A BERT’s Eye View: Identification of Irish Multilanguage Expressions Using Pre-Trained Language Models

Abigail Walsh, Teresa Lynn, Jennifer Foster

PARSEME Shared Task 1.2

- (Ramisch et al. 2020)
- Automatic identification of unseen verbal MWEs
- 14 languages
  - First time including Irish (Walsh et al. 2020)
- 9 systems: 7 open track & 2 closed track
  - 4 systems used Pre-trained language models

Background

- Pre-trained language models have seen widespread use in many NLP applications
- Monolingual language models have been shown to give better model performance than multilingual language models for certain tasks
- We compare results obtained using a multilingual language model (mBERT) (Devlin et al. 2019) with an Irish monolingual language model (gBERT) (Barry et al. 2022) for the task of automatic identification of verbal MWEs (vMWEs) in Irish

Challenges for Irish dataset

- Systems performed the most poorly on Irish data
- Many labels used (7 compared to average of 5 across languages)
- High ratio of unseen vMWEs (69% compared to average of 33% across languages)
- Small number of training and tuning examples (226 compared to average of 3645 across languages)
- Irish vMWEs exhibit high degree of variability

Model Instability

- Known issue in Transformer architecture
- Training a model with 10 random seed values selected
- Instead, two random seed values are used
- Known issue in automatic identification of verbal MWEs (Ramisch et al. 2020)

Experiment Series 1

Hyperparameter tuning

- Number of layers
  - Range [0–12]
  - Training on all 12 layers gives better performance
- Number of epochs
  - Range [5–40]
  - Training on all 12 layers gives better performance
  - Exception was models with all 12 layers frozen: required larger learning rate

Random seed

- 20 trials using best performing hyperparameters
- Random seed values selected evenly from 5–100
  - It was found that combining a batch size of 2 and a learning rate of 2e-4 results in a model that does not predict any MWEs
  - Instead, two random seed experiments devised for gBERT models using best learning rate (gBERT-12-rate) and best batch size (gBERT-12-batch)

Experiment Series 2

Addressing dataset challenges

Experiment 1

- Using baseline data
- Comparing across 3 labelling schemes

Experiment 2A

- Addressing label complexity by merging two MWE labels
  - LVC.full and LVC.cause → LVC
  - VPC.full and VPC.sem → VPC

Experiment 2B

- Addressing label complexity by merging all MWE labels
  - Any category → VPC

Labelling Schemes

- Sentence: “He did a lot of study and research”
- Annotating sentence

<table>
<thead>
<tr>
<th>Category</th>
<th>MWE</th>
<th>Freq</th>
<th>mBERT</th>
<th>Freq</th>
<th>gBERT</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>55</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gair</td>
<td>25</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dian</td>
<td>23</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dianumh</td>
<td>16</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bain</td>
<td>12</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eirigh</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amach</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iubhair</td>
<td>8</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10 most frequently labelled words for mBERT-optimised and gBERT-optimised models.

Conclusions

- Results demonstrate that monolingual pre-trained language models can achieve surprisingly good results even on small datasets
- Instability is an issue, particularly with small datasets
- Can be combated through training for more epochs, and careful selection of learning rate
- Addressing dataset challenges shows inconclusive results
- Similarity with alternative labelling schemes
- Possible that effects would be more noticeable using different hyperparameters and larger datasets