sciendo Data and Information Management, 2020; 4(3): 177–190

Research Article Open Access

Daqing He*, Zhendong Wang, Khushboo Thaker, Ning Zou

Translation and Expansion: Enabling Laypeople Access to the COVID-19 Academic Collection

https://doi.org/10.2478/dim-2020-0011
received May 11, 2020; accepted May 31, 2020.

Abstract: Academic collections, such as COVID-19 Open Research Dataset (CORD-19), contain a large number of scholarly articles regarding COVID-19 and other related viruses. These articles represent the latest development in combating COVID-19 pandemic in various disciplines. However, it is difficult for laypeople to access these articles due to the term mismatch problem caused by their limited medical knowledge. In this article, we present an effort of helping laypeople to access the CORD-19 collection by translating and expanding laypeople's keywords to their corresponding medical terminology using the National Library of Medicine's Consumer Health Vocabulary. We then developed a retrieval system called Search engine for Laypeople to access the COVID-19 literature (SLAC) using open-source software. Utilizing Centers for Disease Control and Prevention's FAQ questions as the basis for developing common questions that laypeople could be interested in, we performed a set of experiments for testing the SLAC system and the translation and expansion (T&E) process. Our experiment results demonstrate that the T&E process indeed helped to overcome the term mismatch problem and mapped laypeople terms to the medical terms in the academic articles. But we also found that not all laypeople's search topics are meaningful to search on the CORD-19 collection. This indicates the scope and the limitation of enabling laypeople to search on academic article collection for obtaining high-quality information.

Keywords: COVID-19, laypeople, information retrieval, consumer health vocabulary, translation and expansion process

***Corresponding author: Daqing He,** School of Computing and Information, University of Pittsburgh, Pittsburgh, PA. USA.
Email: dah44@pitt.edu
Zhendong Wang, Khushboo Thaker, Ning Zou, School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA

1 Introduction

The novel coronavirus (COVID-19) pandemic, which started at the end of 2019, has been causing great suffering in human society. According to the World Health Organization, the total confirmed cases of COVID-19 globally on April 17, 2020, was 2,078,605, including 139,515 deaths[1], so many peoples' lives have been significantly affected by the pandemic (World Health Organization, 2020). Along with the rapid spread of the virus among different human societies, a huge amount of information has been generated, among which many are misinformation (such as claims on various effective drugs and remedies) or disinformation (such as conspiracy theories on the origin of the virus as biological weapons) that causes more harm and panic (Brennen, Simon, Howard, & Nielsen, 2020). Therefore, the global health crisis also turned itself into a global information crisis (Xie et al., 2020).

As the pandemic affects not only peoples' health and daily life but also their jobs and communities, people naturally have many questions regarding the virus and the pandemic, and desire answers with deep insights to help them cope with the situations. However, since COVID-19 is a new virus that human beings have not encountered before (Guo et al., 2020), not many well-organized and authoritative resources are available to provide high-quality answers to peoples' questions.

Fortunately, there is rich academic literature about coronavirus and its related family accumulated over the years. Recently, researchers have been rapidly publishing academic papers regarding COVID-19 to disseminate their knowledge and enable global collaboration in combating the virus (Kousha & Thelwall, 2020). For example, the authors conducted a quick search of "COVID-19" with the publication year to be "2020" on Google Scholar on April 17, 2020, and the search yielded 14,700 returned results, indicating a rich published literature being available for exploration. Recently, Allen Institute for Artificial

1 https://covid19.who.int

Intelligence (AI) partnered with other groups prepared and distributed the COVID-19 Open Research Dataset (CORD-19), which contains "over 51,000 scholarly articles, including over 40,000 with full text, about COVID-19 and the coronavirus family of viruses" (Wang et al., 2020)[2].

We believe that it is of great benefit for laypeople to access these articles because many of their questions can be answered through accessing these academic publications. There are several reasons for our belief. First, these articles represent the latest achievements in knowledge about COVID-19, and at the same time, these academic papers, which have to go through a rigorous peer-review process, contain high-quality scientific information. Therefore, the information obtained from these articles can be useful for laypeople to combat the misinformation and disinformation problems plaguing the online social media platforms. Second, these articles could be the basis for many government policies and expert suggestions. Laypeople's access to the information extracted from these papers can provide supporting authoritative evidence for them to better understand the policies and recommendations.

However, prior research (Bhavnani, 2002) suggested that online health information seeking can be problematic for laypeople because health-related information often requires certain domain-specific knowledge that could be beyond laypeople's understanding of the medical topics (Hanbury, 2012). Consequently, these articles and their content might be too difficult for laypeople. This difficulty might be particularly severe in the context of laypeople seeking COVID-19 related health information since it is an infectious disease caused by a new coronavirus introduced to humans for the first time (Guo et al., 2020). Laypeople either do not know the existence of these articles or do not know how to search for them.

The goal of our work is to explore the mechanisms for enabling laypeople to access the scientific articles regarding COVID-19. Specifically, we aim to design and develop an automatically terminology translation and expansion (T&E) process between laypeople's vocabulary and professional medical terminology. We model this problem as a classic information retrieval issue called "term mismatch problem" (Zhao & Callan, 2012). Unlike medical professionals, laypeople often do not have adequate knowledge about COVID-19. They often could not describe their COVID-19-related information needs using the same professional medical terms that commonly appear in academic medical articles. Our project aims to resolve this term mismatch problem, and specifically, we explore the following research questions:

- RQ1: How online resources can be integrated to provide the T&E capabilities between laypeople queries and professional medical content in academic articles?
- RQ2: How well such T&E capabilities can help search within the COVID-19 collection?

The contributions of our study include the rapidly developed search engine called the SLAC system (a Search engine for Laypeople to access the COVID-19 literature) for enabling laypeople to gain access to the COVID-19 academic collection, and the focus of automatically translating and expanding laypeople's queries to search on medical content in academic articles.

The remaining sections of this article are organized as follows: First, Section 2 introduces the related work and Section 3 presents the SLAC system. To test the SLAC system, we designed a set of experiments, and Section 4 presents the design details, and Section 5 provides the result analysis. Finally, Section 6 concludes the future work.

2 Related Work

Due to different generation processing of query terms and document terms, there is often the problem of term mismatch between those used in the query by the users and those used in the documents by the authors (Zhao & Callan, 2012). This is particularly problematic when there is a big difference between the domain knowledge of the users and that of the authors, such as laypeople searching for academic articles. This term mismatch problem is very common in the medical domain (Zhang, & He, 2017, 2018).

Many methods have been developed to combat this problem. Initially, manual indexing of documents with control vocabulary terms was adopted, and reference librarians were the people who manually converted laypeople's information needs with corresponding control vocabulary terms (Marchionini, 1997). But in the era of automatic indexing and full-text search engines, the common methods were changed to utilize query expansion through either user-provided relevant documents, relevant documents identified through interactive relevance feedback or extracted terms returned automatically through pseudo relevance feedback (Salton & Buckley, 1990; Harman, 1992; Marchionini, 1997). More recently, word embedding has been applied to expand the query terms with potentially relevant synonyms too (Diaz, Mitra, & Craswell, 2016; Zhang & He, 2018).

2 https://pages.semanticscholar.org/coronavirus-research

COVID-19 pandemic is a global public health crisis (Heymann & Shindo, 2020; Lipsitch, Swerdlow, & Finelli, 2020). It has triggered great interest in the context of telemedicine (Hollander & Carr, 2020), mental health (Liu et al., 2020), and other healthcare topics.

COVID-19 pandemic is also a global information crisis (Xie, 2020), and many information science researchers are working on novel technologies for combating the information crisis. For example, Wang, Ng and Brook (2020) worked on the big data and information technology side of COVID-19. Misinformation on social media platforms have been a great problem, and it is much worse in COVID-19 pandemic (Cinelli et al., 2020). Therefore, many studies worked on identifying the balance between combating misinformation and providing constant information access (Pandey, Gautam, Bhagat, & Sethi, 2020).

Artificial Intelligence (AI) has been widely applied in many areas; therefore, there are studies for applying AI in combating COVID-19 (Bullock, Pham, Lam, & Luengo-Oroz, 2020), and for building neural information retrieval engines to return high-quality search results (Zhang, Gupta, Nogueira, Cho, & Lin, 2020). Another important work is to build up data collections so that modern information technology that needs a large amount of training data can be developed. Some example datasets include a social media dataset built on Twitter data (Chen, Lerman, & Ferrara, 2020), and the CORD-19 provided by the Allen Institute for AI in collaboration with many other institutions (Wang et al., 2020).

3 The SLAC System

The SLAC system is a search engine for laypeople to access the CORD-19. Its goal is to better support laypeople to obtain high-quality authoritative information from COVID-19-related academic literature. As stated, the main problem for laypeople to access academic literature, particularly medical academic literature, is that they often do not have the adequate medical knowledge to master the right methods for accessing academic papers, including searching with the right query terms, and recognizing the relevance of the returned articles to their information needs. Consequently, SLAC has two important features. The first one is the capability of automatically translating query terms using medical domain knowledge so that laypeople's terms can be mapped to Unified Medical Language System (UMLS) medical concepts and their corresponding terms. The second feature is the integrated search engine that can automatically take users' queries, translate to appropriate expanded queries, and fetch back potentially relevant documents for users to read.

In the remainder of this section, we present these two features in detail.

3.1 T&E with Medical Knowledge

Academic medical documents are written for medical experts and consist of specialized medical terminologies (medical concepts) often unknown to laypeople (Kindig, Panzer, & Nielsen-Bohlman, 2004). This makes it difficult for laypeople to retrieve relevant academic documents using the corresponding medical concepts (Gu et al., 2019). For example, "animal rat" may be mentioned as "chiroptera" in medical documents so that the query containing "rat" would not be able to retrieve documents with term "chiroptera."

Our approach to overcome this "term-mismatch" problem is the T&E process. The translation part helps to transform laypeople's words to medical specific terms, and the expansion part connects to other laypeople's terms that can be used to express the same medical concept. All these computations are performed using Consumer Health Vocabulary (CHV), a part of UMLS' Metathesaurus (Schuyler, Hole, Tuttle, & Sherertz, 1993).

3.1.1 Collaborative CHV

UMLS is a meta-thesaurus for providing a standard tool to access medical concepts (Bodenreider, 2004). UMLS consists of an amalgamation of health and medical ontologies and vocabularies, and collaborative CHV is a part of UMLS.

CHV is an open-source and collaborative language resource of biomedical terminologies used by laypeople (or health consumers) (Zeng et al., 2007). It consists of laypeople terms and medical jargon extracted from query logs of an online health information resource called MedlinePlus[3]. MedlinePlus is a service provided by the National Library of Medicine (NLM) for patients and their families to access medical knowledge. CHV consists of 158,518 laypeople terms that map to 57,819 unique UMLS terms[4]. For example, the medical term "Myocardial Infarction" contains 26 laypeople variations that include "Heart Attack", "Heart Infarction", "Coronary Attack", and so on.

3 https://medlineplus.gov/
4 http://www.consumerhealthvocab.org/

Table 1
Examples of Laypeople Terms and Corresponding Medical Concepts Identified Via the T&E Process

Original Query Term	Laypeople Term	Medical Concepts
SARS	SARS virus	SARS coronavirus, SARS-Cov, Severe Acute Respiratory Syndrome
MERS	MERS Virus	middle east respiratory syndrome, MERS
hcq	Hcq	Hydroxychloroquine, chloroquine
Rat	Rat, brown rat	rattus norvegicus
Bat	bat	chiroptera
Pneumonia	Pneumonia, lung disease	pulmonary inflammation, pneumonitis, Lung inflammation
malaria	malaria	paludism

Because CHV provides the mapping between laypeople's terms and UMLS medical concepts, we view it as the bridge for translating laypeople's query terms to specialized UMLS medical terms. However, as Keselman et al. (2008) pointed out, CHV is not an extensive resource and does not cover all the UMLS concepts. We accept this limitation and leave to the future work to utilize more comprehensive and recently generated laypeople's resources (Pylieva, Chernodub, Grabar, & Hamon, 2018; Gu et al., 2019).

3.1.2 T&E of Laypeople's Query Terms

As stated, each term from the laypeople's query goes through the T&E process using CHV as the bridge. The T&E process is performed using MetaMap Application Programming Interface (API) [5]. To increase the coverage of the original query terms, after the initial CHV terms are returned through the API, we further explore other UMLS terms and laypeople terms in CHV that can be the variations of these initial returned CHV terms.

One issue that we have to deal with when working on CHV terms is that some laypeople terms from CHV contain spelling errors. We recognize and remove these terms with errors. The T&E process happens in the background automatically so that the user does not have to be concerned about this step. Table 1 shows examples of laypeople's terms and corresponding medical concepts returned through the T&E process.

3.2 Indexing and Searching the CORD-19 Collection

3.2.1 Indexing the CORD-19 Collection

CORD-19 is provided by the Allen Institute for AI in collaboration with The White House Office of Science and Technology Policy, the NLM, the Chan Zuckerburg Initiative (CZI), Microsoft Research, and Kaggle (Wang et al., 2020). It contains 51,045 academy documents from PubMed Central, World Health Organization, and preprint corpora like bioRxiv and medRxiv. The included articles are publications and preprints on COVID19 and related historical coronaviruses such as SARS and MERS (Wang et al., 2020). The dataset received great attention from the academic community. It has been "viewed more than 1.5 million times and downloaded over 75K times in the first month of its release" (Wang et al., 2020).

Each document in CORD-19 has rich metadata and 80% of the full text was available. When constructing the index of CORD-19, we selected the two metadata fields: abstract and title, as well as the full text when available. The tokenization was performed using "Unicode Text Segmentation algorithm" tokenizer, and all tokens were then normalized to lowercase when possible. This tokenizer performed light preprocessing on the text. Besides case folding, it only removes standard stop words but does not perform stemming.

3.2.2 Search Engine for CORD-19

To speed up the development of the SLAC system, we built our search engine based on several open-source projects.

5 https://metamap.nlm.nih.gov/

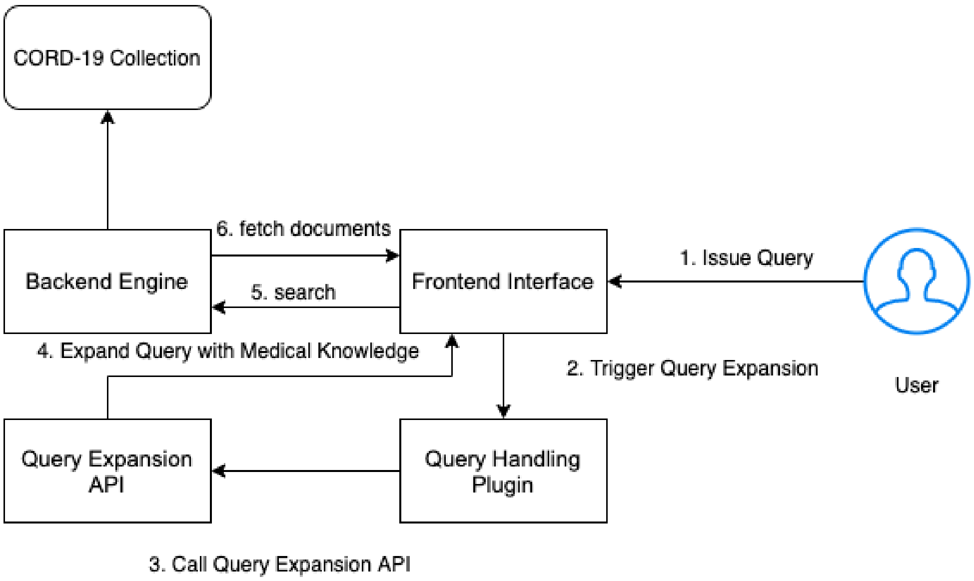

Figure 1. The Architecture and Data Flow of the SLAC System

Figure 1 shows the architecture and data flow of the SLAC system, which has four key components:

- **Backend engine**: An ElasticSearch-based backend engine is adopted to search on the CORD-19 Collection. We choose ElasticSearch since its ready to use and can be flexible in building and customizing a search system (Gormley & Tong, 2015). Two ranking models are selected in the backend engine: BM25 and Language Model with Dirichlet smoothing. There is a switch in the frontend interface to select the ranking model.
- **Frontend interface:** A Kibana-based frontend interface is adopted. Kibana is the dashboard for ElasticSearch that provides a ready-to-use search interface (Gupta, 2015). It provides a Boolean search capability at the search box, which is useful in defining the exact meaning of the query entered.
- **Query expansion API:** A python-based RESTful API that is connected to MetaMap API, which accepts the original query and returns the expanded version with medical knowledge. This component helps to complete the T&E process.
- **Query handling plugin:** A chrome plugin that can read the search query from the Kibana search interface, then call the Query Expansion API and use the expanded query to search.

Figure 2 shows the screenshot of the SLAC system's frontend Interface. The Boolean query shown asks for documents relevant to the COVID-19's origin and the first transmission. Four document surrogates with query terms highlighted are displayed.

Besides shortening the system development time, building the SLAC system with existing open-source software also enables us to take advantage of the software engineering side of these software packages. For example, the response time of the SLAC system to a given query with about 10 query terms can be maintained at around 30 ms. It takes 1.13 s on average for Query Expansion API to expand one term. To speed up this part, we implemented a cache for each expanded term in Query Expansion API to make sure each unique term will only be expanded for one time. It helps to reduce the average response time to a query with 54 query terms to be around 225 ms.

4 Experiment Design

4.1 Laypeople's Information Needs

As stated, many studies on the virus and the treatment of the virus have been published during the COVID-19 pandemic. At the same time, there has been a large amount of social media messages generated that could represent the information needs of laypeople on the virus and the pandemic. However, there has not been a study on laypeople's information needs on COVID-19 yet, which motivated us to look for an alternative source for laypeople's information needs.

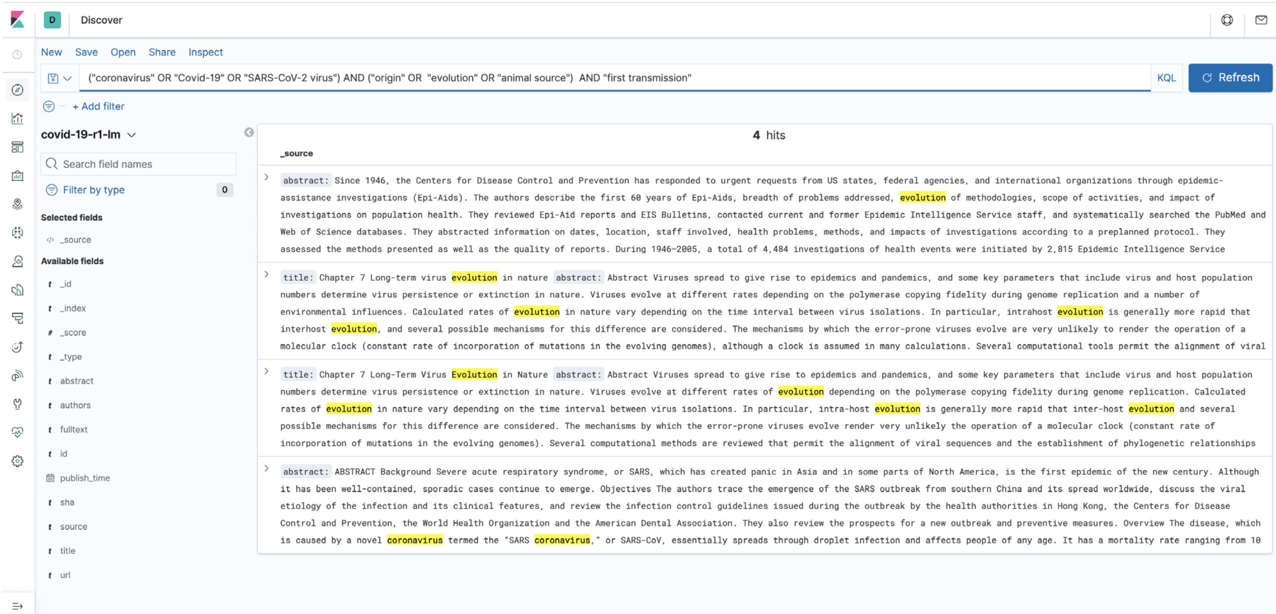

Figure 2. The Screenshot of the SLAC's Frontend Interface

Chi, He and Jeng (2020) found that laypeople reckon health-specific webpages that contributed by professionals as their primary visited/adopted online health information sources. This motivated us to explore Centers for Disease Control and Prevention (CDC) and many other government/society-based institutions as the primary information sources for laypeople. Consequently, we adopted COVID-19 FAQ pages published by CDC[6] to represent laypeople's common needs on COVID-19. In total, we randomly selected three FAQ questions from each of 11 topics posted at CDC's COVID-19 FAQs site to build up the laypeople's information needs (see Table 2 for 33 search questions). The questions are directly copied from CDC's FAQs without any editing. In CDC's COVID-19 FAQs page, each question is provided with the answer. Because we will explore the question of whether or not the answers are useful, we included CDC's answers to the questions when we manually developed the search questions. Table 3 shows the details of search question #8, which contains the CDC topic, the question from the CDC FAQ, and the CDC's answer to the question.

6 https://www.cdc.gov/coronavirus/2019-ncov/faq.html

4.2 Experiment Runs

The goal of our experiment is to explore the impacts of the T&E process for converting laypeople query terms to relevant medical terms so that relevant documents written in medical terminology in the CORD-19 collection can be matched and returned to the laypeople queries. Therefore, we established three baselines for exploring upper and low bound respectively. The details of these runs are:

– *Plain query PQ run (i.e., Low Baseline):* the terms from the original search questions were used without any expansion. Only a set of stopwords were removed, and the queries were written in Boolean format with AND and OR operators linking all the query terms.
– *Expanded query EQ run (i.e., Experiment Run)*: the original queries from the search questions were automatically expanded with our T&E process using CHV.
– *Oracle Plain OP run (i.e., High Baseline 1)*: the query terms in this run were not only based on the original search question but also contained the keywords extracted from the answers provided by CDC. The idea is that the keywords in the answers could be useful in finding relevant documents in CORD-19 collection. Because these keywords from the answers usually do not appear in the search questions, we call this run a *Oracle run*, which could be one of the high baselines.
– *Expanded oracle OE run (i.e., High Baseline 2))*: the query terms in this run were based on the T&E process

Table 2
33 Search Questions Converted from CDC COVID-19 FAQs

CDC Topic	ID	Question
COVID-19 Basics	0	What is a novel coronavirus?
	2	Why might someone blame or avoid individuals and groups create stigma because of COVID-19?
	4	Why do some state's COVID-19 case number sometimes differ from what is posted on CDCs website
How COVID-19 Spreads	8	How does the virus spread?
	10	Can someone who has had COVID-19 spread the illness to others?
	15	What is community spread?
How to Protect Yourself	18	Am I at risk for COVID-19 in the United States
	20	What should I do if I have had close contact with someone who has COVID-19?
	23	Is it okay for me to donate blood?
COVID-19 and Children	24	What is the risk of my child becoming sick with COVID-19?
	27	Should children wear masks?
	29	What steps should parents take to protect children during a community outbreak?
School Dismissals and Children	30	While school's out can my child hang out with their friends?
	32	While school's out will kids have access to meals?
	34	While school's out limit time with older adults including relatives and people with chronic medical conditions?
Preparing Your Home and	35	How can my family and I prepare for COVID-19?
Family for COVID-19	37	What should I do if someone in my house gets sick with COVID-19?
	41	Should I make my own hand sanitizer if I can't find it in the stores?
In case of an Outbreak	42	What should I do if there is an outbreak in my community?
in Your Community	43	Will schools be dismissed if there is an outbreak in my community?
	44	Should I go to work if there is an outbreak in my community?
Symptoms and Testing	46	What are the symptoms and complications that COVID-19 can cause?
	47	Should I be tested for COVID-19?
	48	Where can I get tested for COVID-19?
Higher risk	50	Who is at higher risk for serious illness from COVID-19?
	52	How were the underlying conditions for people considered higher risk of serious illness with COVID-19 selected?
	56	Are people with disabilities at higher risk?
COVID-19 and Funerals	57	Am I at risk if I go to a funeral or visitation service for someone who died of COVID-19?
	60	What should I do if my family member died from COVID-19 while overseas?
	61	My family member died from COVID-19 while overseas? What are the requirements for returning the body to the United States?
COVID-19 and Animals	62	Can I get COVID-19 from my pets or other animals?
	65	Should I avoid contact with pets or other animals if I am sick with COVID-19?
	69	What precautions should be taken for animals that have recently been imported from outside the United States for example by shelters rescues or as personal pets?

Table 3
Search Question 8, Which is Converted from a Question at CDC COVID-19 FAQs.

[Search ID]:	**8**
[CDC Topic]:	How COVID-19 Spreads?
[Question]:	How does the virus spread?
[Answer]:	The virus that causes COVID-19 is thought to spread mainly from person to person, mainly through respiratory droplets produced when an infected person coughs or sneezes. These droplets can land in the mouths or noses of people who are nearby or possibly be inhaled into the lungs. Spread is more likely when people are in close contact with one another (within about 6 feet). COVID-19 seems to be spreading easily and sustainably in the community ("community spread") in many affected geographic areas. Community spread means people have been infected with the virus in an area, including some who are not sure how or where they became infected.

applied to the query terms in the *OP* run. The idea is that the terms extracted from CDC answers could still be laypeople terms so it makes sense to go through our T&E process to remove the term mismatch issues. It helps to establish an even higher upper bound.

4.3 Ground Truth and Evaluation Measures

Our experiment follows the Cranfield tradition of evaluating information retrieval systems (Voorhees & Harman, 2005), so we treat the CORD-19 dataset as the document collection, the 33 questions we adopted from CDC COVID-19 FAQs as the search questions, and the only thing missing was the ground truth annotation.

To obtain the ground truth, we adopted TREC's pooling methodology (Spark-Jones, 1975). For each search question, our four runs generated one ranked list respectively. We fetched the top 10 relevant documents from each run and constructed a pool of 791 unique documents after removing duplicates. Then two of the authors acted as human annotators to evaluate the documents with a relevance score between 1 and 5 (1 means "not relevant" and 5 means "highly relevant"). Both annotators are information scientists who do not have strong medical domain knowledge. However, since all our search questions have CDC's answers to act as reference, both annotators did not find serious problems that prevent them from the annotation. Four random search questions were selected at the beginning and the middle stage of the annotation to calculate the inter-annotator agreement. The Weighted Kappa value was calculated at each stage when the common search question's pool was annotated. The Kappa values obtained were 0.809 and 0.741, indicating a reliable agreement. Because of the high agreement, the rest of the search question pools were evenly divided and annotated independently by one of the two annotators.

Follow the common practice of evaluating web search engines (Agichtein, Brill, & Dumais, 2006), we used Normalized Discounted Cumulative Gain (*nDCG*@10) as the evaluation measure (Järvelin & Kekäläinen, 2002). Since we have 33 search questions to generate queries, we use *t*-test for statistical significant testing.

5 Result Analysis and Discussions

Our experiment results show three interesting insights. We present them in detail in the following subsections.

5.1 The T&E Process is Helpful

5.1.1 Exhibited Benefits of the Current T&E

As shown in Table 4, queries directly extracted from the CDC questions (i.e., the *PQ* run) could achieve *nDCG*@10 at 0.543. With the help of the T&E process, the *EQ* run could improve the search performance measured by nDCG@10 to 0.728. When we examined other evaluation measures such as *precision*@5 and *precision*@10, we observed the same improvement. *Precision*@5 and *precision*@10 for the *EQ* run are 0.436 and 0.373, both of them are higher than that of the *PQ* run (0.382 and 0.348, respectively). The statistical tests confirmed that there are significant differences between the *nDCG*@10 results of *PQ* and that of *EQ* (p¡ 0.05), but not for that of *precision*@5 and *precision*@10.

We further observed that the main impact that the T&E process achieved is to greatly increase the number of queries that can return at least one documents. Due to term mismatch problems, only 19 out of 33 queries in the *PQ* run found matched documents (not necessarily relevant documents) from the CORD-19 collection. This

Table 4
Experiment Results

	PQ run	*EQ* run	*OP* run	*OE* run
Avg Num Query Terms per Query	5.76	23.70	10.00	54.03
Num Queries with > 0 Returned Docs	19	26	12	29
Avg Relevance Score	1.97	2.09	1.46	2.28
Avg nDCG@10	0.543	0.728	0.350	0.806
Avg nDCG@10 on non-0 queries	0.944	0.924	0.964	0.917
Precision@5	0.382	0.436	0.297	0.418
Precision@10	0.348	0.373	0.267	0.373

number increases to 26 in the *EQ* run, which shows that the T&E process is helpful.

5.1.2 Limitations of the Current T&E

However, we do notice two limitations of the current T&E process. First, as shown in the first row of Table 4, the T&E process could greatly expand the number of query terms into the queries, but it does sometimes bring noise into the queries during the process. If we examine the results of *PQ* and *EQ* on the average *nDCG*@10 on those 11 queries that have returned documents, the average *nDCG*@10 value for *PQ* is 0.944, which is higher than that of *EQ* (i.e., 0.924). This demonstrates that future improvement on the T&E process is needed to make the outcomes more accurate.

Secondly, the current T&E process still fails to translate or expand many query terms. We found that only 42.929% of the query terms in *PQ* were translated and/or expanded when the queries in *EQ* were constructed. It was even lower for expanding queries in *OP* to that in *OE*, where only 33.528% query terms in *OP* have corresponding CHV returns. However, once a query term goes through the T&E process, many terms can be added. For example, the average added terms from *PQ* to *EQ*4 is 17.91 words per query, and that from *OP* to *OE* is 44.02 words per query. We further notice that translation can contribute between 20.84% (i.e., 9.18 words in the case of *OP* to *OE*) and 29.70% (i.e., 5.32 words in the case of *PQ* to *EQ*) of the added terms, whereas the rest come from expansion with variations of laypeople terms.

5.1.3 Case Study on Search Question 8

To illustrate further the differences among the four runs and the impacts of the T&E process, we conducted a case study on search question 8, whose request information is presented in Table 3.

Table 5 shows that the T&E process can find many extra terms for T&E so that the query length for *EQ* and *OE* is much longer than that of *PQ* and *OP* respectively. Most of the expanded query terms are relevant to the original query term, so they are in good quality. However, in this particular case, T&E did not generate better retrieval performance. As shown in Table 5, both runs without T&E have better *nDCG*@10 value than the corresponding run with T&E, which is *PQ*:0.991 versus *EQ*:0.922 and *OP*:0.952 versus *OE*:0.917.

We then examined the top 10 returned documents for each run and looked at their relevance scores as well as the overlap among the runs. As shown in Table 6, all four runs managed to return many high relevance score documents (i.e., relevance score 4) in the top 10 positions. However, *PQ* has the highest number of documents with score 5 overall and at top 3, that is why its *nDCG*@10 is the highest.

We further notice that despite many high relevance score documents are returned, there is no overlap between the documents returned by *PQ* and that of *EQ*. Only *OP* and *OE* have four documents shared between them. If we look at the query terms between *PQ* and *EQ* (see Table 5), all query terms in *PQ* are also in *EQ*. So we studied further on this issue.

After analyzing the impacts of various query terms, we recognized that the issue might come from the T&E of query term *covid-19* in *PQ*, which is changed to *(”corona infection virus” OR ”corona infections virus” OR ”coronavirus” OR ”corona virus” OR ”genus: coronavirus” OR ”covid-19” OR ”coronavirus infections”)*. This T&E introduced too many changes so that there are no overlap documents between *PQ* and *EQ* in top 10 ranks at all. To test my hypothesis, we composed a new run for search question 8 called *EQ*’ where the query is changed to be the one marked as *EQ*’ in Table 5. The top 10 returned documents for this new run is presented in the *EQ*’ column in Table 6. This time, 8 out of 10 top-ranked documents are shared between *PQ* and *EQ*’, and one of the two new documents in *EQ*’ is highly relevant (52*zjm9jt* at score 5) too.

Consequently, this case study shows that using T&E, these two runs perform quite differently. However, as stated in Section 5.1.2, noisy terms can be introduced through the T&E process. We will explore better coverage and more accurate T&E methods in the future.

Table 5
Queries for Different Runs for Search Question 8.

Run (nDCG)	Query
PQ (0.991)	”spreads” AND ”virus” AND ”spread” AND ”covid-19”
EQ (0.922)	(”smear- instruction imperative” OR ”spreads” OR ”spread”) AND (”viridae” OR ”viruses” OR ”virus”) AND (”smear-instruction imperative” OR ”spread”) AND (”corona infection virus” OR ”corona infections virus” OR ”coronavirus” OR ”corona virus” OR ”genus: coronavirus” OR ”covid-19” OR ”coronavirus infections” OR ”coronaviruses” OR ”coronavirus infection”)
OP (0.952)	”spread” AND ”covid-19” AND (person to person OR ”close contact” OR ”six feet”) AND (”respiratory droplets” OR ”coughs” OR ”sneezes” OR ”inhaled into lungs”)
OE (0.917)	(”smear-instruction imperative” OR ”spread”) AND (”corona virus” OR ”corona infections virus” OR ”coronavirus” OR ”coronaviruses” OR ”coronavirus infection” OR ”covid-19” OR ”coronavirus infections” OR ”genus: coronavirus” OR ”corona infection virus”) AND (person to person OR (”contact” OR ”close contact” OR ”close” OR ”closed” OR ”contact with” OR ”closing” OR ”contacting”) OR (”six feet” OR ”feet, unit of measurement” OR ”ft” OR ”six” OR ”feet”)) AND ((”respiratory droplets” OR ”respiratory”) OR (”cough, ctcae” OR ”coughs” OR ”cough”) OR (”sneezing” OR ”sneezes” OR ”sneeze”) OR (”inhal” OR ”inhalation” OR ”lung structure” OR ”in breathing” OR ”inspiration” OR ”inspirations” OR ”inhaled” OR ”pulmonary” OR ”inhalations” OR ”lung” OR ”lung structures” OR ”inspired” OR ”inhaling” OR ”inspir” OR ”lungs” OR ”breathing inspiration” OR ”inspiration function” OR ”breathing” OR ”inspiratory” OR ”breathing in” OR ”inhaled into lungs” OR ”respiratory aspiration”))
EQ’ (0.999)	(”smear - instruction imperative” OR ”spreads” OR ”spread”) AND (”viridae” OR ”viruses” OR ”virus”) AND (”smear- instruction imperative” OR ”spread”) AND ”covid-19”

Table 6
Top 10 Returned Documents for the Runs of Search Question 8. Returned Documents are Marked as “DOCID (RelScore).” Documents Whose ID is Bold are the Ones that Shared with Multiple Runs. Documents Marked with “” are Shared between PQ and EQ’.*

rank	*PQ* run	*EQ* run	*OP* run	*OE* run	*EQ’*
1	qz9tgl83 (5)*	b518n9dx (3)	**yy7abob9 (4)**	**yy7abob9 (4)**	qz9tgl83 (5)*
2	8ozauxlk (5)*	djuomhww (5)	**zndtddty (5)**	**m5h19hy6 (5)**	yg5posts (5)*
3	yg5posts (5)*	0lyxvex0 (4)	**m5h19hy6 (5)**	**lasv4e6a (4)**	oee19duz (5)*
4	xfjexm5b (4)*	zpaqd5vd (4)	fu8ndhdo (4)	iv753tly (1)	8ozauxlk (5)*
5	9em5tjya (4)*	pidar1gz (3)	52zjm9jt (2)	k4lzwfge (3)	52zjm9jt (5)
6	oee19duz (5)*	nn15iyqd (4)	bpukqctg (5)	c9ts2g7w (3)	xcacty89 (4)*
7	xcacty89 (4)*	8lku99jc (5)	**lasv4e6a (4)**	sl6gsjz4 (2)	bbmcenpy (3)
8	dxtbp4kd (5)	s155i4e9 (4)	0hrmk77p (5)	39tg92sa (2)	xfjexm5b (4)*
9	ztl54g6q (5)*	ztcyvsoi (4)	ycrrsr5c (3)	**zndtddty (5)**	ztl54g6q (5)*
10	lioj0tkn (3)	smmrl5i6 (1)	k2ixwz9w (4)	hmy8fs3g (5)	9em5tjya (4)*

5.2 Knowing the Answers is Only Helpful to a Certain Extent

Our oracle plain run *OP* was generated based on the keywords extracted from the CDC questions *AND* the keywords extracted manually from the answers provided by CDC on those questions. The motivation of this run is that the keywords in the answers could help identify extra relevant documents that would miss by the queries generated from the original questions.

However, as shown in Table 4, the results are opposite to our original thinking. The average *nDCG@*10 for the *PQ* run is 0.543, which is much higher than that of the *OP* run (i.e., 0.350), even though there is no statistically significant difference between the two results (p = 0.103).

There are several possible explanations behind this result. First, as shown in Table 4, the number of queries that return at least one document has dropped from 19 in *PQ* to 12 in *OP*. This shows that the term mismatch problem became even more problematic with more keywords added from the answers. The queries in *OP* became too restricted.

However, those keywords from answers did provide extra benefits when those queries were helped by the T&E process. The *OE* run, whose queries contain the extra keywords from the T&E process performed on the queries of *OP*, achieved significant improvement over the results of *OP* on all evaluation measures. Its *nDCG*@10 is 0.806, which is significantly higher than that of *OP* (p ¡ 0.01). The number of queries with at least one returned document is also increased to 29 from 12. All of these demonstrated the benefits of having T&E process, which was even more helpful to remove the term mismatch problem in the search.

The benefits of having extra keywords from the answers of CDC FAQs also showed at the differences between *OE* and *EQ*. The average value of *nDCG*@10 was increased from 0.728 to 0.806, and the number of queries with at least one returned document was increased from 26 to 29. But we did not find any statistical significance between the results of *EQ* and *OE*.

All these results demonstrated that knowing the answers for CDC questions could be useful in improving the search performance, but their benefits have certain limitations. Having more keywords from the answers made the laypeople queries to be too restricted to find relevant academic documents in the CORD-19 collection. It is only after the T&E process removes the term mismatch problem, could the queries with answer keywords to be very effective in bring relevant academic documents.

5.3 Scope and Limitation of Translating Laypeople's Terms

Our experiments on the 11 CDC topics further demonstrated that not all laypeople's information needs are meaningful to search on academic collections like CORD-19. As shown in Table 7, the 11 topics can be classified into three categories.

5.3.1 Topics That Are Meaningful to Search in CORD-19

The first category includes the topics that are meaningful to search on CORD-19. It includes the CDC topics such as "COVID-19 spreads", "How to protect yourself", "COVID-19 and Children", "In case of an Outbreak in Your Community", "Symptoms and Testing" and "Higher risk." They all share the characteristics of that high *nDCG*@10 score (> 0.6) even at the *PQ* run.

Looking at these topics, it seems that they would be the topics appearing in academic papers, and therefore, it makes sense to search on an academic paper collection like CORD-19. On these topics, the T&E process can still make some degree of improvement, such as on topics "COVID-19 and Children" (from 0.643 to 0.905), and "In case of an Outbreak in Your Community" (from 0.878 to 0.930). Maybe there are still some words that need translation for laypeople's knowledge. Therefore, their initial knowledge on the possible query terms could not be that useful in the search for relevant academic papers.

Some other topics in this category show strong performance even on the *PQ* run, which probably indicates that these are the topics that are most often discussed among the laypeople, and therefore the gap between academic terms and laypeople terms is not that wide. This, thus, makes no changes on *nDCG*@10 when using T&E on the topic such as "COVID-19 spreads", "How to protect yourself", "Symptoms and Testing" and "Higher risk."

5.3.2 Topics That T&E Process Can Help

The second category includes those topics that are initially not meaningful, but the T&E process can be very helpful. The example CDC topics in this category include "Preparing Your Home and Family for COVID-19", "COVID-19 and Funerals", "COVID19 and Animals." The shared characteristics include that these topics all have bad *nDCG*@10 value in the *PQ* run. These topics might not be discussed very often between laypeople and academic communities; therefore, laypeople would not know the right words to use for the search. However, these are still valid topics to search on an academic collection like CORD-19, therefore, using the T&E process, relevant documents in CORD-19 can be returned, and often to the top of the search results.

5.3.3 Topics Unsuitable to Search in CORD-19

The third category are those topics that are not useful to search in CORD-19 even using the T&E process. A good example of the topic in this category is "School Dismissals and Children." This might be a topic to search for local

Table 7
Average nDCG@10 Values for Different CDC Topics

CDC topics	*PQ* run	*EQ* run	*OP* run	*OE* run
COVID-19 Basics	0.325	0.297	0.301	0.305
COVID-19 Spreads	0.952	0.930	0.965	0.947
How to Protect Yourself	0.933	0.905	0.667	0.923
COVID-19 and Children	0.643	0.905	0.295	0.879
School Dismissals and Children	0	0	0	0.611
Preparing Your Home and Family for COVID-19	0.315	0.924	0.318	0.912
In Case of an Outbreak in Your Community	0.878	0.930	0.314	0.931
Symptoms and Testing	0.981	0.959	0.332	0.589
Higher Risk	0.948	0.937	0.328	0.860
COVID-19 and Funerals	0	0.620	0	0.982
COVID-19 and Animals	0	0.601	0.333	0.927

Table 8
nDCG@10 Scores for Questions in CDC Topic "COVID-19 Basics"

Questions	*PQ* run	*EQ* run	*OP* run	*OE* run
0. What is a novel coronavirus	0.975	0.892	0.902	0.914
2. Why might someone blame or avoid individuals and groups create stigma because of COVID-19?	0	0	0	0
4. Why do some state's COVID-19 case number sometimes differ from what is posted on CDC's website?	0	0	0	0

government site rather than an academic collection because the content is more on policy than on science. This is why the T&E does not help at all. Only when the process is applied to the queries containing keywords from the CDC answers, could the results improved (see the difference between *OE* and *OP* on this topic in Table 7.

An interesting surprise is the topic "COVID-19 Basics." This topic should be the one that there is some level of communication between laypeople and the academics, and therefore the *nDCG@*10 score for this topic is low (at 0.325). But the T&E process does not help at all (the *nDCG@*10 score of this topic on *EQ* is lower at 0.297). Knowing more on the answers (i.e., *EQ*) and performing T&E the answer queries (i.e., *OE*) does not help at all.

We, therefore, went back to look at the three selected questions in the topic "COVID19 Basics." As shown in Table 8, only question 0, which is about the "novel coronavirus" is a topic that would be discussed in the academic articles, whereas the other two would not be. This is why question 0 performed well in all runs, but the *nDCG@*10 scores of the other two are all 0. These two questions are out of the score of the academic collection like CORD-19; therefore, it is not surprising that the T&E process did not help at all for them.

6 Conclusion

The COVID-19 pandemic has been causing great suffering in the whole world. As the pandemic affects not only individuals' health and daily life, but also their jobs and communities, people naturally have many questions regarding the virus and the pandemic. Fortunately, there is rich academic literature about coronavirus and its related family accumulated over the years. At the same time, researchers have been publishing various academic papers regarding COVID-19 quickly to disseminate their knowledge about the virus. For example, CORD-19 contains over 51,000 scholarly articles regarding COVID-19 and the related coronavirus. Not only these academic articles represent the latest achievements on knowledge about COVID-19, but also they, by going through a rigorous peer-review process, have a high chance to contain high-quality information and are free from the misinformation/disinformation problem that plagues the online social media platforms. However, laypeople have difficulties to access the academic literature on COVID-19 because they might not possess adequate domain-specific knowledge to access the health-related knowledge information.

This motivated us to conduct a study of exploring techniques for helping laypeople to access COVID-19 academic collection such as CORD-19. We developed the SLAC system, which has the T&E process utilizing UMLS CHV. Then utilizing CDCs FAQ questions as the surrogates, we developed a set of questions that laypeople could be interested in to search on the CORD-19 collection. Our experiment results show that the T&E process indeed helped to map laypeople's terms to the medical terms in the academic articles, which achieved significantly better search performance. But we also find that not all topics listed by CDC are meaningful laypeople's needs to search on the CORD-19 collection, thus indicates the scope and the limitation of enabling laypeople to search on academic article collection for obtaining high-quality information. Our results also demonstrate some limitations of our T&E process too.

As part of the future work, we will explore more advanced technology for performing the T&E process for laypeople's information needs and also obtain more diverse information resources for ensuring comprehensive coverage of laypeople's information needs.

References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19-26. doi:10.1145/1148170.1148177

Bhavnani, S. K. (2002, April). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. *CHI'02 Extended Abstracts on Human Factors in Computing Systems,* 610-611. doi:10.1145/506443.506508

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic acids research, 32*(suppl_1), D267-D270. doi:10.1093/nar/gkh061

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of Covid-19 misinformation. *Reuters Institute*.Retrieved from https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

Bullock, J., Pham, K. H., Lam, C. S. N., & Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against covid-19. *ArXiv Preprint*. .arXiv:2003.11336.

Chen, E., Lerman, K., & Ferrara, E. (2020). Covid-19: The first public coronavirus twitter dataset. *ArXiv Preprint.* arXiv:2003.07372.

Chi, Y., He, D., & Jeng, W. (2020). Laypeople's source selection in online health information-seeking process. *Journal of the Association for Information Science and Technology.* doi: 10.1002/asi.24343

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., ... & Scala, A. (2020). The covid-19 social media infodemic. *ArXiv Preprint.* arXiv:2003.05004.

Diaz, F., Mitra, B., & Craswell, N.(2016, August). Query expansion with locally-trained word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol 1),,367-377, doi: 10.18653/v1/P16-1035

Gormley, C., & Tong, Z. (2015). *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. Sebastopol, CA: O'Reilly Media, Inc.

Gu, G., Zhang, X., Zhu, X., Jian, Z., Chen, K., Wen, D., ... & Lei, J. (2019). Development of a consumer health vocabulary by mining health forum texts based on word embedding: Semiautomatic approach. *JMIR Medical Informatics, 7*(2), e12704. doi:10.2196/12704

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., ... & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. *Military Medical Research, 7*(1), 1-10.

Gupta, Y. (2015). *Kibana essentials.* Birmingham, UK: Packt Publishing Ltd.

Hanbury, A. (2012). Medical information retrieval: An instance of domain-specific search. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1191-1192,doi:10.1145/2348283.2348537

Harman, D. (1992). Relevance feedback revisited. *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ,1-10. doi:10.1145/133160.133167

Heymann, D. L., & Shindo, N. (2020). COVID-19: What is next for public health?. *The Lancet, 395*(10224), 542-545.

Hollander, J. E., & Carr, B. G. (2020). Virtually perfect? Telemedicine for COVID-19. *New England Journal of Medicine, 382*(18), 1679-1681.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422-446.

Keselman, A., Smith, C. A., Divita, G., Kim, H., Browne, A. C., Leroy, G., & Zeng-Treitler, Q. (2008). Consumer health concepts that do not map to the UMLS: Where do they fit?. *Journal of the American Medical Informatics Association, 15*(4), 496-505.

Kindig, D. A., Panzer, A. M., & Nielsen-Bohlman, L. (Eds.). (2004). *Health literacy: a prescription to end confusion*. Washington, DC: National Academies Press.

Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts.*ArXiv Preprint.* arXiv:2004.10400 doi:10.1162/qss_a_00066

Lipsitch, M., Swerdlow, D. L., & Finelli, L. (2020). Defining the epidemiology of Covid-19—studies needed. *New England journal of medicine, 382*(13), 1194-1196.

Liu, S., Yang, L., Zhang, C., Xiang, Y. T., Liu, Z., Hu, S., & Zhang, B. (2020). Online mental health services in China during the COVID-19 outbreak. *The Lancet Psychiatry, 7*(4), e17-e18.
Marchionini, G. (1997). *Information seeking in electronic environments* (No. 9). Cambridge, England: Cambridge university press.
Pandey, R., Gautam, V., Bhagat, K., & Sethi, T. (2020). A machine learning application for raising wash awareness in the times of covid-19 pandemic. *ArXiv Preprint* arXiv:2003.07074.
Pylieva, H., Chernodub, A., Grabar, N., & Hamon, T. (2018, November). Improving automatic categorization of technical vs. Laymen medical words using FastText word embeddings. *The 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018)*, pp. 1-10.
Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American society for information science, 41*(4), 288-297.
Schuyler, P. L., Hole, W. T., Tuttle, M. S., & Sherertz, D. D. (1993). The UMLS metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association, 81*(2), 217.
Spark-Jones, K. (1975). Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory.* University of Cambridge.
Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 63). Cambridge: MIT press.
Wang, C. J., Ng, C. Y., & Brook, R. H. (2020). Response to COVID-19 in Taiwan: Big data analytics, new technology, and proactive testing. *Jama, 323*(14), 1341-1342.
Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Mooney, P. (2020). CORD-19: The Covid-19 Open Research Dataset. *ArXiv Preprint.* arXiv: 2004.10706.
World Health Organization. (2020). Coronavirus disease 2019 (covid-19): Situation report, 72. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200401-sitrep-72-covid-19.pdf?sfvrsn=3dd8971b_2
Xie, B., He, D., Mercer, T., Wang, Y., Wu, D., Fleischmann, K. R., ... & Lee, M. K. (2020). Global health crises are also information crises: A call to action. Journal of the *Association for Information Science and Technology.* doi: 10.1002/asi.24357
Zeng, Q., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A., ... & Ngo, L. (2007). Term identification methods for consumer health vocabulary development. *Journal of medical Internet research, 9*(1), e4.
Zhang, D., & He, D. (2017). Enhancing clinical decision support systems with public knowledge bases. *Data and Information Management, 1*(1), 49-60.
Zhang, D., & He, D. (2018, March). Can word embedding help term mismatch problem?–A result analysis on clinical retrieval tasks. In Chowdhury G., McLeod J., Gillet V., Willett P. (Ed.s). *Transforming Digital Worlds. IConference 2018. Lecture Notes in Computer Science* (Vol. 10766, pp. 402-408). Cham: Springer.
Zhang, E., Gupta, N., Nogueira, R., Cho, K., & Lin, J. (2020). Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *ArXiv Preprint.* arXiv:2004.05125.
Zhao, L., & Callan, J. (2012). Automatic term mismatch diagnosis for selective query expansion. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 515-524. doi:10.1145/2348283.2348354