Creating a Myxoma virus knowledgebase annotated with BioBERT, a pre-trained biomedical language processing model for biomedical text-mining, to characterize genomic variation of virus strains

Yu-Chih Hsiao[1,2], Chris Yoo PhD[2]

[1]Biomedical Informatics, College of Health Solutions, Arizona State University, Scottsdale, AZ;
[2]Systems Imagination, Inc, Tempe, AZ

Abstract

Here we describe the application of BioBERT, a biomedical text-mining model to automate the extraction of information from public, scientific data sources in the construction of an internal knowledgebase of Myxoma virus research findings. This construct, in combination with other manually-curated data sets from the scientific literature and open access databases, will provide researchers with a highly explorable, graph-based database for new insight discovery about Myxoma virus biology. This research demonstrates the effectiveness of using available programming languages and deep learning techniques to rapidly construct and populate purpose-built biomedical research knowledgebases to enable new insight discoveries.

Introduction

BERT: Natural language processing (NLP) and text mining approaches have been used previously to identify valuable information from published literature. Bidirectional Encoder Representations from Transformers (BERT) showed powerful performance in a wide range of NLP research. Unlike other language models, BERT is designed to train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all learning layers[1]. Users of BERT can use the model which has been trained from immense resource by developers and fine-tune their model based on their target, rather than use their own hardware and time to train the model.

BioBERT: The performance of BERT leaves room for improvement because the word distribution of general corpuses of text, which prior NLP algorithms have been trained on, are not well suited to biomedical literature, creating challenges for the application of generalized text mining models in biomedical knowledge engineering uses. To address this, BioBERT was developed as a model specifically trained for biomedical text (Figure 1)[2].

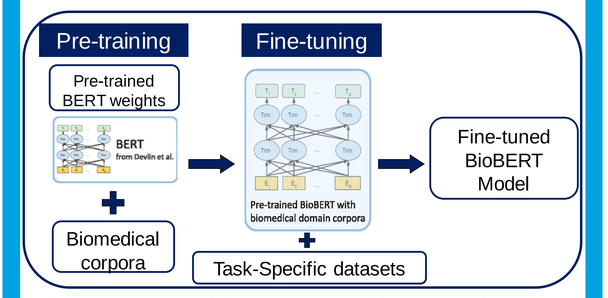


Figure 1. Overview of the pre-training and fine-tuning of BioBERT.

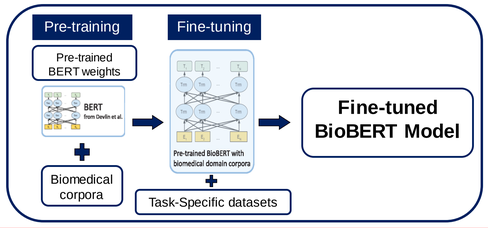
BERN: BERN is a neural Biomedical named Entity Recognition and multi-type Normalization tool [3]. BERN uses BioBERT named entity recognition(NER) models and recognize known entities and discover new entities. It provides Web-based service for researchers who want to discover and tag entities in raw text and even PubMed articles with PubMed ID.

Myxoma virus: We decided to focus on scientific knowledge on the Myxoma virus, due to the fact that the data would be of a manageable size and the types of knowledge, the data would present enough variability for us to determine the accuracy of BioBERT application.

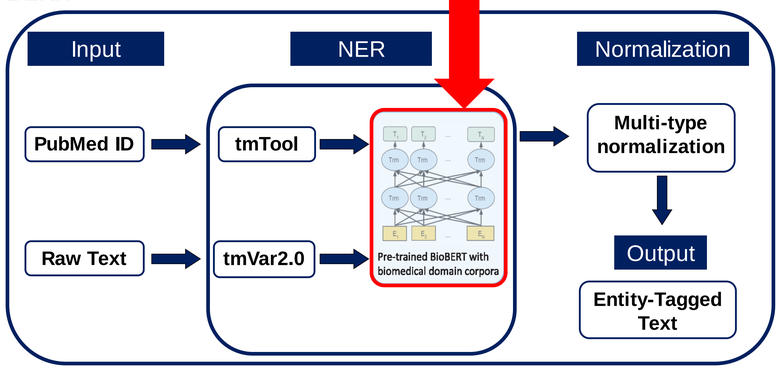
Methods and Materials

The Biomedical Text-mining Pipeline

1) **Pre-training of model**
Using pre-trained weights of BERT as published in the BERT repository by Google which was pretrained on English Wikipedia and Books Corpus[4].
2) **Fine-tuning of model**
Using Speices-800 datasets to train and test model[5]. This dataset is focus on species names which can be used in NER.



3) **BERN**



MATERIALS

NVIDIA DGX-1 Workstation

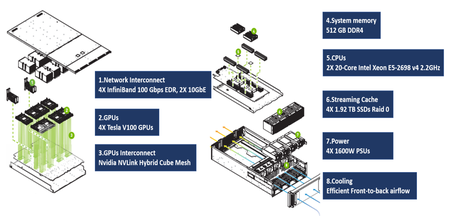


Figure 2. DGX-1 with V100 system components.

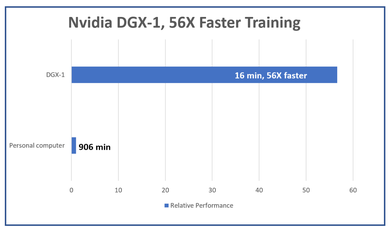


Figure 3. Relative performance based on pre-training time of BioBERT NER model.

Results

Model Training Result

Model	Total Tokens	Total Phrases	Phrases found	Phrases correct	Precision	Recall	F-1 score
BioBERT trained model	42298	767	827	**592**	**71.58%**	**77.18%**	**74.28**
Non-trained model	42298	767	5346	25	0.47%	3.26%	0.82

Figure 4. Performance evaluation result of the trained model and the **non-trained** model applied to scientific data

Results

Named Entity Recognition result

```
30940649|t|Myxoma virus M013 protein antagonizes NF-kappaB and inflammasome pathways via distinct structural motifs.
30940649|a|Among the repertoire of immunoregulatory proteins encoded by myxoma virus, M013 is a viral homologue of the viral pyrin domain-only protein (vPOP) fami
signaling pathways by direct binding to ASC1 and NF-kappaB1, respectively. In this study, a three-dimensional homology model of the M013 pyrin domain (PYD) was bu
electrostatic map of the M013 PYD is the presence of a negatively region consisting of numerous aspartate and glutamate residues in close proximity. Single-site m
biological significance of charge complementarity in the M013-ASC-1 interaction was further confirmed by functional assays of caspase-1 activation and subsequent
N-terminal PYD, and it is enriched in positively charged residues. Deletion of the tail of M013 significantly inhibited the interactions between M013 and NF-kappa
inflammatory cytokines. These results demonstrate that vPOP M013 exploits distinct structural motifs to regulate both the inflammasome and NF-kappaB pathways.
30940649    0    25    Myxoma virus M013 protein    gene    CUI-less
30940649    0    6    Myxoma    disease MESH:D009232|BERN:106924701
30940649    0    12    Myxoma virus    species NCBI:txid10273
30940649    38    47    NF-kappaB    gene    MIM:164011|HGNC:7794|Ensembl:ENSG00000109320|BERN:324174102
30940649    167    179    myxoma virus    species NCBI:txid10273
30940649    181    185    M013    gene    CUI-less
30940649    214    245    viral pyrin domain-only protein gene    CUI-less
30940649    247    251    vPOP    gene    CUI-less
30940649    279    283    M013    drug    CUI-less
30940649    279    291    M013 protein    gene    CUI-less
30940649    344    353    NF-kappaB    gene    MIM:164011|HGNC:7794|Ensembl:ENSG00000109320|BERN:324174102
30940649    394    398    ASC1    gene    MIM:604501|HGNC:12310|Ensembl:ENSG00000103671|BERN:325476802
30940649    403    413    NF-kappaB1    gene    CUI-less
30940649    486    503    M013 pyrin domain    gene    CUI-less
30940649    505    508    PYD    gene    CUI-less
30940649    551    554    PYD    gene    CUI-less
30940649    637    641    M013    drug    CUI-less
30940649    642    645    PYD    gene    CUI-less
30940649    708    717    aspartate    drug    CHEBI:72314|BERN:283828103
30940649    722    731    glutamate    drug    CHEBI:14321|BERN:295829803
30940649    786    795    aspartate    drug    CHEBI:72314|BERN:283828103
30940649    800    809    glutamate    drug    CHEBI:14321|BERN:295829803
30940649    858    863    ASC-1    gene    MIM:604501|HGNC:12310|Ensembl:ENSG00000103671|BERN:325476802
30940649    926    930    M013    drug    CUI-less
30940649    931    936    ASC-1    gene    MIM:604501|HGNC:12310|Ensembl:ENSG00000103671|BERN:325476802
30940649    995    1004    caspase-1    gene    MIM:147678|HGNC:1499|Ensembl:ENSG00000137752|BERN:325264602
30940649    1055    1059    M013    gene    CUI-less
30940649    1089    1090    C    drug    CHEBI:30039|BERN:305421203
30940649    1122    1123    N    drug    CHEBI:35555|BERN:308520903
30940649    1133    1136    PYD    gene    CUI-less
30940649    1213    1217    M013    gene    CUI-less
30940649    1267    1271    M013    gene    CUI-less
30940649    1276    1286    NF-kappaB1    gene    CUI-less
30940649    1427    1436    vPOP M013    gene    CUI-less
30940649    1432    1436    M013    drug    CUI-less
30940649    1511    1520    NF-kappaB    gene    MIM:164011|HGNC:7794|Ensembl:ENSG00000109320|BERN:324174102
```

1. Title
2.Abstract
5.Entity
6. Entity ID
7. Category
3.PubMed ID
4. Location

Figure 5. Result of NER mission by using BioBERT and BERN. There are four different entity categories, gene/protein, drug/chemical, disease and species. The targets are myxoma virus, related genes/proteins and drug/chemicals. In the title it only shows the article is related to M013 protein, NF-kappa B and inflammasome pathways. BioBERT and BERN discovered different genes which is related to myxoma virus in the abstract.

Conclusion

Text mining is a powerful and rapidly developing technology that can be useful for helping humans find relevant information in very large datasets. Text mining models pre-trained on biomedical corpora have been shown to be promising. This study provides solid evidence that text-mining could be effectively used on a specific subset of biomedical research to rapidly provide knowledgebases for new insight discovery. For virus research and research in other areas where the knowledge may be less well known or disseminated, text-mining can play a critical role in bringing together disparate sources into a single research resource. By combining appropriate tools, we are enabling the discovery of useful information more efficiently than current search methods.

References

1. Devlin J, Chang M, Lee K, Toutanova K et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA. pp. 4171–4186. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423.
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So C et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019
3. Kim D, Lee J, So C, Jeon H, Jeong M, Choi Y et al. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. IEEE Access. 2019;7:73729-73740.
4. naver/biobert-pretrained [Internet]. GitHub. 2020 [cited 3 April 2020]. Available from: https://github.com/naver/biober t-pretrained/releases
5. HCMR S. SPECIES - Organism Name Identification in the Scientific Literature [Internet]. Species.jensenlab.org. 2020 [cited 3 April 2020]. Available from: https://species.jensenlab.org/.

Acknowledgements

This research was supported by Systems Imagination. I would like to acknowledge Chris Yoo for his mentorship and guidance throughout the project. I am also grateful to Pieter Derdeyn for his expertise and comments on the earlier stage of research.