Emotion-aware Dialogue Agent for Consumer Health Question Answering

Kalpani Anuradha Welivita
Supervisor: Dr. Pearl Pu Faltings
Human Computer Interaction Group, IC, EPFL

***Abstract*—Consumers struggle to find reliable and succinct answers for their health queries by having to traverse through long lists of documents returned by traditional search engines. Automated consumer health question answering systems address this concern by providing direct and concise answers to consumer health questions. Consumer health questions tend to contain a lot of ambiguities due to the gap between consumer health vocabulary and medical terminology. But existing systems are mostly non-interactive and limit in their ability to resolve ambiguities present in consumer queries by interactively conversing with the user. They also lack in emotion-aware capabilities to respond to consumer health questions in a natural and empathetic manner.**

We present EMA (Empathetic Medical Agent) that can succinctly answer consumer health questions in a conversational manner. It is capable of interactively engaging with the user to resolve any ambiguities detected in consumer queries, enabling it to provide direct and concise answers to the questions asked. We also present our plan on extending the capabilities of EMA to respond in an emotion-aware empathetic manner, making the interaction more natural and human-like.

***Index Terms*—consumer health question answering, dialogue agents, information retrieval, machine learning comprehension, query disambiguation, empathetic response generation**

I. Introduction

A vast amount of information related to health is available on the web. Web search engines and online health information platforms such as WebMD[1] and MayoClinic[2] enable consumers to navigate through a collection of related documents and find answers to their health questions. Often the answer to such a question is a short span of text that can be found on one of the documents retrieved. But having to search through a long list of retrieved documents to find an answer consumes a lot of time and annoys the consumer—especially during emergency situations. This implies the need for information retrieval systems that can succinctly answer consumer health questions.

Several work have already focused on the task of automatic consumer health question answering [1], [2], [3], [4]. They have been able to achieve above average scores at the TREC LiveQA 2017 medical task, which addresses automatic answering of consumer health questions received by the U.S. National Library of Medicine [5]. enquireMe [6] is the only system found in the literature that performs this task in a multi-turn conversational manner.

Previous studies on automatic consumer health question answering have several limitations:

1) Because of the naiveness and lack of domain expertise, it is often the case that consumer health questions are ill-formed and contain a lot of ambiguity. For example, a consumer may mistype a medical term; describe a medical condition in his/her own words without referring to the exact medical term; or not clearly convey what information he/she requires regarding a medical condition. This implies the importance of interactively engaging with consumers to refine their queries and get them validated. But existing work lack such mechanisms to eliminate question ambiguity and increase the accuracy of the answers retrieved.
2) They also fail to identify the affect associated with consumer health questions. Emotions play an important role in the healthcare setting. A study conducted by S. Sundararajan and V. Gopichandran [7] to evaluate the emotional intelligence of medical students in India shows that positive emotions such as empathy, comfort and rapport have a positive influence on the relationship between the doctor and the patient—eliciting patient satisfaction. It has also been found that emotion-awareness increases user satisfaction and enhance the system-user interaction [8]. These studies imply the importance of having empathetic question answering dialog agents in healthcare.

To address these gaps, we designed EMA (Empathetic Medical Agent), an emotion-aware dialog agent to answer consumer health questions in an empathetic manner. Currently, EMA answers questions using a large unstructured collection of documents composed of credible health related articles from WebMD, the most popular source of health information in the United States. It has the ability to ask follow-up questions to clarify any potential errors or ambiguities in consumer health questions. This makes it possible for EMA to understand the exact information need of the consumer and filter out any irrelevant answers retrieved. Research is on going on affective/emotional open-domain neural response generation to extend the capabilities of EMA to carry out open-domain conversations or chitchat with the users while responding in an emotion-aware empathetic manner.

This report is structured as follows. Section II reviews different question answering approaches, query disambiguation and affective response generation techniques used in the literature.

[1] www.webmd.com
[2] www.mayoclinic.org

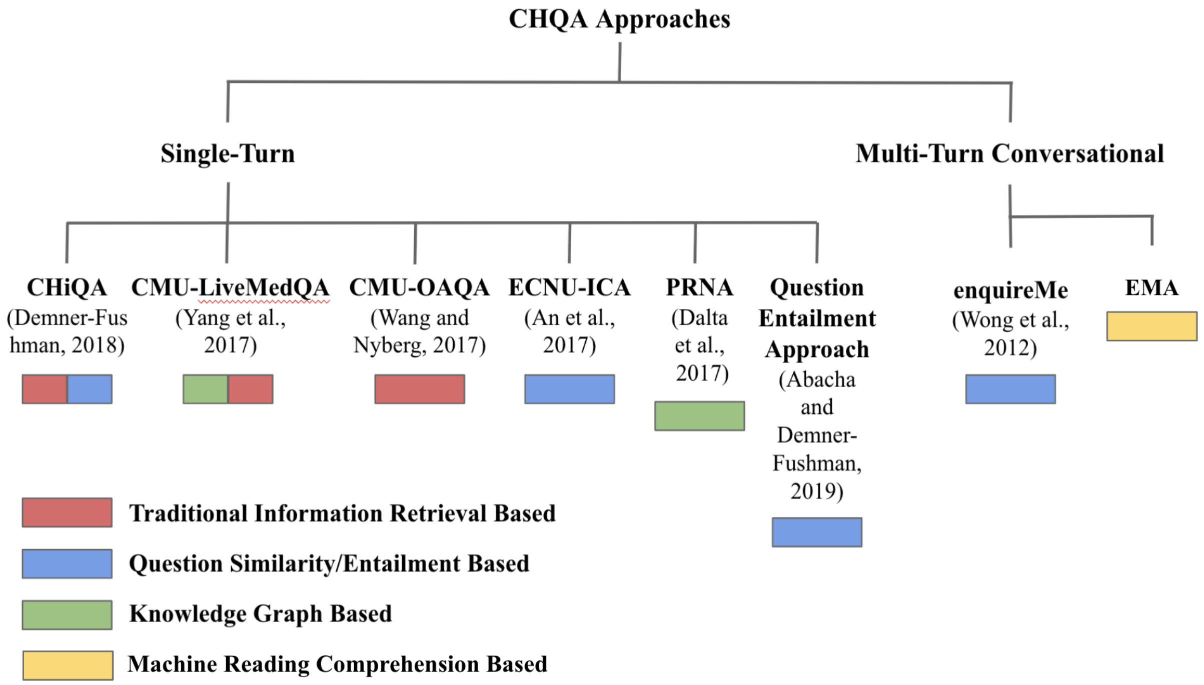


Fig. 1. Classification of consumer health question answering approaches.

Section III discusses the design and architecture of EMA, including its approach to question answering, input understanding and dialogue management and query disambiguation. In section IV, we present the results obtained for question answering and a case study carried out to give the readers a taste of the current capabilities of EMA. Section V elaborates on the results and discusses the limitations. In section VI, we discuss our plan to extend the capabilities of EMA to carry out empathetic open-domain conversations with the user. Section VII concludes the report highlighting our main contributions.

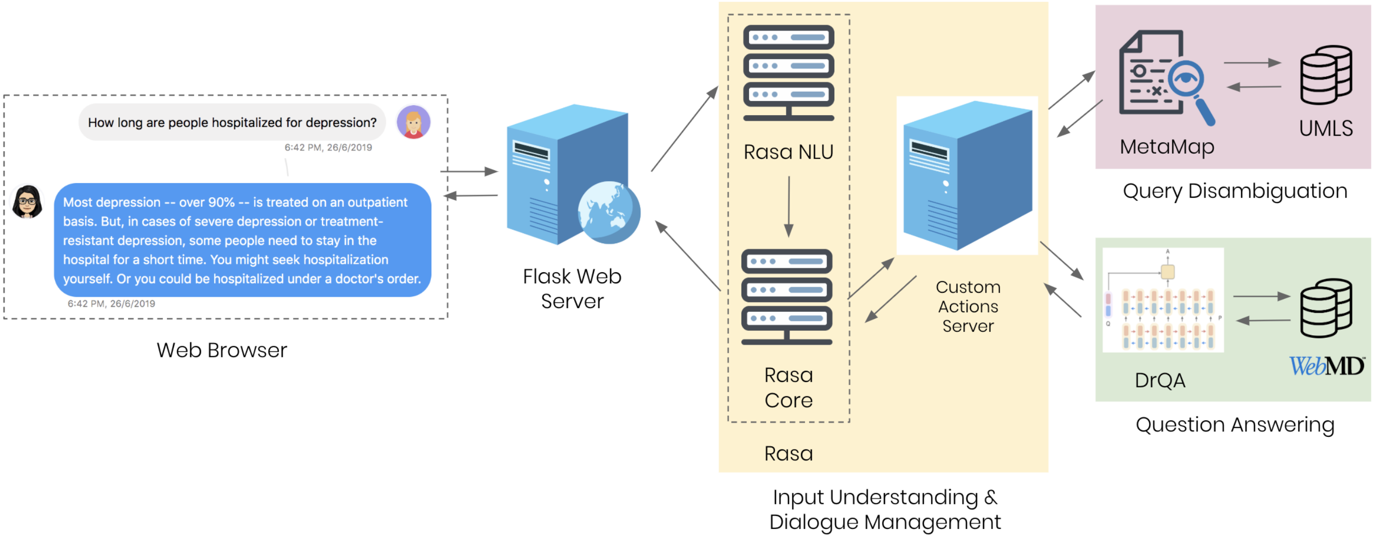
II. Literature Review

There have been a number of systems developed in the literature that specifically address the task of consumer health question answering. The medical question answering task organized by the TREC 2017 LiveQA track [5] provides an open benchmark to compare single-turn consumer health question answering approaches. The task was organized in the scope of the consumer health question answering project[3] of the U.S. National Library of Medicine (NLM) to address automatic answering of consumer health questions received by the U.S. NLM. It motivated a number of research groups such as Carnegie Mellon University's Open Advancement of Question Answering (CMU-OAQA) [1], East China Normal University Institute of Computer Application (ECNU-ICA) [2], Philips Research North America (PRNA) [3] and Carnegie Mellon University's LiveMedQA (CMU-LiveMedQA) [4] to develop approaches focusing on consumer health question answering. Further, the US National Library of Medicine itself has come up with the online available Consumer Health Information and Question Answering (CHIQA) system [9], and a novel question entailment based consumer health question answering approach [10]. enquireMe [6] is a multi-turn contextual approach to answer consumer health questions.

The above approaches can be categorized into 3 based on the techniques used for question answering as: (1) traditional information retrieval based approaches; (2) question similarity/entailment based approaches; and (3) knowledge graph based approaches. Traditional information retrieval based approaches formulate queries by analyzing question text and retrieve candidate answers using search engines or TF-IDF based scoring methods [9], [1]. They employ neural or non-neural scoring techniques or weighted combinations of them to rank the candidate answers and select the best among them. Question similarity/entailment based approaches assume that the answer to a given question is the best answer of the most similar or the most entailed question that already have associated answers [2], [10], [6]. They retrieve similar questions having associated answers, using search in Q&A websites or TF-IDF based scoring methods and select the best answer by ranking the retrieved questions based on the question similarity or the question entailment score calculated using neural or non-neural scoring techniques. Knowledge graph based approaches build and maintain a knowledge graph consisting of health related information such that it preserves the medical domain-specific concept hierarchy [3], [4]. They extract medical concepts from the question text and traverse along the knowledge graph to find the correct answer. There are also works on hybrid methods that combine the strengths of more than one of the above approaches [9], [4]. The works that belong to each of these categories are illustrated in Figure 1. EMA uses a novel approach that uses machine reading comprehension of text to answer consumer health questions.

Machine reading comprehension (MRC) combined with

[3] lhncbc.nlm.nih.gov/project/consumer-health-question-answering



How long are people hospitalized for depression?

6:42 PM, 26/6/2019

Most depression -- over 90% -- is treated on an outpatient basis. But, in cases of severe depression or treatment-resistant depression, some people need to stay in the hospital for a short time. You might seek hospitalization yourself. Or you could be hospitalized under a doctor's order.

6:42 PM, 26/6/2019

Fig. 2. The overall architecture of EMA.

information retrieval is a popular technique used for answering open-domain factoid questions. Systems such as DrQA [11] and FlowQA [12] can be cited as popular examples for such systems. MRC is the task of answering questions after reading a short text or story. Because of recent developments in deep learning such as attention-based and memory-augmented neural networks [13], [14] and release of new training and evaluation datasets such as SQuAD [15], QuizBowl [16], CoQA [17] and QuAC [18], this field has been able to make considerable progress. We test how MRC combined with traditional information retrieval is able to answer consumer health questions, which are non-factoid in nature and answers of which span across several paragraphs of text.

Work has been previously carried out on syntactic and semantic disambiguation of terminology used in healthcare [19], [20], [21], [22]. The studies have utilized Unified Medical Language System (UMLS) Metathesaurus [23] specialist lexicon and semantic information to obtain a domain-specific source of dictionary terms and meanings. UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. The Metathesaurus is the largest component of UMLS, which is a large biomedical thesaurus that is organized by concept, or meaning, and links similar names for the same concept from nearly 200 different vocabularies.

Recently, there have been a lot of work targeting affective neural response generation. For example, MojiTalk [24] utilizes conditional variational autoencoders trained on a large corpus of Twitter conversations that include emojis to generate high-quality abstractive but emotionally appropriate conversation responses. Asghar et al. [25] use a cognitively engineered word-level affective dictionary based on the Valence-Arousal-Dissonance (VAD) affective notation [26] to augment traditional word embeddings used in neural conversational models with a 3D affective space. They also introduce affective loss functions that augment the standard cross entropy loss used for training neural conversational models with affective objectives; and affectively diverse beam search for decoding, which injects affective diversity in to responses generated.

III. EMA: Empathetic Medical Agent

A. System Overview

EMA is a browser compatible dialogue agent built on top of several components for: (1) question answering; (2) input understanding and dialogue management; and (3) query disambiguation. We make use of the DrQA system described by Chen et al. [11] to traverse through our document collection and find spans of text that best answer a question. To understand user input and manage dialogue, we use Rasa[27], an open source machine learning framework. Query disambiguation is managed using custom actions written by ourselves with the help of MetaMap [28], which detects and maps medical terms identified in text to related concepts in the UMLS Metathesaurus [23]. Rasa communicates back and forth with the custom actions server for question answering and query disambiguation and the output is delivered to the web application frontend through a Flask web server. The overall system architecture of EMA including how communication happens between components is indicated in Figure 2.

Majority of our work lies in preparing data and training the DrQA system to answer consumer health questions; training Rasa machine learning models to understand user input and manage dialogue; and building the pipeline from taking in user queries to delivering concise answers, resolving any ambiguities present in them by interactively conversing with the user. The following subsections describe each of the components in EMA in detail.

B. Question Answering

The architecture of the question answering component of EMA, which is based on DrQA [11] is illustrated in Figure 3. It mainly consists of 2 components: (1) the document retriever for finding the most relevant articles for a given question from

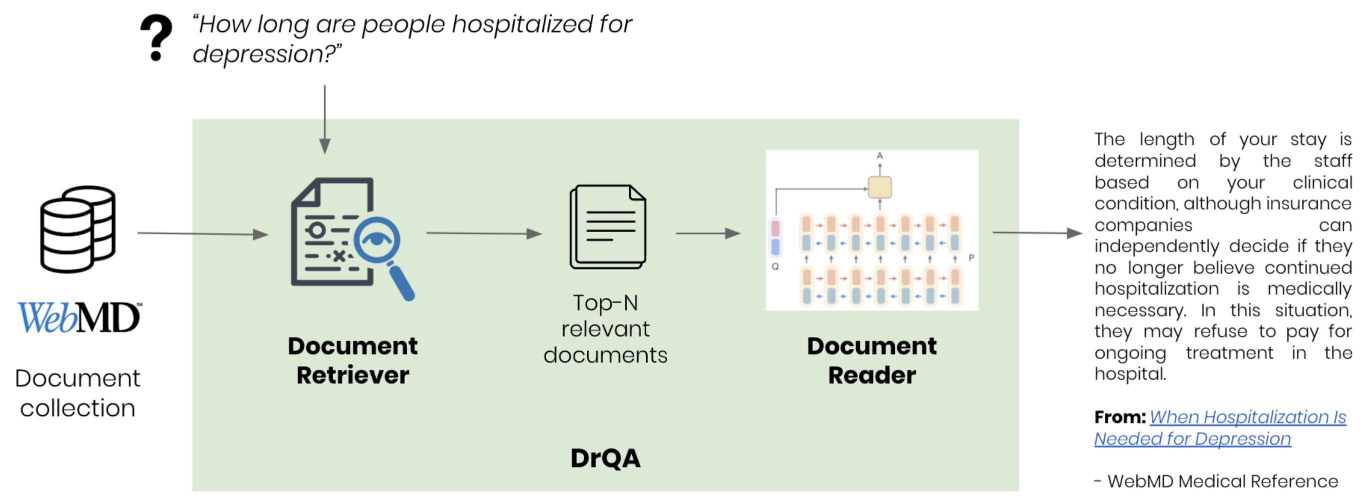

“How long are people hospitalized for depression?”
WebMD
Document collection
Document Retriever
Top-N relevant documents
Document Reader
DrQA
The length of your stay is determined by the staff based on your clinical condition, although insurance companies can independently decide if they no longer believe continued hospitalization is medically necessary. In this situation, they may refuse to pay for ongoing treatment in the hospital.
From: When Hospitalization Is Needed for Depression
- WebMD Medical Reference

Fig. 3. The architecture of the question answering system of EMA based on DrQA [11].

a large collection of documents; and (2) the document reader, a neural machine comprehension model for extracting answer spans from a document or a small collection of documents. The DrQA system is originally used to answer open-domain factoid questions. With certain modifications, we were able to utilize this system to answer consumer health questions, which are usually non-factoid in nature and the answers of which span across several paragraphs of text. The following subsections describe this process in detail.

1) Data Collection and Preprocessing: We crawled WebMD to scrape health related documents spanning all A-Z health topics in WebMD (Asthma, Oral Health, Prostate Cancer, Weight loss and Obesity etc.) and formed an unstructured collection of documents comprising of 10,745 unique articles belonging to 130 health topics. This forms our knowledge source.

We also crawled WebMD for frequently asked questions and their expert reviewed answers, along with the reference WebMD articles those answers are based on. Initially we were able to scrape 44,265 Q&A pairs but after removing duplicates and answers without reference sources the number of Q&A pairs went down to 38,524.

To feed the Q&A pairs to train the neural machine comprehension model described in DrQA, the answers of the pairs have to be direct spans of text extracted from the reference source. However, most of the answers in the Q&A pairs extracted from WebMD were either modified spans of text from the source article or combinations of spans of text taken from different places in the source article. Therefore, we followed our own “answer mapping” process to map each answer to the most matching span of text in the corresponding source article. First, we detect the exactly matching 3-grams between the answer text and the source text and concatenate the ones which are less than a distance of 20 tokens from each other. Then we expand the spans of text formed by concatenation by adding the words that are to the left and right of the test spans using wordnet synonyms and words that are close in edit distance with the words in the original answer. Out of these, we select the text spans having a Jaccard similarity greater than 0.5 with the original answer. After filtering those answer spans, we were left with 35,715 Q&A pairs to feed to the machine comprehension model.

2) Document Retriever: The document retriever uses a simple inverted index lookup followed by term vector model scoring to find the top-N articles that are most relevant to a question. Given a question, first it looks-up the articles in our collection that contain the question terms using simple inverted index lookup, which takes the local word order in to account using bi-gram features. It uses the hashing of Weinberger et al. [29] to map the bi-grams to 2^{24} bins to make it efficient in speed and memory. This is followed by term vector model scoring where the question and articles are compared as TF-IDF weighted bag-of-words vectors, outputting the top-N related articles ranked according to their relevance. We select the top 5 related articles and feed them to the document reader along with the question to select a span of text that best answers the question.

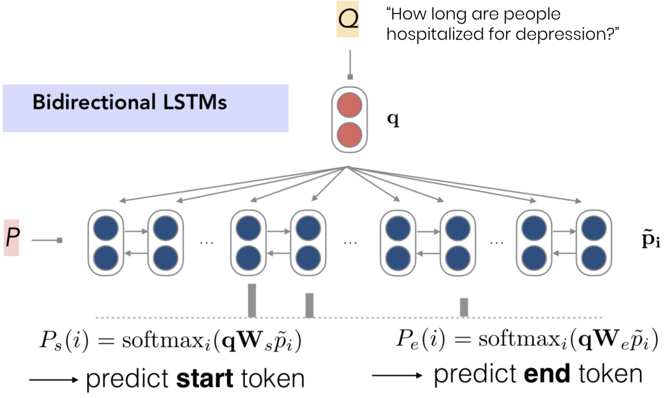

Q
“How long are people hospitalized for depression?”
Bidirectional LSTMs
q
P
$\tilde{\mathbf{p}}_i$
$P_s(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_s\tilde{p}_i)$
predict start token
$P_e(i) = \text{softmax}_i(\mathbf{q}\mathbf{W}_e\tilde{p}_i)$
predict end token

Fig. 4. Bidirectional long-short term memory neural network architecture used in the document reader. Figure derived from [11]

3) Document Reader: The document reader is basically a multi-layer bi-directional long-short term memory (bi-LSTM) neural machine comprehension model trained to detect an answer span given a question and a related document or a small collection of related documents. The bi-LSTM neural network architecture of the document reader is shown in Figure 4. Given an article containing n paragraphs having m tokens $(p_1, ..., p_m)$ and a question q consisting of l tokens $(q_1, ..., q_l)$, it finds a span A in the document that best answers the question. First, the paragraphs are encoded by passing the word embeddings of the paragraph tokens through a multi-layer bi-LSTM network.

$$(\vec{p}_1, ..., \vec{p}_m) \xrightarrow{\text{RNN}} (\boldsymbol{p}_1, ..., \boldsymbol{p}_m) \tag{1}$$

$\boldsymbol{p}_i$ is taken as the concatenation of each layer's hidden units and is expected to encode the useful context information around token p_i. Similarly, another recurrent neural network is applied on top of the word embeddings of the question tokens and the resulting hidden units are combined into one single encoding vector $\boldsymbol{q}$ as follows:

$$(\vec{q}_1, ..., \vec{q}_l) \xrightarrow{\text{RNN}} \boldsymbol{q} \tag{2}$$

$$\boldsymbol{q} = \sum_{j=1}^{l} b_j \vec{q}_j \tag{3}$$

where b_j, given by the following equation, encodes the importance of each question word.

$$b_j = \frac{\exp(\boldsymbol{w}.\vec{q}_j)}{\sum_{k=1}^{l} \exp(\boldsymbol{w}.\vec{q}_k)} \tag{4}$$

where $\boldsymbol{w}$ is the weight vector. Then the paragraph encodings and the question encoding are fed as input to two independently trained RNN classifiers to predict the two ends of the answer span. A bilinear term is used to capture the similarity between $\boldsymbol{p}_i$ and $\boldsymbol{q}$ and compute the probabilities of each token being start and end as:

$$P_{start}(i) \propto \exp(\boldsymbol{p}_i \boldsymbol{W}_s \boldsymbol{q}) \tag{5}$$

$$P_{end}(i) \propto \exp(\boldsymbol{p}_i \boldsymbol{W}_e \boldsymbol{q}) \tag{6}$$

The best span from token i to token i' that maximizes the probability $P_{start}(i) \times P_{end}(i')$ and $i \leq i' \leq (i + 1000)$ is predicted as the answer.

C. Understanding User Input and Managing Dialogue

The Rasa machine learning framework used to understand user input and manage dialogue consists of two main modules: (1) natural language understanding (NLU) module for understanding user input; and (2) core module for holding conversations and deciding what to do next.

In Rasa NLU, incoming messages are processed by a sequence of components. These components are executed one after another in a so-called processing pipeline. There are components for pre-processing, intent classification, entity extraction, and others. For intent classification, we used the Rasa built-in tensorflow embedding classifier, which is based on the StarSpace neural embedding model [30], and trained it on a custom built dataset to identify 9 basic user intents: greet; query; define; refine; suicidal; affirm; deny; thanks; and goodbye and multiple combinations of them. It trains word embeddings from scratch and so is able to adapt to domain specific messages well. One of the limitations in using this classifier is that it requires more training data than classifiers that use pre-trained word embeddings, to generalize well to the target task. But the ability to recognize multiple intents is one of the major reasons for using this classifier over other available classifiers that use pre-trained embeddings.

The Rasa core uses a probabilistic machine learning model trained on annotated conversations or stories to decide which action to take next given the history of a conversation. It differs from other libraries that rely on hand-crafted rules, which are not capable of scaling beyond simple conversations. Since, we do not have a database of conversations annotated with intents and corresponding actions that we want the model to learn from, we used the interactive learning capability of Rasa to generate training data in the Rasa stories format.

The basic end-to-end message handling architecture for answering a simple question (without disambiguation) is shown in Figure 5. When an input message is received, it is passed through the Rasa NLU, which converts it into a dictionary including the original text, the recognized intents, and any entities detected. It is then passed to the Rasa core, in which a tracker object records the input state and pass the current state of the tracker to a policy object to choose which action to take next. The chosen action is logged by the tracker and a response is sent to the user by executing the specified action.

Actions use either hard-coded utterance templates defined by us or external services to generate the responses. We used hard-coded utterance templates to generate simple responses such as: greeting the user; saying thank you, sorry and welcome; asking to rephrase a query; and asking the user whether he/she has any other questions to ask. Custom actions were formulated for detecting and resolving ambiguities and generating answers for questions. When a custom action is predicted, the Rasa core communicates with an external server endpoint that we have specified and modifies the dialogue state based on the information returned. Custom actions are able to access the values of memory variables (slots) and the latest message sent by the user using the tracker object and send responses back to the user through a dispatcher object.

D. Query Disambiguation

EMA is capable of asking simple questions to correct spelling errors related to medical terms such as names of drugs (syntactic disambiguation) and disambiguate the semantic meanings of ambiguous words detected in a query (word sense disambiguation), which enables it to provide concise answers to the questions asked. For this we use the tool MetaMap [28], which detects and maps medical terms identified in text to related concepts in the UMLS Metathesaurus [23]. Using MetaMap we are able to obtain a ranked list of preferred names and their respective semantic types for ambiguous medical terms detected in a question. An example is given in Table I. It is also capable of providing the preferred spellings for the

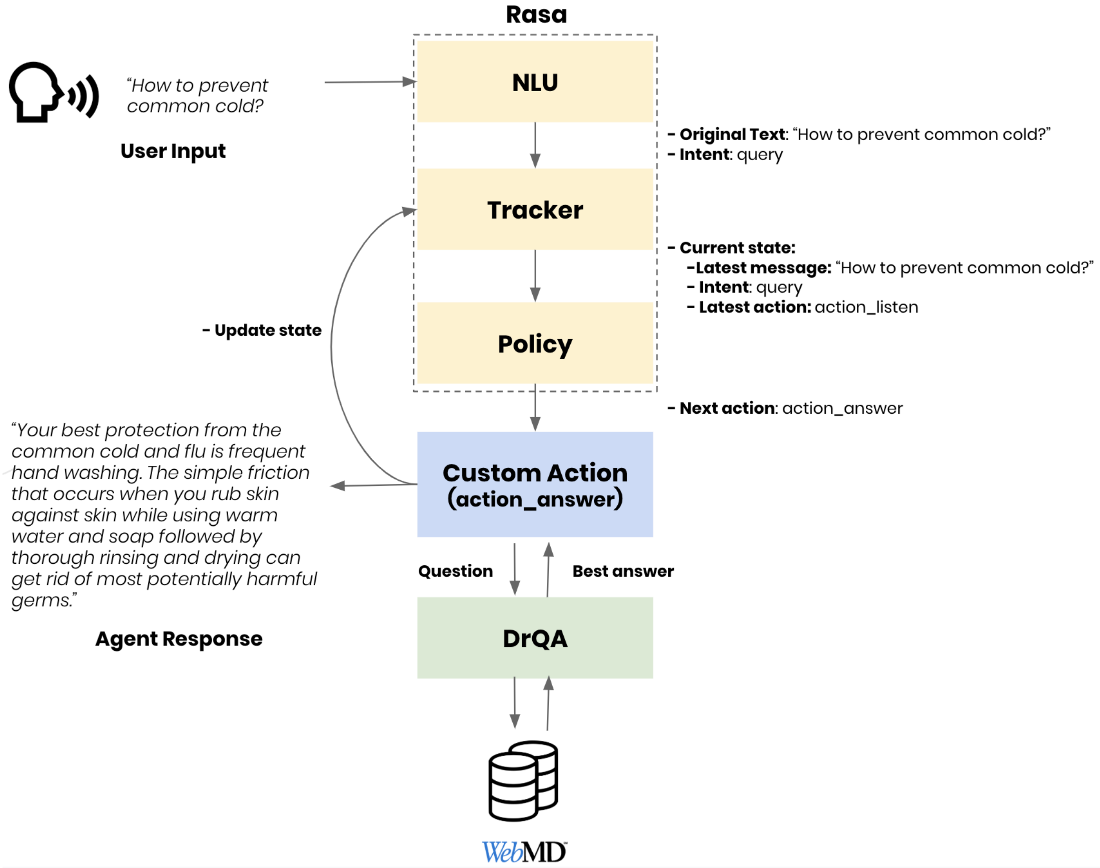

Rasa
"How to prevent common cold?"
User Input
NLU
- Original Text: "How to prevent common cold?"
- Intent: query
Tracker
- Current state:
-Latest message: "How to prevent common cold?"
- Intent: query
- Latest action: action_listen
- Update state
Policy
- Next action: action_answer
"Your best protection from the common cold and flu is frequent hand washing. The simple friction that occurs when you rub skin against skin while using warm water and soap followed by thorough rinsing and drying can get rid of most potentially harmful germs."
Agent Response
Custom Action (action_answer)
Question
Best answer
DrQA
WebMD

Fig. 5. The basic end-to-end message handling architecture for answering a simple question (without disambiguation).

commonly misspelled medical terms such as names of drugs (e.g. "Efexor" vs. "Effexor").

Question: How to get over **cold**?

Preferred name	Score	Semantic type
Cold Temperature	22.51	Natural phenomenon or process
Common Cold	13.05	Disease or syndrome
Upper Respiratory Infections	9.90	Disease or syndrome
Cold Sensation	3.59	Physiologic function

TABLE I
LIST OF PREFERRED NAMES, THEIR SCORES AND SEMANTIC TYPES PROVIDED BY METAMAP FOR THE AMBIGUOUS MEDICAL TERM "COLD" IDENTIFIED IN THE INPUT QUESTION.

For a given question, depending on the number of ambiguities detected either in spelling or in semantic meaning of words, a list of ambiguity types (0-for ambiguous spelling; 1-for ambiguous meaning) and a corresponding list of ambiguous words are maintained as memory variables or slots in the Rasa core. Slots are basically the agent's memory. They act as a key-value store that can be used to store information the user provided as well as information gathered by executing specific actions. The agent resolves each of the ambiguities stored by interactively conversing with the user. The original input question (also maintained as a slot) gets reformulated each time the agent gets an ambiguity resolved and the final reformulated question is sent to the question answering module to obtain an answer.

IV. RESULTS

A. Question Answering

Since the original DrQA system is developed to answer open-domain factoid questions, which are comparatively shorter and less complex than consumer health questions and the answers of which usually consist of one or two tokens; we were expecting the same system trained on consumer health question-answer pairs from WebMD to not perform very well in detecting exact answers for consumer health questions. If this performance lag is significant then the DrQA system would not be quite useful to be used for our context. In order to investigate this, we compared the performance of the document retriever and the document reader components of EMA on the WebMD test Q&A dataset with the performance results given in the DrQA paper [11].

Table II shows the performance of the document retriever for retrieving relevant documents for questions in the WebMD test Q&A dataset compared to retrieving relevant documents for questions in the four factoid Q&A datasets: SQuAD Dev [15]; CuratedTREC [31]; WebQuestions [32]; and WikiMovies [33] on which performance of the document retriever of the original DrQA system was tested. The performance was compared in terms of top-5 accuracy i.e. the percentage of questions for which the ground truth answer appear in one of the top-5 documents returned by the document retriever. For example, achieving a top-5 accuracy of 74.0% on the WebMD test Q&A dataset implies that for 74.0% of the questions in the

WebMD test Q&A dataset, the ground-truth answer appear in one of the top-5 documents returned by the document retriever. As explained in section III-B1, the extracted Q&A pairs from WebMD also contain a reference to the source article each answer is based on. With this, we could easily check whether the returned top-5 documents by the document retriever contained the ground-truth answer or not.

Dataset	Size	Top-5 accuracy (%)
SQuAD Dev [15]	10,570	77.8*
CuratedTREC [31]	694	86.0*
WebQuestions [32]	2,032	74.4*
WikiMovies [33]	9,952	70.3*
WebMD Test Q&A Dataset	3,572	74.0

TABLE II
PERFORMANCE OF THE DOCUMENT RETRIEVER ON WEBMD TEST Q&A DATASET COMPARED TO OTHER FACTOID Q&A DATASETS, IN TERMS OF THE % OF QUESTIONS FOR WHICH THE GROUND TRUTH ANSWER APPEAR IN ONE OF THE TOP 5 DOCUMENTS RETURNED BY THE DOCUMENT RETRIEVER *VALUES ARE OBTAINED FROM CHEN ET AL. [11].

Table III shows the performance of the document reader in detecting the correct answer span given the relevant source document for each question in the WebMD test Q&A dataset compared to that of SQuAD Dev and SQuAD Test factoid Q&A datasets on which performance of the document reader of the original DrQA system was tested. The performance was measured in terms of the F_1-score. For each question-answer pair in the WebMD test Q&A dataset, the question as well as the source article the answer is based on were provided to the document reader and the answer span returned was checked to have a Jaccard similarity greater than 0.5 with the original answer. If the Jaccard similarity with the original answer was greater than 0.5, it was counted as correct. This approach was followed because answers for questions in the WebMD dataset are long and often span across several paragraphs of text, which are highly unlikely to be exactly detected correctly by the document reader. In contrast, to compute the F1-score on SQuAD Dev and SQuAD Test datasets, the authors of DrQA check whether the answers returned by the document reader exactly match the ground-truth answers, since the answers to those questions are usually only one or two tokens long and are comparatively easier for the document reader to detect correctly.

Dataset	Size	Accuracy (F_1-score)
SQuAD Dev [15]	10,570	78.8*
SQuAD Test [15]	694	79.0*
WebMD Test Q&A Dataset	3,572	75.6

TABLE III
PERFORMANCE OF THE DOCUMENT READER ON THE WEBMD TEST Q&A DATASET COMPARED TO SQUAD DEV AND SQUAD TEST FACTOID Q&A DATASETS, IN TERMS OF THE F_1-SCORE. *VALUES ARE OBTAINED FROM CHEN ET AL. [11].

We also tested the case if the user rephrase his/her information need differently, whether the system is still capable of giving contextually similar answers. To test this capability, we compare the top-5 answers given by the system for a certain query with the top-5 answers given for the same query rephrased in a different manner. The results obtained for one of the edge cases that we tested manually is given in Table V.

B. Case Study

In Table IV, we present an example conversation carried out with EMA to give readers a taste of the current capabilities of EMA including query disambiguation and provision of direct and succinct answers for simple consumer health questions.

Case 1	The user does a spelling mistake
Agent:	*Hi, this is Ema. What is your medical problem?*
User:	*Hi ema. What is Efexor?*
Agent:	*By 'Efexor', do you mean 'Effexor'?*
User:	*yeah.*
Agent:	*Venlafaxine (Effexor and Effexor XR) is a newer antidepressant that boosts the levels of norepinephrine and serotonin in the brain. It helps improve mood and concentration. Did you find what you were looking for?*
User:	*yes.*
Agent:	***Oh yay, I'm so happy!!!*** *Any other questions?*
Case 2	The user query contains an ambiguous medical term
User:	*How to prevent cold?*
Agent:	*By 'cold', do you mean 'common cold' or 'upper respiratory infections'?*
User:	*It's common cold.*
Agent:	*Your best protection from common cold and flu is frequent hand washing. The simple friction that occurs when you rub skin against skin while using warm water and soap followed by thorough rinsing and drying can get rid of most potentially harmful germs. Did I answer your question right?*
User:	*nope.*
Agent:	***Oh I'm so sorry.*** *Please rephrase your question or ask any other medical question you have.*
User:	*It's okay. goodbye.*
Agent:	*Have a nice day.*

TABLE IV
AN EXAMPLE CONVERSATION WITH EMA, WHICH SHOWS HOW EMA VERIFIES SPELLING MISTAKES RELATED TO MEDICAL TERMS AND CLARIFIES AMBIGUITIES BEFORE ANSWERING THE QUESTIONS. THE TEXT HIGHLIGHTED IN BOLD ARE AFFECTIVE RESPONSES GENERATED USING HARD-CODED TEMPLATES.

V. DISCUSSION

According to the results in Tables II, the top-5 accuracy of 74.0% achieved by the document retriever on the WebMD test Q&A dataset is comparable to the top-5 accuracy obtained on rest of the factoid Q&A datasets. This implies that the longer and more complex nature of consumer health questions, which may also contain a lot of unnecessary words, does not affect the performance the document retriever in identifying the documents, which contain the correct answer to the question. According to the results in Table III, it can be seen that even though the answers for consumer health questions span across several paragraphs of text, the document reader still returns answers that are reasonably close to the ground-truths in comparison to the SQuAD dev and SQuAD test datasets whose answers are short and often limited to one or two words. This shows that the DrQA system can be extended to answer non-factoid consumer health questions, without any significant drop in performance. It also hints that the same approach can easily be extended to other domains such as restaurant recommendation just by training the model on

Question: How to prevent acne?

Rank	Answer	Answer score	Document score
1	Almost everyone has had at least mild acne at some point. Acne is the most common of all skin disorders, affecting some 40-50 million Americans. A generation or so ago, it was thought that eating too many sweet or greasy foods caused acne; now doctors know much more about why breakouts happen and how to treat	0.71563	39.413
2	often, acne will go away on its own at the end of puberty, but some people still struggle with acne in adulthood. Almost all acne can be successfully treated, however. It's a matter of finding the right treatment	0.46699	39.413
3	over-the-counter acne products don't clear your teen's acne, it may be time to consider prescription treatments. Here's acne treatment information for teens and their parents	0.25619	39.919
4	over-the-counter treatments, like a topical retinoid or products containing benzoyl peroxide or salicylic acid, along with gentle cleansing, don't work for you (give them a chance – it can take 4-12 weeks to improve acne), a dermatologist may be able to help. Severe acne requires aggressive treatment to prevent scarring	0.23263	39.413
5	Clean your face gently, as trauma to the acne breakouts may worsen the acne or cause scarring. When washing your face, use your hands or cotton pads, as any terrycloth or other scrubbing material may cause acne sores to rupture	0.18638	36.098

Question: I wish my facial skin was clear.

Rank	Answer	Answer score	Document score
1	Keep your hair off your face, neck, or other pimple-prone areas. Oil and dirt from your hair can block pores. If you want to wear your hair longer, make sure you keep it clean	41.369	50.618
2	But it's not always easy to pinpoint a specific cause. For example, your eyelids may be chronically dry, red and flaky, but what's to blame: your eyeshadow, eyeliner, makeup remover, or overnight eye cream	31.873	63.855
3	First off, to reduce scarring, never squeeze zits. If a pimple does leave a red or brownish mark it should eventually go away on its own. Be patient – it could take a year or more	13.629	50.618
4	Have a dermatologist check your skin. That's the best way to find out if you have sensitive skin or whether something else is causing your skin condition	13.201	39.805
5	Once you've identified an offending substance, avoid it. Wear gloves or protective clothing to prevent exposing your skin to cleansers, weeds, and other substances during housework or yard work. If your skin makes contact, wash the substance off right away with soap and water	10.887	63.855

TABLE V
TOP-5 ANSWERS GIVEN BY THE DOCUMENT READER FOR THE SAME QUERY PHRASED IN TWO DIFFERENT WAYS.

a domain-specific Q&A dataset and plugging in a domain-specific knowledge source (e.g. restaurant reviews). We also could have compared the performance of the document retriever and document reader against other non-factoid Q&A datasets from other domains such as restaurant booking, to evaluate whether the gap between the colloquial terms used in consumer health questions and the technical terms present in medical knowledge sources used to answer those questions, has as an effect on the overall performance of the system. In place of the Jaccard similarity measure we made use of to detect whether the answers returned by the system were reasonably close to the ground-truths, other measures such as the cosine similarity between the answer embeddings could have been employeed.

According to Table V, it can observed that when the query "How to prevent acne?" is reformulated as "I wish my facial skin was clear.", 2 out of the top-5 answers returned were contextually relevant (answers ranked 1 and 3) for "prevention of acne". Even the other 3 answers can also be considered as contextually relevant to the question asked. This capability of EMA is encouraging since one of the most challenging concerns with respect to consumer health question answering is that most of the time consumers tend to phrase medical queries with the colloquial language used in day today life, which may not contain the correct medical terms to refer to certain conditions (e.g. "acne" in the above case).

The case study presented in Table IV reflects how EMA maintains dialogue, answering questions while interactively conversing with the user to resolve any ambiguity detected. We can see that EMA does show some affective behaviour when an affirmation or a denial is received as response for an answer provided. This is achieved using hard-coded response templates that we have defined in the Rasa core. However, one limitation of using such approach to inject affective behavior to dialogue agents is that even though we can create multiple phrases with similar, but slightly different messages, they can soon become "expected" over the course of interaction. Therefore, it is best to use machine learning based generative approaches to generate affect-aware responses so that the responses are more varying and unpredictable.

VI. Future Work

As future work, we will extend the capabilities of EMA to carry out open-domain conversations or chitchat with the user, while responding in an empathetic manner. To achieve this, we intend to experiment on augmenting seq-to-seq neural conversations models [34] and the recently proposed transformer networks [35] with affective information to come up with our own dialogue model to generate affect-aware responses.

One of the major limitations that stands in our way is the lack of large scale doctor-patient or consumer health dialogue datasets that hinders adopting such data-driven approaches to the domain of consumer health. To address this concern, we have crawled Twitter to curate a multi-turn conversation dataset containing dialogues that convey empathy and are related to the healthcare domain. We used query phrases: "sorry to hear"; "get well soon"; "get better soon"; "speedy recovery"; "fast recovery"; "quick recovery"; "feel better soon"; and "feel good soon" to obtain Twitter conversations having at least one of the above key phrases indicating empathy. We were able to curate a total of 155,454 conversations having altogether 407,357 number of turns. The average number of turns per conversation is 2.6. To compliment this, we also use a dataset of review-response pairs extracted from

TripAdvisor, which consists of 14,504 single-turn dialogues conveying empathy and assurance towards negative comments made by users.

Since, altogether the curated data is still not enough to train a neural conversational model from scratch, we hope to utilize data from the OpenSubtitles dataset [36], which consists of over 1 billion open-domain conversations from movies, to pre-train a model and fine-tune it on the empathetic Twitter conversations and the TripAdvisor datasets using transfer learning. Another idea is to use the pre-trained BERT language model with transformer networks and fine-tune it on the target datasets. These approaches will help the model to perform better at the target task with little risk of overfitting.

To embed affective information in the neural conversational model, we hope to experiment with different affective embeddings such as Valence-Arousal-Dissonance (VAD) [26] based word vectors and DeepMoji [37] embeddings trained on a large Twitter dataset containing over 1.6 billion tweets annotated with 64 different emojis types. Further to increase the diversity of the responses generated, we hope to incorporate affectively diverse decoding strategies such as the one introduced by Asghar et al. [25]. Other mechanisms such as affective attention [38] and loss functions that incorporate affective objectives will also be experimented to improve model performance.

VII. Conclusion

In this report, we discussed in detail the design and architecture of EMA, an empathetic dialogue agent for answering consumer health questions. We showed that the machine reading comprehension based DrQA system used to answer open-domain factoid questions can easily be extended to answer restricted-domain non-factoid questions, the answers of which span across several paragraphs of text. We also discussed how tools such as MetaMap can be utilized to disambiguate the syntactic and semantic ambiguities present in consumer health questions. As future work, we discussed our plan in embedding empathetic response generation capabilities in EMA using neural conversational approaches. Altogether EMA has the potential to succinctly answer consumer health questions in an empathetic and human-like manner nearing to the behaviour of a real medical agent.

References

[1] D. Wang, and E. Nyberg, "Cmu oaqa at trec 2017 liveqa: A neural dual entailment approach for question paraphrase identification", In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC*, pp. 15-17, 2017.

[2] W. An, Q. Chen, W. Tao, J. Zhang, J. Yu, Y. Yang, Q. Hu, L. He and B. Li, "Ecnu at 2017 liveqa track: Learning question similarity with adapted long short-term memory networks", In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC*, 2017.

[3] V. Datla, T. Arora, J. Liu, V. Adduru, S.A. Hasan, K. Lee, A. Qadir, Y. Ling, A. Prakash and O. Farri, "Open domain real-time question answering based on asynchronous multiperspective context-driven retrieval and neural paraphrasing", In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC*, pp. 15-17, 2017.

[4] Y. Yang, J. Yu, Y. Hu, X. Xu and E. Nyberg, "Cmu livemedqa at trec 2017 liveqa: A consumer health question answering system", *arXiv preprint arXiv:1711.05789*, 2017.

[5] A.B. Abacha, E. Agichtein, Y. Pinter and D. Demner-Fushman, "Overview of the Medical Question Answering Task at TREC 2017 LiveQA", In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC*, pp. 15-17, 2017.

[6] W. Wong, J. Thangarajah and L. Padgham, "Contextual question answering for the health domain", *Journal of the American Society for Information Science and Technology*, vol. 63, no. 11, pp. 2313-2327, 2012.

[7] S. Sundararajan and V. Gopichandran, "Emotional intelligence among medical students: a mixed methods study from Chennai, India", *BMC medical education*, vol. 18, no. 1, pp. 97, 2018.

[8] D. McDuff and M. Czerwinski, "Designing Emotionally Sentient Agents", *Commun. ACM*, vol. 61, no. 12, pp. 74-83, 2018.

[9] D. Demner-Fushman, "Clinical, Consumer Health, and Visual Question Answering", In *Annual International Symposium on Information Management and Big Data*, pp. 1-6, 2018.

[10] A.B. Abacha and D. Demner-Fushman, "A Question-Entailment Approach to Question Answering", *arXiv preprint arXiv:1901.08079*, 2019.

[11] D. Chen, A. Fisch, J. Weston and A. Bordes, "Reading wikipedia to answer open-domain questions", *arXiv preprint arXiv:1704.00051*, 2017.

[12] H.Y. Huang, E. Choi and W.T. Yih, "Flowqa: Grasping flow in history for conversational machine comprehension", *arXiv preprint arXiv:1810.06683*, 2018.

[13] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", In *International Conference on Learning Representations (ICLR)*, 2015.

[14] A. Graves, G. Wayne and I. Danihelka, "Neural turing machines", *arXiv preprint arXiv:1410.5401*, 2014.

[15] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text", In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[16] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher and H. DaumeIII, "A neural network for factoid question answering over paragraphs", In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633-644, 2014.

[17] S. Reddy, D. Chen and C.D. Manning, "Coqa: A conversational question answering challenge", *arXiv preprint arXiv:1808.07042*, 2018.

[18] E. Choi, H. He, M. Iyyer, M. Yatskar, W.T. Yih, Y. Choi, P. Liang and L. Zettlemoyer, "Quac: Question answering in context", *arXiv preprint arXiv:1808.07036*, 2018.

[19] H.D. Tolentino, M.D. Matters, W. Walop, B. Law, W. Tong, F. Liu, P. Fontelo, K. Kohl and D.C. Payne, "A UMLS-based spell checker for natural language processing in vaccine safety", *BMC medical informatics and decision making*, vol. 7, no. 1, pp. 3, 2007.

[20] H. Kilicoglu, M. Fiszman, K. Roberts and D. Demner-Fushman, "An ensemble method for spelling correction in consumer health questions", In *AMIA Annual Symposium Proceedings*, vol. 2015, pp. 727, 2015.

[21] A.J. Yepes and A.R. Aronson, "Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation", In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pp. 733-736, 2012.

[22] M. Stevenson and Y. Guo, "Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus", *Journal of biomedical informatics* vol. 43, no. 5, pp.762-773, 2010.

[23] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman and G.O. Barnett, "The unified medical language system: an informatics research collaboration", *Journal of the American Medical Informatics Association*, vol. 5, no. 1, pp. 1-11, 1998.

[24] X. Zhou and W.Y. Wang, "Mojitalk: Generating emotional responses at scale", *arXiv preprint arXiv:1711.04090*, 2017.

[25] N. Asghar, P. Poupart, J. Hoey, X. Jiang and L. Mou, "Affective neural response generation", In *European Conference on Information Retrieval*, Springer, Cham, pp. 154-166, 2018.

[26] A.B. Warriner, V. Kuperman and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas", *Behavior Research Methods*, vol. 45, no. 4, pp. 1191-1207, 2013.

[27] "Rasa: Open source conversational AI", Rasa, 2019. [Online]. Available: https://rasa.com. [Accessed: 27- Jun- 2019].

[28] A.R. Aronson and F.M. Lang, "An overview of MetaMap: historical perspective and recent advances", *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229-236, 2010.

[29] K. Weinberger, A. Dasgupta, J. Langford, A. Smola and J. Attenberg, "Feature hashing for large scale multitask learning, In *International Conference on Machine Learning (ICML)*, pp. 1113-1120, 2009.

[30] L.Y. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes and J. Weston, "Starspace: Embed all the things!", In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[31] P. Baudis and J. Sedivy, "The question answering task in the YodaQA system", In *International Conference of the Cross- Language Evaluation Forum for European Languages*, pp. 222-228, 2015.

[32] J. Berant, A. Chou, R. Frostig and P. Liang, “Semantic parsing on freebase from question-answer pairs”, In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1533-1544, 2013.

[33] A.H. Miller, A. Fisch, J. Dodge, A.H. Karimi, A. Bordes and J. Weston, “Key-value memory networks for directly reading documents”, In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1400-1409, 2016.

[34] O. Vinyals and Q. Le, “A neural conversational model”, in *Proceedings of the 31st International Conference on Machine Learning*, vol. 37, 2015.

[35] T. Wolf, V. Sanh, J. Chaumond and C. Delangue, “Transfertransfo: A transfer learning approach for neural network based conversational agents”, arXiv preprint arXiv:1901.08149, 2019.

[36] J. Tiedemann, “News from OPUS - A collection of multilingual parallel corpora with tools and interfaces” In *Recent Advances in Natural Language Processing*, vol. 5, pp. 237-248, 2009.

[37] B. Felbo, A. Mislove, A. Sgaard, I. Rahwan and S. Lehmann, “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615-1625, 2017.

[38] P. Zhong, D. Wang and C. Miao, “An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss”, arXiv preprint arXiv:1811.07078, 2018.