Knowledge of the Ancestors: Intelligent Ontology-aware Annotation of Biological Literature using Semantic Similarity

PRATIK DEVKOTA[1], DR. SOMYA D. MOHANTY[1], DR. PRASHANTI MANDA[2]

[1] DEPARTMENT OF COMPUTER SCIENCE,

[2] INFORMATICS AND ANALYTICS,

UNIVERSITY OF NORTH CAROLINA AT GREENSBORO

Outline

- Consuming GO ontology to automate annotation of scientific literature
- Span detection and concept normalization
- Data preprocessing
- Model training
- Performance/Results

Can we recognize ontology concepts from text?

Mouse Pachytene Checkpoint 2 (Trip13) Is Required for Completing Meiotic Synapsis

Abstract

In mammalian meiosis, homologous chromosome synapsis is coupled with recombination. As in most eukaryotes, mammalian meiocytes have checkpoints that monitor the fidelity of these processes. We report that the mouse ortholog (Trip13) of pachytene checkpoint 2 (PCH2), an essential component of the synapsis checkpoint in Saccharomyces cerevisiae and Caenorhabditis elegans, is required for completion of meiosis in both sexes. TRIP13-deficient mice exhibit spermatocyte death in pachynema and loss of oocytes around birth. The chromosomes of mutant spermatocytes synapse fully, yet retain several markers of recombination intermediates, including RAD51, BLM, and RPA. These chromosomes also exhibited the chiasmata markers MLH1 and MLH3, and okadaic acid treatment of mutant spermatocytes caused progression to metaphase I with bivalent chromosomes. Double mutant analysis demonstrated that the recombination and synapsis genes Spo11, Mei1, Rec8, and Dmc1 are all epistatic to Trip13, suggesting that TRIP13 does not have meiotic checkpoint function in mice. Our data indicate that TRIP13 is required after strand invasion for completing a subset of recombination events, but possibly not those destined to be crossovers. To our knowledge, this is the first model to separate recombination defects from asynapsis in mammalian meiosis, and provides the first evidence that unrepaired DNA damage alone can trigger the pachytene checkpoint response in mice.

Biological Process Cellular Component

Automated text curation so far ...

Named Entity Recognition

Syntactic Analysis

Machine Learning

Lexical Approaches

Deep Learning

Recent works

Biomedical Concept Recognition Using Deep Neural Sequence Models

Negacy D. Hailu, Michael Bada, Asmelash Teka Hadgu, Lawrence E. Hunter

bioRxiv (2019), DOI: 10.1101/530337

Concept recognition as a machine translation problem

Mayla R Boguslav, Negacy D Hailu, Michael Bada, William A Baumgartner Jr, Lawrence E Hunter

2021 Dec 17, PMID: 34920707, DOI: 10.1186/s12859-021-04141-4

UZH@CRAFT-ST: a Sequence-labelling Approach to Concept Recognition

Lenz Furrer, Joseph Cornelius, Fabio Rinaldi

2019 Nov, DOI: 10.1186/s12859-021-04141-4

GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text

Qile Zhu, Xiaolin Li, Ana Conesa, Cécile Pereira

2018 May 1, PMID: 29272325, DOI: 10.1093/bioinformatics/btx815

Limitations of prior work

Tokens : ‘Mitochondrial’

Ground Truth : **‘GO:0005739’**

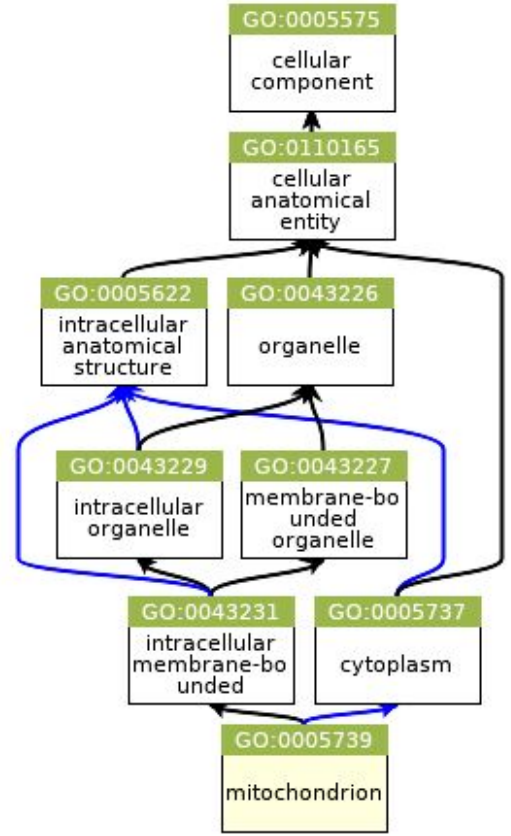
Prediction : **‘GO:0000070’**

‘GO:0043231’

‘GO:0043227’

⋮

‘GO:0005575’



QuickGO - https://www.ebi.ac.uk/QuickGO

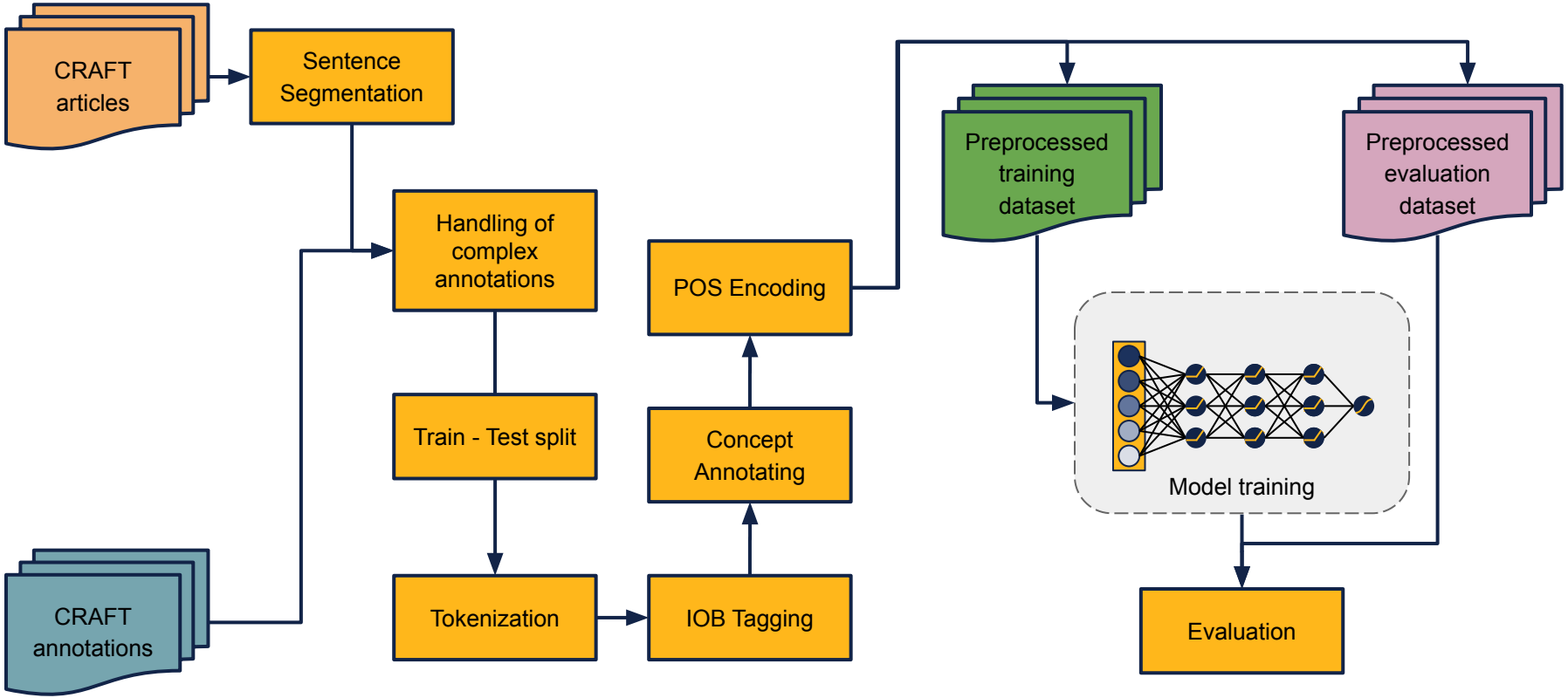
Goal: Develop ontology-aware deep learning architectures for recognizing ontology concepts in text.

Gold standard corpus

CRAFT: THE COLORADO RICHLY ANNOTATED FULL TEXT CORPUS

- 97 articles from the PubMed Central Open Access subset
- 750,479 tokens (34,224 unique tokens)
- 29,015 sentences
- 25,832 concept annotations to Gene Ontology
  - Biological Process (BP)
  - Cellular Component (CC)
  - Molecular Function (MF)

Deep Learning pipelines



IOB format

- Common format for tagging tokens
- Part of span detection
- O represents Outside → not a concept
- B represents Beginning →first word of a phrase
- I represents Inside →all remaining words of the phrase

Our approach: combine span detection and concept normalization in one

If a token is a beginning of a concept and its annotated to ‘GO:X’, we represent the token as ‘B-GO:X’.

Data preprocessing for different annotation formats

- No annotations
- Disjoint annotations
- Overlapping annotations
- Multiple overlapping annotations
- Discontinuous annotations

Data preprocessing

No annotations:

Sentence: Type I fibers are stained dark blue.

Annotations: [ ]

Tokens: ['Type', 'I', 'fibers', 'are', 'stained', 'dark', 'blue', '.']

POS: ['NN', 'NN', 'NNS', 'VBP', 'VBN', 'RB', 'JJ', '.']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'EOS']

Data preprocessing

Disjoint annotations:

Sentence: Well-formed **pedicles** and **spherules** were not evident.

Annotations: 'pedicles' - 'GO:0044316', 'spherules' - 'GO:0044317'

Tokens: ['Well-formed', **'pedicles'**, 'and', **'spherules'**, 'were', 'not', 'evident', '.']

POS: ['JJ', 'NNS', 'CC', 'NNS', 'VBD', 'RB', 'JJ', '.']

IOB Tags: ['O', **'B-GO:0044316'**, 'O', **'B-GO:0044317'**, 'O', 'O', 'O', 'EOS']

Data preprocessing

Overlapping annotations:

Sentence: Having excluded a direct role in **vesicle formation** and membrane fusion,

Annotations: ‘vesicle’ — GO:0031982; ‘vesicle formation’ — GO:0006900

Sentence 1: Having excluded a direct role in **vesicle** and membrane fusion,

Annotations: ‘vesicle’ — GO:0031982

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, 'and', 'membrane', 'fusion', ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0031982'**, 'O', 'O', 'O', 'EOS']

Sentence 2: Having excluded a direct role in **vesicle formation** and membrane fusion,

Annotations: ‘vesicle formation’ — GO:0006900

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, **'formation'**, 'and', 'membrane', 'fusion', ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0006900'**, **'I-GO:0006900'**, 'O', 'O', 'O', 'EOS']

Data preprocessing

Multiple overlapping annotations:

Sentence: Having excluded a direct role in **vesicle formation** and **membrane fusion**,

Annotations: ‘vesicle’ — GO:0031982; ‘vesicle formation’ — GO:0006900; ‘membrane’ — GO:0016020; ‘membrane fusion’ — GO:0061025

Sentence 1: Having excluded a direct role in **vesicle** and **membrane**,

Annotations: ‘vesicle’ — GO:0031982; ‘membrane’ — GO:0016020

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, 'and', **'membrane'**, ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0031982'**, 'O', **'B-GO:0016020'**, 'EOS']

Sentence 2: Having excluded a direct role in **vesicle formation** and **membrane**,

Annotations: ‘vesicle formation’ — GO:0006900; ‘membrane’ — GO:0016020

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, **'formation'**, 'and', **'membrane'**, ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0006900'**, **'I-GO:0006900'**, 'O', **'B-GO:0016020'**, 'EOS']

Data preprocessing

Multiple overlapping annotations:

Sentence: Having excluded a direct role in **vesicle formation** and **membrane fusion**,

Annotations: ‘vesicle’ — GO:0031982; ‘vesicle formation’ — GO:0006900; ‘membrane’ — GO:0016020; ‘membrane fusion’ — GO:0061025

Sentence 3: Having excluded a direct role in **vesicle** and **membrane fusion**,

Annotations: ‘vesicle’ — GO:0031982; ‘membrane fusion’ — GO:0061025

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, 'and', **'membrane'**, **'fusion'**, ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0031982'**, 'O', **'B-GO:0061025'**, **'I-GO:0061025'**, 'EOS']

Sentence 4: Having excluded a direct role in **vesicle formation** and **membrane fusion**,

Annotations: ‘vesicle formation’ — GO:0006900; ‘membrane’ — GO:0016020

Tokens: ['Having', 'excluded', 'a', 'direct', 'role', 'in', **'vesicle'**, **'formation'**, 'and', **'membrane'**, **'fusion'**, ',']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0006900'**, **'I-GO:0006900'**, 'O', **'B-GO:0061025'**, **'I-GO:0061025'**, 'EOS']

Data preprocessing

Discontinuous annotations:

Sentence: The difference between the heart and kidney levels is due to a development delay in **v**/p **formation**.

Annotations: ‘v formation’ — GO:0097084

Transformed

Sentence: The difference between the heart and kidney levels is due to a development delay in **v formation**.

Annotations: ‘v formation’ — GO:0097084

Tokens: ['The', 'difference', 'between', 'the', 'heart', 'and', 'kidney', 'levels', 'is', 'due', 'to', 'a', 'development', 'delay', 'in', **'v'**, **'formation'**, '.']

IOB Tags: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', **'B-GO:0097084'**, **'I-GO:0097084'**, 'EOS']

Results from prior work

Model	Embeddings	F1	Jaccard	Top two F1	Top two Jaccard
LSTM	CRAFT	0.75	0.69	0.82	0.79
	PubMed + PMC	0.69	0.65	0.77	0.75
	GloVe	0.64	0.62	0.73	0.72
	ELMo	0.75	0.76	0.82	0.84
GRU	CRAFT	0.79	0.69	0.85	0.81
	PubMed + PMC	0.68	0.64	0.77	0.75
	GloVe	0.68	0.64	0.79	0.75
	ELMo	**0.78**	**0.78**	**0.84**	**0.85**

https://dl.acm.org/doi/10.1145/3391274.3393636

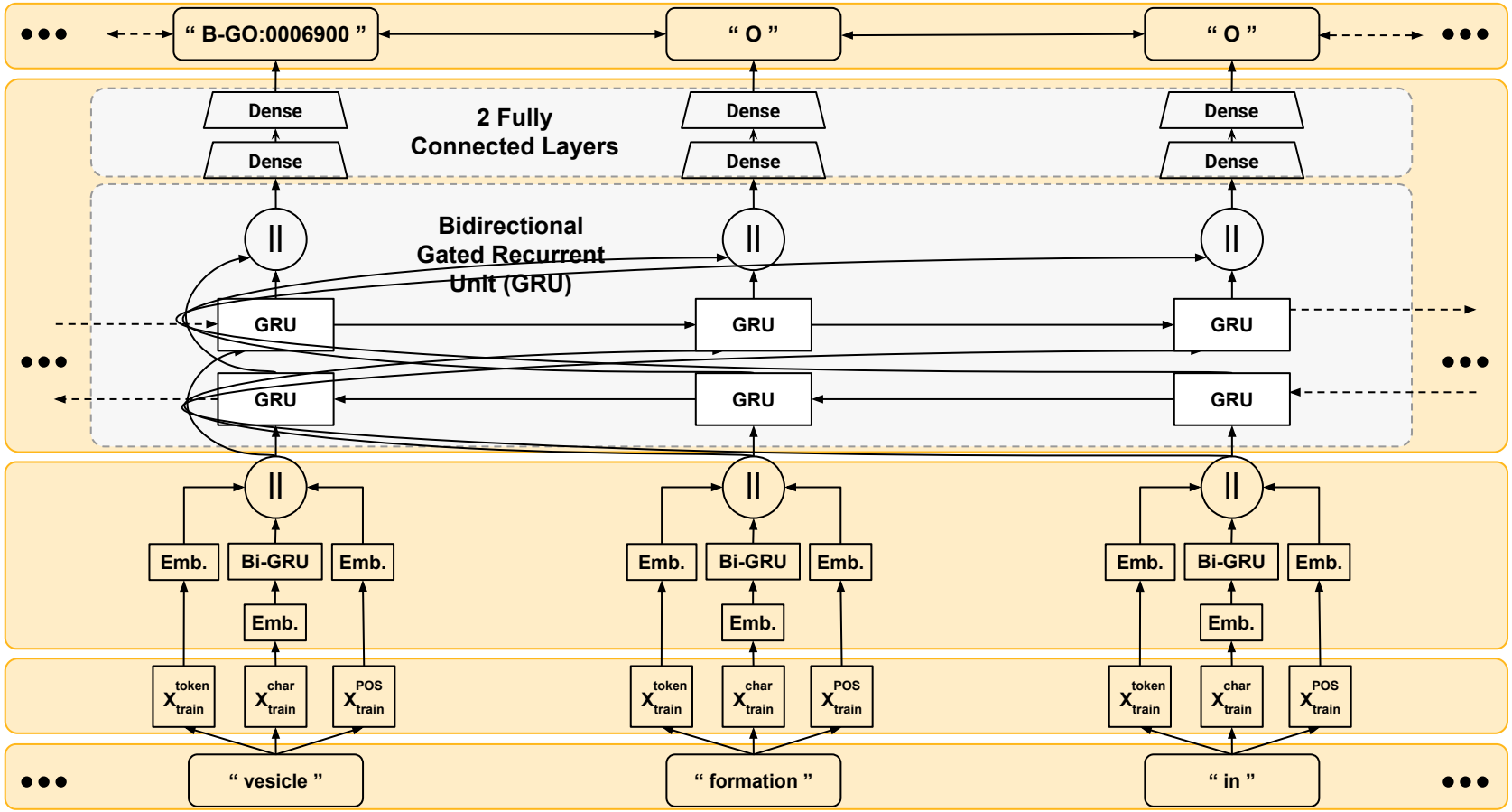
Deep learning algorithm

- Gated Recurrent Unit (GRU)
- Bidirectional Encoder Representations from Transformers (BERT)

Deep learning encoding formats

- CRAFT
- GloVe
- ELMo

Model architecture



Output format

True Label:

	B-GO:0000226	B-GO:0006996	O
B-GO:0000226	[1 ,	0 ,	0]
B-GO:0006996	[0 ,	1 ,	0]
O	[0,	0 ,	1]

Predicted Label:

	B-GO:0000226	B-GO:0006996	O
B-GO:0000226	[0.885 ,	0.098 ,	0.017]
B-GO:0006996	[0.213 ,	0.744 ,	0.043]
O	[0.052,	0.038 ,	0.920]

Target vector representation

Assume that there are only 2 GO concepts:

“GO:0000226”, “GO:0006996”, and “O”.

0 1 2

If our **ground truth** for a sequence is:

[“GO:0000226”, “GO:0006996”, “O”]

General representation:

[[1.0 , 0.0 , 0.0] ,

[0.0 , 1.0 , 0.0] ,

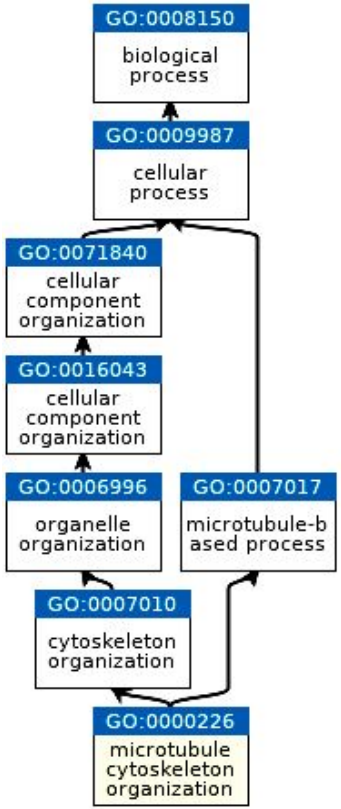
[0.0 , 0.0 , 1.0]]

Ontology aware representation:

[[1.0 , 0.625 , 0.0] ,

[0.625 , 1.0 , 0.0] ,

[0.0 , 0.0 , 1.0]]



QuickGO - https://www.ebi.ac.uk/QuickGO

Target vector representation

Assume that there are only 4 GO concepts: “GO:0000226”, “GO:0016043”, “GO:0006996”, “GO:0016740” and “O”.

0 1 2 3 4

In a general one-hot encoded vector, our ground truth for

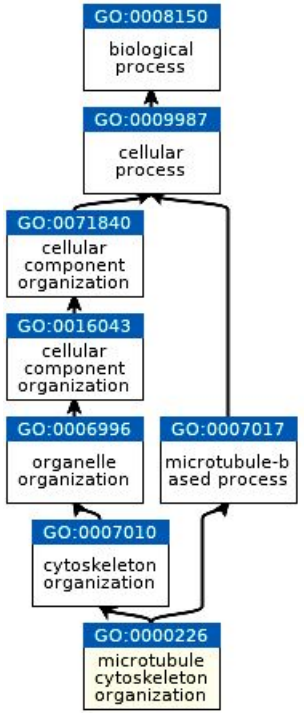
[“GO:0000226”, “GO:0016043”, “GO:0006996”, “GO:0016740” and “O”]

would look like:

With our approach, the ground truth appears like:

```
[ [1,  0,  0,  0,  0],
  [0,  1,  0,  0,  0],
  [0,  0,  1,  0,  0],
  [0,  0,  0,  1,  0],
  [0,  0,  0,  0,  1] ]
```

```
[ [1.0,   0.5,   0.625,  0.0,  0.0],
  [0.5,   1.0,   0.8,    0.0,  0.0],
  [0.625, 0.8,   1.0,    0.0,  0.0],
  [0.0,   0.0,   0.0,    1.0,  0.0],
  [0.0,   0.0,   0.0,    0.0,  1.0] ]
```



Performance evaluation metrics

- Precision
- Recall
- F1 score
- Jaccard semantic similarity

Performance evaluation

Model	Embeddings	F1	Jaccard	Top two F1	Top two Jaccard
Baseline	CRAFT	0.74	0.75	0.82	0.86
	GloVe	0.75	0.76	0.83	0.87
	ELMo	0.79	0.82	0.86	0.90
Ontology aware model	CRAFT	0.80	0.83	0.86	0.91
	GloVe	0.79	0.82	0.86	0.90
	ELMo	0.81	0.84	0.87	0.92

Model	F1	Jaccard
BERT	0.77	0.80
Ontology aware model (ELMo)	0.81	0.84

Future work

- Augmentation from biological sources
  - Bio-Thesaurus, Unified Medical Language System (UMLS)
- Synonymization
  - Using synonyms of under-represented (lower frequency) concepts
- Boosting
  - Boost the probability of a term by taking its subsumer’s probability into consideration

Data and code availability

The **data** used in this work is publicly available at:

https://github.com/UCDenver-ccp/CRAFT/releases/tag/v4.0.1

The **source code** used to generate the results can be found at:

https://github.com/prashanti/intelligentannotation

The **source code** is also archived on Zenodo:

DOI: 10.5281/zenodo.6964353

Acknowledgment

This work is funded by a CAREER grant to Dr. Prashanti Manda from the Division of Biological Infrastructure at the National Science Foundation (#1942727).

Thank You !

p_devkota@uncg.edu