Lister Hill National Center for
Biomedical Communications
An Intramural Research Division of the U.S. National Library of Medicine

A Report to the Board of Scientific Counselors September 2015

Consumer Health Question Answering to Automatically Support NLM Customer Services

Dina Demner-Fushman, MD, PhD
Communications Engineering Branch

Halil Kilicoglu, PhD, CgSB
Kirk Roberts, PhD, CEB
Katherine Masterton, MLIS, NLM/LO/PSD
Ariel Deardorff, MLIS, NLM/LO/OD

Team Members:
Ilya Zavorin, PhD, CEB
Asma Ben Abacha, PhD, CEB
Yassine Mrabet, PhD, CEB

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Contents

Background

The U.S. National Library of Medicine (NLM) receives over 100,000 requests a year from individuals and organizations in the United States and abroad. About half of the requests are for inter-library loans; the second largest category, with over 10,000 requests a year on average, are requests to correct PubMed citations and the third largest category are consumer health questions and questions about drug action mechanisms, indications and side-effects (see Table 1). While NLM staff does their best to reply to all questions in a timely manner, the sheer volume means that response time is somewhat delayed; electronic correspondence (via direct email or the web form) is answered within four business days, while mailed and faxed requests are answered within seven business days.

Table 1 Customer requests received by NLM in 2014

Request Category (top 10 out of 50+)	Requests	% of Total Requests
Document Delivery/ILL	48,662	48%
Reference Questions	7,990	8%
PubMed Correction Requests	6,855	7%
PubMed	5,946	6%
MEDLINEplus Spanish	4,221	4%
Clinicaltrials.gov	4,059	4%
Drug/Product Questions	2,937	3%
MEDLINEplus	2,578	3%
UMLS	2,005	2%
Indexing	1,176	1%

In 2012, given the growth of customer questions and advances in natural language processing research, the Director of the NLM, Dr. Donald A.B. Lindberg, launched a project to build a system that would assist the NLM staff in responding to customer requests. Such a tool would classify customer requests and automatically direct them to the appropriate area of the library so that they could be answered in a more efficient manner. As a research and development division of NLM, the Lister Hill National Center for Biomedical Communications (LHNCBC) was tasked with the development of such a tool. This initiative resulted in development of the Consumer Health Information Question Answering (CHIQA) system.

In this report, we briefly present the overall architecture of the CHIQA system and the results of its integration with the customer support services. We then discuss in depth our research on various aspects of automated question understanding and answering. This includes co-reference resolution, spelling correction, and question decomposition, as well as the answer generation methods needed for a consumer health question answering system.

Objectives

Our long-term objective is to develop an online system that automatically retrieves/generates answers to consumer health questions in real-time.
Our intermediate objectives include:

• Development of methods and publicly available tools that address specific aspects of consumer-health question understanding, such as co-reference resolution, spelling corrections, question decomposition and question frame extraction.

• Development and distribution of annotated collections of questions and question-answer pairs to facilitate question answering research.

• Development of answer extraction and ranking methods.

• Development of back-off strategies such as identification of the question focus and retrieval of web pages relevant to the focus.

Significance

The project's significance is threefold: 1) it facilitates the dissemination of consumer health information by supporting NLM customer services; 2) it advances the state-of-the-art in natural language processing; and 3) it has potential to provide insights into online health information seeking behavior and improve users' experience with online health information seeking.

Methods and Results

We chose iterative design for our highly modular system. This combination of modularity and quick prototyping allows us to refine specific modules while continuously supporting NLM customer services. The overall system architecture is schematically represented in Figure 1. The CHIQA system is inserted between various NLM forms for submitting requests and the pre-existing customer management system, Siebel[1]. The *Request Classifier* module classifies each incoming request as "Health Question", "PubMed Correction Request" or "Other". The "Other" requests, for example, suggestions to add content to MedlinePlus or Genetic Home Reference articles, are forwarded to Siebel directly, whereas the "PubMed Correction Request" and "Health Question" requests are directed to the corresponding CHIQA modules that process the requests, attach the answers and then forward the requests augmented with the answers to Siebel.

[1] Oracle's Siebel Customer Relationship Management System -- http://www.oracle.com/us/products/applications/siebel/overview/index.html

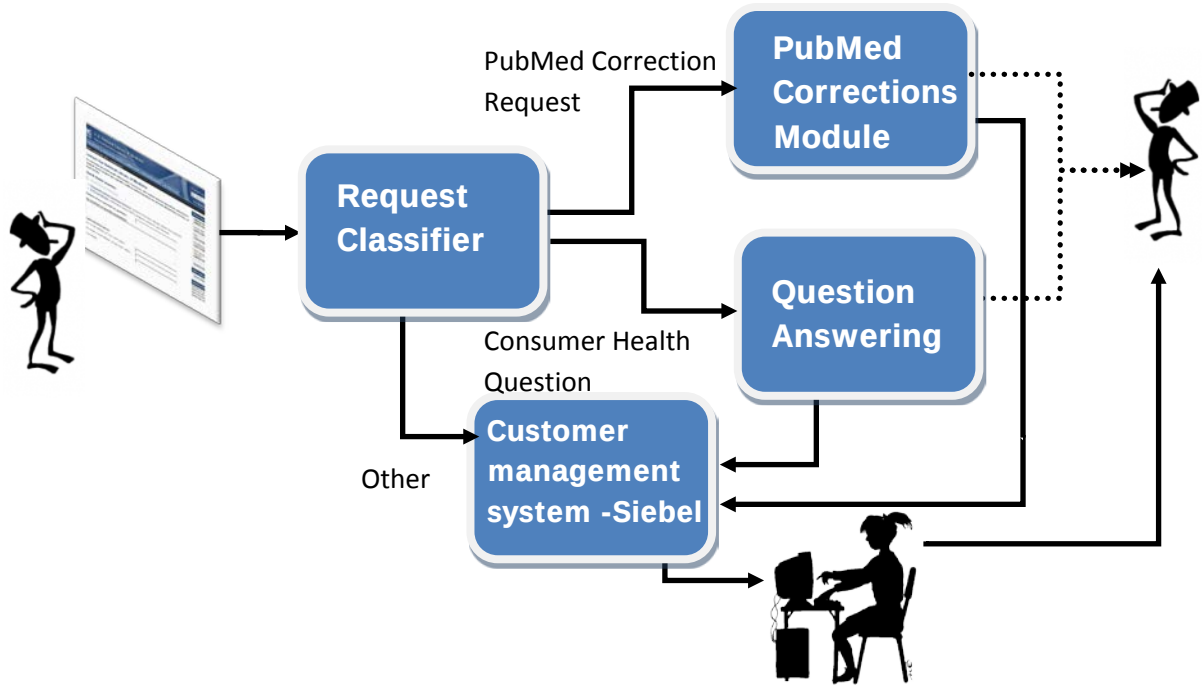

Figure 1 CHIQA process flow. Dashed arrows indicate future direct interactions with customers.

PubMed Citation Correction Requests as Proof of Concept

Because NLM receives a large number of requests about errors in PubMed, we started with the creation of a module to automatically classify and provide responses to these requests. A sample PubMed Citation Correction request is:

In the manuscript: Panminerva Med. 2010 Mar;52(1):67-78.
Renal dysfunction and heart failure. Escobar A, Echarri R, Barrios V.
There is a typographical mistake and Escobar A should be replaced by Escobar C.

An evaluation conducted in 2014 on the Siebel development system showed that the *Request Classifier* and the *PubMed Corrections Module* were ready to support the production system. These modules were deployed and are supporting production customer services since May 2014.

NLM staff have thus far provided manual judgments for the automatically prepared "PubMed Correction Request" answers using five categories:

OK	the request was classified correctly and the system generated a correct response
Error	the request was classified correctly, but the system generated a wrong response
Misfire	the request was misclassified (false positive), therefore the system generated a wrong response
Missed	the request was misclassified (false negative), therefore the system did not attempt to respond
Modified	supplementary information was added to the system response manually

According to the latest evaluation results for *PubMed Corrections Module* shown in Table 2, over 60% of PubMed Correction Requests are now addressed either completely automatically or with some modifications (OK + Modified). Moreover, our goal to emphasize precision over recall results in only 13% of the automatically generated responses (Misfire + Error) that need to be completely replaced.

Table 2 PubMed Corrections module performance evaluated by customer services staff in Q2 FY15 (January – March 2015).

OK	532	59%
Missed	219	24%
Misfire	90	10%
Modified	28	3%
Error	29	3%
Total	898	100%

The PubMed Corrections Module serves as a working proof of concept for the overall CHIQA project. We have successfully integrated our automated system into existing operational tools used by the NLM customer service staff. We hope to continue improving the PubMed Corrections Module, both through improved request processing and through improved data capture methods, such as use of a structured data form. In addition, given this successful proof of concept, we will continue to develop additional modules to classify and respond to additional types of requests, focusing primarily on consumer-health related requests.

Consumer Health Questions at NLM

Although consumers' health information needs are well-studied (primarily using search engine logs analysis) (McCray & Tse, 2003), consumer-health question answering (QA) is a relatively new area, with most of the work focusing on question analysis. Zhang (2010) analyzed health-related questions submitted to Yahoo! Answers and found that these questions primarily described diseases and symptoms (accompanied by some demographic information), were fairly long, dense (incorporating more than one question), and contained many abbreviations and misspellings. In our recent study conducted by Kirk Roberts (manuscript submitted to JAMIA), we set out to analyze if the consumer health questions submitted to NLM differ from those submitted to other sites and from those asked by professionals. In this section, we first present our results of consumer-health questions analysis. We then describe our question answering system that consists of a question understanding module and an answer retrieval and generation module.

Interactive use of online health resources: A comparison of consumer and professional questions

In addition to the collections of health-related questions submitted to NLM by general public and self-identified professionals, we obtained eight online question corpora—four consumer and four professional. We analyzed 40,000 questions using lexical-, syntactic-, and semantic-level natural language processing methods. The five health consumer corpora were: (1) Yahoo! Answers: questions under the category Diseases & Conditions (Surdeanu, 2008) (10,000 questions), (2) WebMD Community: forum to which consumers post questions resulting in conversations that differ in style from the community QA sites (10,000 questions). (3) Doctorspring: a curated QA website where consumers

submit questions to be answered by a health professional for a fee (811 questions). (4) Genetic and Rare Diseases Information Center (GARD): where consumers submit questions to be answered by NIH staff (1,467 questions). (5) NLM Consumer Health Questions (NLMC): questions about diseases, conditions, and therapies submitted to NLM as "General Public" requests (7,164 questions). The five health professional corpora were: (1) Parkhurst Exchange: a curated QA resource for physicians (5,290 questions). (2) Journal of Family Practice: curated questions targeted toward specific cases (601 questions). (3) Clinical Questions: point-of-care questions collected by Ely (1997, 1999) and D'Alessandro (2004) (4,654 questions). (4) PubMed on Tap: point-of-care questions posed during an evaluation of access to PubMed using handheld devices (Hauser, 2007) (521 questions). (5) NLM Professional Health Questions (NLMP): questions similar to NLMC, but the customers self-identify as a "Health Professional" or "Researcher/Scientist" (740 questions).

We found that consumers ask longer questions (37-106 tokens for consumers vs. 11-62 for professionals), though this appears to be primarily an effect of the resource. Among similar resources, the divide was smaller: question answering websites had some variation (37-100 vs. 11-36), while the NLM questions varied little (70 vs. 62). Similar effects can be seen with sentences, where most professional questions had one or two sentences (except NLMP), while consumer questions tended to have three or more (2.8-6.9). Word length was shorter for consumers (4.0-4.7 characters vs. 4.5-5.5), suggesting a less developed vocabulary that perhaps requires more words to describe an information need.

Consumer questions are more readable than professional questions based on the standard readability metrics. The fog index for consumers is lower (9.0-11.9 vs. 12.2-14.8), as is the grade level (6.2-9.2 vs. 9.2-11.9), implying less education is needed to comprehend the questions. Similarly, the reading ease is higher for consumers (49.6-75.4 vs. 32.3-49.9). Consumer-to-consumer websites tend to be the most readable, while professional-to-professional medical journals the least so. However, these readability metrics are primarily concerned with the length of words and sentences, not grammaticality. Human judgments might indicate the opposite, and thus these results indicate more accurate automatic readability metrics need to be devised by the NLP community. Consistent with our expectations, rates of misspelling are higher on consumer questions than on professional questions except for NLMP.

We also evaluated the probability of each corpus being generated by an open domain or medical language model and found that every consumer corpus was judged more probable by the open-domain model, and every professional corpus was judged more probable by the medical model.

To summarize, our results show that the consumers provide different amounts of background information and formulate the questions differently depending on the particular resource, i.e., questions submitted to NLM are distinct, however there are enough similarities in consumer-health questions posted to online resources that methods developed for understanding NLM questions should be useful in processing any consumer health questions.

Question Understanding

This module is a hybrid natural language processing system. It starts with three linguistic preprocessing steps: spelling correction, co-reference resolution, and question decomposition. Next, a knowledge-based module extracts structured question representations (frames) and a supervised machine learning-based module extracts the question focus and assigns the question type. The latter module serves as back-off strategy for questions that were classified as consumer-health questions in the classification step, but for which no frames were generated.

Spelling correction

Kilicoglu H, Fiszman M, Roberts K, Demner-Fushman D. An Ensemble Method for Spelling Correction in Consumer Health Questions. To be presented at AMIA Fall Symposium 2015.

Misspellings can severely hinder automatic question understanding. For example, consider the following question:

> *My mom is 82 years old suffering from anixity and depression for the last 10 years was dianosed early on set deminita 3 years ago. Do yall have a office in Greensboro NC? Can you recommend someone. she has seretona syndrome and nonething helps her.*

Four disorders are mentioned in the question, three of which are misspelled (*anixity* for *anxiety*, *deminita* for *dementia*, and *seretona syndrome* for *serotonin syndrome*). Only after these errors are corrected can our question understanding module properly extract the question frame or focus (theme).

For consumer health information queries, Crowell et al. (2004) used a MedlinePlus frequency-based technique to improve on existing spelling correction tools. Our experience with readily available spelling correction tools revealed that they were not accurate enough on consumer health questions. Therefore, we developed our own spelling correction module (Figure 2).

To develop and evaluate our spelling correction module, we manually corrected a corpus of 472 health-related questions posed to NLM by consumers. The corpus contains 25000 words, 1075 of which are misspelled. We categorized misspellings as follows:

- NON-WORD: the misspelled word does not appear in the dictionary (*physians* for *physicians*)
- REAL-WORD: the misspelled word appears in the dictionary (*leave* for *live*)
- PUNCTUATION: a misspelling error caused by absence of punctuation or a spurious punctuation (*Ive* for *I've*)
- TO-MERGE: a word-break error, where a spurious space is introduced to a word (*on set* for *onset*)
- TO-SPLIT: a word break error, where two adjacent words run together (*knowabout* for *know about*)

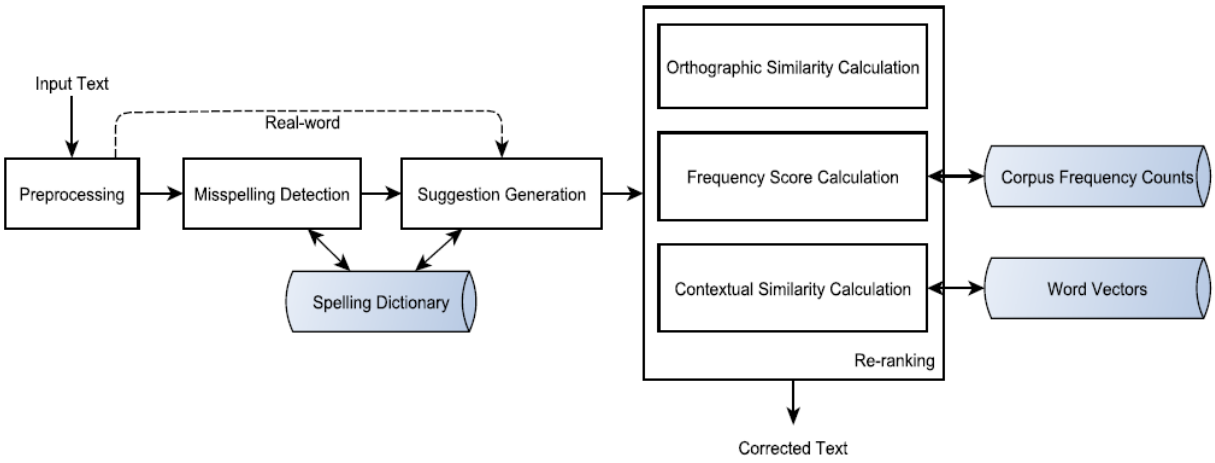

Figure 2 Spelling correction pipeline

In the first step of the error correction pipeline, the preprocessor corrects punctuation and splitting errors using simple rules. The second step detects misspellings using dictionary lookup in the

comprehensive English dictionary distributed with the Jazzy spell checker[2] expanded with tokens extracted from terms in UMLS. The third step generates phonetic and edit distance-based spelling suggestions. Next, the system scores all generated suggestions using orthographic and phonetic similarity between the misspelling and the suggestion, as well as the corpus frequency of the suggestion and its contextual similarity with the text surrounding the misspelling. The contextual similarity score is calculated using word embeddings (Mikolov et al, 2013), a technique based on the distributional hypothesis.

Table 3 Spelling correction evaluation. ALL stands for the best weighted combination of suggestions ranking methods

Method	*Precision*	*Recall*	*F1*
Non-word only			
ESpell (with filtering)	0.53	0.20	0.29
Preprocessing only	0.94	0.33	0.49
W/ Orthographic similarity	0.57	0.52	0.55
W/ Corpus frequency	0.45	0.41	0.43
W/ Context similarity	0.42	0.38	0.40
ALL	0.64	0.58	0.61
Real-word included			
ESpell	0.23	0.26	0.25
ALL	0.57	0.59	0.58
Important for focus/frame only			
ESpell	0.59	0.57	0.58
ALL	0.83	0.70	0.76

To evaluate spelling correction, we used 372 questions in the corpus for training and the remaining 100 for testing. The ESpell algorithm (Wilbur, 2009) used in the PubMed search engine served as strong baseline. The evaluation results shown in Table 3 indicate that our module can adequately suggest corrections important for finding question focus and extracting question frames; however, further improvements might be achieved in the future by using more specialized corpora (e.g., online health forums) as the basis of frequency counts and word embeddings. We are also planning to explore ways to better rank the similarity scores and explore joint spelling correction, since misspelling of adjacent words is common in consumer health questions.

Co-reference Resolution

Kilicoglu H, Fiszman M, Demner-Fushman D. Interpreting consumer health questions: The role of anaphora and ellipsis. In Proceedings of the 2013 ACL BioNLP workshop.

Co-reference resolution is critical in understanding consumer health questions. In a study to classify questions as posed by consumers or professionals, Liu *et al.* (2011) found a significantly higher use of personal pronouns in consumers' questions. Vicedo and Ferrández (2000) have shown that pronominal anaphora resolution improves several aspects of QA systems' performance. Our analysis of consumer health questions confirmed these findings and we developed a knowledge-based co-reference

[2] http://jazzy.sourceforge.net/

resolution module, mainly focusing on anaphora and ellipses. Other than our efforts, anaphora and ellipsis resolution has not been previously attempted in consumer health question understanding.

Our co-reference resolution module identifies two types of anaphoric phenomena: *pronominal anaphora* (including anaphora of personal and demonstrative pronouns) and *sortal anaphora* (anaphora indicated by noun phrases). In the following examples, anaphoric mentions are underlined and their antecedents are in bold:

- Personal pronominal anaphora: *My daughter has just been diagnosed with **Meier-Gorlin syndrome.** I would like to learn more about it ...*
- Demonstrative pronominal anaphora: *We just found out that our grandson has **48,XXYY syndrome**. ... I was wondering if you could give us some information on what to expect and the prognosis for this and ..*
- Sortal anaphora: *I have a 24-month-old niece who has the following symptoms of **Cohen syndrome**: ... I would like seek your help in learning more about this condition*.

The anaphoric expressions are resolved to the most salient terms in the request identified using rules, such as person and number agreement between the anaphoric mention and the antecedent.

Our co-reference resolution methodology is fairly independent of the question understanding process and can be used on various types of biomedical documents. On the other hand, ellipsis resolution is tightly coupled with question understanding; it is only triggered when the frame extraction module (discussed below) fails to identify a question focus while other elements of the frame are identified. Consider the following example:

My child has been diagnosed with pachgyria. What can I expect for my child's future?

Our system recognizes a question about prognosis in the second sentence (as we describe in the next sections), but fails to identify the disease in question or any references to previous mentions of a disease. With ellipsis resolution, we assign the most salient term in the request (*pachgyria*) as the question focus.

We evaluated the co-reference resolution module in the context of question frame extraction (described in the following sections) and found that anaphora/ellipsis resolution significantly improves recall and has a minor negative effect on precision of frame extraction, the overall effect being significantly positive (See Table 4.).

Question Decomposition

Roberts K, Masterton K, Kilicoglu H, Fiszman M, Demner-Fushman D. Annotating Question Decomposition on Complex Medical Questions. LREC 2014.

Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Decomposing Consumer Health Questions. BioNLP 2014.

As mentioned above, consumers pose multi-sentence, complex questions that contain background information and often more than one specific question. Consider the following customer request:

Will Fabry disease affect a transplanted kidney? Previous to the transplant the disease was being managed with an enzyme supplement. Will this need to be continued? What cautions or additional treatments are required to manage the disease with a transplanted kidney?

This request contains three question sentences and one background sentence. The focus (Fabry disease) is stated in the first question but is necessary for a full understanding of the other questions as well. The background sentence is necessary to understand the second question: the anaphor *this* must be resolved to an enzyme treatment. In the same second question, we need to re-construct from the discourse the implicit argument of the predicate *continue* (i.e., *continue after a kidney transplant*). The final question sentence uses a coordination to ask two separate questions (cautions and additional treatments). A decomposition of this complex question would then result in four questions:

1. *Will Fabry disease affect a transplanted kidney?*
2. *Will enzyme treatment for Fabry disease need to be continued after a kidney transplant?*
3. *What cautions are required to manage Fabry disease with a transplanted kidney?*
4. *What additional treatments are required to manage Fabry disease with a transplanted kidney?*

We use supervised machine learning to first assign the following labels to each sentence and its constituents:

- *BACKGROUND* - a sentence that provides useful contextual information, but lacks a question.
- *QUESTION* - a sentence or a clause that contains a question.
- *IGNORE* - a sentence containing nothing of value for understanding of the request.
- *COORDINATION* - a conjunction of two or more questions in a sentence pertaining to one disorder. For example, in "what is the prognosis and life expectancy for multiple sclerosis?" "prognosis and life expectancy" must be recognized as conjunction, whereas in "What is the life expectancy for multiple sclerosis and what can we expect from here?" must be recognized as two questions. This distinction is needed for the decomposition step.
- *EXEMPLIFICATION* - a phrase that introduces an aspect of a more general question, e.g., in "We are interested in learning more about this condition, including the standard course of treatment", the general information request about a condition is augmented with a specific question about the treatments.

First, an SVM categorizes sentences into BACKGROUND, QUESTION, or IGNORE. For QUESTION sentences, phrases containing a trigger word are ranked by an SVM classifier and then the top-ranked candidate is passed through a separate SVM filtering classifier. We evaluated the approach using cross-validation on 1,467 consumer health questions in the GARD collection that we manually annotated with the above labels. We achieved F-scores ranging from 97.5% for overall question recognition to 73.5% for coordination extraction allowing for inexact boundary matches. These results indicate we need to improve question decomposition, potentially through integration of co-reference and implicit argument information, and identification of discourse relations within requests.

Question Type Classification

Roberts K, Masterton K, Fiszman M, Kilicoglu H, Demner-Fushman D. Annotating Question Types for Consumer Health Questions. LREC 2014, BioTxtM workshop

Another important aspect of question understanding is detecting the question type. We proposed the following question types:

- *ANATOMY*: questions about a particular part of the body, such as the location affected by a disease.
- *CAUSE*: questions about the cause of a disease.
- *COMPLICATION*: questions about the problems a particular disease causes.
- *DIAGNOSIS*: questions about diagnostic tests, or methods for determining the difference between possible diagnoses (differential diagnosis).
- *INFORMATION*: requests for general information about a disease.
- *MANAGEMENT*: questions about the management, treatment, cure, or prevention of a disease.
- *MANIFESTATION*: questions about signs or symptoms of a disease
- *OTHER_EFFECT*: questions about the effects of a disease, excluding manifestations and complications.
- *PERSONORG*: questions asking for a person or organization involved with a disease.
- *PROGNOSIS*: questions about life expectancy, quality of life, or the probability of success of a given treatment.
- *SUSCEPTIBILITY*: questions asking how a disease is spread or distributed in a population. This includes inheritance patterns for genetic diseases and transmission patterns for infectious diseases.
- *OTHER*: Identifies disease questions that do not belong to the above types.
- *NOT_DISEASE*: questions that aren't about a disease and thus not yet handled by our system.

To automatically classify question types, we utilized a multi-class SVM trained on the described above manually annotated 1,467 GARD requests, with a total of 2,937 decomposed questions. The cross-validation results shown in Table 4 indicate that for some categories we did not have enough training examples.

Table 4 Question classification results by question type

Question Type	# Annotations	Precision	Recall	F1
Anatomy	12	66.7	16.7	26.7
Cause	119	83.0	78.2	80.5
Complication	32	65.4	53.1	58.6
Diagnosis	229	83.1	75.1	78.9
Information	520	86.3	93.7	89.9
Management	673	91.4	89.7	90.6
Manifestation	103	87.3	86.4	86.8
NotDisease	16	20.0	6.2	9.5
OtherEffect	275	64.7	66.5	65.6
Other	38	63.2	31.6	42.1
PersonOrg	128	87.1	78.9	82.8

Prognosis	313	78.9	79.9	79.4
Susceptibility	420	78.0	86.0	81.8

Question Focus Extraction

Roberts K, Masterton K, Kilicoglu H, Fiszman M, Demner-Fushman D. Annotating Question Decomposition on Complex Medical Questions. LREC 2014.

Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Decomposing Consumer Health Questions. BioNLP 2014.

Question focus extraction is not only essential for question decomposition described above, it is also our back-off strategy for the questions that the frame extraction module fails to parse. We use a 3-step process to extract the focus: (1) a high-recall method identifies potential focus diseases in a request, (2) an SVM ranks the candidates, and (3) the highest-ranking candidate's boundary (the extent of the focus in the question) is modified with a set of rules. To identify candidates for the focus, we use a lexicon constructed from UMLS. UMLS includes very generic terms, such as disease and cancer, which are too general for questions submitted to NLM. We allow these terms to be candidates so as to not miss any focus that does not exactly match an entry in UMLS. When such a general term is selected as the top-ranked focus, the rules described below are capable of expanding the term to the full disease name. To rank candidates, we utilize an SVM with a small number of feature types:

- Unigrams. Identifies generic words such as disease and syndrome that indicate good candidates
- UMLS semantic group
- UMLS semantic type
- Sentence Offset. The focus is typically in the first sentence, and is far more likely to be at the beginning of the request than the end.
- Lexicon Offset. The focus is typically the first disease mentioned.

As mentioned above, we use a number of heuristics to alter the boundary to a more usable focus after it is chosen by the SVM. The rules are applied iteratively to widen the focus boundary until it cannot be expanded any further. If a generic disease word is the only token in the focus, we add the token to the left. Conversely, if the token on the right is a generic disease word, it is added as well. If the word to the left is capitalized, it is safe to assume it is part of the disease's name and so it is added as well. Finally, several rules recognize the various ways in which a disease sub-type might be specified (e.g., *Behcet's syndrome vascular type, type 2 diabetes, Charcot-Marie-Tooth disease type 2C*). We evaluated focus recognition on the 1,467 consumer health questions in the GARD collection described above with both an exact match, where the gold-standard and automatic focus boundaries must line up perfectly, and a relaxed match, which only requires a partial overlap. We achieved 73.6% and 88.1% F-score for exact and relaxed matches respectively.

Question Frame Extraction

We use lexico-syntactic information to extract question frames from each sentence in a request classified as consumer health question. For each question, we aim to construct a frame which consists of the following elements: *type*, *theme*, *question cue, predicate, patient, agent, location and associated_with.* The theme (focus), type and predicate are required for all questions. For all question types, except for general information requests, the question cue is also required. All other frame elements are optional. *Type* refers to the expected answer type, i.e., an aspect of the theme that the question is about (treatment, prognosis, etc.) *Theme* refers to the topic of the question, i.e., its focus, and question cue to the question words (*what, how, are there*, etc.). Theme is restricted to concepts in

the UMLS semantic group Disorders because we are focusing on answering questions about known problems for now. From the question "Can drug remove growth on the neck?", the following frame should be extracted:

Type: *Management*
Theme: *Mass of neck [C0149736]*
Predicate: *remove*
Question cue: *can*
Agent: *drug*

Note that we prefer a more specific term *Mass of neck* to identifying *mass* as theme and *neck* as location. This preference gives us the benefit of automatic UMLS-based query expansion of the correct term when searching for answer candidates. We rely on syntactic dependency relations (e.g., nominal subject, direct object) to link frame indicators (predicates) to their themes and question cues. Two special rules address the following cases:

- If the dependency exists between a predicate of type T and another of type General Information, the General Information trigger becomes a question cue for the frame type T. This handles cases such as "*Is there information regarding prognosis?*"
- If a dependency exists between a trigger T and a patient term P and another between the patient term P and a potential theme argument A, the potential theme argument A is assigned as the theme of the frame indicated by T. This handles cases such as "*What is the life expectancy for a child with Dravet syndrome*?" whereby *Dravet syndrome* is assigned the Theme role for the Prognosis frame indicated by the predicate *life expectancy*.

We evaluated question frame extraction on 54 questions about genetic diseases submitted to NLM in 2012.

Table 5 Question frame extraction results

Frame extraction conditions	# of extracted frames	# of correct frames	Recall	Precision	F-score
A -- Rules with automatic named entity extraction	14	13	0.32	0.93	0.48
A + co-reference resolution	26	22	0.54	0.85	0.66
B -- Rules with gold standard named entities	17	16	0.39	0.84	0.55
B + co-reference resolution	35	33	0.80	0.94	0.86

As can be seen in Table 5, our lexico-semantic patterns have good precision and co-reference resolution significantly helps recall. However, automatic named entity recognition (NER) needs to be improved to achieve better recall. Partially, the existing UMLS-based NER tools fail to recognize themes because of the mismatch between the lay vocabulary for disorders and the UMLS: consumers often rely on lengthy descriptions with a reduced vocabulary when describing symptoms or conditions. We addressed one such frequent way of describing a disease, namely a combination of a generic symptom with a spatial location. We developed a supervised machine learning approach to extract such descriptions and subsequent normalization to UMLS. For this purpose, we generated an annotated corpus of 2,000

sentences with 1,300 spatial relations, and a second corpus of 500 of these relations manually normalized to UMLS concepts (Roberts et al. Automatic Extraction and Post-coordination of Spatial Relations in Consumer Language, to be presented at the AMIA Fall Symposium.) This approach allows us to, for example, map the description "tumors in her brain" to Brain Neoplasms [C0006118]. We plan to incorporate the spatial relations recognition module in CHIQA shortly.

Answer Generation

The research on automatic answering of consumer health questions is even sparser than question understanding research. Luo et al. (2015) use syntactic and semantic analysis to align a new question with the questions previously submitted to NetWellness -- a website through which highly qualified volunteers provide answers to consumer health questions (Marine et al, 2005). Once a question is matched, the corresponding answer can be retrieved from the database (Luo et al., 2015). Others have developed interfaces to help consumers find the right answers by clustering retrieval results (Mu, Ryu, and Lu, 2011) or combining health topics as dynamic and searchable menu items with keyword searches (Cui, Carter, and Zhang, 2014). Since we do not have a readily available database of answers and want to move beyond providing retrieval results, we are developing an approach to finding answers to health questions in online (primarily NLM and NIH) resources. To that end, we developed an approach based on providing sections of MedlinePlus articles for a given question frame.

Informally analyzing the results, we assumed that we might need additional resources, at least the ones directly linked to from MedlinePlus pages for a given disorder. We also wanted to see if we could replace our efforts on harvesting and indexing articles for subsequent answer extraction with issuing search queries to the existing MedlinePlus search engine. To answer these questions, we conducted a study described in the next section.

Answer Retrieval – MedlinePlus Experiments

In this study, we focused on manual retrieval of best possible answers using manually extracted ideal question focus and type. This study establishes an upper bound for an automatic system. To ensure the results of the study can be operationalized, seven reference librarians conducted protocol-driven searches using 300 manually annotated questions. The search protocol is shown below:

- Stage = A, search performed in MedlinePlus using the Focus as the search string
- Stage = B, search performed in MedlinePlus using the Focus and Type as the search string
- Stage = C, search performed in MedlinePlus using any terms from the customers message as the search string
- Stage = D, search performed in MedlinePlus using any terms
- Stage = E, search performed in Google using any terms

In addition to registering what parts of the question were needed to find an answer, we captured the location of the answer:

- Result grade = 1, answers were located within the MedlinePlus database (health topics or encyclopedia articles)
- Results grade = 2, answers were contained within pages linked from MedlinePlus (sites that came up in the MedlinePlus search engine but were from outside providers)

- Results grade = 3, answers were found in authoritative sites (e.g., those with Health on the Net approval seal) that were not linked to from MedlinePlus (and were found via Google)
- Results grade = 4, no authoritative answer could be found
- Results grade = 5, no answer could be found

The MedlinePlus searches were conducted using the MedlinePlus search engine and only the top five results were evaluated. The results of this study are very encouraging: 35% of the answers were located within MedlinePlus.gov and 42% of answers were found in articles that were linked from MedlinePlus, moreover, for 52% of the questions, the librarians used the focus alone to find an authoritative answer. This means that just by using the extracted focus as a search term in the MedlinePlus database a reference librarian was able to retrieve an authoritative answer within the first five search results. Combining the focus and the type led to answering an additional 9% of the requests. These results indicate that we do not need to maintain our own search engine described in the next section and can replace it with search requests to the MedlinePlus search engine, focusing our efforts on answer generation.

Automatic Answer Retrieval and Extraction

Currently, CHIQA uses the Essie search engine developed for ClinicalTrials.gov. We harvested MedlinePlus (MPlus) articles, Genetic Home Reference (GHR) articles and articles in Gene Reviews (GR), as initially our focus was on answering questions about genetic disorders. We developed the following search strategy for translating question frames to searches and extracting answers from these resources: first, the theme or its UMLS synonyms must be found in the title of the document or in the alternative terms for the title disease provided in the "Alternative Name" section of MedlinePlus or its equivalents in other resources. Then, the most appropriate section for a given question type is extracted (see Table 6). Next, the sections are ranked by their predefined weight and the occurrence of search terms and answer type indicators (e.g., life expectancy, quality of life, prognosis, etc. for Prognosis) in the section. The top ranked section is provided as the answer. If the section is too long, the parts that do not contain search terms and answer indicators are trimmed down.

Table 6 QA type/sections alignments. Section weight in parentheses

Question Type	GHR	GR	MPlus
Information	DiseaseDefinition (high)	Summary. Disease characteristics (high) Natural History (medium)	Definition (high) Considerations (medium)
Anatomy	DiseaseMutationGene, DiseaseMutationDescription (medium)	Molecular Genetics (low)	
Treatment		Summary. Management (low) Treatment of Manifestations (high)	Treatment
Prevention		Prevention of Secondary Complications (medium)	Prevention (high)
Diagnosis		Summary. Diagnosis/testing (low) Diagnosis (high)	Signs and tests (high)

		Testing (medium)	
Susceptibility	DiseaseAffectedGroup, (high) DiseaseInheritanceDiscussion (high)	Summary. Genetic counseling(low) Mode of Inheritance (medium) Prevalence (medium)	
Cause	DiseaseEtiology (high) DiseaseMutationDescription (low)	Molecular Genetic Testing (low)	Causes, incidence, and risk factors (medium)
Complications			Complications (high)
Prognosis	DiseaseDefinition (low)		Expectations (prognosis) (high)
Manifestations		Genotype-Phenotype Correlations (low)	Symptoms (high)

An example of a fully answered question is shown below.

I want to know if I have symptoms of cancer, is something seriously wrong?

I have had constant bone pain throughout my body, especially in my legs. I am vitamin d and calcium difficent. I also have had troubles breathing and extreme chest pain along with sharp pains throughout my body. I have several stress fractures in my legs and lower back. I have had a lose of appetite completely and this has been for a couple years. I have experienced some fatigue, and lose weight quickly.

Question Frame: MANIFESTATION of (cancer)

Answer: (http://www.nlm.nih.gov/medlineplus/ency/article/001289.htm)

Symptoms of cancer depend on the type and location of the cancer. For example, lung cancer can cause coughing, shortness of breath, or chest pain. Colon cancer often causes diarrhea, constipation, and blood in the stool.

Some cancers may not have any symptoms at all. In certain cancers, such as pancreatic cancer, symptoms often do not start until the disease has reached an advanced stage.

The following symptoms can occur with most cancers: Chills, Fatigue, Fever, Loss of appetite, Malaise, Night sweats, Weight loss.

Summary and Future Work

To address the growing need for answering consumer health questions, we launched a project aiming to develop an online system that automatically retrieves answers to consumer-health questions. Leveraging our experience and previously developed tools for clinical question answering and biomedical text processing, we made considerable advances towards reaching the project's objectives. We have developed CHIQA – a system that currently classifies customers' requests submitted to NLM and prepares answers for PubMed correction requests -- the largest group of the requests that needs to be handled by NLM customer services. Preparing automatic responses for the second largest part – questions about disorders and therapies is well underway: pending the final testing of the back-off module that classifies requests as consumer health questions, identifies the focus and retrieves appropriate MedlinePlus search results, this module will be deployed to the production customer support system.

In the process of CHIQA development, we generated several annotated collections of consumer-health questions that are publicly available. The co-reference resolution software is undergoing final testing as well and will shortly be publicly available.

Looking forward, we plan to use the collection of 300 question-answer pairs to learn ranking answer candidates. Using answer paragraphs extracted by NLM librarians, we plan to refine answer generation. Evaluating the performance of the knowledge-based frame parser and the statistical focus extractor on the same sets of questions, we observed that each approach has its strengths. We plan to find a way of combining the two approaches that will improve the overall results. Finally, we plan to expand the system to other large sets of questions: questions about pharmacologic actions of drugs, side-effects and ingredients are very frequent, as are questions about clinical trials.

References

Crowell J, Zeng Q, Ngo LH, Lacroix EM. A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries. JAMIA. 2004;11(3):179–185.

Cui L, Carter R, Zhang GQ. (2014) Evaluation of a novel Conjunctive Exploratory Navigation Interface for consumer health information: a crowdsourced comparative study. J Med Internet Res. 10;16(2):e45.

D'Alessandro DM, Kreiter CD, Peterson MW. An Evaluation of Information-Seeking Behaviors of General Pediatricians. Pediatrics, 113:64{69, 2004.

Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER. Analysis of questions asked by family doctors regarding patient care. BMJ,319(7206):358{361, 1999.

Ely JW, Osheroff JA, Ferguson KJ, Chambliss ML, Vinson DC, Moore JL. Lifelong self-directed learning using a computer database of clinical questions. Journal of Family Practice, 45(5):382{388, 1997.

Harabagiu S, Moldovan D, Clark C, Bowden M, Hickl, A Wang P. Employing two question answering systems in TREC-2005. In *Proceedings of Text Retrieval Conference 2005.*

Hickl A, Williams J, Bensley J, Roberts K, Shi Y, Rink B. Question Answering with LCC's CHAUCER at TREC 2006. In *Proceedings of Text Retrieval Conference 2006*

Hauser SE, Demner-Fushman D, Jacobs JL, Humphrey SM, Ford G, Thoma GR. Using Wireless Handheld Computers to Seek Information at the Point of Care: An Evaluation by Clinicians. J Am Med Inform Assoc, 14(6):807{815, 2007.

Liu F, Antieau LD, Yu H. Toward automated consumer question answering: automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, 2011, 44(6): 1032- 1038.

Luo J, Zhang GQ, Wentz S, Cui L, Xu R. SimQ. (2015) Real-Time Retrieval of Similar Consumer Health Questions. J Med Internet Res. 17(2): e43

Marine S, Embi PJ, McCuistion M, Haag D, Guard JR. (2005) NetWellness 1995 - 2005: ten years of experience and growth as a non-profit consumer health information and Ask-an-Expert service. AMIA Annu Symp Proc. 1043.

McCray AT, Tse T. Understanding search failures in consumer health information systems. AMIA Annu Symp Proc. 2003:430-4.

Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. CoRR. 2013;abs/1301.3781.

Mu X, Ryu H, Lu K. (2011) Supporting effective health and biomedical information retrieval and navigation: a novel facet view interface evaluation. J Biomed Inform. 44(4):576-86.

Surdeanu M, Ciaramita M, Zaragoza H. Learning to Rank Answers on Large Online QA Collections. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 719-727, 2008.

Vicedo JL, Ferrández A. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000. pages 555-562.

Wilbur WJ, Kim W, Xie N. Spelling correction in the PubMed search engine. Information Retrieval. Boston. 2006;9(5):543–564.

Zhang Y. Contextualizing Consumer Health Information Searching: an Analysis of Questions in a Social Q&A Community. In Proceedings of the 1st ACM International Health Informatics Symposium (IHI'10), 2010, pages 210-219.