Evaluating Tokenizers Impact on OOVs Representation with Transformers Models

Alexandra Benamar, Cyril Grouin, Meryl Bothua and Anne Vilnat
Université Paris-Saclay, CNRS, LISN, Orsay, France
EDF Lab R&D, Palaiseau, France

1. Problems

1. Pre-trained models (i.e., BERT) have a restrained vocabulary and need to be adapted when working with out-of-vocabulary (OOV).

2. The tokenization of OOVs is purely statistical in pre-trained Transformer models and has a major impact in its representation.

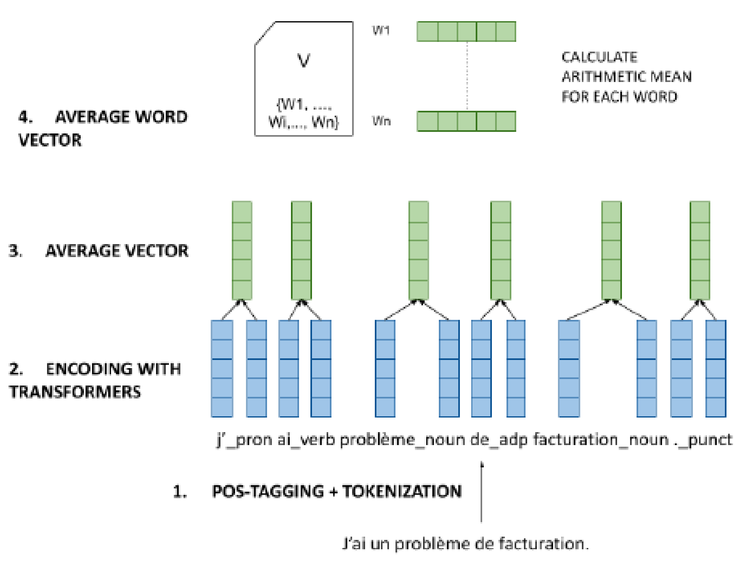
3. The behavior of Transformer models varies regarding the type of OOV to process: **misspelled words** containing typos, **cross-domain homographs** (e.g., "arm" has different meanings in a clinical trial and anatomy), and **new domain-specific terms** (e.g., "eucaryote" in microbiology).

4. Lack of evaluation metrics to compare the semantic of OOVs processed by the models.

2. Our goals

1. To provide a **new evaluation metric** for OOVs processing (i.e. Dice-SU)
2. To **evaluate** quantitatively the performances of Transformer models regarding the **specificities of OOVs**. => We compare the use of vanilla Transformer models with 3 methods to improve the semantics of OOVs: **BERT-POS, adding ELMo representations, and fine-tuning the language models**.

3. BERT-POS



4. Evaluation Metric

- Dice

$$2 \times \frac{n_{t_M(Z)}}{n_{t_M(X)} + n_{t_M(Y)}}$$

- Dice for Sub-Units (Dice-SU)

$$2 \times \frac{\sum_{i=0}^{n_{t_M(Z)}} |t_M(Z)_i|}{\sum_{i=0}^{n_{t_M(X)}} |t_M(X)_i| + \sum_{i=0}^{n_{t_M(Y)}} |t_M(Y)_i|}$$

		cats ("cat"+"s")	snake ("snake")
snakes ("snake"+"s")	Dice	0.50	0.67
	Dice-SU	**0.20**	**0.91**

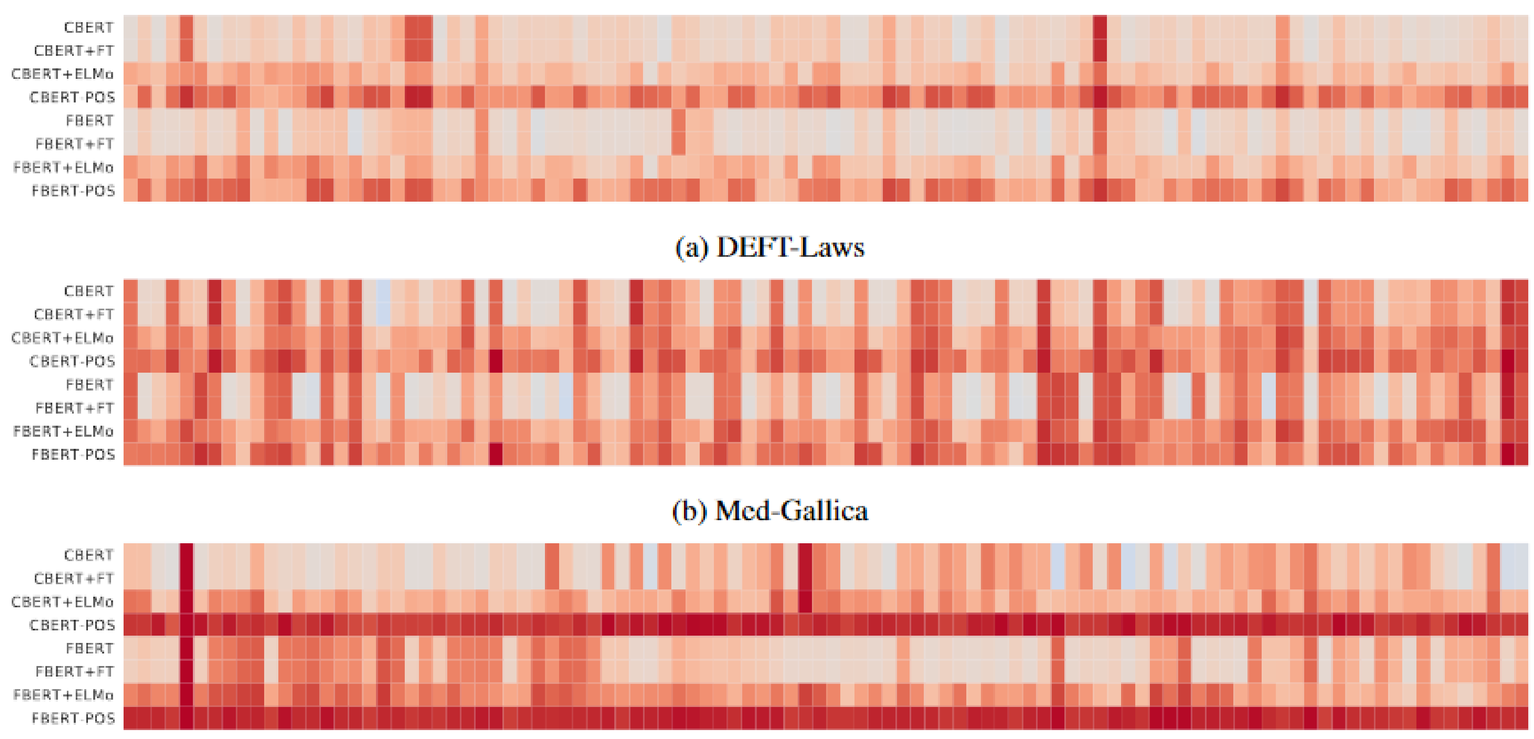
Table 1: Similarity between "snakes" and {"cats", "snake"}

5. French Datasets

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

6. Misspelled Words



(a) DEFT-Laws

(b) Med-Gallica

(c) EDF-Emails

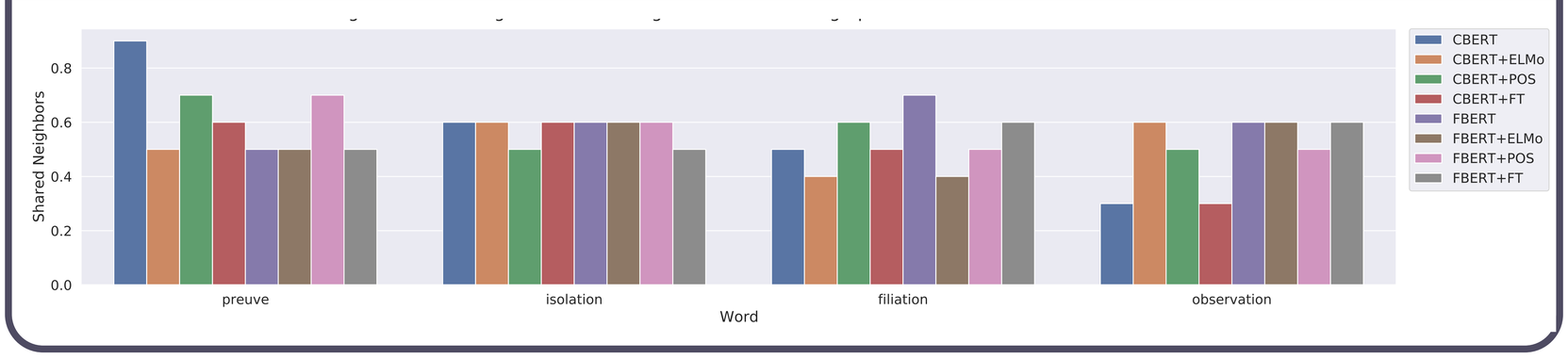
Figure 1: Cosine similarity between 100 random misspelled general-domain words and their correct associate. The x-axis contains the pairs $word_{correct}, word_{misspelled}$.

	Law	Medical	Energy
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 3: Average cosine similarity results between the 100 random selected misspelled words and their correct associates on all datasets

- The performance varies depending on the domain and the type of misspelling.
- **BERT-POS is more efficient** than adding ELMo or fine-tuning the language models for misspelled words;
- We obtain slightly better results with FlauBERT (BPE tokenizer) than CamemBERT (SentencePiece tokenizer).

7. Cross-Domain Homographs



8. Conclusions

- *Dice-SU* is a helpful metric to measure the semantics of OOVs.
- It is easier to improve the representation of new OOVs than OOVs which already exist in the vocabulary.
- Adding information about the structure of sentences is far more effective than fine-tuning.

LREC 2022 Marseille