InfoForager: Leveraging Semantic Search with AMR for COVID-19 Research

Claire Bonial, Stephanie M. Lukin, David C. Doughty, Steven C. Hill, Clare R. Voss
U.S. Army Research Lab
claire.n.bonial.civ@mail.mil

Abstract

This paper examines how Abstract Meaning Representation (AMR) can be utilized for finding answers to research questions in medical scientific documents, in particular, to advance the study of UV (ultraviolet) inactivation of the novel coronavirus that causes the disease COVID-19. We describe the development of a proof-of-concept prototype tool, InfoForager, which uses AMR to conduct a semantic search, targeting the *meaning* of the user question, and matching this to sentences in medical documents that may contain information to answer that question. This work was conducted as a sprint over a period of six weeks, and reveals both promising results and challenges in reducing the user search time relating to COVID-19 research, and in general, domain adaption of AMR for this task.

1 Introduction

UV light can inactivate viruses by making them unable to infect cells, thereby reducing the transmission of viral diseases. While a wealth of literature pertaining to UV inactivation of viruses exists, searching this literature for trustworthy and relevant information can be difficult and inefficient. This difficulty can be especially evident when focusing on diseases such as COVID-19 (caused by the novel coronavirus, SARS-CoV-2[1]), which can be transmitted via aerosols or droplets in a complex process spanning disciplines ranging from physiology to optics to fluid mechanics. There are many unknowns, poorly quantified parameters, and even confusions resulting from differing terminology used. Nonetheless, efficiently finding needed information may be critical in discovering improved methods, such as the use of germicidal UV,[2] to reduce the transmission of COVID-19 and other diseases.

This paper identifies an opportunity for NLP tools to aid in automatically sifting through this mass of documents to find relevant answers to specific and targeted questions from subject matter experts working in the space of UV inactivation of viruses. We introduce InfoForager, a proof-of-concept prototype tool that utilizes semantic understanding and search, going beyond the words in a user question to focus on its *meaning*. By employing a semantic search, we hypothesize that the user will more easily search through medical documents because they do not need to rephrase their questions (for example, into keywords) to conform to the system's search limitations and capabilities. InfoForager first parses a user research question into Abstract Meaning Representation (AMR) (Banarescu et al., 2013), then compares the resulting AMR query to a collection of medical research papers already parsed into AMR. All AMR query-sentence pairs in each paper are scored for their semantic similarity, and InfoForager returns the highest-ranking answer sentence and the source document.

Given the urgent nature of this research, we allotted six weeks during which four NLP researchers worked with two UV inactivation researchers to perform a shallow pass through the semantic searching problem space. Additionally, we worked with test users to obtain an understanding of the system requirements

[1]SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2, first identified in 2019.

[2]Germicidal UV: Also known as UVC, relating to the UV spectrum between 200 - 280 nm.

Proceedings of the 2nd International Workshop on Designing Meaning Representations, pages 67–77
Barcelona, Spain (Online), December 13, 2020

and developed a prototype framework to address those needs. Section 2 describes the UV inactivation problem space. Our semantic search approach is described in Section 3, and Section 4 details a Wizard-of-Oz prototyping user study of the framework and the initial development of the prototype system. An evaluation of the prototype and the results are discussed in Section 5.

2 Background: UV Inactivation Research

Virus particles (virions) can remain infectious on surfaces (e.g., walls, doorknobs, and masks), and may be contained within sneezed or coughed droplets of airway (e.g., mouth, nose, throat) fluids. Such virions can be inactivated by treating them and particles containing them with sufficient UV light (Sagripanti and Lytle, 2011); however, it remains unclear whether, and to what extent, the virions within dried droplets, or particles, are shielded from UV. Researchers need to know how the optical properties of a particle of such respiratory fluids affect the intensity of UV in a virion within that particle. If the particle of dried respiratory fluids actually protects virions within it from UV, then how much more intense does the UV source need to be to achieve adequate inactivation of the virions? Furthermore, what techniques could be used to increase the effectiveness of UV light for inactivation of SARS-CoV-2 in such particles? Given the tremendous variation in sizes, shapes, and compositions of dried respiratory particles, answering these questions by experimentation alone is far too expensive. Having a computational model to accurately represent intensities within UV-illuminated respiratory particles would be beneficial for testing multiple use-cases, and may be critical in discovering improved methods for using germicidal UV to reduce the transmission of COVID-19 and other diseases. Such modeling requires knowledge of the optical properties of airway fluids, which can be estimated from the concentrations and optical properties of the materials in these fluids. However, to date and to our knowledge, there are no medical papers or reports that include all the optically relevant materials in any airway fluids to create such a model.

We designed InfoForager with the aforementioned challenges in mind. The UV researchers and NLP researchers iteratively refined the general topic of UV inactivation of coronavirus into two main research questions of interest, which were used in developing and testing InfoForager:

Q_1 *Which (bodily) fluids have higher concentrations of SARS-CoV-2 particles?*

Q_2 *What is the range of sizes of respiratory droplets, specifically from coughing and sneezing?*

For the purposes of training participants in our Wizard-of-Oz user study, we introduced a basic question:

Q_t *What types of UV light are used in coronavirus research?*

The UV researchers made clear that they were not seeking a system that returned a single answer to their research questions, as there are often multiple, even conflicting answers in cutting-edge medical research. Instead, they desired a system that returns all relevant results and that also clearly points to where the relevant content is in the document, allowing them to quickly assess where the information they are searching for is discussed in the approach, background, or results section of a paper.

This approach stands in contrast to search systems that the UV researchers have become accustomed to and use primarily for building their initial pool of medical documents, or abstracts when the documents are behind paywalls. For example, PubMed is as an extensive, publicly available, online, searchable database of medical and life sciences research abstracts.[3] In this study, we rely on it as a benchmark resource, given its broad and updating coverage of research papers, and make use of its keyword search, wherein terms from the user query are matched against subject designations for articles, the title, abstract, and author names. If no matches are made, the search terms are broken apart and the search process is repeated with the individual terms. This, indirectly, has led researchers away from posing natural language questions. When backing off to break such questions into keywords, PubMed may return either too many results for very general keywords or no results for more targeted combinations. For example, Q_1 yields zero results, and subsequent alterations of the question, e.g., *Which fluids higher SARS-CoV-2 particles* and *Which fluids higher concentration particles*, yield 1 and 1,157 individual research papers respectively. This approach places the burden on the user to determine which terms in the query are the most important for the system to function.

[3] `https://pubmed.ncbi.nlm.nih.gov`

3 InfoForager Approach: Semantic Parse and Search

InfoForager was designed with the UV researchers' goals, criteria, and the current tool limitations in mind. Given that they expect to conduct an iterative search process (to "forage"), where they may find partial information, including different and possibly discrepant values (especially as new findings emerge), we start with the following working hypothesis: an AMR-based approach can assist in identifying such partial information based on where their questions share semantics with sentences within the documents in their collection. Thus, we aim to create a system that accepts natural language questions, interprets their underlying meaning as AMR, detects a range of answers, and finally points to where that information was found in research papers. Our semantic search interaction is presented in Figure 1. First, a user provides a question, which is semantically parsed into AMR (left panel of Figure 1). Second, a semantic search is conducted that matches the parsed query against a collection of medical research papers that had also been semantically parsed (center panel). Finally, the results are scored and ranked, and the top responses along with the source papers are returned (right panel). The remainder of this section describes the parsing, and search and rank processes.

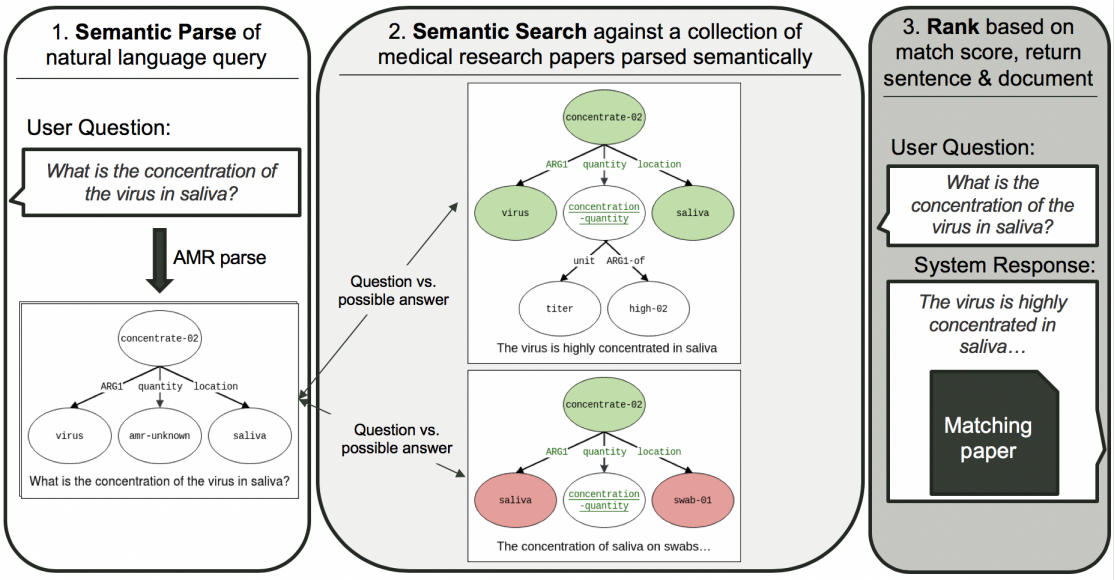


Figure 1: InfoForager overview—(1) User question is automatically parsed into AMR (parsed-query); (2) Parsed-query is compared to a collection of research papers already parsed into AMR (parsed-answers); (3) Matches are ranked and highest ranking sentence is returned with its source document.

3.1 Semantic Parse Using AMR

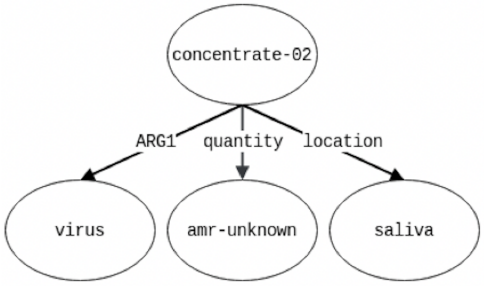
AMR is a directed, acyclic graph (DAG) representation of the meaning of a sentence, in which nodes map to words in the sentence, and edges map to the relations between them. Figure 2 shows the AMR in the text-based Penman notation (a) and the graph notation (b). There is a relatively large and active body of research surrounding AMR, such that there are a variety of parsers for automatically converting natural language text into AMR, including our own work to retrain and adapt various AMR parsers for dialogue systems (Bonial et al., 2019; Bonial et al., 2020). AMR has demonstrated value in biomedical NLP applications in the past (see Section 6), so we elected to explore the use of AMR in the development of a research framework that has the potential to match not only the concepts within a research AMR query, but also the relations between those concepts for more efficient "semantic search."

To obtain AMR parses of our user questions and the medical research document collection, we leverage the parser from Lindemann et al. (2019) based on its high performance after retraining within a new domain in previous research. The parser was retrained on the Linguistic Data Consortium's AMR 3.0 corpora (LDC2020T02) and the freely available Bio-AMR corpus,[4] as well as our own manually annotated dataset of approximately 1,000 AMRs drawn from the Dial-AMR corpus (Bonial et al., 2020).

[4] https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt

```
(a) (c / concentrate-02
      :ARG1 (v / virus)
      :location (s / saliva)
      :quant (a / amr-unknown))
```

(a) AMR in Penman notation



(b) AMR in directed, a-cyclic graph notation

Figure 2: The utterance *What is the concentration of the virus in saliva?* represented in (a) Penman notation and (b) its equivalent, in directed, a-cyclic graph notation.

Given time constraints of this six-week research effort, we opted to rely on our prior evaluations of several AMR parsers in selecting this parser and worked with a limited set of manual ground-truth AMRs for the project assessments.

Once the document collection was tokenized and segmented, we used the retrained parser to obtain a set of AMR graphs for each sentence of the collection. In addition, we parsed Q_t, Q_1, and Q_2, defined in Section 2, into AMR to have automatically-obtained AMRs of the questions. An experienced AMR annotator also provided gold standard, ground-truth AMRs for the questions and portions of the document collection with information relevant to answering the questions. This resulted in a collection of gold-standard AMR parses for the user questions and answers, as well as automatically-obtained AMR parses for the same user questions and answers, henceforth referred to as "gold-queries" and "gold-answers," and "parsed-queries" and "parsed-answers," respectively.

3.2 Semantic Search and Rank Using Graph Matching

To leverage the structured semantic information of AMR for determining if a document contains an answer to a question, we explored how adequately graph matching could serve to find a sentence that best addresses a particular question. We used the Smatch (semantic match) metric (Cai and Knight, 2013), which converts two input AMRs into two sets of node/edge triples, and measures the overlap between two resulting sets of triples. In our study applying Smatch, the AMR parsed-query was compared to all AMR parsed-answers in the document collection, then a list of the sentences ranked by Smatch scores was returned with IDs of documents containing those sentences for the top matches.[5]

To our knowledge, this is the first application of Smatch for semantic overlap of AMRs for a question and a sentence, as it was originally designed to find the closest match between AMRs for two sentences, one a ground-truth and the other a system output. We hypothesized that a strength of our novel application of Smatch would be in its ability to locate relevant information in other sentences that contain not only the same words and concepts as the original question, but also the same semantic relations between those concepts. However, a conceptual shortcoming is that Smatch is not necessarily finding *answers*, but finding the sentences with content most similar to that in the question, as it was designed to do. For example, were the question itself written out in a research paper, then it, and not an answer, would be returned as the best match.[6] Thus, we sought to identify the range or threshold of Smatch scores that would indicate when the AMRs of a question and document sentence were similar enough to capture shared semantics of a question-answer (Q-A) pair. For example, if we compare the parsed-query in Figure 1 for *What is the concentration of the virus in saliva?* to a parsed-answer (*The virus is highly concentrated in saliva*), this Q-A pair receives an Smatch F-score of 78%. Inspection of the Q-A graphs underscores this level of matching: the Q-A graphs are identical, with the exception of the `amr-unknown` node sitting in the position of the question word in the parsed-query; in the answer

[5] Smatch leverages smart initialization and 4 random restarts, where the highest overall score is reported as the final Smatch score. As a result, we observed some variation in Smatch scores when the same graphs are compared in multiple runs.

[6] Given that the UV researchers also track who is working on problems like theirs, even a question match is valuable: the document matched provides provenance to track the authors and institutions where relevant research is being conducted.

graph, it appears as a `concentration-quantity` node with additional modifiers. Thus, we argue that the graph-matching process is a suitable starting point for exploring the effectiveness of AMR in our semantic search framework, and describe in Section 7 other possible approaches.[7]

4 User Studies and Proof-of-Concept Development

Having developed the initial framework for semantic search to address the needs of UV researchers, we conducted a Wizard-of-Oz user study (Section 4.1) to explore how non-expert users interacted as research assistants with the system, while we developed the proof-of-concept prototype (Section 4.2).

4.1 Wizard-of-Oz User Studies

We conducted user studies of InfoForager's semantic search and PubMed, as noted in Section 2, for its keyword-based search. To put both systems on equal footing while InfoForager was under development and lacked a user interface, we deployed a simple chat interface and "Wizard-of-Oz" setup as the same front-end to both. Participants placed their search input in a text box, and a "Search Engine" (a human experimenter) retrieved a summary webpage listing the top five matching research papers from one of the search systems. The participant could then ask to view the individual results for further information in a listed research paper. Participants could remove or re-order words in the research question, but could not add or change words. In this way, for PubMed, we controlled for possible permutations of the research questions, enabling us to run the queries in advance, and return to participants actual PubMed results (as if run in real time). Results were displayed in a ranked list with the article title and authors displayed, along with a snippet of the abstract. This included returning no results when PubMed did not find a match. Similarly, for InfoForager, the "Search Engine" returned a summary webpage listing a ranked set of results with the title of the document, but also the relevant/matching text and what section of the document that text comes from (e.g., abstract, approach, results, etc.). When a user asks to see the source document for a particular summary result, the user is then shown only the abstract for the PubMed system (in keeping with actual follow-on phase in PubMed[8]), whereas for the prototype InfoForager, the user is shown the entire section that the result came from and the matching phrase is outlined in red.

Four subjects without a background in medicine or biology were recruited for the user study. They were told they would be assisting subject matter experts in answering research questions related to the coronavirus. Each participant underwent a training phase to become familiar with both search systems as well as the subject area using Q_t. During training, participants were given ten minutes to attempt to answer the question by searching a collection of medical literature using InfoForager, and then given another ten minutes to search for the answer to the same question using PubMed. Following that phase, there were two main trials, up to twenty minutes each, for answering Q_1 and Q_2. For the main trials, we alternated the order of which search engine was used first; two subjects started with InfoForager, while the others started with PubMed, and then their systems were swapped on the next question.

From the user studies, we observed that participants were able to find answers to their questions more quickly with the InfoForager framework—they felt more confident that they had looked through the collection as extensively as they needed to and therefore tended to declare their task complete at an earlier point. With InfoForager, participants could enter a natural language question, and the matches detected were similar not only in the words used, but also the semantic relationship between those words. In contrast, after discovering that they might receive no results for natural language questions with the PubMed system, participants would quickly turn to a keyword strategy, though this led, at times, to too many results that were not relevant. As a keyword search system, PubMed did not offer explicit indicators for where or why a match occurred. Overall, participants spent more time, in comparison to InfoForager, in reading through the entire returned abstract to determine if it was relevant, or trying different permutations of keywords to see if they could find the right combination to get their answer. Thus, the user studies established motivating evidence for a prototype tool with semantic search, described next.

[7] Several new AMR measures now also exist (Anchieta et al., 2019; Song and Gildea, 2019; Opitz et al., 2020).

[8] For some documents, users may be able click an additional link to access the full text.

4.2 Proof-of-Concept Prototype

After running user studies, we put the anticipated components together into an end-to-end, proof-of-concept system that we could begin to formally evaluate. Prior to any real-time user interaction, the medical document collection of papers are preprocessed, including conversion from PDF to text using the Poppler PDF[9] rendering library, and word tokenization and sentence segmentation using a toolkit developed internally within our team. The documents in our collection had very different formats (e.g., multiple columns or one) which could result in the conversion to text being out of order. Furthermore, the documents had a variety of abbreviations, parenthetical references to figures, and widespread use of footnotes and end notes, all of which could cause errors in the tokenization and sentence segmentation. Thus, it was necessary for us to make adjustments to these components for this medical domain. Then, the document collection was parsed into AMR parsed-answers and the triple formalism required by Smatch. This process is depicted in steps 1 - 6 in Figure 3. During real-time interaction, the user question is parsed into an AMR parsed-query, subsequently converted into a triple representation, and then compared against the triples of the preprocessed collection (steps 7 - 8 in Figure 3).

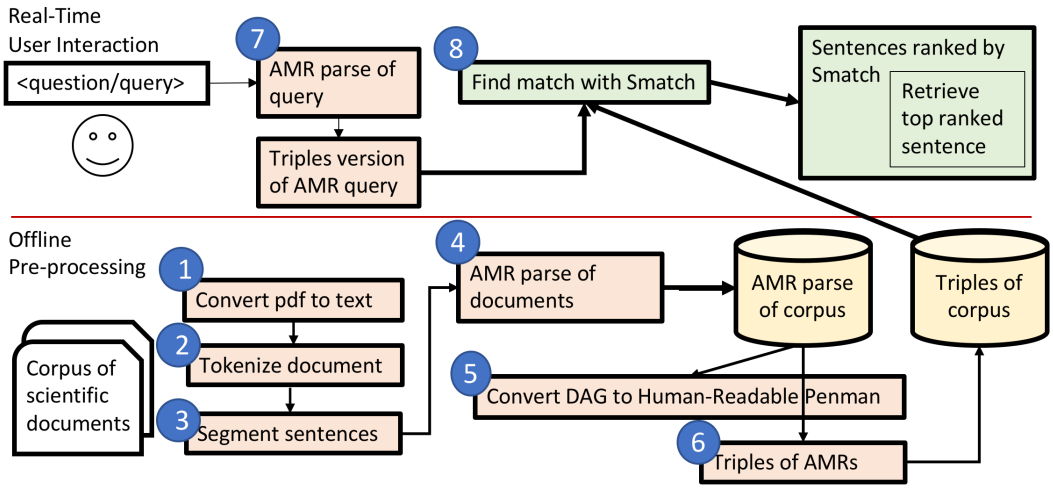


Figure 3: InfoForager Prototype Pipeline

5 Results and Discussion

To make an initial assessment of the validity of our approach, we evaluated the prototype system on the three research questions, for which we had manually identified documents and sentences that best answer the questions. Using these answers and documents as a ground truth, we compared the ranked list of sentences output by our system, thus determining the extent to which our methodology would highly rank and return the same answers as a human researcher.

To focus on the potential for this semantic search approach without introducing too much noise from automatic parsing of the *question* at this stage, we conducted the evaluation using the manually annotated AMR gold-queries, as opposed to the AMR parsed-queries. Note, however, that the target *answer* documents are all automatically parsed. In the subsequent analysis, we compute both 1) gold-query vs. parsed-answer, and 2) gold-query vs. gold-answer, to determine how effective the Smatch metric is in capturing the shared semantic content between query and answer with and without AMR parsing error.

Q$_1$ *Which fluids have higher concentrations of SARS-CoV-2 particles?*
As of June 2020, this research question was judged most urgent by the UV researchers and they found very little supporting literature on the subject—only a single document with several informative sentences in the results section, and one sentence summarizing the information sought:
A$_1$ *Overall, we found higher SARS-CoV-2 titers from saliva than nasopharyngeal swabs from hospital inpatients* (Wyllie et al., 2020).

[9] https://poppler.freedesktop.org/

We were encouraged to find that our system ranked this sentence as an answer third by Smatch F-measure (23%) and fifth by recall (17%). Notably, this sentence has similar phrasing to the question, and as a result, their AMR structures are more similar. The parsed-answer does have a variety of errors, most arising from problems with the `have-degree-91` argument assignment (this is the roleset used to express comparatives in AMR, including "higher"). As noted above, to test for the noise introduced by automatic parsing errors, we also scored the gold-query vs. gold-answer, shown in Figure 4. This gold pair actually receives a slightly lower Smatch F-measure, at 20% (compared to 23% on gold-query vs. parsed-answer). These results made clear that valid query-answer AMR pairs may vary widely in Smatch scores, ranging from 20% to just under 80%, as found for the simple pair in Figure 1.

```
Q1. Which fluids had higher concentrations of virus particles ?
    (h2 / have-degree-91
      :ARG1 (a / amr-unknown
            :domain (f / fluid))
      :ARG2 (h / high-02
            :ARG1 f
            :ARG2 (v2 / virus
                  :consist-of (p / particle
                        :mod (v / virus)
                        :ARG1-of (c / concentrate-02))))
      :ARG3 (m2 / more))

A1. Overall , we found higher SARS-CoV-2 titers from saliva than nasopharyngeal swabs
   from hospital inpatients.
  (f / find-01
    :ARG0 (w / we)
    :ARG1 (h / have-degree-91
            :ARG1 (s3 / saliva
                    :source (i3 / inpatient
                              :location (h6 / hospital)))
            :ARG2 (h5 / high-02
                    :ARG1 s3
                    :ARG2 (v / virus :name (n2 / name :op1 "SARS-CoV-2")
                            :quant (v2 / concentration-quantity
                                    :unit (t2 / titer))))
            :ARG3 (m / more)
            :ARG4 (s5 / swab-01
                    :ARG2 (n3 / nasopharynx
                            :part-of i3)))
    :mod (o / overall))
```

Figure 4: Gold-query Q_1 compared with gold-answer A_1

Q_2 *What is the range of sizes of respiratory droplets, specifically from coughing and sneezing?*
This question had two target documents with several answers that provided information on different aspects of this question. As discussed earlier, the expected InfoForager output need not be a single answer, but rather might include various types of relevant information, which the experts would examine further for source and approach used to obtain that data point. Answers to this question depend on, for example, the methods used to detect the sizes and numbers of droplets, and the location of measurement relative to the droplet source. Because definitions of "droplet" or "aerosol" can depend upon the user, it was expected InfoForager may not find all relevant answers. Answers here include:
A_{2a} *And the geometric mean of droplet size of all the sneezes is 360.1 μm for unimodal distribution and 74.4 μm for bimodal distribution...* (Han et al., 2013).
A_{2b} *...while sneezing produces a greater number of particles than coughing, particles from both activities are of a similar size (a sneeze produces 40,000 - 4,600 particles with 80% of these particles being smaller than 100 μm compared with coughing which produced up to few hundred particles sized between 20 and $>$ 100 μm)* (Gralton et al., 2011).
Although both answers mention "size," they are more complex and distinctly phrased than the question. As a result, our system rankings for this question were weak. For the first target document, two answer sentences ranked in 55th and 128th positions (by Smatch F-measure) out of 1,056 possible matches. For the second target document, one answer sentence ranked 156th of 850 possible matches, while the other received a 0 Smatch score, placing it in the lowest rank where all sentences received a 0 Smatch score.

Given that these low Smatch scores could reflect poor automatic parsing preventing a match, or that the Smatch algorithm does not capture the overlap seen in a Q-A pair, we again examine the gold pair (gold-query vs. gold-answer) for the two answers above (to compare against the gold-query vs parsed-answer). When scored against the gold-query, the gold-answer A_{2a} receives an Smatch f-score of 16%, whereas the parsed-answer A_1 scored significantly higher at 25.8%. For the far longer sentence A_{2b}, its gold-answer receives an Smatch score of 10%, due in large part to its length and complex structure relative to the question. Curiously, the parsed-answer A_{2b} fared far worse, receiving a score of 0. Thus, these A_{2a} and A_{2b} scores for the gold pairs are, in a rough sense, about as challenging to interpret as the A_{2a} and A_{2b} scores comparing the gold-query to their parsed-answers. Significantly more data will be needed to ascertain all the factors in the scoring, but we can see that noise alone from the automatic parse is not the primary cause of low Smatch scores.

Q_t *What types of UV light are used in coronavirus research?*
For this user-study training question, one target answer document and the following answer sentence were identified:
A_t *In conclusion, air disinfection using 254 nm UV-C may be an effective tool for inactivating viral aerosols* (Walker and Ko, 2007).
As with Q_2 above, the phrasing of Q_t leaves open a variety of different answer types, with very different phrasing from the question itself. Nonetheless, our system performed better with this Q-A pair, ranking this parsed answer at the 29th position (Smatch F-measure 24%) out of 406 possible matches, compared to the corresponding gold pair scoring an Smatch of 18%, indicating again, that noise alone from the automatic parse is not bringing down the Smatch score.

To recap, we have seen that this approach using AMR and Smatch can be effective when an answer sentence shares structured semantics with the question. However, when the question is more open-ended, searching for "ranges" and "types," the Smatch approach is much less likely to perform adequately. While these words may on occasion be present in document collections with answers that the UV researchers seek, we realize that InfoForager itself must be augmented to recognize such higher-order questions and, more like a dialogue system with an intelligent agent, potentially engage the user in reformulating their question.

6 Related Work

Previous research has leveraged AMR within the biomedical domain for applications that, like ours, entail natural language understanding, such as information extraction. Garg et al. (2016) use AMR in a graph kernel learning framework to extract biomolecular interactions and report that the use of AMR significantly improves accuracy on this task over surface and syntax-based features, with the best performance achieved by combining AMR and syntax-based features. Notably, this research also furnished the Bio-AMR dataset of 400 manually annotated and 3k automatically annotated AMRs from PubMed scientific articles, which were incorporated into retraining the AMR parser used in InfoForager. Rao et al. (2017) use AMR to identify molecular events/interactions in biomedical text. The authors first define biomedical events of interest as subgraphs within AMR graphs, and develop a neural network-based model that identifies such an event subgraph given an AMR. While the method shows promising results, the authors find that improvements in AMR parsing are needed for further improvement on the task. Wang et al. (2017) were the first to make use of AMR embeddings, in this case along with word and dependency embeddings, to mine for otherwise hidden reports of drug-drug interactions (DDI) in the textual biomedical literature. The best performance the authors report was obtained by combining these three types of embeddings. They also noted that AMR embeddings alone, leveraging the JAMR parser (Flanigan et al., 2014), did not perform adequately for this DDI task, and, like others, attributed this to poor parser performance due to limited medical terms and documents in its training dataset.

Other research that leverages AMR in tasks similar to ours has focused on question-answering. Mitra and Baral (2016) use AMR as an intermediate representation for question-answering tasks designed to test an agent's understanding. The authors find that the addition of a formal reasoning layer significantly

increases the reasoning capability of an agent, and that AMR serves as an effective pivot from natural language to the Answer Set Programming language used for reasoning and inductive logic. AMR is leveraged for a machine reading comprehension (MRC) task for question-answering in Sachan and Xing (2016). Here, the authors first convert a Q-A pair into a single AMR "hypothesis" graph. Additionally, the authors create an AMR for an entire passage, as opposed to the standard single-sentence AMR, by combining at coreference points. With these two AMR graphs, the authors transform the MRC task to a graph containment problem. The authors use a max-margin approach for subgraph matching, reasoning that the hypothesis graph aligns to the passage graph where the question is best answered, not unlike the intuition in InfoForager of applying Smatch to find the best alignment of triples from the query AMR DAG to triples from sentence AMR DAGs in a searched document. The authors obtain competitive results and point out that this approach uniquely allows them to combine evidence from multiple sentences. Although they do not use the AMR formalism per se, Du and Cardie (2020) achieve state-of-the-art performance on the Automatic Content Extraction (ACE) 2005 benchmark event extraction task by making use of template-based questions within a BERT-based question-answering model, suggesting to us that an extension to AMR-based question templates might further improve their approach. The authors experimented with several template strategies for forming the questions targeting the event and its participants. Most relevant to our work, the authors found that the success of their approach was strongly influenced by different questioning strategies, ranging from posing a single keyword as the question to a multi-turn progression of structured questions, and conclude that more natural language questions lead to significantly better performance.

7 Conclusions & Future Work

Our six-week sprint brought together researchers of different backgrounds (UV inactivation and NLP) to identify research requirements, and design and test a prototype framework with software available from other projects. We also identified facts about COVID-19 now more widely known: testing saliva is more effective than nasopharyngeal swabs, given the concentration of the virus is higher in saliva.

We maintain our working hypothesis that the InfoForager framework is promising for facilitating a search system that allows for natural language questions, and in finding answers matching not only on keywords, but the semantic relations between those keywords. However, this process demonstrated that using Smatch for graph matching does not allow us to take full advantage of the semantic structure that AMR offers in this task—namely pinpointing the concept sought in an answer and its direct semantic relations to other concepts. One path we are exploring in the short-term is to have users input their hypotheses in lieu of queries. For example, for Q_1, a user might input a hypothesis like *Saliva has a higher concentration of coronavirus than nasopharyngeal fluid.* Prompting an expert user for a hypothesis could enable us to leverage Smatch more fruitfully to find other graphs more similar overall to the candidate answer, following insights from Sachan and Xing (2016). Although this approach may prove useful, it runs counter to our users' stated preference for posing natural language questions, and would bias the presentation of system results towards the hypothesis graph. This approach would also be inherently limiting for non-expert users with insufficient background to formulate initial hypotheses. The longer-term solution we are exploring follows the related work mentioned above more closely, searching for subgraphs within larger passage graphs that match the query graph (as opposed to scoring the triples of a full query graph against those of a full sentence graph), with supervision from ontological and lexical resources to determine the categories of words in general that could fill certain semantic slots, thereby augmenting what constitutes subgraph matches. For example, for Q_1, we would search for matching subgraph structures where the concepts in the graph could be filled with any phrase for bodily fluids in the relationship of comparison of virus concentrations. Thus, the query would be expanded to a variety of paraphrase alternates, similar to the use of HyTER networks in past AMR research (Dreyer and Marcu, 2012), supporting the full range of matching answer subgraphs. These avenues for future work, in combination with our own findings, encourage us to continue this exploration of semantic search for information foraging, and engage with users to rapidly adapt capabilities for such urgent and dynamically changing situations as the coronavirus pandemic.

References

Rafael T. Anchieta, Marco A. S. Cabezudo, and Thiago A. S. Pardo. 2019. SEMA: an extended semantic evaluation metric for AMR. *arXiv preprint arXiv:1905.12069*.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. LAW*, pages 178–186.

Claire N Bonial, Lucia Donatelli, Jessica Ervin, and Clare R Voss. 2019. Abstract Meaning Representation for human-robot dialogue. *Proceedings of the Society for Computation in Linguistics*, 2(1):236–246.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare R. Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *LREC*.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August. Association for Computational Linguistics.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. *Empirical Methods in Natural Language Processing (EMNLP)*.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland.

Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2718–2726. AAAI Press.

Jan Gralton, Euan Tovey, Mary-Louise McLaws, and William D Rawlinson. 2011. The role of particle size in aerosolised pathogen transmission: a review. *Journal of Infection*, 62(1):1–13.

ZY Han, WG Weng, and QY Huang. 2013. Characterizations of particle size distribution of the droplets exhaled by sneeze. *Journal of the Royal Society Interface*, 10(88):20130560.

M. Lindemann, J. Groschwitz, and A. Koller. 2019. Compositional semantic parsing across graphbanks. In *Proc. of ACL*, pages 4576–4585, July.

Arindam Mitra and Chitta Baral. 2016. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In *Proc. of AAAI*, pages 2779–2785.

Juri Opitz, Anette Frank, and Letitia Parcalabescu. 2020. AMR similarity metrics from principles. *Trans. Assoc. Comput. Linguistics*, 8:522–538.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *Proc. of ACL*, pages 126–135, Vancouver, Canada.

Mrinmaya Sachan and Eric Xing. 2016. Machine comprehension using rich semantic representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 486–492.

Jose-Luis Sagripanti and C David Lytle. 2011. Sensitivity to ultraviolet radiation of lassa, vaccinia, and ebola viruses dried on surfaces. *Archives of virology*, 156(3):489–494.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, Florence, Italy, July. Association for Computational Linguistics.

Christopher M Walker and GwangPyo Ko. 2007. Effect of ultraviolet germicidal irradiation on viral aerosols. *Environmental science & technology*, 41(15):5460–5465.

Yanshan Wang, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Fei Liu, and Hongfang Liu. 2017. Dependency and AMR embeddings for drug-drug interaction extraction from biomedical literature. In *Proc. of ACM-BCB*, pages 36–43, New York, NY, USA.

Anne L Wyllie, John Fournier, Arnau Casanovas-Massana, Melissa Campbell, Maria Tokuyama, Pavithra Vijayakumar, Joshua L Warren, Bertie Geng, M Catherine Muenker, Adam J Moore, et al. 2020. Saliva or nasopharyngeal swab specimens for detection of sars-cov-2. *New England Journal of Medicine*, 383(13):1283–1286.