Background

- Inquire is a research project by Dr. Vinay Chaudhri designing an “intelligent” textbook.
- Inquire’s functionality is dependent on an ontology that encodes key terms, definitions, and their relationships in a structured knowledge base.
- This knowledge base is constructed manually: a time-consuming and difficult process.
- Goal of the project is to build a model to automatically extract this knowledge base.
- We have worked on a simpler version: automatically extracting a textbook’s glossary.

Problem Formulation

Glossary Term Extraction as Token Classification

Input Sentence: Diffusion may be aided by channel proteins
Output Tags: S O O O O B E

S = single word glossary term, O = non-glossary word, B = start, I = interior, E = end of glossary phrase

Glossary Definition Extraction as Sentence Classification

Input Sentence: Diffusion is the process of random ….
Output Label: Definition Sentence (1)

Dataset

We scraped chapter sentences and glossaries from 6 open source science textbooks and the Life Biology textbook being used for the Inquire prototype.

Analysis

Glossary Term Extraction

- Left plot shows false positives split by categories that were identified by a domain expert during the correction process. Most were actually valid terms.
- Right plot shows the fraction of terms that were present in the training data at least once, but not tagged as key terms. This makes it harder for the model to learn more false negatives.

Glossary Definition Extraction

True FN: Pair rule genes divide the embryo into units of two segments each.
Non-Definition FN: Soils have living and nonliving components.

An estimated 45% of false positives are actual definition sentences where the defined word is not bolded as a key term:
The cells in the gastric pits that secrete HCl are called parietal cells.

Model Architectures

Hovey & Ma [1]: Term Extraction
- BLSTM

BERT [2]: Term & Definition Extraction
- BERT Base Encoder
- Max Pool

Anke et al. [3]: Definition Extraction
- CNN

Results

Term Extraction Models

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hovey &amp; Ma</td>
<td>0.432</td>
<td>0.527</td>
<td>0.366</td>
</tr>
<tr>
<td>BERT</td>
<td>0.430</td>
<td>0.541</td>
<td>0.357</td>
</tr>
<tr>
<td>BERT w/ Correction*</td>
<td>0.741</td>
<td>0.708</td>
<td>0.778</td>
</tr>
</tbody>
</table>

Definition Extraction Models

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anke et al.</td>
<td>0.38</td>
<td>0.39</td>
<td>0.37</td>
</tr>
<tr>
<td>BERT</td>
<td>0.44</td>
<td>0.46</td>
<td>0.42</td>
</tr>
<tr>
<td>BERT + Anke et al.</td>
<td>0.47</td>
<td>0.46</td>
<td>0.48</td>
</tr>
</tbody>
</table>

*False positives were corrected by domain expert to account for incomplete glossaries

Conclusions

- Performances are not immediately useable, but we have created a new dataset for the problem and improved on a previous class project’s baseline performance.
- Both sets of models appear to be limited by incomplete tagging and dataset errors.
- The problem formulation is difficult as there is no set rule for what constitutes a key term.
- This is an ongoing research project: advice & feedback appreciated!

References