Institute of Systems Science
National University of Singapore

MASTER OF TECHNOLOGY IN ENTERPRISE BUSINESS ANALYTICS / INTELLIGENT SYSTEMS

Graduate Certificate Online Examination
Semester I 2021/2022

Subject: ***Practical Language Processing***

Instructions for Paper

Date:	Wednesday 10 Nov 2021
Time:	6.30 p.m.
Duration:	Three hours (7.00 p.m. to 10.00 p.m.)
Place:	Online Examination

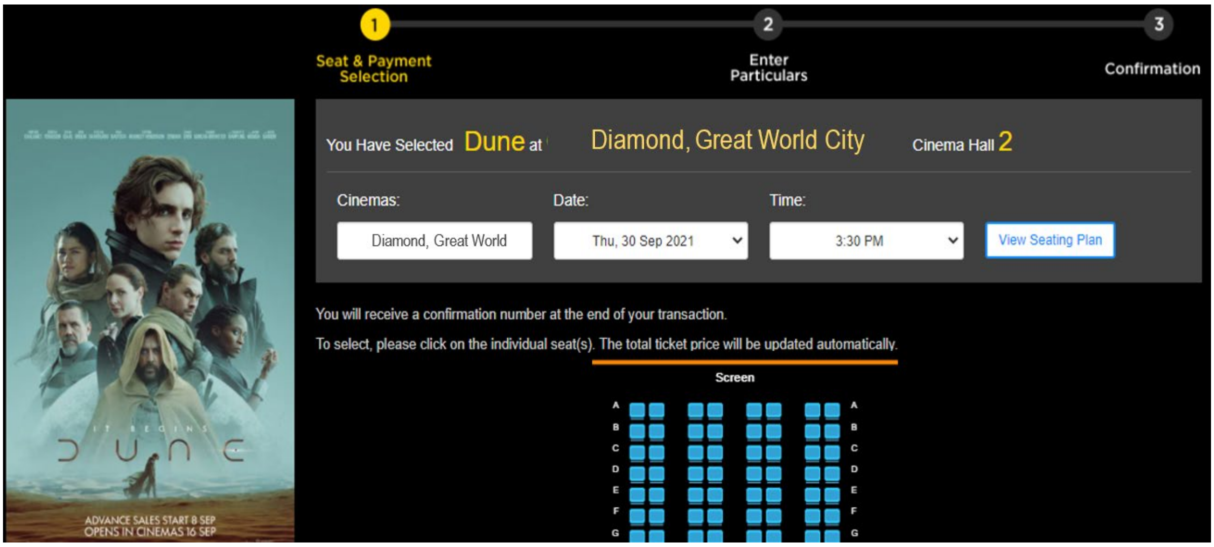
This is an OPEN BOOK examination. This examination paper consists of *one* Section and *two* Questions. You are to answer *ALL* questions. There are a total of *50 Marks* for this paper.

1. Read **ALL** instructions before answering any of the examination questions.
2. There will be only **ONE question paper**, which may be organized in a series of one or more sections (section A, B, C etc.) with/without appendixes. Each section contains one or more questions. You are required to answer **ALL** questions in the **SEPARATE answer booklet(s)**, according to different sections, downloaded from LumiNUS.
3. The first 30 minutes will be reading time, during which you **must not** start answering your answers.
4. Write your NUS Student ID number on the **front page** of the Answer Booklet(s) in the box provided.
5. This is an ***Open Book*** examination. If you wish, you may use reference materials to answer a question. Reference materials can be books, manuals, handouts or notes, including e-notes on PCs/laptops.
6. All answers provided should be in **digital format**. However, you can **hand draw** diagrams on paper using pens and ensure that it is readable as an image. **Insert this image into your answer booklet**. Do not send images as separate files.
7. Non-programmable calculators may be used if required.
8. **Internet access (except LumiNUS and Zoom) using computers of any form e.g. laptops, tablets, smart watches etc. is not permitted during the examination**. Students who are found with suspected academic dishonesty that give them unfair advantage during assessments will be subjected to disciplinary action by the University, as laid out in NUS Code of Student Conduct.
9. State clearly any assumptions you make in answering any question where you feel the requirement is not sufficiently clear.
10. At the end of the examination:
    a) Convert the answer booklet to PDF format and compress them if necessary.
    b) Please name the PDF file with your **Student ID number** prefaced by the abbreviation of the Course and the Section e.g. **PLP_SecA_A0123456X**.
    c) Upload the answer booklet for **each Section separately**.
11. After submission of answer booklet(s), please wait for Proctor to make announcement on the closure of examination.

SECTION A

Question 1 *(Total: 30 Marks)*

StarShows is a website that offers online movie ticket booking services for local cinemas. Currently their users can buy tickets for a movie on show through the website UI (Figure 1). The process is typically as follows: a user starts by selecting the movie he wants to watch, the cinema showing it, and the date and time of the show, from drop-down menus and pick the preferred seats from a seat map; the contact details (phone number and email address) of the user are collected before the user pays for the tickets through a third-party payment system. After the transaction is completed, a confirmation email and an SMS message are sent to the user with the booking details.



You Have Selected Dune at Diamond, Great World City Cinema Hall 2

Cinemas: Diamond, Great World

Date: Thu, 30 Sep 2021

Time: 3:30 PM

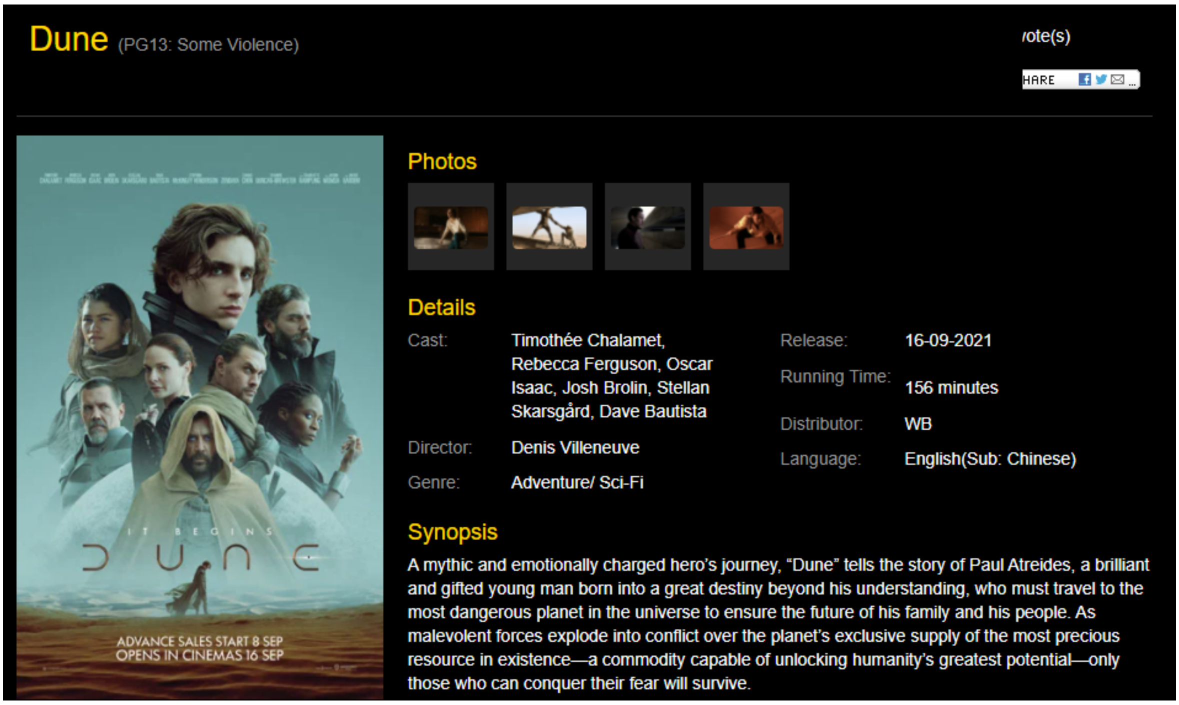
View Seating Plan

You will receive a confirmation number at the end of your transaction.

To select, please click on the individual seat(s). The total ticket price will be updated automatically.

Figure 1 Web UI for ticket booking

The website also provides some basic information about the movies on show or coming soon (Figure 2), such as running time, genre, language, director, cast, and synopsis. Users can browse movie information to help decide which movie to watch.



Dune (PG13: Some Violence)

Photos

Details

Cast: Timothée Chalamet, Rebecca Ferguson, Oscar Isaac, Josh Brolin, Stellan Skarsgård, Dave Bautista

Director: Denis Villeneuve

Genre: Adventure/ Sci-Fi

Release: 16-09-2021

Running Time: 156 minutes

Distributor: WB

Language: English(Sub: Chinese)

Synopsis

A mythic and emotionally charged hero's journey, "Dune" tells the story of Paul Atreides, a brilliant and gifted young man born into a great destiny beyond his understanding, who must travel to the most dangerous planet in the universe to ensure the future of his family and his people. As malevolent forces explode into conflict over the planet's exclusive supply of the most precious resource in existence—a commodity capable of unlocking humanity's greatest potential—only those who can conquer their fear will survive.

Figure 2 Movie Information Page

To further improve the movie ticket booking service and make it a seamless and hassle free experience for the users, ***StarShows*** wants to make it available as a mobile app with a chat bot to facilitate the booking process.

StarShows primarily would like the chat bot to help its users to book movie tickets through chatting, instead of the web UI. Figure 3 shows a sample conversation between the chat bot and a user.

User: *Hey, could you book some movie tickets for me?*

Bot: *Certainly, what would you like to see?*

User: *My brother and I want to see No Time To Die.*

Bot: *So, two tickets then?*

User: *Yes, that's right.*

Bot: *Which cinema would you like to watch it at?*

User: *The one at Vivocity.*

Bot: *Ok, Diamond Vivocity. The available show times for today are 6pm and 9pm. Which one would you prefer?*

User: *9pm.*

Bot: *Alright. Two tickets for No Time To Die at Diamond Vivocity, at 9pm today. Is that right?*

User: *Correct.*

Bot: *Do you have any preference of the seat positions?*

User: *Not really.*

Bot: *I have reserved seats G5 and G6 for you.* (Showing selected seats on Seating Plan)

User: *Ok.*

Bot: *It'll be $20 in total. Please proceed to the payment to confirm the booking.* (Displaying payment link)

User: *Thank you!*

Figure 3 Sample Conversation

Secondly, they also want the bot to be able to answer simple factual questions about current movies (either on show or coming soon) so that users don't need to check the webpage when they want more information. Figure 4 shows some sample questions that they expect from the users.

How long is the movie?

What is its genre?

Is it suitable for young children?

Who is the director of the movie?

What is its age-rating?

What action movies are showing?

Figure 4 Sample User Questions

You are engaged by the company as a consultant to help their development team to build the chat bot.

Answer the following questions based on the information provided above.

a. Your team member, Jason, is very impressed by end-to-end open-domain chatbots like Google's Meena. He proposes to build the movie ticket booking chatbot using similar end-to-end approach, and summarises the outline of his solution in the following table.

1.	Architecture	End-to-end conversational modelling
2.	Approach	Adapting pre-trained transformer model (GPT-2) with conversation data. The model should generate the response based on dialogue input (dialog history)
3.	Training data	Dialogue pairs (*input, response*) extracted from Reddit discussion chains
4.	Response generation using adapted model	Decoding method: beam search.
5.	Model evaluation	Human assessment of model generated responses – is it appropriate, informative and human-like?

Table 1: Solution outline

Critique **each point** in Jason's solution outline. Justify your answer based on the requirements of the given case study.

(10 Marks)

b. The team agrees that there should be an intent called "movie-ticket-booking" for the bot to handle the ticket booking process.

(1) Define the slots for this intent.

(1 Mark)

(2) Draw the high level conversation flow for this intent based on the sample conversation (without the exception handling paths).

(3 Marks)

c. Jason also suggests to tap on the power of pre-trained transformer models like BERT to support the chat bot's function of question answering for movie information. He has found

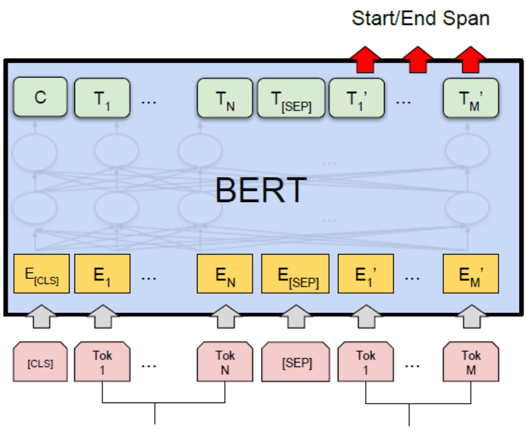


Figure 5 BERT Fine-tuned for Question Answering

a BERT-based model already fine-tuned for the extractive question answering task using SQuAD dataset (as illustrated in Figure 5). He thinks this model should be adequate for the factual questions about movies (samples shown in Figure 4) without further training using movie-specific data. Do you agree with his suggestion? Justify your answer based on the information given in the case study.

(3 Marks)

d. The company reckons that sometimes a user may not specifically know what to watch. They want the chat bot to be able to help the user by arbitrarily suggesting a movie based on a genre preferred by the user, and, more importantly, add one or two sentences explaining why the movie is worth watching. A sample dialogue is shown in Figure 6.

User: *Can you recommend a movie? I like actions and adventures.*

Bot: ***How about "Dune"? It is a faithful adaptation to the fantastic 1965 sci-fi novel by Frank Herbert. A great modern Sci-Fi.***

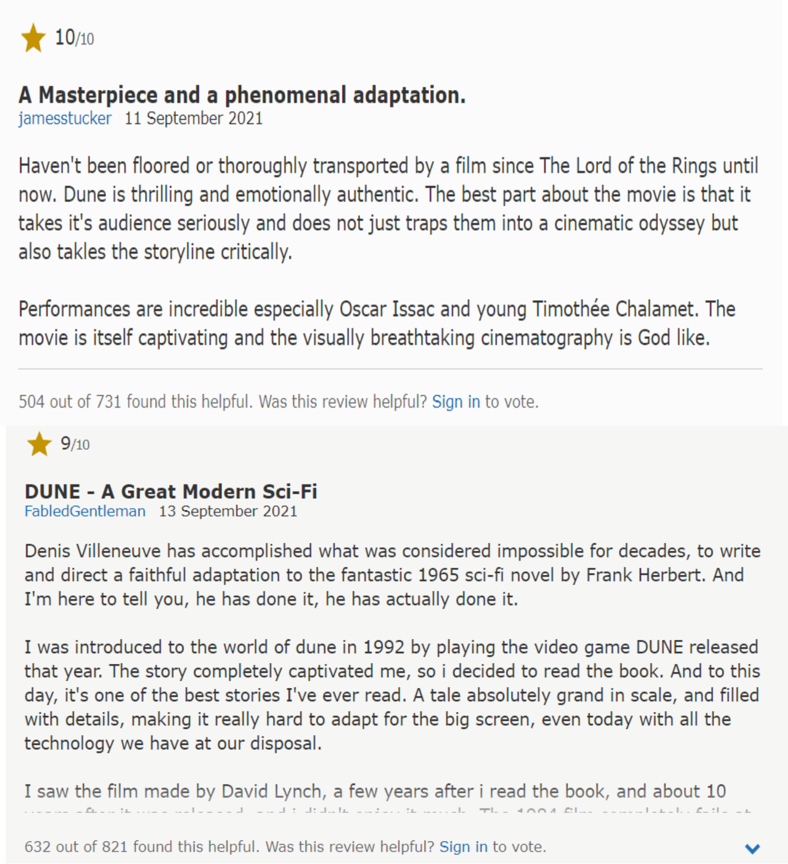
User: *Sounds good! When is the next show?*

Bot: *5:45pm at Dimond Vivocity.*

...

Figure 6 Sample conversation for movie recommendation

The team would like to tap on movie reviews available on IMDB website. Figure 7 shows some sample user reviews.



10/10

A Masterpiece and a phenomenal adaptation.
jamesstucker 11 September 2021

Haven't been floored or thoroughly transported by a film since The Lord of the Rings until now. Dune is thrilling and emotionally authentic. The best part about the movie is that it takes it's audience seriously and does not just traps them into a cinematic odyssey but also takles the storyline critically.

Performances are incredible especially Oscar Issac and young Timothée Chalamet. The movie is itself captivating and the visually breathtaking cinematography is God like.

504 out of 731 found this helpful. Was this review helpful? Sign in to vote.

9/10

DUNE - A Great Modern Sci-Fi
FabledGentleman 13 September 2021

Denis Villeneuve has accomplished what was considered impossible for decades, to write and direct a faithful adaptation to the fantastic 1965 sci-fi novel by Frank Herbert. And I'm here to tell you, he has done it, he has actually done it.

I was introduced to the world of dune in 1992 by playing the video game DUNE released that year. The story completely captivated me, so i decided to read the book. And to this day, it's one of the best stories I've ever read. A tale absolutely grand in scale, and filled with details, making it really hard to adapt for the big screen, even today with all the technology we have at our disposal.

I saw the film made by David Lynch, a few years after i read the book, and about 10

632 out of 821 found this helpful. Was this review helpful? Sign in to vote.

Figure 7 Sample reviews

Based on the above requirement and information given, design a solution for the task of movie recommendation with explanation. Briefly explain the main steps.

(3 Marks)

e. The company also decided to add speech recognition, speech synthesis and speaker recognition functions to the mobile app so that the chatbot can support spoken interaction. Please answer the following questions regarding the speech functions.

(1) In the spoken chatbot, some text may be challenging to the text-to-speech (TTS) system. For the text in Figure 3 and Figure 6, do you identify any text that TTS system needs to normalize in order to read them correctly?

(3 Marks)

(2) The development team tested the speech recognition system. They found that the speech recognition results of some sentences are not recognized correctly. For example, the following errors are identified. If the team is planning to collect some data to improve the recognition accuracy of the speech recognition system, what should they do to improve the system?

No	User's actual speech	Recognition result
1	*could you book some tickets for me?*	*could you book some takes for me?*
2	*what is its age-rating?*	*what is it's a greeting?*
3	*please reserve two tickets for me.*	*Please reserve two packets for me.*

(4 Marks)

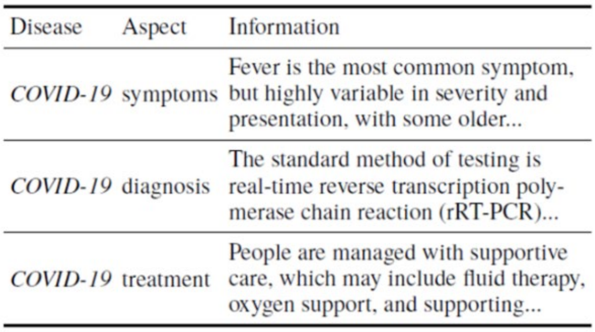
(3) The team is using a voiceprint system to allow the mobile app user to enter his/her account without a password. Suppose the user's account name has already been stored in the app, should speaker identification or speaker verification be used in this case? Please explain the reason and state how this can be done.

(3 Marks)

Question 2 *(Total: 20 Marks)*

Healthcare is one of the major success stories of our times. Medical science has improved rapidly, raising life expectancy around the world. However, as longevity increases, healthcare systems face growing demand for their services, rising costs and a workforce that is struggling to meet the needs of its patients. Building on automation, artificial intelligence (AI) has the potential to revolutionize healthcare and help address some of the challenges set out above. It can increase productivity and the efficiency of care delivery and allow healthcare systems to provide more and better care to more people. AI can help improve the experience of healthcare practitioners, enabling them to spend more time in direct patient care and reducing burnout.

You are currently leading a team to design and develop AI solutions to help healthcare organizations apply cognitive technology to unlock vast amounts of health data and power diagnosis. The product prototype named ***MediBert*** is designed to firstly review and store far more medical information, including medical journal, symptom, and case study of treatment. Then the knowledge about diseases has been extracted from those medical information and represented as an abstraction of tuple <*Disease, Aspect, Information*>. Knowledge of a disease includes information about various aspects of the disease, like the signs, and symptoms, diagnosis, and treatment. As an example, Table 2 highlights several aspects for COVID-19. Eventually the ***MediBert*** system could respond to public users' queries related to any diseases and provide specialised assistance to doctors and nurses.



Disease	Aspect	Information
COVID-19	symptoms	Fever is the most common symptom, but highly variable in severity and presentation, with some older...
COVID-19	diagnosis	The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...
COVID-19	treatment	People are managed with supportive care, which may include fluid therapy, oxygen support, and supporting...

Table 2: **Disease knowledge of *COVID-19* is presented from three aspects: symptoms, diagnosis and treatment**

a. In the progress of processing the medical articles, one major task with great business value is to recognize the disease names from the huge volume of plain text. This will better prepare the system for dealing with user queries by precisely locating the relevant knowledge upon user's question. For example in the sample text: "*Myotonic dystrophy (DM) is caused by a CTG expansion in the 3 untranslated region of the DM gene.*", "*Myotonic dystrophy"* should be labeled as disease name. Your teammate Bob suggests to utilize the exsisting pre-trained named entity recognizer from StanfordNLP and Spacy to generate such labels. Please evaluate the feasibility of his suggestion.

(2 Marks)

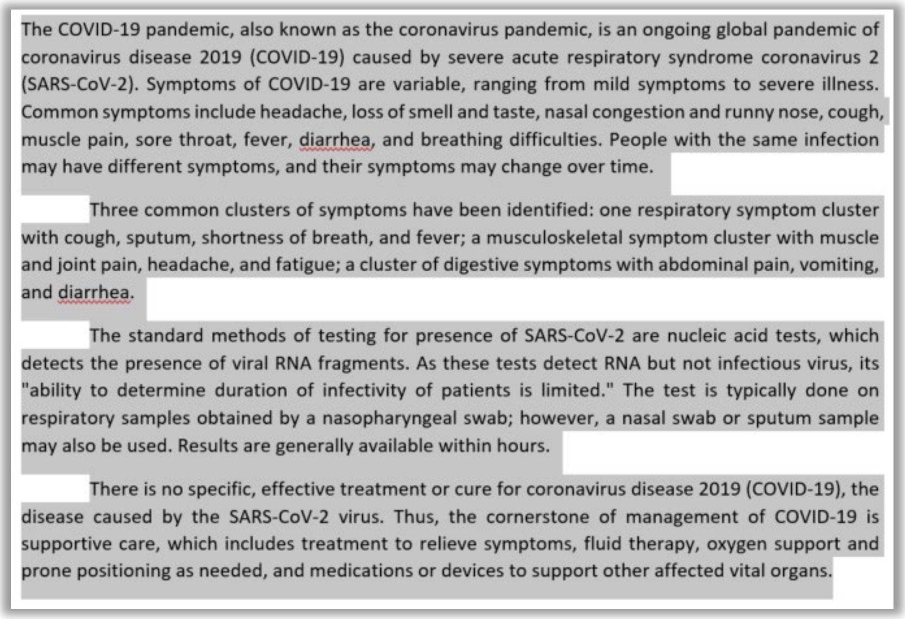
b. In order to train a disease name recognizer using CRF model, the word tokens should be annotated according to their roles in the context. And considering the situation that the role of disease name can span over more than one word tokens, thus there will be three types of labels (*BIO)* introduced as shown in the table below:

Disease Name	*B_Disease*	*I_Disease*
Outside of a Disease Name	*O*	

Bob has finished the data annotation, model training and evaluation, thus reporting 96.6% as the model's accuracy of predicting the three types of labels (*number of correctly predicted labels divided by the total number of labels*). Please make your judgement on the model's performance and Bob's evaluation strategy.

(3 Marks)

c. The following figure (Figure 8) shows an example of medical article describing the disease of COVID-19 from different aspects. To better meet the business requirement of providing detailed information specifically related to a particular aspect of disease, it is critical to break each article into paragraphs and to label each paragraph with a corresponding disease aspect. As an example, the category labels for the four paragraphs in Figure 8 should be *Symptoms, Symptoms, Diagnosis, Treatment.*



The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, ranging from mild symptoms to severe illness. Common symptoms include headache, loss of smell and taste, nasal congestion and runny nose, cough, muscle pain, sore throat, fever, diarrhea, and breathing difficulties. People with the same infection may have different symptoms, and their symptoms may change over time.

Three common clusters of symptoms have been identified: one respiratory symptom cluster with cough, sputum, shortness of breath, and fever; a musculoskeletal symptom cluster with muscle and joint pain, headache, and fatigue; a cluster of digestive symptoms with abdominal pain, vomiting, and diarrhea.

The standard methods of testing for presence of SARS-CoV-2 are nucleic acid tests, which detects the presence of viral RNA fragments. As these tests detect RNA but not infectious virus, its "ability to determine duration of infectivity of patients is limited." The test is typically done on respiratory samples obtained by a nasopharyngeal swab; however, a nasal swab or sputum sample may also be used. Results are generally available within hours.

There is no specific, effective treatment or cure for coronavirus disease 2019 (COVID-19), the disease caused by the SARS-CoV-2 virus. Thus, the cornerstone of management of COVID-19 is supportive care, which includes treatment to relieve symptoms, fluid therapy, oxygen support and prone positioning as needed, and medications or devices to support other affected vital organs.

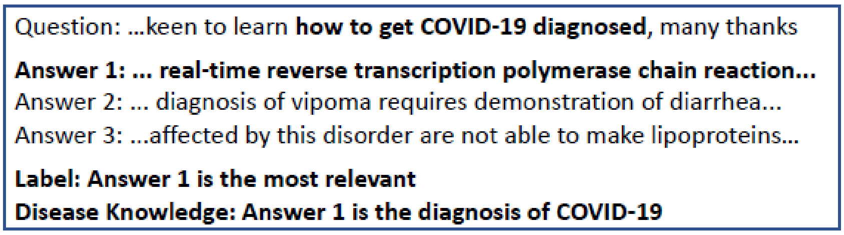
Figure 8 Paragraphs categorized as *Symptoms, Symptoms, Diagnosis, Treatment*

Bob has prepared the dataset and obtained the category labels through crowdsourcing from human annotators. The dataset contains 3500 paragraphs for *Symptoms,* 1500 paragraphs for *Diagnosis,* and 480 paragraphs for *Treatment*. Since the dataset is not balanced, Bob has performed up-sampling strategy through duplicating the training instances from *Diagnosis* and *Treatment,* thus having a balanced dataset with 3500 paragraphs for each category. Then he randomly separated the data into training set and testing set with the ratio of 70% and 30%. The confusion matrix below was reported and the classification model employed is CNN. Please make your judgement on the effectiveness of his experiment and design an enhanced way to improve it.

		Predicted		
		Symptoms	*Diagnosis*	*Treatment*
Actual	*Symptoms*	3410	20	70
	Diagnosis	105	3300	95
	Treatment	30	20	3450

(5 Marks)

d. Consumer health question answering is a critical business goal to achieve. As shown in Figure 9, the ***MediBert*** system is expected to answer the frequently asked questions automatically, by ranking the candidate paragraphs from the huge volume of medical articles and suggesting the most relevant paragraph as answers.



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

Figure 9. Sample question and candidate answers.
Each candidate answer refers to one paragraph of a medical article.

Please design a model to address this business requirement, considering the constraining requirement of providing real-time responses upon a query.

(3 Marks)

e. Instead of returning the whole paragraph as the answer to a user's query (as in d.), in order to further improve the user experience, please design a strategy to provide more precise answers by extracting sentences or phrases from related paragraphs. The constraint on responding time remains as it is.

(3 Marks)

f. As a leading platform in the industry, ***MediBert*** should be distinguished by its AI power of utilising the disease knowledge to provide accurate answers and professional advices to users' enquiry. Unlike the common service provided to the public, an advanced feature of ***MediBert*** is designed for doctors and nurses with specialized requirements. They usually expect that higher quality of information should be returned upon a query, while relatively longer responding time can be tolerated. This requires to re-train the BERT model using the disease knowledge data (illustrated in Table 2), instead of just copying Google's pre-trained models, which will downgrade the performance due to the lack of specialization in medical domain.

Please create your model on top of BERT structure and describe the model pre-training progress. The model should utilize the knowledge of diseases represented by the tuple of *<Disease, Aspect, Information>*.

(4 Marks)