MANDY: Towards a Smart Primary Care Chatbot Application

Lin Ni(✉), Chenhao Lu, Niu Liu, and Jiamou Liu

Department of Computer Science, The University of Auckland,
Auckland, New Zealand
lni600@aucklanduni.ac.nz, jiamou.liu@auckland.ac.nz

Abstract. The paper reports on a proof-of-concept of Mandy, a primary care chatbot system created to assist healthcare staffs by automating the patient intake process. The chatbot interacts with a patient by carrying out an interview, understanding their chief complaints in natural language, and submitting reports to the doctors for further analysis. The system provides a mobile-app front end for the patients, a diagnostic unit, and a doctor's interface for accessing patient records. The diagnostic unit consists of three main modules: An analysis engine for understanding patients symptom descriptions, a symptom-to-cause mapper for reasoning about potential causes, and a question generator for deriving further interview questions. The system combines data-driven natural language processing capability with knowledge-driven diagnostic capability. We evaluate our proof-of-concept on benchmark case studies and compare the system with existing medical chatbots.

Keywords: Medicare chatbot · Patient interview · Natural language processing · AI and healthcare

1 Introduction

Patients arriving at a primary care service sometimes need to wait for a long time before being advised by a doctor [26]. This is often due to high workload and limited resources at the primary care service [7]. To facilitate the process, nurses and other health care staffs usually take the role of patient intake. An incoming patient would be first greeted by a receptionist who carries out an intake inquiry. The receptionist would typically be someone who has a certain level of medical proficiency, and the inquiry involves collecting patient information and understanding the symptoms of the patient. A brief report is generated as outcome of this inquiry to narrow down the causes of the symptoms, so that the doctor may then use minimum effort to perform differential diagnosis [30].

This seemingly robust system still has many shortcomings: Firstly, the medical staffs who carry out patient intake interviews are expected to acquire a good level of medical expertise; this limits the pool of potential candidates and increases the personnel cost. Secondly, at times the staffs need to meet

J. Chen et al. (Eds.): KSS 2017, CCIS 780, pp. 38–52, 2017.
https://doi.org/10.1007/978-981-10-6989-5_4

the demand of a large number of patients and quickly attend to each individual; this increases the risk of losing crucial information in the interview reports. Thirdly, if the intake interview relies on standardized forms or questionnaires, the questions to patients would not be sufficiently personalized to reflect the specific symptoms of individuals, reducing the effectiveness of the interview.

The goal of this paper is to harness the power of artificial intelligence to automate the patient intake process, so that patients receive timely, cost-effective, and personalized healthcare services. To this end, we introduce Mandy, a mobile chatbot who interacts with patients using natural language, understands patient symptoms, performs preliminary differential diagnosis and generates reports.

Despite vast technological advancement, present-day clinics still very much rely on healthcare staff to handle patient intake and carry out initial interviews in a manual way [17]. On the other hand, it is widely viewed that data mining and AI may offer unprecedented opportunities and broad prospects in health [16]. Efforts have been made to deploy humanoid robots (e.g., "Pepper in Belgian hospitals[1]) in hospitals. However, a robot is expensive (e.g. Pepper comes with a price tag of £28000) and would not be able to efficiently cope with a large amount of people. Many industry giants are increasingly investing in AI-enhanced medical diagnosis tools. Notable products include Google DeepMind Health[2], IBM Watson Health[3] and Baidu's Melody[4]. The ambitious goal of these platforms is to allow AI to access and process vast amount of lab test results and genomic data for precision-driven medical diagnosis and predictions.

The novelty of Mandy lies in the fact that it is not directed at precise diagnosis and prediction, but rather, Mandy simply provides a humanized interface to welcome patients and understand their needs, and provide valuable information to physicians for further inquiry. In this way, the system aims to free up the time of healthcare staffs for more meaningful interactions with patients, and help to enable physicians to operate more efficiently.

Mandy is an integrated system that provides a range of functionalities: (1) Mandy provides a patient-end mobile application that pro-actively collects patient narratives of illness and register background information; this may take place at an arbitrary time before the doctor's appointment and at an arbitrary location. (2) Mandy is equipped with natural language processing (NLP) modules that understand patients' lay language, process the patient symptoms, and generate interview questions. (3) Based on interactions during the interview, Mandy will generate a report for the doctor regarding the patient's symptoms and likely causes. (4) Mandy also provides a doctor-end desk-top application for the doctors to check their patients' records and interview reports.

The potential benefits of Mandy are many-fold. Firstly, the system aims to reduce the workload of medical staffs by automating the patient intake process,

[1] http://www.bbc.com/news/technology-36528253.
[2] https://deepmind.com/applied/deepmind-health/, 2017.
[3] https://www.ibm.com/watson/health/, 2017.
[4] http://research.baidu.com/baidus-melody-ai-powered-conversational-bot-doctors-patients/, 2016.

and providing initial reporting to doctors. Secondly, Mandy provides personalized intake service to the patients by understanding their symptom descriptions and generating corresponding questions during the intake interview. Thirdly, by interacting with a chatbot, the patient avoids the need to express his health concerns out loud to people other than the doctor. This also reduces the likelihood of patients not seeking medical help due to shyness or cultural boundaries [28]. Furthermore, many studies have shown that patients tend to be more honest when facing a robot rather than a human health staff [1]. So Mandy is likely to collect truthful information about the patients.

Paper Organization. Section 2 presents related work and identifies insufficiencies with existing AI technology in terms of patient interviews. Section 3 describes system design and core algorithmic modules of Mandy. Section 4 evaluates a proof-of-concept of Mandy by test cases and discusses the results. Finally, Sect. 5 lists some future works which can further improve Mandy.

2 Problem Identification and Related Work

The "overcrowding" issue or long waiting time at emergency units of hospitals and other primary care services has been a world wide challenge [3,10,25]. To cope with the increasing population and an ever increasing demands of patients, a number of countries have implemented targets for reducing waiting time at the healthcare providers, e.g., New Zealand has implemented a "6-hours target" for the waiting time of patients at emergency department since 2009 [15].

Existing patient interview support applications often take the form of expert systems. A common challenge faced by all these applications is the ambiguity and diversity of patient answers. As a result, traditional expert systems usually fail to deliver effective decision support and lacks the flexibility that suits individual needs [14]. An example of AI-driven intake interview assistance system is provided by Warren in [30]. The system sets up complicated rules based on clinical experts' experience and medical knowledge. However, it does not demonstrate capabilities on personalizing the questions to patients and is not able to learn about the individual nature of patients. To apply the system, a clinic needs to provide necessary staffs with sufficient medical background to operate the system. The complicated interface of the system also requires considerable training time, which all adds extra costs to the health provider.

Numerous clinical decision support systems (CDSS) have employed AI technologies in various ways: MYCIN [29] is widely recognized as one of the very first rule-based expert systems that were used for diagnosing infectious diseases. It specialized in bacterial infections and it has been adapted as NEOMYCIN, a teaching and learning platform [8]. Other systems such as INTERNIST-I [22] used a much larger medical knowledge base – obtained from hospital case records – to assist medical personnel in diagnosing internal conditions the patient may have. The system has learning capability on patients' medical history to deliver more accurate results. CDSS technologies have been rapidly developed in the last 10 years. A recent study identified 192 commercially available CDSS

applications in existence [18]. One of the more well-known achievements in this area is from IBM's Watson Health [12]. The system seamlessly combines natural language processing, dynamic learning and hypothesis generation and evaluation to provide useful systems in many key areas such as oncology, genomics, and medical imaging. We remark that most of the CDSS systems are designed to be used by the specialists but not the patients themselves.

Natural language processing has become a prevalent technology and formed an integral part of many IT applications; examples of which include e.g., Siri[5] and Cortana[6]. Chatbot Systems and Spoken Dialogue Systems (SDS) respond with comprehensible sentences and elaborately constructed paragraphs to communicate with the user, which has been adopted in medical domain. The well-known ELIZA [31] was designed to act roughly as psychotherapists. More recently, Florence Bot is a chatbot that reminds patients to take pills regularly[7]. Your.MD[8] and HealthTap[9] are miniature doctors. Studies have verified that SDS could help intervening human habits, to help patients quit smoking [23], or affect their dietary behaviour [9] and physical activity [11]. Others application also used SDS for chronic illness monitor systems, e.g. for hypertensive diabetic [6]. Medical counseling and education is another area which often requires the delivery of SDS [4,5,13]. Among them, Mandy resembles the user experiment of Your.MD the most. Your.MD constructs a Bayesian network with massive medical knowledge to compute the most likely cause of an indisposition. On the other hand, Your.MD has different purpose from Mandy as it is not meant to assist doctors.

3 System Design and Implementation

Figure 1 illustrates the architecture of Mandy. The patient interacts with Mandy through a mobile chatbot. All algorithms are executed and all data are processed in a web services (cloud). This means that all sentences to and from the patients are generated and analyzed in the cloud, respectively. After the intake interview, Mandy scores the patient's record and generate a report regarding the patient's conditions. The doctor can then login into the e-health information management system to access the personalized reports generated for the patient.

Mandy's logic flow simulates a well-established *clinical reasoning process* for differential diagnosis, which consists of a series of well-defined steps [19,24,27]. These steps are guidelines for medical inquiries by a practitioner:

1. *Data acquisition*: Collect patient's history and symptoms, which forms the basis for the initial diagnostic reasoning.
2. *Problem representation*: Summarize the chief complaints of the patient.
3. *Developing differential diagnosis*: Come up with the hypotheses list base on the data acquired.

[5] http://www.imore.com/siri.
[6] https://www.microsoft.com/en/mobile/experiences/cortana/.
[7] Florence Bot, https://florence.chat/.
[8] Your.MD, https://www.your.md/.
[9] HealthTap, https://www.healthtap.com/.

4. *Prioritizing differential diagnosis*: Decide which should be the leading one among the hypotheses list.
5. *Testing hypothesis*: If additional data is required to confirm the hypotheses, order lab tests to take place.
6. *Review and re-prioritize differential diagnosis*: Rule out some diseases and then try to determine the cause of the symptoms. If a diagnosis cannot be drawn, go back to step 3.
7. *Test new hypotheses*: Repeat the process until a diagnosis is produced.

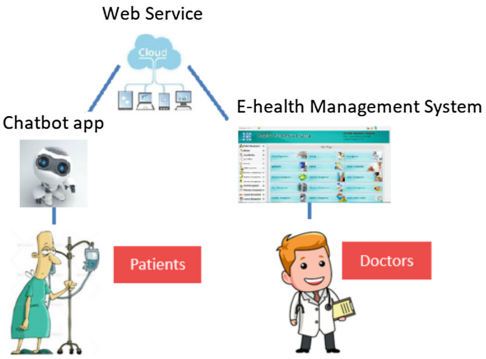


Fig. 1. An illustration of the application scenario and system architecture of Mandy.

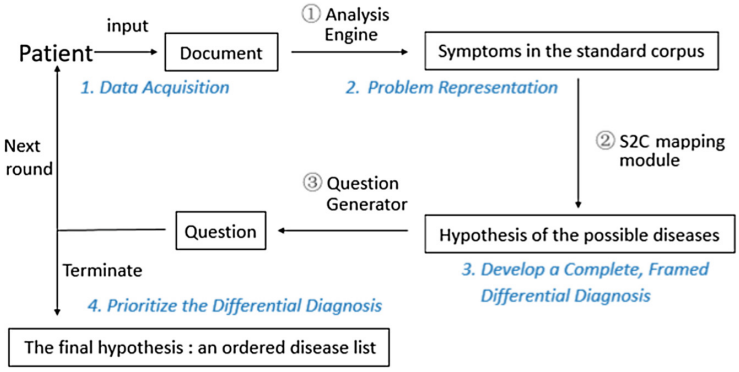


Fig. 2. Main procedure

Figure 2 illustrates the main control flow of Mandy. It simulates Steps 1–4 of the clinical reasoning process above. Mandy starts by asking the patient's chief complaint. After the patient inputs a text in natural language, the *analysis engine* extracts the symptoms in a standard corpus from the patient description

text. In this way, the system gets an accurate problem representation. Then, the *symptoms-to-cause (S2C) mapping module* comes up with a list of hypothetic diseases based on the symptoms provided by the patient's complaint. The system ranks the possibility of the hypothetic diseases. If there is enough information for proposing the final hypothesis list, the procedure will terminate; Otherwise, the *question generator* will produce another question for the patient and repeats the procedure back to the analysis engine.

We next describe the key data structures and algorithmic modules. The internal algorithms of Mandy rely on the following sets:

A *symptom* is a subjective, observable condition that is abnormal and reflects the existence of certain diseases. For ease of terminology, we abuse the notion including also *signs*, which are states objectively measured by others. A patient *feature* is a fact reflecting the patients, age, gender, geographical and demographical information and life styles (e.g. smoking, alcoholic). Mandy uses a set S of words representing standard symptoms and patient features that are extracted from an external knowledge base.

A *disease* is a medical condition that is associated with a set of symptoms. Mandy also uses a set L of standard diseases. The connection between S and L is captured by a matching function $f\colon L \to 2^S$ where each disease $\ell \in L$ is associated with a subset $f(\ell) \subseteq S$.

Example 1. For the diseases "allergies" and "asthma", we have:

f(allergies) = {sneezing, runny nose, stuffy nose, cough, postnasal drip, itchy nose, itchy eyes, itchy throat, watery eyes, dry skin, scaly skin, wheezing, shortness of breath, chest tightness}

f(asthma) = {cough, difficulty breathing, chest tightness, shortness of breath, wheezing, whistling sound when exhaling, frequent colds, difficulty speaking, breathless}

Module I: Analysis Engine. The *analysis engine* understands user's natural language input and extracts a set of symptoms and features from the set S. To implement an effective mechanism that can handle arbitrary language input from the user, we apply Google's *word embedding* algorithm word2vec to map words into a vector space to capture their semantic similarities [2,20,21]. There are two scenarios that word embedding plays a key role: Firstly, when patients describe symptoms in a lay language, the analysis engine picks up keywords and constructs bags of words that represent all patients' symptoms. The algorithm analyzes the most likely diseases by comparing the similarity of the patient's symptoms and all common disease's symptoms. Secondly, when the input words do not appear in the standard corpus S, the analysis engine computes words similarity using a word2vec model, which is pre-trained on a large dataset of medical documents. The words similarity will allow the system to find symptoms in the set S that best align with the input description (Fig. 3).

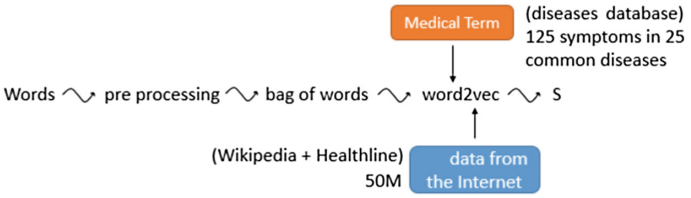


Fig. 3. The algorithmic process of the analysis engine in Mandy.

Example 2. We give two symptoms with their top-10 similar words:

`rash: blisters, itchy, scabies, bumps, hives, ringworm, scaly, bite, flaky, planus`
`nausea and vomiting: dizziness, abdominal pain, nausea, drowsiness, lightheadedness, cramps, sleepiness, vertigo, weakness, bloating`

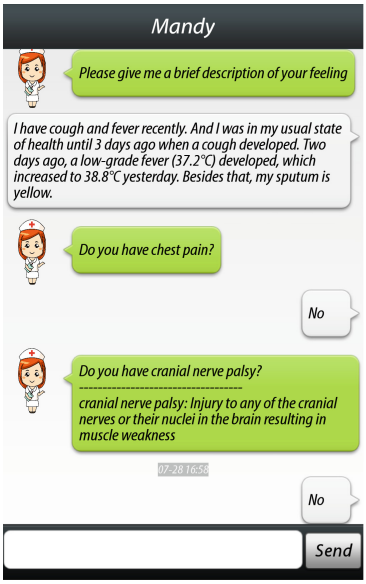
Module II: S2C Mapping. This module takes the set $S' \subseteq S$ of patient's symptoms (output of the analysis engine) as input and computes a hypothesized disease $\ell \in L$ that corresponds to S'. We propose an algorithm, named *Positive-Negative Matching Feature Count* (P − N)MFC, to compare the similarity between S and $f(\ell)$ for all $\ell \in L$. The algorithm runs the following steps: Suppose that we have a set S_+ of positive symptoms of the patient and a set S_- of negative symptoms. Suppose also that the set of diseases L is $\{d_1, d_2, \ldots\}$ and let $S_{d_i} = f(d_i)$ be the set of symptoms corresponding to d_i. For every $d_i \in L$:

1. Calculate $S_+ \cap S_{d_i}$, and let $n_i^+ = |S_+ \cap S_{d_i}|$.
2. Calculate $S_- \cap S_{d_i}$, and let $n_i^- = |S_- \cap S_{d_i}|$.
3. Calculate $\sigma_i = (n_i^+ - n_i^-)$, this is the similarity value of the patient's symptoms with each disease's symptom.

The (P − N)MFC algorithm selects $d_i \in L$ that has the highest σ_i value has the next hypothesis.

Module III: Question Generator. This module takes a list of hypothesized diseases $C \subseteq L$ as input, and generates a new question with a most likely symptom for the patient to confirm. Unless Mandy has obtained enough information to derive a diagnosis, the system will continue to pose new questions to the patient. Note that the input list of hypotheses is the result obtained from S2C Mapping Module; element in the list are ordered by the likelihood according to the current patient info. The output is a symptom that Mandy selects from the knowledge base which represent the most likely symptom the patient has. Mandy will form a question that asks the patient to confirm or reject this symptom. The detailed steps of the algorithm is as follows:

1. Update S_+ and S_- according to the patients input.
2. If S_+ has a new element, perform the (P − N)MFC algorithm to get the most likely disease $\ell \in L$. If $f(\ell) \setminus S_+ \neq \varnothing$, randomly choose one such symptom in $f(\ell)$ but not in S_+ and ask about it in the next question.



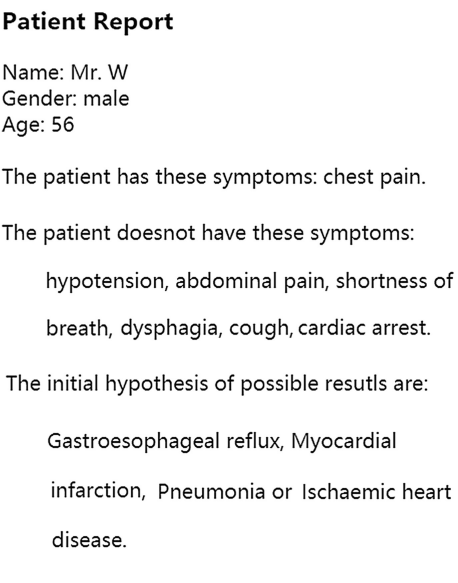


Fig. 4. Left: The app user interface; Whenever users encounter obscure medical terms, the relevant explanation from Merriam-Webster Dictionary can be viewed by clicking the dialog box. Right: The generated initial interview outcome report.

3. If $f(\ell)$ does not contain any symptom not in S_+, the system will analyze patient's input, then choose the most similar symptom in our standard corpus, and use it in the next question.
4. Once the system has got enough information from the patient, it will generate a diagnosis result, list top-most possible diseases which are related to the patient's symptoms.

We deploy a proof-of-concept of Mandy on an Amazon Web Services Cloud[10]. It provides services for both the mobile app version[11] (see Fig. 4) and the PC version[12]. Knowledge about symptoms and diseases is constructed based on external sources[13]. In this proof-of-concept, we select 25 common diseases. The dataset for word2vec to train a word embedding consists of crawled entries from the Internet. Firstly, on Wikipedia[14], the crawler dredges data from the main page of "disease" and visit each medical terminology using hyperlinks. To collect

[10] https://aws.amazon.com/.
[11] https://github.com/lni600/Mandy.git.
[12] http://13.54.91.140:8080/HealthWebApp/ To log in, the user needs to input '`admin`' as both Username and Password.
[13] E.g. online databases such as http://www.diseasesdatabase.com.
[14] https://en.wikipedia.org/.

more colloquial sentences, we also crawled data from Healthline[15]. The collected dataset contains approximately 20,000 web pages on Wikipedia and about 10,000 web pages on Healthline with a size of ≈50 MB.

4 Performance Evaluation

We extracted case studies from a standard medical textbook which contains numerous real-life patient complaints with suggested diagnosis [27]. We evaluate the performance of our proof-of-concepts on four randomly selected disease categories: Chest Pain, Respiratory Infections, Headache and Dizziness. Each case study starts with a patient description and then a list of hypotheses containing valid hypotheses which can be viewed as ground truth results. We investigate the result of Mandy on 11 such case studies.

1. Evaluating the Generated Questions. Mandy is intended to communicate with the patients just like a real healthcare staff. An ideal intake interviewer should pose a list of personalized questions that truthfully reflect the medical conditions of the patient and lead to meaningful information for their treatment. Thus the questions generated by Mandy during an interview amounts to a crucial criterion for its effectiveness.

From the patient description, we recognize main symptoms. We then input only the first symptom to the system and check if the system can generate high-quality questions. We regard the questions which covered the other symptoms as "high-quality" since they are sufficient and important for the doctors to come up with the hypothesis list.

Example 3. One case study includes the following patient description: *"Mrs. G is a 68-year-old woman with a history of hypertension who arrives at the emergency department by ambulance complaining of chest pain that has lasted 6 hours. Two hours after eating, moderate (5/10) chest discomfort developed. She describes it as a burning sensation beginning in her mid chest and radiating to her back. She initially attributed the pain to heartburn and used antacids. Despite multiple doses over 3 hours, there was no relief. Over the last hour, the pain became very severe (10/10) with radiation to her back and arms. The pain is associated with diaphoresis and shortness of breath. The patient takes enalapril for hypertension. She lives alone, is fairy sedentary, and smokes 1 pack of Cigarettes each day. She has an 80 pack year smoking history."* The symptoms extracted from the text are **severe chest pain** associated with **diaphoresis** and **shortness of breath**. To evaluate the generated questions, we only provide "severe chest pain", and see if "diaphoresis" and "shortness of breath" will be asked by Mandy.

After we input "`severe chest pain`" as the answer to the first question, Mandy generated the following interaction. The answers to the questions were obtained from understanding the text description above:

[15] http://www.healthline.com/.

Mandy: Do you have dysphagia? Answer: no
Mandy: Do you have hypotension? Answer: no
Mandy: Do you have cardiac arrest? Answer: no
Mandy: Do you have hyperhidrosis? Answer: yes
Mandy: Do you have fever? Answer: no
Mandy: Do you have abdominal pain? Answer: no
Mandy: Do you have shortness of breath? Answer: yes
Mandy: Do you have nausea and vomiting? Answer: no
Mandy: Do you have productive cough? Answer: no
Mandy: Do you have any other symptoms? Answer: no

Among the 9 questions symptoms, two match exactly as our expected questions. Thus we evaluate the accuracy as: ***Question Accuracy*** = matched questions ÷ expected questions = 2/2 = 100%.

Using the same approach, we calculate question accuracy for six test cases (the other five test cases are all single symptom cases, so they cannot be used to evaluate the question accuracy). See Table 2. Among the six cases, two are from each of Chest Pain and Respiratory Issues, and a single case is from each of Dizziness and Headache. Besides the case for Dizziness which only asks 2 high-quality questions out of the expected 3 ones, the question accuracies for the other cases are all 100%.

2. The Performance of the Diagnosis Module. Another natural evaluation criterion is the diagnosis capability of Mandy. For this, we input the entire paragraph of patient description into the system as the answer to the first question. We then answer subsequent questions manually based on understanding of the case description. When the system has no more questions for the patient, we check if the output hypothesis list from the system matches with the ground truth hypotheses from the book (Table 1).

Table 1. Diagnostic hypotheses for Mr.W.

Diagnostic hypotheses	Clinical clues	Important tests
Leading hypothesis		
Stable angina	Substernal chest pressure with exertion	Exercise tolerance test Angiogram
**Active alternative—	**Most common**	
GERD	Symptoms of heartburn chronic nature	EGD Esophageal pH monitoring
Active alternative		
Musculoskeletal disorders	History of injury or specific musculoskeletal chest pain syndrome	Physical exam Response to treatment

EGD, esophagogastroduodenoscopy; GERD, gastroesophageal reflux disease.

Example 4. The patient Mr.W complained that he felt chest pain with squeezing, sub-sternal pressure while climbing stairs. The only symptom recognized is chest pain. The diagnostic hypotheses including stable angina, GERD, and Musculoskeletal disorders from the guide book Table 8-1 [27] are shown here.

The hypotheses report from our system shows that one out of the four hypotheses is matched with the guide book (GERD). Another hypothesis "Myocardial infarction" (MI) from our system shares the same disease category with "stable angina" from the guide book. We regard MI as correct because it is close enough and "stable angina" does not exist in our current disease corpus.

Table 2. Question and prediction accuracy of Mandy over the case studies.

Disease category	Question accuracy	Prediction accuracy
Respiratory issues	100%	100%
Chest Pain	100%	64%
Headache	100%	25%
Dizziness	66.7%	14%

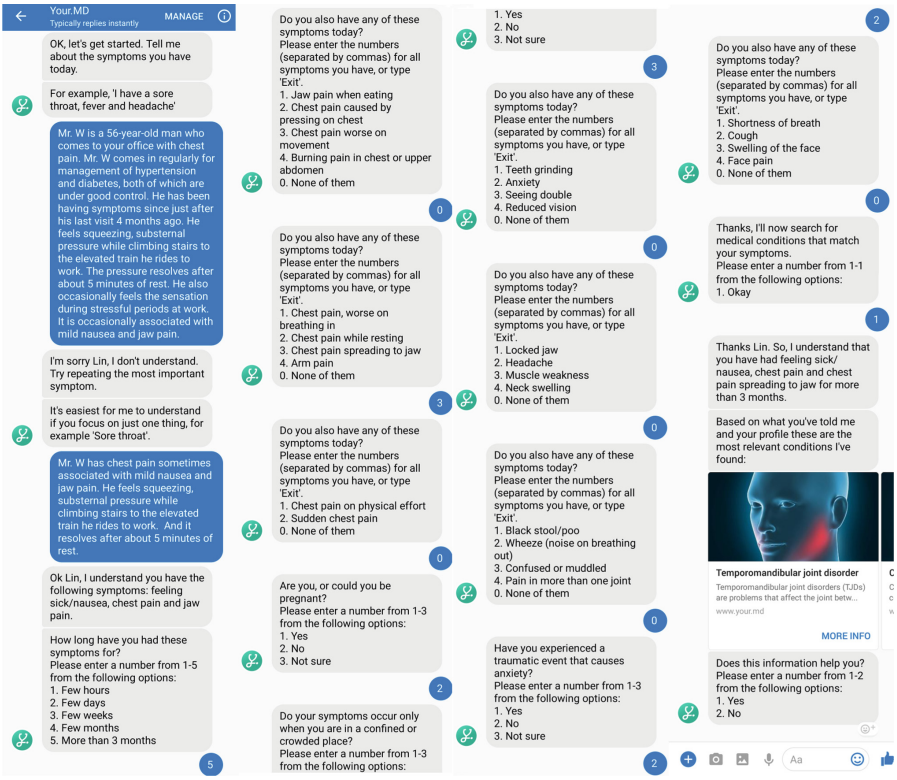


Fig. 5. Mr.W's case on Your.MD

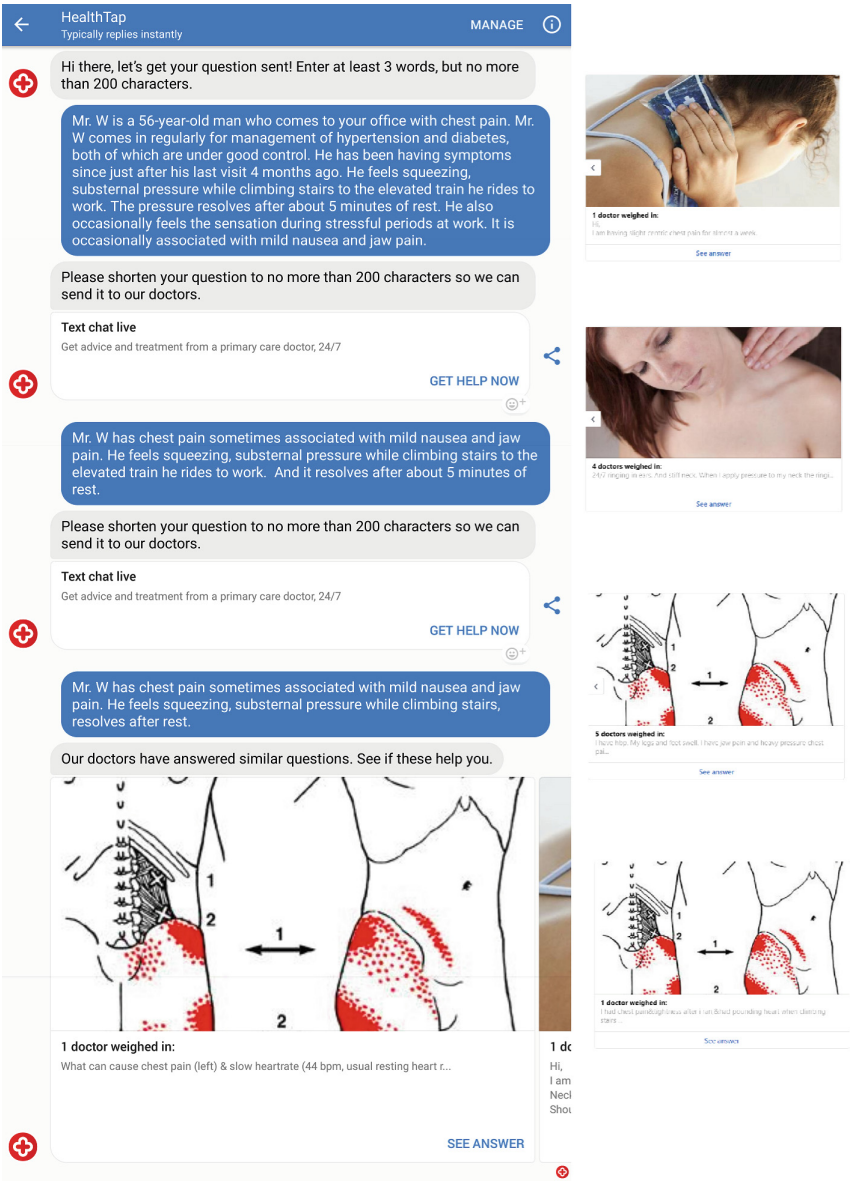


Fig. 6. Mr.W's case on HealthTap

Therefore we conclude that the final accuracy of our system for this case is: ***Prediction Accuracy*** = matched hypotheses from our system/ diagnostic hypotheses in guide book = $2/3 = 67\%$.

Following the same approach, we calculate all the prediction accuracy for the 11 test cases. See Table 2. The low prediction accuracies for Dizziness and

Headache are mainly caused by the lack of training data and knowledge in brain diseases in our system. This can be improved in a future update of the proof-of-concept.

To further evaluate our proof-of-concept, we input Mr.W's case on two well-known existing medical chatbots Your.MD and HealthTap from the Facebook Messenger Bots Platform. The conversations are shown in Figs. 5 and 6. Even including "chest pain" in the description, the results provided by HealthTap were not convincing. Similarly, after Your.MD checked 30 symptoms, the two conclusions were far from the correct one. On this test case, Mandy clearly outperforms these existing chatbots as the questions are related to the symptoms and the hypotheses list also make sense.

5 Conclusion and Future Work

We develop an integrated, intelligent and interactive system called Mandy who is not designed as a diagnostic or clinical decision-making tool but an assistant to doctors. We use word2vec to solve the NLP problem in this particular domain which works well according to our evaluation experiments.

Much further work is needed to improve the capability of our proof-of-concept. Firstly, we need to include more diseases into our system. The total number of human diseases is over 1000, so there is still a lot of work to do. Secondly, a symptom synonym thesauri should also be produced. Then we could generate questions with more understandable symptoms for the patients. Additionally, the symptom thesauri could improve the performance of the trained model. Because more than one target will definitely increase the mapping possibility of patient's doc and standard symptom corpus. Thirdly, the update of our S2C module is necessary. Currently, we only deal with symptoms, due to the lack of proper data, though the function is able to handle other features, such as gender and bad habits. Another data set we desired, is the S2C kind of file with weight for each symptom. Some symptoms are highly likely to lead to some disease more than others. This could greatly improve our system's performance. Additionally, we also plan to add a case-based Incremental learning and reinforcement learning algorithm to enhance the diagnosis accuracy. Besides, the separate modules in our structure make it possible to replace our S2C module like a plug-in with another diagnosis system, if which also provides an ordered hypothesis list. The last but not least, to achieve the ambitious goal, chatting like a human, we need to acquire real life Patient-Doctor conversation data, which will give us more confidence to provide smarter interaction.

Acknowledgement. The first author is partially funded by a scholarship offered by Precision Driven Health in New Zealand, a public-private research partnership aimed at improving health outcomes through data science. Initial progress of the research was reported in the PDH & Orion Health Blog `https://orionhealth.com/global/knowledge-hub/blogs/meet-mandy-an-intelligent-and-interactive-medicare-system/`.

References

1. Ahmad, F., Hogg-Johnson, S., Stewart, D.E., Skinner, H.A., Glazier, R.H., Levinson, W.: Computer-assisted screening for intimate partner violence and control a randomized trial. Ann. Intern. Med. **151**(2), 93–102 (2009)
2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**(Feb), 1137–1155 (2003)
3. Bernstein, S.L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., John McConnell, K., Pines, J.M., Rathlev, N., et al.: The effect of emergency department crowding on clinically oriented outcomes. Acad. Emerg. Med. **16**(1), 1–10 (2009)
4. Bickmore, T., Giorgino, T.: Health dialog systems for patients and consumers. J. Biomed. Inform. **39**(5), 556–571 (2006)
5. Bickmore, T.W., Pfeifer, L.M., Byron, D., Forsythe, S., Henault, L.E., Jack, B.W., Silliman, R., Paasche-Orlow, M.K.: Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J. Health Commun. **15**(S2), 197–210 (2010)
6. Black, L.-A., McTear, M., Black, N., Harper, R., Lemon, M.: Appraisal of a conversational artefact and its utility in remote patient monitoring. In: Proceedings of 18th IEEE Symposium on Computer-Based Medical Systems, pp. 506–508. IEEE (2005)
7. Caley, M., Sidhu, K.: Estimating the future healthcare costs of an aging population in the UK: expansion of morbidity and the need for preventative care. J. Publ. Health **33**(1), 117–122 (2011)
8. Clancey, W.J., Letsinger, R.: NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. Stanford University, Department of Computer Science (1982)
9. Delichatsios, H.K., Friedman, R.H., Glanz, K., Tennstedt, S., Smigelski, C., Pinto, B.M., Kelley, H., Gillman, M.W.: Randomized trial of a "talking computer" to improve adults' eating habits. Am. J. Health Promot. **15**(4), 215–224 (2001)
10. Di Somma, S., Paladino, L., Vaughan, L., Lalle, I., Magrini, L., Magnanti, M.: Overcrowding in emergency department: an international issue. Intern. Emerg. Med. **10**(2), 171–175 (2015)
11. Farzanfar, R., Frishkopf, S., Migneault, J., Friedman, R.: Telephone-linked care for physical activity: a qualitative evaluation of the use patterns of an information technology program for patients. J. Biomed. Inform. **38**(3), 220–228 (2005)
12. High, R.: The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. IBM Corporation, Redbooks (2012)
13. Hubal, R.C., Day, R.S.: Informed consent procedures: an experimental test using a virtual character in a dialog systems training application. J. Biomed. Inform. **39**(5), 532–540 (2006)
14. Hunt, D.L., Haynes, R.B., Hanna, S.E., Smith, K.: Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA **280**(15), 1339–1346 (1998)
15. Jones, P., Chalmers, L., Wells, S., Ameratunga, S., Carswell, P., Ashton, T., Curtis, E., Reid, P., Stewart, J., Harper, A., et al.: Implementing performance improvement in new zealand emergency departments: the six hour time target policy national research project protocol. BMC Health Serv. Res. **12**(1), 45 (2012)
16. Khoury, M.J., Ioannidis, J.P.A.: Big data meets public health. Science **346**(6213), 1054–1055 (2014)

17. Lipkin, M., Quill, T.E., Napodano, R.J.: The medical interview: a core curriculum for residencies in internal medicine. Ann. Intern. Med. **100**(2), 277–284 (1984)
18. Martínez-Pérez, B., de la Torre-Díez, I., López-Coronado, M., Sainz-De-Abajo, B., Robles, M., García-Gómez, J.M.: Mobile clinical decision support systems and applications: a literature and commercial review. J. Med. Syst. **38**(1), 4 (2014)
19. McFillen, J.M., O'Neil, D.A., Balzer, W.K., Varney, G.H.: Organizational diagnosis: an evidence-based approach. J. Change Manag. **13**(2), 223–246 (2013)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
22. Miller, R.A., Pople Jr., H.E., Myers, J.D.: Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. New Engl. J. Med. **307**(8), 468–476 (1982)
23. Ramelson, H.Z., Friedman, R.H., Ockene, J.K.: An automated telephone-based smoking cessation education and counseling system. Patient Educ. Couns. **36**(2), 131–144 (1999)
24. Realdi, G., Previato, L., Vitturi, N.: Selection of diagnostic tests for clinical decision making and translation to a problem oriented medical record. Clin. Chim. Acta **393**(1), 37–43 (2008)
25. Richardson, D.B.: Increase in patient mortality at 10 days associated with emergency department overcrowding. Med. J. Aust. **184**(5), 213–216 (2006)
26. Scheffler, R.M., Liu, J.X., Kinfu, Y., Dal Poz, M.R.: Forecasting the global shortage of physicians: an economic-and needs-based approach. Bull. World Health Organ. **86**(7), 516–523B (2008)
27. Stern, S., Cifu, A., Altkorn, D.: Symptom to Diagnosis an Evidence Based Guide. McGraw Hill Professional, New York City (2014)
28. Taber, J.M., Leyva, B., Persoskie, A.: Why do people avoid medical care? A qualitative study using national data. J. Gen. Intern. Med. **30**(3), 290–297 (2015)
29. Victor, L.Y., Buchanan, B.G., Shortliffe, E.H., Wraith, S.M., Davis, R., Davis, A.R., Scott, A.C., Cohen, S.N.: Evaluating the performance of a computer-based consultant. Comput. Prog. Biomed. **9**(1), 95–102 (1979)
30. Warren, J.R.: Better, more cost-effective intake interviews. IEEE Intell. Syst. Appl. **13**(1), 40–48 (1998)
31. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)