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

PRATIK DEVKOTA¹, SOMYA D. MOHANTY², PRASHANTI MANDA¹

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.
Step 1: Compute **embeddings** for Gene Ontology (GO) concepts.
**Embeddings** is the concept of representing texts and words as vectors of numbers that capture their semantics or meaning.
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.
Mouse Genome Informatics

**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.

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

Random Walk parameters:

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

macromolecule biosynthetic process is a organic substance biosynthetic process
macromolecule biosynthetic process is a macromolecule metabolic process
macromolecule biosynthetic process is a cellular biosynthetic process
Embeddings from Gene Ontology

Word2Vec:

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

Batch size: 50
Epochs: 3

Embeddings in 128 dimensions

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

[Diagram showing the model training process with relevant tokens highlighted and corresponding outputs.]
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 ]
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

<table>
<thead>
<tr>
<th>Tokens:</th>
<th>Characters:</th>
<th>POS:</th>
<th>Y tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ...</td>
<td>...</td>
<td>]</td>
<td>[ ...</td>
</tr>
<tr>
<td>[ ...</td>
<td>...</td>
<td>]</td>
<td>[ ...</td>
</tr>
<tr>
<td>[ ...</td>
<td>...</td>
<td>]</td>
<td>[ ...</td>
</tr>
</tbody>
</table>

CRAFT Embedding

Time Distributed

Bi – Directional GRU

150 Units

Spatial Dropout

Concatenate

Bi – Directional GRU

100 Units

Back propagation with loss from tags

Ground Truth

Tag Only

Architecture
Baseline Model Architecture

- **CRAFT Embedding**
  - **X_token**
    - Time Distributed
  - **X_char**
    - Time Distributed
  - **X_POS**
    - Time Distributed

- **Bi-Directional GRU**
  - **S_t**
  - **S_{t+1}**
  - **S_{t+i}**
  - **S_{t+i-1}**
  - 150 Units

- **Spatial Dropout**

- **Deep**

- **Y_emb**

- **Embedding Only Architecture**

- **Back propagation with loss from embeddings**

### Table

<table>
<thead>
<tr>
<th>pedicles formation</th>
<th>[p, e, d, i, c, l, e, s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>and</td>
<td>[a, n, d]</td>
</tr>
<tr>
<td>-0.13</td>
<td>0.38</td>
</tr>
<tr>
<td>-1.13</td>
<td>0.72</td>
</tr>
<tr>
<td>-0.71</td>
<td>0.32</td>
</tr>
<tr>
<td>-0.14</td>
<td>0.72</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **Y_emb**

- **Yemb**

- **Yemb**

- **Baseline Model Architecture**

- **20**
Cross Connected Model Architecture

Tag → Embedding Architecture

\[ \mathbf{Y}_{\text{emb}} \]

\[ \begin{array}{cccccccccccc}
0.73 & 0.04 & 0.07 & 0.08 & 0.03 & 0.81 & \ldots & 0.08 & 0.01 & 0.09 & 0.01 & \ldots & \ldots & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
-1.01 & 0.59 & \ldots & -0.13 & 0.38 & -1.13 & 0.72 & \ldots & 0.08 & 0.18 & -0.71 & 0.32 & \ldots & \ldots & \ldots \\
\end{array} \]

\[ \mathbf{X}_{\text{train}} \]

\[ \mathbf{X}_{\text{token}} \]

\[ \mathbf{X}_{\text{char}} \]

\[ \mathbf{X}_{\text{POS}} \]

\[ \mathbf{Y}_{\text{tag}} \]

\[ \mathbf{Y}_{\text{emb}} \]

Bi - Directional GRU

150 Units

100 Units

Back propagation with loss from tags and ontology embeddings
Performance evaluation metrics

• Precision
• Recall
• Modified F1 score
• Jaccard semantic similarity
## Model’s performance

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Ontology Embedding F1 Score</th>
<th>Ontology Embedding Similarity Score</th>
<th>Tag F1 Score</th>
<th>Tag Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline Architectures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag – Only (TO)</td>
<td>—</td>
<td>—</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>Ontology Embedding Only (OEO)</td>
<td>0.65</td>
<td>0.74</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Cross – connected Architectures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag to Ontology Embedding (T → OE)</td>
<td><strong>0.80</strong></td>
<td><strong>0.81</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.84</strong></td>
</tr>
<tr>
<td>Ontology Embedding to Tag (OE → T)</td>
<td>0.64</td>
<td>0.75</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Multi – connected Architectures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OE → T → OE</td>
<td>0.78</td>
<td>0.80</td>
<td>0.82</td>
<td>0.83</td>
</tr>
</tbody>
</table>
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!

p_devkota@uncg.edu