Leveraging Biomedical Ontologies to Boost Performance of BERT-Based Models for Answering Medical MCQs

Sahil and P Sreenivasa Kumar

Department of Computer Science and Engineering

Indian Institute of Technology Madras

31st August 2023

Outline of the Talk

1. Introduction
2. MedMCQA Dataset
3. Related Work
4. Proposed Approach for Pre-training BERT
5. Results
6. Conclusions
7. Future work

Ontology

Formal representation of knowledge that defines the concepts, relationships, and properties within a particular domain

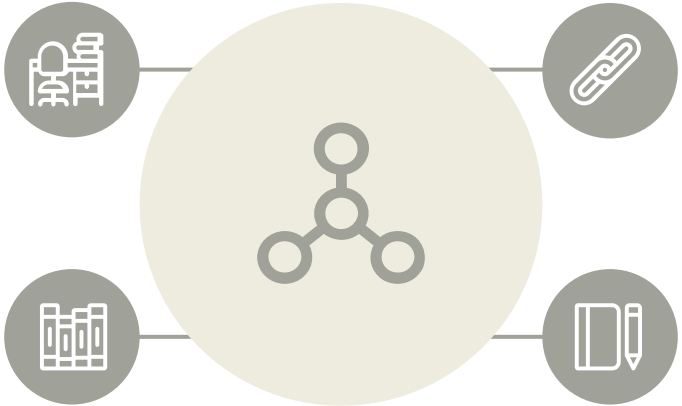
Ontology

Classes

Concepts or categories within a domain

Instances

Specific individual entities within classes



Properties

Relationship between classes

Axioms

Logical statements that define relationships

Biomedical Ontologies

Structured knowledge frameworks designed for organizing and categorizing information in the fields of biology, medicine, and healthcare

1. Gene Ontology : Defines gene functions and relationships
2. Human Phenotype Ontology : Describes phenotypic abnormalities
3. Disease Ontology : Represents diseases and their relationships
4. Foundational Model of Anatomy Ontology : Defines anatomical structures and their spatial relationships
5. Precision Medicine Ontology : Captures genetic, molecular, and clinical data to tailor treatments
6. Dental Ontology : Describes dental anatomy, procedures, and conditions

Biomedical Ontologies - some details

Ontology	Scope	Classes	# Object Properties	# Annotations	# subClass
FMAO Ontology	Anatomy	104721	139	51	262548
Bioassay Ontology	Pharmacology	904	17	34	981
Dental Ontology	Dentistry	2745	62	28	6507
Gene Ontology	Bioinformatics	84108	297	60	192606
Precision Medicine Ontology	Medicine	76155	95	23	122760
Disease Ontology	Pathology	11033	2	53	11063
Paediatrics Ontology	Paediatrics	1771	-	8	1760
HPS Ontology	Physiology	2920	86	34	3143
Mental Disease Ontology	Psychiatry	879	41	102	940

Biomedical Ontologies - Observations

- The primary axioms consist mainly of 'subClassOf', which depict the hierarchical structure
- Biomedical Ontologies are rich in annotation properties, containing valuable medical knowledge
- Synonyms and Definitions were abundantly found within the annotation properties of all the mentioned ontologies
- Our approach involves utilizing subClassOf axioms, along with synonyms and definition properties.

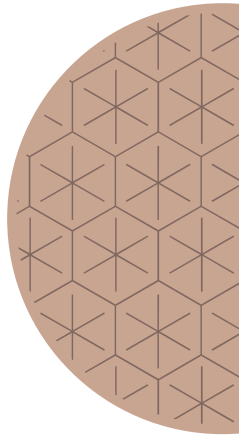
Foundational Knowledge

- Curated Biomedical Ontologies contains foundational biomedical knowledge
- Contains structured concepts, relationships and their synonyms and definitions
- PubMed* abstracts and articles contain advanced research and results
  - Assume reader has foundational biomedical knowledge
  - Do not explicitly mention fundamental knowledge

* https://pubmed.ncbi.nlm.nih.gov

MedMCQA Dataset[1]

- Large-scale Multiple-Choice Question Answering (MCQA) dataset
  - Training Set – 182,822 questions
  - Validation Set – 4,183 questions
  - Test Set – 6,150 questions
- Challenging dataset includes NEET-PG & AIIMS-PG entrance questions
- 2400 healthcare topics and 21 medical subjects
- Ground Truth
  - Test set – not available
  - Validation set – available



1. A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248-260.

MedMCQA Dataset[1] - Example

1. Dentigerous cyst is likely to cause which neoplasia?

   a. Ameloblastoma ✓

   b. Adenocarcinoma

   c. Fibrosarcoma

   d. All of the Above

2. Which one of the following is a muscle splitting incision?

   a. Kocher's incision

   b. Lanz incision

   c. Rutherford–Morrison incision

   d. Pfannenstiel incision ✓

1. A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260.

Ontology-based approaches

- Ontology-based question-answering systems show promise in capturing domain-specific knowledge.
- XMQAS by Midhunlal et al.[2] utilized NLP and ontology-based analysis
- Kwon et al.'s [3] approach employed SPARQL templates for medical knowledge retrieval
- **Limitations:** Template-based approach restricts flexibility

2. M. Midhunlal, M. Gopika, Xmqas-an ontology based medical question answering system, International Journal of Advanced Research in Computer and Communication Engineer- ing 5 (2016) 929-932.
3. S. Kwon, J. Yu, S. Park, J.-A. Jun, C.-S. Pyo, Stroke medical ontology qa system for pro- cessing medical queries in natural language form, in: 2021 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2021, pp. 1649- 1654.

BERT-based model approaches

- BERT-based models comprehend medical terminology and complex questions
- Pretrained on extensive corpora like Pubmed abstracts (3.1B words)
- Following are some BERT-based models, which have shown better results on MCQ answering:

  a. SciBERT [4]

  b. BioBERT [5]

  c. PubmedBERT [6]

4. I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
5. J.Lee,W.Yoon,S.Kim,D.Kim,S.Kim,C.H.So,J.Kang,Biobert:apre-trainedbiomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234- 1240.
6. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1-23.

Looking into a missing piece

How to connect organised Biomedical Ontology knowledge with language models

Previous Approaches

- Limited work on ontologies with neural networks in MCQ answering task
- KLMo[7] incorporates both entities and fine-grained relationships from a Knowledge Graph into language representation learning
- KI-BERT [8]: Infuses knowledge context from ConceptNet and WordNet into Transformer-based language models
- This methods improve entity understanding and representation.
- **Limitations:**

  i. Computational Overhead

  ii. Manual Annotation Efforts

7. L.He,S.Zheng,T.Yang,F.Zhang,Klmo:Knowledgegraphenhancedpretrainedlanguage model with fine-grained relationships, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4536-4542.
8. K. Faldu, A. Sheth, P. Kikani, H. Akbari, Ki-bert: Infusing knowledge context for better language and domain understanding, arXiv preprint arXiv:2104.08145 (2021).

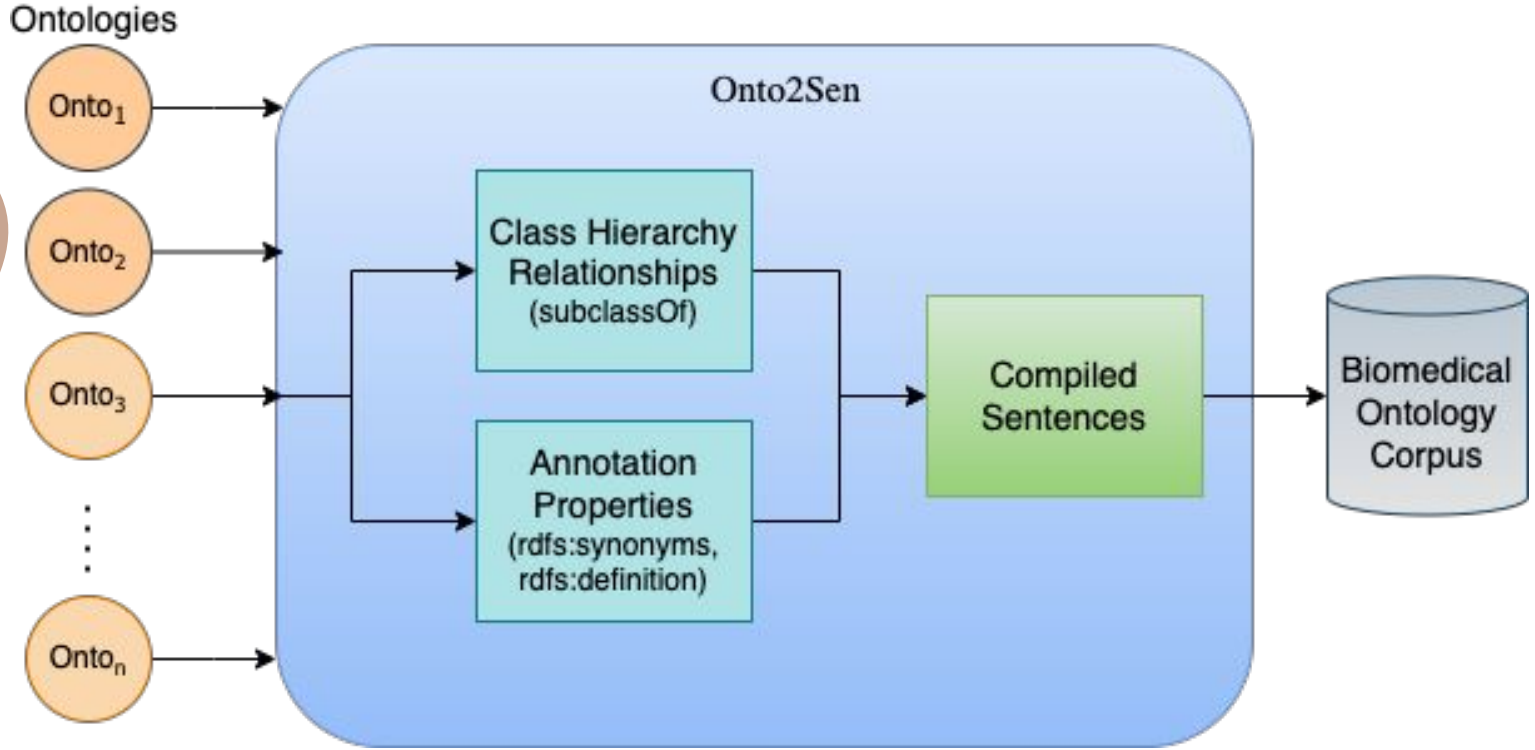
Proposed Approach

Pre-training BERT model on
Biomedical Ontologies

Onto2Sen System

- Generates sentences from the curated ontologies
- Extracts Class hierarchy and Annotation properties using SPARQL queries
- Pre-defined templates are used to form sentences based on the properties
- Generated corpus consists of 20M words (158MB size)

Onto2Sen System



Onto2Sen System Architecture

Onto2Sen Template Example

subClassOf

SPOAN syndrome **is a** Neurodegenerative disease

Synonym

Pterygium **has synonym** Surfer's eye

Definition

Autoimmune Pancreatitis **is defined as** an autoimmune disease of endocrine system that is located in the pancreas.

Corpus Size Comparison

158 MB

20M Words

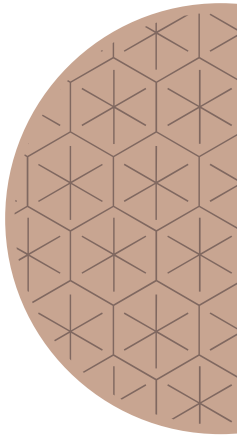
Onto2Sen-generated Biomedical Ontology Corpus

Pre-trained BERT-based models

- Pre-training is crucial for BERT
  - Bidirectional Encoder Representation from Transformers
- BERT learns contextualized word representations from vast unlabeled text data
- Captures semantic relationships using bidirectional transformers

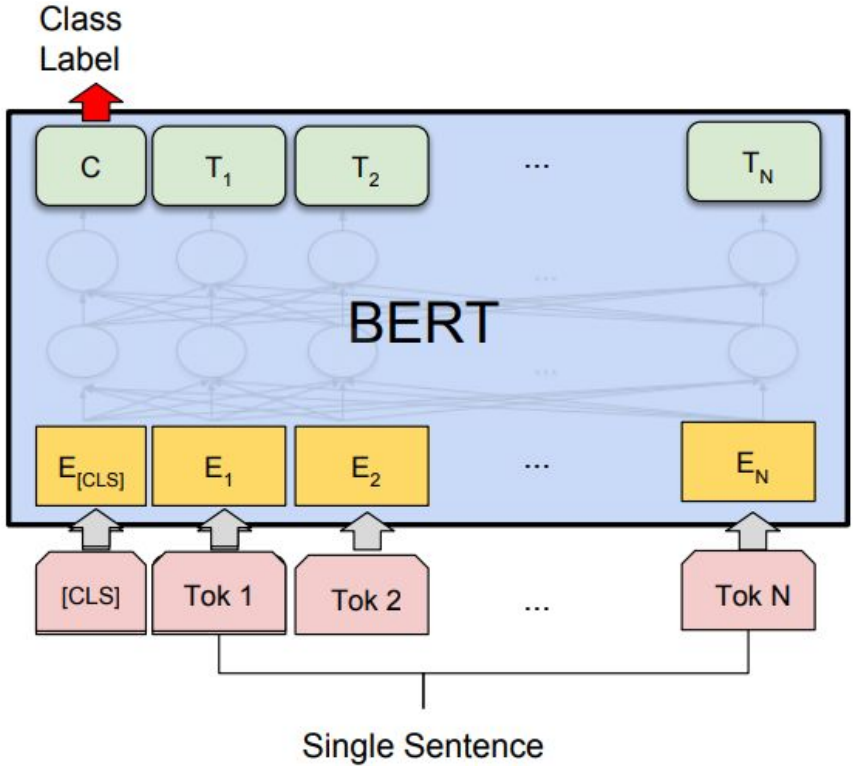
Pre-training on Biomedical Ontologies

- BERT model pre-trained on 9 different Biomedical Ontologies
    - Using corpus generated by Onto2Sen
- Focus on Masked Language Modelling task (MLM)
- Enhances understanding of medical terminologies
- Injects Biomedical Ontology concepts, relationship and properties



BERT-based MLM model

Pre-training BERT on Onto2Sen corpus

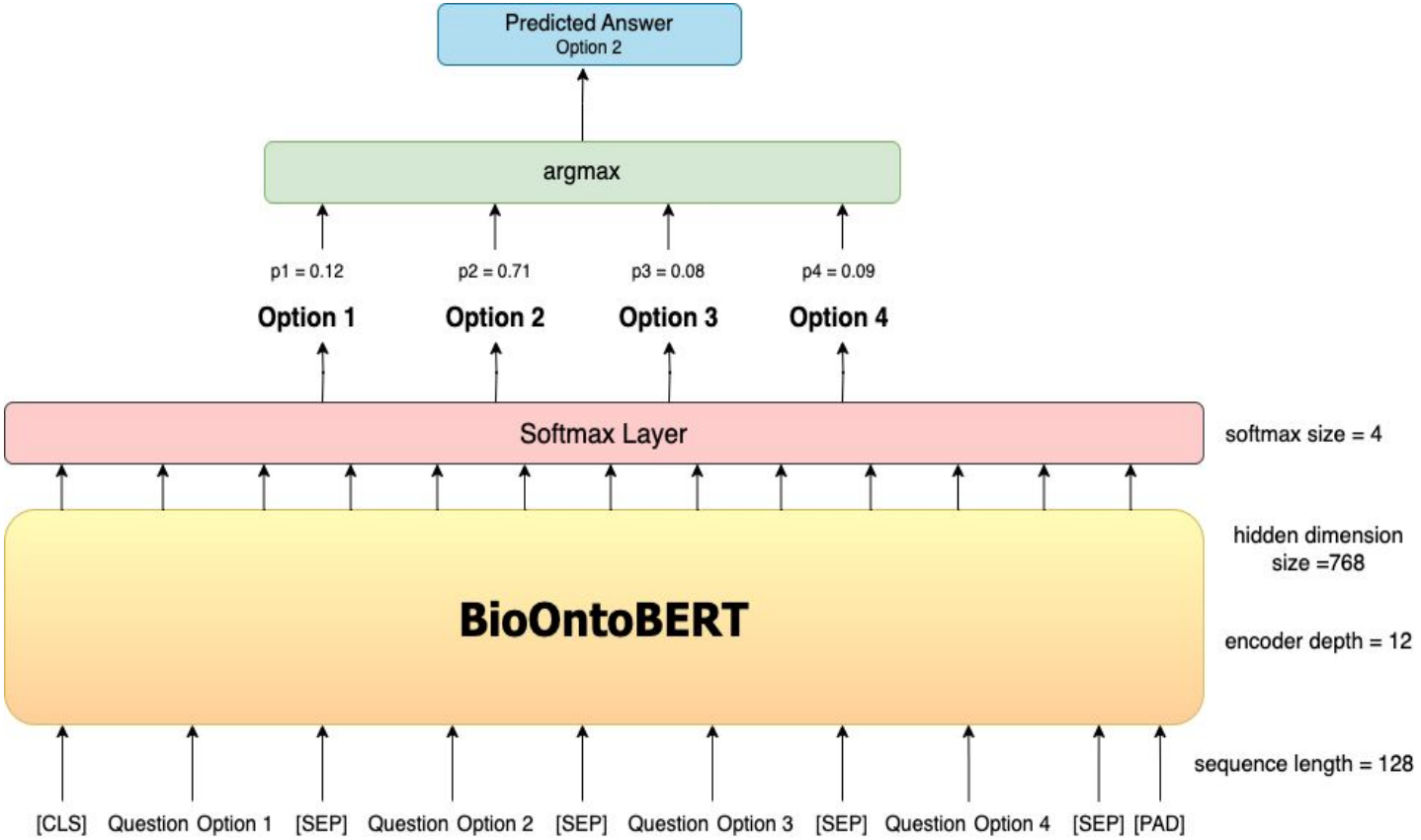


Source - Devlin, et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2.

BioOntoBERT

Pre-trained BERT model on Biomedical Ontologies

Fine-tuning on MCQA task



Results

BioOntoBERT outperformed baseline BERT models including PubmedBERT

Models	Corpus	Text Size	Accuracy
BERT	Wiki + Books	-	35%
BioBERT	PubMed	4.5B Words	38%
SciBERT	PMC + CS	3.2B words	39%
PubmedBERT	PubMed	3.1B words | 21GB	40%
BioOntoBERT (proposed)	Biomedical Ontologies	**20M words | 158 MB**	**42.72%**

Observations

Question : Dentigerous cyst is likely to cause which neoplasia?

a. Ameloblastoma

BioOntoBERT 

b. Adenocarcinoma

PubmedBERT 

c. Fibrosarcoma

d. All of the Above

- Option a, b, c and ‘Neoplasia’ are present in the DOID ontology
- ‘Dentigerous cyst’ is a type of ‘Odontogenic cyst’
- Odontogenic cyst and Odontogenic epithelium are closely related, latter has a reference in DOID ontology

Conclusions

Onto2Sen System

Generates ontology-based sentences

9 Biomedical Ontologies

Captures medical concepts & terminologies from different ontologies

158MB medical corpus

0.70 % of Pubmed Abstracts

BioOntoBERT Model

Pre-trained BERT on Biomedical Ontologies

Foundational Knowledge

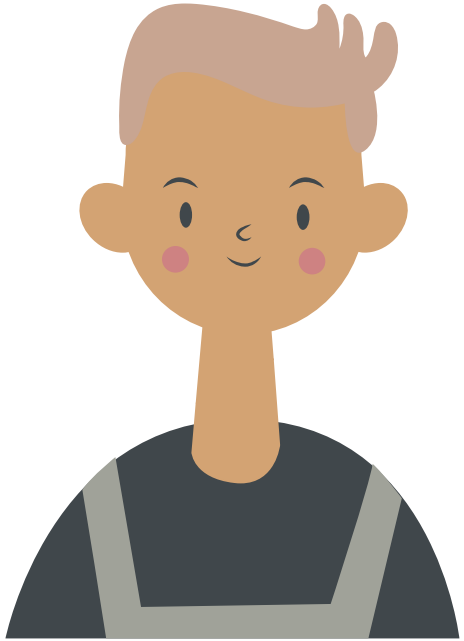
Captures foundational medical knowledge, not covered in medical articles and papers

Outperforms PubmedBERT

Improvement of +2.72 %, while significantly reducing computational costs

Future Work

- Incorporating additional Biomedical Ontologies
- Integrating more properties and exploring different fine-tuning strategies
- Extending the approach to other specialized domains can unlock broader applications in NLP



Thank you