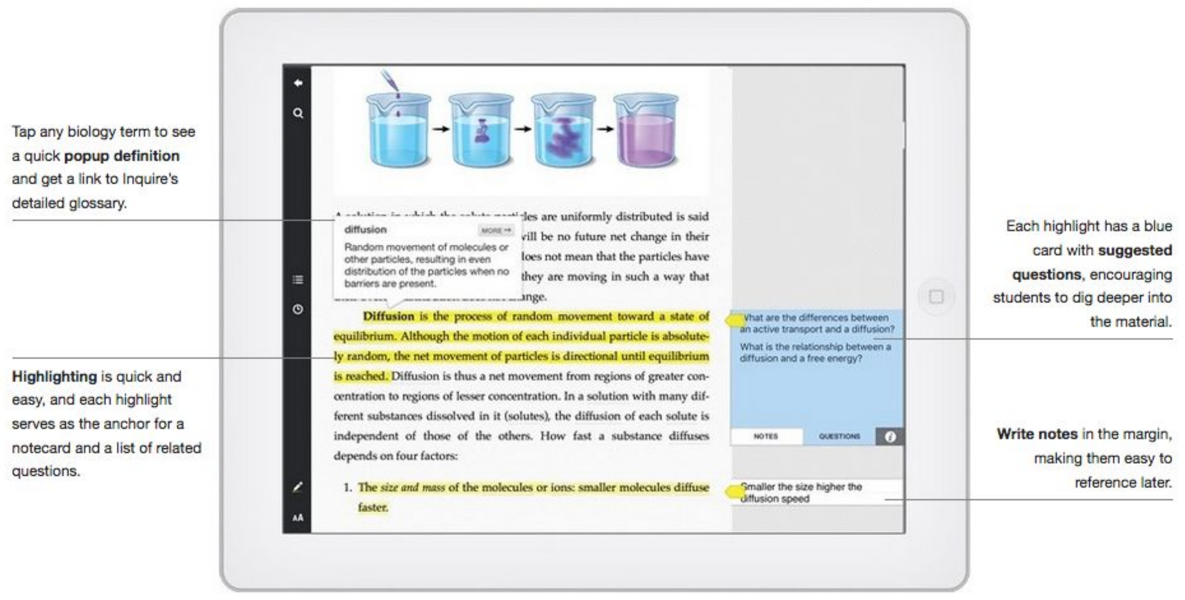
Automatically Extracting Textbook Glossaries Using Deep Learning

Manish Singh, Matthew Boggess, Vinay Chaudhri

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.**



Tap any biology term to see a quick **popup definition** and get a link to Inquire's detailed glossary.

Highlighting is quick and easy, and each highlight serves as the anchor for a notecard and a list of related questions.

Each highlight has a blue card with **suggested questions**, encouraging students to dig deeper into the material.

Write notes in the margin, making them easy to reference later.

diffusion

Random movement of molecules or other particles, resulting in even distribution of the particles when no barriers are present.

Diffusion is the process of random movement toward a state of equilibrium. Although the motion of each individual particle is absolutely random, the net movement of particles is directional until equilibrium is reached. Diffusion is thus a net movement from regions of greater concentration to regions of lesser concentration. In a solution with many different substances dissolved in it (solutes), the diffusion of each solute is independent of those of the others. How fast a substance diffuses depends on four factors:

1. The *size and mass* of the molecules or ions: smaller molecules diffuse faster.

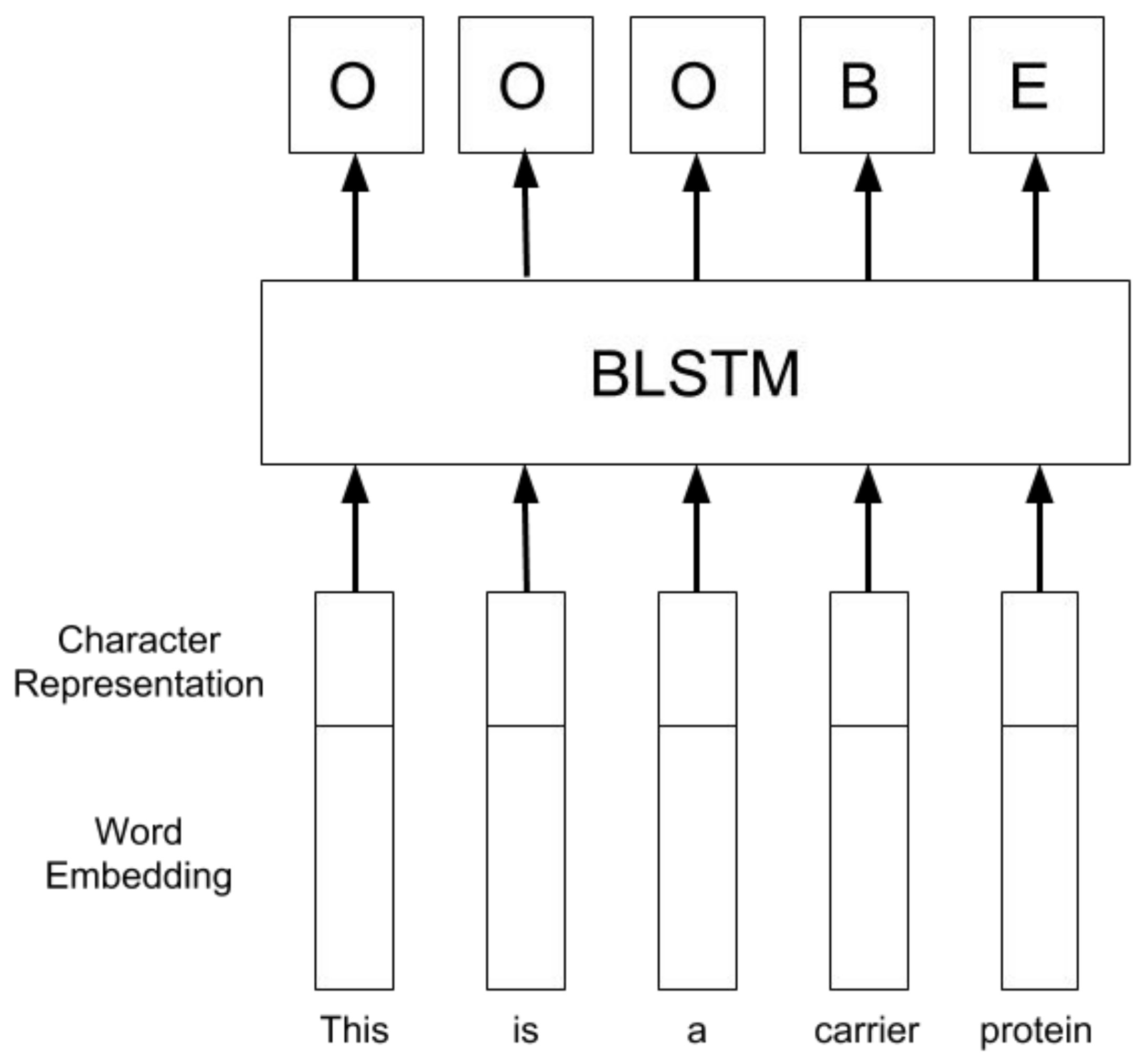
What are the differences between an active transport and a diffusion?

What is the relationship between a diffusion and a free energy?

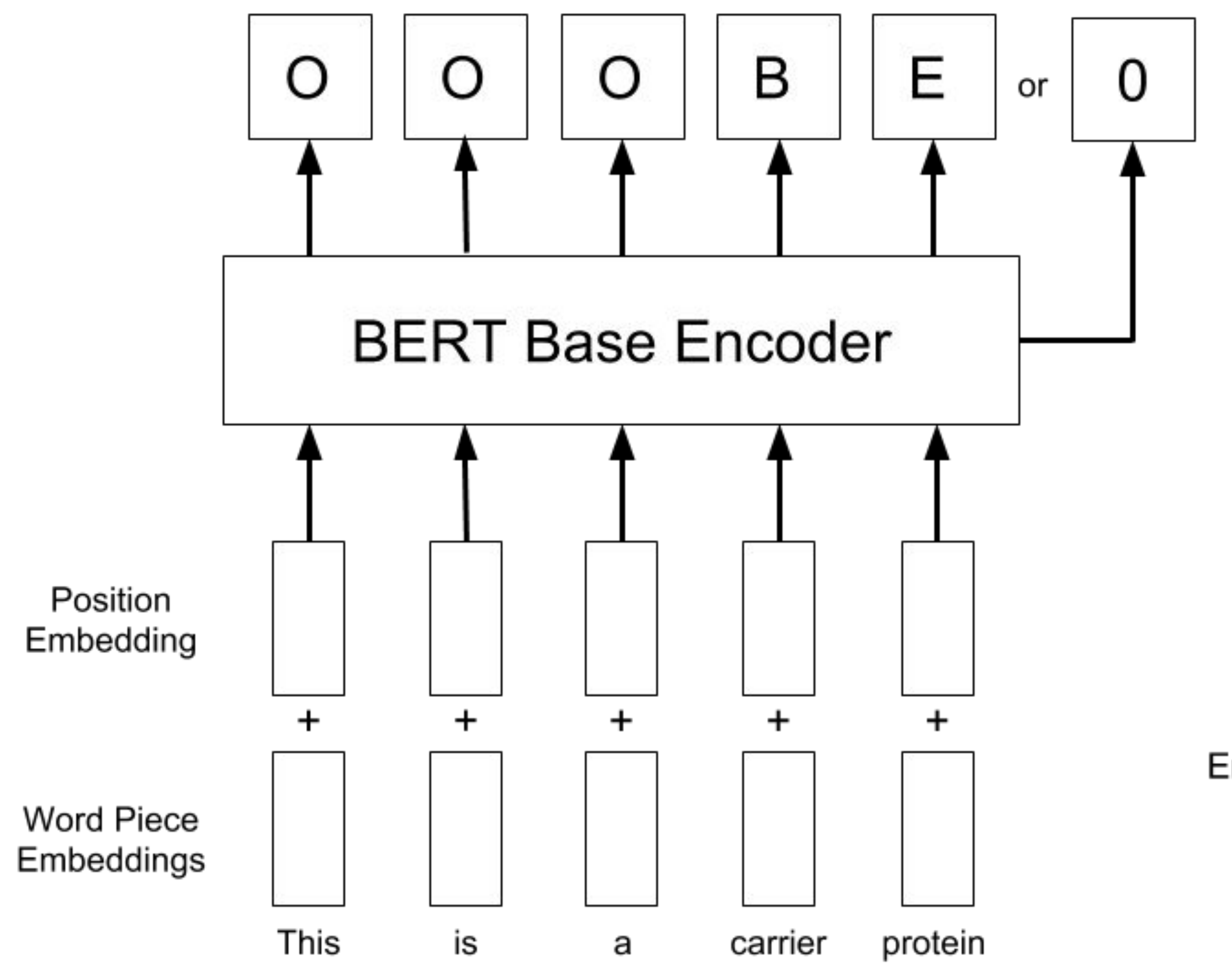
Smaller the size higher the diffusion speed

For more info on Inquire: http://web.stanford.edu/~vinayc/intelligent-life/

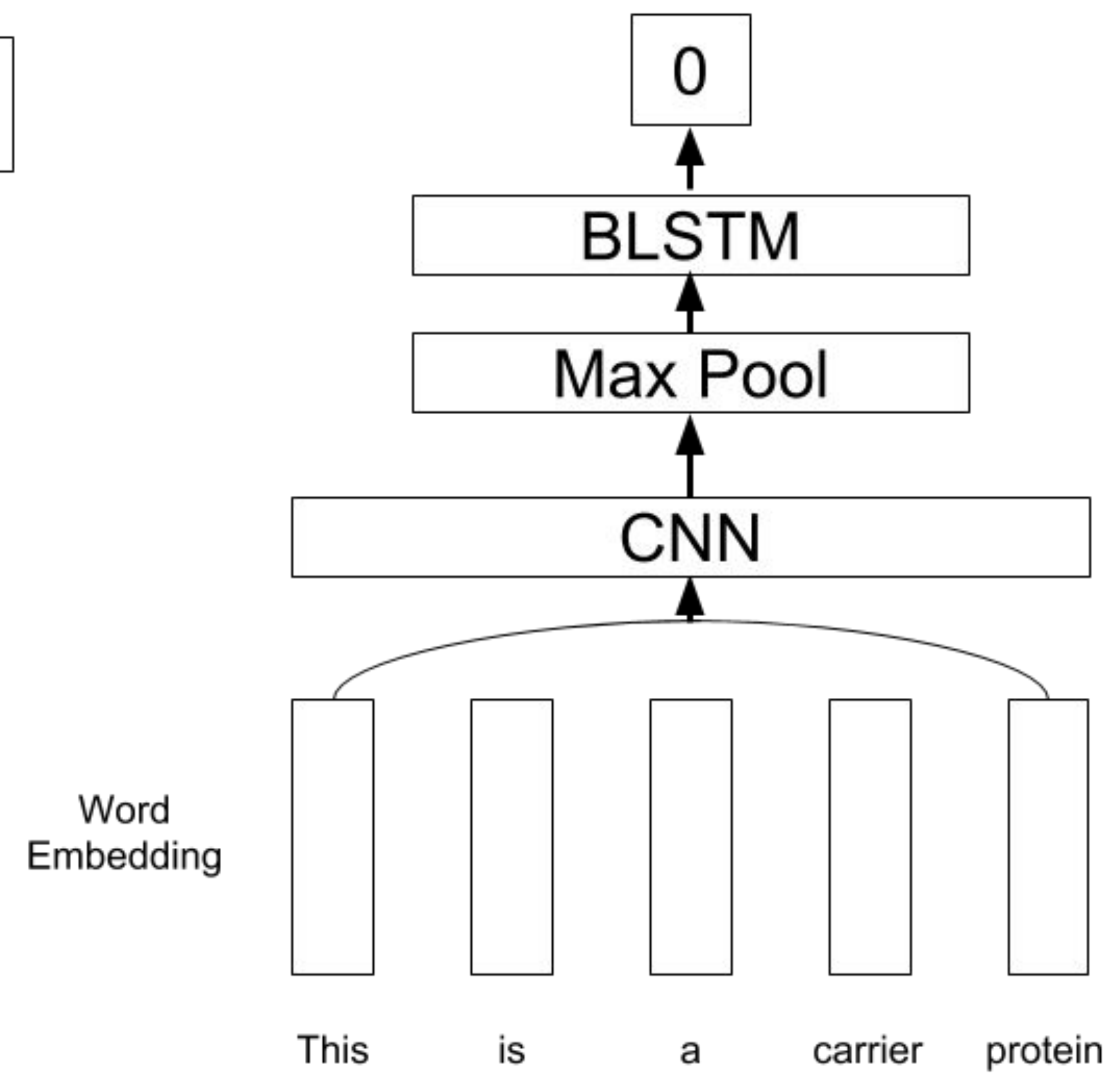
Model Architectures



Hovey & Ma [1]: Term Extraction



BERT [2]: Term & Definition Extraction



Anke et al. [3]: Definition Extraction

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

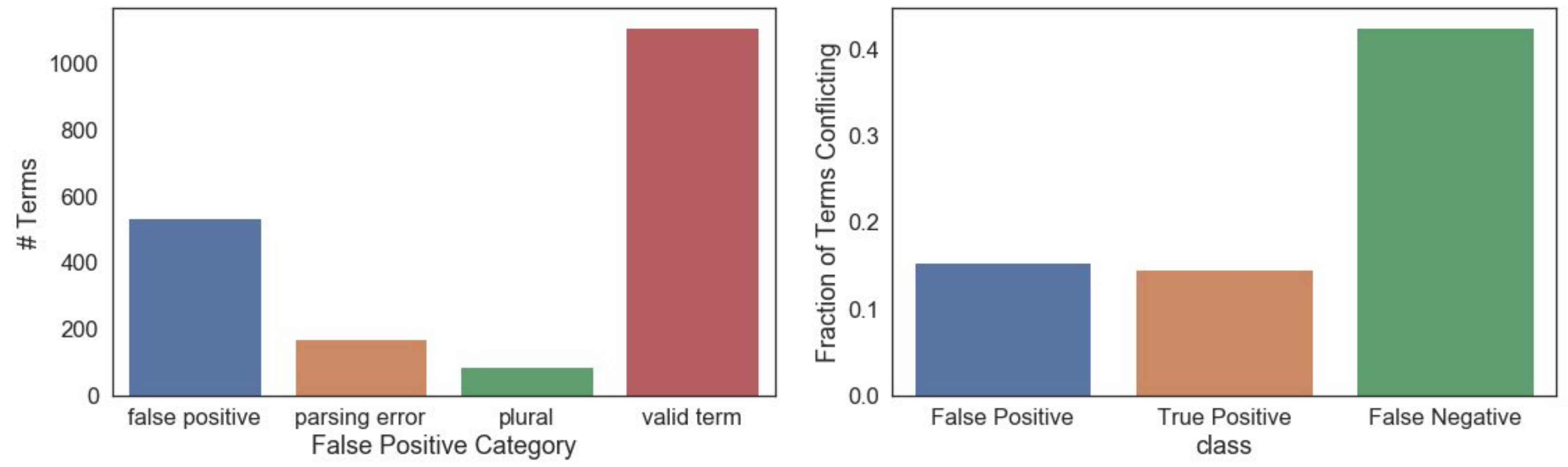
We scraped chapter sentences and glossaries from 6 open source science textbooks and the Life Biology textbook being used for the Inquire prototype.

Textbook	# Sentences	# Terms	# Definitions	Data Split
OS Anatomy	17804	2948	2609	Train/Dev
OS Microbiology	11509	1492	1322	Train/Dev
OS Physics	5500	362	322	Train/Dev
OS Chemistry	8262	677	578	Train/Dev
OS Astronomy	13384	306	493	Train/Dev
OS Biology	20529	1966	2101	Train/Dev
Life Biology	28662	1957	1999	Test

https://openstax.org/subjects/science

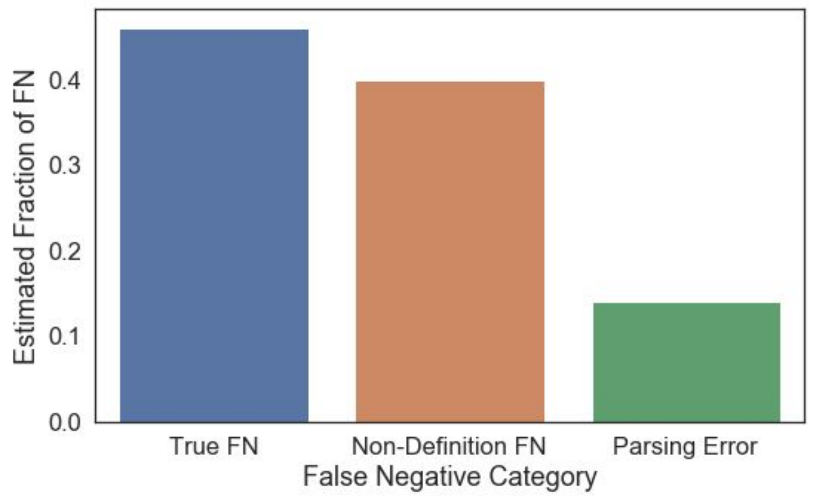
Analysis

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.**

Results

Term Extraction Models			
Model	**F1**	**Recall**	**Precision**
Hovey & Ma	0.432	0.527	0.366
BERT	0.430	0.541	0.357
BERT w/ Correction*	**0.741**	**0.708**	**0.778**
Definition Extraction Models			
Model	**F1**	**Recall**	**Precision**
Anke et al.	0.38	0.39	0.37
BERT	0.44	0.46	0.42
BERT + Anke et al.	**0.47**	**0.46**	**0.48**

*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

[1] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354, 2016.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[3] Luis Espinosa Anke and Steven Schockaert. Syntactically aware neural architectures for definition extraction. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 378–385, 2018.