A study on a virtual agent for dengue fever decision support based on conversations and ontology

a Dissertation Submitted to the
GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF THE
SHIBAURA INSTITUTE OF TECHNOLOGY

by

VU DINH MINH
Student ID: nb17502

in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

SEPTEMBER 2020

Acknowledgments

For the appearance of this dissertation, I would like to convey my heartfelt gratitude and sincere appreciation to all great people who have supported me during the doctoral course.

First of all, I would like to express my deepest gratitude to my supervisor, Professor **Masaomi Kimura**, for his guidance, support, and encouragement throughout my research at Data Engineering Laboratory, Shibaura Institute of Technology. His continued support led me to the right way in the path of research science. His advice and passion for science will always be my guide to my life of scientific research.

I would also like to extend my appreciation to the review committee members. They contributed useful recommendations to improve the quality of my dissertation.

I would also like to thank all members of the Data Engineering Laboratory and my all Vietnamese friends in Japan. They are my good friends who shared with me for studying and enjoying student life during my doctoral course. The enjoyable parties and trip camps are a part of my memories in Japan.

I would like to acknowledge the financial support from Japan International Cooperation Agency (JICA) throughout the ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net) for my doctoral course and my daily life in Japan.

Finally, on this dissertation, I would like to extend my deepest gratitude to my family who always beside me during the years I stayed abroad for studying.

Thanks all for making me an unforgettable experience in my life.

Tokyo, September, 2020

Vu Dinh Minh

SHIBAURA INSTITUTE OF TECHNOLOGY

Abstract

Graduate School of Engineering and Science
Division of Functional Control System

Doctor of Philosophy

by Vu Dinh Minh

Virtual agent is the fundamental term in the artificial intelligence topic, which can communicate to human. For supplying the demand of people, virtual agent has been utilized across various aspects, including the economy, traveling, medical and so on. Especially, a virtual agent in the medical topic has attracted impressive attention of many researchers because of the increasing demand for human health care. Instead of meeting a doctor, people can retrieve useful information through communication with a conversational agent. However, making a virtual agent is one of the most difficult challenges for researchers. The reason is that there are different languages between machines and humans. Besides, the huge medical knowledge needs to be equipped to a virtual agent for working in medical topics. To do so, it is necessary to transfer natural language to the computer language and teach the experience of medicine to a machine; thus, a proper solution is to use a deep learning's technique and an ontology to respond to the question. Regarding the necessity of my dissertation, a medical virtual agent can solve the problem of lacking well-trained doctor in the developing countries such as Vietnam and reduce the overload of hospitals. As a result, the quality of health care services can also be improved.

The main objectives of this research were to solve the problem of natural language understanding by using sentence embedding model and an ontology as well as to classify disease based on the clinical symptoms. This is to supports making conversations in the medical topic.

In this dissertation, I proposed a novel deep learning-based health decision support system for making conversation of diseases between patients and doctors. To do so, I also introduced several systems: intent classification, medical symptom suggestion based on ontology, and disease classification. Through conversation, the proposed system can respond to the related information of disease as well as predict the possibility to have dengue disease of users. From that, the proposed system aims at improving humans healthy and solving the problem of lacking doctors in developing countries.

In my future research, I intend to concentrate on improving the system to expand the ability of diseases prediction. Besides, the ability of natural language understanding and the accuracy of disease classification are the most important. This will be great, if the system will be published on the Internet for evaluation. Finally, a generative virtual agent can communicate to human in open domain which is best target.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC	Area under the ROC Curve
CSE	Context based Sentence Embedding
DL	Deep Learning
FP	False Positive
FN	False Negative
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Units
HSC	Hidden Subspace Clustering
LR	Logistic Regression
LSTM	Long-Short Term Memory
ML	Machine Learning
NLU	Natural Language Understanding
NN	Neural Network
NER	Named-Entity Recognition
POS	Part Of Speech tagging
RDBMS	Relational Database Management System
RNN	Recurrent Neural Network
RFs	Random Forest
Seq2Seq	Sequence To Sequence

SSC	Sparse Subspace Clustering
SSEV	Symptoms Suggestion based on Embedding Vector
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
W2V	Word to vector

Chapter 1

Introduction

This chapter aims to introduce background and motivation so that readers could comprehend the importance of this dissertation, which studies an approach for a virtual agent for dengue fever decision support based on conversations and ontology. Firstly, the existing problem will be presented in this chapter. Secondly, beneficial contributions are suggested and some examples are also provided. Finally, a summary of this chapter and the structure of this dissertation are described.

1.1 Motivation

In order to have strong health, it is important for people to have a good understanding of medical knowledge including information about diseases, treatments, and preventive methods to deal with diseases. Nonetheless, ample knowledge of medical topics is a challenge in the human health care process. Generally, there are two primary approaches for people when they face diseases. The first one is the traditional approach, which is meeting the doctors for health checking. With the doctor's attention, patients will receive the best care service as well as helpful advice in order to recover their health. While science proves that earlier detection is better for curing and recovering [4], there are still cases of patients delaying checking health until their health is in danger. They are usually poor patients with a limited financial ability or business persons with tight schedules, who frequently ignore the clinical symptom. Furthermore, the lack of well-trained doctors and the overloading in health services providers are also factors that cause

the decrease quality of human health care. The second approach that many people tend to do is seeking health information based on the Internet. This approach is quite common for the above-mentioned persons. With a smart portable device with the Internet capability, they could make a diagnosis by themselves at any moment because of the unrestricted time and space of the Internet. However, the results from the Internet need to be assessed carefully because there is not any guaranty for the accuracy of the answers. Without great judgment ability, it is difficult to separate diseases with the same symptoms.

Despite the drawbacks of the approach in using the Internet, it is argued that this approach gives people a good way of checking health. Therefore, there has been an increase in people using the Internet as the decision support system for checking their health. As a statistic, there were 58% of Internet users had used the Internet to search for health information [5] in 2011, and this number had risen to 80.2% on the FDA wave of Health Information National Trend Survey (HINTS) in 2017 [1]. The rapid increase in the percentage of seeking health information on humans is due to the emergence of new serious diseases in the century. For instance, a crisis disease named Severe acute respiratory syndrome (SARS) occurred in 2002 caused 8422 cases with 916 deaths over the world, officially notified by Ministries of health [6]. After several years, this disease has been grown to a novel kind named COVID-19, which has caused significant damage to the whole world in 2020. According to the statistic, about 8 million cases with more than 400 thousand deaths by COVID-19 had been recorded after a few months from the date of break [7]. From those results, finding solutions for the provision of medical knowledge to humans is necessary.

To meet the demand for seeking health information from people, a health decision support system could be considered as a solution. Compared to the other Internet-based methods such as using a search engine system or medical forums, the advantage of this method is that the answer could be responded as quickly and accurately as they are meeting doctors. Indeed, there is confusion for people when they use a search engine system because the mechanism in this system is finding the website which includes keywords in the input question. Meanwhile, although the accuracy of answers in medical forums is higher, those replies take a lot of time, sometimes the questions are never responded. Therefore, many researchers attempt to construct a virtual agent for health decision support in order to answer the requests of people. In general, the health decision support

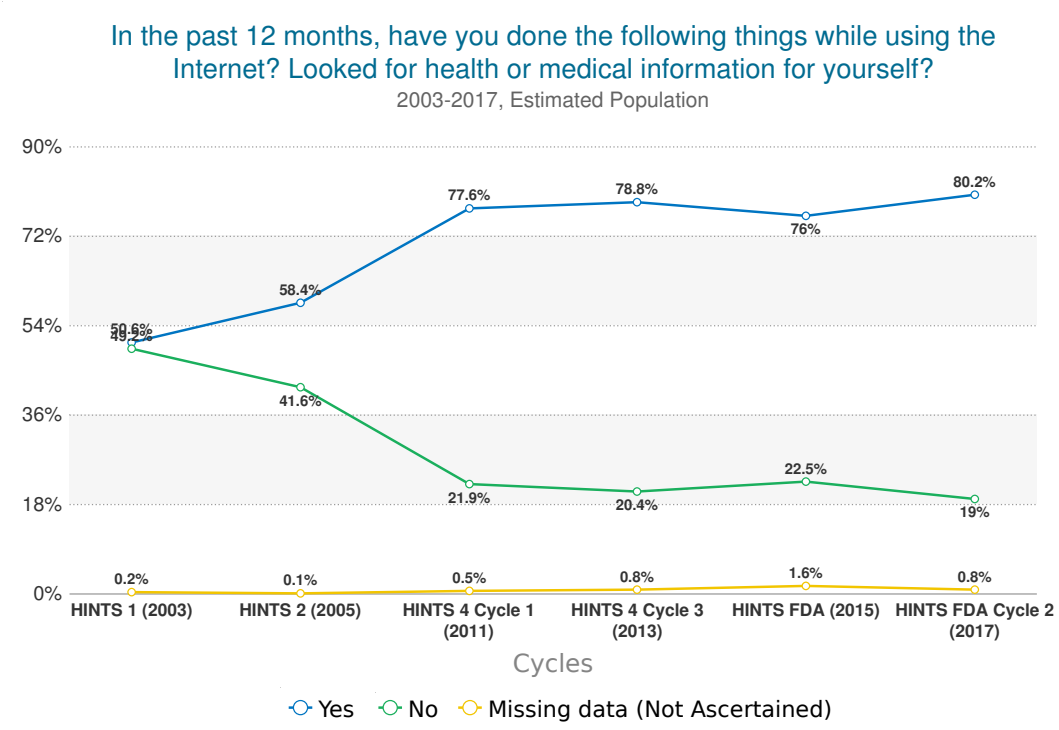

Figure 1.1: Survey on seeking medical information in the Internet [1]

system comprises of two major tasks: information collection and response generation. The first task is responsible for receiving input and extracting meaningful information including intents and symptoms of patients, whereas, the second one is for making a diagnosis based on these symptoms from the first task. For an efficient health decision support system, all useful information related to a patients health needs to be automatically collected for solving the lack of doctors, and a diagnosis should be made with high accuracy. Accordingly, in order to automatically collect all information from a patient, the following requirements should be met: understanding natural language for intent detection, and having knowledge of medicine to ask the equivalent information with diseases from patients. Likewise, the requirement of making health decision ability is needed in order to make a diagnosis. Motivated by these challenges, a new language model based on sentence embedding to represent sentences, and a new method to classify patients with the disease have been proposed in the research presented in this dissertation.

1.2 Challenges in health decision support system

A typical health decision support (HDS) system consists of data collection and disease modeling components as shown in Fig 1.2. The data collection component is responsible for interacting with patients to get the symptoms. The modeling component includes models of diseases that could indicate the state of patients based on the necessary symptoms received from data collection component.

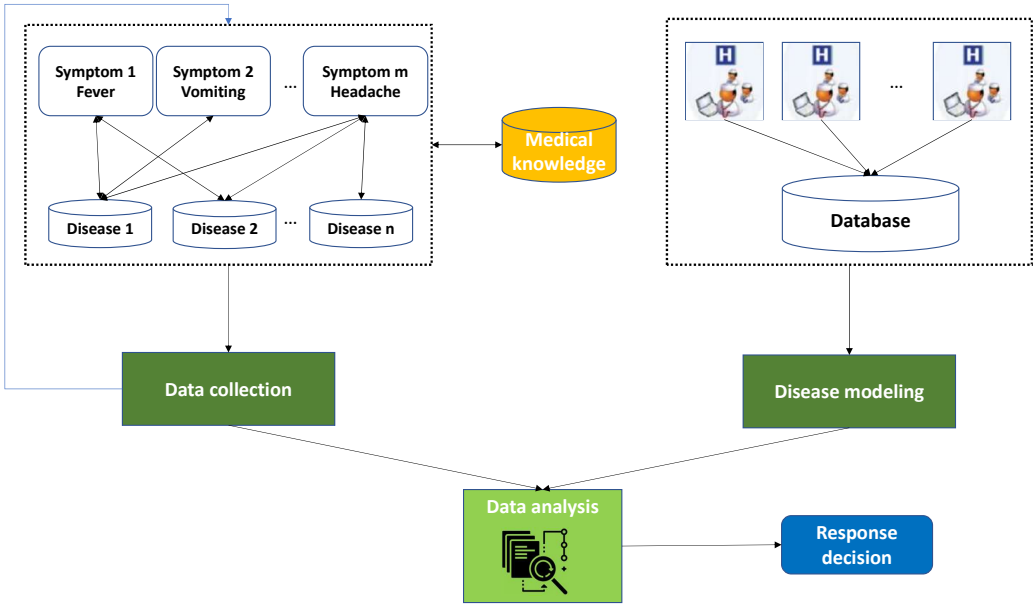

Figure 1.2: A conceptual model of a typical Health Decision Support System

In the data collection component, there are two methods that could be used to collect symptoms of patients (e.g. fever, vomiting, and so on). Firstly, the manual method is that the data collection is done manually by the regular checkup with doctors. Secondly, the automatic one is that patients could interact with the HDS system and submit their information by themselves. In the first method, there is a problem of lacking health care workers which causes overload in hospitals, thereby patients often have to wait in the long queues for their examination. With the second method, the problem of overload is solved but there is difficulty

in designing a good enough interface in order to interact with humans for data collection. Theoretically, there are two approaches for designing interface: rule-based and Artificial intelligent (AI)-based[8].

- The rule-based approach consists of prepared scenarios, and patients submit their necessary symptoms to the system through filling to the fixed application forms. This approach ensures that the information of patients is correctly gathered; however, it has a drawback of scalability when collecting information for many diseases. Specifically, different forms have to be designed for each disease respectively which causes a problem to system development. Furthermore, selecting the appropriate forms is an impossible task for patients with lacking medical knowledge.

- The AI-based approach attempt to make a virtual agent which can understand the intent of human through the conversations. Without depending on the prepared scenarios from imitating human behaviors in a medical checkup, a virtual agent is unrestricted in symptoms collection of various diseases. However, building a virtual agent in medical topics is not easy because of not only the difference of language between humans and machines but also the ample knowledge of medicine. To be more specific, a virtual agent should extract the semantics including the semantic of every vocabulary term within the context of document. Moreover, to get the all necessary symptoms for making a diagnosis, a virtual agent needs to make a list of candidates of diseases based on the provided information. During the conversation, this list will subside until the symptoms are sufficient for diagnosis. To do so, a virtual agent should be equipped with the knowledge of disease and its symptoms. Those above problems lead to the need for a solution of combining natural language understanding and knowledge of medicine.

In the disease modeling component, the diagnosis results of a patient are achieved through a function that maps the input to the output. Regarding the construction of function, disease modeling should be separated into two approaches: knowledge-based and non-knowledge-based. In the first approach, a decision tree could be built based on the rule for disease detection. For instance, patients with symptoms such as fever, vomiting, and skin bleeding could have dengue disease, but without fever symptoms, it is guaranteed that they do not have

dengue fever. This approach is understandable and proven to be extremely effective in several specific cases whenever the difference between disease is obvious. Unfortunately, most diseases are difficult to be identified due to the overlap of symptoms between them. In fact, doctors must also refer to many symptoms and their experience in examination and treatment for making a diagnose. Secondly, the non-knowledge-based approach issue represents the situation in which the probability of one class (e.g., healthy patients) is significantly higher than the other class (e.g., ill patients). In this approach, the dependent on the training data of the classification methods is large because of the mechanism of training. Real-life experiments show that the accuracy of the non-knowledge-based methods is great if one possesses good training data. However, real-life data often includes noise and redundant attributes which cause a decrease in the accuracy of classification. Considering the importance of diagnosis to humans, improving the accuracy of the classification disease is necessary because the wrong results will have unpredictable consequences.

In order to construct a virtual agent in the medical topic for health decision support, I introduced a design of the virtual agent following the AI-based approach as shown in Figure 1.3. Specifically, in natural language understanding(NLU) component, the input sentence of patients will be transformed to the intent by a comparison between input and the predefined sentences, and the entity of symptoms is simply extracted by a dictionary in medical. After that, in order to require more information of other symptoms from patients, the suggestion symptoms component generates the necessary symptoms based on the output of NLU component. This processes are looped until all essential symptoms is collected enough and moving to the making a diagnosis by the last component. For more detail, the components of the virtual agent, which are NLU, symptoms suggestion, and disease classification are in turn to expressed in Section 1.3, Section 1.4 and Section 1.5, respectively.

1.3 Natural language understanding for intent detection and entity extraction

As stated above, intent detection based on natural language understanding can improve the ability of machine in data collection. Specifically, the semantic mean-

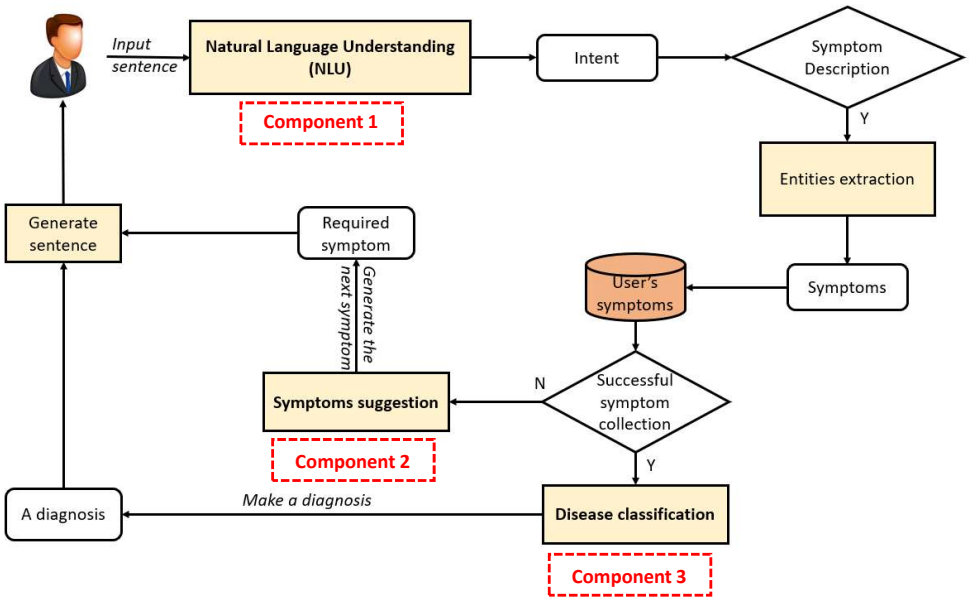

Figure 1.3: The architecture of a conversation-based virtual agent for diseases decision support

ing of sentences is considered as a factor to support the machine for intent detection. Therefore, the new model to extract the semantics of sentences was constructed in this research.

Natural language processing is a challenge in the field of artificial intelligence. Various topics such as part of speech (POS) tagging [9], named-entity recognition (NER) [10], and semantic similarity are being researched to find ways for communication between humans and machines. In order to detect the intents of humans, a comparison of semantics between input and prepared sentences is considered as a good solution because of the scalability of this method. Through the distribution of representation of sentences, a value of semantic similarity can be calculated by cosine distance between them. A famous idea in natural language understanding is a probabilistic-based neural-net language model to help machines learn a distributed representation for words. However, linguistic analysis of vocabulary terms might not be enough for a machine to correctly apply learned knowledge. To succeed in learning human language, a machine must understand further, the semantics of every vocabulary term within the context of the documents. By comparing the semantic of sentences, a machine could satisfy the expectation of humans for the task of intent detection. The accuracy of finding similar semantic sentences is fundamental for evaluating the machine's ability to detect intent. A series of experiments were conducted to confirm the ability of intent detection based on the calculated semantic similarity sentences. The experimental results demonstrated that using the context embedding-based method could achieve favorable results.

1.4 Symptom suggestion based on conversation and ontology

During the medical checkup, doctors have to request more symptoms from patients in order to support the process of making a diagnosis. This process requires doctors to have a medical knowledge for understanding what symptoms need to be detected. For software programs, subsequent symptoms could be detected based on the rules; however, this still has disadvantages because the diseases have too many symptoms and some of them may overlap, which makes an obstacle in system development. Ontology, which is a semantic database which could describe

the relationship between symptoms and diseases could overcome this obstacle. By simulating the process of medical checkup, an ontology-based method automatically detects the necessary symptoms in the conversation, which is proposed in the dissertation. By applying a neural network into the ontology database, the system can detect symptoms in conversations and predict the required subsequent symptoms.

1.5 Subspaces-based hidden features for Dengue classification

In terms of the medical aspect, the accuracy in making diagnoses of the decision support system is one of the important factors. All inaccurate results can have unintended consequences. With the goal of automatically making diagnoses for patients, the decision support system needs to achieve the diagnostic accuracy thresholds. This depends not only on the classification algorithms but also on the training data because of the existence of noise and redundant attributes in the data.

In order to improve the efficiency of disease classification, a model to eliminate redundant attributes as well as to find the relationship between attributes in the data is proposed. From experimental results, the model proved that hidden features corresponding to the correlation of attributes could improve the accuracy of disease classification.

1.6 Limitation

The subject of study for this dissertation is a virtual agent in the medical topic. The focus of this study is to improve the interface between humans and machines. However, the proposed methods were validated in the scenario of an experimental setup with a small dataset only. The scenario of a huge data set can improve the quality of the sentence embedded, which has not been investigated.

In addition, due to the fact that there are numerous diseases in the medical topic(e.g. dengue, cancer, diabetes), our proposals were initially investigated in dengue fever only.

As mentioned in the previous section, our research also concentrated on automatic suggestion of the necessary symptoms. In the research, the collaborative approach using deep learning techniques and ontology was performed. However, the lack of information in associated with a combination of diseases ontology and symptoms ontology is another limitation in the scope of our work.

1.7 Contributions and Thesis Organization

In order to resolve the aforementioned difficulties in building a virtual agent for health decision support, a series of proposals related to understanding language processing and disease classification should be employed. The main contributions of this dissertation can be summarized as follows:

1. The novel model to transform natural language into computer understandable language was proposed for making a conversation between humans and machines. By converting sentences to embedded vectors, the cosine similarity between sentences was determined for their semantic similarity calculation. By applying the pre-trained model to compare input sentences with fixed sentences in the database, thus, intents of the input sentence were identified in order to support the machine to understand the requirements of humans.

2. An approach for symptom suggestion based on conversations was proposed by taking into account the combination of deep learning techniques and ontology. More concretely, a hypothesis of applying neural network to ontology was introduced, which suggests the next necessary symptom be submitted by humans. Thereby, the symptoms of humans are collected and then applied in the data analyzing phase. As a result, by combining with the model mentioned above, the process of data collection has human-likes behaviors.

3. A model for classification diseases was proposed in modeling components for making a diagnosis. In this model, the hidden features are found based on clustering which improved the accuracy of disease classification. As a result, compared to the approach which normally used classification techniques, this proposed approach not only improved the accuracy of classification but also saved the computation time.

The organization of this thesis is illustrated in Figure 1.4 and described as follows:

Chapter 1: Introduction. The motivation and background of this research were described in this chapter. In addition, the research overview and research limitations were also presented. The primary contributions of this research were also concretely summarized in this chapter.

A literature review and background knowledge are provided in Chapter 2 and Chapter 3, respectively. The issues related to two data collection and modeling are thoroughly resolved in detail from Chapter 4 to Chapter 6. Chapter 7 discusses, Chapter 8 concretely concludes the work and figures out future work direction.

Chapter 2: Literature review. In this chapter, I discuss previous researches related to the topics of this dissertation.

Chapter 3: Background. This chapter provides a wide range of background knowledge related to natural language understanding, a virtual agent and disease classification. In this chapter, the mechanism of a virtual agent will be initially expressed, followed by a definition of a virtual agent. Furthermore, the common disease classification model based on machine learning approach is also discussed.

Chapter 4: Calculation of semantic similarity sentences for intent detection. This chapter proposes a method of intent detection based on conversations. The intent of sentences was identified by the comparison of semantic similarity with prepared sentences. Before that, the context-based sentence embedding (CSE) model which is a novel model of representation of sentences, is presented.

Chapter 5: Symptom suggestion based on conversation and ontology for medical examination. This chapter proposes a new method to suggest the necessary symptoms to make a diagnosis. The proposed approach is to combine deep learning technique and ontology in order to establish a topic map function expressing the relation between input symptoms and subsequence appropriate symptoms.

Chapter 6: Dengue classification based on the clinical symptom. This chapter proposes a new method to separate the patients in dengue classification. In this chapter, a review on the methods in dengue classification and the proposed method were clearly presented. The issues of accuracy and computation cost demand a dimension reduction and finding meaningful hidden features from data.

Chapter 7: Discussion. This chapter discusses the work investigated and solutions proposed in this dissertation. Advantages as well as the remaining

issues will be summarized.

Chapter 8: Conclusion and Future Work. This chapter concludes the dissertation by which advantages as well as remained difficulties were discussed. Finally, research directions of great interest for the future work were figured out.

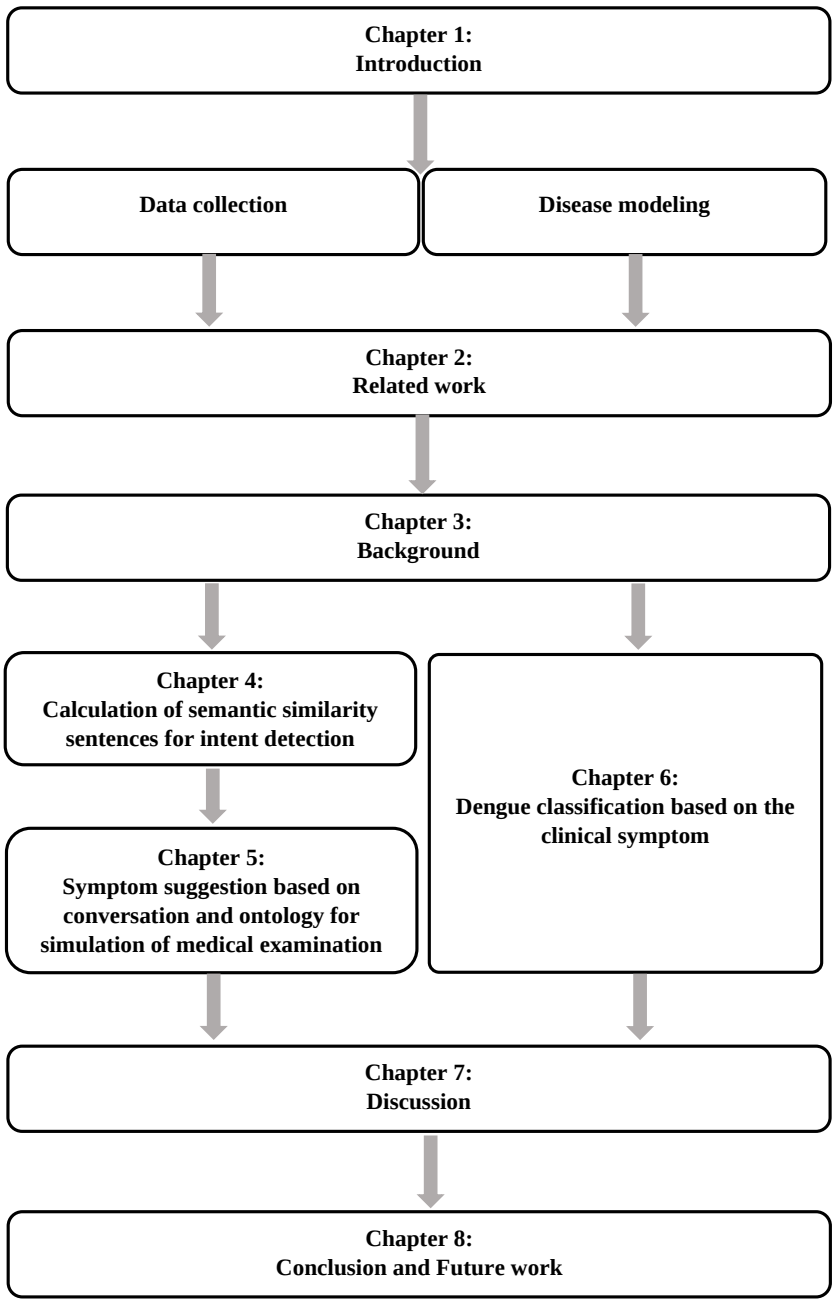

Figure 1.4: Organization of this thesis

Chapter 2

Related works

In the previous chapter, I described the introduction of this dissertation to describe what motivation of this research was and explain what problems were herein addressed, including this researchs objectives and contributions. This chapter aims to provide an idea of the state of art of the corresponding areas. I will review the existing works and thoroughly discuss their solution and problems. Corresponding to the stated challenges, this chapter is organized into four parts: medical agents, intent detection, symptoms suggestion, and disease classification for making a virtual agent in the medical topics.

2.1 Medical agents

A medical agent is a software program, which can perform a diagnosis and provide the treatment based on the provided symptom. The outstanding advantage of medical agent is the ability to immediately connect patients and doctors through the Internet connection. Hence, people can restrict access to hospitals, while early identification of major diseases can be still performed. Due to these benefits, medical agents have been implemented not only in the laboratories but also in the medical industry to provide higher-quality health services. In this section, I review some medical agents from scientific papers found by looking for "medical agents", "health care", "clinical", "chatbot", "survey", and a combination of these terms.

Basically, a medical agent comprised of two primary components: a conversation agent and a diagnostic agent corresponding to the tasks of information

exchange and diagnosis, respectively.

- To perform the diagnostic agent, there are many machine learning algorithms such as Support Vector Machine (SVM) [11], Artificial Neural Network (ANN), Random forest [12], [13] were already applied to classify disease, for example, heart detection, cancer detection, dengue classification, and so on. Depending on the specific dataset of disease, classification algorithms will be selected to achieve the highest accuracy.
- To perform the conversation agent, there are two approaches: rule-based and Artificial Intelligence(AI)-based as I mentioned in Chapter 1. In the first approach, researchers [14], [15], [16], [17], [18] usually utilized the Artificial intelligence markup language (AIML) [19], [20] to create conversational flows for the bot. In the second one, the AI-based agent [21] is designed by many researchers [22], [23], [24], [25], [26], [27] to detect the intent based on the techniques of natural language understanding (NLU), then respond based on the knowledge database.

In general, medical agents in the above papers share the same workflow where the intent firstly needs to be detected, then activities such as requesting additional information or making a diagnosis are controlled by the code component as shown in Figure 2.1. Here, the code component is works as the navigator to call the extender APIs of medical agents based on the intents. From that, the output is generated and respond to humans. However, the template in Figure 2.1 is too simple for comparison between medical agents because the functions are only abstract.

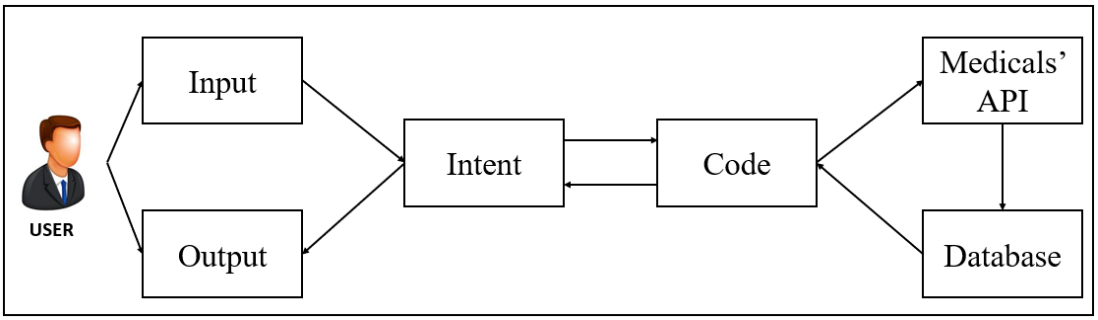

Figure 2.1: The general workflow of medical agents

Therefore, in this dissertation, I proposed a new template for medical agents as shown in Figure 1.3. Especially, different from the previous systems, the proposed

medical agent aims to work with not only a specific disease but also with all diseases. To do that, the proposed system has an additional component to control the input symptoms and suggest the necessary subsequent symptoms as shown in Chapter 1. With the specific criteria corresponding to the important tasks of medical examination, the judgment of medical agents is easier by comparing the efficiency of tasks between agents. Moreover, the proposed medical agent also introduces a framework that includes new methods for increasing efficiency of intent detection, symptom suggestion, and disease classification in order to improve the performance of the whole system. Concretely, the mechanisms of components in the proposed system are as follows.

- Intent detection based on comparison of semantic similarity between sentences instead of pattern matching.
- Symptom suggestion component supports to work with all disease instead of only a specific disease.
- Disease classification based on machine learning algorithms as the previous systems. However, a new model is introduced to improve the accuracy of diagnosis.

2.2 Intent detection

Intent detection is a principle method to support humans in understanding and exchange information with each other. Typically, there are many ways to describe an intent, which depends on a model language. Hence, a model language is essential and play an integral role in numerous applications in natural language processing. As the heart of the application, a model language usually helps the machine understand and carry out the requirement of human. Specifically, for solving the different languages between human and machine, this model is an intermediary to transforms human language into the information understandable for machine and vice versa. In the last decade, with the advent of strong calculating devices, there are many existing methods [3], [28], [29], [30], [31], [32], [33], [34], [35],[36], that acquired impressive achievements. These methods often represent words, sentences based on the embedding vector to capture their semantic

and syntax. Therefore, by computing of the combination of vector, machines can find out a representation of intent of humans.

- One of the first popular approaches in word embedding is Word2vec [3] which aims to compute the semantic relationship of adjacent words within a sentence. This approach relied on either Continuous Bag of Words (CBOW) architecture which uses the bag-of-words context to predict a target word or Skip-gram architecture which predicts the context word based on a given target word.

- Another approach similar to Word2vec is Glove [28], which also uses a context window to simulate word representation. The difference between them is the way of using the context of words. While Word2vec uses context window to make a training set for neural network, Glove utilizes it to create a co-occurrence matrix. In terms of this aspect, Glove might be more comprehensive than Word2vec because it computes based on all dataset, instead of only the context word in Word2vec. However, both models have the drawback in out of vocabulary problem.

- This is one reason why FastText [29] is introduced by Facebook. They suggested a method which divided words into n-gram and training based on these tokens. For instance, the word apple will transform to app, ppl and ple. Thus, this model works well for the rare words and solves the about out of vocabulary problem.

Although word embedding operates well in some tasks, it still has a few issues when performing with sentences. Hence, another research direction was proposed using sentence embedding to learn language representation. Based on auto-encoder architecture, some models attempt to represent sentence by supervised or unsupervised learning. A major advantage of this research direction is keeping the order of word in the sentence.

- First of all, Skip-Thought Vector [34][35] model includes a Recurrent Neural Network (RNN) to map words to sentence vector in the encoder and two RNNs generate the surrounding sentences (previous and next sentence) in the decoder. A major advantage of Skip-Thought and other models in sentence embedding approach compared to word embedding is the order of word in the sentence.

- After Skip-Thought, InferSent a recent approach in sentence embedding has been introduced by Facebook [32] with the same ideology about the representation of sentences. The large difference between the two approaches is that InferSent was a supervised learning approach, instead of unsupervised learning as Skip-Thought. They applied a BiLSTM model [37] to a labeled dataset to predict entailment/contradiction. This approach proves its efficiency with much better results on various tasks than other methods. However, a disadvantage of this approach is the requirement of the high-quality labeled dataset for training. The author used Stanford Natural Language Inference(SNLI) dataset [38] which includes 570k pair of sentences in English to build the semantic sentence.

- Finally, the latest proposed approach recently is Universal Sentence Encoder (USE) [33] which was provided by Google. Like InferSent, this approach also trains on the SNLI dataset but integrated with the unsupervised learning tasks in two different encoders for making sentence representation. The first one is the Transformer-based encoder model which required a high computational resource and got a better accuracy. In which, Transformer [39] is a novel architecture which uses only attention mechanism, instead of a Recurrent neural network. The second one is Deep averaging network (DAN) which utilizes a mechanism where words and n-gram are averaged as the input in a deep neural network.

Although the above approaches highly successfully performed the representation of sentences, there is a major issue in these models: the missing of the influence of context of the sentence. For example, machines cannot select the meaning between two sentences in Table 2.1 if missing the helpful information of the previous sentence. Therefore, we propose a novel approach which covers the sentence and its context for the best performing sentence.

2.3 Symptom suggestion

My focus is to suggest the essential symptoms based on the input symptom. This is an important task in medical examinations between doctors and patients. In order to construct a symptoms suggestion system, there are two primary issues need to be taken into account.

Table 2.1: The influence of the previous sentence to the current sentence.

	Previous sentence	Current sentence	Meaning
Example	I have just bought a new telescope.	I saw a man in the building with a telescope.	There was a man in the building, and I saw him with my telescope.
	I came to my office.	I saw a man in the building with a telescope.	There was a man in the building, who I saw and he had a telescope.

- A reliable source of medical knowledge structure for healthcare, including medical standardized vocabularies in an appropriate design structure.
- A method to effectively exploit the relationship of entities in the database, which supports the machine to determine the situation of a patient with symptoms, thereby, make a suggestion of the required symptoms.

In the first issue, databases are always an integral part of all applications. Especially, in medical applications, knowledge structures for healthcare are studied which provides a solution to many health care challenge of the 21st century. Typically, a Relational Database Management System (RDBMS) [40] such as Server SQL Management Studio (SSMS) [41] or MySQL [42] is a primary method to control information for many areas. However, medical knowledge representation requires the use of standardized vocabularies to ensure both shared understanding between people and interoperability between information systems [43]. Therefore, many researchers utilized an ontology for the representation of medical knowledge [44]. Specifically, an ontology of human disease (DOID) is introduced by Schriml el at. [45] to perform the knowledge about human disease. This ontology includes specific formal semantic rules to express meaningful disease models, thereby additionally providing the related information of the disease. Similarly, another ontology of symptoms (SYMP) is also introduced to describe symptoms and their relations. In the health decision support system, the relationship between disease and symptoms play an important role. To provide this knowledge, a combination of DOID and SYMP, which is an ontology has been developed by Mohammed [46], Mhadhbi [47] is to establish the relationships between diseases and symptoms.

In the second issue, the problem is supporting machine to understand these knowledge. In order to suggest the essential symptoms based on the input symptoms, a method following rule-based approach is frequently used with the role definition is manually defined by humans. In a small database, this approach is extremely efficient under the control of experts in designing the policies of relationships among entities. However, the scalability of the database which can be an obstacle for methods in this approach. Due to the manual designing policies in a database, there is an increase in complexity when we want to apply in a large system. Therefore, another method based on queries to a database which is frequently used. This method is to exploit the relations between entities of semantic databases such as an ontology for finding the highest probability disease, then suggest a symptom in this disease. However, this method has a disadvantage about weights between diseases and symptoms which can be determined as the probability of having diseases if they have symptoms. Specifically, assume that the given input symptoms include S_1 and S_2 where S_1 is a symptom of disease D_1 and S_2 belongs to D_2, the outcome is undefined because of the equivalent probability between these diseases.

Therefore, I proposed a new method which covers the weight of relationship between diseases and symptoms to suggest the necessary subsequent symptom based on the input symptom. For definition, weight can be considered as a probability of having disease with the input symptoms. Based on a comparison of weights between candidate diseases, the efficient of the proposed method is better than previous methods.

2.4 Disease classification

In recent years, machine learning played an important role in medical data analysis for improving the quality of treatment to patients. There are many serious infectious diseases which were detected by the algorithms in Machine learning. In the last two decades, cancer, diabetes, liver, hepatitis and dengue are five popular diseases which have attracted a lot of attention from researchers. Based on the clinical symptoms and the information of data, the researchers have utilized the classification techniques to separate the patients and healthy people in place of doctors. In terms of the result, the output of the classification algorithms is the binary value which represents the state of patients with diseases (0 = False, 1

= True). The common algorithms are Artificial Neural Network (ANN), Nave Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, and Random Forest. However, ANN is used less than other traditional algorithms because this algorithm frequently requires huge data in order to get high accuracy. Meanwhile, the data of patients are usually manually stored as hard copies and the cost of digital converting is also expensive. Therefore, the model with other traditional algorithms are often used to predict the problem of disease classification with clinical symptoms.

For example, in order to detect heart disease [48],[49],[50], [51], Otoom et al. [49] introduced a system for Coronary artery disease detection and monitor. With the Cleveland heart dataset [52] from UCI Machine Learning Repository, they made an experiment with three algorithms Bayes Net, Support Vector Machine and Functional Tree to optimize the classification of disease. The best result in their experiment is the accuracy of 88.3% by SVM method.

In diabetes classification [53], [54], [55], [56], Iyer et al. [54] made an experiment with Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Disease. In their experiment, Decision Tree and Nave Bayes algorithms were applied to predict the state of patients with diabetes. They achieved 79.5652% correctness by Nave Bayes algorithm.

For liver disease, Vijayarani and Dhayanand [57] utilized Support Vector Machine and Nave Bayes for classification. Indian Liver Patient Dataset(ILPD) data set is obtained from UCI. Dataset comprises of 560 instances and 10 attributes. They utilized Matlab to implement Nave Bayes and SVM for liver disease prediction.

Ba-Alwi and Hintaya [58] utilized many data mining algorithms to predict hepatitis disease. The comparison among Naive Bayes, Naive Bayes updatable, FT Tree, K Star, J48, LMT, and Neural Network in hepatitis classification is performed by the Weka tool [59]. In the experiment, Support Vector Machine algorithm achieved the best result.

In dengue classification [60], [61], [62], [63], [64], [65], [66], there are many researchers attempt to apply the traditional algorithms for detecting the patients with dengue. For example, a logistic regression algorithm is utilized by Tuan et al. [61] on the data of 5726 children within 72 hours of fever onset to distinguish patients with dengue illness. This dataset includes 35 attributes/features. 19

features are used out of 35 features. Weka tool was also used for detection. The result of this paper demonstrates the efficiency of statistical methods as a tool for doctors in dengue diagnosis. Thus, the quality of treatment for dengue patients was improved.

Besides, another study by Tanner et al. [62] conducted an experiment using a decision tree algorithm on the data of 1200 patients with fever for the first three days to distinguish patients with dengue illness. In their experiment, compared to the logistic regression algorithm, the decision tree algorithm had the advantage in handling missing values which are commonly encountered in clinical studies.

However, the previous studies often disregard the influence of processing of feature extraction which is one of the important factors to improve the accuracy of classification algorithms. Therefore, we propose a novel approach which aims to find the hidden subspaces for increasing the accuracy of dengue classification. This approach based on the reality of making a diagnosis from doctors. They are not required to investigate all symptoms to make a diagnosis. Instead, they can make a precise diagnosis based on only several crucial symptoms. By successful detecting a small group of dengue patients, the accuracy of dengue classification is also optimized.

Chapter 3

Background

Chapter 2 mentioned the related works and their problems for the corresponding areas of making a virtual agent. In this chapter, I provide a background of a virtual agent and how to build a virtual agent for health decision support, which consolidates knowledge of the subsequent chapters.

3.1 Overview of a virtual agent

3.1.1 Definition of virtual agent

In recent years, virtual agent development has attracted a lot of attention from service providers. Many virtual agents such as ELIZA [67], ALICE [68], IBM WATSON [69], SIRI, GOOGLE assistant, etc.. have been introduced to automatically respond to people like a normal human being. Therefore, the term "virtual agent" has become popular in human-life. In my point of view, the concept of a virtual agent or chatbot in computer science is the fundamental term to describe a software program that is able to communicate with people via textual content or other methods. These programs can automatically respond based on the specific situation without the human factor, thereby improving the quality of services. From those operations, a virtual agent can deliver a number of benefits including the ability to:

- Expand to round-the-cloud automated service for customers, as well as employees, without having to hire full-time workers.

- Respond more quickly and more consistently to customer inquiries or employee requests by scaling the virtual agents to meet anticipated needs.
- Better compliance with standards as virtual agents follow only programmed set of actions.

3.1.2 Types of virtual agents

There are several criteria of virtual agents categories. In this subsection, two primary methods to categorize virtual agents are expressed in order to support the reader to comprehend my virtual agent in details.

- Firstly, I will clarify the classification of virtual agents based on the domain. Virtual agents are implemented in a specific domain such as commercial, customer service, or health care decisions. With these systems, the allowable requirements are related to a certain topic, and the outside requests will be ignored. The advantage of virtual agents in this category is that the quality of responses is frequently favorable. In contrast, the second category is the open domain which includes the virtual agents with the ability to respond to unrestricted requirements. This means that a machine can make a conversation with humans on many topics.
- Secondly, the second category method separates the virtual agents based on the context, which divides virtual agents into single-turn and multi-turn. The single-turn virtual agent is the question and answer system, which is able to reply only to a request and is not capable of remembering the information of conversation. Meanwhile, the second category consists of virtual agents allowed to make a multi-turn conversation as human-like.

3.1.3 Architecture of virtual agents

Generally, the architecture of a virtual agent is usually comprised of three primary components: Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation [2] as shown in Figure 3.1. Besides, there are several additional components such as Speech Recognition (SR) in order to provide the methods to communicate with humans.

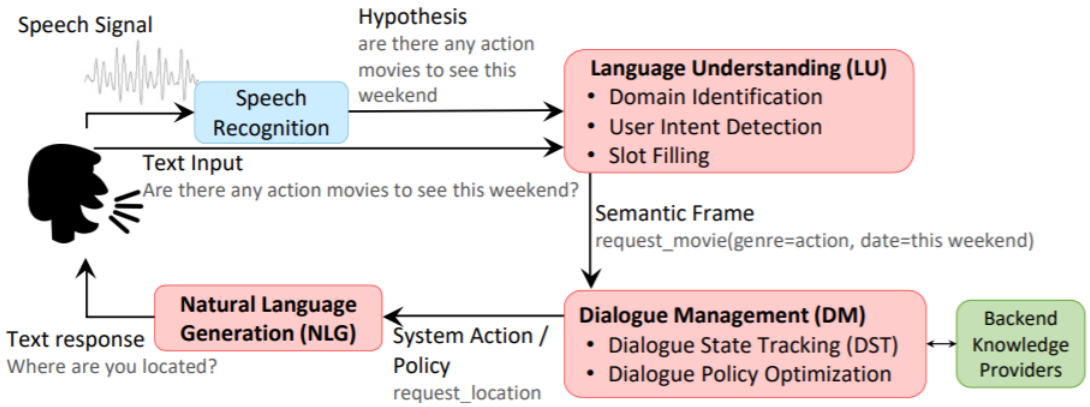

Figure 3.1: The architecture of a virtual agent [2]

3.1.3.1 Natural language understanding

Natural Language Understanding is one of the most important components of a virtual agent, which is responsible for detecting the intent of humans and information extraction. Figure 3.2 and 3.3 show a process to extract the meaningful information from the input of humans.

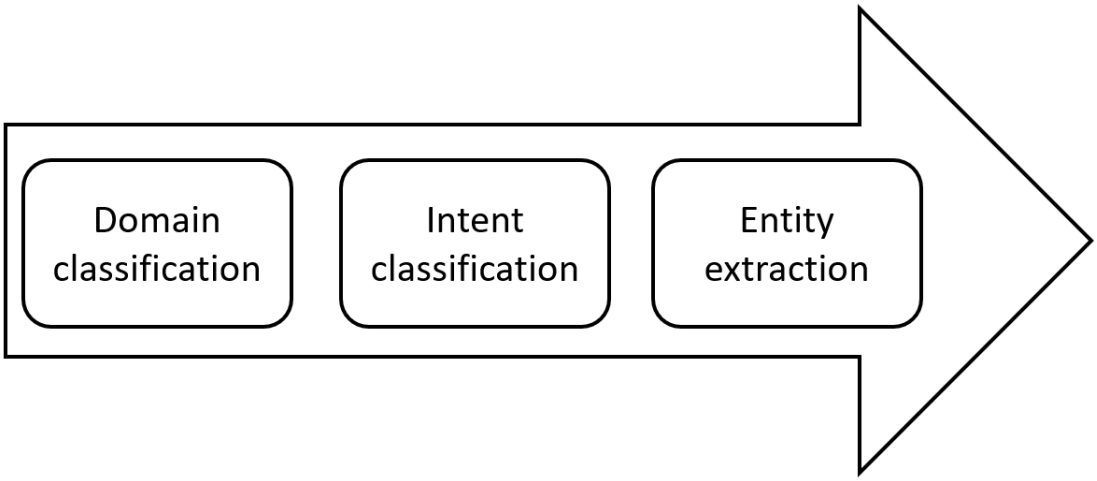

Figure 3.2: The processing of information extraction in Natural Language Understanding(NLU)

- At the first step, a process of detecting domain will assist the machine with shrinking the scope of sentences, thereby, supporting subsequent processes.
- The intent is a requirement of users, which is sent to a virtual agent by a sentence or speech. From that, intent detection is a process of classification

among an input sentence with several predefined sentences. For example, the input of the user is "I have a fever", which indicates that they have a symptom of the disease.

- The second meaningful information, which needs to be extracted from the input of humans is the entities. To collect the necessary symptoms for making a diagnosis, the fever symptom in a sentence "I have a fever" needs to be extracted for supporting a response.

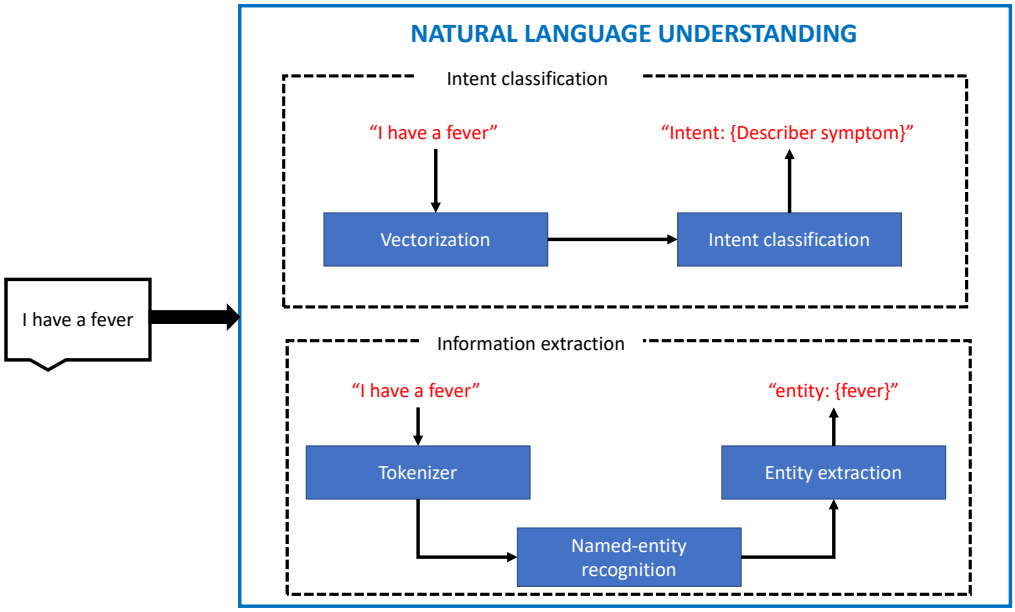

Figure 3.3: The processing of information extraction in NLU

3.1.3.2 Dialogue management

In the multi-turn conversation, a virtual agent is required to remember the context of a conversation, and manage the dialogue states. Therefore, a dialogue management component is designed that is responsible for the smoothness of information exchange between humans and machines. This is because the subsequent dialogue state depends on the previous dialogue states. For example, a virtual agent in medical topic will make a requirement for a necessary symptoms,

which have not been provided by patients. Concretely, a dialogue management component has one of the following rules:

- Receiving the input from natural language understanding component.
- Managing the dialogue states and a dialogue context.
- Transforming the request to the natural language generation component.

3.1.3.3 Natural language generation

Natural language generation is a component which makes a response to humans. The mechanism in this component is a mapping function between the action of dialogue management component and natural sentence. An action of symptoms description and other actions will be replied by the different sentences.

3.1.4 Approaches of construction of virtual agents

Generally, to build the virtual agents, there are two primary approaches: rule-based and neural network-based. The responses in the first approach are usually better because they are manually pre-defined by human. However, the cost of building rule and the poor scalability of rule-based approach are the regrettable disadvantage of itself. In contrast to the first approach, with the advancement of artificial intelligent, machine learning and deep learning, the neural network-based approach has better scalability because a virtual agents can learn by itself. Based on the unstructured conversation dataset, the neural network-based virtual agents system can learn to detect intents, entity of the utterance and make the response.

3.1.5 The effect of a virtual agent in the medical topics

Nowadays, diseases have a major impact on human life because of their widespreadness. With the simple transformation of viruses, the number of patients with diseases has increased rapidly, which frequently causes overload to medical service providers. This causes the problem that doctors often have to treat many patients at the same time within a high-pressure environment, which is one of the reasons

for the mistake in making a diagnosis. Meanwhile, software systems are capable of automatically repeating operations under human control. To share work with doctors, many software systems were built as the assistance tool for increasing the quality of health care services. Specifically, the application in medical topics such as health decisions, or a virtual agent in health decision support has been introduced in recent years, which is a major contribution to reducing the work of doctors.

3.2 A model language based on neural network

The important of a model language were mentioned in Chapter 2, which supports machine to solve the problems in natural language processing. For more clearly, this section investigates the background knowledge related to the classic model based on Encoder-Decoder architecture namely Sequence to Sequence (seq2seq) [70], [71] and the operation principle of Recurrent Neural Network(RNN).

3.2.1 Recurrent Neural Network

To solve the limitation of the traditional neural network, Recurrent neural network(RNN) is introduced with the core idea similar to human thinking [72]. Basically, human thinking always starts with a memory point. For example, in order to comprehend a paragraph, people must understand the words of the paragraph, which means that there is a memory of words to support the task of the paragraph understanding. Similarly, people have misunderstand the specific scenarios in the movies until they get the previous scenarios.

Figure 3.4 depicts the operation principle of RNN. While processing, the previous hidden state is moving to the next step of sequence with a role as a neural network memory, which supports the previous data to be retained. To generate the new hidden state h_1, a combination of previous hidden state h_0 and current input x_1 has to go through a tanh function to ensure the value in the range [-1, 1]. Without this process, some value can be exploded after several transformations, which causes other values to be insignificant.

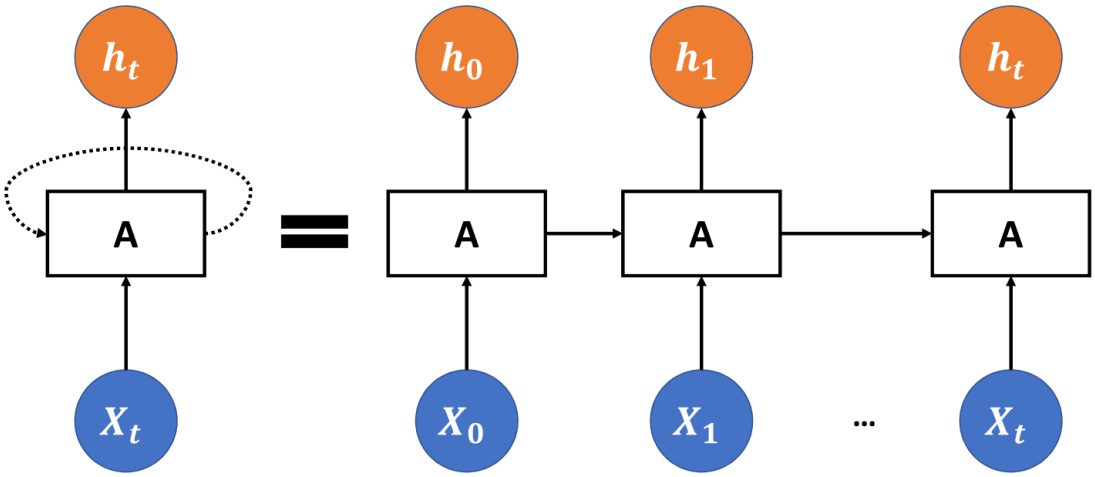

Figure 3.4: The structure of Recurrent Neural Network(RNN).

3.2.2 Sequence to Sequence model

Sequence to sequence (seq2seq) [70], [71] model is a classic model of deep learning and achieved many impressive achievements such as: machine translation, text summarization, image caption, and so on. Typically, the input of seq2seq model is a sequence of words, characters and output is another sequence. Generally, seq2seq model consists of two main components: encoder and decoder as shown in Figure 3.2.

- Encoder: Processing each item in the input sequence, then compiling the obtained information to form a vector context.
- Decoder: After processing the entire input sequence, the encoder will send the context vector to the decoder for processing, the decoder will generate an output.

To transform input to the context vector, there are two steps in encoder.

- Firstly, words have to be converted to vectors by techniques of word embedding.
- Second, these embedding vectors are processed by RNN to generate the context vector.

In decoder, the architecture is similar to encoder, except for only the last layer, which is a softmax function [73], [74] to map the appreciate words.

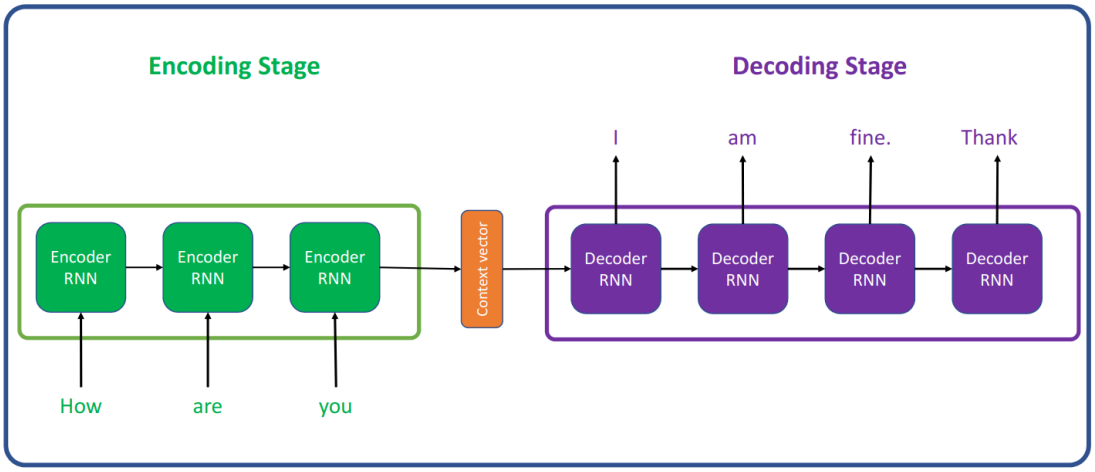

Figure 3.5: Encoder-Decoder architecture (Seq2seq model).

3.3 Ontology

3.3.1 Definition of ontology

In recent years, the term "ontology" has not only been used in laboratories in the field of artificial intelligence but also become popular in many areas of life. From the standpoint of the artificial intelligence industry, an ontology is a description of the concepts and the relationships of those concepts that aim to show a view of the world. In other applied domains of science, an ontology consists of a set of standardized vocabularies or resources of a specific domain, so that researchers can control and exchange knowledge between each other in the most convenient way.

Currently, there are several different definitions of ontology by many researchers such as Gruber [75], Zhong [76], and Sarunya [77]. In this section, I focus on a definition of medical ontology: Ontology is the expression of a set of concepts (objects), in a specific domain (medical) and the relationships between these concepts. Ontology is a compilation of standardizing vocabularies and describes its meaning in an understandable way for machines.

3.3.2 The relationship between symptoms in disease

The relationship between symptoms is one of the reasons that causes extensive medical knowledge. In an ontology, there are many relationships to determine the

correspondence between symptoms. This supports doctors to detect the necessary symptom, which is required for making a diagnosis. A basic relation in ontology of disease and symptoms is 'has_symptom', which can be expressed as in Figure 3.6. From the observation, disease DOID_2355 has seven symptoms, which means the patients are likely to have this disease if they have one of these symptoms. Therefore, a request of SYMP_0000610 could be suggested in the case of a patient having all six symptoms SYMP_0000510, SYMP_0000530, SYMP_0019153, SYMP_0000504, SYMP_0019177, SYMP_0000576

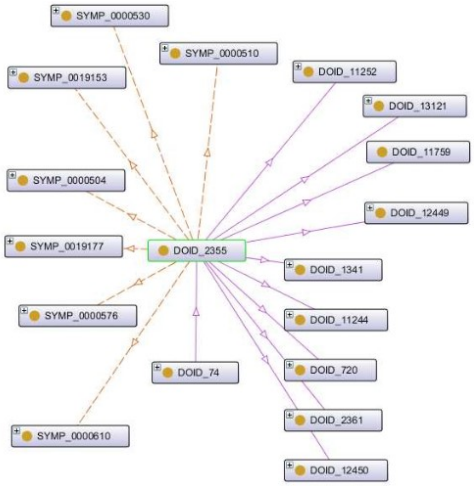

Figure 3.6: Anemia (DOID2355) ISA hierarchy and has symptom relations to its symptoms classes as shown by OntoGraph tab in *Protg*

3.4 Disease classification mechanism of machine learning algorithms

3.4.1 Machine learning

In recent years, machine learning is the popular term with an explosion of technology over the world. There are many problems which were solved by the methods in machine learning such as: automated car, imaging prediction or disease classification. To define the methods in machine learning, this could be a statistical model to generalize the features in training data in order to predict the unknown data. In general, the model in machine learning will be expressed as Figure 3.7. This includes two phases: the training phase and the testing phase. This process is to ensure that the model could work well in the practice. In the model, there are two important parts, which are feature extraction and main algorithm to represent and separate the points of data, respectively.

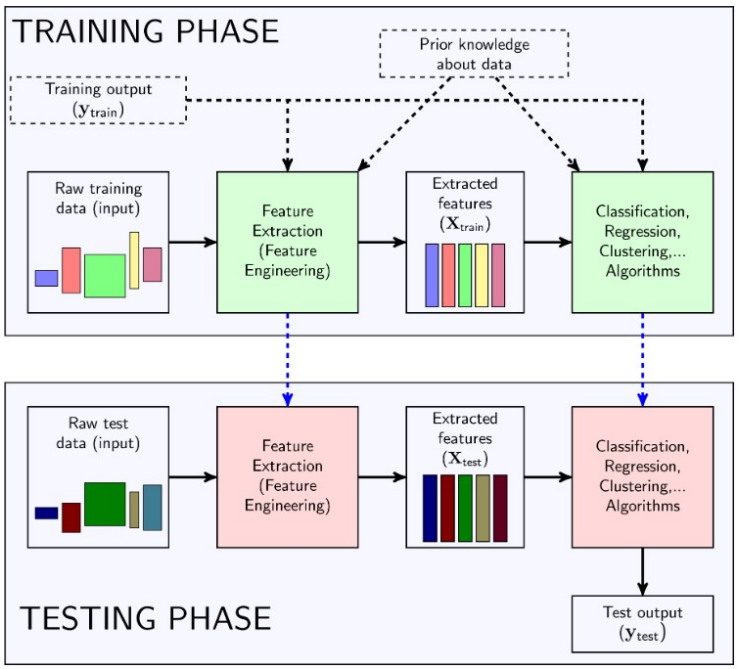

Figure 3.7: The procedure of model in Machine learning

3.4.2 Feature extraction

The main purpose of feature extraction is to transform raw data to suitable data for a specific purpose. The reason is that raw data often includes noise and irrelevant attributes. Hence, removing noise and irrelevant attributes can improve the efficiency of statical models and decrease the computational cost. In the dengue classification task, the raw training input is the clinical symptom of patients and the output is the suitable values for classifying the dengue patients. For example, if the raw input includes too many attributes, the output could be the data without the irrelevant attributes for dengue classification such as gender or name. Regarding the dimension reduction task, there are two main direct approaches: feature transformation and feature selection. In the former approach, by creating combinations of the original attributes, feature transformation techniques [78] summarize a dataset in fewer dimensions. In the latter approach, only the most relevant dimensions from the dataset are selected to reduce dimension by feature selection techniques[79].

3.4.3 Classification algorithms

Classification algorithms are techniques to divide data into categories based on the training example input-output pairs. By optimizing the probability of error lost between prediction and target, the model can learn a function to map an input to output. For example, the square and triangle in Figure 3.8 represent the labeled patients in training data. Using the algorithm in machine learning, we can draw the line which separates patients into two categories such as patients with dengue and healthy people. After the training phase, we labeled the circle which is the non-labeled patients base on their flatness. According to the type of data of these techniques, classification algorithms can be separated into two categories: supervised learning and unsupervised learning.

- In a supervised learning model, the algorithm learns based on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data.
- An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.

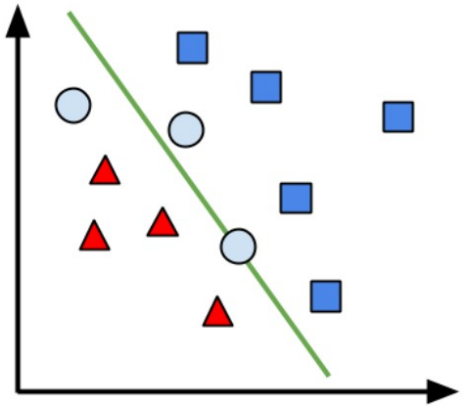

Figure 3.8: The classification problem in Machine learning

3.5 Summary

In this chapter, the background of our study has been briefly presented. Initially, the definition and mechanism of a virtual agent were carefully described. After that, in order to help readers clearly comprehend the approach based on embedding vectors for distributed sentences, a concept of neural-net language model is introduced with principle algorithms. Next, the definition of ontology and the effects of applying ontology in medical knowledge management were also taken into consideration in this chapter. Finally, the mechanism of disease classification based on a machine learning approach was the next focus of this chapter.

Chapter 4

Calculation of semantic similarity sentences for intent detection

In Chapter 3, I mentioned the background of a virtual agent and how to build a virtual agent for health decision support. To support machines to understand the requirement of humans, the process of intent detection is one of the most important factors. As mentioned in previous chapter, the neural network-based approach has better scalability than rule-based approach. In this chapter, I introduce a proposed method based on sentence embedding to calculate semantic similarity between sentences for intent detection. Firstly, I will explain the background of the study and then describe how the method works. I conducted several experiments to evaluate results to prove the validation of the method. Findings will be discussed in a discussion section in this chapter. Finally, I present a conclusion.

4.1 Introduction

For intent detection, a calculation of semantic similarity between sentences is one of the good approaches. By comparing the pre-defined sentences, the intent of the input sentence is identified, which supports machine to understand the requirements. However, estimating the semantic similarity is a difficult task in natural language processing. Unlike humans, it is difficult for a machine to recognize the same meaning among sentences because the structure and syntax of a sentence in human language are too complex. To completely get the intention of a sentence,

a language model is required to cover not only the meaning of words but also the arrangement of these words, and the context of this sentence. In particular, sentences with different words usually do not have the same meaning, and even the ones with same words might not either. For example, the sentence Peter hits Tom and Tom hits Peter share the same words, but the meanings of the two sentences are different. Moreover, the meaning of a sentence is also affected by the context of its conversation. For instance, if a question in the conversation is: "Are you hungry?", the answer could be: "Yes, I am hungry" which is an indication about the condition of the speaker. On the other hand, if the conversation has a question such as "Do you want something to eat?", the purpose of the answer: "Yes, I am hungry" can be considered as a confirmation that he needs food, same as the answer: "Yes, I want something to eat". A language model is often constructed to help a machine solve the complicated problem mentioned above. To represent the distribution of sentences, there are two main research directions: word embedding and sentence embedding. In this subsection, a summary of methods in both approaches is expressed as follow:

- In the former direction, the success of various word embedding models were developed in the past years such as Word2vec [3], Glove [28], FastText [29] and more recently ELMO [30]. The approaches in this direction assume that the representation of sentence is simply calculated by the average of words embedding. With a word embedding model, a baseline method is computing the Bag-of-words (BoW) of word vectors to get sentence representation but does not really have an effective result because BoW rarely ponders the weight of words in the sentence. Consequently, some recent methods, especially smooth-inverse-frequency (SIF) [31], demonstrate the importance of using weighted average and modify them using singular-value decomposition (SVD). However, ignoring the ordering of words in a sentence is still a disadvantage of the methods in this direction. As an example, although Tom hits Peter and Peter hits Tom consist the same words, the meanings of them are different.

- In the latter direction, sentence embedding models based on auto-encoder architecture were developed and has demonstrated that it can address the words ordering. To solve the problem, the methods in this direction usually encode a full sentence to the vector by putting words consecutively

to a neural network. In this direction, there are two types of approaches such as supervised and unsupervised learning-based method. Firstly, the method based on a supervised training task like InferSent [32] or Universal Sentence Encoder (USE) [33] use the Stanford Natural Language Inference (SNLI) [38], [80] labeled dataset to predict entailment/contradiction. By using the same encoder for two sentences with the gold score which is the semantic similarity value manually defined by humans, these above methods have received an impressing result. Nonetheless, the drawback is that these methods must require high quality dataset. Secondly, Skip Thought Vector [34][35] or Quick Thought [36] the unsupervised training-based methods are proposed with the idea using encoder-decoder models for the sentence, and the surrounding sentences of the given sentence. In this approach, the drawback of high quality dataset is solved because those are applied on the unstructured dialogues.

To the best of our knowledge, all prior researches are performing passable representation of sentences but missing an important point about context. In fact, context plays an important role in the represented sentence, and the meaning of a sentence can only be fully understood within the context. Thus, we propose a model with two encoders for the previous sentences and the current sentence, a decoder to express the next sentence. By covering the influence of previous sentences, our model outperformed the previous studies in the unsupervised approach.

4.2 Methodology

Figure 4.1 depicts the architecture of the proposed model. Based on Encoder-Decoder architecture, our model aims to construct the context vector which is a representation of the input sentences. Then, the next sentence will be generated using this context vector. As we mentioned it in section 4.1, our model considers both the current sentence and the previous sentences as the input to obtain the context vector before inserting to the decoder for generating the next sentence. The proposed model is divided into three parts: encoder, context vector extraction and decoder. Firstly, our model utilizes a Gate recurrent unit (GRU) neural network [81] as the encoder to encode the sentences. Throughout GRU, the next

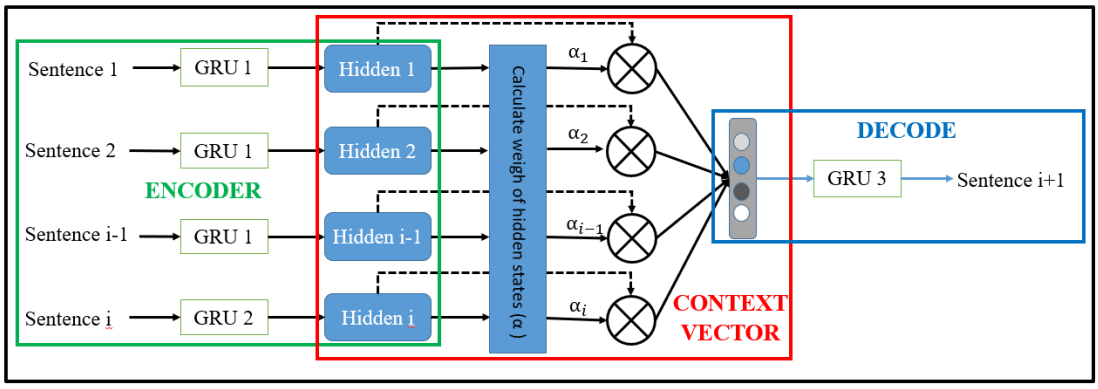

Figure 4.1: Context based sentence embedding model.

word is predicted by the previous words in the sentence. The last hidden state of GRU will be considered as a representation of the input sentence. Furthermore, according to the difference of the effect contributed by the sentences on the next sentence, the current sentence and the past sentences are encoded by two different GRUs. While the meaning of the current sentence is performed to the hidden state by a GRU, another GRU extracts the additional information from the previous sentences to other hidden states for increasing the quality of the representation. In the second part, our model uses the attention mechanism to construct a context vector which represents the meaning of input sentence with its context. Specifically, these hidden states are computed by attention mechanism to get the weight of the input sentences. The average values of weights and hidden states are used to make a context vector. In the last part, the context vector is inserted to eventually generate the next sentence under the operation of the decoder. After the training phrase, the context vector can be considered as a representation of the sentence. More specifically, assume that we have given a list of sentences (S_{i-k} S_i, S_{i+1}) where k is the number of previous sentences. Let W_i^t denote the t^th word for sentence S_i and X_i^t denote for its word embedding. The operation of CSE model will be expressed in three parts: encoder, context extraction and decoder.

4.2.1 Encoder

In the encoder step, we separate the current sentence and the previous sentences. Then, the numerous preceding sentences are encoded by a GRU1. At the same time, GRU2 will transform the current sentence to the hidden states. Regarding

the parameter, both GRUs share the same vocabulary matrix V, but have separate parameters. More specifically, given a sentence, words $W_i^1 \ W_i^N$ in sentence i^{th} will be encoded to the N hidden states by GRU and the last step of the encoding produces h_i^N, which is considered as the representation of sentences. This procedure can be illustrator such as follow:

$$r^t = \delta(W_r x^t + U_r h^{t-1}). \tag{4.1}$$

$$z^t = \delta(W_z x^t + U_z h^{t-1}). \tag{4.2}$$

$$\overline{h}^t = tanh(W x^t + U(r^t \bigodot h^{t-1}). \tag{4.3}$$

$$h^t = (1 - z^t) \bigodot h^{t-1} + z_t \bigodot \overline{h}^t. \tag{4.4}$$

where

- W is the weight matrix in encoder
- U is the weight matrix in encoder
- x^t is word at t^th in sentence
- h^t is the hidden state of the word x^t
- r^t is the reset gate
- z^t is the update gate
- $\bigodot$ denotes a component-wise product
- $\overline{h}^t$ is the proposed state update at time t

All update gates take values between zero and one.

4.2.2 Context extraction

In this step, the given last hidden states h_i^N from encoder step are used to get the context vector. With the definition of the last hidden state which is considered as a representation, the attention mechanism is applied to help the model focus on the important information. More specifically, after encoding the input sentences, h_i^N will represent the first sentence in the dialogue and h_T^N is the hidden state of the current sentence where N is the max length of sentence in the dataset, T is the sum of the number of previous sentences and the current sentence. Using Bahdanaus attention mechanism [82] , the weight of hidden states will be calculated as follows:

$$\begin{aligned} \alpha_j &= align(h_T^N, h_j^N) \\ &= v_a^T tanh(W_a cs + U_a ps), \quad where \quad cs = h_T^N, ps = h_j^N \end{aligned} \tag{4.5}$$

From 4.5, the context vector is expressed such as the following:

$$c = \Sigma_{i=1}^{T} \alpha_i h_i^N. \tag{4.6}$$

where

- c is the context vector
- α_i is the weight of the sentence i^{th}.

4.2.3 Decoder

At the last step, the operation in decoder is the same as in encoder, except for only the first input. To generate the next sentence, instead of token <start>, the context vector is firstly inserted to the decoder. Then, the hidden state of words in the next sentence h_{i+1}^t can be computed as follow:

$$r^t = \delta(W_r^d x^t + U_r^d h^{t-1}). \tag{4.7}$$

$$z^t = \delta(W_z^d x^t + U_z^d h^{t-1}). \tag{4.8}$$

$$\overline{h}^t = tanh(W^d x^t + U(r^t \bigodot h^{t-1}). \tag{4.9}$$

$$h_{i+1}^t = (1 - z) \bigodot h^{t-1} + z_t \bigodot \overline{h}^t. \tag{4.10}$$

where

- W_r^d is the weight matrix in encoder
- U_r^d is the weight matrix in encoder
- W_z^d is the weight matrix in encoder
- U_z^d is the weight matrix in encoder
- h^t is the hidden state of the word x^t
- r^t is the reset gate
- z^t is the update gate
- $(\bigodot)$ denotes a component-wise product
- $\overline{h}^t$ is the proposed state update at time t

Hence, given $t-1$ previous words and the context vector of encoder, we can recognize the word t in the next sentence.

4.3 Experimental results

Our experiment is divided into two parts: training and evaluation on the CSE model. In the training part, Skip-Thought and our model were applied to the DailyDialog [83] which was recently a popular dataset including 13118 daily conversations.

For comparison between Skip-Thought and our model, we utilized the same setting in Table 4.1 for training to predict the next sentence. The loss was computed by the comparison between the prediction and the target sentence.

Table 4.1: The setting in training of sentence embedding

Parameters	Value
Batch size	64
Embedding vector	512
Learning rate	5e-4
Optimization algorithm	Adam [84]
Iterations	1000000

Besides, to evaluate the influence of the previous sentences, we implemented two instances of our model: CSE1 is a model using one previous sentence and CSE2 is a model using two previous sentences.

After the training part, we evaluated Skip-Thought and our model on 1379 pairs of sentences of the STS benchmark dataset [85] and the SICK dataset [86] with 10000 records following the idea proposed by Tai et al. [87] as shown in Figure 4.2 instead of calculating based on cosine similarity. Compared to cosine similarity, this method allows the weights to be learned while cosine similarity applies the same weight for all features. The procedure of this method is carried out as follows. Firstly, a given pair of sentences will be encoded to the representation vector u and v. Then, we extract the relation between these two sentences. In order to do that, the concatenation of the component-wise product $u \cdot v$ and the absolute difference $|u - v|$ are regarded as the features for the given sentences pair. Finally, we train a logistic regression to predict a semantic relatedness for the given sentences pair.

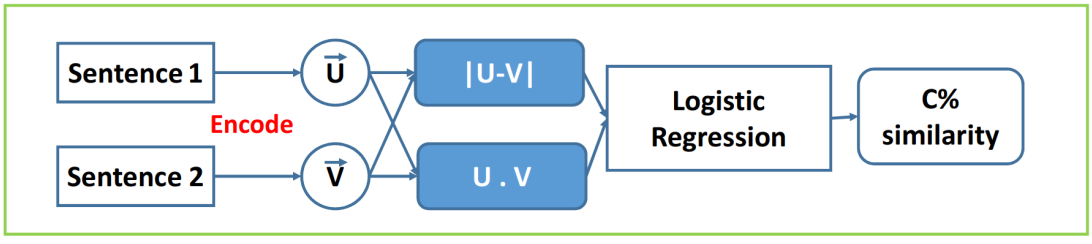

Figure 4.2: The operation of calculating semantic relatedness

For evaluation, Pearson [88] and Spearman correlation are the measurements to estimate the quality of sentence embedding.

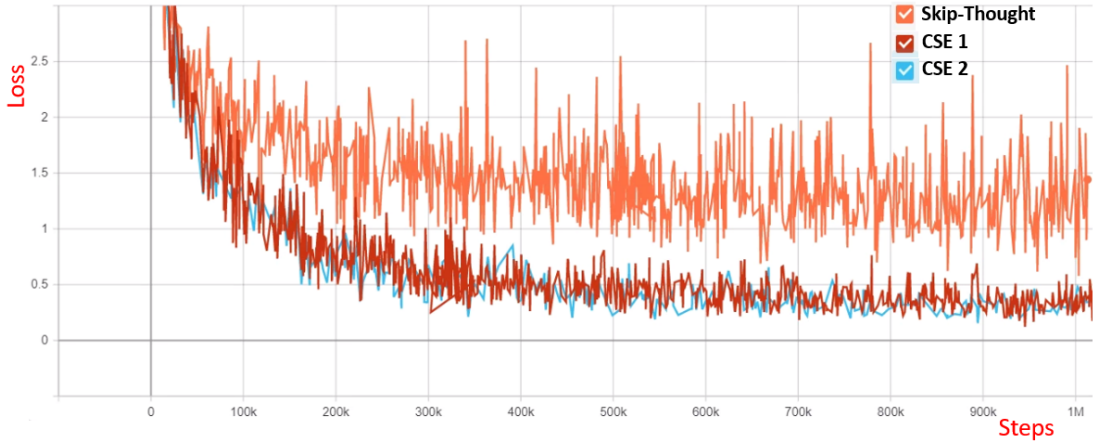

Figure 4.3: Comparison loss of CSE model and Skip-Thought.

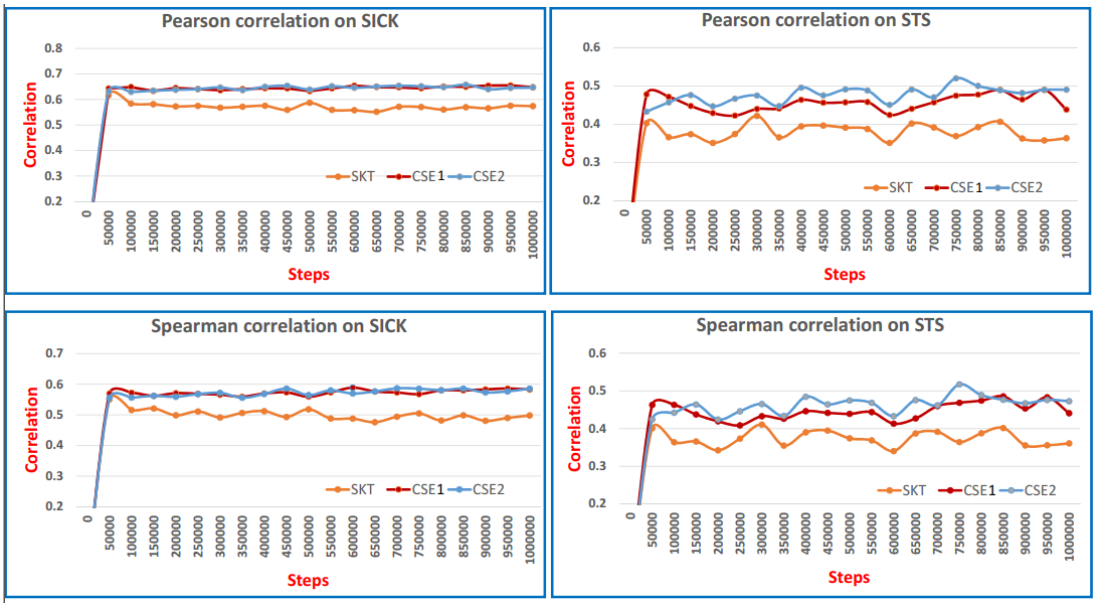

Figure 4.4: Comparison of evaluation of CSE and Skip-Thought

Table 4.2: The best Pearson and Spearman corrlations in 20 checkpoints

Channels	SkipThought	CSE1	CSE2
SICK Pearson	61.66%	65.5%	**65.8%**
SICK Spearman	55.09%	**58.88%**	58.67%
STS Pearson	42.19%	48.98%	**52.01%**
STS Spearman	41.07%	48.59%	**51.74%**

4.4 Discussion

The primary contribution of the proposed model is to find the influence of previous sentences on the meaning of the current sentence. This influence is equivalent to the context of the current sentence which reinforces the clarity of its meaning. This is an important task to separate the sentences having identical words but different meaning. Hence, the representation of sentences is considerably more accurate.

As mentioned, Infersent, USE, and Skip-Thought are the state of art models in calculating semantic similarity between sentences. Therefore, in order to evaluate our model, the comparison with these models is necessary. However, Infersent and USE utilized a different structure with SkipThought and our model so the proposed model was only compared with them through Skip-Thought. Specifically, they used pairs of sentences with labels judged by humans while SkipThought and the proposed model used the conversations including previous sentences. Due to the equivalent accuracy of three models [89] in calculating semantic similarity on SICK and STSBenchmark datasets, the proposed model is expected to be better than all models if it is better than Skip-Thought model. Concretely, Sick and STSBenchmark datasets were used because the Pearson and Spearman correlation of three models above on these datasets is approximated.

In our experiment, our model was shown that it outperformed Skip-Thought model. Figure 4.3 shows that the loss of both models was almost convergence. However, while the prediction of the proposed model was approximated to the target sentence (the loss computed by the predicted and the target sentence was 0.33), the prediction of Skip-Thought model still needed more time to get the best result. In addition, according to the semantic relatedness task, the results in

Table 4.2 indicated that the proposed model also is better. All values of Pearson and Spearman correlation coefficient of our model were higher, especially the difference between the two models in the STSBenchmark dataset is approximate 10%. Besides, Figure 4.4 shows that the worse result in the proposed model also is even better than the best result of Skip-Though.

Regarding the training process, Figure 4.4 also indicates that the early stop can be applied after 100000 steps because the proportion between the highest and the lowest is not too different. In addition, Table 4.1 shows that those models include more information from history which often perform better. With more information, CSE2 achieved three favorable results except for the Spearman correlation on SICK dataset where CSE1 achieved a better result (58.88%). As observed from Figure 4.4, with a large dataset like SICK dataset, the variety in the results of our proposed models two instances are not significantly different. However, with a small dataset such as STS, the model CSE2 is moderately better than CSE1. This can be understood that the model with more information will be more stable.

Although the proposed model demonstrated the efficiency of embedding sentence based on context, the result still needs to be improved. As presented in [34], Skip-Thought could achieve a Pearson correlation 86.76% and Spearman correlation 80.83% on the SICK dataset which higher than our result. However, their experiment was applied on a much higher configuration: a bigger training dataset with 74 million sentences, embedding vector 2400 dimensions, and takes a much larger training time. At this moment, because of the limitation of equipment and time, we cannot make an experiment with a large training dataset. Consequently, the proposed model need to be verified with a large training dataset.

4.5 Summary

In this chapter, we introduced a new approach to perform the representation of sentences. To do this, the proposed model simulated the dialogue with two GRU for the current sentence and previous sentences in an encoder, one GRU in a decoder to predict the next sentence. Then, the evaluation of the quality of sentence embedding is processed on the SICK and STSBenchmark datasets. The experimental result shows that CSE model can transform a sentence to a good

sentence embedding to calculate the semantic similarity. However, the dataset used for training is too small so there are some obstacles in transforming sentences to embedding. In the future, we will apply the proposed model to a large dataset and evaluate model again.

Chapter 5

Symptom suggestion based on conversation and ontology for simulation of medical examination

In Chapter 4, a proposal to assist virtual agents in determining human intent is presented. In this chapter, I will introduce a potential method to teach medical knowledge to machines, thereby, machines can understand what necessary symptoms need to be collected based on the detected intent and the extracted symptoms. Similar to the operation of medical examinations from doctors, machines could collect human symptoms depending on the specific situation.

5.1 Introduction

Theoretically, collecting the symptoms of a patient is a routine process in the health care operations of doctors and plays an important role in making a diagnosis. During conversations, doctors will receive information related to the patient's health by asking them to confirm the status of the symptoms. This process is a loop to confirm all suspected symptoms of patients until doctors have a diagnosis for patients. To carry out the process, doctors must be humans, who are equipped with the medical knowledge to understand what necessary symptoms need to be collected in specific situations. In other words, they need to

Table 5.1: The conversation between doctors and patients.

Doctor: Good morning! Whats wrong with you?
Patient: I have been suffering from fever since yesterday.
Doctor: Did you have a vomiting?
Patient: Yes, I have vomited 2 times.
Doctor: Did you have any skin bleeding?
Patient: Yes, a little in my back.
...

discover hidden information from the relationship between symptoms to find the candidate disease, and then request the missing symptoms from patients. Typically, there are two approaches: the first is followed by the probability of diseases with the input symptoms and the second is followed by the relationship between symptoms. In the first approach, the suggested symptom is a symptom in the most highly probable disease. In the second one, the result is a symptom which has a maximum number of appearance with the input symptoms. In fact, the first approach is more preferable than the second one because it focuses on the patients' disease while the second one does not. This sometimes leads to noises of the collected symptoms from the second approach, which impacts the processing of clinical diagnosis.

However, arranging this knowledge is really a challenge not only for virtual assistants but also for humans. For example, in the conversation in Table 5.1, patients with dengue fever, who have only recognized that they have a symptom of high temperature and do not provide any other information. This causes an obstacle in making a diagnosis because there are various diseases, which also consist the same symptom.

Despite the difficulties of medical knowledge, introducing medical knowledge to virtual agents play a crucial role in symptom collection of patients. For medical virtual agent development, a popular method is to design only a system in a specific domain such as dengue decision support or cancer decision support with fixed symptoms. Through the predefined forms, the necessary symptoms of patients will be simply obtained for making a diagnosis. Unfortunately, this approach often requires additional preprocessing, which is performed by a physi-

cian to clinical classification. Therefore, for the efficiency of the predefined-based approach, it is not enough for a virtual agent in the medical topic. Another approach is to use a query to the semantic database such as an ontology which includes the medical knowledge with diseases, symptoms, and their relationship. By finding a disease based on relations, the necessary subsequent symptoms can be suggested. However, due to the overlapping of symptoms between diseases, it is difficult for machine to determine disease in the case of more than two candidates of diseases. Specifically, if the given input symptoms include S_1 and S_2 where S_1 is a symptom of disease D_1 and S_2 belongs to D_2, the outcome is undefined because of the equivalent probability between these candidate of diseases.

Therefore, a symptom suggestion method based on weights of relations and ontology is introduced in this chapter. Weights are determined as a probability to have a disease if there are symptoms to solve the problem of having more candidate diseases. Hence, the efficiency of a symptoms suggestion system increases.

5.2 Background

Typically, symptoms suggestion system based on the probability of diseases comprised of two steps: ranking candidates' diseases and suggesting a symptom of the highest result. The most difficult part is ranking candidates' disease due to the overlapping of symptoms between diseases. In terms of statistics, the processing of calculating probability of candidates' diseases can be expressed by the Bayesian theory as follows:

$$\begin{aligned} P(D_i|S_1, S_2, ..., S_j) &= \frac{P(D_i)P(S_1, S_2, ...S_j|D_i)}{P(S_1) \times P(S_2) \times ... \times P(S_j)} \\ &= \frac{P(D_i)P(S_1|D_i) \times P(S_2|D_i) \times ... \times P(S_j|D_i)}{P(S_1) \times P(S_2) \times ... \times P(S_j)} \end{aligned} \tag{5.1}$$

where

- $P(D_i)$ is the probability of having disease D_i
- $P(S_j)$ is the probability of having symptom S_j
- $P(D_i|S_j)$ is the likelihood of have disease D_i if having symptoms S_j

- $P(S_j|D_i)$ is the likelihood of have symptoms S_j if having disease D_i

Consider $\mathrm{P}(S_1) \times \mathrm{P}(S_2) \times ... \times \mathrm{P}(S_j)$ as a normalising constant because the probability of symptoms is unchanged, the Equation 5.1 can be converted to the Equation 5.2:

$$P(D_i|S_1, S_2, ..., S_j) = \alpha P(D_i) P(S_1|D_i) \times P(S_2|D_i) \times ... \times P(S_j|D_i) \tag{5.2}$$

Through Equation 5.2, two factors that influence the calculation of the probability of diseases with the input symptoms are identified. The first factor is $\mathrm{P}(D_i)$ to indicate the probability of specific disease, which usually depends on many elements such as environments, location, and prevalence of disease D_i. In this research, we temporarily focus only the prevalence of disease, for example, the probability of popular disease such as Influenza should be higher than a rare disease such as cancer. The second factor is $\mathrm{P}(S_j|D_i)$ which shows the probability of having symptom S_j of disease D_i. While the first factor can be simply computed by statistics of positive cases in total cases of diseases as show in Equation 5.3, calculating the second factor is a challenge due to the overlapping of symptoms between diseases.

$$P(D_i) = \frac{positivecases}{totalcases} \tag{5.3}$$

5.3 Methodology

In this section, the proposed method of this research are introduced in details in order to explain how they connect to the proposal.

Looking back to Section 5.1 in this research, a new method to exploit the relationships between entities in ontology for symptoms suggestion component is proposed. More concretely, the collaborative approach using the ontology and word embedding vectors to clarify a mapping function between input symptoms and the related symptoms is thoroughly discussed in this section. Initially, the entities and their relations are extracted first for constructing the structure of

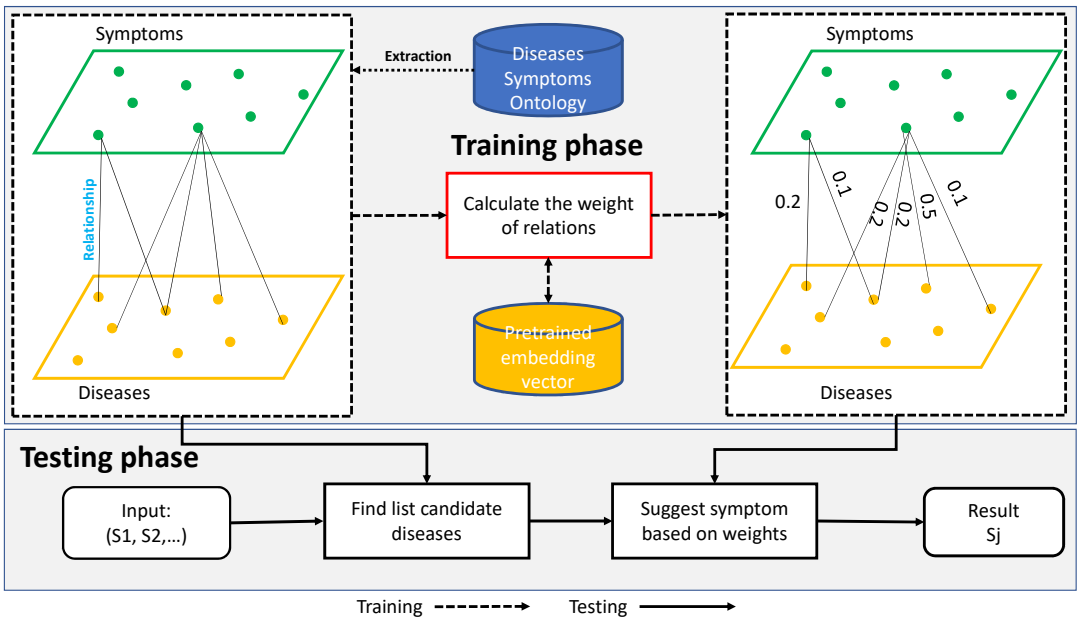

Figure 5.1: The architecture of symptom suggestion method based on weights and ontology

the medical database. Afterward, a method calculates the weights of these relations, which are assigned to these entities to support machines to self-learn the relationship among the entities.

Specifically, Figure 5.1 depicts the architecture of the proposed method. The proposed method is divided into two phases, training phase and testing phase. In training phase, there are two tasks: extraction of entities and their relationships, and calculation of the weight of relations. In the first task, our method collects the entities and their relations by using an ontology of diseases and symptoms. In the second task, our method utilizes a pre-trained word embedding vectors [3] and node embedding vector [90] to calculate the weight of the extracted relations and construct a weight-table. Afterward, in the testing phase, the input symptoms can be considered as a query, which requests to the ontology and weights-table to find the disease, then suggest the subsequent symptom.

5.3.1 Making an ontology

As mentioned in the background section, an ontology is considered as a semantic database, which is used to represent the terms and the relationships between them. Typically, an ontology is constructed by the ontology languages such as XML, RDF, and OWL. In general, to construct an ontology, there are some common steps, such as:

- Determine classes in ontology.
- Arrange classes in a taxonomic hierarchy.
- Determine the properties (slot) and describe the allowable values for these properties.
- Fill the values of instances into the slots.
- The knowledge is then created by defining instances of these classes with their values.

For making disease symptom ontology, it is crucial to have medical knowledge due to the complex taxonomic hierarchy of classes about diseases and symptoms. Basically, the disease symptom ontology includes not only a relationship

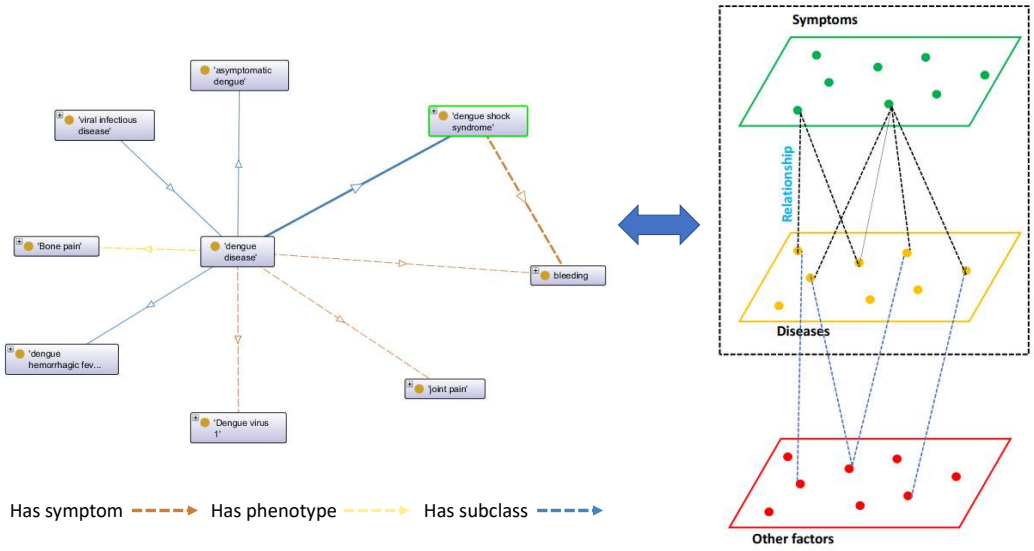

Figure 5.2: The sample relations architecture of ontology

between diseases and symptoms but also other relationships such as "ecologically_related_to" and "has_phenotype" between diseases and other factors as show in Figure 5.2. Therefore, this is usually created by experts in medical topics to ensure the credibility of the knowledge.

5.3.2 Medical knowledge extraction

A disease symptom ontology includes the terms which are entities about diseases and their symptoms, and the relationships between entities. By using queries to the ontology, these medical knowledge should be collected. In a simple model, the only relation is extracted, whose name is "has_symptom" for a description of the relationship between disease and symptom.

Figure 5.3 illustrates a sample ontology with two diseases such as dengue and influenza. The sample output is illustrated in the Table 5.2.

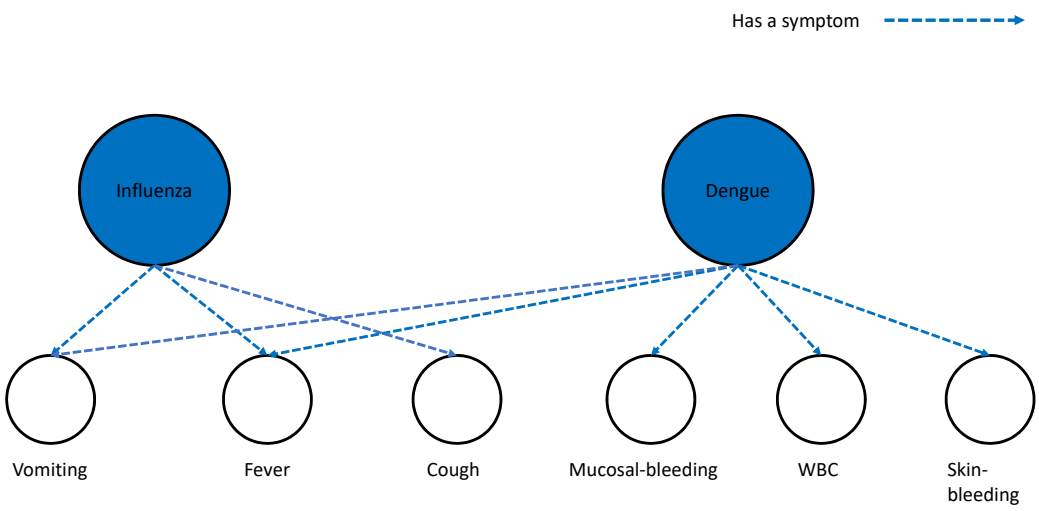

Figure 5.3: The sample ontology with two diseases and their symptoms

Table 5.2: The sample extracted medical knowledge from a dengue disease and their symptoms

Disease	**Symptoms**	**Relations**
Dengue, Influenza.	Vomiting, Fever, Cough, Mucosal-bleeding, WBC, Skin-bleeding.	Dengue-Vomiting, Dengue-Fever, Dengue-Mucosal_bleeding, Dengue-WBC, Dengue-Skin_bleeding, Influenza-Vomiting, Influenza-Fever, Influenza-Cough

5.3.3 Calculate weight of relations based on embedding vectors

After building a medical knowledge database, entities and their relations are input into a function to calculate weights based on embedding vectors. Specifically, a calculating weight of the relationship between a pair of disease and symptom, which is done such as following:

- Transform disease and symptom into embedding vectors by a pre-trained

model.

- Calculate the cosine similarity of these embedding vectors. The result is the weight of a relationship between this pair of disease and symptom.

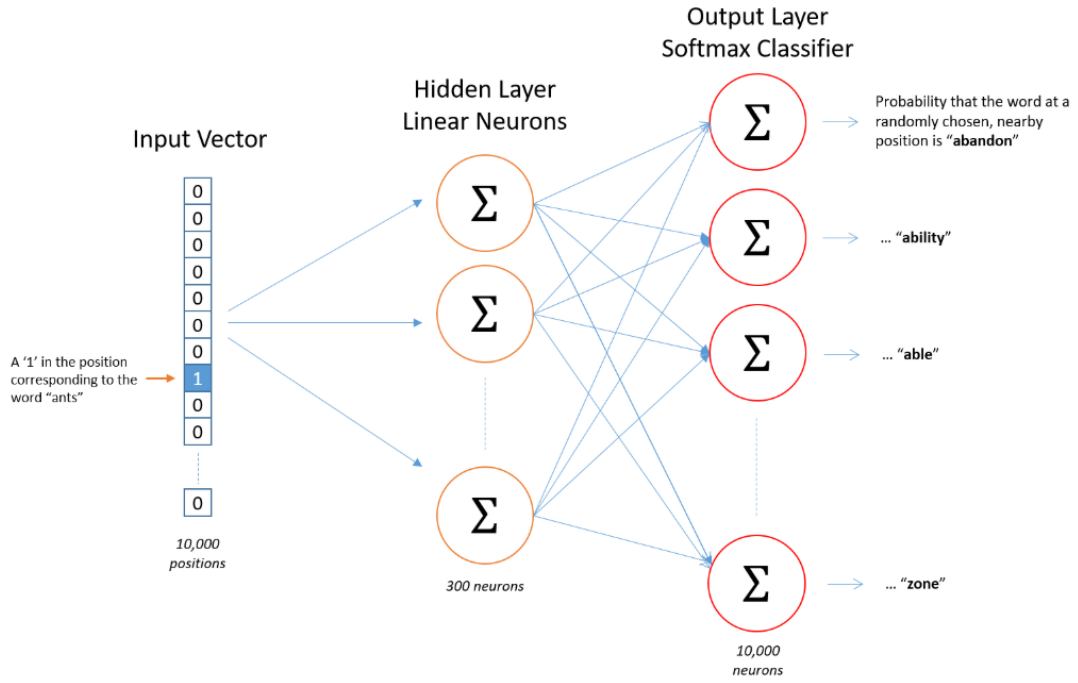

Figure 5.4: The architecture of represent words by embedding vector [3]

The reason for using embedding vectors is that it can represent the semantic and relationship of words in the space, while the ontology-based method cannot determine due to their structure. In terms of theory, the occurrence of diseases and symptoms are in the same context, which indicates the weights of the relationship between them. In other words, the disease and symptoms appear together several times, which represents the strong relationship and the high probability of having the disease with these symptoms. Through using a simple neural network, all words are transformed to embedding vector by a hidden layer and mapped to their context by the activation function at the last layer as shown in Figure 5.4. After training, the outputs of the hidden layer is used to represent the words as the embedding vectors.

In order to improve the quality of embedding vectors, the proposed method utilizes a combination of two pre-trained embedding vectors: word2vec and node2vec. This is because we want to represent not only the semantic but also the structure

between diseases and symptoms by words embedding vectors v^W and node embedding vectors v^G, respectively, as show in Figure 5.5. The redundant attributes of the combination of two embedding vectors is then removed by the Principal Component Analysis (PCA) algorithm [91, 92]. After that, the relationship between diseases and symptoms are measured based on the cosine similarity.

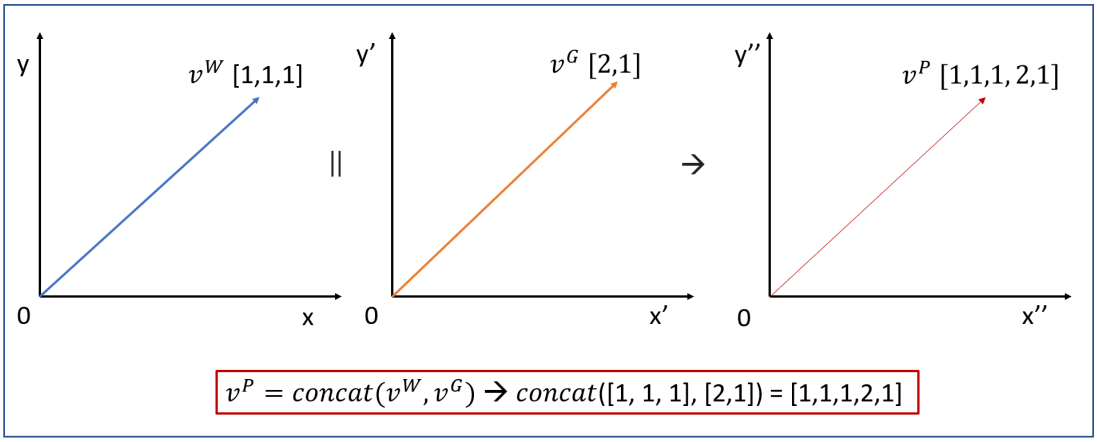

Figure 5.5: The concentrated of word and node embedding vectors

Assume that the input symptoms includes symptom S_1, S_2, and the candidate diseases are $D_1\{S_1, S_2, S_3, S_4\}$, $D_2\{S_1, S_2, S_5\}$. The weights W_{ij} between a pair of disease D_i and symptom S_i should be calculated by the Equation 5.1

$$W_{ij} = \cos\theta_{ij} \tag{5.4}$$

where

- W_{ij} is a weight of relationship between disease D_i and symptom S_j
- $cos\theta_{ij}$ is cosine similarity of disease D_i and symptom S_j

From that, the probability of candidate can be calculated by the Equation 5.2

$$P_i = \frac{\Sigma_{j=1}^{n} W_{ij}}{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n} W_{ij}}. \tag{5.5}$$

where

- P_i is the probability of candidate disease D_i

- W_{ij} is a weight of relationship between disease D_i and input symptom S_j
- n is total of input symptoms
- m is total of candidate diseases

However, the $\cos\theta_{ij}$ between vector disease D_i and vector symptom S_j can not be negative but should ranges from [0, 1]. Therefore, the Equation 5.2 need to convert to the 5.3 such as following

$$P_i = \frac{\Sigma_{j=1}^{n} e^{W_{ij}}}{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n} e^{W_{ij}}}. \tag{5.6}$$

5.3.4 Symptom suggestion based on weights of relations

After constructing a weight-table, in order to suggest the symptoms, there are two steps as follows:

- Finding the candidate diseases by a query into a disease symptom ontology.
- Ranking these candidate by using the weight-table

5.4 Experimental results

The primary objective of the experiment is to simulate the processing of medical examinations between doctors and patients. Considering the knowledge of doctors as the specific sentences which include disease and its symptoms, this process corresponds with the work of mapping between previous words to the subsequent word. This section aims to demonstrate the ability to use the embedding vector in selecting candidates of diseases based on input symptoms, then requiring the necessary symptoms of the highest candidate. In this section, I firstly express an instant of using embedding vectors in separating similar diseases which include the common symptoms. After that, I describe an experiment comprised of making an ontology based on a disease dataset, constructing pre-trained embedding vectors based on this ontology, and use these pre-trained embedding vectors to predict the candidates of diseases. Hence, the efficiency of using embedding vectors in the proposed model is validated by the accuracy of predictions.

Table 5.3: The setting in training of symptoms suggestion

Parameters	**Value**
Dataset1	Human disease ontology [45]
Diseases classes	12480
Symptoms classes	944
Diseases with "has_symptom" relation	456
Symptoms with "has_symptom" relation	131
Dataset2	Kaggle's Disease dataset
Diseases classes	41
Symptoms classes	131
Number of relationships	321
Diseases with "has_symptom" relation	41
Pre-trained embedding vector	GoogleNews-vectors-negative300.bin
Embedding vector	300

In the first experiment, I only applied the proposed model to the existing resources which are a human disease ontology [45] and a provided pre-trained embedding vector by Google with the information as shown in Table 5.3. From the ontology, there are 456 extracted diseases with 131 symptoms and 771 relations. However, only 123 diseases and 102 symptoms can be represented by word embedding vectors due to the difference between science terms and daily words. These extracted entities were applied by the node2vec algorithm and then combined with a pre-trained embedding vector by Google to construct the proposed trained embedding model. After that, we evaluated the proposed model based on a scenario of predicting the highest candidate diseases based on their symptoms. The corpus of input symptoms and the output is extracted from the conversations between doctors and patients, which is collected from the Internet. The result is considered correct when the prediction of the proposed model corresponds with the collected output. Specifically, Table 5.4 illustrates two conversations about dengue and influenza. By the proposed method, we can separate cases between dengue and influenza, which are often confusing due to the overlapping of symptoms between them.

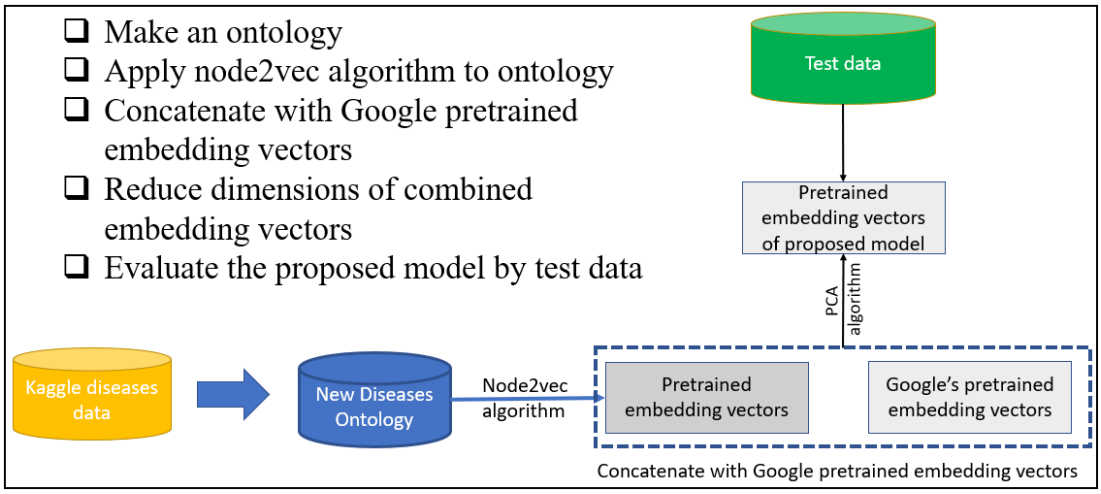

Figure 5.6: The experiment of applying proposed model to Kaggle dataset disease

However, separating two diseases is too simple to show that our method works for more general setting. In the second experiment, the proposed model is applied to the collected dataset from Kaggle, which includes 4920 records of 41 diseases and their symptoms as shown in Figure 5.6. The scenario to validate the proposed model is expressed as follows:

- Firstly, I made an ontology-based on diseases, symptoms, and their relationships. In this step, diseases and symptoms were represented as the nodes and the relationships between them were the edges in the ontology. Concretely, I created 172 nodes and 321 edges.

- Secondly, the constructed ontology was applied by the node2vec algorithm and combined with the Google's pre-trained embedding vectors. In more details, ontology is considered as input of node2vec algorithm with the setting of embedding vectors' dimensions is 200. After that, the output of node2vec was used to combine with the Google's pre-trained embedding vectors. The dimensions of the combined vectors was reduced by the PCA algorithm. As a result, the pre-trained of the proposed model is a model to represent diseases and symptoms by the embedding vectors' size is 300.

- Finally, I made a test dataset by randomly removing a few symptoms from records. Next, this dataset is used to evaluate the performance of the proposed model. Specifically, for each record, there are 41 predictions of candidate diseases. The probabilities are calculated based on the pre-trained

model above and the Equation 5.6. The highest candidate disease is then compared with the actual result in the test dataset to measure the accuracy of predictions.

To clarify the performance of the proposed model, I applied two scenarios of removing symptom in Kaggle dataset: a random symptom and two random symptoms. For each record, I attempt to predict the disease and compare it with the original correct answer. For example, the correct prediction in Table 5.5 is the case of the predicted result and original correct answer are similar. Otherwise, this is the incorrect prediction.

As a result, Table 5.6 indicates that the efficiency of proposed model were high in both scenarios with the accuracy of 99.73% and 98.45%, respectively.

5.5 Discussion

The primary contribution of this study is to provide a method to automatically suggest the subsequent symptom based on the input symptom. To do that, the most important part needs to be considered, which is an overlapping of symptoms between diseases. Without this part, a machine has not determined the disease in the case of more than two candidate of diseases. We saw that, through the experiment in Table 5.4, the provided symptoms of patients are common, which makes a confusing diagnosis between Dengue and Influenza diseases. By calculating the weights between input symptoms and each disease, the probability of each disease was determined. Moreover, by observation an example in Table 5.5, the prediction was wrong when the dischromic-patches symptom is missing, thus, dischromic-patches is an important symptom to detect fungal-infection disease. As a result, the high accuracy of the proposed model in predicting candidate diseases was achieved as shown in Table 5.6, which indicates the efficiency of the proposed model in calculating the weights of diseases and symptoms. Thus, this supports to suggest the missing subsequent symptoms.

The advantage of this method is the scalability when applying to the huge database due to the self-learn mechanism. This is an important point to provide a the large medical knowledge to the model with the overlap of symptom among diseases.

Looking back to the experiment section, there is a problem of difference between science terms and daily words, which is the reason why I had to make an ontology based on the Kaggle dataset but not using the original ontology (Human disease ontology). Specifically, during the processing of experiment, I tried to utilized the original ontology to predict probability of candidates' diseases in the Kaggle dataset but the accuracy is quite low, which is around 18%. This is because there are too much symptoms in records, which does not appear in the original ontology. These symptoms might be the important attributes which support to detect a disease.

Although this method could solve the problem of finding related symptom based on the input symptom, this method needs evaluation with a judgment of participants who works in the medical field. Besides, only a relationship "have_symptom" could not express all medical knowledge of diseases and symptoms. Specifically, Figure 5.2 shows that there are many potential relations inside an ontology such as the "ecologically_related_to", "has_material_basis_in" and "has_phenotype". Therefore, various relations need to be clarified in the future. For example, with the same scenario in Table 5.4, doctors will prefer a dengue candidate disease in the summer. On the contrary, they usually select candidate of Influenza disease if the timing in the winter. Furthermore, the evaluation dataset is still simple and lack of various similar diseases. Therefore, the proposed model needs to be clarified by using a large database.

5.6 Summary

In this chapter, the relationship between disease and symptoms have been modeled for prediction of subsequent symptom. A novel model based on the collaboration between ontology and embedding vector is proposed to map the previous input symptoms to a target. This process can be determined as the ranking of candidate symptoms based on the probability. The experiment result shows that the proposed model can solve a problem of discovering relationship between symptoms.

However, the used ontology includes a small relationship between diseases and symptoms. Therefore, it is necessary for the model to be clarified in a large ontology with more diseases, symptoms, and relations.

Table 5.4: Evaluation on the conversations of dengue and influenza diseases

Dengue disease's conversation	**Influenza disease's conversation**
Patient: Hi Doctor, good morning Doctor: very good morning, how are you, what is the problem? Patient: I am not feeling well, suffering from **fever** for the last three days. Doctor: Do you have a **bleeding**? Patient: yes doctor. But they told me I am not suffering from malaria. Doctor: ok, are you still having a fever? Patient: yes doctor. Doctor: I have a doubt so you should go for a **dengue** test. Patient: is it a necessary doctor? Doctor: yes, because in this season dengue is severely seen in most of the people. Patient: ok doctor, I will give blood for dengue test after that I will come. Doctor: sure don't neglect it is a serious problem when we neglect. Patient: ok doctor I will meet u after getting reports. Doctor: ok take care bye. Patient: bye doctor.	Patient: Hello Doctor. Doctor: Hello, how are you feeling today? Patient: Not very well, Doctor Doctor: Ok, tell me about it Patient: Well, I have a terrible **headache**. Doctor: How about your **throat**? Patient: It hurts a little. Doctor: Ok. Do you have a **cough**? Patient: Yes, I have a cough, too. Doctor: Do you feel weak? Patient: Yes, I get tired very quickly. Doctor: Let me take your temperature. Hmmm, you have a little **fever**. It seems that you have the **flu**. Patient: Oh, thats terrible. Doctor: Dont worry. Take this medicine and rest. Patient: Ok. I understand.
Input symptoms: fever, bleeding Result of proposed method: **Dengue** Probability of Dengue: 53% {fever:0.37, bleeding:0.14} Influenza: 47% {fever:0.37, bleeding:0.08}	Input symptoms: headache, cough, throat, fever Result of proposed method: **Influenza** Dengue: 46% {headache: 0.07, cough: 0.25, throat: 0.04, fever: 0.37} Influenza: 54% {headache: 0.12, cough: 0.31, throat: 0.05, fever: 0.37}

Table 5.5: The examples of predictions based on the proposed model

Each of record	**Input symptoms**	**Removed symptoms**	**Predicted result**	**Original correct answer**
Correct prediction	Nodal-skin-eruptions, dischromic-patches	Itching	Fungal-infection	Fungal-infection
Incorrect prediction	Itching, skin-rash	Dischromic-patches	Drug-reaction	Fungal-infection

Table 5.6: The accuracy of the proposed model in predicting highest candidate of diseases

	A random symptom	**Two random symptoms**
Accuracy	98.67%	97.5%

Chapter 6

Dengue classification based on the clinical symptom

The proposals presented in Chapter 4 and Chapter 5 have efficiently solved the issues of intent detection and symptoms suggestion for making a medical checkup conversation between humans and machines. This chapter introduces a proposed method for the classification of disease, towards the research goal, that is to say, making a diagnosis for patients about dengue fever.

6.1 Introduction

Dengue fever is a severe infectious disease happening in over 150 countries, mostly in tropical and subtropical regions [93]. Primary dengue infection in humans usually leads to a variety of clinical symptoms, such as mild fever, even death in the worst case. The primary transmission vector of the dengue virus is Aedes mosquitoes [94], which often reside in tropical regions. Through biting, they transmit dengue from patients to healthy people. Due to its simple transformation, dengue has become a public health problem that significantly affects human health. More dangerously, until now, there is no specific treatment for dengue, such as the vaccine. The patients usually get only supportive care with analgesics, fluid replacement, bed rest, and acetaminophen to treat the fever. Fortunately, as presented in [95, 96], early detection and access to proper medical care can lower fatality rate to below 1%. Also, the recovery of patients is high if they receive treatment within the first three days.

However, dengue fever is difficult to be diagnosed because the symptoms are similar to other diseases such as roseola or fever virus. In traditional methods, the used solutions are often blood tests. Specifically, the Dengue NS1-Ag assay is a rapid test to determine the NS1 antigen in dengue plasma or serum from patients. Generally, NS1-Ag appears in the blood from day one to day nine of the disease; thus, it is used as a tool to help doctors perform diagnosis. Unfortunately, despite being the gold standard diagnostic, this rapid test is not routinely available in developing countries. The primary reason is due to the lack of well-trained doctors who can perform the NS1-Ag test. Unlike the regular tests, NS1-Ag is easily affected by the external factors from the environment, such as moisture. Because of the carelessness orthe lack of necessary skill from doctors, the result of tests could be wrong. Another reason for the late dengue detection of patients is the expensive cost of this test compared to their financial ability. These poor patients often refuse the test until they are in a dangerous situation.

Meanwhile, Machine learning, including the statistical models, can be a good solution for the above problem. These statistic-based methods share the same advantages, which are rapid and simple to be widely executed in developing countries. Based on the values of clinical blood tests as the input, the algorithms in Machine learning quickly make a prediction of dengue fever for patients. From that, it can reduce the pressure for doctors and increase the ability of early dengue detection for patients.

The greatest obstacle machine learning techniques must deal with is high-dimensional data. Such data include irrelevant attributes and noise, which results in high computational costs. A common solution is a dimensional reduction to remove noise, sparse outlying entries, and missing entries. This approach can be recognized as one of the feature transformations [78] and feature selection techniques[79]. With such approaches, irrelevant dimensions and redundant attributes are removed such that computational time and memory requirements are reduced without affecting accuracy. However, with such methods, the meaning of subspaces may be overlooked.

In order to improve the efficiency of the classification algorithms, we propose the model to reduce data dimension and find hidden features to improve classification accuracy and reduce computational costs. Consequently, dengue fever can be detected with minimal symptoms with high accuracy.

6.2 Methodology

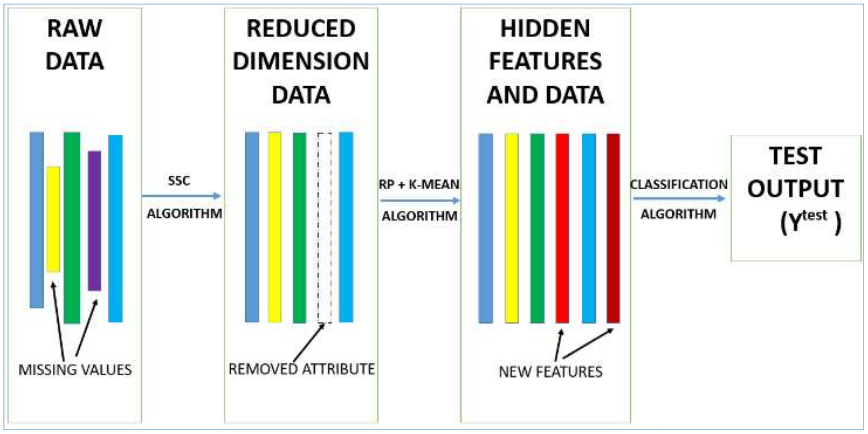

Figure 6.1: The procedure of the HSC model.

The proposed model is called Hidden Subspaces Clustering (HSC). This model aims to find quality information from a small group of attributes for improving the accuracy of classification algorithms. This model includes three main parts: reduction dimensions, finding hidden features and classification for removing redundant attributes, find useful information, and detect disease, respectively. In details, the operations of the model are summarized in Figure 6.1 and Table 6.1. Specifically, since the raw data may contain the records coupled with missing values, preprocessing the dataset is necessary. This process comprises of replacing the missing values with the mean of attributes, performing the normalization, and randomly taking the data for the training process. Then, irrelevant dimensions are removed using the sparse subspace clustering (SSC) algorithm [97], [98]. In the next step, hidden features are identified and combined with the data. Finally, classification algorithms are applied to the dataset to obtain results.

The details of each part are described as follows.

Table 6.1: The HSC model

Model: Hidden Subspace Clustering (HSC)
Input: A union of data points y which lie in a high dimension
1. Cleaning data: replace missing data. 2. Reduce dimensions of data by sparse subspace clustering (SSC) algorithm. 3. Find the hidden dimensions. • Project subspaces to new areas by random projection (RP) algorithm. • Evaluate the subspaces to find meaningful hidden features. 4. A classifier with methods such as logistic regression(LR), Support Vector Machine (SVM), Random Forest (RF) to find the optimal prediction algorithm .
Output: label of class.

6.2.1 Dimensions Reduction

High-dimensional data typically includes irrelevant attributes and noise. This increases the computational costs and negatively impacts classification accuracy. To address this issue, dimension reduction is a crucial solution that can be performed by applying one of the following subspace clustering approaches: iterative, algebraic, statistical and spectral clustering [97, 98, 99, 100]. Each class has its advantages; however, spectral clustering-based algorithms are used more commonly because of their ability to solve noise and outliers in data. Also, knowing the dimensions and number of subspaces is not required. The proposed model can handle noise, missing data for dimension reduction by using a spectral clustering algorithm [97], namely the Sparse Subspace Clustering algorithm (SSC). This approach is appropriate for the target dengue dataset because this dataset is high-dimensional with significant amounts of information and some irrelevant attributes. Besides, this dataset can be reduced without knowing the size of the separated dimensions.

In the SSC algorithm, the number of dimensions is reduced by dividing sub-

spaces into two segments. This algorithm includes two steps. In the first step, a sparse optimization program is used to find numerous other points that belong to the same subspace S_l. In the second step, spectral clustering is applied to a similar graph to realize data segmentation.

Here, the dimension reduction operation finds the cut line between a pair of clusters, where $\{S_l\}_{l=1}^{n}$ is a set of subspaces of $\mathbb{R}^D$ of dimension $\{d_l\}_{l=1}^{n}$ and data points $\{y_i\}i - 1^N$ in the union of n subspaces. The data points are defined by matrix Y in two ways:

$$Y = [y_1, y_2, ..., y_N] \tag{6.1}$$

$$Y = [Y_1, Y_2, ..., Y_n]\Gamma \tag{6.2}$$

Here, Y_l is a collection of Ni data points, which is a representation of subspace S_l, and Γ is an unknown permutation matrix used to arrange the columns of the data points of Y_l belonging to subspace S_l. Let C = $\Gamma^{-1}[c_1^T, c_2^T, c_n^T]^T$ Assume y is a linear combination of Ni data points in the same subspace. From 6.1 and 6.2, we define data point y_i as follows:

$$y_i = Y_i c_i, \qquad e_{ij} = 0 \tag{6.3}$$

where

- $c_i \triangleq [e_{i1}\ , e_{i2},\ ...\ e_{iN}]$ are linear values.
- $e_{ij} = 0$ is a constraint.

However, the representation of y_i is not unique in the subspaces. The SSC algorithm utilizes ci to recognize a data point in the subspace; thus, finding the set of appropriate subspaces is an optimization problem that can be express as follows:

$$\min \|c_i\|_q \quad s.t. \quad y_i = Yc_i, \quad e_{ij} = 0 \tag{6.4}$$

where $\|c_i\|_q$ is l_q-norm of $|c_i| \in \mathbb{R}^N$

- $\|c_i\| > 0$: data point lying in subspace S_l

- $\|c_i\| = 0$: data point does not lying in subspace S_l

After solving the optimization problem, a weighted graph G = (V, E, W) is utilized to organize the sparse points, where V stands for the collection of N nodes of the graph corresponding to N data points, the set of edges connecting them denoted by E, W is asymmetric non-negative and similarity matrix, representing the following edge weights.

$$W = \begin{bmatrix} W_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_n \end{bmatrix} \Gamma \tag{6.5}$$

In the second step, the SSC algorithm applies spectral clustering to matrix W to segment the data.

6.2.2 Finding hidden features

As mentioned previously, the proposed HSC model reduces data dimensionality and finds meaningful hidden features in the dataset. Thus, the proposed model promisingly outperforms existing methods because doctors are not required to investigate all symptoms to make a diagnosis. Instead, they can make a precise diagnosis based on only several crucial symptoms. From the statistical point of view, the traditional methods can be expressed as follows.

$$Y = f(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n \tag{6.6}$$

A patient is determined to have dengue based on the value of Y.

- Dengue (1) if Y > threshold.
- No dengue (0) if Y < threshold.

To improve the accuracy of diagnosis, the proposed model combines detection of a small group of patients with diseases with traditional classification methods. This group will be identified by finding a pair of appropriate clusters in the subspaces. For example, in a given subspace, the data points are divided into two clusters, and these clusters are represented by a binary vector, which includes

value 1 for a cluster of patients with disease and value 0 for a cluster of undefined patients. By comparing to the training data, this subspace can be considered as a meaningful hidden feature h_i if the corresponding clusters of patients are high.

In order to make the diagnosis, the processing described in Figure 6.2.

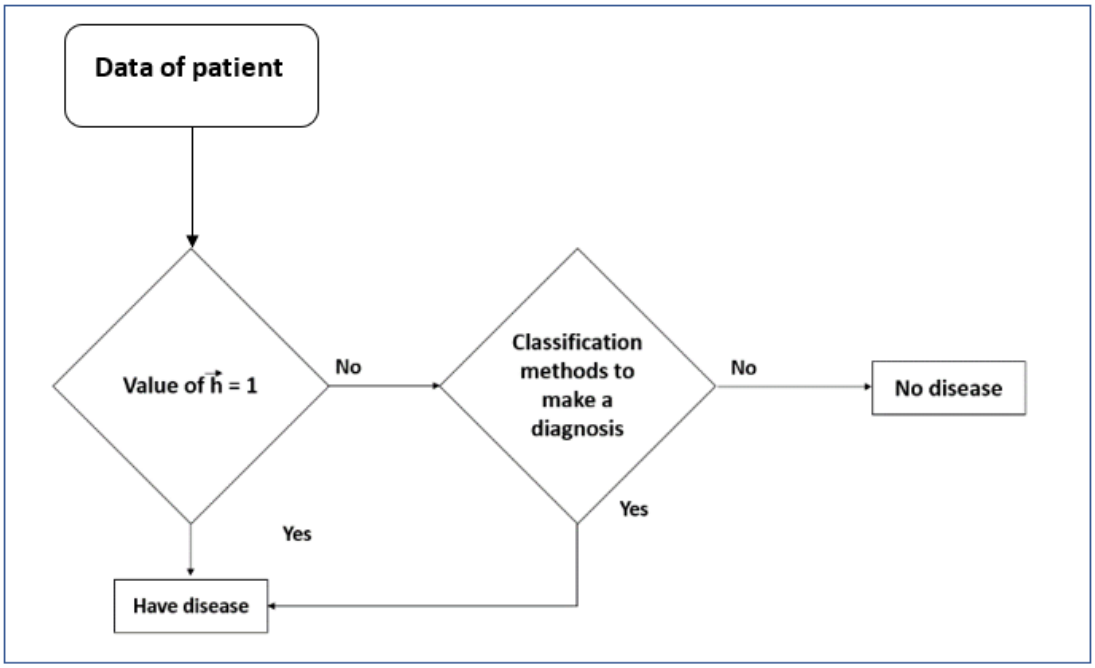

Figure 6.2: The processing of making a diagnosis

From 6.6, the patient will be distinguished as having diseases as follows:

$$y_i = g(f(x)) = f(x_i) + h_i(threshold - f(x_i)) \tag{6.7}$$

where

- y_i = threshold(disease) if $h_i = 1$
- y_i = f(x_i) if $h_i = 0$

Here, y_i is a data point of the disease data $\{y_i\}_{i=1}^{N}$, h_i is the i^{th} element of a vector $\vec{h}$, and f(x_i) is a function to predict disease. The advantage of this approach is that it can clearly identify dengue patients represented by $h_i = 1$, thereby improving the performance of diagnosis.

To find vector $\vec{h}$, we propose a two-step solution. The first step is to find new areas that are a projection of data points with k dimension less than the original data. In the second step, the data points in the new space are clustered. Here, the primary goal is to collect clusters and evaluate the correlation between clusters and the result class in order to find hidden useful dimensions. To get a new subspace, we employ the random projection (RP) algorithm [101] which is a practical and effective method to project data points to another area. The RP algorithm shows that the n × d original data matrix (X) of n-dimensional observations is projected by a d×k random matrix ($\mathbb{R}$) (where k×d) to produce a lower dimensional subspace P of n×k.

$$P_{n\times k} = X_{n\times d} R_{d\times k} \tag{6.8}$$

The output of this step is new areas with different dimensions. After finding the new area, the cluster operation is performed. Here, the Gaussian mixture cluster algorithm [102] is employed for clustering because it can plot high-dimensional data and separated clusters. In addition to that, compared to the other clustering algorithms such as K-Means, the advantages of this algorithm is that accounts for the covariance of clusters. Instead of only considering the distribution of points is circular, GMM can handle well with all shapes of distribution of points. Thus, GMM can distinguish clusters better. This algorithm begins with random Gaussian parameters θ which consists of a centroid μ, a covariance σ, and a mixing probability π to define clusters. The points are fitted to the clusters by the following equation:

$$p(X) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(x_n|\mu, \sigma_k) \tag{6.9}$$

where

- X is data points
- p is the probability
- x_n belong to Gaussian k
- μ defines the center of Gaussian k
- σ_k is a covariance that defines the width of Gaussian k

- π_k is mixing probability that defines how big the Gaussian k is.

Based on the expectation maximization algorithm, Gaussian mixture model optimizes the parameters as the following:

$$\theta^* = argmax_\theta(p(X, \theta)) \tag{6.10}$$

The results obtained by Gaussian mixture model are a pair of clusters which can be the hidden features in the dataset. From that, the hidden features can be selected based on the correlation between the values of vector and the values of the result class in the training set. This correlation is also the accuracy of detecting disease based on a vector which is higher than the accuracy of classification methods. Therefore, it improves the performance of diagnosis.

6.2.3 Application to classification methods

After finding the hidden features by GMM, a dataset which is the combination of data and hidden features is given as the input to a classification algorithm, such as LR, SVMs, and RFs, to identify an optimal prediction algorithm. The difference is the quality of original dataset and dataset with hidden features. By finding a group of dengue patients with a high degree of accuracy, the proposed model helps improve cases that are not found by conventional dengue classification models. Specifically, the number of patients can be calculated as follows.

In traditional method:

$$\begin{aligned} I &= n \times a \qquad s.t \qquad a < 1 \\ &= m \times a + (n - m) \times a \end{aligned} \tag{6.11}$$

In HSC model:

$$I = m + (n - m) \times a \tag{6.12}$$

where

- I: the number of dengue cases can be detected

- n: the total of records.
- m: the dengue cases detected based on hidden features
- a: accuracy of classification algorithm

Due to a < 1, the accuracy of classification in HSC model is better than the traditional methods.

In addition, the proposed model utilizes cross-validation techniques to restrict classification overfitting. All data were divided into K subsets and the learning process of the machine has K times. Each time, one subset is used for testing and the other $K - 1$ is used for training.

6.3 Experimental results

We applied the proposed HSC model to a dataset [61] that includes 5726 children with the following criteria.

- Fever less than 72 hours from onset.
- Attending physician identifies dengue as a possible diagnosis.

After preprocessing the data, dimension reduction was initially performed by the SSC algorithm. Here, the input data are the clinical symptoms on the day of admission to hospital, such as temperature, vomiting, etc. The SSC algorithm was used to obtain a new dataset, i.e., a subspace of the original data. Table 6.2 shows that the SSC algorithm divided the attributes of the original dataset into two segments (SSC dataset) in which Segment 2 is removed. To prove that SSC did not change the classification accuracy of classification methods, we evaluate both datasets by logistic regression algorithm.

The results of the two tests were equivalent. This means that the dimensions of the original dataset were reduced successfully which leads to the reduction of the computational time. Besides, Table 6.2 also demonstrates that SSC did not change the classification accuracy between the original and SSC datasets.

After reducing the dimensions, we attempted to find the hidden features in the SSC dataset. The proposed algorithm selected the hidden features as data points

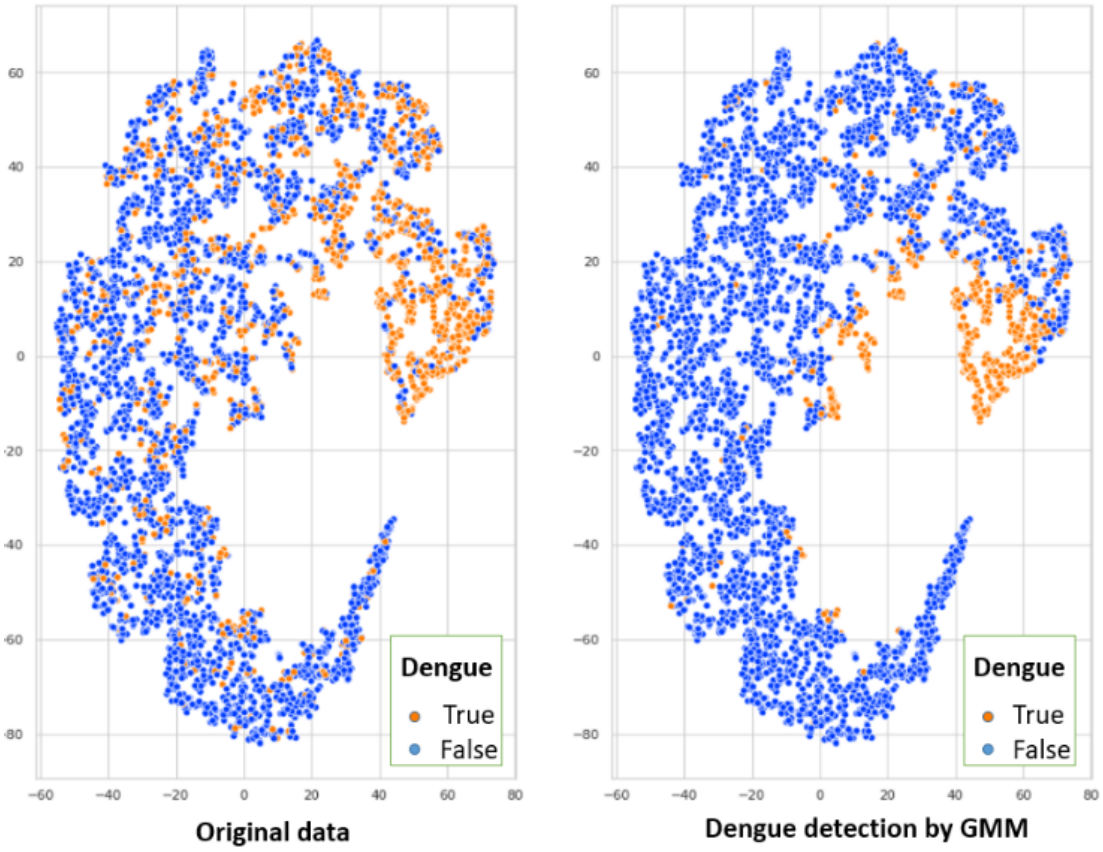

Figure 6.3: The projection of data points.

Table 6.2: The result of SSC algorithm

	Original dataset (%)	Apply SSC to the dataset
Attributes	Day disease, Age, Sex, BMI, Temp, Vomiting, Skin bleeding, Mucosal bleeding, Abdominal, Rash, Flush, Injection, WBC, HCT, Platelet, ALB, AST, CK.	- Segmentation 1: Day disease, Age, Sex, BMI, Temp, Vomiting, Skin bleeding, Mucosal bleeding, Abdominal, WBC, HCT, Platelet, ALB, AST, CK. - Segmentation 2: Rash, Flush, Injection.
Accuracy / Sensitivity of classification (Logistic regression)	80.07% / 73.01%	80.03% / 72.94%

which lying in n dimensions of the subspace, where n < d. Based on the distance and separation of clusters, especially the correlation between this projection of data and the result class in training data, we selected the appropriate features, as shown in Figure 6.3. To select the best method for evaluation of clusters, we conducted experiments using K-mean, Support Vector Machine and mixture Gaussian cluster algorithms. Besides, we use t-SNE [103] which is a visualization method to perform the projection of data to the two dimensions space.

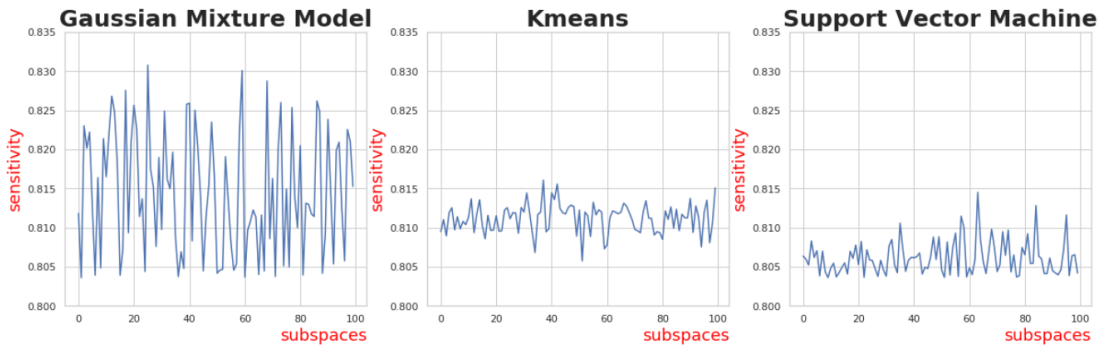

Figure 6.4: The sensitivity of dengue classification in subspaces.

In order to compare the efficiency between K-Means and GMM, we applied classifications algorithms to the original dataset with the hidden features found by each algorithm. Three tests were applied with Logistic regression, Support Vector Machine and Random Forest to evaluate based on the accuracy and sensitivity. These are the crucial values to indicate the proposed model's ability to classify dengue patients and can be calculated as follows:

Table 6.3: The values in classification disease.

		True condition	
		Dengue	No dengue
Prediction	Dengue prediction	True Positive(TP)	False Positive(FP)
	No dengue prediction	False Negative(FN)	True Negative(TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.13}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6.14}$$

- TP: the number of correctly predicted dengue patients
- TN: the number of correctly predicted patients with no dengue
- FP: the number of incorrectly predicted remaining dengue patients
- FN: the number of remaining dengue patients

Table 6.4: The result of applying the HSC model.

	Original dataset + hidden feature found by K-Means		Original dataset + hidden feature found by GMM	
	Accuracy	Sensitivity	Accuracy	Sensitivity
Logistic Regression	81.54%	75.4%	**81.93%**	**75.9%**
SVM	81.52%	78%	**81.9%**	**78.5%**
Random Forest	80.98%	73.4%	**81.58%**	**73.6%**

6.4 Discussion

For dengue fever, early detection plays a very important role in the treatment of patients. In this chapter, a proposed model was presented for solving a problem of shortage of well-train doctors in developing countries, which has been an obstacle with early dengue detection. Based on the process of making a diagnosis from doctors, the proposed model was designed with the following purposes.

The primary purpose of the proposed model is to find underlying features in the subspaces. This is understood as the identification of clusters corresponding to the referenced result class in new areas, then improve the efficiency of classification algorithms. As demonstrated in Equation 6.11 and Equation 6.12, the number of patients could be found by the HSC model is more accurate because this model successfully detected the true patients. In order to find the hidden features, there is many algorithms could be a solution such as K-Means, GMM and Support Vector Machine. Thus, these algorithms were applied in our experiment to find the best method. The sensitivity is the probability of detection of dengue patients which is used as the measure for evaluation of these algorithms. This is because we want to guarantee that patients having dengue are identified. As observed from Figure 6.4, GMM achieved the highest sensitivity with 0.831. These values were higher comparing to the sensitivity of classification algorithms in original data, thereby the patients can be found more accurately. Besides, most of the sensitivity by GMM is better than the highest one of the remain algorithms. The reason is that GMM used a covariance to handle the oblong cluster instead of only the circle as K-Mean, while SVM usually attempts to separate to the two equal parts, which is not suitable because the data is an imbalance. Therefore,

GMM is the most suitable solution for detecting the meaningful hidden features of dengue disease. To observe the efficiency of GMM, t-SNE is used to show data points to the two dimensions area. In Figure 6.4, the left visualization illustrates the original data and the right one shows the clusters of patients by GMM in subspace. In the figure, the orange points represent the dengue patients and the blue points represent the patients with no dengue. From observation, GMM seems to be able to successfully cluster a small group of dengue patients. This means GMM could be a good solution for finding correct dengue patients. Besides, the experiments indicate that SVM was better than other algorithms (78.5%) with the highest sensitivity and Logistic regression was the best option in terms of accuracy (81.93%). Finally, the comparison between K-Means and GMM indicates that using better hidden features make a better result in classification dengue. In Table 6.4, both the accuracy and sensitivity of the model with hidden features found by GMM were better than the other found by K-Means in all classification algorithms.

The second purpose of the proposed model is to reduce data dimensions in order to decrease computational time without affecting classification accuracy. The experimental results indicate that the proposed model successfully reduced the dimensions of the data. The attributes were divided into two separate clusters. By using only important attributes which were identified by the SSC algorithm, the accuracy of classification dengue patients was almost unaffected. This is shown in Table 6.2, the computation on original data with all 19 attributes have an accuracy of 80.07%, and the computation on the data using only Segment 1 is only slightly lower with an accuracy of 80.03%.

In terms of the result of dengue classification, the proposed model is effective at finding the percentage of dengue patients. Specifically, accuracy and sensitivity are the used criteria to evaluate the performance of models where accuracy is the difference between true value and result, while sensitivity measures the proportion of actual positives that are correctly identified. In the first aspect, the accuracy of disease classification algorithms usually ranges from 70% to 90% [104], and the proposed model got 81.93% in terms of accuracy by the Logistic regression algorithm, which can be acceptable. Comparing to doctors, this result was lower because the doctors can utilize more attributes such as having an NS1 antigen or not, thus, patients with no dengue (False Positive) were found better than the proposed model. Especially, with the high specificity of the NS1 test,

doctors could detect 99.2% patients with no dengue [61]. Due to the imbalance of the dengue dataset (1098:4028), the accuracy from doctors usually better than machines. However, in our other test using the NS1 value as input, the proposed model's accuracy was 93%, which was higher than that of the doctors. The reason why we omitted NS1 is because making this test is difficult and it needs to be made by doctors while the purpose of proposed system is to assist patients with making a diagnosis without the support from doctor. Besides, patients having an NS1 antigen might not have dengue fever but another disease. Because of that, the second aspect is the sensitivity that also considered as the criteria in dengue examination. Based on observing Equation 6.14, we can see that the purpose of sensitivity calculation is to increase the rate of accurate finding patients with dengue by decreasing the number of False Negatives values in the Table 6.3. In terms of this aspect, the best sensitivity of the proposed model (**78.5%**) is higher than that of medical tests, such as NS1 and NS1 Ag strip (70.4%) [61]. As mentioned above, dengue fever is a serious infection, which needs to early detected to increase the recoverability of patients. The higher sensitivity of the proposed model supports to find more patients earlier. Consequently, the results of this study contribute to the medical diagnosis of dengue fever.

As mentioned above, the accuracy of the proposed model is lower than the doctors. Based on observing Equation 6.13, the accuracy is dependent not only on the False Negatives (FN) value but also on the False positive (FP) value. Therefore, in order to increase accuracy, we must also decrease the number of False Positive cases in Table 6.3. With the hypothesis of making a diagnosis for patients with no dengue based on a few symptoms, I intend to find more hidden features which have the ability to indicate a small group of patients with no dengue in the future. Hence, the False Positive cases are decreased and the accuracy of the proposed model is also better.

As a result, the proposed model can help doctors make accurate diagnosis. Besides, the simple implementation of the proposed model will solve the lack of well-train doctors in developing countries. Therefore, dengue patients can be early detected, then reduce dengue fatality.

6.5 Summary

In this chapter, we have proposed a model to reduce data dimensions and improve the accuracy of classification methods. To achieve this, firstly the proposed model uses the SSC algorithm to reduce data dimensionality. Consequently, the Random Projection algorithm finds a new projection from subspaces, and the Gaussian Mixture algorithm clusters data points in new areas. We then find and evaluate hidden features based on the distribution among the points in the clusters and the corresponding referenced independent variable. Finally, classification algorithms, such as Logistic Regression, SVM, and Random Forest, are applied to optimize the results.

However, the efficiency of the proposed HSC model with this dataset is not as high as expected due to the limitations of the numerous identifications of actual patients. As mentioned in the discussion section, there is a gap in accuracy between doctors and the proposed model, therefore, the model needs to be improved for increasing accuracy. In order to do that, the proposed model requires to optimize not only the sensitivity but also the specificity of dengue classification. In the future, the identifications of actual patients need to be investigated further to improve the sensitivity of the proposed model. Besides, hidden vectors representing patients with no dengue is also required to be detected to improve the specificity of dengue classification.

Chapter 7

Discussion

This chapter discusses the works investigated and solutions proposed in this dissertation by which advantages as well as the remaining issues will be summarized.

As mentioned in this dissertation, with the serious impact of various diseases and the huge demand for seeking health information of humans, it is essential to have a proper solution in order to improve the quality of current health care services. In recent years, virtual agents in medical topics have attracted large attention from academic communities, resulting in numerous publications. These studies mostly focus on the natural language understanding of machine and the disease classification algorithms. However, besides the advantages, several issues remain waiting to be solved. In this chapter, I will summarize what I have done in this dissertation, including the core findings of each study. Next, I will discuss them and introduce possible ways to improve the studies.

7.1 Findings of this dissertation

Chapter 4: Aiming to make a virtual agent in medical topics, the first problem I focused on is the interface that allows human communication with machines. Due to the different languages, it is necessary to have a method to transform the human language into the computer understandable language. In Chapter 4, I introduced a model to support machines for detecting human intent. By calculating the semantic similarity between sentences, the machine can determine human intent based on predefined sentences. To do this, I first explored the existing methods including 2 approaches: word embedding and sentence embedding.

Considering both approaches, the sentence embedding-based approach was my selection because it can cover the ordering of words. One method of sentence embedding-based approach is SkipThought, which attracted me with the idea of sentence embedding based on the unstructured data. Comparing with other methods, SkipThought has the advantage of scalability due to the explosion of data on the Internet. In order to improve the quality of sentence embedding, I proposed a model which outperformed SkipThought thanks to the ability to cover not only the meaning but also the context of sentences. To prove the proposed model, I conducted experiments of both models in the same setting. The experiment results showed that the proposed model was better not only for the prediction of the next sentence but also for the calculation of semantic similarity between sentences in specific scenario.

Chapter 5: After understanding the human intent, another important criterion in information collection processing is finding the necessary symptoms based on the identified symptom. To achieve this, the equipment of medical knowledge for a virtual agent is needed. The reason is that the provided information from patients is usually not enough for making a diagnosis. In literature, the diseases and symptoms are controlled by a Relational Database Management System (RDBMS) or a semantic database such as Ontology. In order to exploit the knowledge in these database systems, a script-based method often utilized for discovering the relationship between entities. However, the script-based approach has drawbacks due to scalability when applying into a huge database. In that situation, the ontology-based approach provides a better way to predict the subsequent symptoms by a method based on the calculation of weights between disease and symptoms. In this research, an automatic self-learn medical knowledge method is introduced for the prediction of subsequent symptoms from the previous input symptoms. Through the experiment, this method has demonstrated that it can support machines in the criteria of symptoms suggestion.

Chapter 6: The final task in a medical examination is a message that the patient is positive or negative. A health decision support system has an extremely important role in improving the quality of health care, not only for patients but also for doctors. To patients, they are provided a convenient method for checking their health. To doctors, they received the reference information, which supports them in making a diagnosis. Especially, dengue fever is a serious infectious disease, which annually occurs in many countries. In the health decision support

system, accuracy is one of the most important criteria because this affects the treatment of patients. In order to improve the accuracy of disease classification, especially dengue fever, a novel model based on finding meaningful hidden features was introduced in this research. Similar to doctors' examination, this model aims to discover the relationship of a group of symptoms that could identify a true patient. Experimental results showed that hidden features improved the accuracy of disease classification.

From the result analysis of the whole study, essential findings in this dissertation had been found based on the facts of the study. They were listed as below.

- Attention mechanism could calculate the impact of previous sentences on the current sentence.
- The size of the embedding vector should be 512 for representing a sentence.
- Covering context of sentence improves the quality of sentence embedding.
- Using more information from history, which often perform better and stable.
- The training of a sentence embedding-based approach usually take a long time but the early stop can be applied.
- Weights between diseases and symptoms can support the efficiency of symptom suggestion.
- Ontology-based symptom suggestion system has a good scalability.
- The difference between ontology-terms and the daily words is an important issue, which affects the performance of predicting candidate diseases.
- Cosine similarity should be a measurement for calculating the weight between diseases and symptoms.
- Hidden features can improve the accuracy of disease classification.
- The most suitable method to find hidden features is GMM.
- Using SSC algorithm can reduce the dimensions of data without the affection of accuracy.

- The sensitivity of the proposed model is higher than the medical test, which allows finding more patients.

7.2 Discussion and limitations

In the dissertation, I proposed a medical agent works on the English. However, the proposals were designed to work with the unstructured dataset, therefore, implementations of medical agents with other languages such as French or Vietnamese are not too difficult; we only need to pay the attention to the preprocessing for differences between corpus languages. For example, in order to make a pre-trained model for calculating semantic similarity between sentences, we have to prepare the corpus with vocabularies. While English utilizes only a word, Vietnamese sometimes requires two words to describe the meaning. Therefore, we need to concentrate words into a new word in the vocabulary before making a pre-trained model. The medical agent is important for developing countries where there is a shortage of doctors and the overloading in the hospitals. With their outstanding advantages, the medical agents will assist in increasing the quality of healthcare services for humans. Especially, the dengue disease is a serious infection, which usually occurs in tropical countries annually and affects significantly to the human-life. As a result, the dissertation concentrated on the medical agent for dengue decision support.

This dissertation aimed to address many problems that can be solved by each system. The problem of the data collection, which was the core problem, could be addressed by my main system. The critical point needed to be discussed here was how the whole system can be automated to collect the symptoms from patients. The main idea of this research was to propose methods intent detection by calculating the semantic similarity between sentences as well as utilize ontology for symptoms suggestion. The model following the sentence embedding-based approach was introduced for supporting machines to learn the distribution of representation sentences. Besides, a mechanism of symptoms suggestion was also introduced for collecting the necessary symptoms which unprovided by patients.

Unfortunately, the limitation on training data and computational cost caused the suboptimal results of implementation. Specifically, the used training data includes only the 13118 daily conversations, which is smaller than an unstructured

used data by SkipThought with around 74 million sentences. Furthermore, a lack of medical knowledge was an obstacle in discovering the relationship between entities in medical topics. The experiment results was received specifically for dengue fever; future works need to verify other diseases as well.

Chapter 8

Conclusion and Future Work

This chapter concludes the dissertation and figured out directions for future work

8.1 Conclusion

As analyzed in the Chapter 1, the symptoms collection processing and the disease classification processing must be performed by a virtual agent in the medical topic. This has become the motivation for numerous contemporary studies for years. In this research, an improvement in the quality of health care service has been achieved by solving three major issues associated with conversation-based data collection and modeling of diseases. The conclusions of this dissertation are summarized as follows:

Chapter 4 focuses on proposing a model to support machines to learn the distribution of sentences representation for communication with humans. In this chapter, the impact of previous sentences to the current sentence was clarified. Through a calculation of semantic similarity between input and pre-defined sentences, the intent from input of humans could be identified. To prove the efficiency of the proposal model, an experiment of calculation relatedness sentences with the previous model and the proposed model is presented. The experimental result demonstrated that a context-based sentence embedding approach can improve the accuracy in computation semantic similarity between sentences.

Chapter 5 aimed at proposing a collaborative approach using an ontology and a neural network to suggest the necessary symptoms in medical examina-

tion. More concretely, an approach to construct a corpus of medical knowledge is introduced. Furthermore, the relations of entities in this corpus were performed by this model for continuous data. The evaluation results demonstrate that this approach can predict a subsequent symptom based on the input symptom with a percentage of accuracy is favorable. Besides, with the advantage of scalability, it is easy to expand the model on other huge datasets.

Chapter 6 focused on proposing a model to improve the accuracy of classification algorithm in disease dataset. Basically, a disease classification is the process of analyzing the probability of attributes in the dataset. However, the real-life data usually includes noise, and the redundant attribute, which causes a decrease of accuracy in classification algorithms. Therefore, in this chapter, a subspaces-approach is proposed for finding the hidden attributes and removing the redundant attributes. This approach initially starts with preprocessing the data, then move to a reduction of redundant attributes, and finally improve the accuracy of classification by finding meaningful attributes in the projections of data. As the result, by applying the proposed approach, the accuracy of disease classification is increased, while the computation cost is reduced.

In summary, I have proposed a framework in this dissertation with three methods to realize a virtual medical agent that supports patients to make a diagnosis of dengue fever. The virtual medical agent support patients to make a diagnosis throughout the Internet. Therefore, the problem of lacking well-trained doctors and the overload of service providers was solved. Hence, human health has cared better.

8.2 Future Work

Throughout this study, the author continues to develop a virtual agent for disease decision support. Although the proposals have solved several issues of a virtual agent in the medical topic, the performance of each issue needs to be clarified and improved. Some recommendations will be briefly presented as follows:

1. It is necessary for all proposals to be expanded on the huge dataset with many complex components.

2. A public evaluation on the Internet is required for getting feedbacks from users in order to evaluate the performance of disease decision support of virtual agents.

3. More importantly, the interface of communication between humans and machines is a criterion for the evaluation of the intelligence of a virtual agent. This is required for the improvement of the model towards human-like behaviors.

Bibliography

[1] National Cancer Institute. Health information national trends survey (b5a). https://hints.cancer.gov/view-questions-topics/question-details.aspx?red=1&qid=757, 2017. Accessed: 2017.

[2] Yun-Nung Chen, Asli Celikyilmaz, and Dilek Hakkani-Tur. Deep learning for dialogue systems. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 25–31, 2018.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] World Health Organization et al. Dengue and severe dengue. Technical report, World Health Organization. Regional Office for the Eastern Mediterranean, 2014.

[5] Nancy Atkinson, Sandra Saperstein, and John Pleis. Using the internet for health-related activities: findings from a national probability sample. *Journal of medical Internet research*, 11(1):e5, 2009.

[6] National Cancer Institute. Health information national trends survey (b4). https://www.who.int/csr/sars/country/2003_08_15/en/, 2012. Accessed: 2014-02-20.

[7] Worldometers. Coronavirus cases statistical:. https://www.worldometers.info/coronavirus/, 2020. Accessed: 2020.

[8] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th*

International Conference on Information and Education Technology, pages 111–119, 2019.

[9] Paul Schachter and Timothy Shopen. Parts-of-speech systems. *Language typology and syntactic description*, 1:3–61, 1985.

[10] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[13] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[14] Amiya Kumar Tripathy, Rebeck Carvalho, Ajit Puthenputhussery, Nikita Chhabhaiya, and Bijoy Anthony. Mediassistedgesimplifying diagnosis procedure & improving patient doctor connectivity. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–6. IEEE, 2015.

[15] Shafquat Hussain and Ginige Athula. Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 698–703. IEEE, 2018.

[16] Rashmi Dharwadkar and Neeta A Deshpande. A medical chatbot. *International Journal of Computer Trends and Technology (IJCTT)*, 60(1):41–45, 2018.

[17] Hameedullah Kazi, BS Chowdhry, and Zeesha Memon. Medchatbot: An umls based chatbot for medical students. *International Journal of Computer Applications*, 55(17), 2012.

[18] Surya Roca, Jorge Sancho, José García, and Alvaro Alesanco. Microservice chatbot architecture for chronic patient support. *Journal of Biomedical Informatics*, 102:103305, 2020.

[19] Richard Wallace. The elements of aiml style. *Alice AI Foundation*, 139, 2003.

[20] Maria das Graças Bruno Marietto, Rafael Varago de Aguiar, Gislene de Oliveira Barbosa, Wagner Tanaka Botelho, Edson Pimentel, Robson dos Santos França, and Vera Lúcia da Silva. Artificial intelligence markup language: a brief tutorial. *arXiv preprint arXiv:1307.3091*, 2013.

[21] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[22] RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. A graph based chatbot for cancer patients. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 717–721. IEEE, 2019.

[23] Nudtaporn Rosruen and Taweesak Samanchuen. Chatbot utilization for medical consultant system. In *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pages 1–5. IEEE, 2018.

[24] Soufyane Ayanouz, Boudhir Anouar Abdelhakim, and Mohammed Benhmed. A smart chatbot architecture based nlp and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, pages 1–6, 2020.

[25] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. Mandy: Towards a smart primary care chatbot application. In *International Symposium on Knowledge and Systems Sciences*, pages 38–52. Springer, 2017.

[26] S Divya, V Indumathi, S Ishwarya, M Priyasankari, and S Kalpana Devi. A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing*, 3(1):1–7, 2018.

[27] Qiming Bao, Lin Ni, and Jiamou Liu. Hhh: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–10, 2020.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[30] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[31] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

[32] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

[33] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[34] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[35] Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia R de Sa. Rethinking skip-thought: A neighborhood based approach. *arXiv preprint arXiv:1706.03146*, 2017.

[36] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

[37] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[38] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[40] Edgar F Codd. A relational model of data for large shared data banks. In *Software pioneers*, pages 263–294. Springer, 2002.

[41] Jamie MacLennan, ZhaoHui Tang, and Bogdan Crivat. *Data mining with Microsoft SQL server 2008*. John Wiley & Sons, 2011.

[42] Paul DuBois. *MySQL*. Addison-Wesley Professional, 5th edition, 2013.

[43] Richard H Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. *Summit on translational bioinformatics*, 2009:116, 2009.

[44] Grigoris Antoniou and Frank Van Harmelen. *A semantic web primer*. MIT press, 2004.

[45] Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962, 2019.

[46] Osama Mohammed, Rachid Benlamri, and Simon Fong. Building a diseases symptoms ontology for medical diagnosis: an integrative approach. In *The First International Conference on Future Generation Communication Technologies*, pages 104–108. IEEE, 2012.

[47] Linda Mhadhbi and Jalel Akaichi. Ds-ontology: A disease-symptom ontology for general diagnosis enhancement. In *Proceedings of the 2017 International Conference on Information System and Data Mining*, pages 99–102, 2017.

[48] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[49] Ahmed Fawzi Otoom, Emad E Abdallah, Yousef Kilani, Ahmed Kefaye, and Mohammad Ashour. Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, 9(1):143–156, 2015.

[50] K Vembandasamy, R Sasipriya, and E Deepa. Heart diseases detection using naive bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2(9):441–444, 2015.

[51] Vikas Chaurasia and Saurabh Pal. Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol*, 2:56–66, 2014.

[52] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[53] G Parthiban and SK Srivatsa. Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, 3(7), 2012.

[54] Aiswarya Iyer, S Jeyalatha, and Ronak Sumbaly. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*, 2015.

[55] Sanjay Kumar Sen and Sujata Dash. Application of meta learning algorithms for the prediction of diabetes disease. *International Journal of Advance Research in Computer Science and Management Studies*, 2(12), 2014.

[56] V Anuja Kumari and R Chitra. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2):1797–1801, 2013.

[57] S Vijayarani and S Dhayanand. Liver disease prediction using svm and naïve bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, 4(4):816–820, 2015.

[58] Fadl Mutaher Ba-Alwi and Houzifa M Hintaya. Comparative study for analysis the prognostic in hepatitis data: data mining approach. *spinal cord*, 11:12, 2013.

[59] Ian H Witten, Eibe Frank, and Mark A Hall. Practical machine learning tools and techniques. *Morgan Kaufmann*, page 578, 2005.

[60] Bekir Karlik. Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers. *IBU Journal of Science and Technology*, 1(1), 2012.

[61] Nguyen Minh Tuan, Ho Thi Nhan, Nguyen Van Vinh Chau, Nguyen Thanh Hung, Ha Manh Tuan, Ta Van Tram, Nguyen Le Da Ha, Phan Loi, Han Khoi Quang, Duong Thi Hue Kien, et al. Sensitivity and specificity of a novel classifier for the early diagnosis of dengue. *PLoS neglected tropical diseases*, 9(4), 2015.

[62] Lukas Tanner, Mark Schreiber, Jenny GH Low, Adrian Ong, Thomas Tolfvenstam, Yee Ling Lai, Lee Ching Ng, Yee Sin Leo, Le Thi Puong, Subhash G Vasudevan, et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*, 2(3), 2008.

[63] A Fathima and D Manimegalai. Predictive analysis for the arbovirus-dengue using svm classification. *International Journal of Engineering and Technology*, 2(3):521–7, 2012.

[64] Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

[65] Nd Ahmad Tarmizi, Farha Jamaluddin, An Abu Bakar, Zulaiha Ali Othman, Suhaila Zainudin, and Abdul Razak Hamdan. Malaysia dengue outbreak detection using data mining models. *Journal of Next Generation Information Technology (JNIT)*, 4(6):96–107, 2013.

[66] Ana Lisa V Gomes, Lawrence JK Wee, Asif M Khan, Laura HVG Gil, Ernesto TA Marques Jr, Carlos E Calzavara-Silva, and Tin Wee Tan. Classification of dengue fever patients based on gene expression data using support vector machines. *PloS one*, 5(6), 2010.

[67] Joseph Weizenbaum. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[68] Richard S Wallace. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer, 2009.

[69] Bayan Abu Shawar and Eric Atwell. Different measurement metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 89–96, 2007.

[70] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[71] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

[72] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[73] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[74] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[75] Thomas R Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–221, 1993.

[76] Hong Zhong and Li-Min Xia. Ontology-based image retrieval. *Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications)*, 42(17):37–40, 2007.

[77] Sarunya Kanjanawattana. *A method of graph information extraction and retrieval for academic literatures by use of semantic relationships*. PhD thesis, SHIBAURA INSTITUTE OF TECHNOLOGY, 2017.

[78] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[79] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[80] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[81] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[82] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[83] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[85] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[86] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, 2014.

[87] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[88] KENTSTATE University. Spss tutorials: Pearson correlation. `https://libguides.library.kent.edu/SPSS/PearsonCorr`, 2017. Retrieved : 14 May 2017.

[89] Christian S Perone, Roberto Silveira, and Thomas S Paula. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*, 2018.

[90] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[91] Konstantinos I Diamantaras and Sun Yuan Kung. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc., 1996.

[92] Bernhard Flury. *Common principal components & related multivariate models*. John Wiley & Sons, Inc., 1988.

[93] Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, et al. The global distribution and burden of dengue. *Nature*, 496(7446):504–507, 2013.

[94] Maria G Guzman and Eva Harris. Dengue. *The Lancet*, 385(9966):453–465, 2015.

[95] Duane J Gubler. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends in microbiology*, 10(2):100–103, 2002.

[96] Thomas P Monath. Dengue: the risk to developed and developing countries. *Proceedings of the National Academy of Sciences*, 91(7):2395–2400, 1994.

[97] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[98] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004.

[99] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[100] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[101] Xiaoli Z Fern and Carla E Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 186–193, 2003.

[102] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009.

[103] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[104] Sherwin Fernandes, Rutvij Gawas, Preston Alvares, Macklon Femandes, Deepmala Kale, and Shailendra Aswale. Survey on various conversational systems. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–8. IEEE, 2020.