Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4193–4204
Marseille, 20-25 June 2022

Evaluating Tokenizers Impact on OOVs Representation with Transformers Models

Alexandra Benamar, Cyril Grouin, Meryl Bothua, Anne Vilnat
Université Paris-Saclay, CNRS, LISN, Orsay, France
{first name}.{last name}@lisn.upsaclay.fr
EDF Lab R&D, Palaiseau, France
{first name}.{last name}@edf.fr

Abstract

Transformer models have achieved significant improvements in multiple downstream tasks in recent years. One of the main contributions of Transformers is their ability to create new representations for out-of-vocabulary (OOV) words. In this paper, we have evaluated three categories of OOVs: (A) new domain-specific terms (e.g., "eucaryote" in microbiology), (B) misspelled words containing typos, and (C) cross-domain homographs (e.g., "arm" has different meanings in a clinical trial and anatomy). We use three French domain-specific datasets on the legal, medical, and energetical domains to robustly analyze these categories. Our experiments have led to exciting findings that showed: (1) It is easier to improve the representation of new words (A and B) than it is for words that already exist in the vocabulary of the Transformer models (C), (2) To ameliorate the representation of OOVs, the most effective method relies on adding external morpho-syntactic context rather than improving the semantic understanding of the words directly (fine-tuning) and (3) We cannot foresee the impact of minor misspellings in words because similar misspellings have different impacts on their representation. We believe that tackling the challenges of processing OOVs regarding their specificities will significantly help the domain adaptation aspect of BERT.

Keywords: Out-of-vocabulary, Transformer models, Sub-units tokenization, Language models

1. Introduction

The most recent NLP models based on neural networks, such as Long-Short Term Memory (LSTM) networks (Peters et al., 2018), attentive convolution (Yin and Schütze, 2018) and Transformer models (Vaswani et al., 2017), are trained and evaluated on general-domain datasets. Pre-trained Transformer models such as BERT (Devlin et al., 2018) have proven their effectiveness in adapting to multiple NLP tasks and domains. Among the specificities of this architecture, its capacity to adapt to out-of-vocabulary (OOV) terms is a crucial element of its success.

The pre-trained models rely on a Unigram tokenizer algorithm (e.g., SentencePiece or WordPiece) (Kudo and Richardson, 2018) or on the Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2015), which allows the splitting of a word into multiple sub-words. The main idea of dividing words into sub-words is to reduce the vocabulary size by computing frequent sub-units in OOVs. Unfortunately, these tokenization algorithms are purely statistical and lead to semantic information loss when dealing with domain-specific terms. Bostrom and Durrett (2020) showed that BPE is sub-optimal for language models pre-training, as it did not align well with morphology compared to Unigram for English and Japanese. It suggests that sub-units do not contain semantic or syntactic information and that, consequently, the words are poorly represented in the embedding space. In this study, we study the impact of tokenization on the representation of OOV terms.

On top of that, the models have to address user-generated noisy text (e.g., content from social media or e-mails). Typos (e.g., character insertion), smileys, and abbreviations are the most common noise. According to Park et al. (2016), OOVs can be categorized into multiple categories when working on social media (e.g., foreign words, spelling errors, internet slang). Using their typology as inspiration, we will evaluate the robustness of Transformer models for three types of OOVs:

- new domain-specific terms (e.g., "protozoon" and "eucaryote" in microbiology);
- misspelled words containing typos (e.g., "infractus" instead of "infarctus" in medicine);
- cross-domain homographs (i.e., words that are spelled alike but have different meanings) of words existing in the general language (e.g., "arm", either an anatomical part in the general language or a sub-part from a cohort of patients in clinical trials).

By examining these issues, we move towards a deep understanding of the robustness of Transformer models regarding OOVs processing. In particular, we make the following main contributions:

1. We define a statistical measure to quantify how much the tokenization process impacts the position of OOVs in the embedding space. We show that our measure leverages the semantic information shared between clusters of words to evaluate if the OOVs are segmented into coherent units.

2. We evaluate several methods in the literature to help process the semantics of OOVs (i.e., fine-tuning the language models, adding morpho-syntactic information before the encoding, and concatenating the output with external representation). We demonstrate that adding morpho-syntactic context improves the representations for the three categories of OOVs studied in this paper. We conclude that to better process OOVs with Transformer models, adding structural information is more effective than adding semantic information into the embeddings (e.g., with fine-tuning).

3. We analyze the specificities of three types of OOVs (i.e., domain-specific terms, misspelled words, and homographs). We show that if the representation of new OOVs can be improved using various methods, modifying the representation of existing words (homographs) remains challenging. Fine-tuning the models with specific hyper-parameters combined with data augmentation could help resolve this issue.

2. Related Works

2.1. Tokenization into sub-words

Tokenizing texts into sub-tokens / sub-words has received much attention in recent years with the rise of NLP models based on deep learning to generate language models (Pires et al., 2019). For instance, pharmacological entity detection is complex because drugs often recombine independently existing words. Each sub-unit has a singular semantic (e.g., "hydroxychloroquine" may be tokenized into "hydr", "oxy", "chloro", "quine"). Finding the rare recombination in a large corpus can be difficult with a regular tokenizer. However, using the correct sub-words may enhance the processing of the semantics of the words. Gage (1994) and Sennrich et al. (2015) introduced Byte-Pair Encoding (BPE), which relies on a pre-tokenizer that splits the training data into a set of unique words and their frequency. Then, BPE creates a vocabulary consisting of all symbols in the training data and represents each word with its corresponding sequence of characters, plus a final symbol representing the end of the word. BPE then counts the frequency of each possible symbol pair and picks the symbol pair that occurs most frequently. Each merge operation produces a new symbol, which represents a character n-gram. The most frequent character n-grams – which may correspond to whole words – are merged into one symbol. The vocabulary size of the model (i.e., the symbol vocabulary + the number of merges) is a hyperparameter to choose before the training. Unigram is a sub-word tokenization algorithm introduced by Kudo and Richardson (2018) which initializes its vocabulary to a large number of symbols. Next, the algorithm iteratively trims down every symbol to reduce its size. For instance, the initial vocabulary could correspond to all pre-tokenized words and their most frequent sub-words. The unigram tokenization method is not used directly by the models but exists in conjunction with SentencePiece. The specificity of SentencePiece is that it includes the space in the symbol vocabulary before using Unigram to construct the vocabulary. Doing so allows languages that do not use space as a word delimiter to build a suitable vocabulary.

2.2. Impact of Noise on Tokenization

Sun et al. (2020) studied the robustness of BERT towards adversarial inputs containing keyboard typos and demonstrated that BERT has unbalanced attention towards the typos. They also showed that BERT is not robust towards noise on question answering and sentiment analysis tasks. Bagla et al. (2021) further demonstrated that BERT's performance on fundamental NLP tasks like sentiment analysis and textual similarity drops significantly in the presence of simulated noise (spelling mistakes, typos).

3. Transformer Models

In this paper, we will compare two French Transformer models: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) which are trained on large French corpora. We use the standard version (i.e., "Base") for both models containing 12 layers, 12 bidirectional self-attention heads, and 768 hidden units. We provide details about both models in Table 1. We compare the native models with the following variations, aiming to correct OOVs semantic processing:

Fine-tuned CamemBERT/FlauBERT We first fine-tune the language models on the three datasets separately by pursuing the training of the models on new texts. We perform vanilla fine-tuning to compare the models without hyper-parametrizing the models too much.

Concatenation with ELMo We combine a pre-trained contextual embeddings model (Peters et al., 2018) with Transformers to add more context into the embeddings, as suggested by Polatbilek (2020). We use the French pre-trained model provided by Che et al. (2018), with embeddings of dimension 512. We concatenate the representations of ELMo and Transformer models (i.e., CamemBERT or FlauBERT) for each word, obtaining a dimension of 1230.

CamemBERT/FlauBERT-POS (Benamar et al., 2021) Like the original setup, we added morpho-syntactic features into the models, as presented in Figure 1. *Part-of-speech* (POS) tags are used to add morpho-syntactic context into the representations. The authors believed that structural context was essential to improve the semantic processing of OOVs.

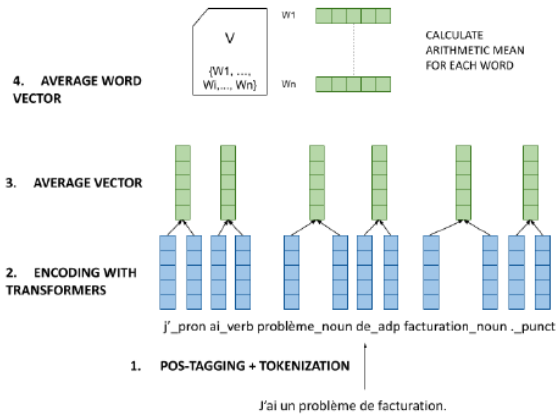


Figure 1: Encoding example with BERT-POS for the sentence *J'ai un problème de facturation*, which could be translated as "I have a billing problem" and tagged as " I_prop have_verb a_det billing_noun problem_noun ._punct"

Model	Tokenizer	*Vocab. Size
CamemBERT-Base	SentencePiece	30 522
FlauBERT-Base	BPE	30 145

Table 1: Transformer Models' description ; *Number of tokens in the vocabulary

4. Datasets

In this section, we present the three datasets used for our experiments[1]. Our objective is to analyze the robustness of the methods to enhance the processing of OOVs in various domains and compare the results with a qualitative evaluation. To provide an in-depth analysis, we analyze the results in three domains: medical, legal, and energy. We provide a descriptive analysis of the datasets in Table 2. We present the distribution of POS-tags in the datasets in Figure 6 and the number of tokens obtained with the pre-trained models (i.e., detailed in Table 1) in Table 6.

Dataset	Domain	#Docs.	#Sents.
Med-Gallica	Medical	942	912 209
DEFT-Laws	Legal	363 721	364 498
EDF-Emails	Energy	79 916	250 923

Table 2: Datasets' description

Med-Gallica The dataset has been collected on the French digital library GALLICA. GALLICA contains a large amount of digitized historical documents such as books or press articles and offers to download the documents after OCR. We selected French documents with plain text access for this experiment and dated between 1887 and 1900. Each document contains metadata, including publication date, author(s), title, and other categories. The dataset consists of medical news extracted from the French medical journal *"Journal de Microbiologie"*. We segmented the long documents into sentences before cleaning them from the noise (whole capital sentences, encoding errors, header removal). All the cleaning has been handmade. The OCR errors created many misspellings in the dataset, which generated OOVs. We will study the impact of OCR errors on the representation of the OOVs.

DEFT-Laws The second dataset consists of concatenated laws extracted from the DEFT'06 training corpus (Azé et al., 2006). After the original extraction of the texts, we observed many misspelled words. The most frequent error is that many accents have disappeared from the original documents.

EDF-Emails The Electricité de France (EDF) emails corpus is a private and anonymized corpus of French customer emails collected between October 2018 and October 2019. We applied the same cleaning techniques as the corpora extracted from Gallica. There are orthographic and syntactic errors (e.g., wrong word order and conjugation errors), smileys, abbreviations, and Energy-specific terms.

5. Evaluation Metrics for Sub-Units

In this paper, we aim to measure the similarity between two words based on their segmentation into sub-words. We propose the two metrics detailed below: *Dice coefficient* adapted to sub-units and *Dice-SU coefficient*. We use the notations detailed in Table 3 to define the metrics.

Notation	Description
X, Y	$X = x_1, \ldots, x_k$ and $Y = y_1, \ldots, y_l$ are strings of length k and l, respectively, composed of symbols of a finite alphabet.
n_X, n_Y	number of n-grams in X and Y.
n_Z	number of common n-grams between X and Y
$t_M(X), t_M(Y)$	tokenization function for X and Y, using the Transformer model M.
$n_{t_M(X)}$,$n_{t_M(Y)}$	total number of sub-words obtained after the tokenization of X and Y.
$n_{t_M(Z)}$	total number of common sub-words between X and Y.
$\|t_M(X)_i\|$,$\|t_M(Y)_i\|$	total number of characters in the i^{th} sub-word of X and Y.
$\|t_M(Z)_i\|$	total number of characters in the i^{th} common sub-word between X and Y.

Table 3: Definitions and Notations

Dice coefficient *Dice coefficient* is a popular word similarity measure to calculate the ratio of the num-

[1]The datasets DEFT-LAWS and MED-GALLICA are provided at https://github.com/alexandrabenamar/evaluating_tokenizers_oov

ber of n-grams shared between two strings and the total number of n-grams in both strings:

$$\text{Dice(X,Y)} = 2 \times \frac{n_Z}{n_X + n_Y} \quad (1)$$

We adapt the *Dice coefficient* by replacing the n-grams with the sub-words (generated during tokenization):

$$\text{Dice(X,Y)} = 2 \times \frac{n_{t_M(Z)}}{n_{t_M(X)} + n_{t_M(Y)}} \quad (2)$$

Dice for Sub-Units (Dice-SU) coefficient *Dice coefficient* is very effective to compare the sub-units between two words. However, generating small sub-words during tokenization can be suboptimal, as some of the sub-tokens might be too small to retain semantic information. For instance, if we measure the *Dice coefficient* between the words *snakes*, *snake* and *cars*, tokenized respectively into "snake+s", "snake" and "car+s", we obtain close results between $Dice(snakes, snake) = 0.67$ and $Dice(snakes, cars) = 0.5$, due to plural markers rather than semantic information. We propose the *Dice for Sub-Units (Dice-SU) coefficient*, a variant of the *Dice coefficient*, which penalizes the small sub-units during tokenization. The higher the value, the bigger the sub-units shared between two words. We propose the following formal definition:

$$\text{Dice-SU}(X,Y) = \frac{2 \times \sum_{i=0}^{n_{t_M(Z)}} |t_M(Z)_i|}{\sum_{i=0}^{n_{t_M(X)}} |t_M(X)_i| + \sum_{i=0}^{n_{t_M(Y)}} |t_M(Y)_i|} \quad (3)$$

Back to the example, we obtain 0.91 of similarity with *DICE-SU*$(snakes, snake)$ and only 0.2 for *DICE-SU*$(snakes, cars)$. We manage to better estimate the shared level of semantic between two words.

6. Experiments

In this paper, we use the partition of OOVs in three categories, as detailed in Section 1.

6.1. Domain-specific OOVs classified into new words

We aim to evaluate the representations of new domain-specific terms generated by CamemBERT and FlauBERT. We conduct this experiment on two domains: legal and medical. We do not use EDF-Emails in that section since these types of OOVs are not frequent in the corpus. We selected ten frequent domain-specific OOVs in DEFT-Laws[2] and Med-Gallica[3]. We computed the five closest associates of each OOV using cosine similarity. Next, we tokenized the OOVs and their associates using a Transformer model (CamemBERT or FlauBERT). We measured and averaged the *Dice coefficient* and the *Dice-SU coefficient* between the OOVs and their associates. The complete procedure is detailed in Algorithm 1. We present the results obtained on DEFT-Laws and Med-Gallica in Figure 2. We observe the distributions of *Dice coefficients* and *Dice-SU coefficients* obtained for the OOVs with each model. A detailed example of the results obtained for the word "discriminatoires" on DEFT-Laws is presented in Table 8.

DEFT-Laws We observe that the *Dice coefficient* is on average higher with CamemBERT than with FlauBERT by 20%. Consequently, the OOVs representations are more affected by the tokenization with CamemBERT than FlauBERT. Moreover, adding morpho-syntactic information drastically reduces the tokenizer's impact on the OOVs representation. FlauBERT-POS is averaging 0% of similarity between the sub-units, meaning that the sub-units of the associates of OOVs are utterly different from the sub-units of the OOVs. Finally, *Dice-SU coefficient* is globally inferior to 0.5%, meaning that the neighbors share small sub-units with the OOVs. Consequently, we assume that the shared sub-units do not hold much semantic information. We note that the fine-tuning did not change the representation of the OOVs. However, adding contextual information (with ELMo or POS) reduced the impact of the tokenizer, especially when adding morpho-syntactic context.

Med-Gallica We obtained similar tendencies on Med-Gallica, except that there are more shared sub-units between the OOVs and their associates. The *Dice-SU coefficient* is higher, with the averaged similarity score comprised between 50% and 70%. In the medical domain, the coverage between the sub-units of OOVs and the sub-units of their closest associates is high, meaning that even though the tokenizer has a significant impact on the representation of OOVs, the selected sub-units are probably full of relevant semantic information.

6.2. Misspelled words

Misspelled words are challenging to detect with models using sub-words such as FastText (Edizel et al., 2019) or BERT (Nayak et al., 2020; Sun et al., 2020). At word-level, such errors may result from incorrect insertion, deletion, or substitution of a character or the transposition of two adjacent characters. At sentence-level, they can correspond to syntactic errors. We conduct this study on three specific types of misspellings, one for each dataset:

- DEFT-Laws: we analyze misspellings on words that exist in the vocabularies of CamemBERT and FlauBERT (e.g., "xonditions" instead of "conditions").

[2] *allegation, fraudulent, minutes of proceedings, deliberate, rule, appearance, discriminatory, regularized, transferee, national, regularized, registered, sealed, and affixed*

[3] *incubation, bacteriological, epileptic, prophylactic, tuberculosis, cauterization, bacillophage, sepsis, hyperesthesia, and anorexia*

Algorithm 1 Unsupervised Evaluation of Tokenizer's Impact in the Embedding Space of OOVs

Input: A domain-specific word w; a dataset containing a vocabulary V of size N; a matrix of embeddings X of shape (N,length of embeddings); a Transformer model M; a similarity measure $sim(t_M(W_1), t_M(W_2))$ between the words W_1 and W_2; a number n of closest associates to evaluate.

Step 1. Compute the n closest associates of w.

$p \leftarrow$ position of w in V
$c_p \leftarrow$ cosine similarity between X_p and X ▷ Similarity between w and the rest of the vocabulary
$a_i \leftarrow \arg\max_{c_p}, i = 1, \ldots, n$ ▷ Get the indices of the top n closest associates
$Y \leftarrow V[a_i]$

Step 2. Tokenize w and its Y associates, and compute the similarity.

$S_n = []$
for y_i in Y **do**
    $S_i \leftarrow sim(t_M(w), t_M(y_i))$ ▷ Similarity between the sub-units of w and y_i
end for
$s \leftarrow mean(S_n)$ ▷ Averaged similarity for the top n associates

Output: s

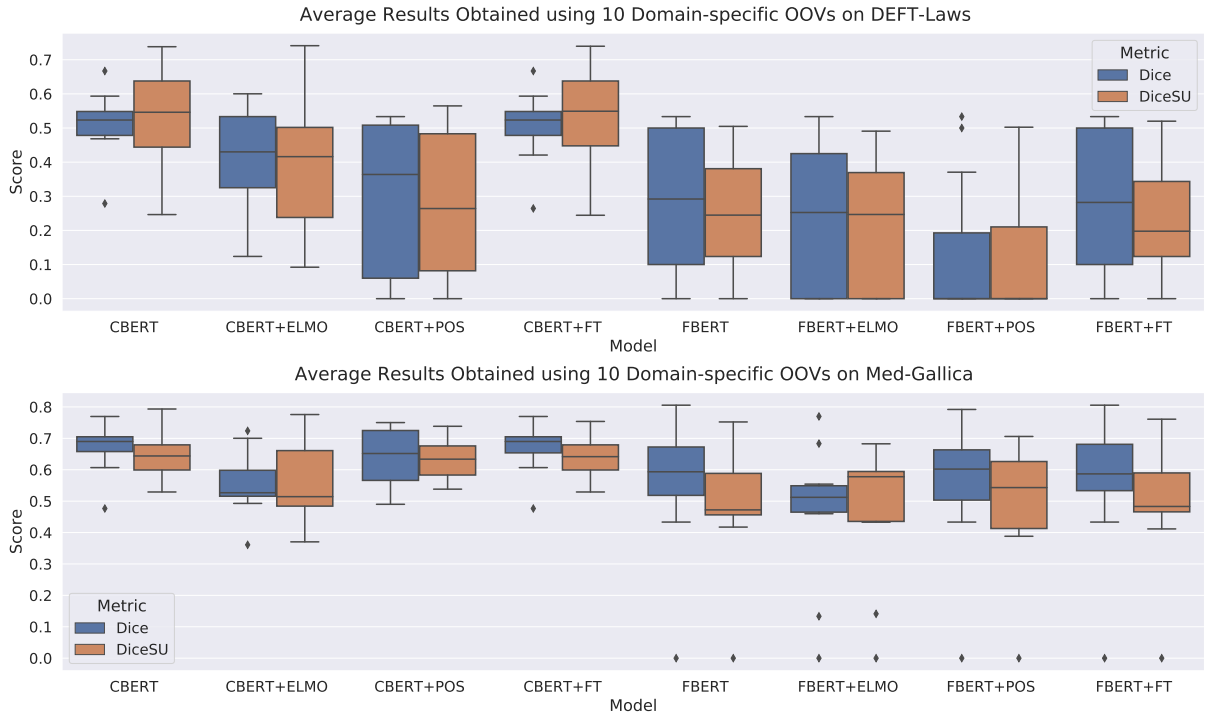
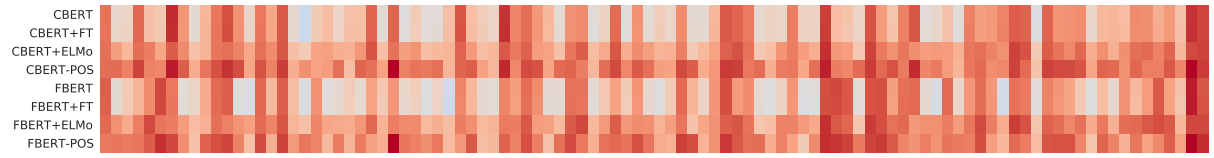


Figure 2: Distribution of *Dice coefficient* and *Dice-SU coefficient* over 10 domain-specific OOVs on DEFT-Laws (top) and Med-Gallica (bottom). We computed the averaged coefficients between each OOV and their five closest neighbors (using cosine similarity). The boxplots contain the ten scores computed (one for each OOV) with a model (e.g., CamemBERT, FlauBERT, etc.) and a metric (Dice or Dice-SU)

- Med-Gallica: we focus on misspellings of domain-specific words not existing in the models' vocabulary (e.g., "injoction" instead of "injection"). OCR errors cause all the misspellings in this dataset.
- EDF-Emails: we evaluate the processing of misspellings on words that are specific to the structure of emails and appear at similar positions in the texts (e.g., "cordiallement" instead of "cordialement", meaning *cordially*).

We randomly extracted 100 misspelled words from each corpus to compare the handling of misspellings by the models. Next, we associated the misspelled words with their correct version. Finally, we computed the cosine similarity between the pairs $\{word_{correct}, word_{misspelled}\}$. The results are presented using heatmaps, where each box represents the cosine similarity between a pair $\{word_{correct}, word_{misspelled}\}$ (i.e., correctly written and its misspelled variation). We present the results obtained on DEFT-Laws, Med-Gallica and EDF-Emails on Figures 3a, 3b and 3c respectively. Table 4 contains the average similarity obtained for the 100 OOVs.

(a) DEFT-Laws



(b) Med-Gallica

(c) EDF-Emails

Figure 3: Cosine similarity between 100 random misspelled general-domain words and their correct associate. The x-axis contains the pairs $word_{correct}, word_{misspelled}$. The redder the box, the higher the similarity between the pair

	Law	Medical	Emails
CBERT	0.19	0.39	0.27
+ELMo	0.32	0.54	0.44
+POS	**0.63**	**0.66**	**0.92**
FBERT	0.15	0.37	0.34
+ELMo	0.34	0.57	0.56
+POS	0.56	0.63	**0.93**

Table 4: Average cosine similarity results between the 100 random selected misspelled words and their correct associates on all datasets

DEFT-Laws CamemBERT performed slightly better by than FlauBERT with 3% of difference between the cosine similarity. While it could mean that the tokenization based on SentencePiece is slightly more effective than BPE in constructing semantic sub-words for misspelled words, the difference between the results of both methods remains too thin to validate the hypothesis. Surprisingly, fine-tuning the language model with part of the data did not change the results for the selected misspelled words. Consequently, the misspelled words were not selected during fine-tuning to add to the vocabularies, neither with BPE nor SentencePiece. Concatenating the output of Transformers with ELMo improved the models, even though these words are probably OOVs in the ELMo model since ELMo was trained on generic data. However, ELMo appears to regroup words with their misspelled versions, using the context of other words in the sentence. The concatenated representation increases the similarity between the misspelled words and their correct version by 19% of similarity with FlauBERT and 13% with CamemBERT. Moreover, adding morpho-syntactic information into the embeddings improves the results even more, with 44% of cosine similarity obtained with CamemBERT and 41% on FlauBERT. We conclude that to process OOVs more efficiently, the models need to better understand the surrounding context without incorporating the OOVs into their vocabularies.

Med-Gallica We observe similar results on the misspelled domain-specific words in the Med-Gallica dataset, corresponding to OCR errors. This task is more complex than the previous one because first, the models must process the domain-specific terms in the first place. Second, they must process the misspelled versions of the words and understand their proximity. Nevertheless, the models have a better semantic understanding of these misspelled words than those in DEFT-Laws. The results are higher than on misspelled words of general-domain, obtaining 39% and 37% of similarity, respectively, with CamemBERT and FlauBERT. Even though we cannot compare the results obtained on DEFT-Laws and Med-Gallica directly since the context of the words are different and so is the type of the OOVs, it is interesting to note the difference between the OOVs. In the medical domain, the Transformers performed poorly, averaging 38% of similarity between the models, but still captured some semantic proximity for the misspelled words compared to the results in the legal domain.

EDF-Dataset On EDF-Dataset, we analyze the misspellings on email-specific words. This study is different from the others because the context of the words surrounding the target will change drastically, but the words' positions in the sentences are similar. In Table 3, we observe that the models perform similarly

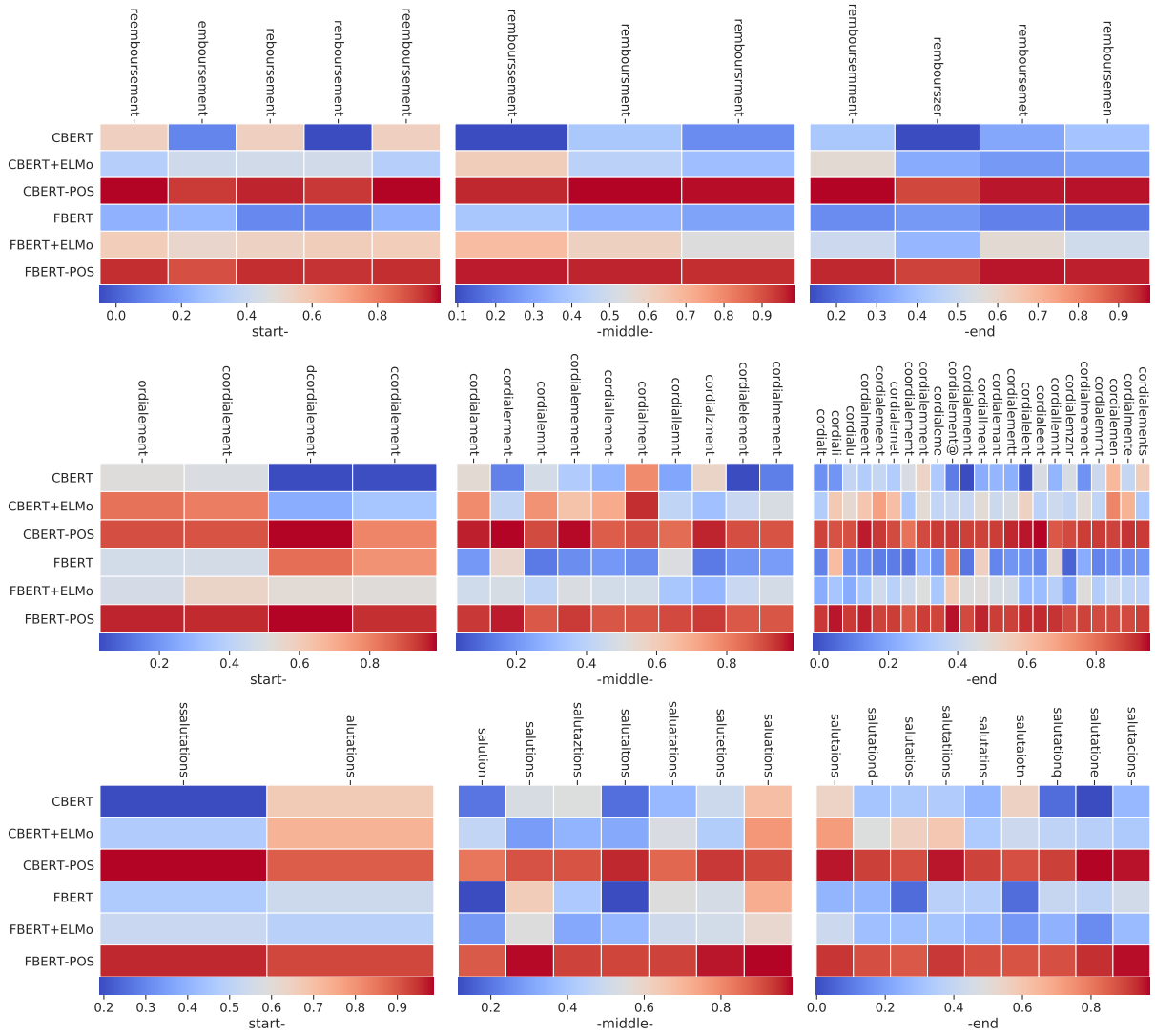


Figure 4: EDF-Emails dataset - Cosine similarity between the words "remboursement" (i.e., *refund*) on top, "cordialement" (i.e., *cordially*) in the middle and "salutations" (i.e., *greetings*) at the bottom and their spelling errors' variants. We distinguish between errors at the start, middle and end of the words (from left to right).

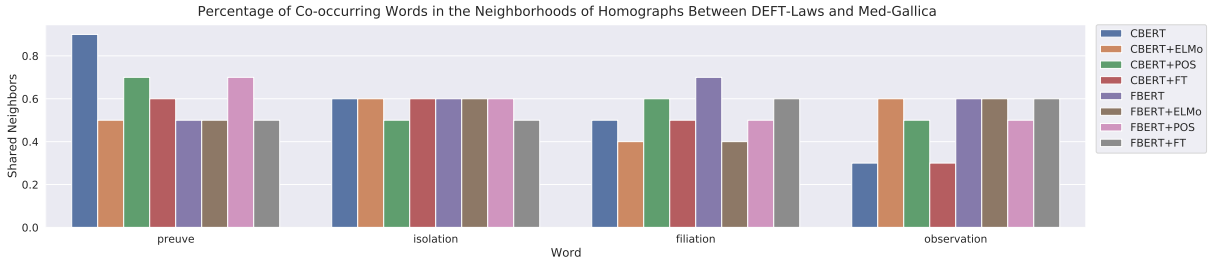


Figure 5: Percentage of co-occurring words in the neighborhood of homographs between DEFT-Laws and Med-Gallica. We compared the n=10 nearest neighbors of each word in both datasets

to other types of misspellings, except that the performances after adding morpho-syntactic context OOVs are even better. Indeed, we obtain 92% of cosine similarity between the pairs $\{word_{correct}, word_{misspelled}\}$ with CamemBERT-POS and 93% with FlauBERT-POS. It represents 65% and 59% of the performance gap, respectively, with CamemBERT and FlauBERT. It is easy to explain why the method performs even better on this task. Indeed, the addition of POS makes it possible to add syntactic context without adding any semantic information. We expect to bring together the selected words thanks to their position in the texts, so this method is effective. Nayak et al. (2020) hypothesized that word-beginning spelling errors were more severe than others because they affected tokenization the most. To test this hypothesis, we use EDF-

Emails because it contains users' misspellings instead of external tools-generated mistakes. We selected three words for their high frequency in the corpora and their high rate of misspelling errors: "cordialement" (*cordially*), "remboursement" (*refund*) and "salutations" (i.e., *greetings*). The results are shown in Figure 4, where misspelled words are separated regarding the position of the error in the word (i.e., start, middle, and end). We cannot observe significant differences between the misspelled errors for these words. However, we state that not all errors of the same type are equivalent. For instance, the word "cordialement" (tokenized into *_cordialement* with CamemBERT) has two similar misspellings: "dcordialement" and "ccordialement" (tokenized into *_c,cord,iale,ment* with CamemBERT), but "dcordialement" (tokenized into *_d,cord,iale,ment* with CamemBERT) is the closest to the original word with all the methods. It is interesting because it means that misspellings can generate different sub-words and that, even with the same tokenization and depending on the characters, the misspelled words cannot be processed similarly. This result is problematic because it cannot be predicted when processing the sub-words.

6.3. Cross-domain Homographs

In this section, we aim to determine if terms specific to a domain are treated contextually by the Transformer models. More precisely, we seek to evaluate the impact of the domain on the processing of homographs specific to this domain. We recover four words with different meanings when used in legal or medical contexts. Table 7 contains the definitions[4] of these terms in both contexts.

To evaluate if the semantics of the words are specific to the specialty domain of the corpora, we recover for each word and each domain the ten closest neighbors, using cosine similarity. Next, the objective is to compare, for each word, its neighborhood in the legal field and the medical field. We compute the co-occurring neighbors between domains using the formula:

$$similarity = \frac{number\ of\ co\text{-}occurences}{n} \quad (4)$$

where n is the number of neighbors we use in the experiments. In these experiments, we use n=10. We motivate our approach by stating that the neighbors of a word, when using embeddings as representation, symbolize the semantic understanding captured by the model. Therefore, the synonyms are supposed to be close in the embedding space. The other words that appear in the same context should also be close in the embedding space. In our experiments, the homographs appear in different contexts with different vocabularies. Consequently, they should not share many neighbors.

Figure 5 presents the similarity we obtain for the four selected words with all the models. We provide more detailed results in Table 9 for the word "preuve". We observe that the words are similar in both datasets as we obtain at least 50% similarity between their neighbors for most of the methods. For instance, we observe a 90% similarity for the word "preuve" with CamemBERT. Looking at the neighbors obtained for "preuve", we observe that apart from context-specific words in law (e.g., *testimonials, to convince, counterparty*), we obtain non-specific terms in the neighborhood for both datasets. It may be because the training data for CamemBERT and FlauBERT is generic. However, fine-tuning the language models with DEFT-Laws and Med-Gallica did not improve the results. It is unexpected, considering that the context should have helped the models process a better semantic of the words. We hypothesize that while fine-tuning can improve the representation of new words, such as specific terms or misspelled words, it is not enough to change the representation of words existing in the initial vocabulary. In this experiment, we used a vanilla baseline to provide a general analysis of the models. Naturally, these results could be improved by specific tuning of the hyperparameters. Then, we observe that neither the training data of the models nor the tokenizers appear to impact the results since CamemBERT and FlauBERT generate similar results. Finally, we show that no model has continuously distinguished the semantics of the words between the domains.

7. Conclusion and Discussion

In this paper, we aimed to evaluate the impact of the tokenizers on the representation of OOVs.

We proposed a *Dice-SU*, a new metric for evaluating the tokenizer's impact on OOV representation. The method is based on the shared sub-units between an OOV and its neighbors in the embedding space. Moreover, it incorporates the notion of semantics inside the sub-units and penalizes the small sub-units. We demonstrated that we could not predict the representation of misspelled words since the models' tokenizers are based on a statistic segmentation rather than a linguistic one. It constitutes a significant issue considering that similar mistakes cannot be resolved the same way. Moreover, we showed that it is easier to improve the representation of new OOVs than for OOVs, which exist in the general domain (homographs). Finally, we demonstrated that adding information about the structure of sentences (i.e., POS tags) is far more effective than learning new words (fine-tuning). In this paper, we could not demonstrate the efficiency of fine-tuning to improve the representation of OOVs. In future work, we would like to apply our findings to a French e-mails corpus to determine if adding morpho-syntactic information improves the results of an automatic classification task. Moreover, we will study the improvement in downstream NLP tasks (e.g., sentiment analysis) using the improved embeddings for OOVs.

[4]Most of the definitions were taken from the TLFi dictionary: `https://www.cnrtl.fr/definition/`

8. Bibliographical References

Azé, J., Heitz, T., Mela, A., Mezaour, A., Peinl, P., and Roche, M. (2006). Présentation de deft'06 (defi fouille de textes). *Proceedings of DEFT*, 6(1):3–12.

Bagla, K., Kumar, A., Gupta, S., and Gupta, A. (2021). Noisy text data: Achilles' heel of popular transformer based NLP models. *CoRR*, abs/2110.03353.

Benamar, A., Bothua, M., Grouin, C., and Vilnat, A. (2021). Easy-to-use combination of pos and bert model for domain-specific and misspelled terms. In *NL4IA Workshop Proceedings*.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Che, W., Liu, Y., Wang, Y., Zheng, B., and Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

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

Edizel, B., Piktus, A., Bojanowski, P., Ferreira, R., Grave, E., and Silvestri, F. (2019). Misspelling oblivious word embeddings. *arXiv preprint arXiv:1905.09755*.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco et al., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: des modèles de langue contextualisés pré-entraînés pour le français. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, pages 268–278. ATALA; AFCP.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Nayak, A., Timmapathini, H., Ponnalagu, K., and Gopalan Venkoparao, V. (2020). 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*, pages 1–5, Online, November. Association for Computational Linguistics.

Park, S., Fazly, A., Lee, A., Seibel, B., Zi, W., and Cook, P. (2016). Classifying out-of-vocabulary terms in a domain-specific social media corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2971–2975.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Polatbilek, O. (2020). *Enriching Contextual Word Embeddings with Character Information*. Ph.D. thesis, Izmir Institute of Technology (Turkey).

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., and Xiong, C. (2020). Adv-BERT: BERT is not robust on misspellings! generating nature adversarial samples on BERT. *arXiv preprint arXiv:2003.04985*.

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.

Yin, W. and Schütze, H. (2018). Attentive convolution: Equipping CNNs with RNN-style attention mechanisms. *Transactions of the Association for Computational Linguistics*, 6:687–702.

Appendix

We provide a glossary of translations for the tables in the Appendix :

Word	Translation
arbitraire	arbitrary
contradictoire	contradictory
contraignant	binding
discriminatoire	discriminatory
exhaustives	détaillée
inacceptables	unacceptable
inefficaces	ineffective
injustifiés	unjustified
restrictives	restrictive
néfastes	harmful
monotone	monotonous
souhaitables	desirable
unilatérales	unilateral
preuve	evidence
convaincre	to convince
particularité	particularity
certitude	certainty
vérité	truth
témoignages	testimony
enregistrement	registration
vestiges	remnants

Table 5: Translation of French words used in the Appendix

Dataset	CamemBERT	FlauBERT
Med-Gallica	21 156	35 517
DEFT-Laws	16 742	23 972
EDF-Emails	18 991	26 352

Table 6: Number of tokens for each dataset using pretrained models

We provide examples of the neighbors for the word ”discriminatoires” :

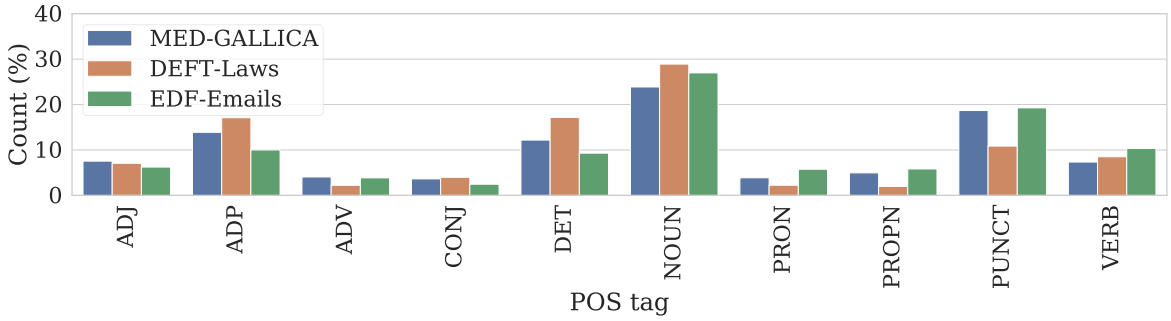


Figure 6: Distribution of POS in the datasets

Word	Legal	Medical
Preuve (*Proof or Evidence*)	Material element (e.g., contractual document, certificate) which demonstrates, indicates, proves the truth or the reality of a situation of fact or law: proof of a crime.	Scientific evidence is evidence used to support or disprove a theory or hypothesis in science.
Filiation (*Filiation*)	Legal relationship between parents and their children.	The continuity of the different forms of life, resulting from each other.
Observation (*Observation or Monitoring*)	The act of complying with a rule, a law, a regulation.	The scientific process of investigation consists of the careful examination of a fact, of a process, intending to know it better, understanding it, and excluding any action on the phenomena studied.
Isolement	Separation of an individual – or a group of individuals – from other members of society.	Bacteria and virus culture technique allowing them to be separated within a contaminated product.

Table 7: Definitions of homographs in a legal context and a medical context

Model	Top candidates	Dice	Dice-SU
C-BERT	discriminatoire, discrimination, restrictives, injustifiées, inacceptables	0.33	0.22
C-BERT + ELMo	discriminatoire, souhaitables, restrictives, exhaustives, contraignants	0.53	0.25
C-BERT + POS	discriminatoire, restrictives, injustifiés, arbitraires, unilatérales	0.53	0.25
F-BERT	discriminatoire, non-discriminatoires, discriminations, non-discriminatoire, discrimination	0.13	0.18
F-BERT + ELMo	discriminatoire, non-discriminatoires, restrictives, non-discriminatoire, inefficaces	0.13	0.18
F-BERT + POS	discriminatoire, contradictoires, contraignantes, néfastes, monotone	0.0	0.0

Table 8: DEFT-Laws - Top 5 candidates for the word ”discriminatoires” (i.e., plural of *discriminatory*) using cosine similarity. We averaged the *Dice coefficient* and *Dice-SU coefficient* obtained between ”discriminatoire” and each of its five candidates

Model	Laws - Top candidates	Medical - Top candidates
C-BERT	preuves, confirmation, convaincre, prouve, prouvent, prouvé, particularité, démonstration, prouvant, certitude	preuves, preuveên, confirmation, convaincre, prouver, prouvent, prouvé, particularité, prouvant, démonstration
+ELMo	preuves, certitude, conclusion, confirmation, justification, promesse, différence, contrepartie, démonstration, possibilité	preuves, preuveên, certitude, conclusion, démonstration, vérité, chance, trace, justification, particularité
+POS	preuves, justification, mécanismes, résultant, préservation, arguments, démonstration, corrélation, témoignages, certitude	preuves, mécanismes, justification, enregistrement, arguments, démonstration, vestiges, corrélation, témoignages, renforce
+FT	preuves, confirmation, prouvent, prouvé, convaincre, certitude, démontrer, démonstration, prouver, prouvant	preuves, preuveên, convaincre, prouvé, prouvent, prouvant, confirmation, témoignent, particularité, prouve

Table 9: Example of associates for the word *"preuve"* on DEFT-Laws (left) and Med-Gallica (right) with four models: CamemBERT, CamemBERT+ELMo, CamemBERT-POS and CamemBERT+Fine-tuning