Named Entity Recognition of IEEE Abstracts
Paulina Toro Isaza, John Alling, You Wu, Justin Clark

The Problem
The Institute for Electrical and Electronic Engineering publishes 30% of the world’s literature in electrical and electronic engineering and computer engineering. These 5.2 million plus documents are available through IEEE Xplore.

NER Multi-Class vs. Multi-Label
The inherited annotations and models were for mutually exclusive (multi-class) entities. As per IEEE’s request, we moved to the multi-label task to accommodate examples as shown:

NER Results
For our capstone project at Harvard University’s Institute of Applied Computational Science, we used natural language processing models such as spaCy and BERT to perform named entity recognition on abstracts published by IEEE. All training was done in PyTorch and data was stored in a RedisML data warehouse.

Motivation
Discoverability
Help users discover methods, organizations, and products without needing to know what keyword to search beforehand.

NER Modeling Approaches

spaCy Baseline
- CNN-based
- Non-contextual embeddings from GloVe
- Pretrained on NER task of Common Crawl of internet
- Pretrained entities: person, location, and organization
- Inherited from a previous IEEE team

BERT-Base
- 12-layer transformer
- Bidirectional
- Attention-based
- Contextual embeddings
- Pretrained token masking task on Book Corpus and Wikipedia text
- Same architecture as BERT-Base
- Pretrained token masking task on academic papers
- 82% from medical
- 18% from computer science

SciBERT
- Same architecture as BERT-Base
- Pretrained token masking task on academic papers
- 82% from medical
- 18% from computer science

The Importance of Good Labeled Data
Our NER model performance was mostly capped by the quality of our labeled entities. In order to be “good”, labeled data should be:

Consistent
If human annotated labels are not consistent, the model tries to learn from conflicting examples. We could not expect our model to outperform the agreement of human annotators.

Expert-Driven
Our models tended to perform worse for fields in which we, the annotators, had little or no expertise: materials science, environmental science, and some social sciences.

NER Modeling Approaches

spaCy
Baseline

BERT-Base

SciBERT
- Same architecture as BERT-Base
- Pretrained token masking task on academic papers
- 82% from medical
- 18% from computer science

Coreference Resolution (Coref)
In order to be “good”, labeled data should be:

Coref Modeling Approach

Coreference Resolution (Coref)

Coref Modeling Approach

Baseline

BERT-Base

SciBERT
- Same architecture as BERT-Base
- Pretrained token masking task on academic papers
- 82% from medical
- 18% from computer science

The Importance of Good Labeled Data
Our NER model performance was mostly capped by the quality of our labeled entities. In order to be “good”, labeled data should be:

Consistent
If human annotated labels are not consistent, the model tries to learn from conflicting examples. We could not expect our model to outperform the agreement of human annotators.

Expert-Driven
Our models tended to perform worse for fields in which we, the annotators, had little or no expertise: materials science, environmental science, and some social sciences.

Microsoft Academic Graph

Abstract text and metadata such as author and field of study for more than 250 million papers

NER Results

Training Loss: BERT-Base

Training Loss: SciBERT

F1 Score: 10-Fold Cross Validation

<table>
<thead>
<tr>
<th></th>
<th>spaCy*</th>
<th>BERT-Base</th>
<th>SciBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>0.07</td>
<td>0.67 ± 0.014</td>
<td>0.69 ± 0.011</td>
</tr>
<tr>
<td>Organization</td>
<td>0.34</td>
<td>0.64 ± 0.017</td>
<td>0.65 ± 0.076</td>
</tr>
<tr>
<td>Product</td>
<td>0.21</td>
<td>0.61 ± 0.042</td>
<td>0.57 ± 0.048</td>
</tr>
<tr>
<td>None</td>
<td>-</td>
<td>0.97 ± 0.002</td>
<td>0.97 ± 0.001</td>
</tr>
</tbody>
</table>

* spaCy trained and tested only on multi-class task.