Classification and Comparison of Gene-Disease Associations using BioBERT and PubMedBERT on GeneRIF Data

A. Midorikawa* Y. Sakurai T. Kimura K. Kawahara (midorikawa@w-fusion.co.jp)

WorldFusion Co., LTD. 103-0014 Tokyo, Japan.

Abstract

In this research, we explore gene-disease associations using a two-stage fine-tuning approach with BERT-based models. Initially, we fine-tune the models without domain-specific data, and subsequently, we conduct a second round of fine-tuning incorporating domain-specific knowledge from the Life Science Knowledge Bank (LSKB). Our study leverages the GeneRIF dataset, a rich resource with functional annotations for genes and associated diseases. We classify gene-disease associations into 11 hypotheses using the Natural Language Inference (MNLI) technique and observe remarkable accuracy improvements through this strategic fine-tuning process.

Introduction

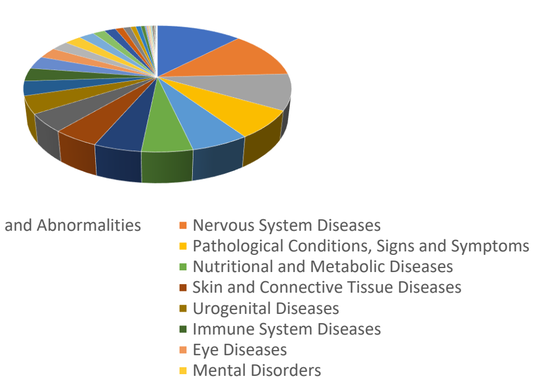
This study embarks on an exploration of gene-disease associations, employing a multi-stage fine-tuning approach with state-of-the-art BERT-based models for Natural Language Inference (MNLI). BioBERT[1)] and PubMedBERT[2)] serve as the foundational language models. Our methodology involves a two-stage fine-tuning process: first, the models undergo fine-tuning without domain-specific data, and then, in the second stage, we incorporate domain-specific knowledge from the Life Science Knowledge Bank (LSKB)[4)].

Methods

GeneRIF Data Overview

We began our analysis with GeneRIF data, which is a valuable resource for gene annotations. Our dataset, retrieved on June 14, 2022, comprises 1,589,973 entries. These entries encompass 108,378 unique genes and span a diverse taxonomy, including Human (20,220 genes), Mouse (14,305 genes), Rat (7,545 genes), and Plant (9,781 genes) origins. The dataset also contains annotations for 21,681 distinct diseases, showcasing a wide range of disease classes.

Disease Class in GeneRIF



Fine-Tuning Datasets

To fine-tune our models, we initially utilized non-domain-specific datasets. Examples from these datasets, including entailment, neutral, and contradiction instances, are presented in Table 1 for the first round of fine-tuning. In the second round, we incorporated domain-specific knowledge from LSKB, resulting in significant model improvement.

Table 1: Data Examples of Fine-tuning

1st round Fine-tuning Data Example		
Entailment	PREMISE:	How do you know? All this is their information again.
	Hypothesis:	This information belongs to them.
Neutral	PREMISE:	He turned and smiled at Vrenna.
	Hypothesis:	He smiled at Vrenna who was walking slowly behind him with her mother.
Contradiction	PREMISE:	but that takes too much planning
	Hypothesis:	It doesn't take much planning.
2nd round Fine-tuning Data Example		
Entailment	PREMISE:	Regarding Asparagine degradation, It also has asparagine hydrolase activity.
	Hypothesis:	Asparagine hydrolase activity can be used to degrade Asparagine.
Neutral	PREMISE:	Regarding Sucralfate, Sucralfate is also used to treat mucositis.
	Hypothesis:	Marrow cells can be used to treat diseases.
Contradiction	PREMISE:	Regarding Oxymetholone, It is used mainly in the treatment of anemias.
	Hypothesis:	Obesity is a common symptom of Rubinstein-Taybi syndrome 2.

MNLI System Environment

Our experiments were conducted within a robust system environment to ensure reliable and consistent results. The hardware specifications of the system used for fine-tuning and analysis are as follows:

- **BioBERT Ver1.1 / PubMedBERT base**
- **CPU**: Intel(R) Core(TM) i7-10700F CPU @ 2.90GHz
- **System Memory**: 64GB
- **Graphics Card**: NVIDIA RTX A4000 (Memory: 16GB)
- **SSD**: 500GB

The software stack included:

- **Operating System**: Oracle Linux Server release 8.5
- **Python Version**: 3.10.2
- **Keras Version**: 2.11.0
- **TensorFlow Version**: 2.11.0
- **Transformers Library Version**: 4.32.1

Hypothesis Classification

We categorized our hypotheses into six major classes, each addressing different aspects of gene-disease relationships. Table 2 provides a concise list of these hypotheses, their corresponding IDs, and classes.

Table 2: Hypothesis and classification

ID	Hypothesis	Class
Hypo 01	"gene" is the cause of "disease"	Etiology
Hypo 02	"gene" associates "disease"	Etiology
Hypo 03	"gene" may become the biomarker of "disease"	Biomarker
Hypo 04	negative expression of "gene" is the cause of "disease"	Altered Expression
Hypo 05	negative expression of "gene" is the risk of "disease"	Altered Expression
Hypo 06	positive expression of "gene" is the cause of "disease"	Altered Expression
Hypo 07	positive expression of "gene" is the risk of "disease"	Altered Expression
Hypo 08	Mutations in "gene" are associated with "disease"	polymorphism
Hypo 09	"gene" polymorphism associates "disease"	polymorphism
Hypo 10	inhibition of "gene" is the cause of "disease"	Inhibition
Hypo 11	enhancement of "gene" is the cause of "disease"	Enhancement

Results

Etiology Hypothesis Class

For the Etiology hypothesis class, we observed notable improvements following the second round of fine-tuning. Table 3 details the changes in the number of entailment instances and the accuracy before and after fine-tuning.

Table 3: Comparison of the results for Etiology Hypothesis Class

	# of Entailment				Accuracy	
	1st Model		2nd Model		1st Model	2nd Model
Both Entailment	264,798	36.1%	19,297	20.5%	98.0%	90.0%
Entailment BioBERT only	152,029	20.7%	51,277	54.4%	12.0%	88.0%
Entailment PubMedBERT_only	316,383	43.2%	23,628	25.1%	30.0%	74.0%
Total	733,210		94,202		46.7%	84.0%
Total (+Both Non-Entailment)	2,020,529		606,419			

Polymorphism Hypothesis Class

In the Polymorphism hypothesis class, we also observed substantial enhancements in accuracy after domain-specific fine-tuning. Table 4 presents the alterations in the number of entailment instances and the accuracy before and after fine-tuning.

These results underscore the impact of domain-specific fine-tuning on BERT-based models, particularly in the classification of gene-disease relationships across various hypotheses and classes.

Table 4: Comparison of the results for Polymorphism Hypothesis Class

	# of Entailment				Accuracy	
Entailment 総数	1st Model		2nd Model		1st Model	2nd Model
Both Entailment	128,530	39.6%	4,754	20.5%	72%	86%
Entailment BioBERT only	93,358	28.8%	19,077	54.4%	38%	80%
Entailment PubMedBERT_only	102,642	31.6%	8,395	25.1%	44%	86%
Total	324,530		32,226		51%	84%
Total (+Both Non-Entailment)	690,127		250,778			

Conclusion

Our research showcases the potential of AI and domain-specific fine-tuning in classifying gene-disease associations. The highly accurate models developed in this study hold promise for extracting insights from biological datasets, facilitating discoveries in the life sciences.

References

1. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* (09 2019). https://doi.org/10.1093/bioinformatics/btz682
2. Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." *Volume 1 (January 2021)*, 1, Article 1, 24 pages. https://doi.org/10.1145/3458754
3. GeneRIF. https://www.ncbi.nlm.nih.gov/gene/about-generif
4. LSKB. https://www.lskb.jp/
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "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)*, 4171-4186.