JRIRD at the NTCIR-16 FinNum-3 Task: Investigating the Effect of Numerical Representations in Manager’s Claim Detection

Shunsuke Onuma and Kazuma Kadowaki
The Japan Research Institute, Limited
Abstract

- Participate in Manager’s Claim Detection (English subtask) of FinNum-3
  - Claim detection: judges whether a target numeral is in a manager’s claim or not
  - Numerical category classification: classifies a target numeral into one of 12 categories

- Investigate the performance of the claim detection task with various numerical representations

- Experiment on two task settings
  - Claim detection only
  - Joint learning
    - claim detection & numerical category classification
Our Approach for Manager’s Claim Detection

- Use five pre-trained language models and fine-tune them
  - BERT (base), BERT (large), FinBERT, RoBERTa (large), T5 (large)
- Preprocess the input texts with the following numeral formats:

<table>
<thead>
<tr>
<th>Format</th>
<th>Example: Fiscal Year 2018 Fourth Quarter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask</td>
<td>Fiscal Year [MASK] Fourth Quarter</td>
</tr>
<tr>
<td>Marker</td>
<td>Fiscal Year [NUM] 2018 [NUM] Fourth Quarter</td>
</tr>
<tr>
<td>Digit</td>
<td>Fiscal Year [NUM] 2 0 1 8 [NUM] Fourth Quarter</td>
</tr>
<tr>
<td>Scientific (sig1)</td>
<td>Fiscal Year [NUM] 2 [EXP] 3 [NUM] Fourth Quarter</td>
</tr>
<tr>
<td>Scientific (sig4)</td>
<td>Fiscal Year [NUM] 2 . 0 1 8 [EXP] 3 [NUM] Fourth Quarter</td>
</tr>
</tbody>
</table>

- Expect: *Digit* and *Scientific* help language models better recognize numerals
  - *Digit* splits numerals into each digit (avoids subwording numerals)
  - *Scientific* indicate significant digit(s) and magnitude of each numeral
Training Method

• Split train dataset into 5 folds
  • train dataset : valid dataset = 4 : 1 → 5 train/valid datasets
• Fine-tune a language model for each of 5 train/valid datasets
  • Grid search for best hyperparameters
• Average the predictions from 5 models for final prediction
  • Voting for T5 and soft average for other models
Select Models

• Select models for submission
  • joint learning setting
  • Best score in each model of BERT (large), RoBERTa and FinBERT
    • Macro-F1 score (dev) of the claim detection task
    • Experiment using T5 is not conducted before submitting

• Submit models
  1. BERT (large) with *Marker*
  2. RoBERTa with *Scientific (sig4)*
  3. FinBERT with *Marker*
### Results: Effect of Numerical Formats

Macro-F1 (test) for the claim detection task on joint learning:

<table>
<thead>
<tr>
<th>Model</th>
<th>BERT (base)</th>
<th>BERT (large)</th>
<th>FinBERT</th>
<th>RoBERTa</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask</td>
<td>0.895</td>
<td>0.899</td>
<td>0.893</td>
<td><strong>0.904</strong></td>
<td>0.896</td>
</tr>
<tr>
<td>Marker</td>
<td>0.903</td>
<td><strong>0.908</strong></td>
<td>0.910*3</td>
<td>0.904</td>
<td>0.893</td>
</tr>
<tr>
<td>Digit</td>
<td><strong>0.911</strong></td>
<td>0.902</td>
<td>0.901</td>
<td>0.897</td>
<td>0.900</td>
</tr>
<tr>
<td>Scientific (sig1)</td>
<td>0.900</td>
<td>0.897</td>
<td>0.899</td>
<td>0.901</td>
<td><strong>0.903</strong></td>
</tr>
<tr>
<td>Scientific (sig4)</td>
<td>0.904</td>
<td>0.903</td>
<td><strong>0.911</strong></td>
<td>0.895*2</td>
<td>0.901</td>
</tr>
</tbody>
</table>

**Score**: best score in each pretrained model

*: submitted models

#### Results

- Numerals are informative
  - Formats other than *Mask* were best for each models (except RoBERTa)
- Best formats depend on models
  - We need further experiment to investigate the effect of formats
### Results: Effect of Joint Learning

**Improvement of macro-F1 for the claim detection task by joint learning:**

<table>
<thead>
<tr>
<th></th>
<th>BERT (base)</th>
<th>BERT (large)</th>
<th>FinBERT</th>
<th>RoBERTa</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask</td>
<td>0.011</td>
<td>0.014</td>
<td>0.006</td>
<td>0.001</td>
<td>−0.002</td>
</tr>
<tr>
<td>Marker</td>
<td>0.011</td>
<td>0.013</td>
<td>0.017</td>
<td>0.003</td>
<td>−0.005</td>
</tr>
<tr>
<td>Digit</td>
<td>0.009</td>
<td>0.003</td>
<td>0.008</td>
<td>−0.005</td>
<td>−0.002</td>
</tr>
<tr>
<td>Scientific (sig1)</td>
<td>0.014</td>
<td>−0.004</td>
<td>0.008</td>
<td>−0.008</td>
<td>0.005</td>
</tr>
<tr>
<td>Scientific (sig4)</td>
<td>0.009</td>
<td>0.002</td>
<td>0.017</td>
<td>−0.013</td>
<td>0.004</td>
</tr>
</tbody>
</table>

**Results**
- Improve constantly in small models: BERT (base) and FinBERT
- **Not consistent** in large models: BERT (large), RoBERTa and T5
  - Our setting of joint learning might not be optimal

*Red: negative effect*
Conclusion

• Investigate the performance of the claim detection task with various numerical formats in the FinNum-3

• Results
  • Numerals are informative in the claim detection task
  • Best numerical formats depends on the models and settings
  • Joint learning is effective in some cases

• Future works
  • Statistical analysis for the effect of formats
  • Investigating optimal setting of joint learning