Using ontology embeddings with deep learning architectures to improve prediction of ontology concepts from literature

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

[1] INFORMATICS AND ANALYTICS,

UNIVERSITY OF NORTH CAROLINA AT GREENSBORO

[2] UNITED HEALTHCARE

Outline

- Automation of ontology annotation of scientific literature
- Deep Learning for Named Entity Recognition
- Deep Learning for Ontology Embeddings
- Information augmentation with ontology embeddings
- Performance/Results
- Discussion

Recent works

A Gated Recurrent Unit based architecture for recognizing ontology concepts from biological literatures

Pratik Devkota, Somya D. Mohanty, Prashanti Manda

2022, DOI: 10.1186/s13040-022-00310-0

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

Pratik Devkota, Somya Mohanty, Prashanti Manda

2022

Ontology-powered Boosting for Improved Recognition of Ontology concepts from Biological literatures

Pratik Devkota, Somya Mohanty, Prashanti Manda

2023, DOI: 10.5220/0011683200003414

Goal: Develop deep learning architectures that capture context from **both** scientific literatures and Gene Ontology structures using **embeddings**.

Methodology

Two-step process:

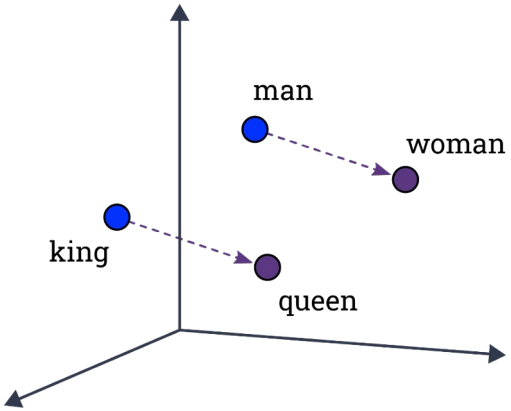
1. Compute **embeddings** for all Gene Ontology concepts
2. Train **deep learning models** with the information from the training dataset as well as semantic relationship from ontology hierarchy.

Methodology

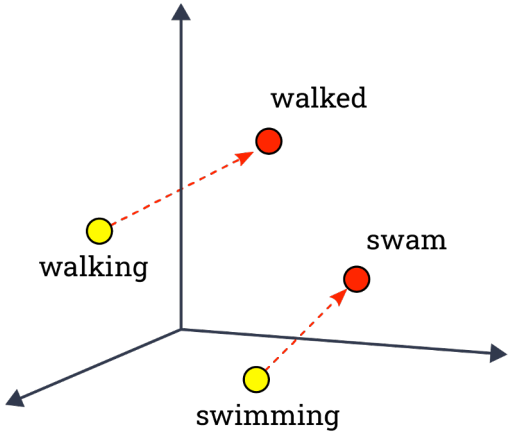
Step 1: Compute **embeddings** for Gene Ontology (GO) concepts.

Embeddings

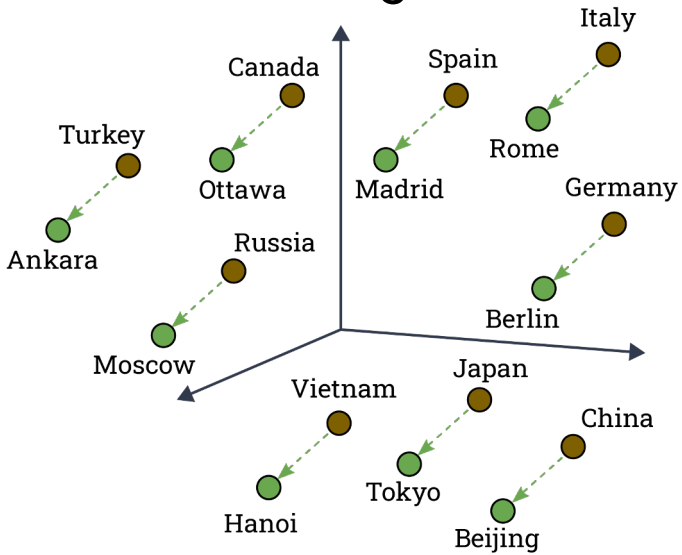
Embeddings is the concept of representing texts and words as vectors of numbers that capture their semantics or meaning.



Male-Female

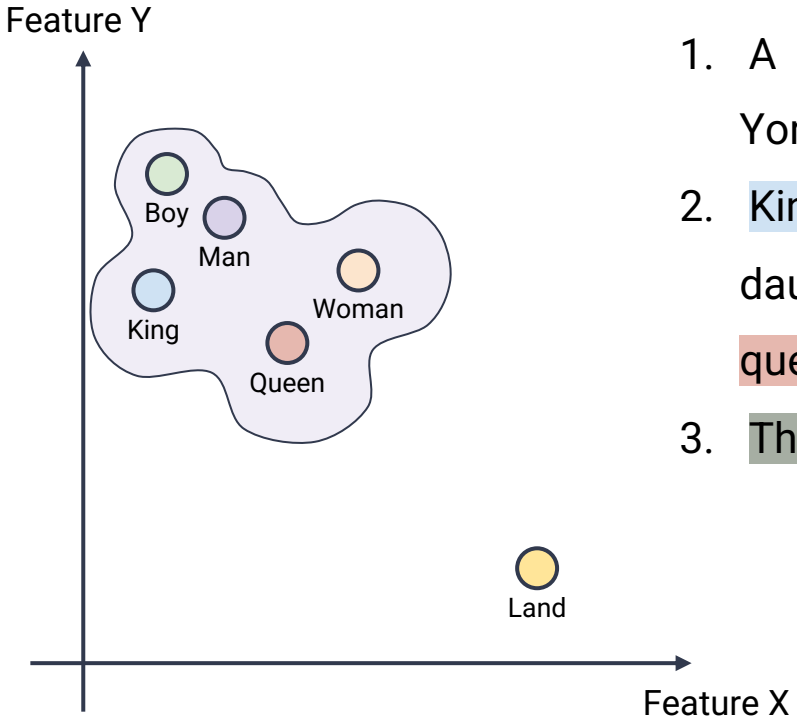


Verb Tense



Country-Capital

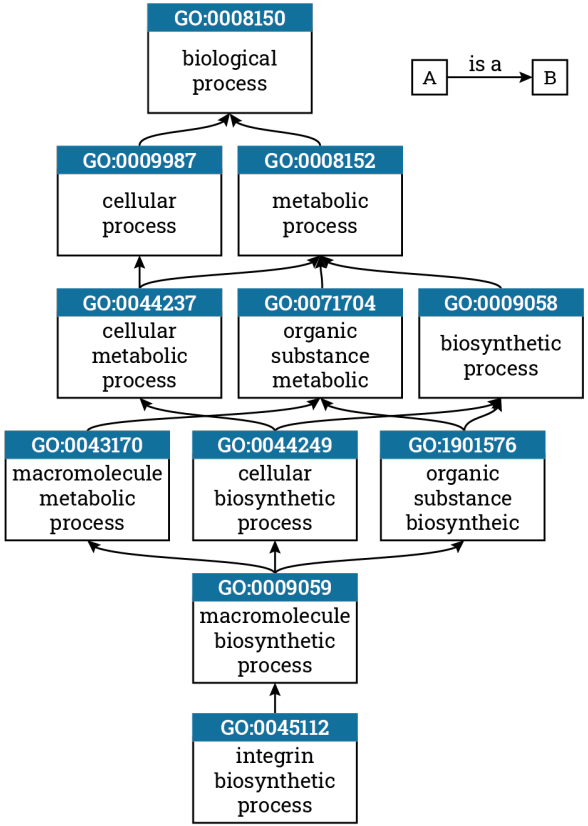
Embeddings from context



Projection in 2D plane

1. A man from Chicago married a woman from New York.
2. King Aldric, of Valeria married Princess Elara, daughter of King Adrian of Lunaria. Elara is now the queen of Valeria.
3. They gave birth to a beautiful baby boy, Prince Cedric.

Embeddings from Gene Ontology

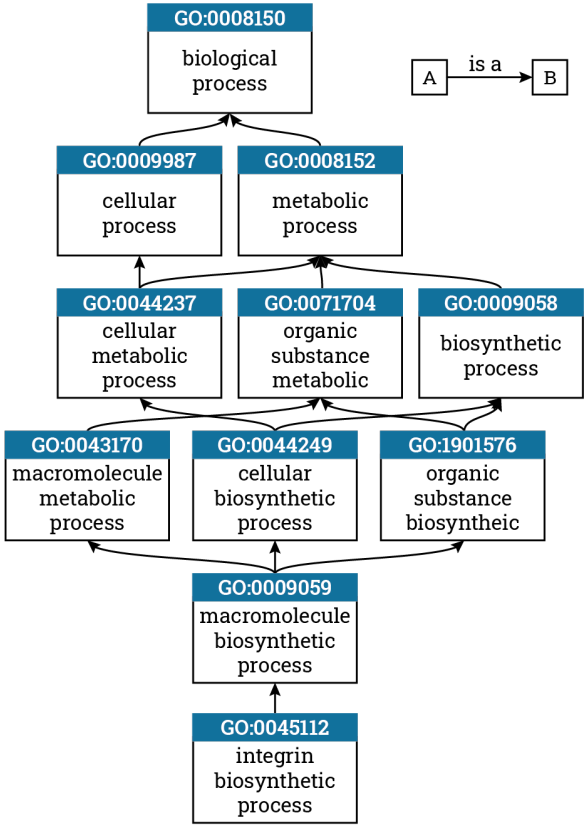


QuickGO - https://www.ebi.ac.uk/QuickGO

Node2Vec algorithm:

1. Use biased random walks to generate sequence of ontology concepts.
2. Use the generated sequences as input to deep learning algorithm (word2vec) for the generation of embedding vectors.

Embeddings from Gene Ontology

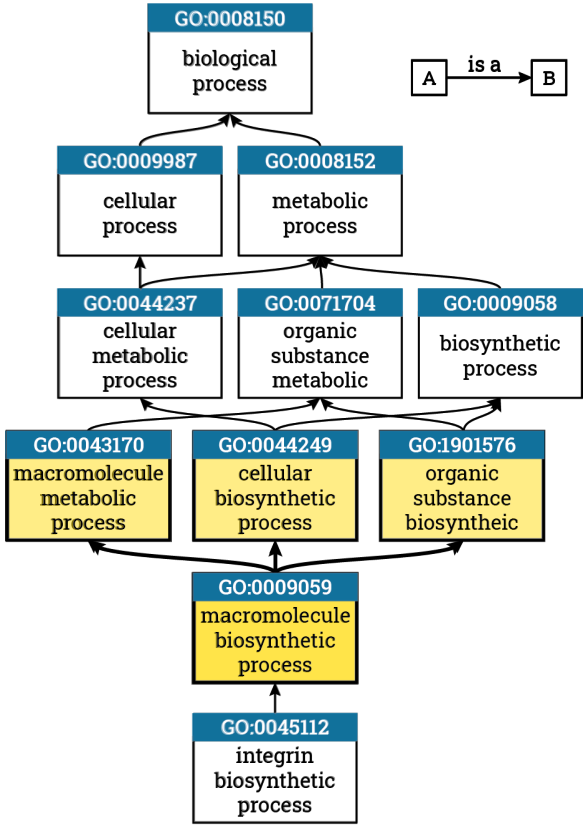


QuickGO - https://www.ebi.ac.uk/QuickGO

Random Walk parameters:

1. walk length ⇒ # nodes to explore
2. walk number ⇒ # samples
3. p ⇒ probability, 1/p, of returning to source
4. q ⇒ probability, 1/q, of moving further away from source node

Embeddings from Gene Ontology



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

Random Walk parameters:

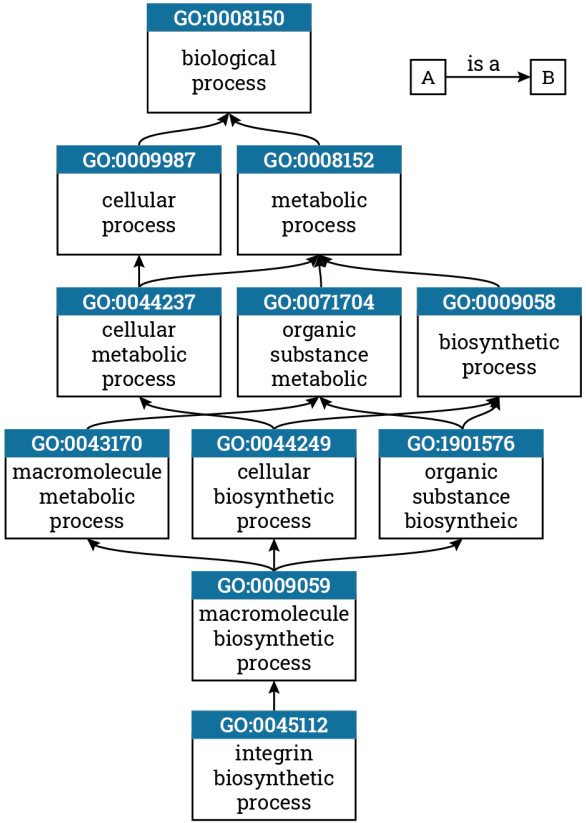
1. walk length ⇒ # nodes to explore
2. walk number ⇒ # samples
3. p ⇒ probability, 1/p, of returning to source
4. q ⇒ probability, 1/q, of moving further away from source node

macromolecule biosynthetic process is a organic substance biosynthetic

macromolecule biosynthetic process is a macromolecule metabolic process

macromolecule biosynthetic process is a cellular biosynthetic process

Embeddings from Gene Ontology



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

Word2Vec:

Batch size: 50

Epochs: 3

GO:0045112	-1.01	0.59	-0.13	...	0.35	0.38
GO:0009059	-1.11	0.91	0.16	...	0.18	0.01
GO:0043170	-0.82	0.72	-0.32	...	0.08	0.18
GO:0044237	-0.86	0.49	-0.06	...	0.37	-0.09
GO:0009987	-0.75	0.36	-0.53	...	-0.17	0.84
GO:0008150	-0.56	0.80	-0.81	...	0.03	0.55

Embeddings in 128 dimensions

Methodology

Step 2: Train **deep learning models**.

Training dataset

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)

Data Preprocessing

1. Each sentence in the article is an **input sequence.** The sequence is broken down as list of words called tokens.

Sentence: Well formed pedicles and spherules were not evident.

Tokens:

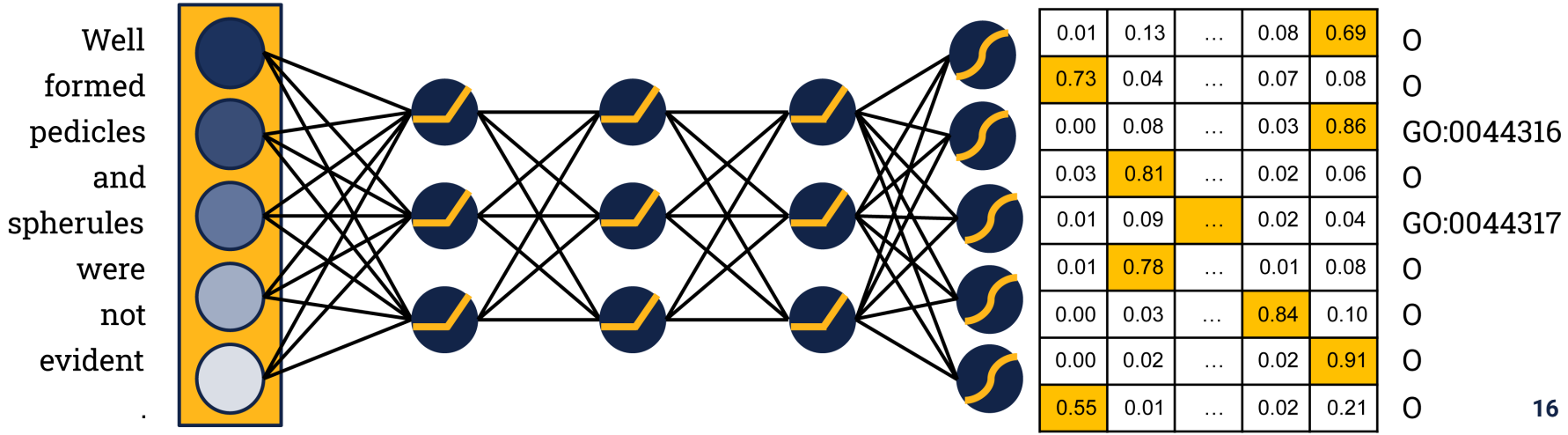
[	Well	formed	Pedicles	and	spherules	were	not	evident	.	]

Model Training

2. For each token, we specify whether it represents a concept or not.

Sentence: Well formed pedicles and spherules were not evident.

Tokens:	[	Well	formed	**pedicles**	and	**spherules**	were	not	evident	.	]
Outputs:	[	O	O	**GO:0044316**	O	**GO:0044317**	O	O	O	O	]

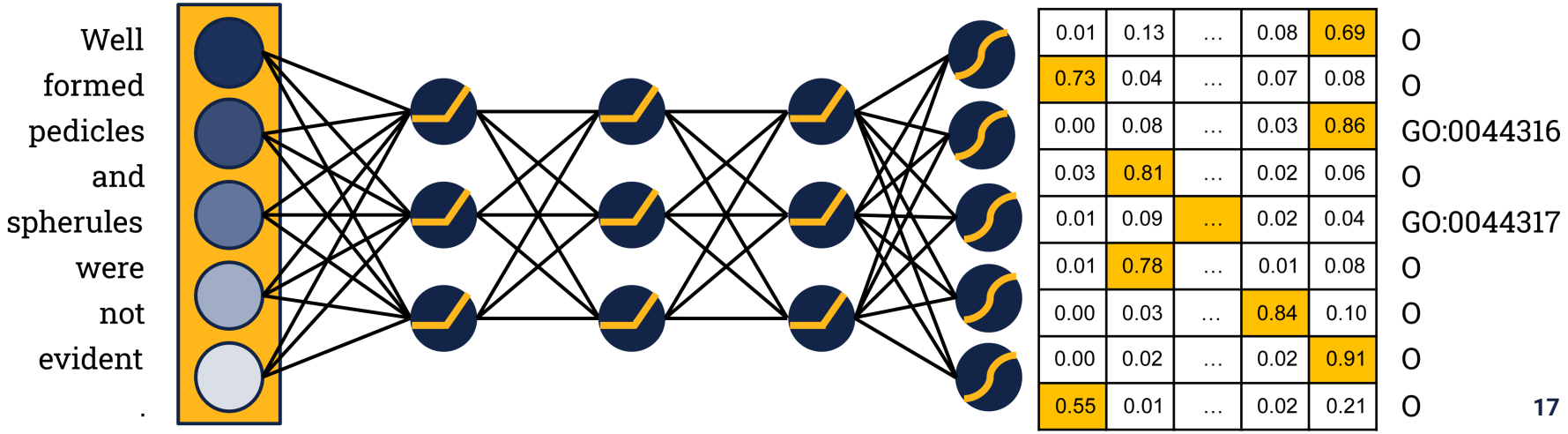


Model Training

2. For each token, we specify whether it represents a concept or not.

Sentence: Well formed pedicles and spherules were not evident.

Tokens:	[	Well	formed	**pedicles**	and	**spherules**	were	not	evident	.	]
Outputs:	[	O	O	**GO:0044316**	O	**GO:0044317**	O	O	O	O	]

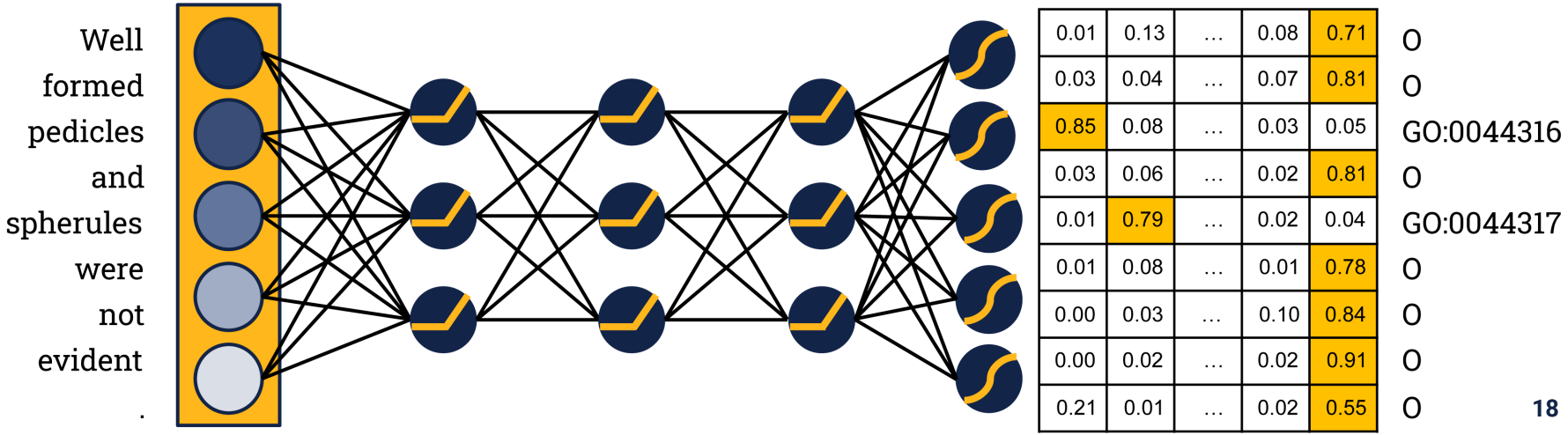


Model Training

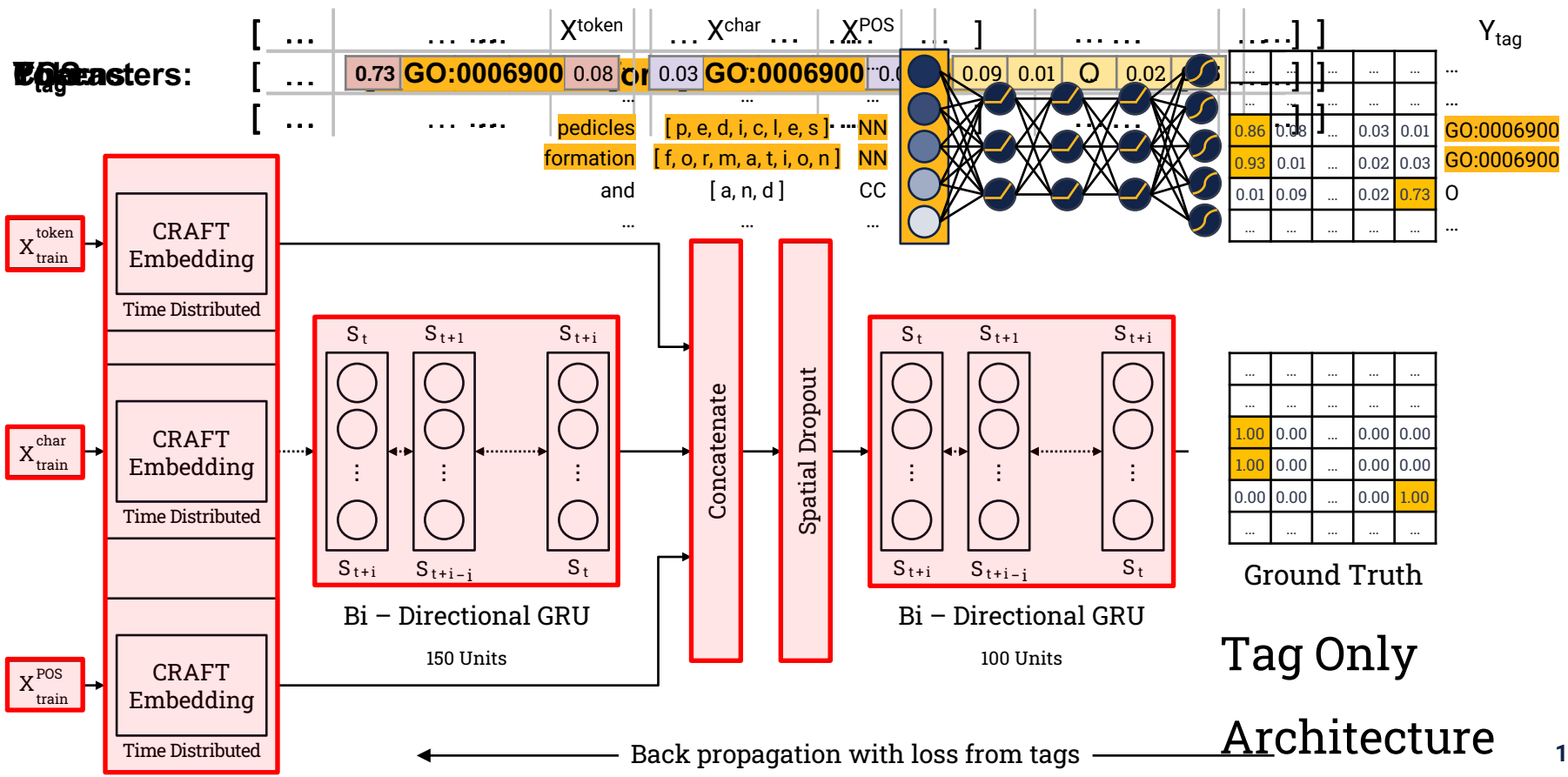
2. For each token, we specify whether it represents a concept or not.

Sentence: Well formed pedicles and spherules were not evident.

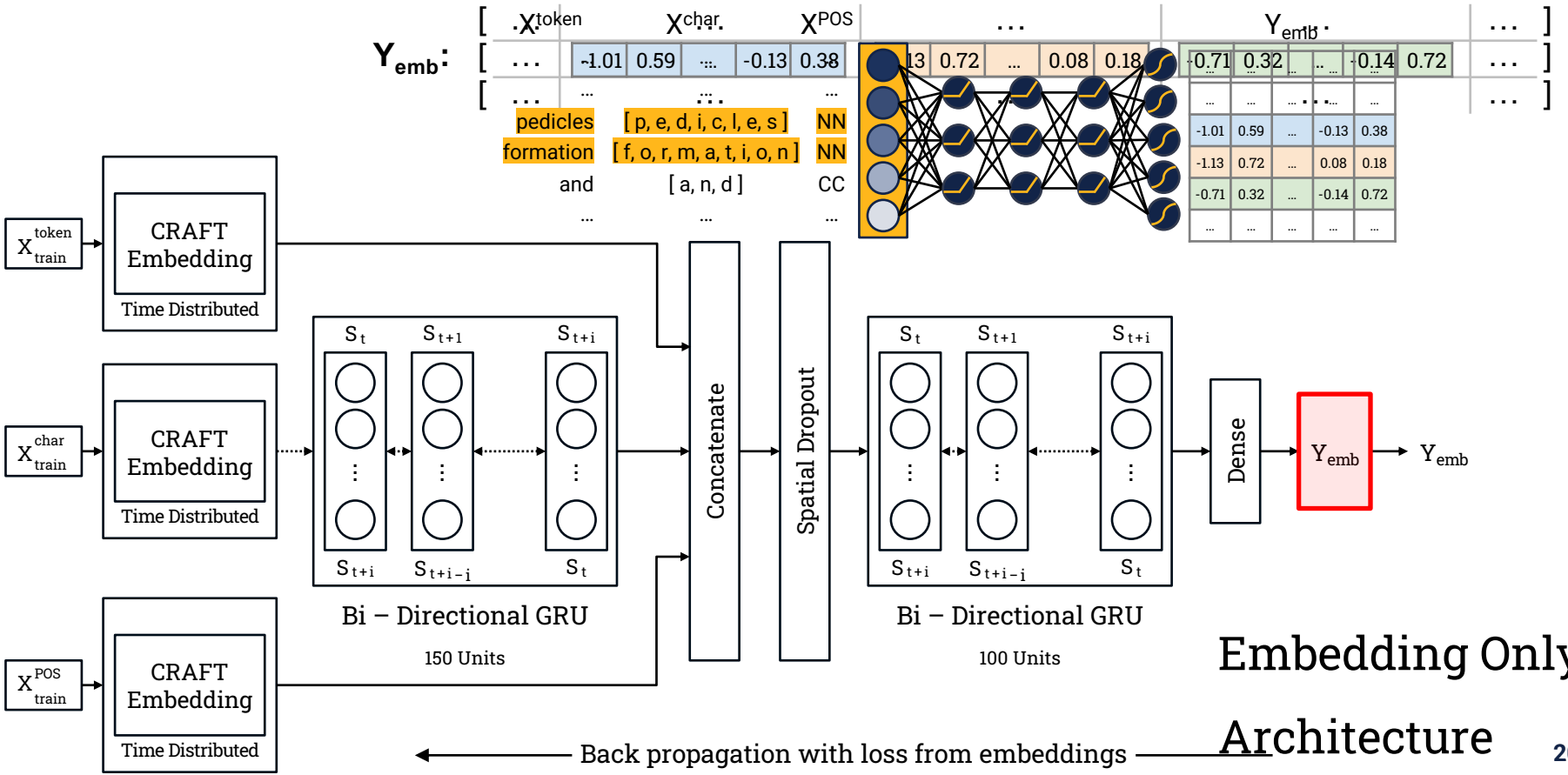
Tokens:	[	Well	formed	**pedicles**	and	**spherules**	were	not	evident	.	]
Outputs:	[	O	O	**GO:0044316**	O	**GO:0044317**	O	O	O	O	]



Baseline Model Architecture



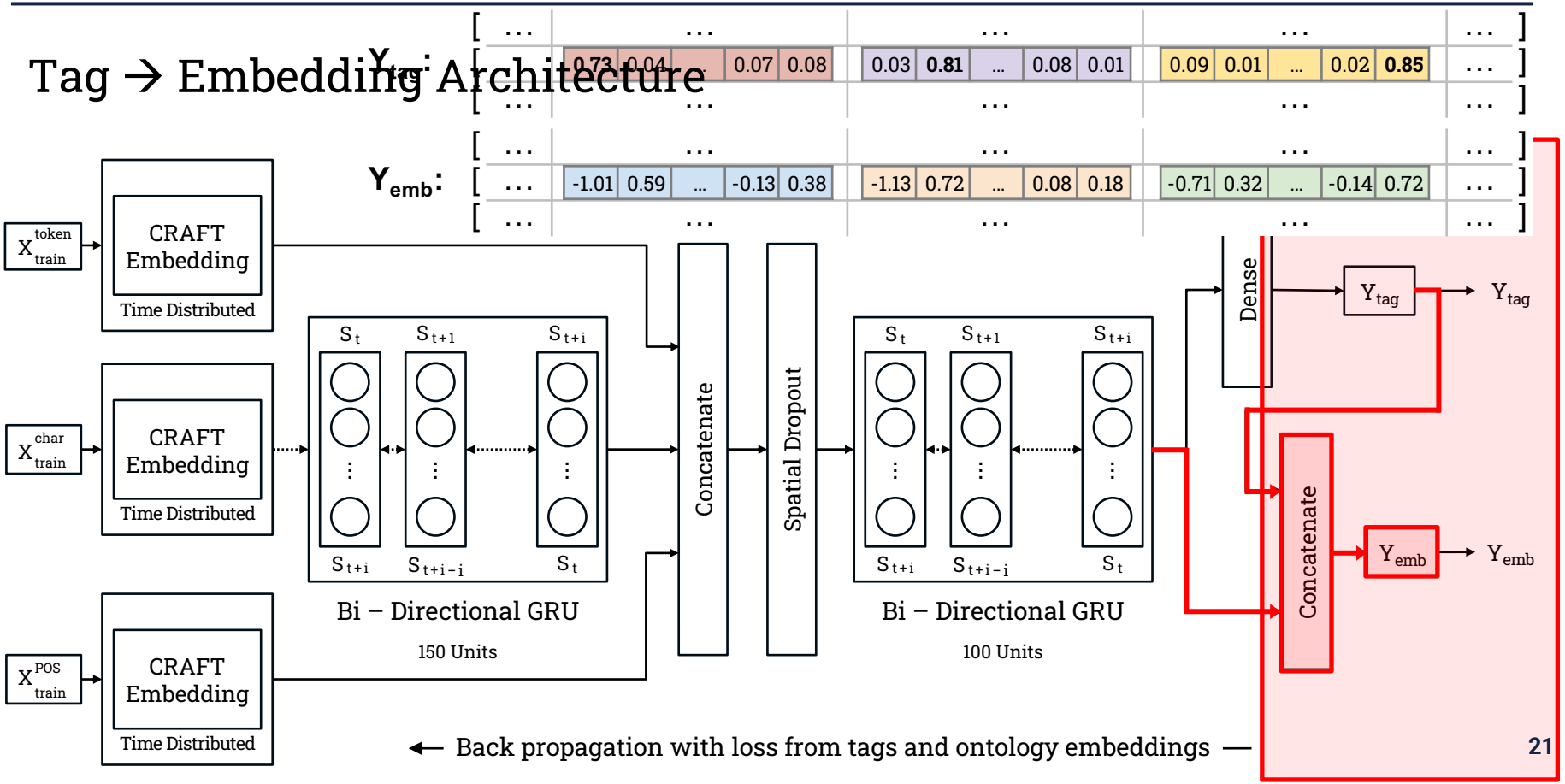
Baseline Model Architecture



Embedding Only Architecture

Cross Connected Model Architecture

Tag → Embedding Architecture



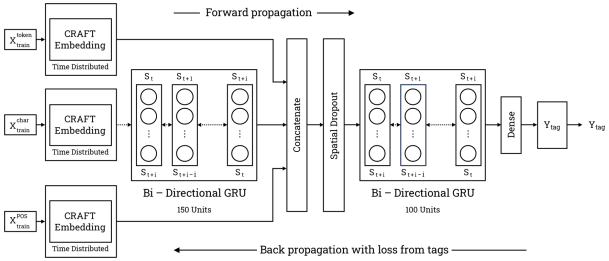
Performance evaluation metrics

- Precision
- Recall
- Modified F1 score
- Jaccard semantic similarity

Model's performance

Architecture	Ontology Embedding F1 Score	Ontology Embedding Similarity Score	Tag F1 Score	Tag Similarity Score
Baseline Architectures				
Tag – Only (TO)	—	—	0.80	0.83
Ontology Embedding Only (OEO)	0.65	0.74	—	—
Cross – connected Architectures				
Tag to Ontology Embedding (T → OE)	**0.80**	**0.81**	**0.83**	**0.84**
Ontology Embedding to Tag (OE → T)	0.64	0.75	0.83	0.84
Multi – connected Architectures				
OE → T → OE	0.78	0.80	0.82	0.83

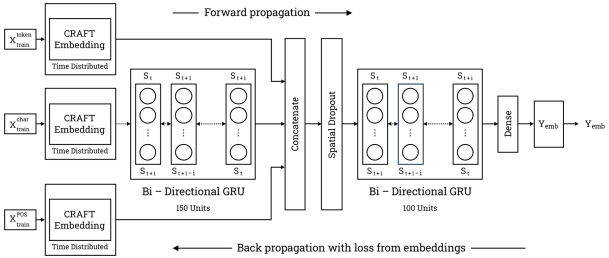
Discussion



Baseline Tag Only Architecture			
Onto Emb F1: –	Onto Emb Sem: –	Tag F1: 0.80	Tag Sem: 0.83

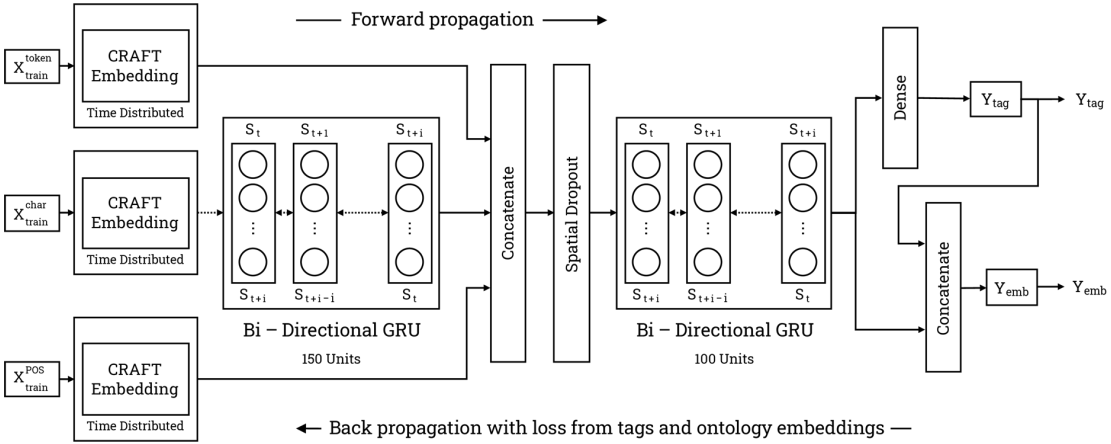
Good Performance but limited predictability

Can only predict 1000/47000 GO concepts



Baseline Embedding Only Architecture			
Onto Emb F1: 0.65	Onto Emb Sem: 0.74	Tag F1: –	Tag Sem: –

Higher predictability but with poor performance



Cross Connected Tag to Embedding Architecture			
Onto Emb F1: 0.80 (23%▲)	Onto Emb Sem: 0.81 (9.4%▲)	Tag F1: 0.83 (3.8%▲)	Tag Sem: 0.84 (1.2%▲)

Improved performance and higher predictability

Future Works

Employing Large Language Models (LLMs) for:

- Improved prediction of ontology annotations
- Implicit understanding of how ontologies are structured

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 !

CONTACT ME

p_devkota@uncg.edu