LLM-based Entity Extraction Is Not for Cybersecurity

Maxime Würsch, EPFL & CYD Campus
Context

- Software present everywhere
  - Contains vulnerabilities
- Technologies evolves fast
  - Increase security gap
- Need to stay up to date
  - Reduce risk of attack

- Common approach is bibliometrics search
  - Using entities extraction
  - Comparing through embedding space
- Emergence of LLM-based entities extraction
  - Need evaluation about performance
Aim

• Measure performance of entities extractors
  • Compare between LLM-based and not
  • Similitudes and differences between models

• Relevance of the method to classify documents

• Are LLM-based entity extractors suited for scientific bibliometrics?
Large Language Model (LLM)

• Attention since late 2022, with conversational agent's public trials

• Term come from ELMo LLM in 2018
  • At least 100M parameters and 1B tokens

• Now goes to more than trillions of parameters

• Small LLMs
  • Less resources demanding
  • Weaker version of larger model
  • Reduce version of failure modes
Methods (dataset information)

• Comes from arXiv, until late 2022, Computer Science (cs) category
• Subset of the cs category
• Keep only English text
  • XLM-RoBERTa model
• Remove preamble and references

<table>
<thead>
<tr>
<th>arXiv selected cs listings</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs.CR</td>
</tr>
<tr>
<td>cs.NI</td>
</tr>
<tr>
<td>cs.CC</td>
</tr>
<tr>
<td>cs.LO</td>
</tr>
<tr>
<td>cs.DS</td>
</tr>
<tr>
<td>cs.IT</td>
</tr>
<tr>
<td>cs.CL</td>
</tr>
<tr>
<td>cs.AI</td>
</tr>
</tbody>
</table>
Methods (models)

- 4 major types
- Document segmented to fully fit the attention windows
- At most 100 entities extracted
  - Select by highest confidence

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Refs</th>
<th>Entities/Doc</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy Large*P</td>
<td>[17]</td>
<td>99.3 ± 6.93</td>
<td>Noun Extractor</td>
</tr>
<tr>
<td>spaCy Transformer*P</td>
<td>[17]</td>
<td>99.3 ± 6.97</td>
<td></td>
</tr>
<tr>
<td>Yake*P</td>
<td>[5]</td>
<td>19.9 ± 1.97</td>
<td>Keyphrase Extractor</td>
</tr>
<tr>
<td>KeyBERT*P</td>
<td>[15]</td>
<td>99.3 ± 7.25</td>
<td></td>
</tr>
<tr>
<td>KBIR kpcrowd</td>
<td>[23, 25]</td>
<td>96.9 ± 14.6</td>
<td></td>
</tr>
<tr>
<td>KBIR inspec</td>
<td>[23, 37]</td>
<td>76.4 ± 27.7</td>
<td></td>
</tr>
<tr>
<td>BERT-base-uncased</td>
<td>[11]</td>
<td>44.7 ± 24.0</td>
<td></td>
</tr>
<tr>
<td>BERT-base-uncased</td>
<td>[11]</td>
<td>43.3 ± 23.3</td>
<td>NER+CON R</td>
</tr>
<tr>
<td>XLM-RoBERTa-base Onconotes 5</td>
<td>[40, 18]</td>
<td>36.4 ± 23.4</td>
<td>NER+NUM</td>
</tr>
<tr>
<td>ELECTRA-base conll03</td>
<td>[8, 38]</td>
<td>39.9 ± 25.0</td>
<td></td>
</tr>
<tr>
<td>BERT-large-cased conll03</td>
<td>[11, 38]</td>
<td>41.7 ± 24.9</td>
<td></td>
</tr>
<tr>
<td>BERT-large-uncased conll03</td>
<td>[11, 38]</td>
<td>33.5 ± 23.3</td>
<td></td>
</tr>
<tr>
<td>DistilBERT-base-uncased conll03</td>
<td>[39, 38]</td>
<td>37.7 ± 24.8</td>
<td></td>
</tr>
<tr>
<td>RoBERTa-large conll03</td>
<td>[24, 38]</td>
<td>28.7 ± 21.1</td>
<td></td>
</tr>
<tr>
<td>XLM-RoBERTa-large conll03</td>
<td>[14, 38]</td>
<td>26.0 ± 19.5</td>
<td></td>
</tr>
<tr>
<td>BERT COCA-docusco</td>
<td>[11, 20]</td>
<td>99.6 ± 6.11</td>
<td>TokC</td>
</tr>
</tbody>
</table>
Methods (visualisations)

• Hierarchical clustering
  • Embedded with SpaCy
  • Average cosine distance
  • Identify similitude between extractor

• 2D Projection
  • Subsample data: reduce processing time and number of point
  • 6 embeddings:
    • SpaCy, GloVe, Fasttext, Word2Vec, BERT-large, GPT-2
  • 4 low-dimensional projection
    • Linear, spectral, t-SNE, UMAP
  • Show if themes can be detected in an unsupervised way
Results and Discussion

• Performance mainly defined by architecture and fine-tuned dataset

• Dataset not based on scientific texts
  • Conll03

• => Not suited for scientific articles
Results and Discussion

- Cosine similarity of embedding do not perform well to cluster themes
  - Even with 2D embedding algorithm that tend to overfit

- Exception with NER
Umap projection of Spacy using RoBERTa-large conll03 (NER)
Results and Discussion

• Cosine similarity highly dependent of embedding space
• Important change with different embedding and algorithm
Conclusion

• LLM-based entity extraction seems not suited for concept-oriented bibliometrics in scientific article

• Work only on arXiv cs category

• Nouns extraction seems more robust
Thanks for your attention!

Questions?