FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français

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

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

FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français

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, benjamin.lecouteux, laurent.besacier, didier.schwab}@univ-grenoble-alpes.fr
{vincent.segonne@etu, bcrabbe@linguist}.univ-paris-diderot.fr, alexandre.allauzen@espci.fr

RÉSUMÉ

Les modèles de langue pré-entraînés sont désormais indispensables pour obtenir des résultats à l'état-de-l'art dans de nombreuses tâches du TALN. Tirant avantage de l'énorme quantité de textes bruts disponibles, ils permettent d'extraire des représentations continues des mots, contextualisées au niveau de la phrase. L'efficacité de ces représentations pour résoudre plusieurs tâches de TALN a été démontrée récemment pour l'anglais. Dans cet article, nous présentons et partageons FlauBERT, un ensemble de modèles appris sur un corpus français hétérogène et de taille importante. Des modèles de complexité différente sont entraînés à l'aide du nouveau supercalculateur *Jean Zay* du CNRS. Nous évaluons nos modèles de langue sur diverses tâches en français (classification de textes, paraphrase, inférence en langage naturel, analyse syntaxique, désambiguïsation automatique) et montrons qu'ils surpassent souvent les autres approches sur le référentiel d'évaluation FLUE également présenté ici.

ABSTRACT

FlauBERT : Unsupervised Language Model Pre-training for French.

Language models have become a key step to achieve state-of-the art results in many NLP tasks. Leveraging the huge amount of unlabeled texts available, they provide an efficient way to pre-train continuous word representations that can be fine-tuned for downstream tasks, along with their contextualization at the sentence level. This has been widely demonstrated for English. In this paper, we introduce and share FlauBERT, a model learned on a very large and heterogeneous French corpus. We train models of different sizes using the new CNRS *Jean Zay* supercomputer. We apply our French language models to several NLP tasks (text classification, paraphrasing, natural language inference, parsing, word sense disambiguation) and show that they often outperform other pre-training approaches on the FLUE benchmark also presented in this article.

MOTS-CLÉS : FlauBERT, FLUE, BERT, français, modèles de langue, évaluation, classification de textes, analyse syntaxique, désambiguïsation lexicale, inférence en langue naturelle, paraphrase.

KEYWORDS: FlauBERT, FLUE, BERT, French, language model, NLP benchmark, text classification, parsing, word sense disambiguation, natural language inference, paraphrase.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 268–278.
HAL : hal-02784776.

1 Introduction

En 2018, l'introduction de représentations linguistiques profondes contextuelles, obtenues à partir de textes bruts, a conduit à un changement de paradigme pour plusieurs tâches du TALN. Alors que les approches fondées sur des représentations continues telles que word2vec (Mikolov *et al.*, 2013) ou GloVe (Pennington *et al.*, 2014) apprennent un vecteur unique pour chaque mot, les modèles introduits récemment produisent des *représentations contextuelles* qui dépendent de la séquence de mots d'entrée complète. Initialement fondées sur des réseaux neuronaux récurrents (Dai & Le, 2015; Ramachandran *et al.*, 2017; Howard & Ruder, 2018; Peters *et al.*, 2018), ces approches ont peu à peu intégré des modèles *Transformer* (Vaswani *et al.*, 2017) comme c'est le cas pour GPT (Radford *et al.*, 2018), BERT (Devlin *et al.*, 2019), XLNet (Yang *et al.*, 2019b), RoBERTa (Liu *et al.*, 2019), ALBERT (Lan *et al.*, 2019) et T5 (Raffel *et al.*, 2019). L'utilisation de ces modèles pre-entraînés a permis des avancées de l'état-de-l'art pour de nombreuses tâches du TALN. Cependant, ceci a surtout été montré pour l'anglais, même si des variantes multilingues sont également disponibles, prenant en compte plus d'une centaine de langues dans un seul modèle : mBERT (Devlin *et al.*, 2019), XLM (Lample & Conneau, 2019), XLM-R (Conneau *et al.*, 2019). Dans cet article [1], nous décrivons notre méthodologie pour construire et partager FlauBERT (**F**rench **La**nguage **U**nderstanding via **B**idirectional **E**ncoder **R**epresentations from **T**ransformers), un modèle BERT pour la français. FlauBERT surpasse le modèle multilingue mBERT dans plusieurs tâches. Nous proposons aussi un référentiel d'évaluation nommé FLUE (**F**rench **L**anguage **U**nderstanding **E**valuation) similaire au benchmark GLUE (Wang *et al.*, 2018) pour l'anglais. FlauBERT et FLUE sont disponibles en ligne pour la communauté TALN. [2]

Étant donné l'impact des modèles contextuels pre-entrainés, plusieurs auteurs ont récemment publié des modèles disponibles pour d'autres langues que l'anglais. Par exemple, ELMo (Peters *et al.*, 2018, ELMo) existe pour le portugais, le japonais, l'allemand et le basque, [3] tandis que BERT a été récemment entraîné pour plusieurs langues (allemand, chinois, espagnol, finnois, italien, néerlandais, suédois). [4] Pour le français, en parallèle du modèle que nous proposons, une équipe jointe INRIA et Facebook a développé CamemBERT (Martin *et al.*, 2019). Une autre tendance considère des modèles estimés sur plusieurs langues avec un vocabulaire commun, comme par exemple une version de BERT multilingue pour 104 langues [5]. Mentionnons également les travaux récents utilisant des données parallèles comme LASER (Artetxe & Schwenk, 2019) pour 93 langues, XLM (Lample & Conneau, 2019) et XLM-R (Conneau *et al.*, 2019) pour 100 langues.

Par ailleurs, l'existence d'un référentiel d'évaluation tel que GLUE (Wang *et al.*, 2018) pour l'anglais est très utile pour stimuler des recherches reproductibles. Les bonnes performances obtenues avec des modèles contextuels pre-entrainés sur la plupart des tâches de TALN couvertes par GLUE ont conduit à son extension. SuperGLUE (Wang *et al.*, 2019) est un nouveau référentiel construit sur les mêmes principes, incluant un ensemble de tâches plus difficiles et variées. Une version chinoise de GLUE [6] est aussi développée pour évaluer la performance du modèle sur cette langue. À ce jour, nous n'avons pas connaissance d'un tel référentiel pour le français, d'où la proposition détaillée en section 3.

1. Cet article est une version traduite et raccourcie de l'article de Le *et al.* (2019), accepté à LREC 2020.
2. https://github.com/getalp/Flaubert
3. https://allennlp.org/elmo
4. Une liste de modèles, en constante évolution, est disponible sur https://huggingface.co/models
5. https://github.com/google-research/bert
6. https://github.com/chineseGLUE/chineseGLUE

2 Apprentissage du modèle FlauBERT

Données d'apprentissage et pré-traitements Nous agrégeons 24 sous-corpus de types divers (wikipedia, livres, *Common Crawl*, ...). Nos trois sources principales sont (1) les textes monolingues des campagnes d'évaluation WMT19 (Li *et al.*, 2019, 4 sous-corpus), (2) les textes en français de la collection OPUS (Tiedemann, 2012, 8 sous-corpus), (3) le projet Wikimedia [7] (8 sous-corpus). La taille totale (non compressée) des textes ainsi agrégés est de 270GB. Après un prétraitement consistant en différents filtrages (enlever les phrases très courtes, les séquences de numéros ou d'adresses électroniques, etc.), une normalisation de l'encodage des caractères, et une tokenisation à l'aide de Moses (Koehn *et al.*, 2007), nous obtenons un corpus de 71GB. Notre code pour télécharger et pré-traiter les données est publiquement disponible. [8]

	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	CamemBERT	$\text{FlauBERT}_{\text{BASE}}$/$\text{FlauBERT}_{\text{LARGE}}$
Langue	Anglais	Anglais	Français	Français
Données d'apprentissage	13 GB	160 GB	138 GB†	71 GB‡
Objectifs de pré-entraînement	NSP et MLM	MLM	MLM	MLM
Nombre total de paramètres	110 M	125 M	110 M	138 M/ 373 M
Tokenisation	WordPiece 30K	BPE 50K	SentencePiece 32K	BPE 50K
Masque	Statique + sous-mots	Dynamique + sous-mots	Dynamique + mot entier	Dynamique + sous-mot

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

TABLE 1 – Comparaison entre FlauBERT et d'autres modèles de langue pré-entraînés.

Objectif de l'entraînement et optimisation Le pré-entraînement du Bert original consiste en deux tâches supervisées : (1) un *modèle de langue masqué* (MLM) qui apprend à prédire des jetons masqués de façon aléatoire ; et (2) une *prédiction de la prochaine phrase* (NSP - *Next Sentence Prédiction*) dans laquelle le modèle apprend à prédire si `B` est une phrase qui suit effectivement `A`, étant donné une paire de phrases d'entrée `A,B`.

Devlin *et al.* (2019) a observé que la suppression de NSP nuit considérablement aux performances sur certaines tâches. Cependant, le contraire a été démontré dans des études ultérieures, notamment Yang *et al.* (2019b, XLNet), Lample & Conneau (2019, XLM), et Liu *et al.* (2019, RoBERTa). [9] Par conséquent, nous avons utilisé seulement l'objectif MLM dans FlauBERT.

Pour optimiser cette fonction objectif, nous avons suivi Liu *et al.* (2019) et utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec les paramètres suivants :

— $\text{FlauBERT}_{\text{BASE}}$: étapes de mise en route (ou *warm up*) de 24k, taux d'apprentissage maximal de $6\text{e}{-4}$, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1\text{e}{-6}$ et perte de poids de 0,01.

— $\text{FlauBERT}_{\text{LARGE}}$: étapes de mise en route de 30k, taux d'apprentissage maximal de $3\text{e}{-4}$, $\beta_1 = 0,9$, $\beta_2 = 0,98$, $\epsilon = 1\text{e}{-6}$ et perte de poids de 0,01.

Modèles et configuration d'apprentissage Nous utilisons la même architecture que BERT (Devlin *et al.*, 2019). Un vocabulaire de 50K unités sous-lexicales est construit en utilisant l'algorithme *Byte Pair Encoding* (Sennrich *et al.*, 2016, BPE). Nous entraînons deux principaux modèles (transformers

7. https://meta.wikimedia.org/w/index.php?title=Data_dumps&oldid=19312805
8. https://github.com/getalp/Flaubert
9. Liu *et al.* (2019) ont émis l'hypothèse que l'implantation originale de BERT pourrait avoir supprimé la fonction de coût associée au NSP tout en conservant le format d'entrée consistant en des paires de phrases.

bi-directionnels) : FlauBERT$_{\text{BASE}}$ (12 blocs de dimension cachée 768, 12 têtes pour l'attention) et FlauBERT$_{\text{LARGE}}$ (24 blocs de dimension cachée 1024, 12 têtes). Le critère d'apprentissage est de type *masked language model* : il consiste à prédire des tokens d'une phrase ayant été préalablement et aléatoirement masqués. FlauBERT$_{\text{BASE}}$ est appris sur 32 GPU Nvidia V100 SXM2 32 GB en 410h et FlauBERT$_{\text{LARGE}}$ est appris sur 128 de ces mêmes GPU en 390h.

3 FLUE

Le référentiel d'évaluation FLUE est composé de 7 tâches correspondant à différents niveaux d'analyse (syntaxique, sémantique) du traitement automatique du français.

Classification de texte Le corpus d'analyse de sentiments translingue CLS (Prettenhofer & Stein, 2010) est constitué de critiques issues du site Amazon pour trois catégories de produits (livres, DVD et musique) en quatre langues : anglais, français, allemand et japonais. Chaque échantillon contient une critique associée à une note allant de 1 à 5. Suivant Blitzer *et al.* (2006) et Prettenhofer & Stein (2010), les évaluations avec 3 étoiles sont écartées et la note est binarisée avec un seuil de 3. Pour chaque catégorie de produit, nous construisons des ensembles d'apprentissage et de test qui sont équilibrés. Les données de test contiennent ainsi 2000 avis en français.

Identification de paraphrases Cette tâche consiste à identifier si des paires de phrases sont sémantiquement équivalentes ou non. PAWS-X est un ensemble de données multilingues pour l'identification des paraphrases (Yang *et al.*, 2019a). Il s'agit de l'extension de la tâche PAWS (Zhang *et al.*, 2019) pour l'anglais à six autres langues : français, espagnol, allemand, chinois, japonais et coréen. Yang *et al.* (2019a) ont utilisé la traduction automatique pour créer les corpus de ces autres langues mais les ensembles de développement et de test pour chaque langue sont traduits manuellement. Nous prenons à nouveau la partie française pour FLUE.

Natural Language Inference (NLI) Cette tâche, également connue sous le nom de reconnaissance d'implications textuelles (RTE), considère une prémisse (p) et une hypothèse (h) et consiste à déterminer si p implique, contredit ou n'implique ni ne contredit h. Le corpus *Cross-lingual NLI Corpus* (Conneau *et al.*, 2018, XNLI) étend l'ensemble de développement et de test du corpus *Multi-Genre Natural Language Inference* corpus (Williams *et al.*, 2018, MultiNLI) à 15 langues. Les ensembles de développement et de test pour chaque langue consistent en 7 500 exemples annotés manuellement, soit un total de 112 500 paires de phrases annotées avec les étiquettes *entailment*, *contradiction* ou *neutre*. FLUE intègre la partie française de ce corpus.

Analyse syntaxique et étiquetage morphosyntaxique Nous considérons deux tâches d'analyse syntaxique : analyse en constituants et en dépendances, ainsi que l'étiquetage morphosyntaxique. Pour cela, nous utilisons le *French Treebank* (Abeillé *et al.*, 2003), une collection de phrases du *Monde* annotées manuellement en constituants et dépendances syntaxiques. Nous utilisons la version de ce corpus de la campagne d'évaluation SPMRL 2014 décrite par Seddah *et al.* (2013), qui contient 14759, 1235 et 2541 phrases pour respectivement l'entraînement, le développement et l'évaluation.

Désambiguïsation lexicale des verbes et des noms La désambiguïsation lexicale consiste à assigner un sens, parmi un inventaire donné, à des mots d'une phrase. Pour la désambiguïsation lexicale de verbes, nous utilisons les données de FrenchSemEval (Segonne *et al.*, 2019). Il s'agit d'un corpus d'évaluation dont les occurrences de verbes ont été annotées manuellement avec les sens de Wiktionary. [10] Pour la désambiguïsation lexicale des noms, nous utilisons la partie française de la tâche de désambiguïsation multilingue de SemEval 2013 (Navigli *et al.*, 2013). Nous adaptons l'inventaire de sens de BabelNet utilisé par Navigli & Ponzetto (2010) pour WordNet 3.0 (Miller, 1995), en convertissant les étiquettes de sens lorsqu'une projection est présente dans BabelNet, et en les supprimant dans le cas contraire. Ce processus de conversion donne un corpus d'évaluation composé de 306 phrases et 1 445 mots français annotés en sens WordNet, et vérifiés manuellement. Les données d'apprentissage sont obtenues par transfert selon la méthode décrite par Hadj Salah (2018), qui consiste à traduire des corpus annotés en sens puis transférer leurs annotations. Nous rendrons disponibles à la fois nos données d'entraînement et d'évaluation.

4 Expériences et résultats

Dans cette section, nous présentons les résultats de FlauBERT sur le référentiel d'évaluation FLUE. Nous comparons les performances de FlauBERT avec BERT multilingue (Devlin *et al.*, 2019, mBERT) et CamemBERT (Martin *et al.*, 2019) sur toutes les tâches. Nous comparons également avec le meilleur modèle non contextuel pour chaque tâche. Nous utilisons les bibliothèques open-source XLM (Lample & Conneau, 2019) et Transformers (Wolf *et al.*, 2019). Nous renvoyons à Le *et al.* (2019) pour une description détaillée des expériences.

Classification de texte Nous avons suivi le processus de réglage fin (*fine tuning*) standard de BERT (Devlin *et al.*, 2019). Le bloc de classification ajouté au dessus du model BERT est composé des couches suivantes : dropout, linéaire, activation $\tanh$, dropout et linéaire. Les dimensions de sortie des couches linéaires sont respectivement égales à la taille cachée du Transformer et au nombre de classes (2). La valeur de dropout a été fixée à 0.1. Nous entraînons le modèle pendant 30 époques, par lots de 8 exemples. Nous testons 4 valeurs de *learning rate* ($1e{-}5$, $5e{-}5$, $1e{-}6$ et $5e{-}6$). Nous utilisons comme ensemble de validation un échantillon aléatoire de 20% des données, pour sélectionner le meilleur modèle. Le tableau 2 présente l'exactitude finale sur l'ensemble de test pour chaque modèle. Les résultats mettent en évidence l'importance d'un modèle monolingue en français pour la classification des textes : CamemBERT et $\text{FlauBERT}_{\text{BASE}}$ surpassent largement mBERT.

Identification de paraphrases La configuration de cette tâche est presque identique à la précédente, la seule différence étant que la séquence d'entrée est maintenant une paire de phrases `A, B`. La performance finale de chaque modèle est indiquée dans le tableau 2. On peut observer que notre modèle monolingue français ne fonctionne que légèrement mieux qu'un modèle multilingue (mBERT), ce qui pourrait être attribué aux caractéristiques de l'ensemble de données PAWS-X. En effet, cet ensemble de données contient des paires de phrases avec une forte proportion de chevauchement lexical, ce qu'un modèle multilingue peut détecter aussi bien qu'un modèle monolingue.

10. Version du 20-04-2018 incluse dans le jeu de donnée.

Tâche	Classification			Paraphrase	NLI	Constituants		Dépendances		Désambiguïsation	
Section	Livres	DVD	Musique							Noms	Verbes
Mesure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
État de l'art ant.	91.25[c]	89.55[c]	93.40[c]	66.2[d]	80.1/**85.2**[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15[c]	86.9[c]	86.65[c]	89.3[d]	76.9[f]	87.5	98.1	89.5	85.86	56.5	44.9
CamemBERT	**93.40**	**92.70**	94.15	89.8	81.2	88.4	**98.2**	91.37	88.13	56.1	**51.1**
FlauBERT$_{\text{BASE}}$	**93.40**	92.50	**94.30**	**89.9**	81.3	**89.1**	98.1	**91.56**	**88.35**	54.9/**57.9**[g]	47.4

TABLE 2 – Résultats finaux sur les tâches de FLUE. [a]Kitaev *et al.* (2019). [b]Constant *et al.* (2013). [c]Eisenschlos *et al.* (2019, MultiFiT). [d]Chen *et al.* (2017, ESIM). [e]Conneau *et al.* (2019, XLM-F BASE/LARGE). [f]Martin *et al.* (2019). [g]Utilise FlauBERT$_{\text{LARGE}}$. [h]Segonne *et al.* (2019).

Natural Language Inference (NLI) Comme cette tâche a également été considérée par Martin *et al.* (2019, CamemBERT), nous utilisons la même configuration expérimentale pour que nos résultats soient comparables. L'entrée du modèle pour cette tâche est aussi une paire de phrases. Nous présentons la performance pour chaque modèle dans le tableau 2. Les résultats confirment la supériorité des modèles français par rapport aux modèles multilingues (mBERT) pour cette tâche. FlauBERT$_{\text{BASE}}$ fonctionne légèrement mieux que CamemBERT. Les deux dépassent clairement XLM-R$_{\text{BASE}}$, bien qu'ils ne puissent pas dépasser XLM-R$_{\text{LARGE}}$. Il convient de noter que XLM-R$_{\text{LARGE}}$ employait une architecture beaucoup plus profonde.

Analyse syntaxique en constituants et étiquetage morphosyntaxique Nous réalisons de manière conjointe l'analyse en constituants et l'étiquetage morphosyntaxique, à l'aide de l'analyseur [11] décrit par Kitaev & Klein (2018) et Kitaev *et al.* (2019). La table 2 présente les résultats. Sans pré-entraînement, nous reproduisons le résultat de Kitaev & Klein (2018). FastText n'améliore pas les résultats. CamemBERT obtient un meilleur résultat que mBERT, grâce à son entraînement monolingue. FlauBERT obtient un encore meilleur résultat (+0.7). Les trois analyseurs utilisant un modèle de langue contextuel obtiennent des résultats similaires en étiquetage morphosyntaxique (98.1-98.2).

Analyse syntaxique en dépendances Pour l'analyse en dépendances, nous utilisons une réimplémentation de l'algorithme de Dozat & Manning (2017) avec décodage par arbre couvrant de poids maximal. L'analyseur prend en entrée des phrases étiquetées en partie du discours. Nous utilisons les tags prédits fournis par la campagne SPMRL. Notre représentation lexicale est une concaténation de plongements lexicaux et de plongements de tags appris avec le reste du modèle d'analyse sur le French Treebank ainsi que d'un vecteur préentrainé. Les résultats sont donnés en table 2. Tous les modèles utilisant les vecteurs BERT font au moins aussi bien que l'état de l'art sur cette tâche et les deux modèles monolingues sont état de l'art avec les vecteurs FlauBERT$_{\text{BASE}}$ qui donnent un résultat marginalement meilleur que les vecteurs CamemBERT. On remarque également que les deux modèles monolingues apportent des résultats substantiellement meilleurs que le modèle mBERT multilingue.

Désambiguïsation lexicale des noms Nous utilisons le réseau de neurones décrit par Vial *et al.* (2019a,b) dont le code est fourni. [12] Il prend, en entrée, les vecteurs issus d'un modèle de langue

11. https://github.com/nikitakit/self-attentive-parser
12. https://github.com/getalp/disambiguate

pré-entraîné, qui restent fixes, puis il est composé de plusieurs couches d'encodeur *Transformer* et d'une couche linéaire qui sont entraînées. La couche linéaire réalise une projection sur l'ensemble des *synsets* vus pendant l'entraînement. Enfin, le *synset* qui obtient le plus haut score est choisi. Nous donnons le résultat issu d'un ensemble de 8 modèles entraînés indépendamment, qui moyenne la sortie du *softmax*. Dans la Table 2, on observe d'abord des performances largement meilleures avec les modèles BERT qu'avec des vecteurs statiques. mBERT obtient de meilleures performances que CamemBERT ainsi que FlauBERT$_{\text{BASE}}$, ce que nous pensons être dû à la nature translingue des corpus d'entraînement, mais FlauBERT$_{\text{LARGE}}$ obtient les meilleurs résultats sur la tâche.

Désambiguïsation lexicale des verbes Nous suivons la méthode décrite par Segonne *et al.* (2019). Nous utilisons les plongements contextuels fournis par les modèles FlauBERT/mBERT/CamemBERT pour les representations vectorielles des occurences (l'inventaire de sens et données d'évaluation). Nous comparons également nos résultats à une représentation plus simple qui consiste à moyenner les plongements lexicaux des mots entourant le mot cible. Pour cette expérience nous avons utilisé les plongements lexicaux issus de FastText avec une fenêtre de mots de taille 5. Les résultats de nos expériences sont présentés dans la table 2. On observe que l'utilisation des modèles BERT pour cette tâche apporte un gain conséquent par rapport à l'état de l'art, les meilleurs résultats étant obtenus par CamemBERT. De plus, nos expériences confirment l'intérêt des modèles spécifiquement entraînés sur le français puisque les deux modèles CamemBERT et FlauBERT$_{\text{BASE}}$ surpassent le modèle multilingue mBERT.

5 Conclusion

Nous avons présenté et partagé FlauBERT, un ensemble de modèles de langues pre-entraînés pour le français, accompagné de FLUE, un dispositif d'évaluation. FlauBERT obtient des résultats à l'état de l'art sur un certain nombre de tâches de TALN. Il est aussi compétitif avec CamemBERT (Martin *et al.*, 2019) – un autre modèle pour le français développé en parallèle – bien qu'il ait été entraîné sur presque deux fois moins de données textuelles. Nous espérons que cette contribution stimulera les recherches sur le TALN en français. [13]

6 Remerciements

Ce travail a bénéficié du programme « Grand Challenge Jean Zay » (projet 100967) et a également été partiellement soutenu par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). Nous remercions Guillaume Lample et Alexis Conneau pour leur soutien technique pour l'utilisation du code XLM.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht. DOI : 10.1007/978-94-010-0201-1_10.

13. FlauBERT est notamment disponible sur https://huggingface.co/models.

ARTETXE M. & 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.

BLITZER J., MCDONALD R. & PEREIRA F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, p. 120–128 : Association for Computational Linguistics.

CHEN Q., ZHU X., LING Z.-H., WEI S., JIANG H. & 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)*, p. 1657–1668.

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

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S., SCHWENK H. & STOYANOV V. (2018). Xnli : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2475–2485.

CONSTANT M., CANDITO M. & 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. & LE Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems*, p. 3079–3087.

DEVLIN J., CHANG M.-W., LEE K. & 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)*, p. 4171–4186.

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

EISENSCHLOS J., RUDER S., CZAPLA P., KARDAS M., GUGGER S. & HOWARD J. (2019). Multifit : Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.

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. HAL : tel-02139438.

HOWARD J. & 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)*, p. 328–339.

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

KITAEV N., CAO S. & 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*, p. 3499–3505, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1340.

KITAEV N. & 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), p. 2676–2686, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1249.

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*, p. 177–180.

LAMPLE G. & 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. & SORICUT R. (2019). Albert : A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

LI X., MICHEL P., ANASTASOPOULOS A., BELINKOV Y., DURRANI N., FIRAT O., KOEHN P., NEUBIG G., PINO J. & SAJJAD H. (2019). Findings of the first shared task on machine translation robustness. *Fourth Conference on Machine Translation (WMT19)*, p. 91–102.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & 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, p. 3111–3119, USA : Curran Associates Inc.

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

NAVIGLI R., JURGENS D. & 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)*, p. 222–231.

NAVIGLI R. & 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*, p. 216–225 : Association for Computational Linguistics.

PENNINGTON J., SOCHER R. & 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. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, p. 2227–2237.

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

RADFORD A., NARASIMHAN K., SALIMANS T. & 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. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

RAMACHANDRAN P., LIU P. & LE Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 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. & 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*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

SEGONNE V., CANDITO M. & 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*, p. 259–270.

SENNRICH R., HADDOW B. & 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)*, p. 1715–1725.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

VIAL L., LECOUTEUX B. & SCHWAB D. (2019a). Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL)*, Toulouse, France. HAL : hal-02092559.

VIAL L., LECOUTEUX B. & SCHWAB D. (2019b). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, Wroclaw, Poland. HAL : hal-02131872.

WANG A., PRUKSACHATKUN Y., NANGIA N., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. R. (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & 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*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WILLIAMS A., NANGIA N. & 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)*, p. 1112–1122.

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

YANG Y., ZHANG Y., TAR C. & 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. & LE Q. V. (2019b). Xlnet : Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*.

ZHANG Y., BALDRIDGE J. & HE L. (2019). 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)*, p. 1298–1308.