019). Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N.,

Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *In EMNLP*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Pham, N.-Q., Nguyen, T.-S., Niehues, J., Muller, M., and Waibel, A. (2019). Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377*.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Prettenhofer, P. and Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Ramachandran, P., Liu, P., and Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and Villemonte de la Clergerie, E. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Segonne, V., Candito, M., and Crabbé, B. (2019). Using wiktionary as a resource for wsd: the case of french verbs. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 259–270.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Sérasset, G. (2012). Dbnary: Wiktionary as a lmf based multilingual rdf network. In *Language Resources and Evaluation Conference, LREC 2012*.

Skadins, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1850–1855.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Styler, W. (2011). The enronsent corpus. *Technical Report 01-2011. University of Colorado at Boulder Institute of Cognitive Science, Boulder, CO*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199.

Vial, L., Lecouteux, B., and Schwab, D. (2018). UFSAC: Unification of Sense Annotated Corpora and Tools. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, May.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, Wroclaw, Poland.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R.

(2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., and Chao, L. S. (2019b). Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xu, H., Liu, Q., van Genabith, J., and Zhang, J. (2019). Why deep transformers are difficult to converge? from computation order to lipschitz restricted parameter initialization. *arXiv preprint arXiv:1911.03179*.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019a). Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019b). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*.

Zhang, H., Dauphin, Y. N., and Ma, T. (2019a). Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321*.

Zhang, Y., Baldridge, J., and He, L. (2019b). Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.