## General Context & Motivation

**Transfer Learning**

- Main idea: knowledge acquired on a first task can be successfully transferred to other tasks, improving final performance.
- Some important milestones:
  - ULMFiT (Baevski et al., 2019)
  - NeMo (see paper for details)
  - Contextualized Embeddings from a Language Model for feature extraction
  - Bert (Devlin et al., 2019)
  - LM-Transformer (Balle-Tilley, 2021)

**Specialized Domains**

- Technical domains (e.g., medical domain) that often have specific vocabularies, writing styles, etc.
- Require specialized models.
- **BERT in Specialized Domains**
  - Off-the-shelf pre-trained models (often, general domain)
  - Adapted by re-training on specialized corpora (e.g., biomedical literature)
  - Keep the original vocabulary (often, general domain)
  - Possible interface with target domain

**Methodology**

- **Objective:** analyze the impact of using general-domain WordPieces vs. specialized WordPieces.
- **Proposed Approach**
  - Given BERT's original general-domain WordPieces
  - Given a specialized corpus in the target domain
  - Contextualized domain-specific WordPieces
    - Compare adapted BERT's original WordPieces
    - Train BERT from scratch using the specialized WordPieces
    - Compare a model trained on domain-specific WordPieces

## Experimental Setup

### How Does BERT Tokenize?

- BERT's tokenization system (simple + WordPiece tokenization)
  - A single tokenization (e.g., [CLS] [SEP] [MASK] [UNK])
  - 4 WordPiece tokenization (towards De-Off Vocabulary tokens)

### BERT & the WordPieces

- Effect on Downstream Performance

<table>
<thead>
<tr>
<th>General BERT</th>
<th>Medical BERT</th>
<th>How Does BERT Tokenize?</th>
</tr>
</thead>
<tbody>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>General BERT: clinical natural language inference (pair classification)</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
<tr>
<td>General BERT</td>
<td>Medical BERT</td>
<td>How Does BERT Tokenize?</td>
</tr>
</tbody>
</table>

### Experimental Setup

- We use the following corpora for pretraining:
  - General: MNLI-Multilingual data (base-uncased) 
  - Medical: SNIPS + EMNLP (34,300 tokens)

### Results: Number of Splits

<table>
<thead>
<tr>
<th>Device</th>
<th>Corpora</th>
<th></th>
<th>of documents</th>
<th>of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI-Multilingual</td>
<td>15.17 tokens</td>
<td>44.5 tokens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNIPS + EMNLP</td>
<td>14.5 tokens</td>
<td>54.5 tokens</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Results: Quality of Splits

<table>
<thead>
<tr>
<th>Device</th>
<th>Corpora</th>
<th></th>
<th>of documents</th>
<th>of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI-Multilingual</td>
<td>12.17 tokens</td>
<td>41.5 tokens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNIPS + EMNLP</td>
<td>11.5 tokens</td>
<td>51.5 tokens</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Conclusion

**Main Takeaways**

- Using general-domain WordPieces vs. specialized WordPieces.
- Training from scratch using a specialized vocabulary/corpus.
- Overall better than re-training from general BERT.
- May not be worth the additional pre-training cost.
- Limitations: experiments conducted for a single domain & language and missing experiments (e.g., General vs. English).