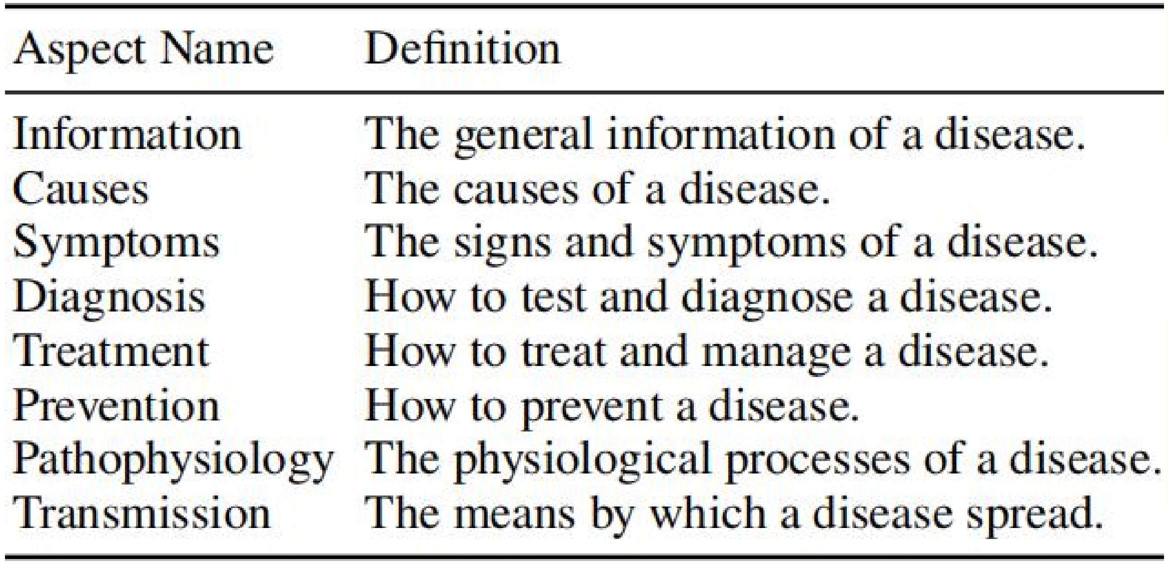
Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition

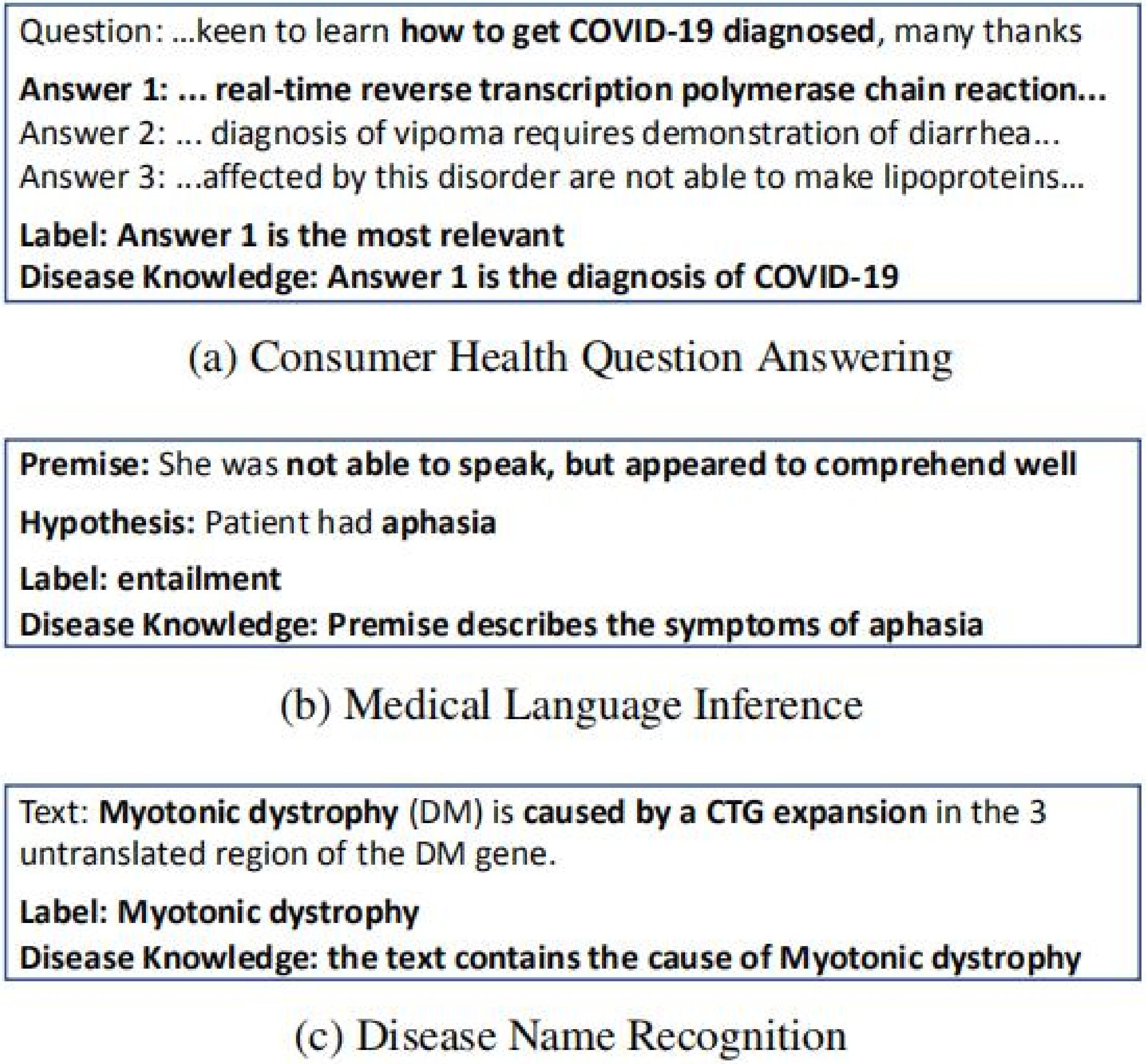
Yun He[1], Ziwei Zhu[1], Yin Zhang[1], Qin Chen[2], James Caverlee[1]
[1]Texas A&M University, College Station, USA
[2]Fudan University, Shanghai, China
{yunhe, zhuziwei, zhan13679, caverlee}@tamu.edu
qin_chen@fudan.edu.cn

Related Work

- Knowledge-Enriched BERT
- Biomedical BERT
- Biomedical Knowledge Integration Methods with UMLS



Aspect Name	Definition
Information	The general information of a disease.
Causes	The causes of a disease.
Symptoms	The signs and symptoms of a disease.
Diagnosis	How to test and diagnose a disease.
Treatment	How to treat and manage a disease.
Prevention	How to prevent a disease.
Pathophysiology	The physiological processes of a disease.
Transmission	The means by which a disease spread.



Question: ...keen to learn **how to get COVID-19 diagnosed**, many thanks

Answer 1: ... real-time reverse transcription polymerase chain reaction...
Answer 2: ... diagnosis of vipoma requires demonstration of diarrhea...
Answer 3: ...affected by this disorder are not able to make lipoproteins...

Label: Answer 1 is the most relevant
Disease Knowledge: Answer 1 is the diagnosis of COVID-19

(a) Consumer Health Question Answering

Premise: She was **not able to speak, but appeared to comprehend well**

Hypothesis: Patient had **aphasia**

Label: entailment

Disease Knowledge: Premise describes the symptoms of aphasia

(b) Medical Language Inference

Text: **Myotonic dystrophy** (DM) is **caused by a CTG expansion** in the 3 untranslated region of the DM gene.

Label: Myotonic dystrophy

Disease Knowledge: the text contains the cause of Myotonic dystrophy

(c) Disease Name Recognition

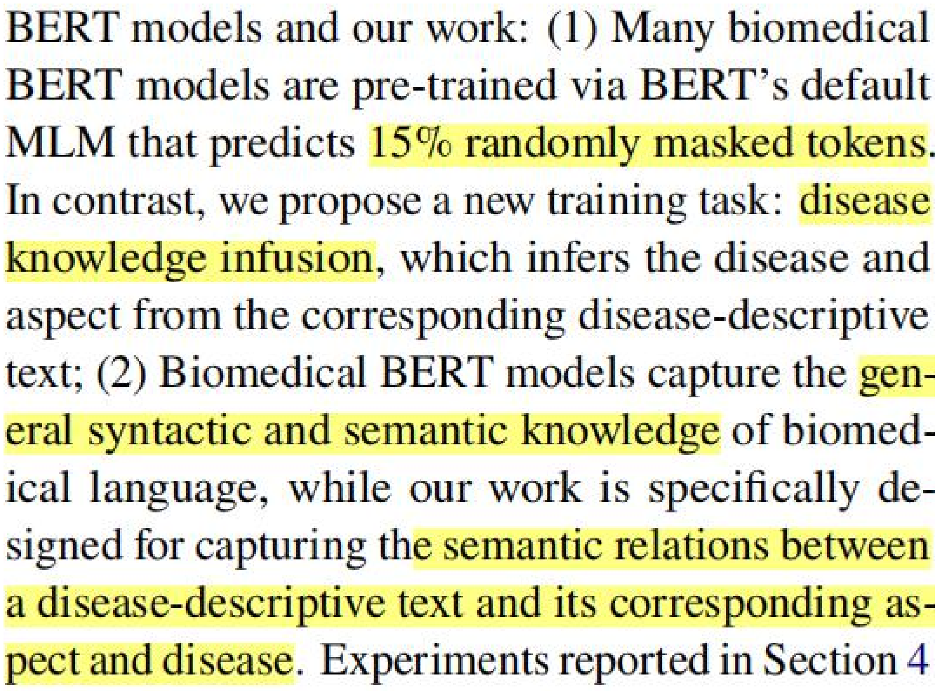
Summary

- Proposed Method: **Disease Knowledge Infusion Training**
- Goal: integrate BERT-like pre-trained language models with disease knowledge to achieve better performance on a variety of medical domain tasks including answering health questions, medical language inference, and disease name recognition.
- The approach is guided by three questions:

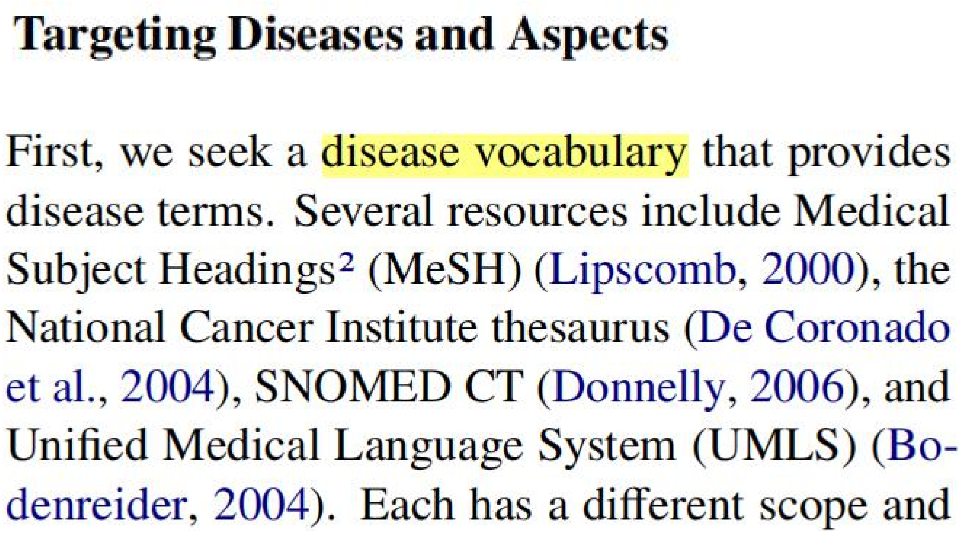
✓ Which diseases and aspects should we focus on?

✓ How do we infuse disease knowledge into BERT-like models?

✓ What is the objective function of this training task?



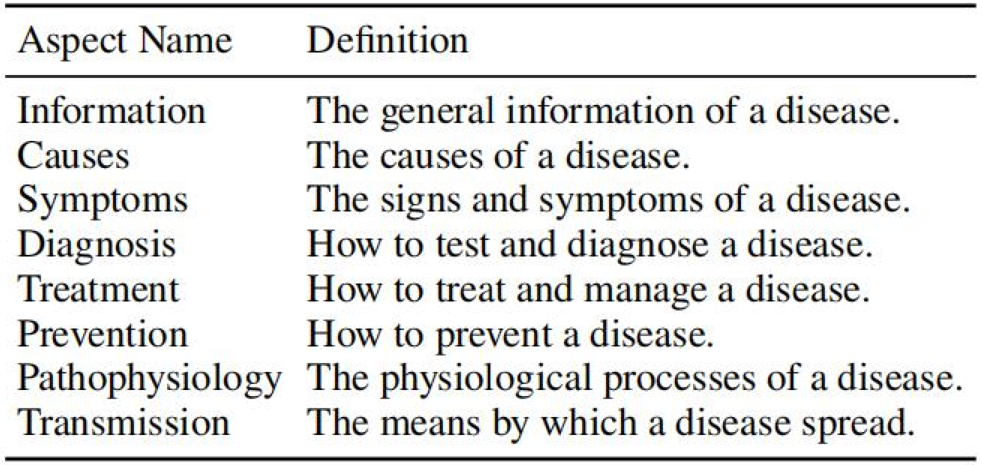
BERT models and our work: (1) Many biomedical BERT models are pre-trained via BERT's default MLM that predicts 15% randomly masked tokens. In contrast, we propose a new training task: disease knowledge infusion, which infers the disease and aspect from the corresponding disease-descriptive text; (2) Biomedical BERT models capture the general syntactic and semantic knowledge of biomedical language, while our work is specifically designed for capturing the semantic relations between a disease-descriptive text and its corresponding aspect and disease. Experiments reported in Section 4



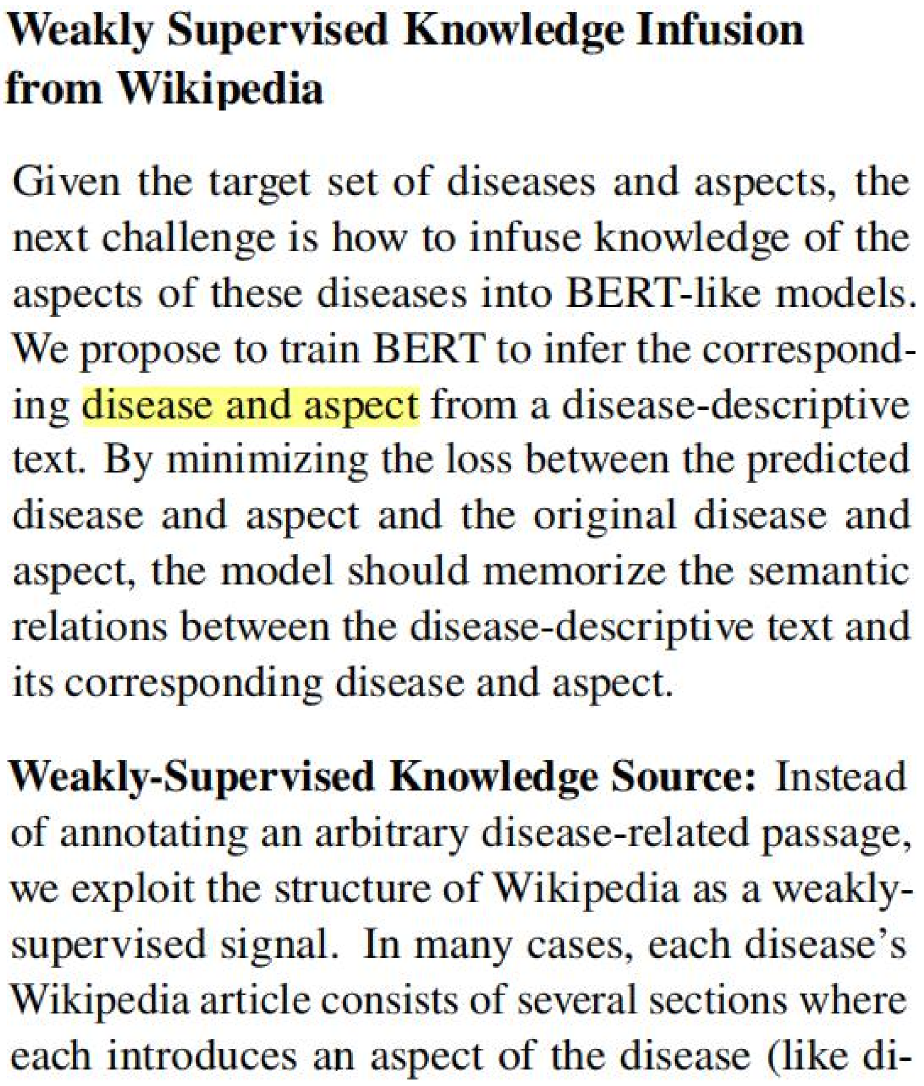
Targeting Diseases and Aspects

First, we seek a disease vocabulary that provides disease terms. Several resources include Medical Subject Headings[2] (MeSH) (Lipscomb, 2000), the National Cancer Institute thesaurus (De Coronado et al., 2004), SNOMED CT (Donnelly, 2006), and Unified Medical Language System (UMLS) (Bodenreider, 2004). Each has a different scope and

5,853 total disease terms
14,617 passages



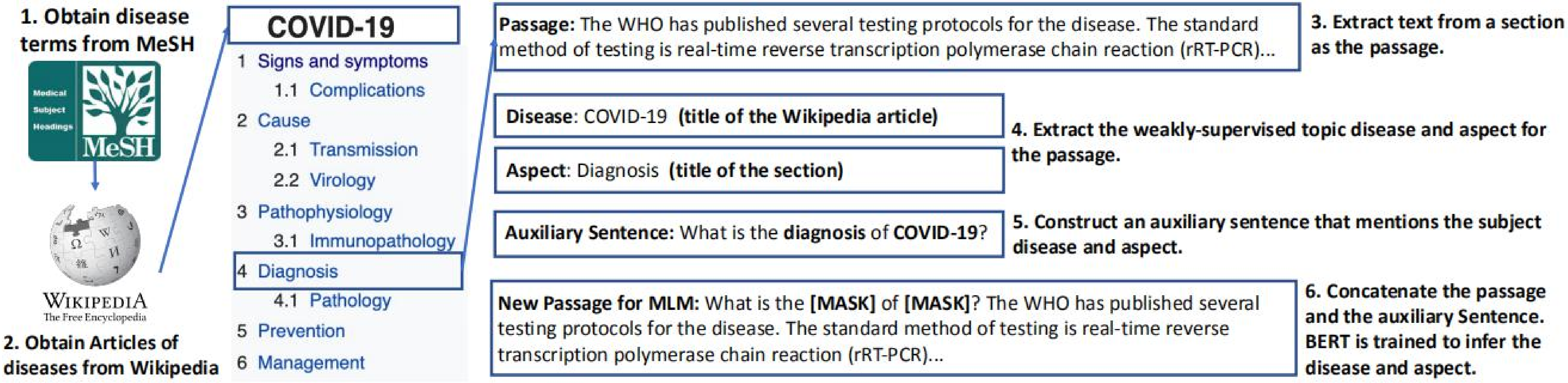
Aspect Name	Definition
Information	The general information of a disease.
Causes	The causes of a disease.
Symptoms	The signs and symptoms of a disease.
Diagnosis	How to test and diagnose a disease.
Treatment	How to treat and manage a disease.
Prevention	How to prevent a disease.
Pathophysiology	The physiological processes of a disease.
Transmission	The means by which a disease spread.



Weakly Supervised Knowledge Infusion from Wikipedia

Given the target set of diseases and aspects, the next challenge is how to infuse knowledge of the aspects of these diseases into BERT-like models. We propose to train BERT to infer the corresponding disease and aspect from a disease-descriptive text. By minimizing the loss between the predicted disease and aspect and the original disease and aspect, the model should memorize the semantic relations between the disease-descriptive text and its corresponding disease and aspect.

Weakly-Supervised Knowledge Source: Instead of annotating an arbitrary disease-related passage, we exploit the structure of Wikipedia as a weakly-supervised signal. In many cases, each disease's Wikipedia article consists of several sections where each introduces an aspect of the disease (like di-



1. Obtain disease terms from MeSH

2. Obtain Articles of diseases from Wikipedia

Passage: The WHO has published several testing protocols for the disease. The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...

3. Extract text from a section as the passage.

Disease: COVID-19 **(title of the Wikipedia article)**

Aspect: Diagnosis **(title of the section)**

4. Extract the weakly-supervised topic disease and aspect for the passage.

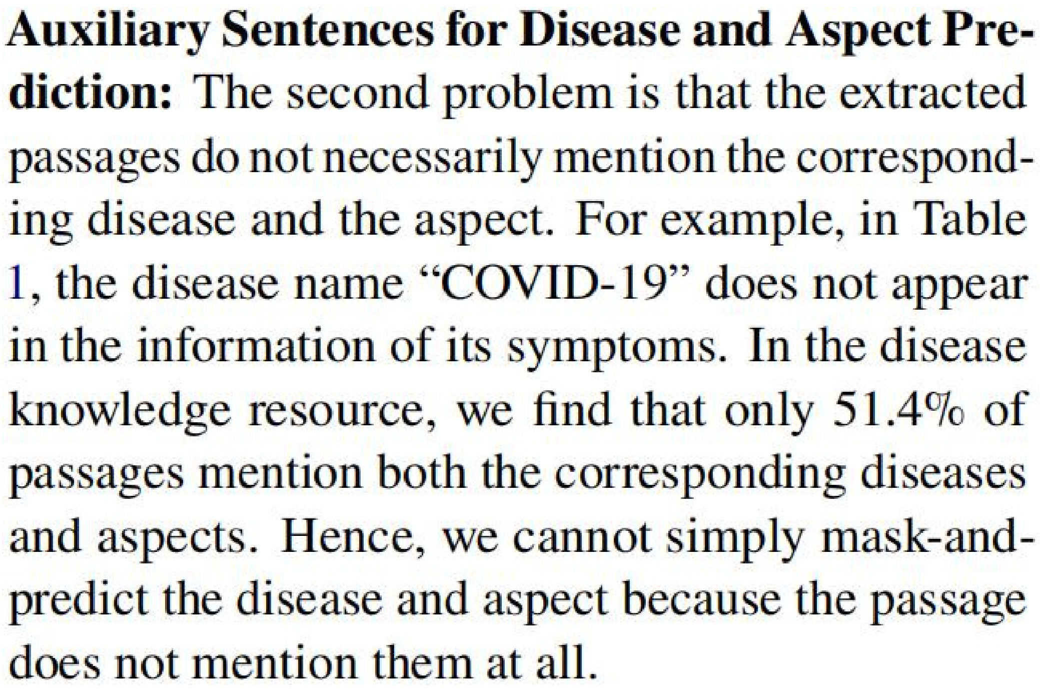
Auxiliary Sentence: What is the **diagnosis** of **COVID-19**?

5. Construct an auxiliary sentence that mentions the subject disease and aspect.

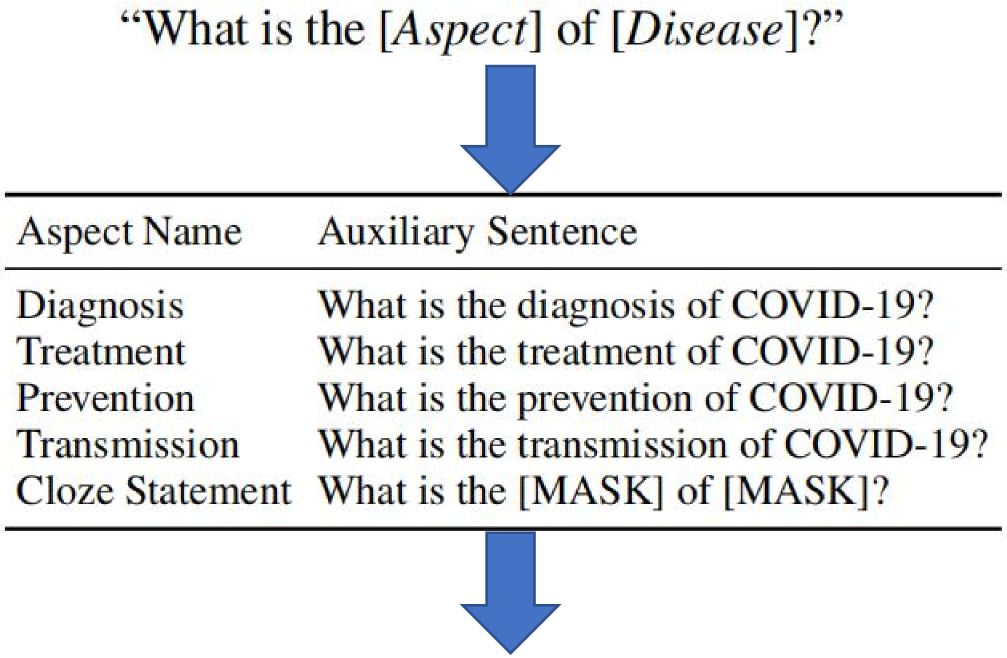
New Passage for MLM: What is the **[MASK]** of **[MASK]**? The WHO has published several testing protocols for the disease. The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...

6. Concatenate the passage and the auxiliary Sentence. BERT is trained to infer the disease and aspect.

Figure 2: Disease Knowledge Infusion Training: An example with COVID-19.

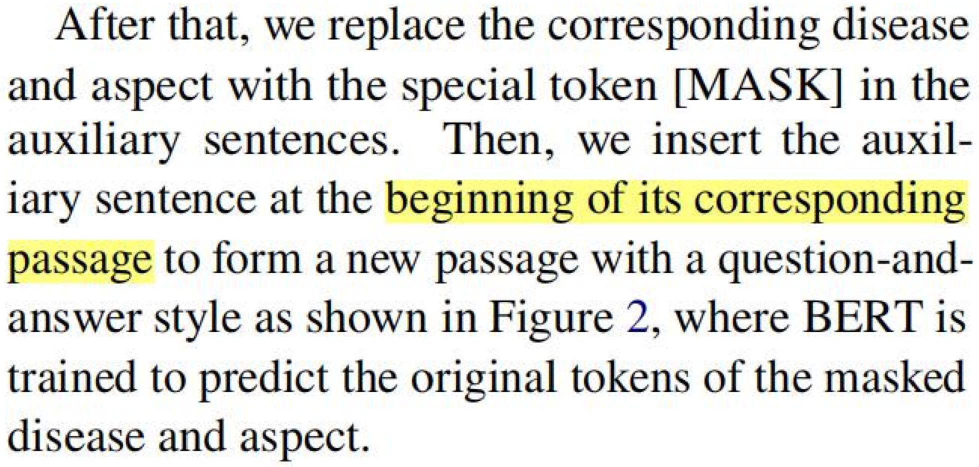


Auxiliary Sentences for Disease and Aspect Prediction: The second problem is that the extracted passages do not necessarily mention the corresponding disease and the aspect. For example, in Table 1, the disease name "COVID-19" does not appear in the information of its symptoms. In the disease knowledge resource, we find that only 51.4% of passages mention both the corresponding diseases and aspects. Hence, we cannot simply mask-and-predict the disease and aspect because the passage does not mention them at all.

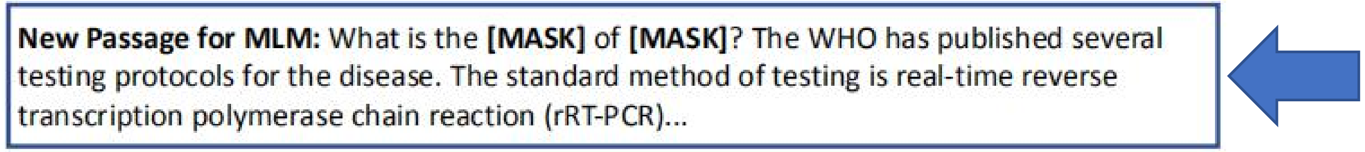


"What is the [*Aspect*] of [*Disease*]?"

Aspect Name	Auxiliary Sentence
Diagnosis	What is the diagnosis of COVID-19?
Treatment	What is the treatment of COVID-19?
Prevention	What is the prevention of COVID-19?
Transmission	What is the transmission of COVID-19?
Cloze Statement	What is the [MASK] of [MASK]?



After that, we replace the corresponding disease and aspect with the special token [MASK] in the auxiliary sentences. Then, we insert the auxiliary sentence at the beginning of its corresponding passage to form a new passage with a question-and-answer style as shown in Figure 2, where BERT is trained to predict the original tokens of the masked disease and aspect.



New Passage for MLM: What is the **[MASK]** of **[MASK]**? The WHO has published several testing protocols for the disease. The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...

Training Objective and Details

Finally, we show the objective function of disease infusion training. Since most disease names are out of BERT vocabulary, the WordPiece tokenizer (Wu et al., 2016) will split these terms into sub-word tokens that exist in the vocabulary. For example, “COVID-19" will be split into 4 tokens: “co", “vid", “-" and “19". Formally, let $X = (x_1, ..., x_T)$ denote a sequence of T tokens that are split from a disease name where x_t is the t-th token. The original cross-entropy loss is to get the conditional probability of a masked token as close as possible to the 1-hot vector of the token:

$$\mathcal{L}_{disease} = -\sum_{t=1}^{T} log\ p(x_t|passage) \tag{1}$$

$$p(x_t|passage) = \frac{exp(z_t)}{\sum_{z \in \mathcal{V}} exp(z)} \tag{2}$$

$$z_t = \mathbf{w} \cdot \mathbf{y}_t + b \tag{3}$$

$$\mathcal{L}_{disease} = -\sum_{t=1}^{T} logp(x_t|passage) + \frac{\beta}{\sum_{t=1}^{T} z_t} \tag{4}$$

$$\mathcal{L}_{aspect} = -log\ p(a|passage)$$

$$\mathcal{L} = \mathcal{L}_{disease} + \mathcal{L}_{aspect}$$

Experiments

- **BERT-base** architecture (12 layers and 768 hidden embedding size with 108M parameters)
- **ALBERT-xxlarge** (12 layers and 4096 hidden embedding size with 235M parameters
- **BioBERT**(the first BERT pre-trained on biomedical corpora), initialized with BERT's pre-trained parameters (108M) and then further trained over PubMed abstracts (4.5B words) and PubMed Central full-text articles (13.5B words)--BioBERT v1.1
- **ClinicalBERT**(initialized from BioBERT v1.0), based on discharge summaries of clinical notes: Bio-Discharge Summary BERT
- **BlueBERT**, firstly initialized from BERT-base and further pre-trained over a biomedical corpus of PubMed abstracts and clinical notes
- **SciBERT**,firstly initialized from BERT-base model and pre-trained on a random sample of the full text of 1.14M papers from Semantic Scholar, with 18% ofpapers from the computer science domain and 82% from the biomedical domain.

Tasks	Consumer Health Question Answering						NLI	NER	
Datasets	MEDIQA-2019			TRCEQA-2017			MEDNLI	BC5CDR	NCBI
Metrics(%)	Accuracy	MRR	Precision	Accuracy	MRR	Precision	Accuracy	F1	F1
BERT	64.95	82.72	66.49	74.61	56.17	52.55	75.95	83.09	85.14
BERT + disease*	66.40↑	83.33↑	68.94↑	75.33↑	56.41↑	54.01↑	77.29↑	83.47↑	86.81↑
BlueBERT	65.13	81.50	67.35	74.26	48.40	52.55	82.21	85.73	87.78
BlueBERT + disease	68.47↑	81.17	71.57↑	77.59↑	50.96↑	57.62↑	83.90↑	86.30↑	87.79↑
ClinicalBERT	67.30	84.78	70.59	77.00	52.56	56.62	81.50	84.90	87.25
ClinicalBERT + disease	69.02↑	88.94↑	69.84	78.90↑	54.97↑	60.40↑	81.65↑	85.63↑	87.22
SciBERT	68.47	84.47	68.07	77.23	54.57	57.54	80.94	86.16	87.24
SciBERT + disease	73.35↑	85.44↑	76.28↑	79.02↑	56.57↑	59.57↑	82.14↑	86.34↑	88.30↑
BioBERT	68.29	83.61	72.78	77.12	49.84	57.25	81.86	85.99	87.70
BioBERT + disease	72.09↑	87.78↑	74.40↑	78.43↑	54.76↑	58.45↑	82.21↑	86.52↑	87.14
ALBERT	76.54	88.46	81.41	75.09	**58.57**	53.03	85.48	84.28	87.56
ALBERT + disease	**79.49**↑	90.00↑	**84.02**↑	**80.10**↑	57.21	**62.40**↑	**86.15**↑	84.71↑	87.69↑
SOTA*	78.00	**93.67**	81.91	77.23	54.57	57.54	84.00	**87.15**	**89.71**

* SOTA, state-of-the-art as of May 2020, to the best of our knowledge.

* “ + disease" means that we train BERT via disease knowledge infusion training before fine-tuning.

Table 6: Ablation Study on MEDIQA-2019

Variants	Accuracy	MRR	Precision
Default	79.49	90.00	84.02
- Auxiliary Sentence	78.23	90.89	78.10
- Aspect Prediction	78.41	89.06	80.00
- Disease Prediction	72.90	85.72	79.44
15% Randomly Masked Tokens	77.06	87.33	85.18