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
\[ 2 \times \frac{nt_M(Z)}{nt_M(X) + nt_M(Y)} \]
• Dice for Sub-Units (Dice-SU)
\[ 2 \times \frac{\sum_{i=0}^{nt_M(Z)} |t_M(Z)_i|}{\sum_{i=0}^{nt_M(X)} |t_M(X)_i| + \sum_{i=0}^{nt_M(Y)} |t_M(Y)_i|} \]

5. French Datasets
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>#Docs.</th>
<th>#Sents.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Med-Gallica</td>
<td>Medical</td>
<td>942</td>
<td>912 209</td>
</tr>
<tr>
<td>DEFT-Laws</td>
<td>Legal</td>
<td>363 721</td>
<td>364 498</td>
</tr>
<tr>
<td>EDF-Emails</td>
<td>Energy</td>
<td>79 916</td>
<td>250 923</td>
</tr>
</tbody>
</table>

6. Misspelled Words

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.