Bridging the Gap Between Consumers’ Medication Questions and Trusted Answers
Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan and Dina Demner-Fushman
Lister Hill National Center of Biomedical Communications
National Library of Medicine, Bethesda, MD

Abstract
This paper addresses the task of answering consumer health questions about medications. To better understand the challenge and needs in terms of methods and resources, we first introduce a gold standard corpus for Medication Question Answering created using real consumer questions. The gold standard consists of six hundred and seventy-four question-answer pairs with annotations of the question focus and type and the answer source. We first present the manual annotation and answering process. In the second part of this paper, we test the performance of recurrent and convolutional neural networks in question type identification and focus recognition. Finally, we discuss the research insights from both the dataset creation process and our experiments. This study provides new resources and experiments on answering consumers’ medication questions and discusses the limitations and directions for future research efforts.

Keywords: Health Informatics, Data Collection, Natural Language Processing

Introduction
Digital information and the World Wide Web make it possible to find information on demand, helping both patients and professionals to find relevant and valuable medical information. With these benefits come growing concerns about the potential misinformation of patients and non-expert consumers. Harmful consequences include self-diagnosis and unfounded anxiety about common symptoms after reviewing online health information (i.e., cyberchondria) [1]. The Pew Research Center reports that 72% of U.S. internet users (over 200 million) have gone online in the past year specifically for health-related information. Of that group, 77% say that their research started with search engines such as Google, while only 13% say they began at a specialized website such as WebMD [2]. Safe and efficient search alternatives are needed to protect non-expert consumers from misleading health information found online. Specialized Question Answering (QA) systems are a potential solution for the medical domain and the consumers’ need for reliable health information [2]. Several research efforts tackled the problem of QA in the (bio)medical domain and highlighted the challenges related to question understanding, answer retrieval, and ranking [3-6]. To the best of our knowledge, no studies were previously dedicated to QA about medications.

While several previous studies included such questions as a part of larger and more heterogeneous QA tasks, we argue that questions about medications have specific characteristics that need to be analyzed and evaluated in a focused study. These aspects include (i) different question types (e.g. action, storage, usage time, stopping, and tapering), (ii) more conditional answers (e.g. ingredients and appearance depend often on the manufacturer), (iii) different answer sources having a specific data structure (e.g. websites recommended by the FDA such as DailyMed [3]), and (iv) potentially more difficult questions (e.g. drug alternatives, questions about not formalized/written medical knowledge such as usage time), and a higher sensitivity to error (e.g. drug interactions that cannot be confirmed online).

Several efforts focused on methods for medical question answering [7-10] as well as the creation of relevant datasets. For instance, Kilicoglu et al. [11] introduced a dataset of 2,614 medical questions annotated with the question focus, type and triggers of the question types. Ben Abacha et al. [12] organized a medical QA task at the TREC 2017 LiveQA track and published training and testing datasets including consumer health questions received by the NLM, annotations and reference answers retrieved by medical informatics experts. The coverage of questions about medications in the above collections, however, was insufficient to train and develop efficient systems to answer questions about medications.
medications. In this paper, we present a study of consumers’ medication questions with the following contributions:

1. The development and publication of a manually annotated dataset of medication question-answer pairs based on real consumer questions submitted to MedlinePlus.
2. A synthesis of the manual annotation effort summarizing the insights obtained by a variety of experts from annotating and answering those questions.
3. New experiments using deep learning networks trained specifically for the tasks of identifying the main focus in the user’s question, and the question type.

In what follows, we first describe our annotation methodology and the baseline approaches in the Methods section. We present the statistics and characteristics of the developed dataset and the empirical results in the Results section.

Methods

In this section, we describe the guidelines used in the manual annotation process and our first empirical evaluation methods based on the dataset.

Data Creation

Selecting Consumer Questions about Drugs. We selected anonymized consumer questions submitted to MedlinePlus. We first performed Medical Entity Recognition using MetaMapLite [13]. We restricted the recognized entities to the following UMLS semantic types associated with medications: Antibiotic [antb], Clinical Drug [clnd], Neuroreactive Substance or Biogenic Amine [nsba], Pharmacologic Substance [phsu], Steroid [strd], and Vitamin [vita]. Finally, we manually selected the questions that (i) were deemed understandable and potentially answerable and (ii) have a drug name as focus. Figure 1 presents a word cloud of the most frequent terms in the selected consumer health questions.

Annotating the Questions. In a study conducted recently on consumer health question answering, Deardoff et al. [14] showed that for 62% of the questions, it was possible for librarians to find an answer in the top 5 search results in MedlinePlus using only the focus and question type. Given the importance of these two elements for QA, we, therefore, focused our efforts on manually annotating each question with:

- Question focus (always a Drug name in this dataset),
- Question type (e.g. Dose, Interaction, Side effects).

Searching for Reference Answers. For each answerable question, annotators had to retrieve manually a correct and complete reference answer (with its URL and section title).

Correct with regards to the question’s explicit and implicit information (e.g. a question about a drug in the UK, a specific form or dose), and extracted from reliable websites or scientific papers.

Complete with regards to all possible answers (e.g. all doses, ingredient lists from all manufacturers), and preferably written in a consumer-friendly language.

To select our answer sources, we followed the FDA’s guidelines [15] which prefer written questions and prioritizes the following sources in the order of priority:

2. Other NIH or U.S. government websites.
3. Other trustworthy websites (e.g., the Mayo Clinic) or academic institutions’ websites.
4. Other websites returned by a Google search.

Four annotators participated in the manual annotation and answering process. Then, a medical doctor and an expert in question answering reconciled the annotations of the question types and validated the retrieved answers.

Baseline Methods

Focus Recognition. We adapted, extended and evaluated Bi-directional Long Short-Term Memory (Bi-LSTM) networks on the task of recognizing the question focus (i.e., main drug name) according to the state-of-the-art architecture proposed by Xuezhe and Hovy [15]. The network includes a first Bi-LSTM network to build character-level embeddings, and a second Bi-LSTM taking as input both word embeddings built from the UMLS and pre-trained embeddings built with GloVe [16] and the character-level embeddings built during training with the first Bi-LSTM layer. The token labels were generated with a final Conditional Random Fields (CRF) layer. Our UMLS embeddings consist of binary vectors, where each word is tagged as the Beginning, Inside, Outside, End, or Single token (BIOES) of each semantic concept in the UMLS.

Question Type Identification. We implemented and evaluated a Convolutional Neural Network (CNN) on the task of identifying the question type (e.g., dosage, usage, contraindications). The input embeddings include UMLS BIOES embeddings and a randomly initialized vector of 128 dimensions updated with back-propagation during training.

Answer Retrieval. We conduct a first qualitative study on answer retrieval using twenty questions randomly selected from our dataset and the CHiQA question answering system. CHiQA is the first online medical QA system for consumer health questions from reliable sources such as NIH websites (e.g., MedlinePlus, GARD, NCI), Mayo Clinic and DailyMed. The system relies on different machine learning and knowledge-based methods to recognize the question’s focus and type [17-18] and uses the extracted information to retrieve answers with the Lucene search engine and a question-entailment recognition approach [19-20]. Figures 2 and 3 present the first answers returned by CHiQA to two questions selected randomly from the gold standard.

![Figure 2– First Answer Returned by CHiQA to the Question: “How to Stop Taking Bisoprolol?”](https://chiqa.nlm.nih.gov)
Results

Characteristics of the Dataset

The final gold standard contains 674 question-answer pairs with their associated annotations. These annotations include 25 question types, reported with examples in Table 1, and the answer sources, summarized in Figure 4. In particular, on the 674 answers, DailyMed was used to answer 290 questions, MedlinePlus for 128 questions, and other websites were used to answer 256 questions (e.g. cdc.gov, mayoclinic.org, health.harvard.edu, and PubMed abstracts and articles). Table 2 presents token-and-sentence-level statistics about the questions and the answers in the dataset.

Table 1 – Question Types in the Gold Standard

<table>
<thead>
<tr>
<th>Question Type</th>
<th>#</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>112</td>
<td>what type of drug is amphetamine?</td>
</tr>
<tr>
<td>Dose</td>
<td>70</td>
<td>what is a daily amount of prednisolone eye drops to take?</td>
</tr>
<tr>
<td>Usage</td>
<td>61</td>
<td>how to self inject enoxaparin sodium?</td>
</tr>
<tr>
<td>Side Effects</td>
<td>60</td>
<td>does benazapril aggravate hepatitis?</td>
</tr>
<tr>
<td>Indication</td>
<td>55</td>
<td>why is pyridostigmine prescribed?</td>
</tr>
<tr>
<td>Interaction</td>
<td>51</td>
<td>can i drink catalam when i drink medrol?</td>
</tr>
<tr>
<td>Action</td>
<td>39</td>
<td>how xarelto affects in the process of homeostasis?</td>
</tr>
<tr>
<td>Appearance</td>
<td>38</td>
<td>what color is 30mg prednisone?</td>
</tr>
<tr>
<td>Usage/time</td>
<td>36</td>
<td>when is the best time to take lositensin?</td>
</tr>
<tr>
<td>Stopping/tapering</td>
<td>31</td>
<td>how to come off citalopram?</td>
</tr>
<tr>
<td>Ingredient</td>
<td>28</td>
<td>what opioid is in the bupropion patch?</td>
</tr>
<tr>
<td>Action/time</td>
<td>23</td>
<td>how soon does losartan affect blood pressure?</td>
</tr>
<tr>
<td>Storage and disposal</td>
<td>13</td>
<td>in how much temp bce vaccine should store?</td>
</tr>
<tr>
<td>Comparison</td>
<td>11</td>
<td>why is losartan prescribed rather than a calcium channel blocker?</td>
</tr>
<tr>
<td>Contraindication</td>
<td>11</td>
<td>if i am allergic to sula can i take glipiide?</td>
</tr>
<tr>
<td>Overdose</td>
<td>10</td>
<td>what happens if your child ate a tylenol tablet?</td>
</tr>
<tr>
<td>Alternatives</td>
<td>8</td>
<td>what medicine besides statins lower cholesterol?</td>
</tr>
<tr>
<td>Usage/duration</td>
<td>7</td>
<td>how long should i take dutasteride?</td>
</tr>
</tbody>
</table>

Table 2 – Statistics about the Questions and Answers

Average number of tokens per question 7.16
Average number of tokens per answer 69.06
Average number of sentences per answer 3.23
Percentage of questions tokens present in answers 34.72%

Evaluation Results

We report the average performance on 5 runs and the variation range for focus recognition and question type identification.

Focus Recognition. The Bi-LSTM-CRF network was implemented with Python 3.5 and Tensorflow 1.4. We trained the network with a random subset containing 80% of the data, using 10% for hyperparameter tuning and the remaining 10% for the final test. All data were represented in the BIOES token annotation format. The results on the test set are summarized in Table 3. We used a learning rate of 0.01, a dropout of 0.5, a batch size of 40, and the Adam optimizer [21] to minimize the CRF-based loss function.

Table 3 – Bi-LSTM-CRF Results for Focus Recognition

<table>
<thead>
<tr>
<th>Results (%)</th>
<th>F1</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact entity match</td>
<td>74.07 [73.8-74.3]</td>
<td>78.12</td>
<td>70.42</td>
</tr>
<tr>
<td>Partial entity match</td>
<td>90.37 [89.7-90.9]</td>
<td>95.31</td>
<td>85.92</td>
</tr>
</tbody>
</table>

Question Type Identification. Our CNN network was also implemented using Python 3.5 and Tensorflow 1.4. Data were split according to same 80/10/10 percent subsets for training, development and test. To have enough training samples per class, we reduced the question types into the 14 most common types by aggregating the subtypes into their hypernym types: Information, Dose, Usage, Tapering, Interaction, Side effects, Indication, Action, Ingredient, Alternatives, Contraindication, Comparison, Manufacturing, and Appearance. We used a learning rate of 0.005, a dropout of 0.6, a batch size of 40, and the Adam optimizer to minimize the softmax-based loss function. The CNN network achieved an average accuracy of 75.7% on 5 runs with a variation in the [73.8, 77.6]% range. Despite using fixed random seeds for Tensorflow in both experiments, the relatively small size of the training data made them more

Figure 3 – First Answer Returned by CHiQA to the Question: “How to Self Inject Enoxaparin Sodium?”

Figure 4 – Websites Used to Answer the Drug Questions
susceptible to the non-deterministic implementation of GPU reduction operations and the Adam optimizer in TensorFlow\(^8\).

**Answer Retrieval.** In our qualitative study of the answers returned to 20 random questions from the dataset, CHiQA found the correct answer in the top four results in 35% of the cases, only related answers for 35% of them and irrelevant answers for the remaining 30%. While this limited evaluation must be taken with caution, our independent observations from the annotation process also hint that classical QA systems may not be the best fit for medication questions.

We discuss these insights and potential ways to improve answering questions about medications in the following section.

<table>
<thead>
<tr>
<th>ID</th>
<th>Question (Q) / Answer (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>(Q) “what does prednisone do to the body?” (A) [depends on the manufacturer] PINK</td>
</tr>
<tr>
<td>1b</td>
<td>(Q) “what would a normal dose be for valacyclovir?” (A) “Common limitations exist for screening benzodiazepines when using traditional immunoassay (IA) tests. IA testing for benzodiazepines often targets nordiazepam and oxazepam (…)” AND “Some commonly prescribed drugs have limited cross-reactivity (…)”</td>
</tr>
<tr>
<td>1c</td>
<td>(Q) “how much graval to kill you?” (A) “Pharmacological therapy can influence morbidity and mortality in severe chronic obstructive pulmonary disease (COPD). Long-term domiciliary oxygen therapy (LTOT) improves survival in COPD with chronic hypoxaemia. Oral steroid medication has been associated with improved survival in men and increased mortality in women, while inhaled steroid medication has been associated with a reduction in the exacerbation rate (…)”</td>
</tr>
<tr>
<td>1d</td>
<td>(Q) “time should take memantine?” (A) “Following a 10 mg oral dose of hydrocodone administered to five adult male subjects, the mean peak concentration was 23.6 ± 5.2 ng/mL. Maximum serum levels were achieved at 1.3 ± 0.3 hours and the half-life was determined to be 3.8 ± 0.3 hours.”</td>
</tr>
</tbody>
</table>

**Discussion**

**Difficulty and Ambiguity of Medication Questions**

In our annotation and answering efforts, questions about medications showed traits common to consumer health questions such as linguistic ambiguity due to misspellings, or wrong grammar leading to unclear meaning or multiple interpretations. For instance, in Table 4, example 1a, it is unclear whether the question is about the side effects or the action of “prednisone”. In addition, a distinct pattern emerged with many questions lacking essential information or context, making them too underspecified [22] to answer. For instance, in example 1b, finding a relevant answer about the right dose of “Valacyclovir” requires additional information about the patient and the condition or purpose of administration. Circumlocution [23] was another phenomenon that we encountered, i.e., the use of many words or general terms to express a medical term. For instance, in example 1c, the medical term “toxic dose” or “overdose” is replaced with the expression “how much X to kill you?”. In several cases, we also encountered a knowledge barrier when the requested medical information/knowledge was not formalized or written online (e.g. Table 4, 1d).

**Complexity of Manual Answer Retrieval**

Our annotators reported conditional answers frequently. These include answers that depend on the manufacturer, on the disease, or on patient information. For instance, in Table 4, 2a, four different answers are possible according to the manufacturer of phenytoin. Distributed answers were also encountered in several cases. These are answers that can be formed only by combining different text snippets from different answers and/or sources (e.g., Table 4, 2b). Many answers also required an understanding of expert terminology. These include answers that need to be translated to consumer-friendly language and questions that required rephrasing to expert language in order to find relevant answers (e.g., Table 4, 2c). Other answers could not be found without expert inference based on background knowledge (e.g., Table 4, 2d). External resources like eHealthMe were needed for specific types of questions such as Interaction questions. Answering some of the questions was also time consuming even for medical experts. For example, an hour was necessary to answer the question “is it alright to use fluticasone when using oxygen?” from PubMed.

**Challenges of Automating Medication QA**

Prior to our study, the lack of gold standard datasets for medication QA was the major bottleneck in automatic QA for drug questions. This new dataset opens new opportunities for both qualitative studies and quantitative evaluation of QA systems. In addition, systems relying on big training data can use it both (i) as a development set for fine-tuning the hyperparameters and testing different architectures and (ii) as a test benchmark.

In question understanding, our baseline networks achieved an encouraging performance despite the limited training data, with (i) 74% F1 score in question focus recognition for exact span matching and 90% for partial span matching, and (ii) 75.7% accuracy in identifying the question type. Several improvements can be considered for future developments, such as a richer set of embeddings, or a relevant language model. A more fine-grained adaptation of the UMLS can also be applied by restricting to the list of relevant semantic types either through pre-filtering or through trainable masks.

In answer retrieval, insights from both our annotation process and our CHiQA-based evaluation provide additional guidance on the relevant solutions and ways of improvement for medication question answering. In particular:

- Medical text translation [24] and simplification [25] are often needed to find relevant answers and make the retrieved answers readable for non-expert users.
- More data and resources are needed to cover information about drug interactions and usage guidelines. Such information can be extracted from both the scientific literature and clinical sources [26].
- Conditional answers require different solutions such as providing a list of answers or interacting with the user in a dialogue-based approach.

\(^8\) [https://www.twosigma.com/insights/article/a-workaround-for-non-determinism-in-tensorflow](https://www.twosigma.com/insights/article/a-workaround-for-non-determinism-in-tensorflow)
Due to the frequency of these conditional answers, fully unsupervised approaches are less likely to succeed than approaches based on (advanced) inference methods that can rely on explicit context semantics.

Conclusions

We studied consumers’ questions about medications. We created a new gold standard corpus for question answering about drugs that we shared in the scope of this paper. We presented statistics, insights and conclusions based on the manual annotation process, deep learning experiments, and preliminary evaluation of automatic answer retrieval. We hope that this new benchmark and initial experiments will foster new approaches and additional community efforts in addressing the growing need for reliable information about medications online.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, NIH.

References


Address for Correspondence

Asma Ben Abacha, PhD, asma.benabacha@nih.gov
Research Scientist
U.S. National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894, USA.