Neural Authorship Attribution and Obfuscation

The Pennsylvania State University

Dongwon Lee, Ph.D.

dongwon@psu.edu

HIATUS / Jan 19, 2022
Penn State PIKE Team

- Looking for a HIATUS team to join!

- PI with 11 Ph.D. students
  - https://pike.psu.edu/

- Active research in Data Science and AI areas

- # pubs in top CS venues for last 3 years
  - **Data Science**: KDD (6), ICDM (4), WWW (4), CIKM (4), SIGIR (1), ICDE (1)
  - **AI**: AAAI (4), AAMAS (1)
  - **NLP**: ACL (1), EMNLP (2), NAACL (1)
  - **HCI**: CHI (2), CSCW (1)
Old Problem, New Spin!

Neural Authorship

Authorship Attribution and Obfuscation

Hugging Face

Tasks
- Fill-Mask
- Question Answering
- Summarization
- Table Question Answering
- Text Classification
- Text2Text Generation
- Token Classification
- Translation
- Zero-Shot Classification
- Sentence Similarity

Models
- gpt2
- distilgpt2
- xlnet-base-c
Team’s Expertise

- **EMNLP 2020 & 2021**
  - **TuringBench**: public benchmark environment to study neural Authorship Attribution (AA) problem
    - Created 20 corpus on news genre (200K)—ie, 19 generated by language models and 1 human-written
  - Compared 10 AA detection models

<table>
<thead>
<tr>
<th>AA Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.5893</td>
<td>0.6053</td>
<td>0.5847</td>
<td>0.6147</td>
</tr>
<tr>
<td>SVM (3-grams)</td>
<td>0.7124</td>
<td>0.7223</td>
<td>0.7149</td>
<td>0.7299</td>
</tr>
<tr>
<td>WriteprintsRFC</td>
<td>0.4578</td>
<td>0.4851</td>
<td>0.4651</td>
<td>0.4943</td>
</tr>
<tr>
<td>OpenAI detector</td>
<td>0.7810</td>
<td>0.7812</td>
<td>0.7741</td>
<td>0.7873</td>
</tr>
<tr>
<td>Syntax-CNN</td>
<td>0.6520</td>
<td>0.6544</td>
<td>0.6480</td>
<td>0.6613</td>
</tr>
<tr>
<td>N-gram CNN</td>
<td>0.6909</td>
<td>0.6832</td>
<td>0.6665</td>
<td>0.6914</td>
</tr>
<tr>
<td>N-gram LSTM-LSTM</td>
<td>0.6694</td>
<td>0.6824</td>
<td>0.6646</td>
<td>0.6898</td>
</tr>
<tr>
<td>BertAA</td>
<td>0.7796</td>
<td>0.7750</td>
<td>0.7758</td>
<td>0.7812</td>
</tr>
<tr>
<td>BERT-Multinomial</td>
<td>0.8031</td>
<td>0.8021</td>
<td>0.7996</td>
<td>0.8078</td>
</tr>
<tr>
<td>RoBERTa-Multinomial</td>
<td><strong>0.8214</strong></td>
<td><strong>0.8126</strong></td>
<td><strong>0.8107</strong></td>
<td><strong>0.8173</strong></td>
</tr>
</tbody>
</table>
Team’s Expertise

- Authorship Obfuscation (AO) in the Wild
  - What do real human-written AO look like?
  - What are AO less likely to be detected?
Contact Information

Dongwon Lee, Ph.D.

dongwon@psu.edu

https://pike.psu.edu/