Draft: A Medical Conversational Agent

Serena Yeung

1 Introduction

In this project, we endeavor to build a conversational agent capable of providing assistance to MedWhat users on medical topics. As an example, a user may begin a conversation by asking the question, *"Can I drink alcohol with my malaria pills?"* The agent could then ask a series of followup questions, such as *"What is the name of your malaria pills?"*, and use the acquired information to answer: *"I would not recommend drinking alcohol with Malarone pills. The pills may cause dizziness which can worsen with alcohol."*

There are two key technical components to this project. The first is developing an agent capable of conversing fluently in natural language. Given a prompt such as *"Should I go see the doctor?"*, an appropriate conversational response would be *"Yes, you should go see the doctor."* An inppropriate response would be *"My name is Sam."* The second component is accurate and reliable question answering. In other words, it is important that all statements are not only conversationally appropriate, but also medically accurate. This can be addressed through a combination of careful agent design (Sec. 4), and adding a safety valve that all medical statements are self-contained, e.g. *"If you have taken Malarone and are experiencing severe dizziness, you should go see your doctor."*

Following, I first discuss related work. I then give a brief overview of the datasets available to MedWhat, and describe our approach to a reliable medical conversational agent.

2 Related Work

Recurrent neural networks (RNNs) are neural network models that operate over sequential data and have shown strong ability for modeling natural language [4, 6, 8, 9, 11, 13]. In the context of conversations and question-answering, RNN-based sequence to sequence models [11, 13] obtain state-of-the-art performance. An RNN is used to read an input sequence (e.g. a question) one word at a time, and encode the sequence into a state vector that compresses the semantic information. This semantic state is then used to initialize a second RNN that decodes the semantics to produce an appropriate response, again output one word at a time. The model is trained through reading many examples of question-answer or statement-response pairs. For our medical conversational agent, we can leverage these models to tackle our first technical challenge of fluent conversation. However, since the agent will learn to speak in the style and manner of the training examples, we will use medical conversation data collected and owned by MedWhat to train an agent with the unique character of a medical professional.

[1, 2, 3, 5, 10, 12, 14] have tackled the challenge of knowledge-based conversation and question answering. While this is still an active area of research, recent advancements include using attention and memory networks to reason over the semantics of questions and statements or facts [3, 5, 10]. We may incorporate these ideas to enhance our model. However, our goal in this project differs from many of these works which attempt to learn facts from reading, then reason about questions to answer them with as high accuracy as possible, i.e. reading comprehension. Instead, we introduce a novel model that addresses the real-world conversational medical agent setting where 100% accuracy is a constraint, not an objective. To achieve this, we formulate our model to use conversation with a user towards the objective of determining the matching (or most similar) medical question that we have a known, expert-provided answer for, and returning that answer with the corresponding to context to the user.

3 Data

MedWhat currently has 3 sources of data relevant to medical questioning:

- 1 million medical question-answer pairs: A dataset of 1 million layperson-generated questions, and corresponding medical expert answers, crawled from medical websites. We can attempt to match user questions to this database and provide corresponding precise answers. However, the answers are quite specific (so often, an exact match may not be found), and there is significant noise in the dataset. We will therefore use this as a reserve resource as needed (Sec. 4.2).
- Medical questions: A large set of medical questions, with no answers. This dataset does not provide medical knowledge that can be used to respond to questions, but it provides data for modeling the distribution of questions that the agent can expect from users.
- Medical articles: A large set of medical articles. This provides medical facts outside the context of question-answering. While these facts could potentially be used to answer user questions, the level of the language is quite advanced and requires a high level of reading comprehension. Utilizing this data in a high-accuracy manner is best left for future work, not the initial implementation of our agent.

Additionally, MedWhat is planning to collect a dataset of high tens to hundreds of thousands of medical conversations similar to what it can expect from users. With proper constraints on the manner of data collection (such as ending each conversation with self-contained facts or advice), this will be the most useful dataset for training our agent and will be used for the core conversational agent (Sec. 4.1).

4 Approach

Given the data available to MedWhat in Sec. 3, we will take the following approach to developing a conversational and medically accurate agent. We define a conversation episode as beginning with a question or statement from a human. The agent has access to a large database (e.g. hundreds of thousands) of medical facts and advice. The goal of the agent is then to converse with the human to decide the most useful fact or advice from its database, and provide that information back to the human.

4.1 Conversational Agent

Since MedWhat users will be allowed to converse in unconstrained natural language, we will use an artificial intelligence (specifically deep learning) approach to interpret users' words and reason on what they are asking about. The core of the agent will be a novel model that uses sequence to sequence learning to converse with users in the character of a medical professional, and an additional recurrent network to integrate acquired information from the conversation sequence. This additional semantic recurrent network takes as input the encoded semantic states from the conversation network, and uses it to inform both the next follow-up question to ask as well as the most useful medical fact for the user. At each step of the conversation (one time step of the recurrent network), a probability distribution is maintained over the facts in the database, and when the agent is sufficiently confident about what the user is asking, it can respond with the information most likely to be useful. Fig. 1 shows an overview of this conversational agent.

4.2 Reserve Question-Answer Matching

The conversational agent will be trained using tens to hundreds of thousands of curated medical conversations owned by MedWhat. In addition to this data, we also have access to approximately one million medical question-answer pairs without conversations, crawled from the web. While this data is not curated and therefore less informative (e.g. questions may be overly specific), we can still use it as a reserve source of potentially useful medical information in case the conversational model is not able to satisfy a particular user question using the smaller, curated database. Specifically, we can take two approaches to match a conversation with a user to the most likely question-answer pair in the reserve dataset.

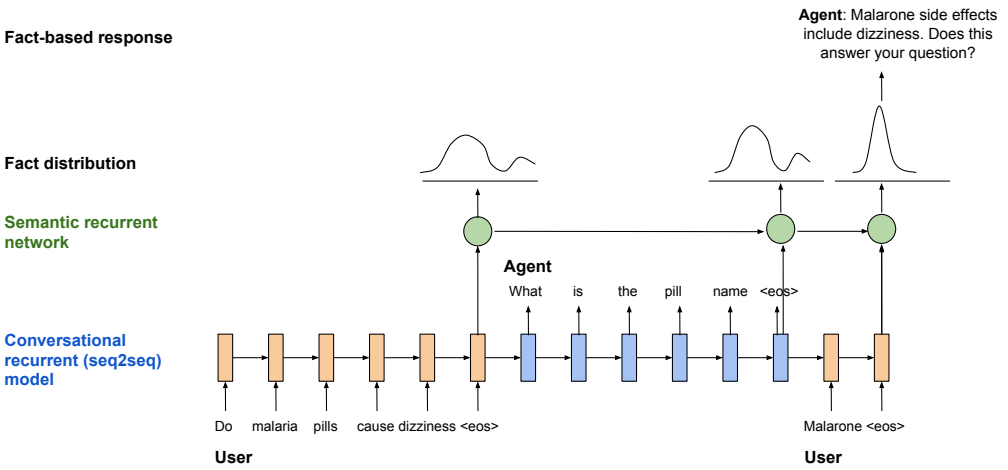

Figure 1: Core conversational agent. A sequence-to-sequence network is used to generate conversational responses to user text. An additional recurrent network takes the encoder states from the sequence-to-sequence network as input, and integrates the semantic information from these states to update its belief (a probability distribution) over the most useful expert-generated medical fact or advice to return to the user. When the network is sufficiently confident or a maximum conversational length has been reached, the corresponding, self-contained fact or advice is returned to the user. All returned statements are self-contained to ensure reliability. For example, "Malarone side effects include dizziness" is self-contained, but "Yes, that can cause dizziness" is not.

The first is question-question matching using unsupervised deep learning, in particular word vector embedding [7]. The word vector embedding approach looks at large streams of text (existing models have used Wikipedia, and we can additionally add medical text from our data sources), and uses this to learn an embedding space where words that are not necessarily the same but have similar meaning or context are closer together. We can map conversations as well as the one million reserve questions into this embedding space, choose the closest question to a user conversation, and return that question-answer pair as potentially useful information to the user.

The second approach is question-answer matching using supervised deep learning. Given our dataset of one million questions and matching answers, we can train a recurrent neural network-based model to predict the correct answer for a given question. In order to do this, the model must learn to interpret the semantic information in a question to match it with the most likely answer. Given a new, previously unseen question (or conversation), we can then use this model to retrieve the answer with the most similar semantic information, in other words likely to discuss the same medical topic. While this is not a strong enough guarantee that the answer is a correct medical response to the specific question the user is asking, returning the answer with the corresponding prompt question from the dataset may still provide useful information on the topic.

These two approaches of question-question matching and question-answer matching from the reserve dataset can be used to suggest similar questions with answers that the user may be interested in if the core conversational agent was not able to provide satisfactory assistance.

5 Plan

A rough implementation timeline is as follows:

- **May - July 2016**: MedWhat will collect the dataset of medical conversations using a data collection company. Meanwhile, Eric and Julian will work on getting initial results for the reserve question-answer matching. Eric has begun work on the question-question mapping approach, and Julian on the question-answer mapping approach.
- **Aug 2016 - Dec 2016**: The conversational data should now be available. Eric and Julian will work together on implementing the initial model. One person may focus on the sequence

to sequence conversational component, and the other on the RNN-based fact selection component.

- **Jan 2017 - on**: Iterate and push results on both the conversational agent and reserve question-answering components, and integrate the full system together.

References

[1] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, page 6, 2013.

[2] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *ACL (1)*, pages 1415–1425, 2014.

[3] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692, 2015.

[6] A. Karpathy, J. Johnson, and F.-F. Li. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[9] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[10] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439, 2015.

[11] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[12] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber, and P. Cimiano. Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648. ACM, 2012.

[13] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[14] M. Yahya, K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum. Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 379–390. Association for Computational Linguistics, 2012.