Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis

Halil Kilicoglu, Marcelo Fiszman, Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, MD, USA
{kilicogluh, fiszmanm, ddemner}@mail.nih.gov

Abstract

While interest in biomedical question answering has been growing, research in consumer health question answering remains relatively sparse. In this paper, we focus on the task of consumer health question understanding. We present a rule-based methodology that relies on lexical and syntactic information as well as anaphora/ellipsis resolution to construct structured representations of questions (frames). Our results indicate the viability of our approach and demonstrate the important role played by anaphora and ellipsis in interpreting consumer health questions.

1 Introduction

Question understanding is a major challenge in automatic question answering. An array of approaches has been developed for this task in the course of TREC Question Answering evaluations (see Prager (2006) for an overview). These collectively developed approaches to question understanding were successfully applied and expanded upon in IBM's Watson system (Lally *et al.*, 2012). Currently, Watson is being retargeted towards biomedical question answering, joining the ongoing research in domain-specific question answering (for a review, see Simpson and Demner-Fushman, 2012).

Much research in automatic question answering has focused on answering well-formed factoid questions. However, real-life questions that need to be handled by such systems are often posed by lay people and are not necessarily well-formed or explicit. This is particularly evident in questions involving health issues. Zhang (2010), focusing on health-related questions submitted to Yahoo Answers, 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. For example, consider the following question posed by a consumer:

(1) my question is this: I was born w/a esophagus atresia w/dextrocardia. While the heart hasn't caused problems,the other has. I get food caught all the time. My question is...is there anything that can fix it cause I can't eat anything lately without getting it caught. I need help or will starve!

It is clear that the person asking this question is mainly interested in learning about treatment options for his/her disease, in particular with respect to his/her esophagus. Most of the textual content is not particularly relevant in understanding the question (*I need help or will starve*! or *I get food caught all the time*). In addition, note the presence of anaphora (*it* referring to *esophagus atresia*) and ellipsis (*the other has* *[caused problems]*), which should be resolved in order to automatically interpret the question. Finally, note the informal *fix* instead of the more formal *treat*, and *cause* instead of *because*.

The National Library of Medicine® (NLM®) receives questions from consumers on a variety of health-related topics. These questions are currently manually answered by customer support services. The overall goal of our work is to assist the customer support services by automatically interpreting these questions, using information retrieval techniques to find relevant documents and passages, and presenting the information in concise form for their assessment.

In this paper, we specifically focus on question understanding, rather than information re-

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 54–62,
Sofia, Bulgaria, August 4-9 2013.

trieval aspects of our ongoing work. Our goal in question understanding is to capture the core aspects of the question in a structured representation (*question frame*), which can then be used to form a query for the search engine. In the current work, we primarily investigate and evaluate the role of anaphora and ellipsis resolution in understanding the questions. Our results confirm the viability of rule-based question understanding based on exploiting lexico-syntactic patterns and clearly demonstrate that anaphora and ellipsis resolution are beneficial for this task.

2 Background

Despite the growing interest to biomedical question answering (Cairns *et al.*, 2012; Ni *et al.*, 2012; Bauer and Berleant, 2012), consumer health question answering remains a fairly understudied area of research. The initial research has focused on the analysis of consumer language (McCray *et al.*, 1999) and the types of questions they asked. Spink *et al.* (2004) found that health-related queries submitted to three web search engines in 2001 were often advice seeking and personalized, and fell into five major categories: general health, weight issues, reproductive health and puberty, pregnancy/obstetrics, and human relationships. Observing that health queries constituted no more than 9.5% of all queries and declined over time, they concluded that the users turn more to the specialized resources for the answers to health-related questions. Similar to the findings of Zhang (2010), Beloborodov *et al.* (2013) found that diseases and symptoms were the most popular topics in a resource similar to Yahoo Answers, Otvety@Mail.Ru. They analyzed Otvety@Mail.Ru questions by mapping questions to body parts and organs, applying Latent Dirichlet Allocation method with Gibbs sampling to discover topics, and using a knowledge-based method to classify questions as evidence-directed or hypothesis-directed.

First efforts in automated consumer health question processing were to classify the questions using machine learning techniques. In one study, frequently asked questions about diabetes were classified according to two somewhat orthogonal taxonomies: according to the "medical type of the question" (Causes, Diagnostic, Prevention, Symptoms, Treatment, etc.) and according to the "expected answer type" (Boolean, Causal, Definition, Factoid, Person, Place, etc.) (Cruchet *et al.*, 2008). Support Vector Machine (SVM) classification achieved an F-score in low 80s in classifying English questions to the expected answer type. The results for French and medical type classification in both languages were much lower. Liu *et al.* (2011) found that SVM trained to distinguish questions asked by consumers from those posed by healthcare professionals achieve F-scores in the high 80s - low 90s. One of distinguishing characteristics of the consumer questions in Liu *et al.*'s study was the significantly higher use of personal pronouns (compared to professional questions). This feature was found to be useful for machine learning; however, the abundance of pronouns in the long dense questions is also a potential source of failure in understanding the question.

Vicedo and Ferrández (2000) have shown that pronominal anaphora resolution improves several aspects of the QA systems' performance. This observation was supported by Harabagiu *et al.* (2005) who have manually resolved coreference and ellipsis for 14 of the 25 scenarios in the TREC 2005 evaluation. Hickl *et al.* (2006) have incorporated into their question answering system a heuristic based question coreference module that resolved referring expressions in the question series to antecedents mentioned in previous questions or in the target description. To our knowledge, coreference and ellipsis resolution has not been previously attempted in consumer health question understanding.

Another essential aspect in processing consumer questions is defining a formal representation capable of capturing all important points needed for further processing in automatic query generation (in the systems that use document passage retrieval to find a set of potential answers) and answer extraction and unification. Ontologies provide effective representation mechanisms for concepts, whereas relations are better captured in frame-like or event-related structures (Hunter and Cohen, 2006). Frame-based representation of extracted knowledge has a long-standing tradition in the biomedical domain, for example, in MedLEE (Friedman *et al.*, 1994). Demner-Fushman *et al.* (2011) showed that frame-based representation of clinical questions improve identification of patients eligible for cohort inclusion. Demner-Fushman and Abhyankar (2012) extracted frames in four steps: 1) identification of domain concepts, 2) extraction of patient demographics (e.g., age, gender) and social history, 3) establishing dependencies between the concepts using the Stanford dependency parser (de Marneffe *et al.*, 2006), and 4) adding concepts not involved in the relations to the

frame as a list of keywords. Event-based representations have also seen increasing use in recent years in biomedical text mining, with the availability of biological event corpora, including GENIA event (Kim *et al.*, 2008) and GREC (Thompson *et al.*, 2009), and shared task challenges (Kim *et al.*, 2012). Most state-of-the-art systems address the event extraction task by adopting machine learning techniques, such as dual composition-based models (Riedel and McCallum, 2011), stacking-based model integration (McClosky *et al.*, 2012), and domain adaptation (Miwa *et al.*, 2012). Good performance has also been reported with some rule-based systems (Kilicoglu and Bergler, 2012). Syntactic dependency parsing has been a key component in all state-of-the-art event extraction systems, as well. The role of coreference resolution in event extraction has recently been acknowledged (Kim *et al.*, 2012), even though efforts in integrating coreference resolution into event extraction pipelines have generally resulted in only modest improvements (Yoshikawa *et al.*, 2011; Miwa *et al.*, 2012; Kilicoglu and Bergler, 2012).

Coreference resolution has also been tackled in open domain natural language processing. State-of-the-art systems often employ a combination of lexical, syntactic, shallow semantic and discourse information (e.g., speaker identification) with deterministic rules (Lee *et al.*, 2011). Interestingly, coreference resolution is one research area, in which deterministic frameworks generally outperform machine learning models (Haghighi and Klein, 2009; Lee *et al.*, 2011).

In contrast to coreference resolution, ellipsis resolution remains an understudied NLP problem. One type of ellipsis that received some attention is *null instantiation* (Fillmore and Baker, 2001), whereby the goal is to recover the referents for an uninstantiated semantic role of a target predicate from the wider discourse context. A semantic evaluation challenge that focused on null instantiation was proposed, although participation was limited (Ruppenhofer *et al.*, 2010). Gerber and Chai (2012) focused on implicit argumentation (i.e., *null instantiation*) for nominal predicates. They annotated a corpus of implicit arguments for a small number of nominal predicates and trained a discriminative model based on syntactic, semantic and discourse features collected from various linguistic resources. Focusing on a different type of ellipsis, Bos and Spenader (2011) annotated a corpus of verb phrase ellipsis; however, so far there have been little work in verb phrase ellipsis resolution. We are also not aware of any work in ellipsis resolution in biomedical NLP.

3 Methods

We use a pipeline model for question analysis, which results in frame annotations that capture the content of the question. Our rule-based method begins with identifying terms (named entities/triggers) in question text. Next, we recognize anaphoric mentions and, if any, perform anaphora resolution. The next step is to link frame triggers with their *theme* and *question cue* by exploiting syntactic dependency relations. Finally, if frames with implicit arguments exist (that is, frames in which *theme* or *question cue* was not instantiated), we attempt to recover these arguments by ellipsis resolution. In this section, we first describe our data selection. Then, we explain the steps in our pipeline, with particular emphasis on anaphora and ellipsis. The pipeline diagram is illustrated in Figure 1.

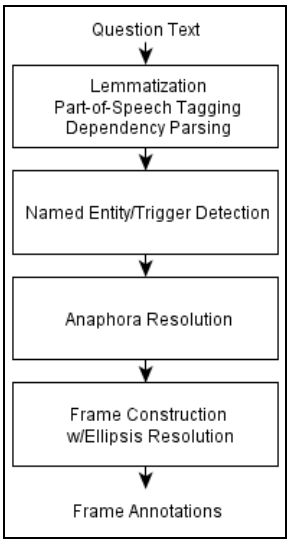

Figure 1. The system pipeline diagram

3.1 Data Selection and Annotation

In this study, we focused on questions about genetic diseases, due to their increasing prevalence. Since the majority of the consumers' questions submitted to NLM are about treatment and prognosis, we selected mainly these types of questions for our training set. Note that while these questions mostly focused on treatment and prognosis, some of them also include other types of questions, asking for general information or about diagnosis, etiology, and susceptibility (thus, confirming the finding of Zhang (2010)). The majority of selected questions were asked by real consumers in 2012. Due to our interest in genetics questions, we augmented this set with

some frequently asked questions from the Genetic and Rare Disease Information Center (GARD) [1]. Our selection yielded 32 treatment and 22 prognosis questions. An example treatment question was provided earlier (1). The following is a training question on prognosis:

(2) They have diagnosed my niece with Salla disease. I understand that this is a very rare disease and that its main origin is Finland. Can you please let me know what to expect? My niece is 7 years old. It has taken them 6 years to finally come up with this diagnosis.

We used training questions to gain linguistic insights into the problem, to develop and refine our methodology, and as the basis of a trigger/question cue dictionary.

After the system was developed, we selected 29 previously unseen treatment-focused questions posed to GARD for testing. We annotated them with target frames (41 instances) using *brat* annotation tool (Stenetorp *et al.*, 2012) and evaluated our system results against these frames. 29 of the target frames were treatment frames. Additionally, there were 1 etiology, 6 general information, 2 diagnosis, and 3 prognosis frames.

3.2 Syntactic Dependency Parsing

Our question analysis module uses typed dependency relations as the basis of syntactic information. We extract syntactic dependencies using Stanford Parser (de Marneffe *et al.*, 2006) and use its collapsed dependency format. We rely on Stanford Parser for tokenization, lemmatization, and part-of-speech tagging, as well.

3.3 Named Entity/Trigger Detection

We use simple dictionary lookup to map entity mentions in text to UMLS Metathesaurus concepts (Lindberg, 1993). So far, we have focused on recognizing three mention categories: problems, interventions, and patients. Based on UMLS 2007AC release, we constructed a dictionary of string/concept pairs. We limited the dictionary to concepts with predefined semantic types. For example, all problems in the dictionary have a semantic type that belongs to the Disorders semantic group (McCray *et al.*, 2001), such as Neoplastic Process and Congenital Abnormality. Currently our dictionary contains approximately 260K string/concept pairs.

Dictionary lookup is also used to detect triggers and question cues. We constructed a trigger and question cue dictionary based on training data and limited expansion. The dictionary currently contains 117 triggers and 14 question cues.

3.4 Recognizing Anaphoric Mentions

We focus on identifying two types of anaphoric phenomena: *pronominal anaphora* (including anaphora of personal and demonstrative pronouns) and *sortal anaphora*. The following examples from the training questions illustrate these types. 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.*

To recognize mentions of personal pronominal and sortal anaphora, we mainly adapted the rule-based techniques outlined in Kilicoglu and Bergler (2012), itself based on the deterministic coreference resolution approach described in Haghighi and Klein (2009). While Kilicoglu and Bergler (2012) focused on anaphora involving gene/protein terms, our adaptation focuses on those involving problems and patients. In addition, we expanded their work by developing rules to recognize demonstrative pronominal anaphora.

3.4.1 Personal Pronouns

Kilicoglu and Bergler (2012) focused on only resolving *it* and *they*, since, in scientific article genre, resolving other third person pronouns (*he, she*) was less relevant. We currently recognize these two pronouns, as well. For personal pronouns, we merely tag the word as a pronominal anaphor if it is tagged as a pronoun and is in third person (i.e., *she, he, it, they*).

3.4.2 Demonstrative Pronouns

We rely on typed syntactic dependencies as well as part-of-speech tags to recognize demonstrative pronominal anaphora. A word is tagged as demonstrative pronominal anaphor if it is one of *this, that, those,* or *these* and if it is not the de-

[1] https://rarediseases.info.nih.gov/GARD/

pendent in a *det* (determiner) dependency (in other words, it is not a pronominal modifier). Furthermore, we ensure that the pronoun *that* does not act as a complementizer, requiring that it not be the dependent in a *complm* (complementizer) dependency.

3.4.3 Sortal Anaphora

In the current work, we limited sortal anaphora to problem terms. As in Kilicoglu and Bergler (2012), we require that the anaphoric noun phrases not include any named entity terms. Thus, we allow *the syndrome* as an anaphoric mention, while blocking *the Stickler syndrome*. To recognize sortal anaphora, we look for the presence of *det* dependency, where the dependent is one of *this, that, these, those,* or *the*.

Once the named entities, question cues, triggers, and anaphoric mentions are identified in a sentence, we collapse the syntactic dependencies from the sentence to simplify further processing. This is illustrated in Table 1 for the sentence in (3).

(3) My partner is a carrier for Simpson-Golabi-Behmel syndrome and her son was diagnosed with this rare condition.

Dependencies before	**Dependencies after**
amod (syndrome, simpson-golabi-behmel)	*prep_for(carrier, simpson-golabi-behmel syndrome)*
prep_for(carrier,syndrome)	
det(condition,this)	*prep_with (diagnosed, this rare condition)*
amod(condition, rare)	
prep_with(diagnosed, condition)	

Table 1: Syntactic dependency transformations

3.5 Anaphora Resolution

Anaphora resolution is the task of finding the antecedent for an anaphoric mention in prior discourse. Our anaphora resolution method is again based on the work of Kilicoglu and Bergler (2012). However, we made simplifying assumptions based on our examination of the training questions. First observation is that each question is mainly about one salient topic (problem) and anaphoric mentions are highly likely to refer to this topic. Secondly, the salient topic often appears as the first named entity in the question. Based on these observations, we did not attempt to use the relatively complex, semantic graph-based resolution strategies (e.g., graph distance) outlined in that work. Furthermore, we have not attempted to address *set-instance anaphora* or *event anaphora* in this work, since we did not see examples of these in the training data.

Anaphora resolution begins with identifying the candidate antecedents (problems, patients) in prior discourse, which are then evaluated for syntactic and semantic compatibility. For pronominal anaphora, compatibility involves person and number agreement between the anaphoric mention and the antecedent. For sortal anaphora, number agreement as well as satisfying one of the following constraints is required:

- *Head word constraint:* The head of the anaphoric NP and the antecedent NP match. This constraint allows *Wolf-Hirschhorn Syndrome* as an antecedent for *this syndrome,* matching on the word *syndrome*.
- *Hypernymy constraint:* The head of the anaphoric NP is a problem hypernym and the antecedent is a problem term. Similar to gene/protein hypernym list in Kilicoglu and Bergler (2012), we used a small list of problem hypernym words, including *disease, disorder, illness, syndrome, condition,* and *problem.* This constraint allows *Simpson-Golabi-Behmel syndrome* as an antecedent for *this rare condition* in example (3).

We expanded number agreement test to include singular mass nouns, so that plural anaphora (e.g., *they*) can refer to mass nouns such as *family, group, population*. In addition, we defined lists of gendered nouns (e.g., *son, father, nephew,* etc. for male and *wife, daughter, niece,* etc. for female) and required gender agreement for pronominal anaphora.

After the candidate antecedents are identified, we assign them *salience scores* based on the order in which they appear in the question and their frequency in the question. The terms that appear earlier in the question and occur more frequently receive higher scores. The most salient antecedent is then taken to be the coreferent.

3.6 Frame Construction

We adapted the frame extraction process based on lexico-syntactic information outlined in Demner-Fushman *et al.* (2012) and somewhat

modified the frames to accommodate consumer health questions. For each question posed, we aim to construct a frame which consists of the following elements: *type*, *theme*, and *question cue*: *theme* refers to the topic of the question (problem name, etc.), while *type* refers to the aspect of the theme that the question is about (treatment, prognosis, etc.) and question cue to the question words (*what, how, are there*, etc.). Theme element is semantically typed and is restricted to the UMLS semantic group Disorders. From the question in (1), the following frame should be extracted:

Treatment	*fix*
Theme	*Esophageal atresia (Disease or Syndrome)*
QCue	*Is there*

Table 2: Frame example

We rely on syntactic dependencies to link frame indicators to their themes and question cues. We currently search for the following types of syntactic dependencies between the indicator mention and the argument mentions: *dobj* (direct object), *nsubjpass* (passive nominal subject), *nn* (noun compound modifier), *rcmod* (relative clause modifier), *xcomp* (open clausal complement), *acomp* (adjectival complement), *prep_of, prep_to, prep_for, prep_on, prep_from, prep_with, prep_regarding, prep_about (*prepositional modifier cued by *of, to, for, on, from, with, regarding, about*, respectively). Two special rules address the following cases:

- If the dependency exists between a trigger 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.*.' where there is a *prep_regarding* dependency between the General Information trigger '*information*' and the Prognosis trigger '*prognosis*'. This results in '*information*' becoming the question cue for the Prognosis frame.
- 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 *life expectancy*.

3.6.1 Ellipsis Resolution

The frame construction step may result in frames with uninstantiated themes or question cues. If a constructed frame includes a question cue but no theme, we attempt to recover the theme argument from prior discourse by ellipsis processing. Consider the question in (4) and the frame in Table 3 extracted from it in previous steps:

(4) They have diagnosed my niece with Salla disease. …Can you please let me know what to expect? …

Prognosis	*expect*
Theme	-
QCue	*what*

Table 3: Frame with uninstantiated Theme role

In the context of consumer health questions, the main difficulty with resolving such cases is recognizing whether it is indeed a legitimate case of ellipsis. We use the following dependency-based heuristics to determine the presence of ellipsis:

- Check for the presence of a syntactic dependency of one of the types listed in Section 3.5, in which the frame trigger appears as an element. If such a dependency does not exist, consider it a case of ellipsis.
- Otherwise, consider the other element of the dependency:
    - If the other element does not correspond to a term, we cannot make a decision regarding ellipsis, since we do not know the semantics of this other element.
    - If it corresponds to an element that has already been used in creating the frame, the dependency is accounted for.
- If all the dependencies involving the frame trigger are accounted for, consider it a case of ellipsis.

In example (4), the trigger *expect* is found to be in an *xcomp* dependency with the question cue *know*, which has already been used in the frame. Therefore this dependency is accounted for, and we consider this a case of ellipsis. On the other hand, consider the example:

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

As in the previous example, the Theme role of the Prognosis frame indicated by *expect* is uninstantiated. However, it is not considered an ellip-

tical case, since there is a *prep_for* dependency between *expect* and *future*, a word that is semantically unresolved.

Once the presence of ellipsis is ensured, we fill the Theme role of the frame with the most salient term in the question text, as in anaphora resolution.

In rare cases, the frame may include a theme but not a question cue. This may be due to a lack of explicit question expression (such as in the question '*treatment for Von Hippel-Lindau syndrome.'*) or due to shortcomings in dependency-based linking of frame triggers to question cues. If no fully instantiated frame was extracted from the question, as a last resort, we construct a frame without the question cue in an effort to increase recall.

4 Results and Discussion

We extracted frames from the test questions and compared the results with the annotated target frames. As evaluation metrics, we calculated precision, recall, and F-score. To assess the effect of various components of the system, we evaluated several scenarios:

- Frame extraction without anaphora/ellipsis resolution (indicated as A in Table 4 below)
- Frame extraction with anaphora/ellipsis resolution (B)
- Frame extraction without anaphora/ellipsis resolution but with gold triggers/named entities (C)
- Frame extraction with anaphora/ellipsis resolution and gold triggers/named entities (D)

The evaluation results are provided in Table 4. In the second column, the numbers in parentheses correspond to the numbers of correctly identified frames.

	# of frames	*Recall*	*Precision*	*F-score*
A	14 (13)	0.32	0.93	0.48
B	26 (22)	0.54	0.85	0.66
C	17 (16)	0.39	0.84	0.55
D	35 (33)	0.80	0.94	0.86

Table 4: Evaluation results

The evaluation results show that the dependency-based frame extraction method with dictionary lookup is generally effective; it is precise in identifying frames, even though it misses many relevant frames, typical of most rule-based systems. On the other hand, anaphora/ellipsis resolution helps a great deal in recovering the relevant frames and only has a minor negative effect on precision of the frames, the overall effect being significantly positive. Note also that the increase in recall without gold triggers/named entities is about 40%, while that with gold triggers/named entities is more than double, indicating that accurate term recognition contributes to better anaphora/ellipsis resolution and, in turn, to better question understanding.

The dictionary-based named entity/trigger/question cue detection is relatively simple, and while it yields good precision, the lack of terms in the corresponding dictionary causes recall errors. An example is given in (6). The named entity *Reed syndrome* was not recognized due to its absence in the dictionary, causing two false negative errors.

(6) A friend of mine was just told she has Reed syndrome… I was wondering if you could let me know where I can find more information on this topic. I am wondering what treatments there are for this, …

Similarly, dependency-based frame construction is straightforward in that it mostly requires direct dependency relations between the trigger and the arguments. While the two additional rules we implemented redress the shortcomings of this straightforward approach, there are cases in which dependency-based mechanism is still lacking. An example is given in (7). The lack of a direct dependency between *treatments* and *this condition* causes a recall error. A more sophisticated mechanism based on dependency chains could recover such frames; however, such chains would also increase the likelihood of precision errors.

(7) Are people with Lebers hereditary optic neuropathy partially blind for a long period of time …. ?Are there any surgical treatments available to alter this condition or is it permanent for life?

Anaphora/ellipsis processing clearly benefited our question understanding system. However, we noted several errors due to shortcomings in this processing. For example, from the sentence in (8), the system constructed a General Information frame with the trigger *wonder* and the Theme argument *central core disease*, which caused a false positive error.

(8) After 34 years of living with central core disease, …. My lower back doesn't seem to work, and I wonder if I will ever be able to walk up stairs or run.

The system recognized that the trigger *wonder* had an uninstantiated theme argument, which it attempted to recover by ellipsis processing. However, this processing misidentified the case as legitimate ellipsis due to the dependency relations *wonder* is involved in. A more sophisticated approach would take into account specific selectional restrictions of predicates like *wonder;* however, the overall utility of such linguistic knowledge in the context of consumer health questions, which are often ungrammatical and not particularly well-written, remains uncertain.

Our anaphora resolution method was unable to resolve some cases of anaphora. For example, consider the question in (6). The anaphoric mention *this topic* corefers with *Reed syndrome*. However, we miss this anaphora since we did not consider *topic* as a problem hypernym in scenario D, in which gold named entities are used.

5 Conclusions and Future Work

We presented a rule-based approach to consumer health question understanding which relies on lexico-syntactic information and anaphora/ellipsis resolution. We showed that lexico-syntactic information provides a good baseline in understanding such questions and that resolving anaphora and ellipsis has a significant impact on this task.

With regard to question understanding, future work includes generalization of the system to questions on topics other than genetic disorders (e.g., drugs) and aspects (such as complications, prevention, ingredients, location information, etc.) and broader evaluation. We also plan to automate dictionary development to some extent and address misspellings and acronyms in questions. We have been extending our frames to include ancillary keywords (named entities extracted from the question) that are expected to assist the search engine in pinpointing the relevant answer passages, similar to Demner-Fushman and Abhyankar (2012). We will also continue to develop our anaphora/ellipsis processing module, addressing the issues revealed by our evaluation as well as other anaphoric phenomena, such as recognition of pleonastic *it*.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Michael A. Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Human Genomics*, 6(1):17.

Alexander Beloborodov, Artem Kuznetsov, Pavel Braslavski. 2013. Characterizing Health-Related Community Question Answering. In *Advances in Information Retrieval*, 680-683.

Brian L. Cairns, Rodney D. Nielsen, James J. Masanz, James H. Martin, Martha S. Palmer, Wayne H. Ward, Guergana K. Savova. 2011. The MiPACQ clinical question answering system. In *AMIA Annual Symposium Proceedings*, pages 171-180.

Sarah Cruchet, Arnaud Gaudinat, Célia Boyer. 2008. Supervised approach to recognize question type in a QA system for health. *Studies in Health Technology and Informatics*, 136:407-412.

Marie-Catherine de Marneffe, Bill MacCartney, Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449-454.

Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F. Loane, Bastien Rance, François-Michel Lang, Nicholas C. Ide, Emilia Apostolova, Alan R. Aronson. 2011. A Knowledge-Based Approach to Medical Records Retrieval. In *Proceedings of Text Retrieval Conference 2011*.

Dina Demner-Fushman and Swapna Abhyankar. 2012. Syntactic-Semantic Frames for Clinical Cohort Identification Queries. *Lecture Notes in Computer Science, 7348*:100-112.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources.*

Carol Friedman, Philip O. Alderson, John HM Austin, James J. Cimino, and Stephen B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2): 161-174.

Matthew S. Gerber and Joyce Y. Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics*, 38(4): 755-798.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP 2009*, pages 1152-1161.

Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, Andrew Hickl, Patrick Wang. 2005. Employing two question answering systems in TREC-2005. In *Proceedings of Text Retrieval Conference 2005.*

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Ying Shi, Bryan Rink. 2006. Question Answering with LCC's CHAUCER at TREC 2006. In *Proceedings of Text Retrieval Conference 2006*.

Lawrence Hunter and Kevin B. Cohen. 2006. Biomedical language processing: what's beyond PubMed? *Molecular Cell*. 21(5):589-94.

Halil Kilicoglu and Sabine Bergler 2012. Biological Event Composition. *BMC Bioinformatics*, 13 (Supplement 11):S7.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, Akinori Yonezawa. 2012. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13(Supplement 11):S1.

Adam Lally, John M. Prager, Michael C. McCord, Branimir Boguraev, Siddharth Patwardhan, James Fan, Paul Fodor, Jennifer Chu-Carroll. 2012. Question analysis: How Watson reads a clue. *IBM Journal of Research and Development*, 56(3):2.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the CoNLL-2011 Shared Task*, pages 28-34.

Donald A.B. Lindberg, Betsy L. Humphreys, Alexa T. McCray. 1993. The Unified Medical Language System. *Methods of information in medicine*, 32(4): 281-291.

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

David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, Christopher Manning. 2012. Combining joint models for biomedical event extraction. *BMC Bioinformatics*, 13 (Supplement 11): S9.

Alexa McCray, Russell Loane, Allen Browne, Anantha Bangalore. 1999. Terminology issues in user access to Web-based medical information. In *AMIA Annual Symposium Proceedings*, pages 107–111.

Alexa McCray, Anita Burgun, Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proceedings of Medinfo*, 10(Pt1): 216-220.

Makoto Miwa, Paul Thompson, Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759-1765.

Yuan Ni, Huijia Zhu, Peng Cai, Lei Zhang, Zhaoming Qui, Feng Cao. 2012. CliniQA: highly reliable clinical question answering system. *Studies in Health Technology and Information*, 180:215-219.

John M. Prager. 2006. Open-domain question answering. *Foundations and Trends in Information Retrieval*, 1(2):91-231.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of EMNLP 2011*, pages 1-12.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45-50.

Matthew S. Simpson and Dina Demner-Fushman. 2012. Biomedical Text Mining: a Survey of Recent Progress. *Mining Text Data 2012*:465-517.

Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, H. Cenk Ozmutlu. 2004. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*. 21(1):44-51.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii. 2012. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Sessions at EACL 2012*, pages 102-107.

Paul Thompson, Syed A. Iqbal, John McNaught, Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349.

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

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, Yuji Matsumoto. 2011. Coreference Based Event-Argument Relation Extraction on Biomedical Text. *Journal of Biomedical Semantics*, 2 (Supplement 5):S6.

Yan Zhang. 2010. 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)*, pages 210-219.