Named Entity Recognition of IEEE Abstracts

AC 297r
Final Presentation
May 12, 2021

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

IACS
INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

IEEE
Advancing Technology for Humanity

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

IEEE
Advancing Technology for Humanity

IEEE Xplore®
Digital Library

World's largest organization of technical professionals

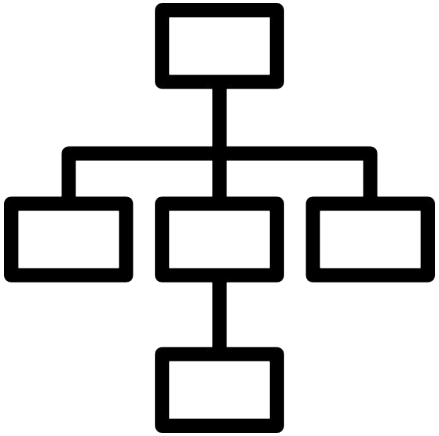
Publishes **30%** of world's literature in:

- Electrical and electronic engineering
- Computer engineering

5.2 million plus documents

- Journals
- Conferences
- Standards
- eBooks
- And more...

The Problem

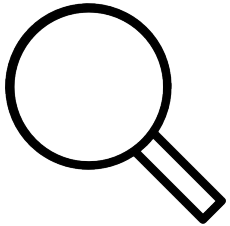


unstructured text data

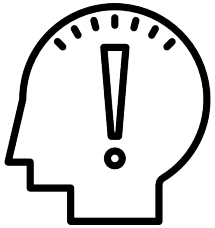
structured text data

The Problem

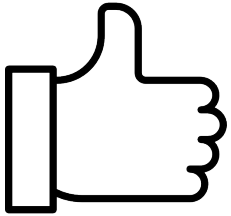
Business Use-Cases



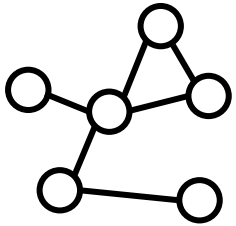
Search and discoverability



Organization innovation profiles



Recommender systems



Knowledge graphs

Named Entity Recognition (NER)

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

PS

Adobe Photoshop

Adobe PS

Photoshop

Convolutional Recurrent Neural Network

conv-RNN

Convolutional Recurrent Neural Network (conv-RNN)

Data

	Microsoft Academic Graph	IEEE Xplore® Digital Library	Inherited Annotations	IACS Annotations
Papers	250 million	5.2 million		
Abstracts	73 million	4.8 million	1322	1050
Year of Publication	1800-2022	1879-2021	1899-2018	1879-2021
Fields of Study (electrical engineering, computer science, materials science, etc.)	19	19	16	19

Training Data

Stratified Sampling

Year of Publication

Year	Percent
2016-2021	50%
2001-2015	30%
Before 2000	20%

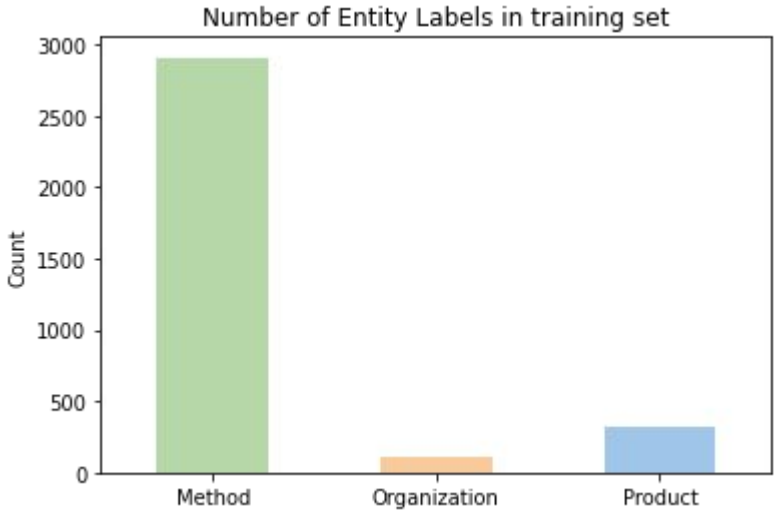
Field of Study

Top 5 Fields	Percent
Computer Science	46%
Engineering	18%
Materials Science	12%
Mathematics	10%
Physics	7%

- Sampled proportionately to all fields

Manual Annotations

	Abstracts
Total labeled	1050
Train & Validation	800
Test	200
Shared (to measure inter-annotator agreement)	50



- ~2 minutes per annotation
- ~43 combined hours
- ~10 hours per annotator

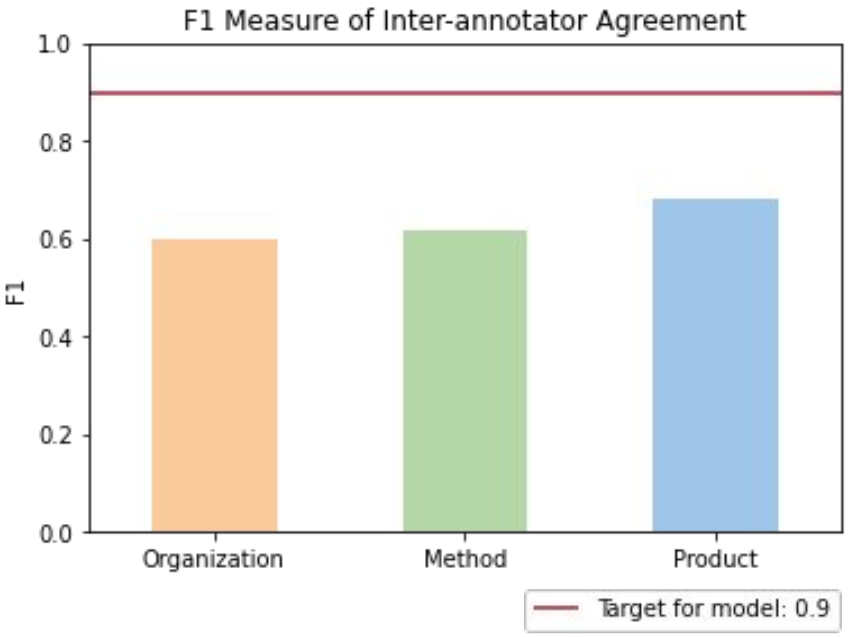
Annotation Guidelines

- What is a Method, Organization, or Product?

	Method	Organization	Product
Guiding questions	*How was this research done?*	*Are there people working here?*	*Can this be (potentially) sold as a commodity?*
Examples	K-Nearest-Neighbors (KNN) Optimization Quantitative analysis	Harvard IEEE Google	MATLAB Tensorflow C++

Inter-annotator Agreement

Entity Class	F1 Score
Method	0.62
Organization	0.60
Product	0.68



NER Task:

Multi-Label

Multi-Label

Multi-Class → Multi-Label

BERT-Base

No hyperparameter tuning

		F1 Score				
Task	**Loss**	**Avg**	**Org**	**Method**	**Product**	**None**
Multi-Class	Cross Entropy	0.58	0.45	0.61	0.51	0.98
Multi-Label	Binary Cross Entropy*	0.64	0.43	0.62	0.55	0.97

* Threshold for positive class: 0.4

NER Models

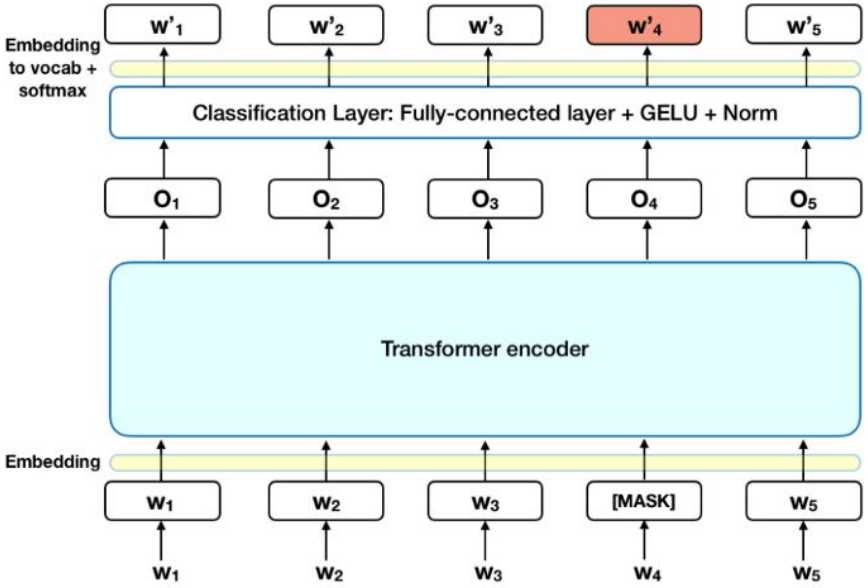
NER:
Inherited spaCy Model

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

	Abstracts
Train	1197
Validation	133

NER:
BERT-Base Model



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

J. Devlin, M. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectionaly Transformers for Language Understanding”

NER:
BERT-Base Model

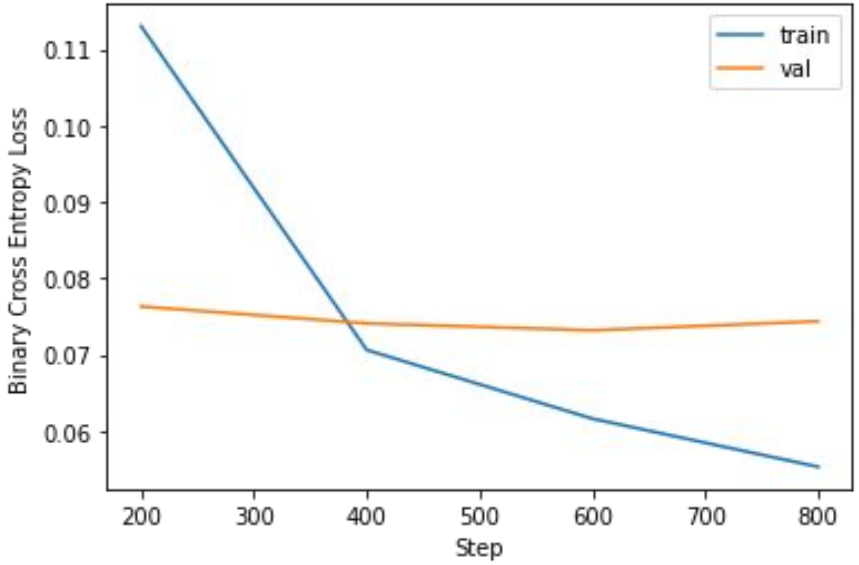
Data: IACS Annotations

	Abstracts
Train & Validation	800
Test	200

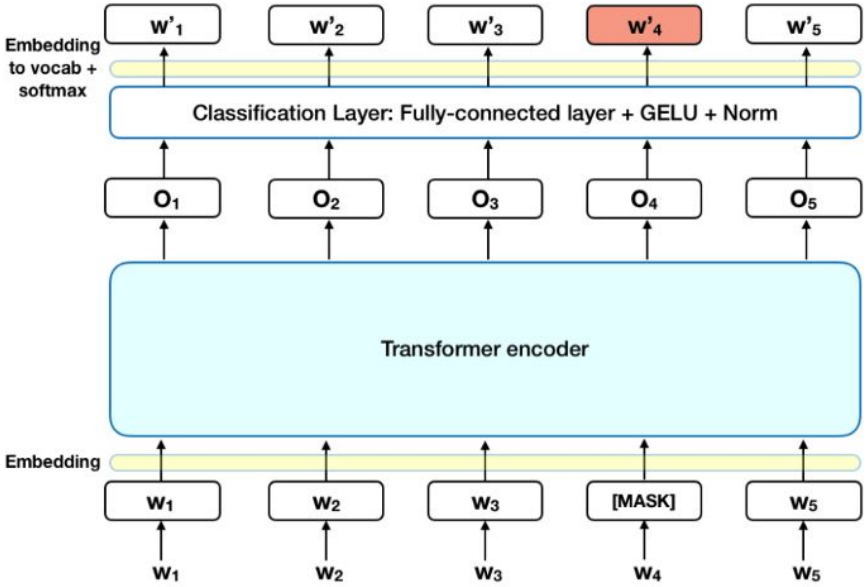
Hyperparameter Tuning

Optimizer: AdamW
Learning rate: 9.9e-5
Weight decay: 0.0024
Dropout prob: 0.13
Epochs: 4
Batch size: 4

Train & Validation Loss



NER:
SciBERT Model



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

I. Beltagy, K. Lo, A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text”

NER:
SciBert Model

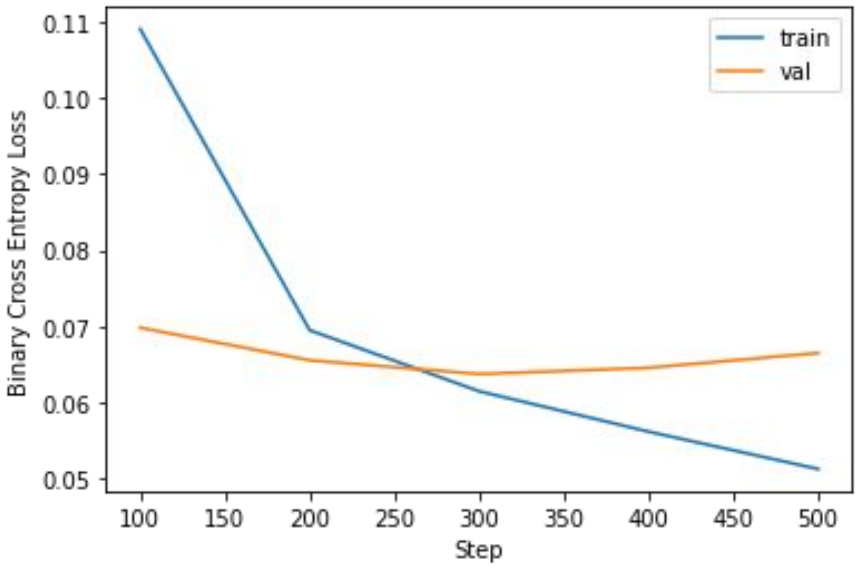
Data: IACS Annotations

	Abstracts
Train & Validation	800
Test	200

Hyperparameter Tuning

Optimizer: AdamW
Learning rate: 1.8e-4
Weight decay: 8.2e-8
Dropout prob: 0.27
Epochs: 5
Batch size: 8

Train & Validation Loss



NER:
Model Comparisons

F1 Score: 10-Fold Cross Validation

	spaCy (reported)*	spaCy*	BERT-Base	SciBERT
Method	0.40	0.07	0.67 ± 0.014	0.69 ± 0.011
Organization	0.42	0.34	0.64 ± 0.117	0.65 ± 0.076
Product	0.40	0.21	0.61 ± 0.042	0.57 ± 0.048
None			0.97 ± 0.002	0.97 ± 0.001

*spaCy models only trained and validated on multi-class task. spaCy model trained on previous team annotations and validated on IACS annotations.

NER:
Good Predictions

This is a Method, this is an Organization, this is a Product.		
Human	**BERT-Base**	**SciBERT**
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 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 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.	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.	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

False Positives			
BERT		SciBERT	
-	5.6%	-	5.4%
)	2.3%	)	2.5%
(	1.9%	(	2.5%

False Negatives			
BERT		SciBERT	
-	4.3%	-	4.6%
##s	2.7%	)	3.2%
)	2.7%	##s	2.2%

***How to read the tables**: 5.6% of False Positives predicted by BERT is the token '-'*

NER: Sample Errors

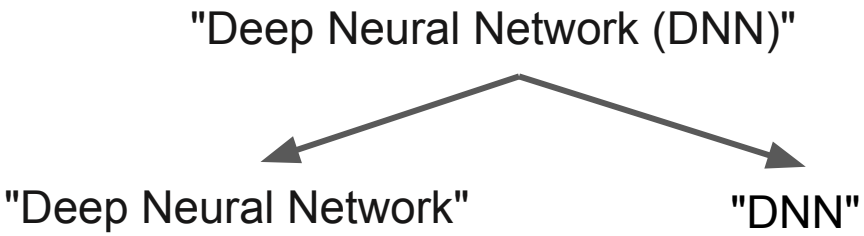
This is a Method, this is an Organization, this is a Product.		
Human	**BERT-Base**	**SciBERT**
We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...	We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...	We present an all-digital measurement circuit that enables wafer-level test and characterization of phase-locked loop (PLL) response...
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...	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...	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...

Coreference Resolution

Coreference Resolution: Naive

"Deep Neural Network (DNN)"

Coreference Resolution: Naive



Coreference Resolution: Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"

"DNN"

Coreference Resolution: Naive

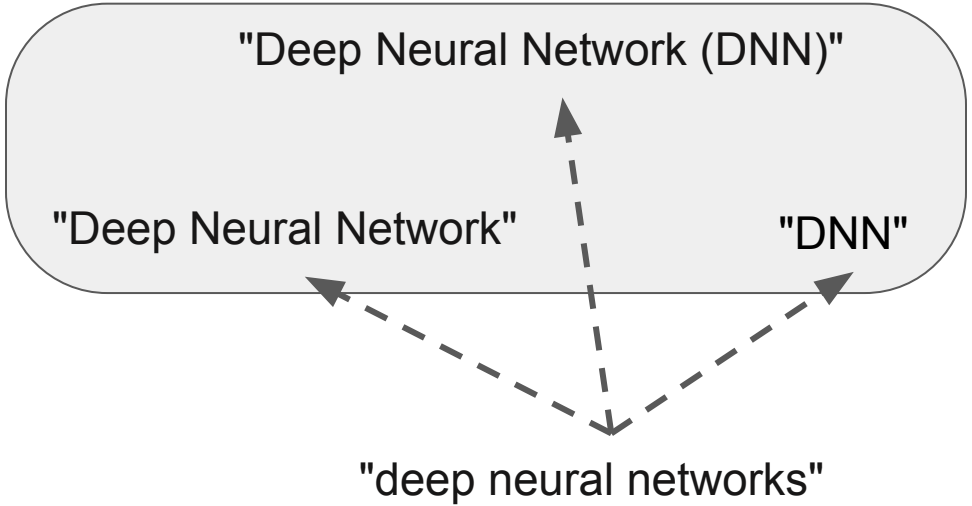
"Deep Neural Network (DNN)"

"Deep Neural Network"

"DNN"

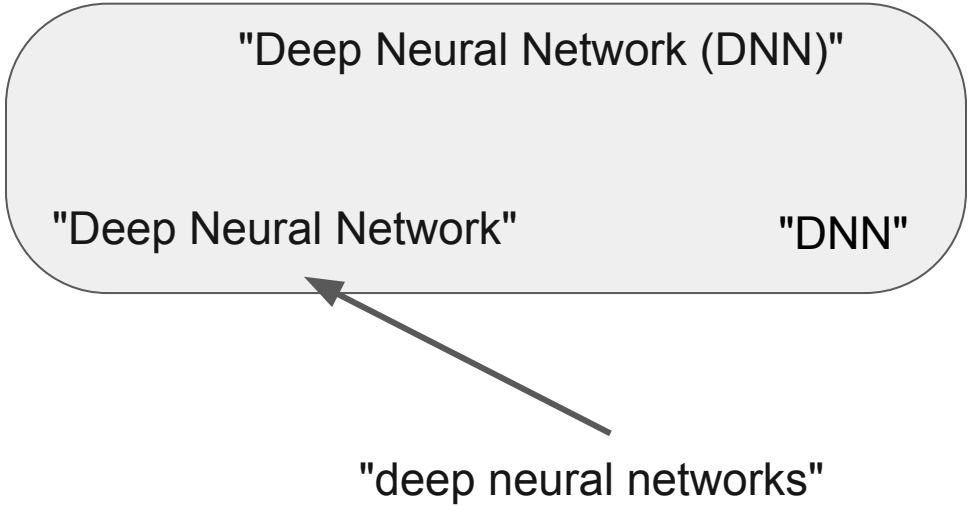
"deep neural networks"

Coreference Resolution: Naive



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

Coreference Resolution: Naive



Coreference Resolution: Naive

"Deep Neural Network (DNN)"

"Deep Neural Network"

"DNN"

"deep neural networks"

Coreference Resolution: Naive

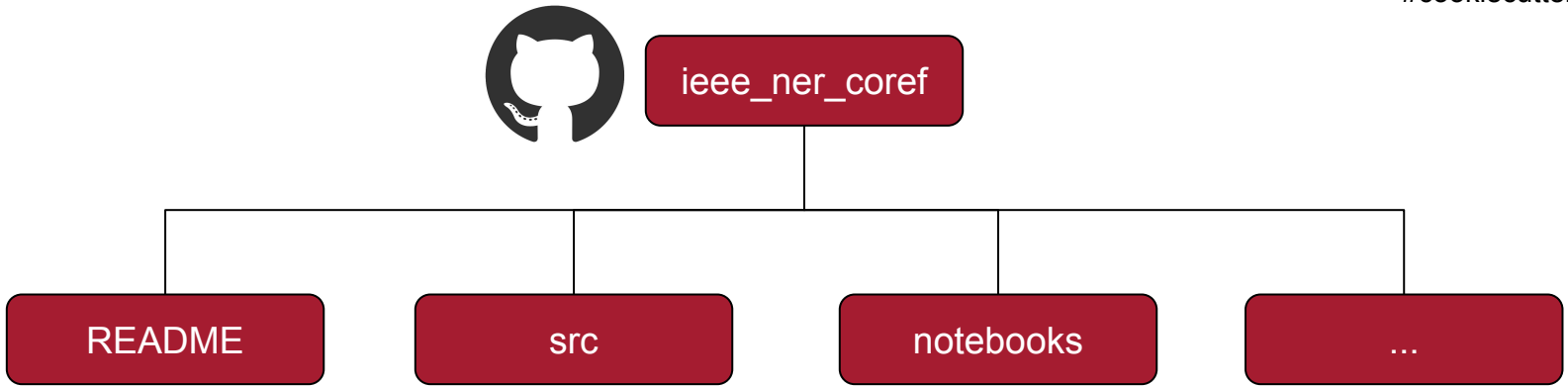
- No “ground truth”
- Clusters are validated manually

	Labels	Unique	Clusters	Mis-clustered	Accuracy
Method	3,752	1,315	615	17	**98.7%**
Organization	118	30	21	1	**96.7%**
Product	406	104	68	1	**99.0%**

Preparing for Future Use

Packaging

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



- Requirements
- Examples
- Retraining Guidance

- Models
- Classes & Methods
- Data

- Model Training Walkthrough

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

https://github.com/iacs-capstone-ieee/ieee_ner_coref

Packaging

```
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)
```

NER Example

https://github.com/iacs-capstone-ieee/ieee_ner_coref#named-entity-recognition-quickstart

Packaging

```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)
```

Coreference Resolution Example

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

Documentation

NER and Coref for IEEE
0.1

Search docs

- Getting Started
  - Requirements
  - Setup
  - Full Documentation
  - Named Entity Recognition Quickstart
  - Coreference Resolution Quickstart
  - Retraining
  - Running Tests
  - Project Organization
- Settings
- Main Classes
- Named Entity Recognition
- Coreference Resolution
- Commands

» Getting Started View page source

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:

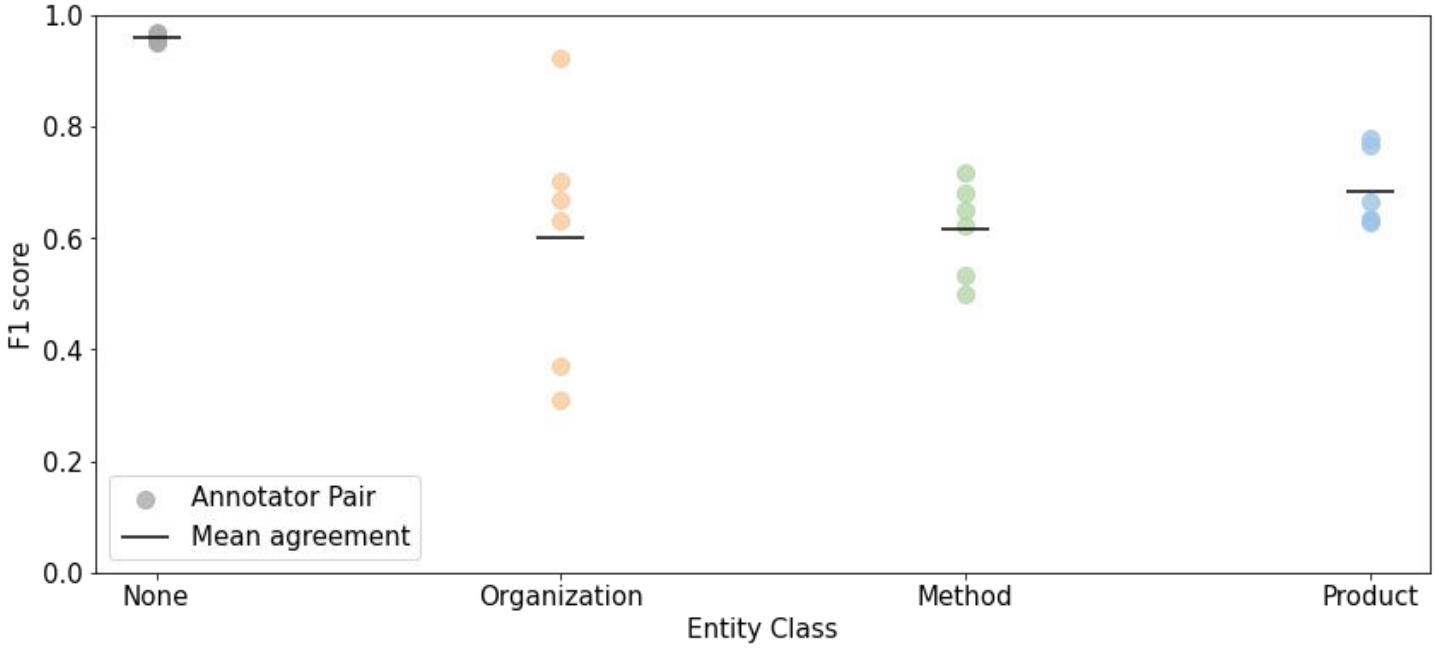
brk.mn/ieee_ner_coref

The Importance of Good Labeled Data

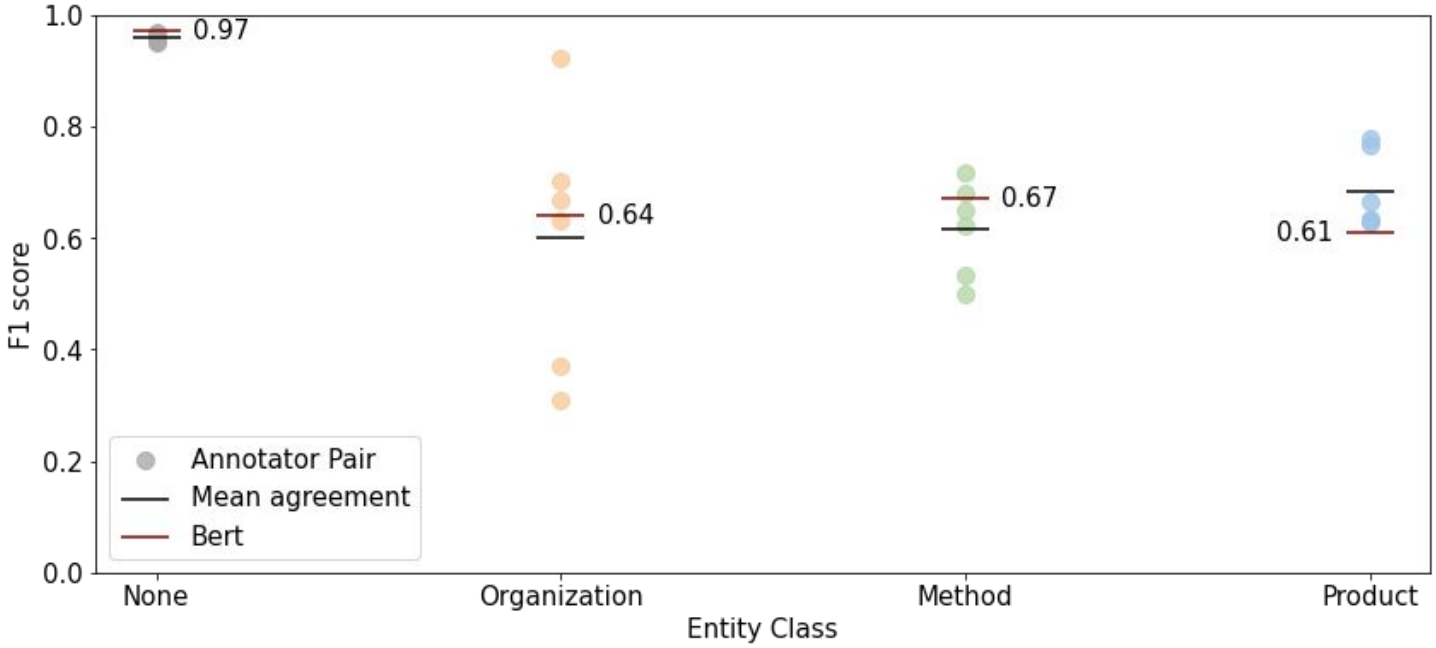
Inter-annotator Agreement (IAA)

"True" annotator	"Model" annotator	F1 Score
A	B	**0.69**
	C	**0.72**
	D	**0.66**
B	C	**0.69**
	D	**0.84**
C	D	**0.69**

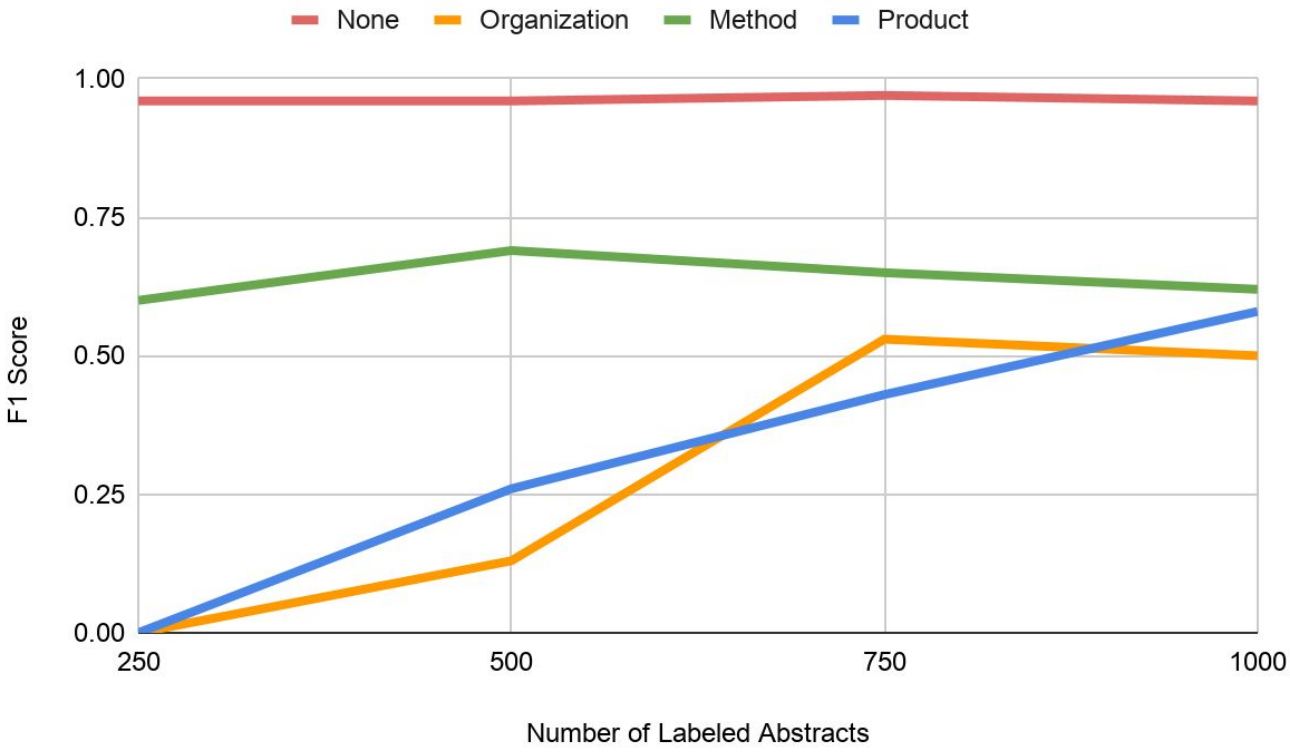
IAA by Entity Class



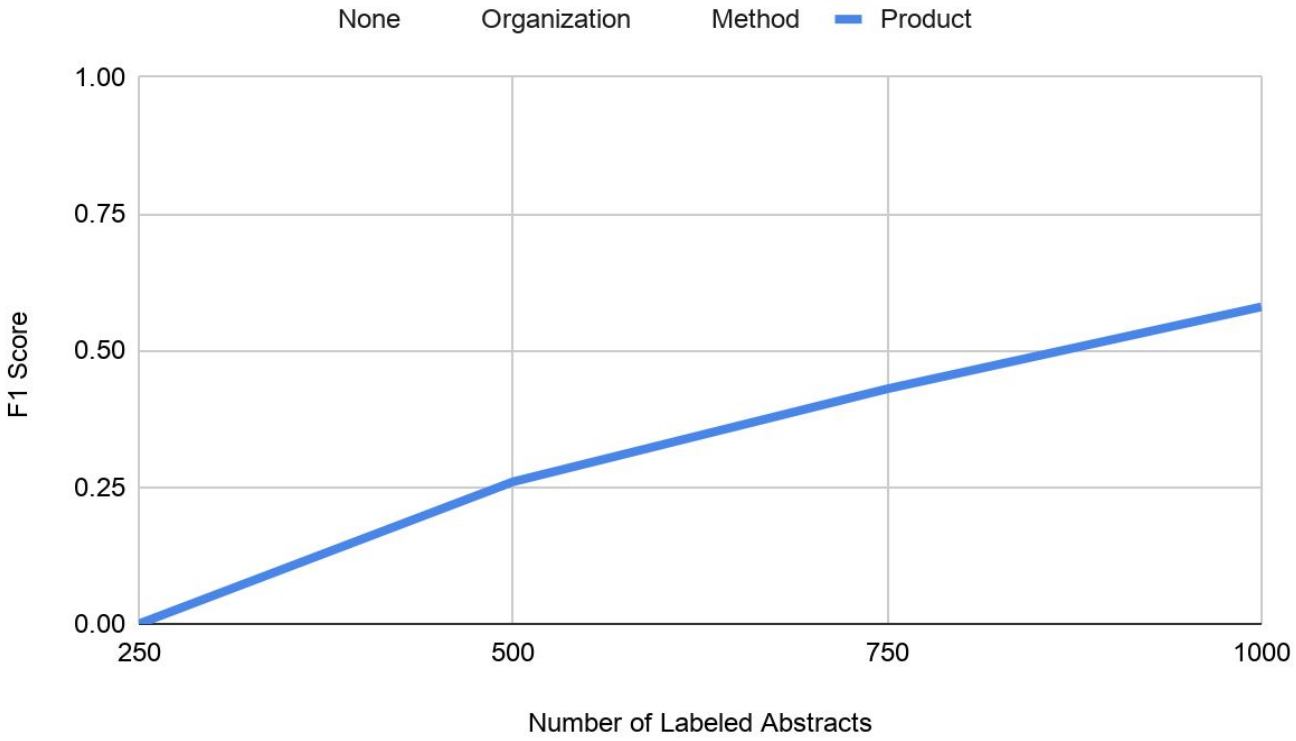
Annotator Pair vs. BERT-Base F1 Scores



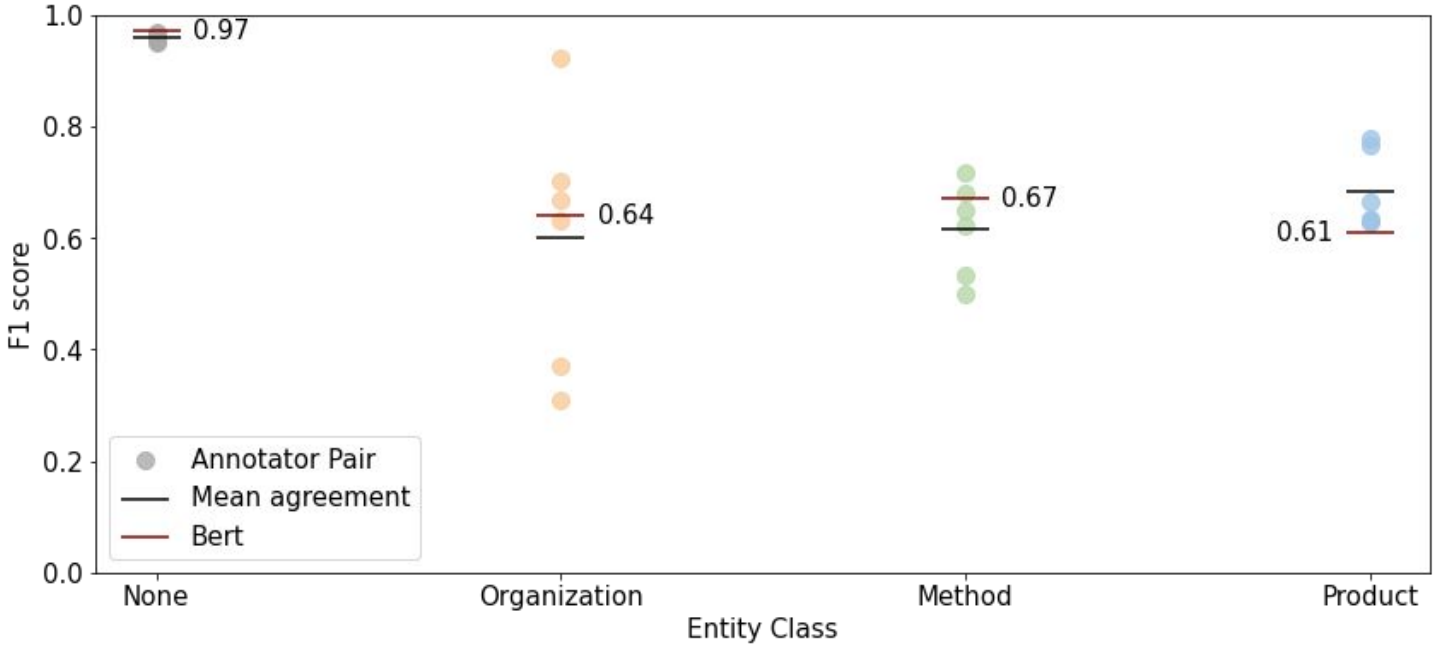
F1 Score vs. Dataset Size



F1 Validation Score vs. Dataset Size



Annotator Pair vs. BERT-Base F1 Scores

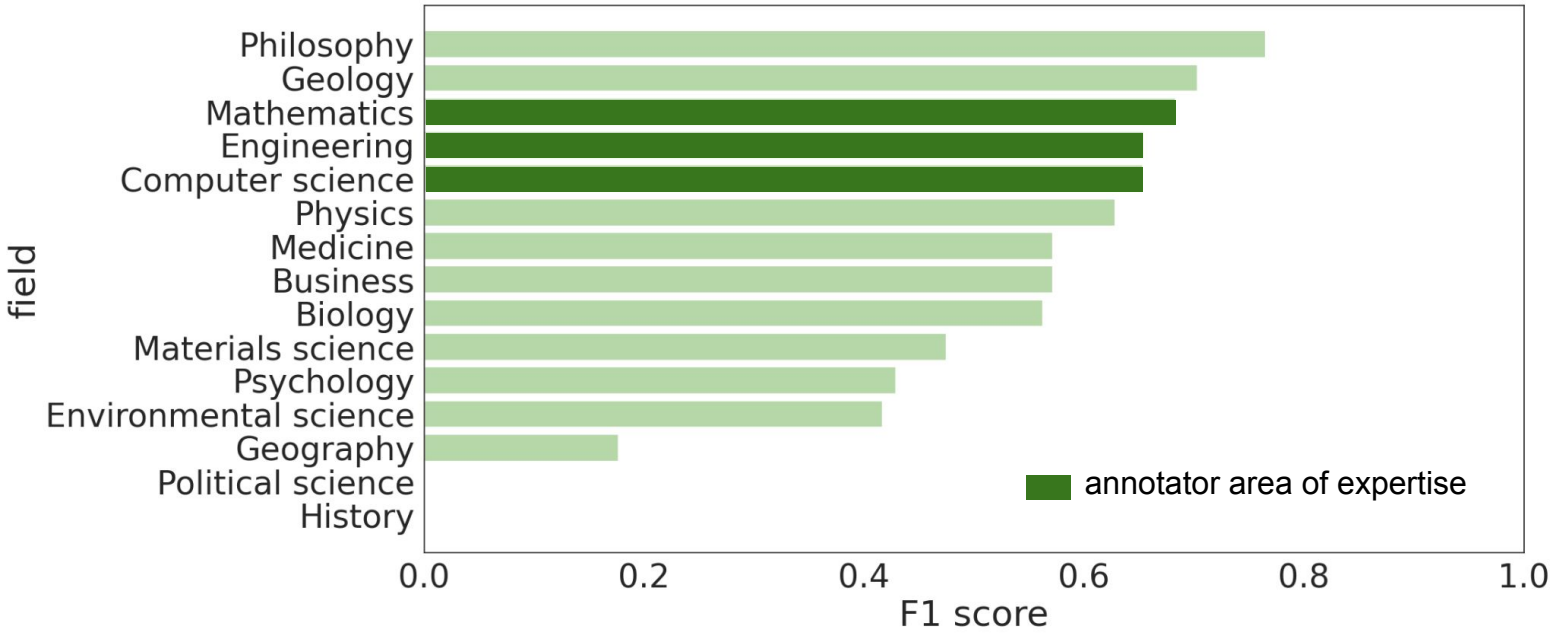


“Good” Labeled Data is...

Consistent

Expert-Driven

BERT-Base F1 Scores by Field of Study



“Good” Labeled Data is...

Consistent

Expert-Driven

The Answer

NER and Coref for IEEE

0.1

Search docs

» Future Directions View page source

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 muddies 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

[1] G. Hripcsak and A. S. Rothschild, “Agreement, the F-Measure, and Reliability in Information Retrieval,” *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, May 2005, doi: 10.1197/jamia.M1733.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” October 2018, arXiV: 1810.04805v2

[3] I. Beltagy, K. Lo, A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615-3620, November 2019, doi: 10.18653/v1/D19-1371

[4] R.J. Bayardo, Y. Ma, R. Srikant, “Scaling Up All Pairs Similarity Search,” *WWW ‘07: Proceedings of the 16th International Conference on World Wide Web,* pp. 131-140, May 2007, doi: 10.1145/1242572.1242591