Automatically Classifying Question Types for Consumer Health Questions

Kirk Roberts, PhD, Halil Kilicoglu, PhD, Marcelo Fiszman, MD, PhD,
Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD

Abstract

We present a method for automatically classifying consumer health questions. Our thirteen question types are designed to aid in the automatic retrieval of medical answers from consumer health resources. To our knowledge, this is the first machine learning-based method specifically for classifying consumer health questions. We demonstrate how previous approaches to medical question classification are insufficient to achieve high accuracy on this task. Additionally, we describe, manually annotate, and automatically classify three important question elements that improve question classification over previous techniques. Our results and analysis illustrate the difficulty of the task and the future directions that are necessary to achieve high-performing consumer health question classification.

Introduction

Questions posed in natural language are an intuitive method for retrieving medical knowledge. This is especially true for consumers, who may both lack medical background knowledge and be untrained in clinical problem solving techniques. The types of questions consumers ask differ as well. Medical question answering systems targeted to professionals often utilize the PICO structure,[1,2] which is particularly useful for comparative treatment questions. Conversely, consumers often ask for general information about a disease they have been diagnosed with, potential diseases associated with their symptoms, the range of potential treatment options, the typical prognosis, and how they might have acquired the disease.[3] Additionally, the most appropriate type of answer for a consumer is qualitatively different than that for a medical professional. The PICO-based methods typically draw answers from the latest medical research, such as that available in Medline®. Since consumers often lack the necessary medical background knowledge, providing them with the latest research may not aid their understanding. Instead, answers should be taken from consumer-oriented resources such as MedlinePlus® and other medical encyclopedias. Because of the differing nature in the type of questions and the most appropriate answers, consumer health question answering systems necessarily diverge from approaches that target medical professionals. In this work, we tackle a critical aspect of consumer question answering: automatically classifying the type of question asked by the consumer. The question type can then be used to identify the most appropriate resource to retrieve answers.

The National Library of Medicine® (NLM®) provides a medical information service, largely targeted toward consumers, as part of its library operations. Confirming the findings of Zhang,[3] the requests NLM receives from consumers contain questions that are different in nature from those posed by medical professionals. The consumer health questions concentrate less on the technical details, and more on general information, such as prognosis, treatment, and symptom information for diseases. Furthermore, consumer health requests often contain multiple questions, with significant amounts of co-reference and ellipses. For example:

- *I have recently been diagnosed with antisynthetase syndrome. Could you please provide me with information on antisynthetase syndrome? I am also interested in learning about prognosis, treatment, and clinical trials.*
- *Although I have not been diagnosed with trimethylaminuria, I have been having a foul odor for about 2 years. Can you tell me more about this condition? How can I be tested? Is there a cure for trimethylaminuria?*

This work is part of a larger project to automatically provide feedback to consumers with appropriate resources for their medical questions. Currently, NLM answers these questions manually, which means that a consumer question might not be answered for several days. By performing this process automatically, consumers can have their questions answered immediately. The process of extracting and representing individual questions from the larger request has already been discussed,[4,5] as well as the handling of co-reference and ellipses,[6] but the difficult problem of accurately classifying the individual consumer health questions has, to our knowledge, not yet been studied.

In this work, we discuss our initial attempt to automatically classify consumer health questions. While many of the questions are not strictly "factoid" questions in the style of the TREC question answering competitions,[7] in

this paper we focus on identifying the structural question elements typically found in factoid questions. Our results demonstrate that recognizing these elements with high accuracy can indeed aid in automatic question classification, but further work is necessary to make significant improvements in classifying consumer health questions. In summary, the primary contributions of this work are:

(i) presenting methods for automatically classifying consumer health questions,
(ii) demonstrating the importance of understanding the key semantic elements of each question, and
(iii) discussing what further work is needed to achieve high accuracy classification of consumer health questions.

Methods

We begin by describing our question types and the data they are annotated on. Next, we describe previous approaches to question classification to provide useful baseline methods. Finally, we describe additional annotations and features proposed in this work for consumer health question classification.

A. Question Types

We classify questions using the following question types. In many respects, the question types are generalizable to non-consumer medical questions. A more complete description of each question type, along with annotation rules and many more examples, can be found in Roberts et al.[8] To illustrate how the question types correspond to encyclopedic sections, for each question type we list the sections of MedlinePlus where answers are most likely to be found (though our system uses several other sources as well).

(1) ANATOMY: Identifies questions asking about a particular part of the body, such as the location affected by a disease. (Answers to ANATOMY questions are typically found in the "Anatomy/Physiology" section.)
   – *Does IP affect any areas inside the body, such as internal organs?*

(2) CAUSE: Identifies questions asking about the cause of a disease. This includes both direct and indirect causes, such as factors that might increase the susceptibility to a disease. ("Common Causes")
   – *Can pregnancy trigger such an issue?*

(3) COMPLICATION: Identifies questions asking about the problems a particular disease causes. This primarily focuses on the risks faced by patients with the disease and does not include the signs/symptoms of a disease. ("Related Issues")
   – *Are carriers of cystic fibrosis at a higher risk for other health conditions?*

(4) DIAGNOSIS: Identifies questions asking for help making a diagnosis. Our question answering system is not designed to provide a direct diagnosis. However, the DIAGNOSIS type does include questions asking about diagnostic tests, or methods for determining the difference between possible diagnoses (differential diagnosis). ("Diagnosis", "Testing")
   – *Can Bloom syndrome be detected before symptoms appear?*

(5) INFORMATION: Identifies questions asking for general information about a disease. This includes type information about diseases, e.g., whether two disease names represent the same disease, or if one disease is a type of another disease. ("Definition", "Description")
   – *Can you please provide me with general information about hyper IgD syndrome?*

(6) MANAGEMENT: Identifies questions asking about the management, treatment, cure, or prevention of a disease. ("Treatment", "Prevention")
   – *Are there new therapies for treatment of pili torti?*

(7) MANIFESTATION: Identifies questions asking about signs or symptoms of a disease. ("Symptoms")
   – *Are there any physical characteristics associated with the disorder?*

(8) OTHEREFFECT: Identifies questions asking about the effects of a disease, excluding signs/symptoms (MANIFESTATION) or risk factors (COMPLICATION). When the question requires medical knowledge to understand a given effect is actually a MANIFESTATION or COMPLICATION, it is instead classified as an OTHEREFFECT. ("Symptoms", "Related Issues")
   – *Can tuberous sclerosis affect eye closure?*

Request:
I have been recently diagnosed with antisynthetase syndrome. Could you please provide me with information on antisynthetase syndrome? I am also interested in learning about prognosis, treatment, and clinical trials.
Decomposed Questions:
Q1) Could you please provide me with information on antisynthetase syndrome?
Q2) I am also interested in learning about prognosis.
Q3) I am also interested in learning about treatment.
Q4) I am also interested in learning about clinical trials.

Table 1: Question Decomposition Example.

(9) PERSONORG: Identifies any question asking for a person or organization involved with a disease. This can include medical specialists, hospitals, research teams, or support groups for a particular disease. Answers to PERSONORG questions are typically not found in MedlinePlus articles, but may be found in links on MedlinePlus pages (especially for support groups).
   – *I would like information about which doctors could treat me.*

(10) PROGNOSIS: Identifies questions asking about life expectancy, quality of life, or the probability of success of a given treatment. ("Expectations")
   – *I would like to know what the life expectancy is for people with this syndrome.*

(11) SUSCEPTIBILITY: Identifies 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. ("Mode of Inheritance", "Prevalence")
   – *Are people of certain religious backgrounds more likely to develop this syndrome?*

(12) OTHER: Identifies disease questions that do not belong to the above types. This includes non-medical questions about a disease, such as requests for financial assistance or the history of a disease. Answers to OTHER questions are typically found in the "Definition" and "Description" sections, though by definition these answers may be found anywhere in the encyclopedic entry.
   – *How did Zellweger Syndrome get its name?*

(13) NOTDISEASE: Identifies questions that are not handled by our question answering system.
   – *Is there any information about what genes are on chromosome 20q12?*

B. Data

As a source for consumer health questions about diseases, we used 1,467 publicly available requests on the Genetic and Rare Diseases Information Center (GARD) website. It may be argued that, since these requests are about rare diseases, they do not constitute a representative sample of disease questions. While this may be true, GARD requests are still appropriate for this task because the aspects of diseases (treatment, prognosis, etc.) that the requests are concerned with are shared between all diseases. Additionally, since there are far fewer online resources for rare diseases, these questions may be more reflective of the types of questions consumers actually ask than other disease question sources.

The GARD requests contain multiple question sentences, and many of the question sentences contain requests for different types of information, thus the questions were annotated for syntactic decomposition as described in Roberts et al.[4] Table 1 illustrates an example of question decomposition. This decomposition process results in 2,937 questions from the 1,467 GARD requests. Additionally, these requests are annotated with a FOCUS, typically one per request, the disease the consumer is interested in learning more about. In Table 1, the FOCUS is *antisynthetase syndrome*. Since our question answering approach primarily utilizes medical encyclopedias, the FOCUS is often the encyclopedic entry, while the question type is the section of the entry in which the answer is found. For cause-and-effect questions, for example, if the FOCUS is the cause, the answer is likely to be found in the effect section, and vice versa.

The process of annotating the thirteen question types on the GARD data is described more thoroughly in Roberts et al.[8] In this paper, we describe the first automatic classification methods on this data. Additionally, we have manually annotated additional elements, described below, not discussed in any of our previous work.

C. Previous Question Classification Approaches

To investigate how classifying consumer health questions differs from previous forms of question classification, we present four previous machine learning-based methods and their corresponding features. These methods provide solid baseline approaches as well as a reliable set of features from which to choose for consumer health question classification. In some cases, as described below, we slightly alter features from their original versions due to either resource availability or appropriateness. In other cases, the original feature description might not have been completely clear, and thus we chose the best interpretation based on our data.

Li and Roth[9] presented the first machine learning method for classifying questions by their answer type. Answer types specify the semantic class of the answer, as opposed to our question types which specify the search strategy (typically, the section of an encyclopedic entry). In many cases, the answer type and question type are identical.

- f_{LR1}: Words. This is the classic "bag-of-words"feature. We use the caseless form of each token
- f_{LR2}: Part-of-speech tags
- f_{LR3}: Phrase chunks
- f_{LR4}: Head chunks. The first noun phrase chunk and the first verb phrase chunk in the question
- f_{LR5}: Named entities. Li and Roth used named entity types such as PERSON and LOCATION, which make less sense for our data. Instead, we use UMLS semantic types
- f_{LR6}: Semantically related words. Here Li and Roth utilized lexicons that would capture similar words such as *actor* and *politician* for the HUMAN:INDIVIDUAL type. Instead, we approximate this feature using the cue phrases from Kilicoglu et al.[6]

Yu and Cao[10] presented a method for classifying into a similar set of classes to our own, but tailored for questions posed by physicians instead of consumers.

- f_{YC1}: Words, the same as f_{LR1}
- f_{YC2}: Stemmed (lemmatized) words
- f_{YC3}: Bigrams
- f_{YC4}: Part-of-speech tags, the same as f_{LR2}
- f_{YC5}: UMLS concepts. This feature is useful to recognize UMLS synonyms
- f_{YC6}: UMLS semantic types, the same as f_{LR5}

Liu et al.[11] presented a method for distinguishing consumer health questions from professional questions. Their features are designed more to capture the lexical tendencies of consumers and professionals against all types of questions, but we utilize them here because their work is one of the first studies in automatically classifying consumer questions.

- f_{LAY1}: Words, the same as f_{LR1}
- f_{LAY2} - f_{LAY4}: Minimum, maximum, and mean word length in the question
- f_{LAY5}: The question length in words
- f_{LAY6} - f_{LAY8}: Minimum, maximum, and mean inverse document frequency from Medline 2010. Instead, we use a 2013 version of Pubmed Central

Patrick and Li[12] presented a method for classifying clinical questions about information within an EHR. The questions in their data were posed by medical staff in an ICU. They classify clinical questions into a taxonomy and a set of template questions to facilitate answer retrieval within an EHR system. Patrick and Li used SNOMED-CT, while we use all of UMLS.

- f_{PL1}: Unigrams, same as f_{LR1}
- f_{PL2}: Lemmatized unigrams, same as f_{YC2}
- f_{PL3}: Bigrams, same as f_{YC3}
- f_{PL4}: Lemmatized bigrams
- f_{PL5}: Interrogative word. Typically, a WH-word (e.g., `what`, `how`) or verb (`is`, `could`)
- f_{PL6}: Interrogative word + next token

- f_{PL7}: UMLS semantic types, same as f_{LR5}. Patrick and Li also use a version of this feature limited to the SNOMED category "observable entity". We do not include this specific feature as it has no clear analogue nor recognizable need in our data
- f_{PL8} + f_{PL9}: verb-subject relations in both the original (f_{PL8}) and lemmatized (f_{PL9}) form. While Patrick and Li use the Enju parser, we use the Stanford dependency parser[13] In the question, "*What treatments have you recommended?*", f_{PL8} would be `nsubj(recommended,you)`, f_{PL9} would be `nsubj(recommend,you)`
- f_{PL10} + f_{PL11}: verb-object relations in both the original (f_{PL10}) and lemmatized (f_{PL11}) form, again using the Stanford dependency parser. From the question above, f_{PL10} would be `dobj(recommended,treatments)`, f_{PL11} would be `dobj(recommend,treatment)`
- f_{PL12} - f_{PL15}: Features f_{PL8} - f_{PL11}, but with the subject/object being replaced by `UMLS` if it is a UMLS concept
- f_{PL16} - f_{PL19}: Features f_{PL8} - f_{PL11}, but with the subject/object being replaced by its semantic type if it is a UMLS concept
- f_{PL20}: WordNet synonyms and antonyms of key terms found by analyzing the data. We approximate this by using a WordNet-expanded version of f_{LR6}

Patrick and Li employ three classifiers that each utilize a sub-set of these features. Their unanswerable question classifier (UQC) uses f_{PL2}, f_{PL13}, and f_{PL15}. Their answerable question taxonomy classifier (QTC) uses f_{PL2}, f_{PL13}, f_{PL15}, and f_{PL20}. Their genetic template classifier (GTC) uses f_{PL2}, f_{PL5}, f_{PL6}, f_{PL7}, and f_{PL8}.

D. Question Stems

The *question stem* is the span of text that introduces the question. It is often referred to as the WH-word or interrogative word, though it need not be a classic WH-word (e.g., "*Are the treatments effective?*") or even a single word. It provides a useful signal for indicating the question structure, while also indicating distributional differences in the question types. For example, a *where* question is far more likely to have a PERSONORG or ANATOMY type than a PROGNOSIS or CAUSE type. Examples of question stems are:

- **What** *may be the underlying cause?*
- *At* **what** *age progressive hemifacial atrophy typically present?*
- *If she needs hormonal treatment,* **which** *medication may be the safest choice?*
- **Could** *you please let me know the name of the blood test used to diagnose a vitiligo carrier?*

Not all WH-words are question stems, however, as it depends upon their syntactic function:

- **Are** *there any other recommended treatments other than what my doctor has already tried for this condition?*
- **Have** *any side effects been reported in patients who use these medications?*

We annotated all 2,937 questions with a single-token question stem. 99.2% of questions were judged to have a question stem, and in 76.6% of these the question stem was the first word. The ten most common question stems in the GARD data are *what*, *is*, *how*, *can*, *are*, *if*, *does*, *do*, *looking*, and *could*.

We use a two-step machine learning method for automatically recognizing question stems. First, a question-level SVM determines whether or not a question has a question stem using a bag-of-words model where the first word is intentionally cased and other words are lower-cased. Second, a token-level SVM is used to rank the words in the sentence, with the top-ranked word being chosen as the question stem. This ranker uses the current word (cased as before), lemma, part-of-speech, previous word, and next word as features.

Once recognized, the question stem allows for several useful features during question classification.

- f_{QS1}: Uncased Question stem
- f_{QS2}: Uncased question stem plus the next token
- f_{QS3}: Uncased question stem plus the next two tokens
- f_{QS4}: Uncased question stem plus the next three tokens
- f_{QS5}: Whether or not the question stem is in a pre-defined set of boolean stems
- f_{QS6} + f_{QS7}: Question tokens (uncased + stemmed) ignoring those tokens before the question stem

- f_{QS8} + f_{QS9}: Versions of f_{QS6} and f_{QS7}, respectively, that account for word order after the question stem
- f_{QS10} + f_{QS11}: Similar to f_{QS6} and f_{QS7}, but only over the range of the WHNP in the syntactic parse tree
- f_{QS12} + f_{QS13}: Similar to f_{QS6} and f_{QS7}, but only over the range of the clause in the syntactic parse tree

E. Answer Type Terms

The *answer type term* is the noun phrase that specifies the expected answer type. Due to its importance, a significant amount of question answering research has focused on describing and automatically identifying answer type terms.[14,15] In the traditional TREC-style factoid questions, it is typically found immediately after the question stem *what* in one of several predictable syntactic positions:

- *What* **disease** *does she have?*
- *What* **symptoms** *usually occur?*
- *What is the* **prognosis** *for this disease?*

In our data, however, answer type terms may occur in less predictable locations, or be syntactically dominated by a more general word:

- *Could you tell me some of the* **symptoms** *of cardiomyopathy hypogonadism collagenoma syndrome?*
- *How can I obtain information about* **treatment options** *for FIBGC?*

Additionally, since our interest is in question types instead of answer types, we relax the definition from an *explicit* answer type term to an *implicit* answer type term. For instance:

- *Can women have* **symptoms** *of glucose 6 phosphate dehydrogenase (G6PD) deficiency?*
- *Where can I get information on the official* **recommendations** *for pregnant women?*

Neither question above has an explicit answer type term, as their question stems convey the answer type (boolean for *Can*, location for *Where*). However, implicitly these questions are asking for symptoms and recommendations, respectively. This indirect question style is a major characteristic of our data, and reflects a less formal querying style. We thus use implicit answer type terms for classifying questions, which are more semantic in nature than the syntactically predictable explicit answer type terms. Thus our answer type term annotations resemble something closer to our question types and further from Li and Roth's answer types.

We annotated all 2,937 questions with a single-token answer type term. 62.4% of questions were judged to have an answer type term. The ten most common answer type terms in the GARD data are *information*, *treatment*, *symptoms*, *prognosis*, *more*, *treatments*, *chances*, *testing*, *risk*, and *expectancy*.

We use the same two-step process to identify answer type terms as we do question stems. Once recognized, the answer type term allows for many useful features:

- f_{ATT1}: Whether or not an answer type term is present
- f_{ATT2} + f_{ATT3}: The uncased and lemmatized answer type term token
- f_{ATT4}: The WordNet hypernyms of the answer type term token
- f_{ATT5} + f_{ATT6}: The uncased and lemmatized words in the answer type term's noun phrase
- f_{ATT7}: The WordNet hypernyms of the words in the answer type term's noun phrase
- f_{ATT8}: The full noun phrase of the answer type term

F. Answer Type Predicates

Not all questions specify their answer type with a question stem or nominal answer type term. Especially in our data, often the answer type is specified with a verb or adjective. To differentiate these from answer type terms, we refer to the verb or modifier span as the *answer type predicate*. Examples of answer type predicates are:

- *How is this condition* **treated***?*
- *If so, what is the likelihood he could* **pass** *it on to his next child?*
- *Is this a* **genetic** *disorder?*

The first two examples above illustrate verbal answer type predicates, while the third illustrates and adjective answer type predicate. The second example also shows a case where an answer type term (underlined) and an answer type

predicate both exist in the same question. In this case, the answer type predicate is more useful in determining a question type of SUSCEPTIBILITY, but often both are necessary to make a proper decision.

We annotated all 2,937 questions with a single-token answer type predicate. 54.2% of questions were judged to have an answer type predicate. The ten most common answer type predicates in the GARD data are *treated*, *have*, *is*, *diagnosed*, *causes*, *affect*, *genetic*, *cause*, *tested*, and *help*.

We use the same two-step process to identify answer type predicates as question stems and answer type terms. The feature that utilize answer type predicates (f_{ATP1} - f_{ATP7}) correspond exactly with the answer type term features. The only exception is that adjectives in WordNet, unlike verbs and nouns, cannot have hypernyms.

Collectively, we refer to the question stem, answer type term, and answer type predicate as the *key question elements*.

G. Classifier

To automatically classify question types, we utilize a multi-class SVM[16]. Due to the wide variety of features described in this paper, we use an automatic feature selection technique[17] to choose the best sub-set of features. Because many of these features convey redundant information, using every feature would not only be slower, but would result in over-training and a drop in accuracy of 3-4%. Ideally, we would have sufficient annotated data to perform these experiments on a development set and provide a final evaluation on a held-out test set. The main goal of this paper, however, is to explore the utility of various machine learning features on this problem. The feature selection technique relies on a 5-fold cross validation on the full data set using a different view (a shuffled split) of the data from that used in the Results section to ensure more generalizable results. Furthermore, since the key question elements are trained on the same data, we utilize stacking to ensure the relevant features are representative of the automatic output.

In addition to replicating the existing feature sets above, we propose three additional feature sets chosen with automatic feature selection:

- CQT_1: The best features without key question element features: f_{LR1}, f_{PL6}, f_{LR5}, f_{LAY7}, f_{LAY8}, f_{LAY4}
- CQT_2: The best features using the automatic key question elements: f_{QS4}, f_{QS7}, f_{QS10}, f_{ATT4}, f_{ATP3}, f_{ATP8}, f_{LR6}
- CQT_3: The same features as CQT_2 using the *gold* key question elements

Feature set CQT_1 allows us to determine a baseline without any key question element involvement. Feature set CQT_2 provides an expectation of how well our overall method would perform in practice. Finally, feature set CQT_3 allows us to establish a ceiling for the utility of key question elements based on the gold annotations.

Results

The results for our key question element classifiers on a 5-fold cross validation are shown in Table 2(a). Question stem recognition is the easiest task with an F_1-measure of 96.16, largely due to the limited vocabulary and the frequency of the first word acting as the question stem. The answer type term recognizer was the next best at 84.08, likely because the answer type terms are nouns that often occur in specific contexts. Finally, the answer type predicate recognizer has the most difficulty with an F_1-measure of 76.81. The fact that only around half of questions have an answer type predicate, while almost every question has at least one verb or adjective, likely leads to the lower score. While it is certainly possible to improve the automatic recognition of these question elements, below we discuss why dramatic improvements to the accuracy of these methods might not result in similar improvements to the overall question type classifier.

The results for question classification are shown in Table 2(c). The baseline bag-of-words model performs at 76.9%. The previous question classification techniques barely out-perform the baseline, with one system even under-performing this baseline. Feature set CQT_1 demonstrates slight improvements (an increase of 0.6 points) can be made over Yu and Cao's feature set by automatic feature selection. The key question element features (CQT_2), however, show a larger improvement of 2 percentage points. When gold question elements are used (CQT_3), this improves to 4 points over Yu and Cao's method and 5.5 points over the bag-of-words baseline. This ceiling, however, appears fairly low considering we used gold question elements. Originally this was thought to be a result of annotation inconsistency, but upon further analysis several interesting sources of error emerge, discussed below. Finally, Table 2(b) provides details of how well CQT_3 classifies each question type. The F_1 scores are largely reflective of the imbalances in the

	Precision	Recall	F_1
Question Stems	95.97	96.36	96.16
Answer Type Terms	84.47	83.69	84.08
Answer Type Predicates	76.90	76.71	76.81

(a) Results for automatic detection of key question elements.

Question Type	# Annotations	Precision	Recall	F_1
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

(b) Detailed question classification results by question type using CQT_3 feature set.

	Accuracy
Bag-of-words	76.89%
Li and Roth (2002)	77.45%
Yu and Cao (2008)	78.43%
Liu et al. (2011)	76.37%
Patrick and Li (2012) – UQC	77.41%
Patrick and Li (2012) – QTC	77.76%
Patrick and Li (2012) – GTC	77.76%
CQT_1	79.01%
CQT_2	80.40%
CQT_3	82.42%

(c) Results for automatic classification of question types.

Min # Elements	# Questions	Accuracy
1	2,814	82.16
2	2,606	84.92
3	2,481	86.50
4	2,353	87.46
5	2,286	88.15
10	1,919	90.20
25	1,538	92.39

(d) Sparsity Experiment. Demonstrates how accuracy of CQT_3 model increases as questions with uncommon question elements are removed.

Table 2: Experiments

data, though OTHEREFFECT appears particularly difficult while PERSONORG, CAUSE, and MANIFESTATION appear particularly simple relative to their relative frequencies in the data.

Discussion

The fact that using gold question elements only raised the score slightly is somewhat surprising. These are the elements of a question that humans intuitively look for while annotating. The semantic gap between answer types and question types does not fully explain the 17.6% error rate when using gold elements. We performed a detailed error analysis to understand the major types of errors made when using gold elements.

The first major type of error involves word sense ambiguity, where a question element could have multiple possible interpretations. For instance:

- *My lower back doesn't seem to work, and I wonder if I will ever be able to* **run**. PROGNOSIS
- *Does CREST syndrome* **run** *in families?* SUSCEPTIBILITY
- *Can you send me a* **link** *concerning hereditary fructose intolerance?* INFORMATION
- *Is there a* **link** *between MELAS and a person who is not really strong?* OTHEREFFECT

In the first two examples, different senses of the answer type predicate *run* are used. In the first, *run* corresponds to the WordNet sense defined as "move fast by using one's feet". This corresponds to inquiring about a disease's impact on a part of one's lifestyle, which we define as PROGNOSIS. In the second case, *run* corresponds to the WordNet sense meaning "occur persistently". The question is therefore asking whether the disease is passed genetically and is thus a SUSCEPTIBILITY question. In our data, the answer type predicate *run* corresponds more with SUSCEPTIBILITY and thus the first question was mis-classified. The next two examples illustrate different senses of the answer type term *link*. The first *link* refers to a website and should be classified INFORMATION. The second *link* is referring to a causal connection between a disease and a possible effect. This question was annotated as OTHEREFFECT. In our data, the answer type term *link* corresponds more with OTHEREFFECT, and so the first *link* was mis-classified. Clearly, some form of word sense disambiguation (WSD) might prove helpful, but WSD has been shown to be a very difficult and highly domain-dependent task[18]. We therefore leave the task of WSD in consumer health questions to future work.

The word sense problem illustrates a key insight into how consumer health questions differ from professional questions. Consumers often lack the terminological familiarity to pose questions in an unambiguous manner. For example, a health professional might have written the second and fourth questions above as:

Answer Type Terms	information, treatment, symptoms, prognosis, more, treatments, chances, testing, risk, expectancy, research, chance, options, cure, test, people, studies
Answer Type Predicates	treated, have, is, diagnosed, causes, affect, genetic, cause, tested, help, treat, associated, having, rare, inherited

Table 3: Answer type terms and answer type predicates with at least 25 instances in our data.

- *Is CREST syndrome* **hereditary***?*
- *Is physical weakness a* **manifestation** *of MELAS?*

In these questions, the key question elements are entirely unambiguous, and should be easily recognized by automatic classifiers. Instead, ambiguous words like *run* and *link* are sufficiently common in the training data to result in misclassification of questions where those words might have actually been the best choice of terminology. Note that this is a different terminological problem from that discussed by McCray et al.[19] In their work, terminology issues arose from the way diseases were specified. For question classification, the disease itself is largely unimportant since we are trying to classify which aspect of the disease the user is interested in.

A second important type of error when using gold question elements has to do with data sparsity: if a question element only appears once or twice, there is not sufficient evidence for its proper question type. For instance:

- *I'm looking for a* **dermatologist** *in my area who has experience with this condition.* PERSONORG
- *What is the* **remedy** *of mixed connective tissue disorder?* MANAGEMENT

While clearly unambiguous, *dermatologist* and *remedy* each only appear once in our data. Resources like WordNet can be utilized to recognize a *dermatologist* is a *doctor*, and a *remedy* is a *cure*, and indeed f_{ATT4} does just this. Yet hypernym features also introduce a good deal of noise, and often aren't highly trusted by the classifier. We can explore the impact of sparsity on question classification by simply removing questions with uncommon elements. Table 2(d) shows this experiment. It is relatively safe to say that having at least 10 occurrences of a question element in the data removes the effects of sparsity, in which case the error rate is reduced by almost half.

Finally, the sparsity experiment provides a useful mechanism for exploring the role of the semantic gap between an answer type (the form or class of an answer) and the question type (which is more related to the topic). Table 3 lists the elements that occur at least 25 times in the data. Clearly, some are ambiguous (e.g., *have*, *options*), but many of the seemingly unambiguous words do not correspond to one specific question type. This is due to the semantic difference between answer types and question types. For example, the answer type predicate *tested* is used in 18 DIAGNOSIS, 12 SUSCEPTIBILITY, and 2 PERSONORG questions. In each case *tested* is being used in the same sense. Furthermore, the answer type term *people* is used in 11 SUSCEPTIBILITY, 9 OTHEREFFECT, 6 PERSONORG, 2 MANAGEMENT, and 2 PROGNOSIS questions. For instance:

- *How can I be* **tested** *for this condition?* DIAGNOSIS
- *Who in the family needs to be* **tested** *for the carrier gene for MLD?* SUSCEPTIBILITY
- *How many* **people** *are affected by Alexander disease?* SUSCEPTIBILITY
- *I would like to contact other* **people** *with epidermolysis bullosa acquisita.* PERSONORG

The first use of *tested* is in a question asking for a diagnostic test, and thus a DIAGNOSIS question. The second question asks *who* requires testing because they are genetically susceptible. In the third question, *people* is being used as a unit of measure for prevalence (SUSCEPTIBILITY), while the fourth question uses *people* to imply other sufferers (PERSONORG). As could be seen in these examples, in order to recognize question types with much higher accuracy than the systems presented here, methods will need to be developed to understand the role an answer type plays in determining the consumer health question's class.

Conclusion

We have presented a method to automatically classify consumer health questions into one of thirteen question types for the purpose of supporting automatic retrieval of medical answers. We have demonstrated that previous question classification methods are insufficient to achieve high accuracy on this task. Additionally, we described, annotated, and classified three important question elements that improve question classification over previous techniques. Our results showed small improvements on a difficult task. We concluded by motivating importance of overcoming word sense ambiguity, data sparsity, and the answer type/question type semantic gap for future work.

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

References

1. Dina Demner-Fushman and Jimmy Lin. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103, 2007.
2. Connie Schardt, Martha B. Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics & Decision Making*, 7(16), 2007.
3. Yan Zhang. 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*, 2010.
4. Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Annotating Question Decomposition on Complex Medical Questions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2598–2602, 2014.
5. Kirk Roberts, Z, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. Decomposing Consumer Health Questions. In *Proceedings of the 2014 BioNLP Workshop*, pages 29–37, 2014.
6. Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In *Proceedings of the 2013 BioNLP Workshop*, pages 54–62, 2013.
7. Ellen M. Voorhees. Overview of TREC 2004. In *Proceedings of the Thirteenth Text Retrieval Conference*, 2004.
8. Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Annotating Question Types for Consumer Health Questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, 2014.
9. X. Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.
10. Hong Yu and YongGang Cao. Automatically Extracting Information Needs from Ad Hoc Clinical Questions. In *Proceedings of the AMIA Annual Symposium*, 2008.
11. Feifan Liu, Lamont D. Antieau, and Hong Yu. Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, 44(6), 2011.
12. Jon Patrick and Min Li. An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45:292–306, 2012.
13. Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.
14. Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. Enhanced Answer Type Inference from Questions using Sequential Models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.
15. Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question Classification using Head Words and their Hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 927–936, 2008.
16. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
17. Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.
18. Roberto Navigli. Word sense disambiguation: A survey. In *ACM Computing Surveys*, volume 41, pages 1–69, 2009.
19. Alexa T. McCray, Russell F. Loane, Allen C. Browne, and Anantha K. Bangalore. Terminology Issues in User Access to Web-based Medical Information. In *Proceedings of the AMIA Annual Symposium*, pages 107–111, 1999.