Named Entity Recognition of IEEE Abstracts

AC 297r
Final Presentation
May 12, 2021

Paulina Toro Isaza, John Alling, You Wu, Justin Clark
Outline

1. The Problem
2. Task Overview
3. Training Data
4. Multi-Label Named Entity Recognition
5. Models for Named Entity Recognition
6. Coreference Resolution
7. Software Packaging
8. Importance of Good Labeled Data
The Context

World’s largest organization of technical professionals

Publishes **30%** of world’s literature in:
- Electrical and electronic engineering
- Computer engineering

**5.2 million** plus documents
- Journals
- Conferences
- Standards
- eBooks
- And more...
The Problem

unstructured text data → structured text data
The Problem

Business Use-Cases

Search and discoverability
Organization innovation profiles
Recommender systems
Knowledge graphs
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 Redshift data warehouse.
Coreference Resolution

MIT
Massachusetts Institute of Technology

Adobe Photoshop
Adobe PS
Photoshop

Convolutional Recurrent Neural Network
conv-RNN
Convolutional Recurrent Neural Network (conv-RNN)
## Data

<table>
<thead>
<tr>
<th></th>
<th>Microsoft Academic Graph</th>
<th>IEEE Xplore® Digital Library</th>
<th>Inherited Annotations</th>
<th>IACS Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Papers</strong></td>
<td>250 million</td>
<td>5.2 million</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abstracts</strong></td>
<td>73 million</td>
<td>4.8 million</td>
<td>1322</td>
<td>1050</td>
</tr>
<tr>
<td><strong>Year of Publication</strong></td>
<td>1800-2022</td>
<td>1879-2021</td>
<td>1899-2018</td>
<td>1879-2021</td>
</tr>
<tr>
<td><strong>Fields of Study</strong></td>
<td>19</td>
<td>19</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>(electrical engineering, computer science, materials science, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training Data
## Stratified Sampling

### Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-2021</td>
<td>50%</td>
</tr>
<tr>
<td>2001-2015</td>
<td>30%</td>
</tr>
<tr>
<td>Before 2000</td>
<td>20%</td>
</tr>
</tbody>
</table>

### Field of Study

<table>
<thead>
<tr>
<th>Top 5 Fields</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>46%</td>
</tr>
<tr>
<td>Engineering</td>
<td>18%</td>
</tr>
<tr>
<td>Materials Science</td>
<td>12%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>10%</td>
</tr>
<tr>
<td>Physics</td>
<td>7%</td>
</tr>
</tbody>
</table>

- Sampled proportionately to all fields
# Manual Annotations

<table>
<thead>
<tr>
<th></th>
<th>Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total labeled</td>
<td>1050</td>
</tr>
<tr>
<td>Train &amp; Validation</td>
<td>800</td>
</tr>
<tr>
<td>Test</td>
<td>200</td>
</tr>
<tr>
<td><strong>Shared</strong> (to measure inter-annotator agreement)</td>
<td>50</td>
</tr>
</tbody>
</table>

- ~2 minutes per annotation
- ~43 combined hours
- ~10 hours per annotator
## Annotation Guidelines

- **What is a Method, Organization, or Product?**

<table>
<thead>
<tr>
<th>Method</th>
<th>Organization</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guiding questions</td>
<td>How was this research done?</td>
<td>Are there people working here?</td>
</tr>
<tr>
<td>Examples</td>
<td>K-Nearest-Neighbors (KNN) Optimization Quantitative analysis</td>
<td>Harvard IEEE Google</td>
</tr>
</tbody>
</table>

---
Inter-annotator Agreement

<table>
<thead>
<tr>
<th>Entity Class</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>0.62</td>
</tr>
<tr>
<td>Organization</td>
<td>0.60</td>
</tr>
<tr>
<td>Product</td>
<td>0.68</td>
</tr>
</tbody>
</table>
NER Task: Multi-Label
Multi-Label

IEEE Xplore

Organization  Product
## Multi-Class → Multi-Label

**BERT-Base**

No hyperparameter tuning

<table>
<thead>
<tr>
<th>Task</th>
<th>Loss</th>
<th>Avg</th>
<th>Org</th>
<th>Method</th>
<th>Product</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Class</td>
<td>Cross Entropy</td>
<td>0.58</td>
<td>0.45</td>
<td>0.61</td>
<td>0.51</td>
<td>0.98</td>
</tr>
<tr>
<td>Multi-Label</td>
<td>Binary Cross Entropy*</td>
<td>0.64</td>
<td>0.43</td>
<td>0.62</td>
<td>0.55</td>
<td>0.97</td>
</tr>
</tbody>
</table>

* Threshold for positive class: 0.4
NER Models
NER: Inherited spaCy Model

spaCy

- 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

<table>
<thead>
<tr>
<th></th>
<th>Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>1197</td>
</tr>
<tr>
<td>Validation</td>
<td>133</td>
</tr>
</tbody>
</table>
NER: BERT-Base Model

- 12-layer transformer
- Attention-based
- Bi-directional
- Pretrained token masking task on Books Corpus and Wikipedia text

Data: IACS Annotations

<table>
<thead>
<tr>
<th></th>
<th>Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train &amp; Validation</td>
<td>800</td>
</tr>
<tr>
<td>Test</td>
<td>200</td>
</tr>
</tbody>
</table>

Hyperparameter Tuning
- Optimizer: AdamW
- Learning rate: 9.9e-5
- Weight decay: 0.0024
- Dropout prob: 0.13
- Epochs: 4
- Batch size: 4
NER: SciBERT Model

- Same architecture as BERT
- Pretrained token masking task on academic text
  - Medical: 82%
  - Computer Science: 18%

NER: SciBert Model

Data: IACS Annotations

<table>
<thead>
<tr>
<th></th>
<th>Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train &amp; Validation</td>
<td>800</td>
</tr>
<tr>
<td>Test</td>
<td>200</td>
</tr>
</tbody>
</table>

Hyperparameter Tuning
Optimizer: AdamW
Learning rate: 1.8e-4
Weight decay: 8.2e-8
Dropout prob: 0.27
Epochs: 5
Batch size: 8
**NER: Model Comparisons**

### F1 Score: 10-Fold Cross Validation

<table>
<thead>
<tr>
<th></th>
<th>spaCy (reported)*</th>
<th>spaCy*</th>
<th>BERT-Base</th>
<th>SciBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td>0.40</td>
<td>0.07</td>
<td>0.67 ± 0.014</td>
<td>0.69 ± 0.011</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>0.42</td>
<td>0.34</td>
<td>0.64 ± 0.117</td>
<td>0.65 ± 0.076</td>
</tr>
<tr>
<td><strong>Product</strong></td>
<td>0.40</td>
<td>0.21</td>
<td>0.61 ± 0.042</td>
<td>0.57 ± 0.048</td>
</tr>
<tr>
<td><strong>None</strong></td>
<td></td>
<td></td>
<td>0.97 ± 0.002</td>
<td>0.97 ± 0.001</td>
</tr>
</tbody>
</table>

*spaCy models only trained and validated on multi-class task. spaCy model trained on previous team annotations and validated on IACS annotations.
The aim of this study was to detect sleep stages of human by using EEG signals. In accordance with this purpose, discrete wavelet transforms (DWT) and empirical mode decomposition (EMD) were separately used for feature extraction...

The prototype has been developed using App Inventor 2, PHP and MYSQL. A mock-up database of the victim personal information is stored in MYSQL.
# NER: Failure Modes (Method)

## Most Commonly Mistaken Tokens

<table>
<thead>
<tr>
<th>False Positives</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT</td>
<td>SciBERT</td>
</tr>
<tr>
<td>-</td>
<td>5.6%</td>
<td>5.4%</td>
</tr>
<tr>
<td>)</td>
<td>2.3%</td>
<td>2.5%</td>
</tr>
<tr>
<td>(</td>
<td>1.9%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>False Negatives</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT</td>
<td>SciBERT</td>
</tr>
<tr>
<td>-</td>
<td>4.3%</td>
<td>4.6%</td>
</tr>
<tr>
<td>##s</td>
<td>2.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>)</td>
<td>2.7%</td>
<td>##s</td>
</tr>
<tr>
<td>)</td>
<td>2.7%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

*How to read the tables:* 5.6% of False Positives predicted by BERT is the token ‘-’
This is a Method, this is an Organization, this is a Product.

<table>
<thead>
<tr>
<th>Human</th>
<th>BERT-Base</th>
<th>SciBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...</td>
<td>We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...</td>
<td>We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...</td>
</tr>
<tr>
<td>Internet of Things(IoT) is an important part of the new generation of information technology and an important stage of development in the era of informatization. As a next generation network, Information Centric Network (ICN) has been introduced into the IoT...</td>
<td>Internet of Things(IoT) is an important part of the new generation of information technology and an important stage of development in the era of informatization. As a next generation network, Information Centric Network (ICN) has been introduced into the IoT...</td>
<td>Internet of Things(IoT) is an important part of the new generation of information technology and an important stage of development in the era of informatization. As a next generation network, Information Centric Network (ICN) has been introduced into the IoT...</td>
</tr>
</tbody>
</table>
Coreference Resolution
Coreference Resolution: Naive

"Deep Neural Network (DNN)"
Coreference Resolution:
Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"  "DNN"
Coreference Resolution: Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"  "DNN"
Coreference Resolution: Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"  "DNN"

"deep neural networks"
Coreference Resolution: Naive

R.J. Bayardo, Y. Ma, R. Srikant, “Scaling Up All Pairs Similarity Search”
Coreference Resolution: Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"  "DNN"

"deep neural networks"
Coreference Resolution:
Naive

"Deep Neural Network (DNN)"
"Deep Neural Network"
"deep neural networks"
"DNN"
Coreference Resolution: Naive

- No “ground truth”
- Clusters are validated manually

<table>
<thead>
<tr>
<th>Method</th>
<th>Labels</th>
<th>Unique</th>
<th>Clusters</th>
<th>Mis-clustered</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>118</td>
<td>30</td>
<td>21</td>
<td>1</td>
<td>96.7%</td>
</tr>
<tr>
<td>Product</td>
<td>406</td>
<td>104</td>
<td>68</td>
<td>1</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

| Organization | Method | 3,752 | 1,315 | 615 | 17 | 98.7% |
| Product      | 406    | 104   | 68    | 1  | 99.0% |
Preparing for Future Use
Packaging

Project based on the cookiecutter data science project template. #cookiecutterdatascience

README
- Requirements
- Examples
- Retraining Guidance

src
- Models
- Classes & Methods
- Data

notebooks
- Model Training Walkthrough

...  
- Docs
  - Annotation Guidelines
- Reports
- Tests
- Etc.

https://github.com/iacs-capstone-ieee/ieee_ner_coref
from src.ieee_ner_coref import EntityRecognizer
from src.models import Document

paper_ids = set([3101806155, 2990891550, 3023781467, 3039827942])
docs = Document.from_paper_ids(paper_ids)
annotated_docs = EntityRecognizer(method='best').recognize_entities(docs)
Packaging

Coreference Resolution Example

```python
from src.ieee_ner_coref import EntityClusterer

# Get annotated documents from somewhere
annotated_docs = ...
all_entities = [entity for doc in annotated_docs
               for entity in doc/entities]
entity_groups = EntityClusterer(method='best').cluster(all_entities)
```

https://github.com/iacs-capstone-ieee/ieee_ner_coref#coreference-resolution-quickstart
Documentation

Getting Started

Here is how to get up and running with named entity recognition and coreference resolution (entity clustering) for IEEE abstracts.

Requirements

This project makes a few assumptions:

1. Documents are stored in Amazon Redshift using a specific subset of the Microsoft Academic Graph schema.
2. An Amazon S3 bucket exists that contains the assets (pretrained models and data) for this project.
3. The project is running on a machine that has CUDA capability. This isn't strictly required, but running models will be much slower in its absence.

Setup

1. Using your AWS credentials, create an AWS profile in ~/.aws/credentials to allow downloading assets from S3.

brk.mn/ieee_ner_coref
The Importance of Good Labeled Data
### Inter-annotator Agreement (IAA)

<table>
<thead>
<tr>
<th>&quot;True&quot; annotator</th>
<th>&quot;Model&quot; annotator</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.66</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.84</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>0.69</td>
</tr>
</tbody>
</table>
IAA by Entity Class

![Graph showing IAA by Entity Class with F1 score on the y-axis and Entity Class on the x-axis. The graph includes data points for Annotator Pair and Mean agreement.](image)
Annotator Pair vs. BERT-Base F1 Scores
F1 Score vs. Dataset Size

Number of Labeled Abstracts

F1 Score

- None
- Organization
- Method
- Product
Annotator Pair vs. BERT-Base F1 Scores
“Good” Labeled Data is...

- Consistent
- Expert-Driven
BERT-Base F1 Scores by Field of Study

- Philosophy
- Geology
- Mathematics
- Engineering
- Computer science
- Physics
- Medicine
- Business
- Biology
- Materials science
- Psychology
- Environmental science
- Geography
- Political science
- History

F1 score

annotator area of expertise
“Good” Labeled Data is...

- Consistent
- Expert-Driven
The Answer

Future Directions

There are a number of ways in which this project could be improved and extended. Here we outline some of them.

Refining Annotation Guidelines

We believe the current performance bottleneck in this project is due to inconsistencies between annotators. One way to reduce that inconsistency is to more precisely define the entity classes. We spent significant time on this already, but others may have more luck. One could also define more entity classes to help sharpen the somewhat fuzzy boundaries that currently exist. For example, the follow are entity classes we encountered and found similar to existing classes:

- **Tasks** - Sometimes "image segmentation" is a method that is being used to accomplish a specific task, say "object recognition." But sometimes "image segmentation" is the overarching task that the paper is trying to accomplish, and the authors may create a new method or algorithm for achieving that task. In those papers, "image segmentation" is mentioned but is never actually used as a method in and of itself. It may be helpful to draw the distinction between papers that try many methods to accomplish "image segmentation" and papers that use "image segmentation" to accomplish something else.

- **Devices** - As a lot of IEEE papers relate to hardware, there are a lot of mentions of generic hardware components like "bipolar transistors" or "IEC 1107 optical ports." Neither of these are products as we've defined them, but they don't look completely distinct from products either. Boundaries here can get fuzzy, especially with annotators who are not familiar with the domain.

- **Natural Phenomena** - In some papers, authors are engineering solutions that rely on specific natural phenomena, e.g. Van der Waals forces, specific kinds of radiation, or certain properties of light. These are sometimes used to accomplish specific tasks, which muddles their
Special Thanks to...

IEEE Team:
- Lavanya Savam
- Milad Hosseinipour
- Jeremy Brent
- David Goldstein
- Andrew Sproul
- Zibin Guan

AC 297r Capstone Teaching Staff:
- Chris Tanner
- Isaac Slavitt

Thank you!
Attributions

IACS Logo and Shield © Institute for Applied Computational Science at Harvard University

IEEE and IEEE Xplore Logos © The Institute of Electrical and Electronics Engineers

Icons and Graphics © Flaticon, Freepik, srip, wanicon, geotatah, Becris, and Paulina Toro Isaza
References


