CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation

Xie-Sheng Hong, Jia-Jun Lee, Shih-Hung Wu
Chaoyang University of Technology
Taichung, Taiwan

Mike Tian-Jian Jiang
Zeals Co, Ltd
Tokyo, Japan
Outline

• Task Define and Difficulty
• Method: Deep Learning and BERT
• Proposed Runs
• CYUT-1
• CYUT-2
• CYUT-3
• Model Configurations
• Additional Runs
• Official Runs and Additional Runs
• Discussion
• Conclusions and Future Work
Task Define and Difficulty

• Task Define
  • To predict whether a number is a "claim" in a given description
  • A binary classification task

• Difficulty
  • Training dataset imbalance.
    • Analyst's Report (Chinese data): 1 : 0.3
    • Earnings Conference call (English): 1 : 0.14
  • Model trained by this may bias towards specific label
Method: Deep Learning and BERT

• Deep Learning:
  - Using deep learning method to solve the problem

• BERT (Bidirectional Encoder Representations from Transformers)[1]:
  - Pretrained Model
  - Fine-tune and transfer learning
  - Using variant of BERT
Proposed Runs

• CYUT-1: MacBERT / RoBERTa with BiLSTM
  • Baseline for our all systems
  • Data resampling

• CYUT-2: AWD-LSTM
  • Based on CYUT-1.
  • Replace BiLSTM with AWD-LSTM
  • Data resampling

• CYUT-3: Additional data
  • Based on CYUT-1.
  • GPT-2 generates additional data
CYUT-1 : BiLSTM and Data Resampling

• Pretrained Model
  • Analyst's Report (Chinese data): MacBERT\textsuperscript{[6]}
  • Earnings Conference Call (English data): RoBERTa\textsuperscript{[7]}

• The classical classifier
  • BiLSTM

• Data Resampling
  • Repeatedly extract data from lesser number of labels (which is 1).
  • Adjust the data ratio of the 2 labels to 1 : 1
CYUT-2: Replace BiLSTM with AWD-LSTM

- AWD-LSTM (ASGD Weight-Dropped LSTM)[5]
  - It is a variant of LSTM, which is a weight-dropped LSTM
  - It drop part of data of weight matrix between the hidden states in the LSTM

- Advantages
  - 1. It prevents the overfitting problem of traditional LSTM
  - 2. It minimizes the effect on the training speed

- Data Resampling
CYUT-3: GPT-2 Data Generation

• Not to use data resampling
  • Using GPT-2\(^4\) generate additional data that label must be 1 (in-claim)

• GPT-2 Model used:
  • Analyst's Report (Chinese data): CLUECorpussmall, trained by CLUECorpus2020 dataset
    • Dataset: news, wiki, comments from Amazon, etc.
  • Earnings Conference call (English data): Original GPT-2 model without additional training
CYUT-3: GPT-2 Data Generation

- A very intuitive way to generate text with GPT-2:
  - A fixed text + A random number input GPT-2 => Generate subsequent text

- Example:
  - Analyst’s Report (Chinese data):
    - Input: 我們推測會上升 X (We predict an increase of X)
    - Output(Fixed length is 100): 我們推測會上升 X%．明天早晨大跌... (We predict an increase of X% and a big fall tomorrow morning...)

  - Earnings Conference call (English data):
    - Input: We anticipate a X increase
    - Output(Fixed length is 50): We anticipate a X increase in the number of cases with...

- Notice: X meaning a random number (range: 0 - 1000)
CYUT-3: GPT-2 Data Generation

• Amount and rate of data generation
  • Analyst’s Report (Chinese data):
    • Amount of generation: 2200
    • Amount of data: 999: 3220 to 3199 : 3220
    • Data rate: 1 : 0.3 to about 1 : 1

  • Earnings Conference call (English data):
    • Amount of generation: 4000
    • Amount of data: 1039 : 7298 to 5039 : 7298
    • Data rate: 1 : 0.14 to about 1 : 0.7
# Model Configurations

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT Model</td>
<td>macbert-base or Roberta-base</td>
</tr>
<tr>
<td>Batch Size</td>
<td>4 or 8</td>
</tr>
<tr>
<td>Max Length</td>
<td>512</td>
</tr>
<tr>
<td>Optimizer</td>
<td>AdamW</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>2e-5</td>
</tr>
</tbody>
</table>
Model Configurations: Learning Rate Schedule

- Cosine schedule is better linear schedule for our model in this task.
  - Learning rate warmup
  - Dynamically adjusted learning rate

Ref. https://huggingface.co/docs/transformers/main_classes/optimizer_schedules#transformers.get_cosine_schedule_with_warmup.num_cycles
Additional Runs

• Additional run is mainly modified based on CYUT-3:
  • Add the text type used for GPT-2 text generation

  • Adjust the amount of additional text generated
    • Increase or decrease the amount of additional text, etc.

• No change
  • Only BERT + BiLSTM, not use data augmentation method including data resampling at all
## Official Runs and Additional Runs

<table>
<thead>
<tr>
<th>Run</th>
<th>Macro-F1</th>
<th>Micro-F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYUT-2</td>
<td>86.76%</td>
<td>91.73%</td>
<td>90.32%</td>
</tr>
<tr>
<td>CYUT-3</td>
<td>88.20%</td>
<td>92.16%</td>
<td>88.76%</td>
</tr>
<tr>
<td>CYUT-1</td>
<td>88.80%</td>
<td>92.11%</td>
<td>87.34%</td>
</tr>
<tr>
<td>No Change</td>
<td>88.75%</td>
<td>92.52%</td>
<td>89.30%</td>
</tr>
<tr>
<td>More Data and Seeds</td>
<td>89.23%</td>
<td>92.86%</td>
<td>89.92%</td>
</tr>
<tr>
<td>More Seeds</td>
<td>89.30%</td>
<td>93.14%</td>
<td><strong>91.66%</strong></td>
</tr>
<tr>
<td>1000 Data</td>
<td>89.97%</td>
<td>93.16%</td>
<td>89.52%</td>
</tr>
<tr>
<td>More Data</td>
<td><strong>90.24%</strong></td>
<td><strong>93.43%</strong></td>
<td>90.31%</td>
</tr>
</tbody>
</table>
## Official Runs and Additional Runs

<table>
<thead>
<tr>
<th>Run</th>
<th>Macro-F1</th>
<th>Micro-F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYUT-1</td>
<td>85.53%</td>
<td>94.67%</td>
<td>79.82%</td>
</tr>
<tr>
<td>More Seeds</td>
<td>85.93%</td>
<td>95.00%</td>
<td>80.74%</td>
</tr>
<tr>
<td>More Data</td>
<td>86.73%</td>
<td>95.93%</td>
<td>84.78%</td>
</tr>
<tr>
<td>More Data and Seeds</td>
<td>87.17%</td>
<td>95.76%</td>
<td>83.25%</td>
</tr>
<tr>
<td>1000 Data</td>
<td>87.28%</td>
<td>95.73%</td>
<td>83.15%</td>
</tr>
<tr>
<td>CYUT-2</td>
<td>87.49%</td>
<td>95.64%</td>
<td>82.39%</td>
</tr>
<tr>
<td>CYUT-3</td>
<td>87.88%</td>
<td>96.43%</td>
<td>87.25%</td>
</tr>
<tr>
<td>No Change</td>
<td>88.15%</td>
<td>96.22%</td>
<td>85.03%</td>
</tr>
</tbody>
</table>
Discussion

• We Found:
  • Raising the number of additional texts may improve the model, but is not an absolute factor
  • “No Change” is the best system in Earnings Conference Call data

• We assume that:
  • The quality of additional texts may be more important than the amount
  • Poor data quality can be counterproductive
  • Whether data Resampling is causing overfitting?
Discussion

• There are advantages and disadvantages among systems in each category
  • Build a large multi-model system that leverages the strengths of each system
Conclusions and Future Work

• Data augmentation technique for imbalance dataset.
  • Data resampling
  • Data generation

• Official and Additional Run.
  • Analyst's Report (in Macro-F1):
    • Official: 88.80%
    • Additional: 90.24%

  • Earnings Conference Call (in Macro-F1):
    • Official: 87.88%
    • Additional: 88.15%
Conclusions and Future Work

• Pay attention
  • Additional text data quality is more important than amount.
  • Possible overfitting

• Future Work:
  • Improving the quality of data generation
    • T-5 model[2]、GPT-3 model[3], etc.
  • Adjusting the BERT model
  • Building multi-model system make the right talent for the right place
Reference


Thank You!

Please let me know if you have any questions.