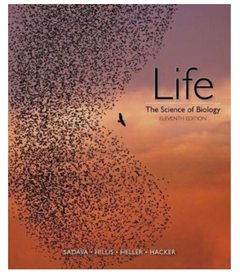
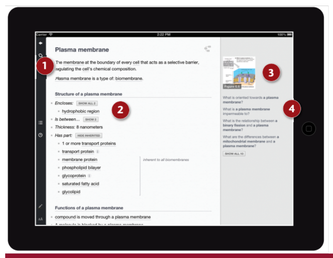
Featureless Deep Learning Methods for Automated Key-Term Extraction

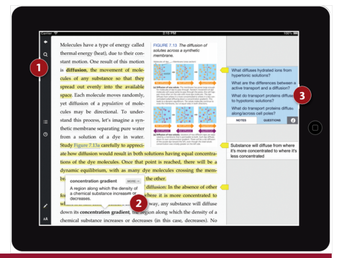
Nicholas Bowman, Robbie Jones, Kush Khosla, Project Mentor: Dr. Vinay Chaudhri
Departments of Computer Science and Mathematics, Stanford University



Motivation and Goals

- Automate knowledge base creation for the *Inspire* intelligent biology textbook [3]
- Identifying important technical terms in the text is the first step of ontology construction
- Current process for term extraction is slow and labor-intensive and done entirely by hand by domain experts.
- This work aims to automate the process of key-term extraction from textbooks, aiming for high precision in key-term identification

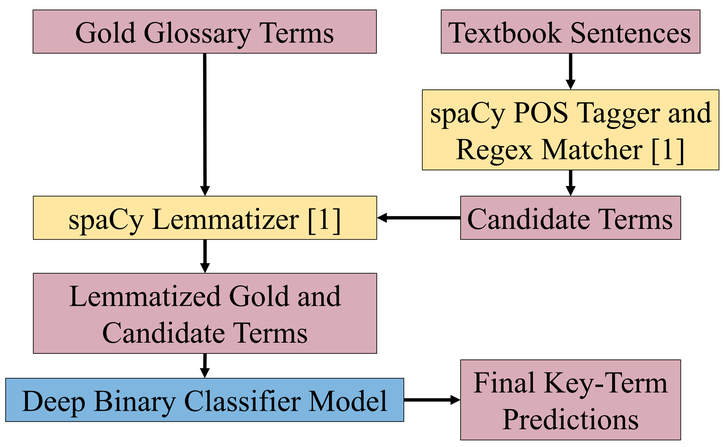




Problem Definition

Our task is defined simply as follows: given a corpus of text T and the set $N(T)$ of all n-grams contained in T, output a set $\mathcal{K} \subseteq N(T)$ that contains the key-terms of T. It is important to note that the definition of a "key-term" varies from application to application in the domain of key-term extraction. In our case, a key-term is considered to be one that appears in the glossary of the textbook, which serves as a brief dictionary of significant domain-specific terms.

Processing Pipeline



Implementation

We implemented three different architectures of binary classifier for this task: a supervised deep CNN, a supervised fully-connected neural network, and a semi-supervised co-training pipeline that unifies the two previous models.

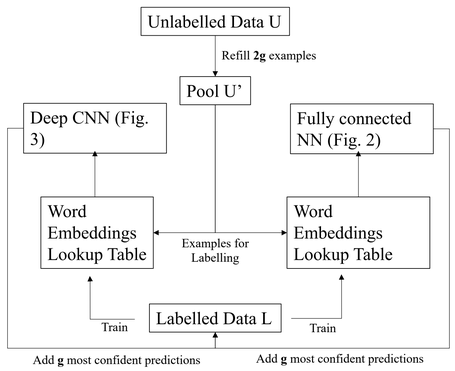


Figure 1: Co-training Pipeline Architecture

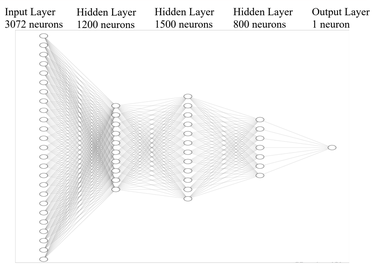


Figure 2: Fully Connected NN Architecture

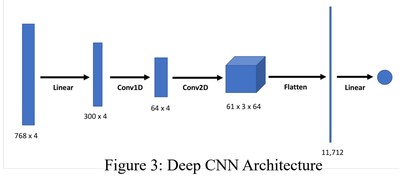


Figure 3: Deep CNN Architecture

Results

Biology (OpenStax)			
Model	**Precision**	**Recall**	**F1**
CNN	0.62	0.32	0.43
FC	0.35	0.22	0.27
Co-Train	0.13	0.44	0.20
Baseline	0.25	0.29	0.26
Microbiology (OpenStax)			
Model	**Precision**	**Recall**	**F1**
CNN	0.64	0.23	0.34
FC	0.0	0.0	N/A
Co-Train	0.17	0.41	0.24
Baseline	0.21	0.39	0.27
Biology + *Microbiology* (OpenStax)			
Model	**Precision**	**Recall**	**F1**
CNN	0.61	0.18	0.28
FC	0.0	0.0	N/A
Co-Train	0.15	0.40	0.22
Baseline	0.16	0.19	0.17
Life: The Science of Biology (Sadava)			
Model	**Precision**	**Recall**	**F1**
CNN-B	0.27	0.20	0.23
CNN-B SSL	0.1	0.6	0.17
Co-Train	0.08	0.5	0.14
Baseline	0.15	0.46	0.23

Table 2: Precision, Recall and F1 for various experiments

Table 2 summarizes important automatic metrics on all three of our textbook corpora

Table 3 summarizes the results of human evaluation tests that were administered to undergraduate and PhD students studying biology

False Positive Evaluation (models trained on *Biology*)			
Model	**New Precision**	**New Recall**	**New F1**
CNN	0.8	0.42	0.55
FC	0.69	0.43	0.53
Power Ranking			
Textbook	**Baseline**	**Our Model**	**Glossary**
Biology	24	32	34
Microbiology	16	36	38

Table 3: Results from Human Evaluation

Analysis

- Across the board, the CNN has the best performance. We hypothesize this is because it is able to leverage the structure embedded within the word vectors more efficiently than the fully connected network.
- For co-training, because the seed set had a higher +/- ratio than the true distribution it predicted positive much more often.
- Most experiments conducted on *Life* use networks trained elsewhere, in order to simulate our real use case: using a pre-trained model on a new textbook to automatically extract the glossary.
- In power ranking human evaluation model, domain experts were unable to distinguish our predicted terms from true predicted glossary terms
- Human false positive tagging gave huge boosts in reported precision and recall

Datasets

- Data was extracted from 3 different biology textbooks, two of which are open-source textbooks provided by the Openstax foundation

Table 1: Token Statistics Across Textbook Corpora

Textbook Name	Total # of Tokens	Total # Gold
Openstax Biology	392,311	2,307
Openstax Microbiology	403,189	1,744
Sadava Life	772,294	1,902

Methodology and Evaluation

- Once our model has produced its predictions of what the key-terms of a corpus of text are, we evaluate the quality of the extracted terms using both automatic and human evaluation standards
- The three automatics metrics that we use are the standard metrics of precision, recall, and F1 score.
  - Precision $= \frac{TP}{TP+FP}$
  - Recall $= \frac{TP}{TP+FN}$
  - F1-score $= \frac{2 * precision * recall}{precision+recall}$
- The three standards of human evaluation used are:
  - Human tagging of false-positive
  - Human tagging of conflicting predictions between two models (direct comparison)
  - Human ranking of groups of glossary terms predicted by different models

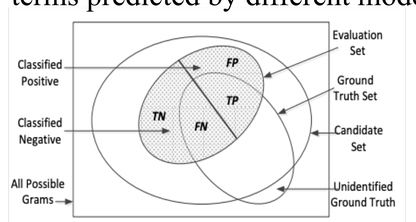


Figure 4: Relationships in TP, TN, FP, and FN for Term Extraction [2]

References

1. Matthew Honnibal. spaCy: Industrial-Strength Natural Language Processing. URL: https://spacy.io
2. Rui Wang, Wei Liu, and Chris McDonald. "Featureless Domain-Specific Term Extraction
with Minimal Labelled Data".
3. Vinay Chaudhri et al. *Inquire* Biology: A Textbook that Answers Questions