Discovering Content through Text Mining for a Synthetic Biology Knowledge System

Mai H. Nguyen
Gaurav Nakum
Jiawei Tang
Xuanyu Wu
University of California, San Diego
La Jolla, CA, USA
{mhnguyen,gnakum,jit072,xuw057}@ucsd.edu

Bridget T. McInnes
Nicholas E. Rodriguez
Virginia Commonwealth University
Richmond, VA, USA
{btmcinnes,rodriguezne2}@vcu.edu

Eric Young
Kevin Keating
Worcester Polytechnic Institute
Worcester, MA, USA
{emyoung,kwkeating}@wpi.edu

1 INTRODUCTION

The field of synthetic biology has seen exciting growth in the last few years. Though the amount of data and publications has increased tremendously, the numerous available data sources are fragmented, and locating relevant data for genetic design is a challenging task. For example, finding biological part performance and sequence data remains a manual process of sifting through articles and supplemental material. To address this, we are developing a synthetic biology knowledge system that integrates disparate data and publication repositories to deliver effective and efficient access to available information.

Scientific articles contain a wealth of information about experimental methods and results on biological designs. Due to its unstructured nature and multiple sources of ambiguity and variability, however, extracting information from text is a difficult task. We are exploring various text mining approaches to identify concepts and entities in published articles in order to link each article to other elements in our knowledge system.

This paper describes our work using named entity recognition (NER), a sub-field of text mining, to mine existing literature. The goal of NER is to locate and classify named entities present in text into pre-defined categories. For synthetic biology, examples of such categories are names of genes, vectors, and regulatory elements. NER in biology domains has additional challenges due to the pace of new named entities being added, lack of naming convention, lengthy names, presence of special characters, and frequent and variable use of abbreviations.

2 METHODS

Deep neural network approaches have been applied to NER on biomedical texts. Specifically, state-of-the-art approaches use Long Short-Term Memory (LSTM) [2] with Conditional Random Field (CRF) [3] models and Transformers [6]. In our experiments, we use two deep neural network models developed for biomedical NER, namely HUNER [7] and BioBERT [4].

HUNER (Humboldt-Universität Named Entity Recognition) uses a combination of bidirectional LSTMs and CRFs. LSTMs are a type of deep learning model known as recurrent neural networks that can learn sequential data. Recurrent neural networks have an internal state to retain context from previous inputs, enabling them to learn sequences of data such as speech and text. In a bidirectional LSTM, each input sequence is presented both forwards and backwards so that context before as well as after the word being modeled are captured. A CRF is a discriminative probabilistic graphical model that models the conditional distribution of output variables given observed values. CRFs can also take context into account in making predictions for a data sample, making it ideal for predicting sequential data. In HUNER, a bidirectional LSTM is used to encode forward and backward contexts for the input word, which are then concatenated and fed to a CRF to predict the NER tag for the word.

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) uses another type of deep neural network model called a Transformer. Transformers are also designed to learn sequence data. Instead of relying on recurrent connections, however, Transformers use an attention mechanism to weigh the relevance of each input in producing the output. BioBERT is pre-trained on top of BERT [1], a general-purpose language representation model. This pre-training was conducted over PubMed abstracts and PubMed Central full-text articles to adapt to biomedical text mining tasks. In our work, the BioBERT output is fed into a simple feed forward neural network for the final NER prediction.

3 DATA

HUNER Data. The HUNER dataset consists of 34 different corpora covering five entity types: Chemicals, Cell Lines, Genes/Proteins, Species, and Diseases. The data was partitioned into 60% training, 10% validation, and 30% testing.
ACS Data. The American Chemical Society (ACS) Data comprises of full text articles from the Synthetic Biology Journal. The data set contains 1,545 articles between the years 2011 and 2019.

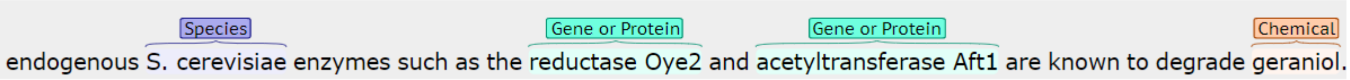
endogenous S. cerevisiae enzymes such as the reductase Oye2 and acetyltransferase Aft1 are known to degrade geraniol.

Figure 1: Annotations discovered by our NER model in an ACS article

4 RESULTS

Results on HUNER Test Data. Table1 shows the F-1 scores for HUNER and BioBERT on a subset of the HUNER dataset. F-1 is the harmonic mean between the precision and recall, where precision calculates how many instances are predicted correct out of all instances and and recall calculates how many instances are correctly predicted out of all the correct instances that should have been predicted. The results show that BioBERT obtained a higher F-1 score for all of the datasets except for OSIRIS.

Table 1: HUNER and BioBERT F-1 scores on HUNER Data

Entity Type	Corpus	BioBERT	HUNER
cellline	Gellus	0.924	0.714
	JNLPBA	0.818	0.649
	CLL	0.883	0.730
chemical	SCAI Chemicals	0.931	0.778
	CHEBI	0.877	0.804
	CHEMDNER patent	0.915	0.855
	CHEMDNER	0.920	0.889
	CDR	0.937	0.929
disease	miRNA	0.894	0.823
	NCBI Disease	0.923	0.854
	CDR	0.903	0.837
	SCAI Disease	0.855	0.801
gene	BioCreative II GM	0.898	0.779
	miRNA	0.807	0.697
	Variome Gene	0.928	0.823
	IEPA	0.913	0.824
	BioInfer	0.932	0.846
	DECA	0.731	0.688
	OSIRIS	0.811	0.874
species	Variome Species	0.823	0.701
	s800	0.834	0.725
	miRNA	0.955	0.909

Preliminary results on ACS Data. Our ultimate goal is to extract this type of information from the ACS dataset. To determine the efficacy of applying NER on biology-specific corpora to identify synthetic biology entities, we used our NER model to predict the entity types on 100 randomly selected ACS articles. Figure 1 shows entity mentions found by NER in an article. Each mention is associated with a single entity type. The BRAT annotation software [5] was used to create and view annotations. Figure 2 shows a word cloud that illustrates how often the Chemical types were mentioned in the ACS dataset. The larger the term is in the word cloud, the more often it was identified by NER in the set of ACS articles. Annotations found by the NER model will be reviewed and enhanced by domain experts to create a more refined dataset for fine tuning the model.

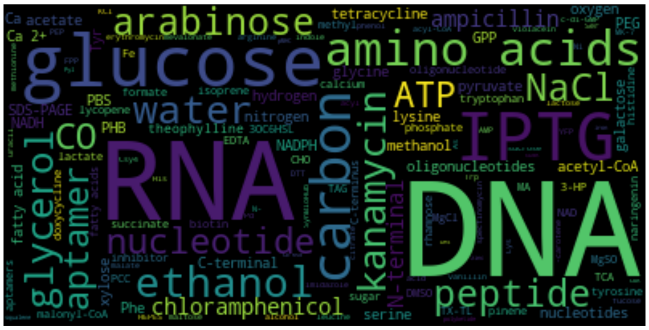

Figure 2: Word cloud of Chemical entities found by NER

5 DISCUSSION

This work presents the application of deep neural network models to identify entities mentioned in scientific articles. The approach described here to extract information about the contents of an article can be used to link publications to data in our knowledge system. The integration of disparate data sources will allow researchers to effectively and efficiently locate related work, enabling maximal leverage of previous research, and has the potential to greatly accelerate exploration and discovery of synthetic biology research.

ACKNOWLEDGEMENTS

This work was funded by the National Science Foundation under Grants No. 1939885, 1939951, and 1939860.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[3] Lafferty, J., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 36*, 4 (2020), 1234–1240.

[5] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the 13th EACL Conference (Demo)* (2012), pp. 102–107.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[7] Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M., and Leser, U. Huner: improving biomedical ner with pretraining. *Bioinformatics 36*, 1 (2020), 295–302.