HAL open science

Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms

Alexandra Benamar, Meryl Bothua, Cyril Grouin, Anne Vilnat

To cite this version:

Alexandra Benamar, Meryl Bothua, Cyril Grouin, Anne Vilnat. Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms. NL4IA Workshop Proceedings, Nov 2021, Milan, Italy. hal-03474696

HAL Id: hal-03474696

https://hal.science/hal-03474696v1

Submitted on 10 Dec 2021

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

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

Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms

Alexandra Benamar[1,2], Meryl Bothua[2], Cyril Grouin[1], and Anne Vilnat[1]

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France,
[first name].[last name]@lisn.upsaclay.fr
[2] EDF R&D, Palaiseau, France [first name].[last name]@edf.fr

Abstract. In this paper, we present BERT-POS, a simple method for encoding syntax into BERT embeddings without re-training or fine-tuning data, based on Part-Of-Speech (POS). Although fine-tuning is the most popular method to apply BERT models on domain datasets, it remains expensive in terms of training time, computing resources, training data selection and re-training frequency. Our alternative works at the preprocessing level and relies on POS tagging sentences. It gives interesting results for words similarity regarding out-of-vocabulary both in terms of domain-specific words and misspellings. More specifically, the experiments were done on French language, but we believe that they would be similar on others.

Keywords: Natural Language Processing · Language Models · Semantic Similarity · Out-of-Vocabulary Words · Part-Of-Speech

CEUR-WS.org/Vol-3015/paper132.pdf

1 Introduction

For a variety of Natural Language Processing (NLP) tasks, state-of-the-art results have been reported with generic pre-trained language models, such as BERT [2] and other BERT-like models [14,19] or task-specific such as GPT [23] designed for automatic text generation. In these approaches, the pre-trained language models are applied to downstream machine learning tasks using task-specific fine-tuning. Currently, Transformer models [29] are trained on different sets of generic data (i.e., books, news, Wikipedia, etc.) and are not adapted to domain datasets, both in terms of vocabulary or syntactic structure. Therefore, these models are not intended to be used as is but should be tailored to specific data sets. At the word level, two types of out-of-vocabulary (OOV) words must be correctly processed: application-specific and misspelled words. In this paper, we propose a novel method to improve semantic understanding of domain-specific data. To do this, we present BERT-POS, an easy-to-use technique to integrate external morpho-syntactic context into BERT-like architectures. The proposed method combines BERT with an automatic preprocessing stage which saves

[1] Copyright ©2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

computing time (i.e., Fast learning) and energy consumed (i.e., Green AI). The use of syntax combined with contextual models enable the addition of contextual characteristics in corpora that are difficult to process. The addition of morpho-syntactic information allows to compensate for the difficulties related to the processing of OOVs by integrating a knowledge of sentence structure. BERT-POS is based on a pre-training technique that is not only robust on the processing of domain-specific terms but also on misspelled terms. This study is conducted on a French dataset through the CamemBERT model [19].

2 Related Work

Fine-tuning language models The problem of adapting the language models was studied [24] and suggested that combining BERT with other neural networks obtained better results than fine-tuning BERT-like models, which was favored in other studies [16,27,33,3]. Specific models are shown to perform best when they are specific to the textual genre studied (i.e., SciBERT [1] and BioBERT [15]). However, pre-training BERT-like models can be computationally expensive and require having a dataset representative of the target data.

Words segmentation Some studies have shown that the decisions made by BERT tokenizers are difficult to explain when splitting words [25]. It was demonstrated that the processing of domain-specific OOV terms is strongly impacted by the splitting of the input terms of the model, leading to a significant decrease in the semantic understanding of the words [20]. Recent works on misspelling generation [26,28] proved that BERT is not robust on misspellings and performed significantly worse on downstream tasks.

Overcoming OOVs in BERT Several studies have worked on overcoming domain specific OOVs and misspellings in BERT. For instance, [4,17] proposed to construct representations at the character-level and obtained promising results for domain-specific terms. Other studies have tried to add external features to deal with misspellings such as a word-recognition module [22] or other strategies [5,8].

3 Proposed Method

In this section, we propose BERT-POS, a preprocessing method for encoding morpho-syntactic information into BERT-like embeddings which does not require a complementary phase of fine-tuning [4,15]. Figure 1 presents the processing chain of our method. For this experiment, we chose CamemBERT because it used SentencePiece [13], which was easy to use when working with re-constructing words from sub-units. Nevertheless, we assume that this work could be easily applied to architectures that use WordPiece [32] such as BERT. First, the dataset was split into sentences or sequences of words when the sentences were too difficult to distinguish. Empirically, we split the documents into sequences of 150 tokens. The POS tagging step consists of concatenating each word with its POS using "_"

character. Here is an example of annotating a sentence containing n words and m POS tags: $word_1_pos_a$, $word_2_pos_b$, $word_3_pos_a$, ..., $word_n_pos_m$. This annotation technique is commonly used for non-contextual models to disambiguate polysemous words which differ in their grammatical category. Here, our objective is to force the addition of morpho-syntactic information in the embeddings. For a given sentence, if the SentencePiece tokenizer does not recognize a word, it splits it into known sub-units. This creates problems with new sentence structures containing a lot of small words. In parallel, we encode a vector for each word and a vector for each POS tag. For every word, a vector is generated by computing the sum of the sub-vectors associated with the sub-tokens of the words. The same process is done with tags and subtags. We made sure that all the POS tags were not recognized as words so that a unique embedding is re-constructed for each tag. Finally, we computed the average of each occurrence of the pairs $\{word_i, pos_j\}$ to construct an unique vector for each word of the corpus.

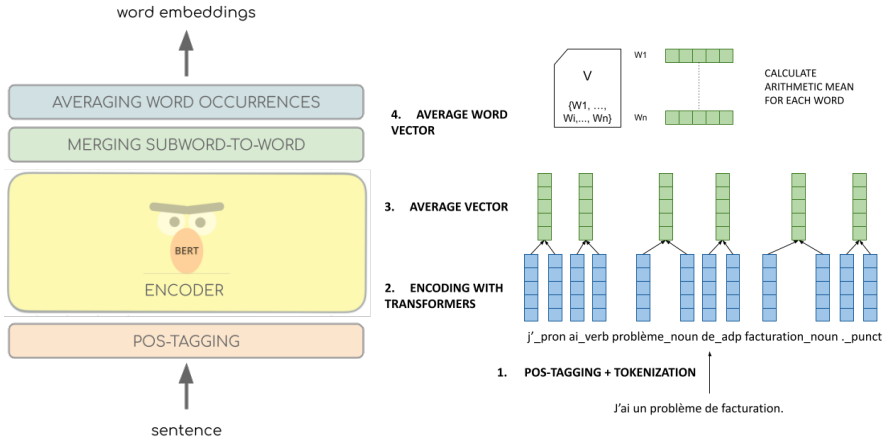


Fig. 1: BERT-POS Framework (left) and one encoding example for the sentence *J'ai un problème de facturation*, which could be translated as "I have a billing problem" and tagged as " I_prop have_verb a_det billing_noun problem_noun ._punct" (right)

4 Datasets

Before detailing our experiments and results, we present our datasets containing French emails in Table 2a. Both datasets are made-up of French email messages:

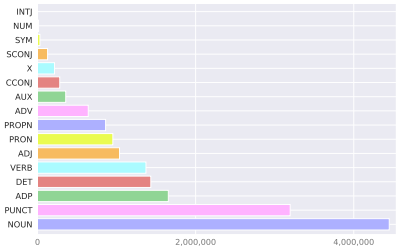
- EASY [21]: a subset of the corpora was extracted to only collect the emails. The dataset is annotated in syntactic relations.
- EDF-Emails[3]: anonymized customer emails extracted from October 2018 to October 2019. This dataset is more difficult to process, since it contains emails with different formality levels, containing spelling and syntactic errors.

[3] This work is part of a broader study for Electricité De France (EDF) with the aim of improving a classification system. EDF is the leading electricity supplier in France.

Moreover, it contains Energy-specific vocabulary which can be existing words in French or words belonging to the specific domain. Table 1 contains several examples of misspellings, SMS language and domain terms that exist in the corpus. The distribution of POS tags in this corpus, obtained with spaCy[4] [7], is described in Figure 2b.

Dataset	Domain	# *docs*
EASY	Diverse genres	765
EDF-Emails	Customer emails	99 993

(a) Description of the datasets



(b) Distribution of POS tags in EDF-Emails

Fig. 2: Datasets' content and POS tags distribution.

Email	**Translation**
Bonjour je suis PERSON je envoye un message pour ve dire cset possible pour peyer la facture peu à peu pasque je pas bouceaoup l' argent .. S' il vous plait . Merci	*Hello I am PERSON I sen a message to tell ye ist possible for peying the bill step by step becose I not alot on money. Please. Thank you*
Bonjour , Nous souhaitons être informés et bénéficiés de votre offre Mes jours Zen et mes jours Zen plus . Dans l' attente de votre retour par téléphone Cordialement PERSON	*Hello, We would like to be informed and benefited about your My Zen days and my Zen days plus offer. Waiting for your return by phone Regards PERSON*
Bonjour Je voulais savoir comment cela se passe comme je vous ai fait parvenir un chèque énergie de 48€ ??? . . . Cordialement ☺☺	*Hello I wanted to know how it goes as I sent you an energy check of 48 € ??? . . . Regards* ☺☺

Table 1: Examples of emails in EDF-Emails dataset with translations in English. PERSON: anonymized name; red: syntactic errors; violet: domain-specific expressions; orange: smileys

5 Transformer Models

Table 2 presents the pre-trained CamemBERT models used for the experiments, without any fine-tuning. To study the impact of training datasets on performance, we use four CamemBERT models[5] which differ by the datasets used during training:

[4] We randomly selected and manually annotated the first 300 tokens of EASY and EDF-Emails datasets and compared the results obtained with spaCy (*fr_core_news_lg*) to calculate a POS tagging accuracy for the respective datasets: 0.95 and 0.83.

[5] We worked with the models implemented in the *transformers* library [31]. The models were downloaded on May 2021.

- Oscar [19] is a set of monolingual corpora extracted from Common Crawl. It was selected using a classification model for each language following the approach of [6] based on FastText [12]. The classifier was previously pre-trained on Wikipedia, Tatoeba and SETimes, and covering 176 languages.
- CCNet [30] is a dataset extracted from Common Crawl but with a different filtering from that of Oscar. It was built with a language model using Wikipedia, thus allowing it to filter out noise (code, tables, etc.). CCNET thus contains documents longer on average than Oscar.
- Wikipedia is a homogeneous corpus in terms of genre and style which was preprocessed using WikiExtractor.

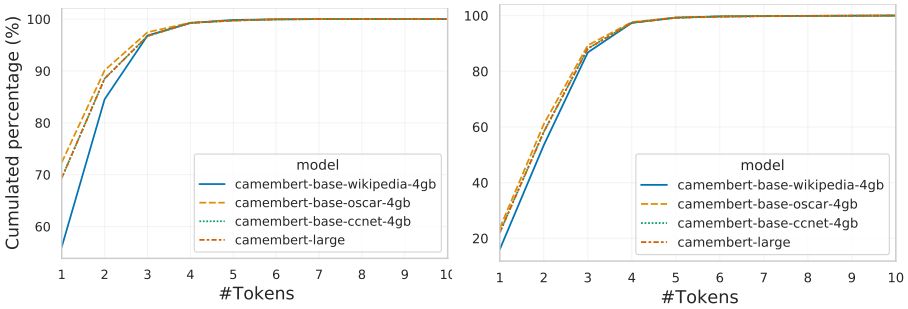
Models	*#layers*	Dataset	Size (GB)
camembert-base-oscar-4gb	12	Oscar	4
camembert-base-ccnet-4gb	12	CCNet	4
camembert-base-wikipedia-4gb	12	Wikipedia	4
camembert-large	24	CCNet	135

Table 2: CamemBERT models' description

6 Experiments

In this section, we aim to assess the impact of the training dataset on language models, to analyze its importance in terms of quality and distance towards the applicative dataset.

6.1 Tokenization Problems on Misspelled and Domain Terms



(a) EASY (b) EDF-Emails

Fig. 3: Cumulative percentage of the number of sub-tokens obtained for each word of the vocabularies

Figure 3 presents the differences between our datasets and the training datasets of CamemBERT's models, presented in Section 5. For each word of the vocabularies, we compute the number of tokens obtained by the models presented

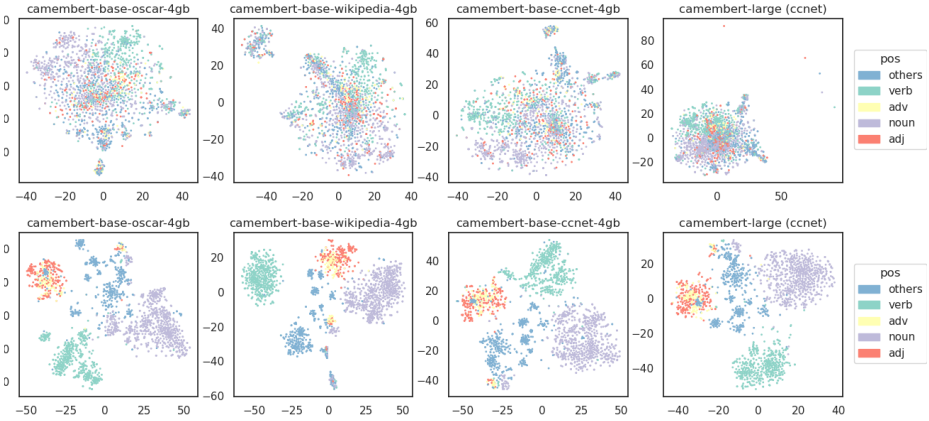
in Table 2. The more tokens are obtained for a single word, the less the model is semantically accurate. We note that for EASY and EDF-Emails, the Wikipedia corpus is the furthest one regarding lexical proximity. It could be explained because the dataset is the lexically poorest from the ones extracted from the web or because our domains of applications are more present in the web-extracted corpora. This result is very relevant, because it shows that the level of cleanliness of the learning corpus (i.e., construction of sentences, order of words, etc.) is not more important than the proximity to the application corpus. Moreover, there is no differences when using CamemBERT using OSCAR than CCNet, which implies that the pre-processing step of CCNet does not have any impact on our datasets. Therefore, we will not use CamemBERT's CCNet model in further analysis. The vocabulary of the EASY dataset is known, at best, at 70% while the one from EDF-Emails is only understood at 20%. Those major differences are expected to be seen while computing similarity, as discussed in Section 6.3. Examples of tokenization with CamemBERT's models is presented in Table 3, using four frequent words in EDF-Emails: domain-specific (i.e., meter, linky and refund) and Emails-specific (i.e., cordially). The domain-specific words exist in all models' vocabularies, except for "linky" (i.e., a French electric meter proposed by EDF), which does not exist in general French language. Interestingly, we observe that the Wikipedia model tokenize this word differently than the others. The segmentation of OOV words is purely based on statistics rather than linguistic properties [20]. This can lead to a loss of semantics when reconstructing words after their tokenization. Indeed, we expect to obtain different words surrounding "linky" when using CamemBERT's Wikipedia compared to the others, due to the sub-units obtained following the tokenization.

Word	**Model**	**Tokens**
linky	*wikipedia*	["_l","in","ky"]
	others	["_l","ink","y"]
remboursement	*wikipedia*	["_rem","bour","s","ement"]
	others	["_remboursement"]
cordialement	*wikipedia*	["_cord","iale","ment"]
	others	["_cordialement"]

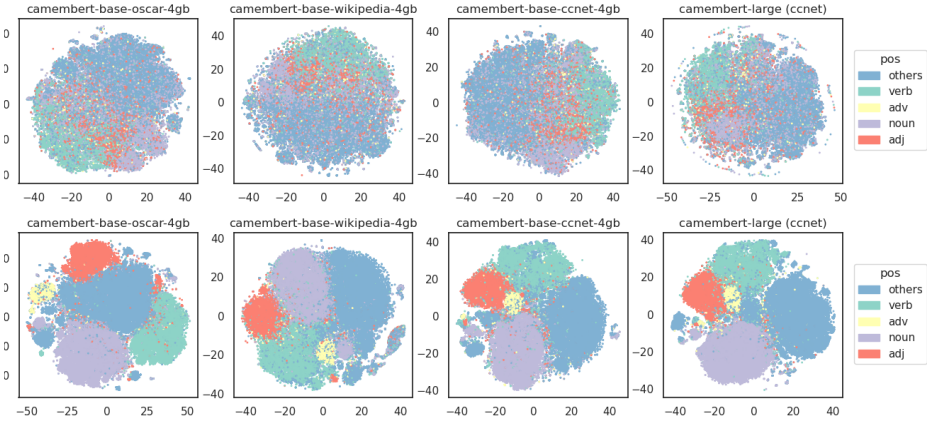
Table 3: Tokenization of domain-specific terms with CamemBERT models (i.e., SentencePiece tokenizer) on EDF-Emails. The models presented in Table 2 based on OSCAR, CCNet and the large model obtained the same results and are referred to as *others*

6.2 Visualizing Differences in Global Structure

Figure 4 presents the significant impact of applying BERT-POS on vocabulary distribution by visualizing the extracted representations of words from EASY and EDF-Emails datasets. The representations are visualized by t-SNE [18]. As expected, we demonstrate that word vectors from our approach are more separable regarding POS categories than those from CamemBERT. This indicates



(a) EASY Dataset - CamemBERT (top) and CamemBERT-POS (bottom)



(b) EDF-Emails Dataset - CamemBERT (top) and CamemBERT-POS (bottom)

Fig. 4: t-SNE visualization of words with CamemBERT

that we managed to cluster syntactically similar words together by adding POS features into CamemBERT before encoding data. To validate our observations, we carried out a k-means clustering with Euclidean distance. We use two metrics to evaluate clustering results objectively: purity and Normalized Mutual Information (NMI). Given that, we do not seek to obtain a single representative cluster of each morpho-syntactic category but several clusters, the purity metric is particularly interesting in this study. We perform k-means clustering 10 times on EDF-Emails, and on each implementation randomly generate the initial seeds. We select the number of clusters with the elbow method [11]. The results are detailed in Table 5 and highlight that the size of training data does not modify the syntactic representation of terms. There are two possible explanations for this: 1) the small dataset contains representative examples of the larger one or 2) a small dataset is sufficient to model syntactic properties of sentences, as computed by CamemBERT.

Word	Model	Neighbors
Train set: OSCAR		
linky (proper noun)	CBERT	**linkys, linkie, linké, linked, linkl**
	CBERT-POS	ginko, zac, cbe, log, installateur
	Fine-tuning	**linkie**, **linked**, **linkdy**, **linké**, **linkys**
remboursement (noun) *refund*	CBERT	règlement, débit, transfert, retrait, rétablissement *settlement, debit, transfer, withdrawal, reinstatement*
	CBERT-POS	services, intervention, règlement, télépaiment, besoin *services, intervention, payment, telepayment, need*
	Fine-tuning	règlement, informée, surtout, non, gratuit *settlement, (is) informed, mostly, no, free*
cordialement (adv) *cordially*	CBERT	merci, bonne, ph, obtenez, sincère *thanks, good, ph, (you) get, sincere*
	CBERT-POS	**cordiallement**, chaleureusement, sincèrement, infiniment, remerciant **cordiallly, warmly, sincerely, infinitely, thanking*
	Fine-tuning	restant, si, merci, quelle, bonne *remaining, yes, thank you, which, good*
Train set: Wikipedia		
linky, (proper noun)	CBERT	**linki**, **linkin**, **linke**, **linkey**, **linké**
	CBERT-POS	**linki**, ld, **li**, **link**, log
	Fine-tuning	**linki**, **lindky**, **linly**, **lynky**, **linxy**
remboursement (noun) *refund*	CBERT	**remboursements**, **remboursment**, **rembourssement**, **remboursés**, **remboursable** *refunds, *refnd, *refuund, (they were) reimbursed, refundable*
	CBERT-POS	**remboursements**, **rembousementt**, reglement, règlement, régularisations *refunds, refundd, *règulations, regulations, regularizations*
	Fine-tuning	**remboursements**, **remboursemment**, **remboursrment**, **remboursemenr**, **reimboursement** *refunds, *refundds, *refnd, *refundr, reimbursement*
cordialement (adv) *cordially*	CBERT	**cordialemement**, **cordialment**, **cordialemment**, **cordiales**, **cordialementt** **cordialylly, *cordialy, *cordiallly, *cordiales, *cordiallyy*
	CBERT-POS	**cordiallement**, **cordiales**, franchement, amicalement, chaleureusement **cordiallly, *cordiales, frankly, kindly, warmly*
	Fine-tuning	**cordialment**, **cordiales**, **cordiallement**, **cordiale**, **cordialelent** **cordialy, *cordiales, *cordially, *cordial, *cordiallially*

Table 4: First 5 neighbors of frequent words using CamemBERT, CamemBERT-POS and CamemBERT after fine-tuning. : translated word containing spelling mistakes. Words in **bold** share the same root as the input word. Pronouns in translated verbs indicates their conjugation in French. CBE: Electronic Blue Counter - GINKO: Enedis' Information System serving the Linky smart meter - ZAC: Joint Development Zone - Sub: abbreviation for "subdivision"

		# *clusters*								
		13			*14*			*15*		
Model	**Metric**	CBERT	CPOS	FT	CBERT	CPOS	FT	CBERT	CPOS	FT
oscar	NMI	.150	.481	.164	.151	**.498**	.165	.153	.496	.163
	Purity	.518	.838	.584	.521	.853	.589	.521	**.862**	.589
wiki.	NMI	.128	**.484**	.164	.122	.470	.165	.124	.462	.157
	Purity	.495	**.862**	.592	.490	.836	.601	.490	.838	.592
large	NMI	.130	**.519**	-	.130	.515	-	.131	.513	-
	Purity	.555	.869	-	.559	.877	-	.560	**.882**	-

Table 5: K-means clustering after t-SNE on EDF-Emails. We compare the quality of the results (i.e., using NMI and *purity* metrics), according to the clustering of morpho-syntactic categories, between CBERT (i.e., CamemBERT), CPOS (i.e., CamemBERT-POS) and FT (i.e., Fine-Tuned CamemBERT, see Section 6.4)

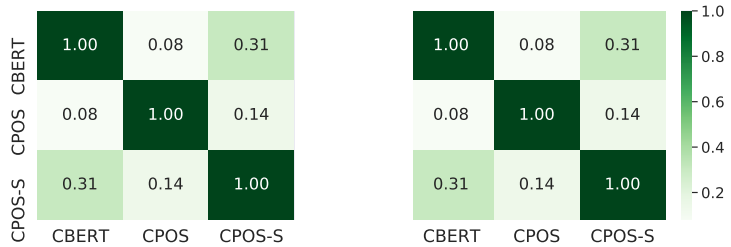
6.3 Comparing local neighborhoods

Both models demonstrate semantic and syntactic sensitivity regarding word similarity. It is observed through comparing the nearest associates for a given word on EDF-Emails dataset, as presented in Table 4. We use the EDF-Email dataset because it contains more noise than general domain. Nevertheless, we computed similar results with the EASY dataset, as shown in Table 7. We computed cosine similarity between frequent words and the rest of the vocabulary to evaluate the neighbors surrounding these words obtained with both models. Applying the *camembert-base-wikipedia-4gb* model on EDF-Emails allows to generate strong similarities between terms which share the same root, or which are spelling variants of existing words. On the contrary, using the *camembert-base-oscar-4gb* model produces clusters of synonyms or words that appear in a similar context. Most of the time, CamemBERT finds similar words according to word structure: it associates verbs with their conjugated forms while not always respecting the proximity regarding the tense of the verbs. However, CamemBERT-POS enhances the possibility of regrouping words that appear in the same context: synonyms and antonyms. However, two distinct phenomena are observed. First, the term "linky", which does not resemble any word in the general field, is now associated with other very specific domain terms, such as another type of electric meter or even meter installation areas. Second, these domain-specific terms are not chosen randomly and have close links, indicating that CamemBERT-POS does not only cluster random OOVs together but keeps the meaning of the terms. Therefore, the proposed method avoids relying on the tokenization step as much by adding morpho-syntactic context. To quantify the differences between the neighbors generated by CamemBERT and CamemBERT-POS, we use comparative metrics. We implemented the Jaccard distance [9], which estimates how dissimilar two sets are by computing the number of intersecting elements in two sets. To calculate the distance, the first 50 neighbors obtained by each method were used and we computed the dissimilarity between the sets of neighbors obtained with CamemBERT and CamemBERT-POS. We averaged

the similarities obtained for the hundred most frequent words in the corpus. As shown in Figure 5, both models generate significantly different neighbors with a Jaccard similarity averaging 0.08, confirming that CamemBERT-POS drastically changes words representation.

Word	CBERT	CBERT-POS
kikou	idem, grâce, mauvaise, félicitations, ok	**salut**, cool, **bonjour**, **bonsoir**, félicitations
cool	sb, **ok**, combien, gaffe, quant	**ok**, joueur, **okidoki**, **super**, gaffe
salut	**bonjour**, moi, ok, hello, oui	cool, **bonsoir, bonjour**, félicitations, **hello**

Table 6: First 5 neighbors of words written in familiar language in EASY dataset using *camembert-base-wikipedia-4gb*. Words in **bold** are synonyms



(a) Oscar (b) Wikipedia

Fig. 5: Jaccard similarity for the 50 closest neighbors of the 100 most frequent words.

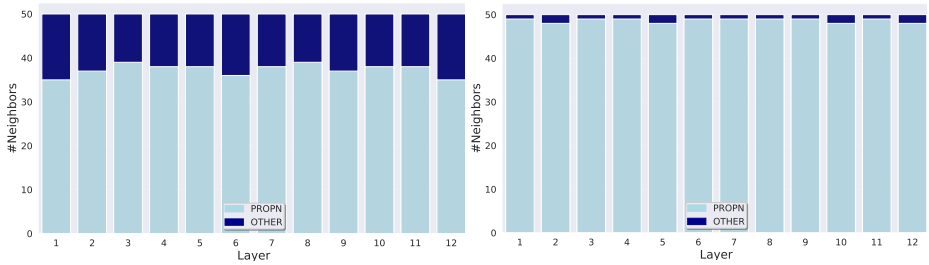
6.4 Fine-Tuning

We aim to compare the results obtained with CamemBERT-POS regarding OOV terms with CamemBERT after fine-tuning the language model. Our implementation follows the fine-tuning example released in the BERT project to use a vanilla baseline to compare against. All hyperparameters remain as default values. We trained the model on two Epochs, using 100,000 Emails. The results are presented in Table 4. Surprisingly, the results obtained after fine-tuning are not that different from the ones with CamemBERT. It mostly generates spelling variations in OOV's neighborhood. For this application, fine-tuning does not seem adequate when working with domain-specific data when we aim to deal with emerging terms in a context of poor writing. As we do not intend to re-train the model frequently, the process of adding external and automatic features is more adapted to our application study. Furthermore, Table 5 shows that fine-tuning the language model slightly improved the processing of morpho-syntactic words.

6.5 Ablation Study

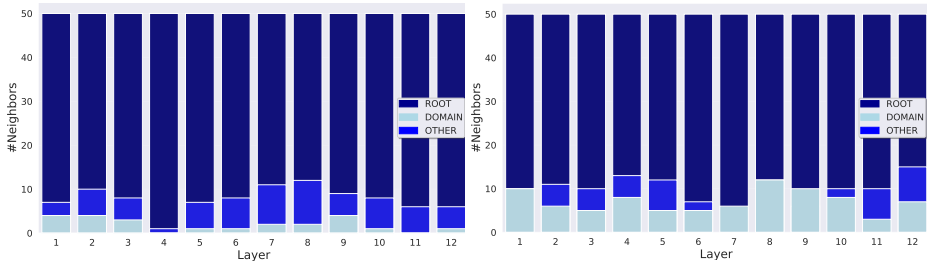
Layer selection BERT encodes multiple types of characteristics depending on the network layer used to represent sentences [10]: the first layers encode morpho-

syntactic information better than higher layers. To evaluate the impact of the choice of the layer in our evaluation, we observe the differences of neighbors for the word "linky" for the dataset EDF-Emails with the model *camembert-base-wikipedia-4gb* in Figures 6 and 7. We note that the neighbors caracteristics remain consistent from one layer to another. CamemBERT-POS regroup similar POS tags together and reduce the distance between semantically close words. With CamemBERT-POS, the new interesting neighbors are either related to electrical offers ("smart", "blue", "green", etc.), other electrical meters (SMA, CBE, meter, etc.) or installation companies (Scopelec, ENEDIS, etc.).



(a) CBERT (b) CBERT-POS

Fig. 6: Closest first 50 neighbors of "linky" computed using cosine similarity divided in two categories: neighbors that are proper nouns and others



(a) CBERT (b) CBERT-POS

Fig. 7: Closest first 50 neighbors of "linky" computed using cosine similarity divided in three categories: neighbors that share the same root as "linky" (ROOT), terms that are relevant and domain-specific (DOMAIN) and others (OTHER)

Number of POS tags A final experiment was carried out to determine whether a high level of POS knowledge was required, or if only certain POS were relevant. To answer this, we built CamemBERT-POS-Small and calculated the neighbors as before. We chose the most important morpho-syntactic categories with regards to semantics: nouns, verbs, adjectives and adverbs. Results are shown in Table 7. At first sight, we notice that this method does not answer the problem of tokenization with the Wikipedia model as well as CamemBERT-POS for these

words. Interestingly, we observe that the cloud is less altered with this method than with the complete CamemBERT-POS, as shown in Figure 5. Yet, we obtain other very relevant synonyms for domain words like "meter" and "refund". We conclude that CamemBERT-POS requires having a fine-grained knowledge of the syntax to get around the processing of OOV terms. However, the word cloud can be impacted by adding a few relevant tags. The addition of these tags allows to obtain interesting clusters of semantically close neighbors.

Word	Oscar	Wikipedia
linky	**linki**, **link**, compteur, **linkie**, lotissement **linki, *link, meter, *linkie, subdivision*	**linki**, **link**, lot, lotissement, **li** **linki, *link, sub, subdivision, *li*
remboursement	paiement, réglement, rattrapage, retrait, règlement *payment, *séttlement, catch-up, withdrawal, settlement*	**remboursements**, **rembourssement**, **remboursment**, **remboursez**, **remboursemen**, *refunds, *refuund, *refnd, (you) repay, *refun*
cordialement	bisous, re, bref, client, heureusement *kiss, re, anyway, customer, fortunately*	**cordiallement**, **corialement**, **cordilement**, sincère, sincerement **cordiallly, corially, cordilly, sincere, sincerely*

Table 7: First 5 neighbors of frequent words using CamemBERT-POS-Small, presented in Section 6.5. *: translated word containing spelling mistakes. Words in **bold** share the same root as the input word.

7 Conclusion and Future Work

We studied the effect of syntactic noise (i.e., spelling mistakes) and domain-specific vocabulary in French textual data on the performance of CamemBERT. We further show that, on a difficult corpus, the proximity between words is drastically impacted by the tokenization of OOV words. To address the problem of noisy vocabulary (i.e., OOV), we propose BERT-POS, a method that reduces the impact of tokenization while processing OOV terms. Our work stands out from the literature in two ways. First, the combination of morpho-syntactic markers and language models remains a very limited field of research, in which our work fits. Even though BERT is a contextual model, new words can alter the structure of the sentences entering the model. External markers (i.e., morpho-syntactic markers) allow sentences to be re-structured when they become too fragmented. Second, we offer a model that does not require re-training or fine-tuning and is easy to set up, which is, to our knowledge, the first such model built with a goal of improving tokenization issues. In our future work, we want to evaluate the impact of adding syntax on different tasks, by conducting a large number of experiments on different domain datasets. This will allow us to assess the robustness of our method in different domains and on several tasks.

References

1. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
3. El Boukkouri, H.: Ré-entraîner ou entraîner soi-même? stratégies de pré-entraînement de bert en domaine médical. In: Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). pp. 29–42 (2020)
4. El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., Tsujii, J.: CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6903–6915. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.609, https://aclanthology.org/2020.coling-main.609
5. Fukuda, N., Yoshinaga, N., Kitsuregawa, M.: Robust Backed-off Estimation of Out-of-Vocabulary Embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4827–4838. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.434, https://aclanthology.org/2020.findings-emnlp.434
6. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893 (2018)
7. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303, https://doi.org/10.5281/zenodo.1212303
8. Hu, Y., Jing, X., Ko, Y., Rayz, J.T.: Misspelling correction with pre-trained contextual language model. In: 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). pp. 144–149. IEEE (2020)
9. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)
10. Jawahar, G., Sagot, B., Seddah, D.: What Does BERT Learn about the Structure of Language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3651–3657. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1356, https://www.aclweb.org/anthology/P19-1356
11. Joshi, K.D., Nalwade, P.: Modified k-means for better initial cluster centres. International Journal of Computer Science and Mobile Computing **2**(7), 219–223 (2013)
12. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
13. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of

the 2018 EMNLP: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-2012, http://aclweb.org/anthology/D18-2012
14. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbe, B., Besacier, L., Schwab, D.: FlauBERT: Unsupervised Language Model Pre-training for French. In: LREC. Marseille, France (2020), https://hal.archives-ouvertes.fr/hal-02890258
15. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
16. Li, F., Jin, Y., Liu, W., Rawat, B.P.S., Cai, P., Yu, H.: Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. JMIR medical informatics **7**(3), e14830 (2019)
17. Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., Hu, G.: Charbert: Character-aware pre-trained language model. arXiv preprint arXiv:2011.01513 (2020)
18. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research **9: 2579–2605** (2008)
19. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.645, https://aclanthology.org/2020.acl-main.645
20. Nayak, A., Timmapathini, H., Ponnalagu, K., Venkoparao, V.G.: Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words. In: Proceedings of the First Workshop on Insights from Negative Results in NLP. pp. 1–5 (2020)
21. Paroubek, P., Robba, I., Vilnat, A., Ayache, C.: Data, annotations and measures in easy the evaluation campaign for parsers of french. In: LREC. pp. 315–320. Citeseer (2006)
22. Pruthi, D., Dhingra, B., Lipton, Z.C.: Combating adversarial misspellings with robust word recognition. arXiv preprint arXiv:1905.11268 (2019)
23. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
24. Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., Stent, A., Kumar, Y., Shah, R.R., Zimmermann, R.: Keyphrase extraction as sequence labeling using contextualized embeddings. Advances in Information Retrieval **12036**, 328 (2020)
25. Singh, J., McCann, B., Socher, R., Xiong, C.: BERT is not an interlingua and the bias of tokenization. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 47–55. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-6106, https://aclanthology.org/D19-6106
26. Srivastava, A., Makhija, P., Gupta, A.: Noisy text data: Achilles' heel of bert. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 16–21 (2020)
27. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics. pp. 194–206. Springer (2019)

28. Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., Xiong, C.: Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv preprint arXiv:2003.04985 (2020)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
30. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, É.: Ccnet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4003–4012 (2020)
31. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
32. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
33. Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M., Lin, J.: Data augmentation for bert fine-tuning in open-domain question answering. arXiv preprint arXiv:1904.06652 (2019)