Annotating Question Decomposition on Complex Medical Questions

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Bethesda MD, USA
kirk.roberts@nih.gov, ddemner@mail.nih.gov

Abstract

This paper presents a method for annotating question decomposition on complex medical questions. The annotations cover multiple syntactic ways that questions can be decomposed, including separating independent clauses as well as recognizing coordinations and exemplifications. We annotate a corpus of 1,467 multi-sentence consumer health questions about genetic and rare diseases. Furthermore, we label two additional medical-specific annotations: (1) background sentences are annotated with a number of medical categories such as symptoms, treatments, and family history, and (2) the central focus of the complex question (a disease) is marked. We present simple baseline results for automatic classification of these annotations, demonstrating the challenging but important nature of this task.

Keywords: question decomposition, question answering, medical language processing

1. Introduction

Natural language questions provide an intuitive interface for querying medical knowledge. This is especially true for non-experts, often referred to as *consumers*. While medical experts are trained to both understand diseases and the medical problem solving process, consumers generally seek broader, multi-faceted medical information. This results in questions that may be far more elaborate, often containing contextual background information such as a patient's family history, diagnoses, symptoms, and comorbidities. Their information needs often yield multiple related questions that, when combined with background information, are substantially more complex than the questions most existing question answering systems are designed to handle. For instance:

(1) I have an infant daughter with Coffin Siris Syndrome. I am trying to find information as well as connect with other families who have an affected child.

(2) My mother has recently been diagnosed with Myofibrillar myopathy. Her first symptoms became apparent about 2.5 years ago when she developed weakness in one foot. She currently has little movement in her legs, and her hands, arms, speech and breathing are affected. What is the prognosis and life expectancy for this condition and what can we expect from here?

(3) A family member has recently been diagnosed with hemangiopericytoma of the skull and a lesion in the lung and possibly the kidney. We are interested in learning more about this condition, including the standard course of treatment.

We refer to these as *complex questions*, as they contain multiple questions related by a central theme or focus. In order to process these questions, their underlying sub-questions, background elements, and focus need to be recognized. We refer to this task as a form of *question decomposition*, as it enables the decomposition of long, complex medical questions into concise questions that may reference appropriate background information. The information needs of the question can then be re-formulated into more traditional medical questions. For the questions above, these decomposed questions would be:

(1′) I am trying to find information about Coffin Siris Syndrome.
I am trying to connect with other families who have an affected child with Coffin Siris Syndrome.

(2′) What is the prognosis for Myofibrillar myopathy?
What is the life expectancy for Myofibrillar myopathy?
What can we expect from here with Myofibrillar myopathy?

(3′) We are interested in learning more about hemangiopericytoma.
We are interested in learning the standard course of treatment for hemangiopericytoma.

In this paper we present a manually annotated corpus of 1,467 consumer health questions related to genetic and rare diseases. The goal for this corpus is to enable the training and evaluation of automatic techniques for decomposing complex medical questions. To demonstrate the efficacy of the corpus and to get a general sense for the difficulty of the task, we further evaluate easily reproducible baselines that combine rule-based and machine learning (ML) based methods.

The rest of this paper is organized as follows. Section 2 outlines previous work in medical question answering, question decomposition, and text simplification, as well providing background information on how this work fits into our larger goals for medical question answering. Section 3 describes our annotations and provides further examples. Section 4 describes our annotated corpus, including annotation statistics and inter-annotator agreement numbers. Section 5 evaluates simple baseline methods for automatically recognizing each of the discussed annotations in order to estimate the level of difficulty for this task. Finally, Section 6 summarizes our contributions and proposes future work.

2. Background

The need for efficient access to medical knowledge has led to significant study of medical question answering. Demner-Fushman and Lin (2007) propose a question-answering system based on filling in PICO (Patient, Intervention, Comparison, Outcome) slots instead of natural language questions. Andersen et al. (2012) created a medical question corpus focused on diabetes and heart disease. They concentrate, however, on concise questions, such as those easily convertible to an SQL query. Yu and Cao (2008) identify "general topics" for complex medical questions posed by physicians. These topics correspond roughly to question types, such as requesting information about diagnosis, tests, or treatments. Similar to our data, they may identify multiple information needs, but they only perform a question-level multi-labeling and do not identify the actual question boundaries. For example, in Question (2) their method would identify that there is both a test and treatment information need, but not indicate the span of text indicating these questions and any question-specific information that may be associated with it. Nor do they identify the focal concept to use to find treatment, testing, or other information about. Since they do not decompose the complex question into more concise questions, this limits the information retrieval strategy to largely keyword-based techniques where the keywords for each sub-question are drawn from the entire complex question. In addition, they do not include a general information topic, which is very common in consumer-posed questions. The AskHERMES system (Cao et al., 2011) utilizes this question analysis method in a web-based medical question answering system targeted toward physicians.

Outside of the medical question answering domain, question decomposition has been approached in a variety of ways. Lacatusu et al. (2006) demonstrates an example of *syntactic* question decomposition, utilizing the syntactic structure of a complex question to identify sub-questions. Hartrumpf (2008) performs a deep *semantic* question decomposition, relying upon the question already being converted into a semantic logical form. From this logical representation, it is then possible to identify sub-questions that may be asked in isolation, acting as a filter on answers to the fully-specified question. Harabagiu et al. (2006) performs question decomposition through a random walk over a knowledge base of questions created from a corpus. The random walk increases the relevant information presented to a multi-document summarization system. Of all of these methods, our data most closely resembles the syntactic question decomposition, yet there are several semantic considerations that we discuss below.

At its most general level, question decomposition can be seen as a form of text simplification (Bott et al., 2012; Drndarević et al., 2013). While question decomposition has the aim of concise questions answerable by traditional automatic question answering systems, text simplification seeks to simplify text for the sake of human consumption. Nevertheless, many of the operations presented below, such as recognizing coordinating clauses, are also part of text simplification tasks.

This work is part of a larger project with the aim of automatically answering consumer health questions. The National Library of Medicine (NLM) receives thousands of medical questions each year, largely from individuals outside the medical profession. Our question answering system categorizes and routes user requests according to a library-specified taxonomy. For questions about diseases, our system identifies the most appropriate consumer resource that answers the user's question. Typically, this answer is in the form of a page or section from MedlinePlus (Schnall and Fowler, 2013), a consumer-oriented medical encyclopedia. In previous work (Kilicoglu et al., 2013) we have addressed the importance of ellipses and co-reference in analyzing consumer-posed questions. In this work we concentrate on a different aspect of these complex questions: identifying the explicit questions asked by consumers in long, complex medical questions.

3. Complex Question Decomposition

The goal of question decomposition is to turn a complex, multi-faceted question into a list of related, concise questions that can be answered by more traditional question answering systems (such as those of Demner-Fushman and Lin (2007) and Cao et al. (2011)). To accomplish this, we propose a largely syntax-based method of annotating complex questions with their sub-questions and background information.

Question sentences can be syntactically decomposed in one of two ways. Question (2) shows how sentences can be syntactically split into separate questions by recognizing that the coordinating conjunction *and* splits independent clauses (*What is the... and what can we...?*) with syntactically self-contained questions. Alternatively, questions may be decomposed by extracting phrase level coordinations, typically either a noun phrase (NP) or verb phrase (VP). Consider the following:

(4) I am an adult woman who has been recently diagnosed with Ménétrier disease. My symptoms seem to be worsening. I am seeking information about the symptoms, cause(s), prognosis, genetic association, and treatment. I am eager to hear from you and appreciate your assistance.

Here, a multiple coordination is used to indicate five separate questions:

(4′) I am seeking information about the symptoms of Ménétrier disease.

I am seeking information about the cause(s) of Ménétrier disease.

I am seeking information about the prognosis of Ménétrier disease.

I am seeking information about the genetic association of Ménétrier disease.

I am seeking information about the treatment of Ménétrier disease.

To capture both types of syntactic decomposition, we propose six annotation types, explained below and illustrated in Table 1.

Question (1) S_1: [I have an infant daughter with [Coffin Siris Syndrome]$_{\text{FOCUS}}$.]$_{\text{BACKGROUND(DIAGNOSIS)}}$ S_2: [I am trying to [find information as well as connect with other families who have an affected child]$_{\text{COORDINATION}}$.]$_{\text{QUESTION}}$
Question (2) S_1: [My mother has recently been diagnosed with [Myofibrillar myopathy]$_{\text{FOCUS}}$.]$_{\text{BACKGROUND(DIAGNOSIS)}}$ S_2: [Her first symptoms became apparent about 2.5 years ago when she developed weakness in one foot.]$_{\text{BACKGROUND(SYMPTOM)}}$ S_3: [She currently has little movement in her legs, and her hands, arms, speech and breathing are affected.]$_{\text{BACKGROUND(SYMPTOM)}}$ S_4: [What is the [prognosis and life expectancy]$_{\text{COORDINATION}}$ for this condition]$_{\text{QUESTION}}$ [and what can we expect from here?]$_{\text{QUESTION}}$
Question (3) S_1: [A family member has recently been diagnosed with [hemangiopericytoma]$_{\text{FOCUS}}$ of the skull and a lesion in the lung and possibly the kidney.]$_{\text{BACKGROUND(DIAGNOSIS)}}$ S_2: [We are interested in learning more about this condition, [including the standard course of treatment]$_{\text{EXEMPLIFICATION}}$.]$_{\text{QUESTION}}$
Question (4) S_1: [I am an adult woman who has been recently diagnosed with [Ménétrier disease]$_{\text{FOCUS}}$.]$_{\text{BACKGROUND(DIAGNOSIS)}}$ S_2: [My symptoms seem to be worsening.]$_{\text{BACKGROUND(SYMPTOM)}}$ S_3: [I am seeking information about the [symptoms, cause(s), prognosis, genetic association, and treatment]$_{\text{COORDINATION}}$.]$_{\text{QUESTION}}$ S_4: [I am eager to hear from you and appreciate your assistance.]$_{\text{IGNORE}}$

Table 1: Decomposed Questions

(1) BACKGROUND - a sentence-level annotation that indicates useful contextual information, but lacks a question. There are several sub-types of background information that we annotate as well: COMORBIDITY, DIAGNOSIS, FAMILY_HISTORY, LIFESTYLE, SYMPTOM, TEST, and TREATMENT. Since a sentence can contain more than one type of background information (or a type other than the seven listed), we consider these sub-types as attributes of BACKGROUND sentences.

(2) QUESTION - a sentence- or clause-level annotation that indicates a question. When a conjunction links two clauses, the conjunction is included with the second question as in Question (2) in Table 1.

(3) COORDINATION - a phrase-level annotation that spans the set of decomposable items, such as that in Questions (1), (2), and (4) in Table 1.

(4) EXEMPLIFICATION - an phrase-level annotation that spans an optional item, such as Question (3) in Table 1.

(5) IGNORE - a sentence-level annotation indicating nothing of value is present. See Question (4) in Table 1.

(6) FOCUS - a NP-level annotation indicating the theme of the complex question. Useful when filling ellipsis and substituting anaphora as in Question (2). For more information, see Kilicoglu et al. (2013). In the data discussed in this paper, this is always a disease.

BACKGROUND types are annotated at the sentence level for convenience and consistency, as many of the instances of background concepts do not easily lend themselves to precise boundaries (see Forbush et al. (2013)). The choice to separate clause-level QUESTION and phrase-level COORDINATION annotations enables easier construction of decomposed questions. In both cases, questions are only considered decomposable (i.e., a question sentence is split into multiple QUESTIONs or a phrase is annotated with a COORDINATION) when semantically valid decomposed questions would result. Consider the following two question sentences:

(5) Can this disease be cured or can we only treat the symptoms?

(6) Are males or females worse affected?

While Question (5) contains two "*Can...*" questions and Question (6) contains the coordination "*males or females*", both questions are actually providing a choice between two alternatives and decomposing them would alter the semantic nature of the original question. Thus both would be considered a single QUESTION with no COORDINATIONs.

4. Annotated Corpus

Here we provide a brief description of the annotated corpus and the process of creating it. The Genetic and Rare Diseases Information Center (GARD) maintains a website[1] that provides information on many diseases, as well as a contact form where consumers can ask medical questions. We collected 1,467 complex questions from this resource for annotation.

Each GARD request was annotated by two of three annotators, a computer scientist (KR), a medical librarian (KM), and an MD (MF). After the initial annotation and prior to conflict resolution, inter-annotator agreement numbers were calculated (see Table 2). The inter-annotator agreement numbers were particularly low. Individual annotators, however, were fairly consistent, implying intra-annotator agreement would be quite high. This suggests that relatively predictable differences existed between annotators. For the most part, QUESTION splitting, COORDINATION, and EXEMPLIFICATION disagreements were largely a function of one annotator missing the annotation. Because of this, an additional effort was made during the resolution process to go back through the data to reduce the number of missed annotations. For example, while only 41 EXEMPLIFICATIONs were found by at least one annotator during the initial annotation process, after resolution there were 53 total EXEMPLIFICATIONs due to instances where both annotators missed. As can be seen by the baseline results discussed below, resolution drastically improved the quality of the annotations, so inter-annotator agreement is not an indicator of the upper bound of each sub-task.

After annotation, we found an average of 1.7 decomposed questions per complex question, with 0.25 COORDINATIONs and 0.04 EXEMPLIFICATIONs per complex ques-

[1] http://rarediseases.info.nih.gov/gard

Annotation	# Items	# Disagreements	Accuracy	Kappa
Sentence Class	4,115	100	97.5	95.0
Background Class	1,643	455	71.8	63.5
Question Split	2,465	76	41.7	
Coordination	371	184	46.8	
Exemplification	41	15	53.1	
Focus	1,540	135	90.8	

Table 2: Initial inter-annotator agreement statistics.

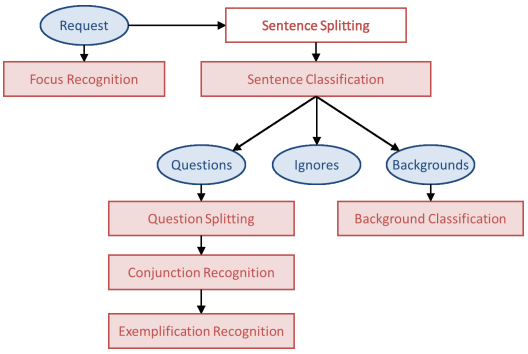

Figure 1: Question Decomposition Architecture

Task	Score
Sentence Classification	97.1%
Background Classification	67.3 F_1
Question Splitting	90.0 F_1
Coordination Recognition	71.9 F_1 (Relaxed) 31.5 F_1 (Exact)
Exemplification Recognition	70.5 F_1 (Relaxed) 64.5 F_1 (Exact)
Focus Recognition	88.9 F_1 (Relaxed) 55.9 F_1 (Exact)

Table 3: Baseline Results

tion. Additionally, there were an average of 1.2 BACKGROUND sentences and 0.03 IGNORE sentences per complex question. The BACKGROUND sentences included 23 COMORBIDITY, 690 DIAGNOSIS, 151 FAMILY_HISTORY, 13 LIFESTYLE, 320 SYMPTOM, 61 TEST, and 137 TREATMENT sentences. Finally, there were an average of 1.03 FOCUS annotations per complex question. A complex question typically has more than one FOCUS when the consumer is asking about the interaction or relation between two diseases.

The dataset is publicly available from our project webpage.[2]

5. Baseline Experiments

We propose a baseline method for automatically decomposing complex questions according to the annotations presented above. The architecture of the system is shown in Figure 1. Given a complex question, we first perform tokenization, sentence segmentation, and syntactic parsing with the Stanford Parser (Klein and Manning, 2003). Sentences are then classified using an SVM (Fan et al., 2008) with bag-of-words features into BACKGROUND, QUESTION, and IGNORE. For BACKGROUND sentences, we use binary SVM classifiers, with bag-of-words and bag-of-bigrams features, to identify the seven attributes. For QUESTION sentence splitting, we use a rule-based method based on the syntactic parse tree to identify clause-separating conjunctions followed by a question word (e.g., *what*, *how*). These rules were constructed by examining the syntactic parse trees of a handful of examples from the data, and did not involve any tuning on the entire dataset. For COORDINATION recognition, we use similar rules based on the syntactic parse tree to find NPs containing conjunctions. Additionally, simple three-word phrases where the left and right conjunct match parts-of-speech (e.g., "*prognosis and cause*", "*want and need*") are considered COORDINATIONS. For EXEMPLIFICATION, a small set of rules were used in combination with a set of trigger words/phrases (e.g., *such as*, *including*). When a trigger is the head of a prepositional or adjective phrase, the entire phrase is considered an EXEMPLIFICATION. Finally, to identify the FOCUS, we recognize UMLS (Lindberg et al., 1993) terms in the complex question, then use an SVM to rank these. As features for the ranker, we use the candidate words, UMLS category information (CUI and semantic type), the sentence offset, and the term's offset (the first UMLS term mentioned is often the FOCUS of the question). The results of these baselines are presented in Table 3. For the machine learning-based components, 5-fold cross validation is performed for evaluation.

The baseline method for BACKGROUND classification performed poorly, though it clearly performed better on the more common annotations (e.g., DIAGNOSIS classification has an F_1 of 80.5). It is clear the bag-of-words model is insufficient, especially as it cannot account for negation.

While the syntactic parse tree rules performed relatively well at identifying separate QUESTIONS, COORDINATION and EXEMPLIFICATION recognition were quite poor. The gap between the relaxed scores (where any overlap counts as a match) and the exact score (where a complete overlap is required) is related to the difficulty of syntactic parsing of coordinations. Coordination resolution is one of the most difficult tasks in syntactic parsing (Ogren, 2010). The false positives for coordination recognition typically involve coordinations that were semantically inseparable, as in Question (7), where the question requires all the conjuncts to retain its original purpose. False negatives usually involved conjuncts that were not NPs, such as Question (8). Relaxing from an NP rule produced worse results.

(7*) Can *lupus, antiphospholipid syndrome, and ITP antibodies* occur together?

(8) I have prolidase deficiency and would like to *receive as much information as possible as well as contact other sufferers*.

[2] http://lhncbc.nlm.nih.gov/project/consumer-health-question-answering/

The false positives for exemplification recognition indicate the rules cannot distinguish between an exemplifying phrase and the primary verb phrase, as in Question (9). While false negatives often involved exemplifications that did not employ trigger words, as in Question (10).

(9*) I am *particularly interested in learning more about genetic testing for the syndrome*.

(10) I am looking for any info about heavy metal toxicity, *symptoms, treatment, outcomes*.

The baseline FOCUS recognition method performed reasonably well using the relaxed method, while most of its loss with the exact metric came from not correctly recognizing the full span of the FOCUS. This is largely due to slight differences in how UMLS encodes concepts and how they were encountered in our data. For instance, UMLS contains the concept "*Ehlers Danlos, hypermobility type*", but not "*Ehlers Danlos hypermobility type*". Instead, the system guessed simply "*Ehlers Danlos*". Hopefully, solving minor differences such as this can dramatically improve automatic FOCUS recognition.

6. Conclusion

We have presented an annotated corpus and baseline methods for decomposing complex medical questions posed by consumers. We employ a largely syntactic approach to question decomposition, recognizing both clause-level and NP-level conjunctions. We also recognize several contextual elements within the complex question that provide useful background information. We then presented the details of our annotated corpus. Finally, we proposed and evaluated baseline methods for performing decomposition on complex medical questions.

For future work, we plan to develop ML-based methods for recognizing coordinations and exemplifications to overcome the difficulties of relying solely on the syntactic parse tree. Furthermore, we plan to improve upon the ML methods for the remaining annotations, notably by including negation, co-reference, and discourse information.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We would additionally like to thank Stephanie M. Morrison and Janine Lewis for their help accessing the GARD data.

7. References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., and Wedekind, J. (2012). Creation and use of Language Resources in a Question-Answering eHealth System. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2536–2542.

Bott, S., Saggion, H., and Mille, S. (2012). Text Simplification Tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1665–1671.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44:277–288.

Demner-Fushman, D. and Lin, J. (2007). Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–500.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Forbush, T. B., Gundlapalli, A. V., Palmer, M. N., Shen, S., South, B. R., Divita, G., Carter, M., Redd, A., Butler, J. M., and Samore, M. (2013). "Sitting on Pins and Needles": Characterization of Symptom Descriptions in Clinical Notes. In *AMIA Summit on Clinical Research Informatics*, pages 67–71.

Harabagiu, S., Lacatusu, F., and Hickl, A. (2006). Answer Complex Questions with Random Walk Models. In *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 220–227.

Hartrumpf, S. (2008). Semantic Decomposition for Question Answering. In *Proceedings on the 18th European Conference on Artificial Intelligence*, pages 313–317.

Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2013). Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In *Proceedings of BioNLP*, pages 54–62.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Lacatusu, F., Hickl, A., and Harabagiu, S. (2006). Impact of Question Decomposition on the Quality of Answer Summaries. In *Proceedings of LREC*, pages 1147–1152.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System: An informatics research collaboration. *Methods of Information in Medicine*, 32(4):281–291.

Ogren, P. (2010). Improving Syntactic Coordination Resolution Using Language Modeling. In *Proceedings of the NAACL HLT Student Research Workshop*, pages 1–6.

Schnall, J. G. and Fowler, S. (2013). MedlinePlus.gov: Quality Health Information for Your Patients. *American Journal of Nursing*, 113(9):64–65.

Yu, H. and Cao, Y. (2008). Automatically Extracting Information Needs from Ad Hoc Clinical Questions. In *Proceedings of the AMIA Annual Symposium*, pages 96–100.