Semantic Shifts Reveal the Multipurpose Use of Potential COVID-19 Treatments

Baitong Chen¹(✉), Qi Yu², Yi Bu³, and Ying Ding⁴,⁵

¹ Department of Library, Information and Archives, Shanghai University, Shanghai, China
² School of Management, Shanxi Medical University, Taiyuan, Shanxi, China
³ Department of Information Management, Peking University, Beijing, China
⁴ School of Information, University of Texas, Austin, TX, USA
⁵ Dell Medical School, University of Texas, Austin, TX, USA

Abstract. When the world is rushing to test potential COVID-19 treatments, the limited resources and the overwhelming information demand the fast understanding of the promising drugs. This paper examines the multipurpose use of potential COVID-19 treatments by mining scientific papers on COVID-19 and related historical coronavirus research. Semantic shifts of the treatment-related entities are recognized to present their various applications in practice. The results identify 10 multipurpose entities. For selected entities, a detailed interpretation is given via text mining analysis about their possible use and rationale in different situations.

Keywords: COVID-19 · Semantic shifts · Treatment-related entities

1 Introduction

The COVID-19 pandemic poses serious challenges to the medical and health systems of countries worldwide, triggering intense global R&D activity to test potential treatments. Since the outbreak of the disease, more than 300 clinical trials have been launched or prepared in order to develop a cure or vaccine [1]. Questions have been raised that whether such a rush of tests is beneficial when the resources are limited. It has been reported that the study of remdesivir, a potential COVID-19 medication, was terminated prematurely in China because it was difficult to enroll patients. Given that so little is known about the disease, mining COVID-19 and coronavirus related literature might be an alternative way to narrow down the scope of trials.

The clinical trials against COVID-19 are mostly based on the current potential treatments, which are mainly repurposed drugs. The amount of the scientific literature that indicate the applications of these drugs is rich but overwhelming, resulting in difficulties for scholars to go through manually.

To help narrow down the trial scopes and facilitate researchers to overcome the information overload, this paper investigates the multipurpose use of potential COVID-19 treatments based on the semantic shifts of the treatment-related entities through text mining. We start by extracting topics from the coronavirus-related literature corpus using
a topic model. The semantic shifts of the treatment-related entities are indicated by the entity-topic distributions aggregated from the training results of the model. Then the multipurpose use of the entities is presented through the semantic shifts and the relevant documents. Finally, we summarize the study and explain the possible limitations.

2 Methods

2.1 Data Set and Topic Extraction

The data is collected from the COVID-19 Open Research Dataset (CORD-19) [2]. CORD-19 is a free resource of scientific papers on COVID-19 and related historical coronavirus research (e.g., SARS, MERS, etc.). The 2020-03-27 released version of CORD-19 is used in this study, containing a total of 45,774 coronavirus-related documents.

The biomedical entities are extracted from the titles and abstracts of the CORD-19 documents, using the named entity annotation method provided by PubTator Central [3]. Two treatment related entity types, Chemicals and Genes, are selected for topic extraction. Prior to extracting topics, all entities are lemmatized to unify the inflection of word forms. Entities with only one letter or appear in only one document are removed.

The Latent Dirichlet Allocation (LDA) [4] is applied for extracting topics from the corpus. LDA is a three-layer Bayesian model that is now widely used in discovering latent topic themes in collections of documents. The LDA model represents each document with a probability distribution over topics, in which each topic is represented as a probability distribution over words. For a detailed explanation of the algorithm, refer to, e.g., reference by Blei [5]. The Gensim library [6] is used for implementing the LDA model, in which the parameter alpha is set to “auto” to learn an asymmetric prior, and eta is set to “None” as the standard value proposed by Gensim.

The number of topics extracted was determined according to the results of the coherence test [7]. The test results showed that when topic number $k = 20$, the coherence score across topics was the highest, and therefore, the number of topics extracted was eventually determined as 20. The sensitivity test regarding the effects of changing the topic number is presented in the end of the method section.

2.2 Detecting Semantic Shifts of Entities

In linguistics, the meaning of a concept $C$ at some moment in time $t$ is defined as a triple $(\text{label}_t(C), \text{int}_t(C), \text{ext}_t(C))$, where $\text{label}_t(C)$ is a String, $\text{int}_t(C)$ refers to the intension of $C$ and is a set of properties, and $\text{ext}_t(C)$ refers to the extension of $C$ and is a subset of the universe. Either of the three elements in the triple changes will cause the meaning change of a concept [8]. In this study, the semantic shifts of an entity are detected from the topic level. The semantic shifts of an entity refer to the extension change, relating to different applications of the entity when distributed in different topics. The topic level semantic shifts are represented by the entity-topic probability distributions from the following two perspectives.
The Over-Time Topic Distribution of Entities. The topic distribution of an entity in a document is parameterized by the variational parameter $\phi$ in the LDA model [9]. The $\phi$ value indicates the likelihood of an entity belonging to a topic in terms of a particular document. Each entity in a document has $k$ topic $\phi$ values corresponding to the $k$ topics extracted (Table 1). After normalizing by the frequency of the entity in the document, the sum of an entity’s $\phi$ values is equal to 1. The same entity from two different documents usually has two different sets of $\phi$ values. For a selected entity, we calculate its average $\phi$ value for each topic in each year. The average $\phi$ values represent the topic probability distribution for the entity in that year. The over-time topic distribution of entities is obtained as the $\phi$ values change over the years.

| Document ID | Year | Entity | Topic 1 | Topic 2 | … | Topic ($k−1$) | Topic $k$ | Sum |
|-------------|------|--------|---------|---------|---|--------------|----------|-----|
| 100         | 2000 | $x$    | $\phi_1$ | $\phi_2$ | … | $\phi_{k−1}$ | $\phi_k$ | 1   |

The Overall Topic Distribution of Entities. Despite the time tags (in years), the topic distribution of an entity can be obtained through the integration of the per-document entity-topic distribution (normalized $\phi$ values). The per-document topic distributions of a target entity are summed and averaged (divided by the number of the documents that the entity appears in), thus producing the overall topic distribution of the entity.

2.3 Detecting Multipurpose Entities from Potential COVID-19 Treatments

The entity list regarding the potential treatments for COVID-19 is obtained from the evidence table provided by the American Society of Health-system Pharmacists (ASHP). The topic distributions of entities indicate their semantic diversities in the topic level, which are associated with the multipurpose use of the potential treatments. For treatments that both proposed in the evidence table and existing in the CORD-19 corpus, the topic distribution of the treatment-related entity is examined to grasp its multipurpose use.

The degree of multipurpose use of entities is measured by information entropy [10]. The entropy of entity $X$ with a distribution over $k$ topics $\{x_1, …, x_k\}$ can be expressed as:

$$H(X) = - \sum_{i=1}^{k} P(x_i) \log P(x_i)$$  \hspace{1cm} (1)

where $P(x_i)$ is the value for the corresponding topic $i$ in the entity-topic distribution. With a larger entropy value, the entity will tend to be distributed in more topics, indicating

1 https://www.ashp.org/-/media/assets/pharmacy-practice/resource-centers/Coronavirus/docs/ASHP-COVID-19-Evidence-Table.ashx.
a higher degree of multipurpose use. When the entity belongs to only one topic, the entropy reaches its minimum value of zero.

From the treatment-related entity list, entities are selected as multipurpose entities when distributed in at least two topics with a probability larger than 0.05, which equals to 1 divided by the topic number 20. The excluded entities in the treatment list are those with only one topic probability exceeds 0.05, when the probabilities of the rest topics are all below the threshold. These excluded entities stably belong to one particular topic, whose use or rationale tends to be explicit and single-purpose.

### 2.4 Interpreting the Multipurpose Use of Entities

The multipurpose use of potential COVID-19 treatments is indicated by the semantic shifts of entities through the overall and over-time topic distribution. For a target entity, the overall topic distribution will be presented in the first place. To interpret the related topics, the most relevant document for each topic is identified based on the document-topic distribution obtained from the LDA training results. Regarding a target entity, the most relevant document of a related topic is the one that contains the entity and has the highest probability of the given topic.

The over-time entity-topic distribution reveals the topic trends of the entity in different years. The relevant documents that have been identified previously usually matches with the peaks of the over-time trends. They are the most relevant documents within the whole time span, thus will definitely be the most relevant ones in the peak years.

If there are more notable peak years in the over-time trends other than the publishing year of the relevant documents extracted before, the most relevant documents of the other notable peaks will be identified and additional discussion for those peaks will be added.

### 2.5 Sensitivity Test When Changing the Number of Topics

We reran our topic model with alternative values of topic number to test the effects when changing topic numbers. The results show that the quality of the entity-topic distribution depends on the coherence of the topic extraction results. When the coherence score across topics are low, the topics are overlapping with each other, resulting in poor accuracy of the multipurpose use indicated by the entity-topic distribution.

It is worth noting that when changing topic numbers, the number of topics is reduced or increased simultaneously for all entities. The entropy scale is the same for all the entities no matter how the topic number is changed. The entity selection threshold (=1/topic number) also changes synchronously with the topic number.

In summary, our choice of the topic number of 20 gives the best coherence score, which is a relatively reliable value for the analysis of the current corpus.

### 3 Results and Discussion

#### 3.1 Entropies of Entities

From the COVID-19 treatment-related entity list, 10 entities that meet with the threshold are finally identified. The entropy values are presented in Fig. 1 in descending order, along
Semantic Shifts Reveal the Multipurpose Use

with their document frequencies. A Kendall’s rank correlation test has been applied to test the correlation relationship between the entropy and document frequency based on data of all the entities in the dictionary. The result is weak positive ($\tau = 0.02052$, $p$-value $= 0.06157$), indicating the association is not practically relevant. To be specific, a high entropy value indicating a potential of multipurpose use is not necessarily associated with the entity appearing in large amount of documents. Entities that only have been studied in limited documents might also possess a considerable chance to have multipurpose usage which may lead to their repurposing for treatments of emerging viruses.

![Fig. 1. Document frequency and entropy of entities](image)

Most of the selected entities in Fig. 1 are antiviral agents that at least in vitro evidence of activity against various viruses, including coronaviruses. These antivirals include Lopinavir, Ritonavir, Arbidol, Favipiravir, Hydroxychloroquine and Remdesivir. Azithromycin is a supporting agent that has been used as adjunctive therapy in the management of certain respiratory diseases (e.g., influenza). Other agents including Niclosamide, Losartan and Nitazoxanide are currently with hypothetical benefit that may have a protective effect in treatment of COVID-19.
3.2 Semantic Shifts Indicating the Multipurpose Use of Entities

The semantic shifts based on the overall and over-time topic distribution of the selected entities are presented in this section. The multipurpose use of each entity is interpreted through the relevant documents as below. Since this research is still in progress, two of the representative entities are presented in this short paper.

**Nitazoxanide.** Nitazoxanide belongs to the antiprotozoal class, structurally similar to niclosamide [11, 12]. Its topic distribution is presented in Fig. 2. The most relevant documents are presented in Table 2 (topics with a proportion larger than 0.05 (=1/topic number) are displayed). In the headers of Table 2, *Tid* means Topic id, and *PMID* means PubMed ID.

![Fig. 2. The entity-topic distribution of nitazoxanide](image)

**Table 2.** Relevant documents of nitazoxanide

| Tid | Year | Relevant doc                                                                 | PMID     |
|-----|------|------------------------------------------------------------------------------|----------|
| 4   | 2018 | Impact of confinement housing on study end-points in the calf model of cryptosporidiosis | 29694356 |
| 13  | 2019 | The FDA-Approved Oral Drug Nitazoxanide Amplifies Host Antiviral Responses and Inhibits Ebola Virus | 31402258 |
| 19  | 2016 | Nitazoxanide, a new drug candidate for the treatment of Middle East respiratory syndrome coronavirus | 27095301 |
| 9   | 2019 | Dissolution Advantage of Nitazoxanide Cocrystals in the Presence of Cellulosic Polymers | 31881696 |

Nitazoxanide was originally developed and commercialized as an antiprotozoal agent, but was later identified as a broad-spectrum antiviral drug and has been repurposed for treatment of influenza and other viral respiratory infections (Table 2, Tid 19). It
exhibits in vitro activity against various viruses, including MERS-CoV and other coronaviruses. Nitazoxanide is also a promising drug that inhibits Ebola virus replication and broadly amplifies the host innate immune response to viruses (Tid 13). Nonetheless, further clinical trials, including dose-ranging trials, and evaluation of combination therapy with other potential antivirals are still needed.

Two cocrystals derived from nitazoxanide, namely, nitazoxanide-glutaric acid and nitazoxanide-succinic acid, may be effective in the treatment of important parasitic and viral diseases (Tid 9). For treatment of diarrhea (Tid 4), nitazoxanide is the only approved chemotherapeutic therapy against cryptosporidiosis, a leading cause of diarrhea in children below five years old. But note that the drug has poor efficacy in HIV positive children.

Figure 3 presents the over-time semantic shifts of nitazoxanide. In general, the publishing years of the most relevant documents in Table 2 match with the peaks of the over-time trends in the semantic-shift graph. However, there is another notable peak in topic 4 in the year of 2014. The related paper is presented in Table 3. Again, it emphasizes that nitazoxanide is a first-in-class broad-spectrum antiviral drug that inhibits a broad range of influenza A and B viruses including influenza A(pH1N1) and the avian A(H7N9) as well as viruses that are resistant to neuraminidase inhibitors. In cell culture assays, it also inhibits the replication of a broad range of other RNA and DNA viruses.

![Fig. 3. Over-time semantic shifts of nitazoxanide](image)

| Tid | Year | Relevant doc                                      | PMID       |
|-----|------|--------------------------------------------------|------------|
| 4   | 2014 | Nitazoxanide: A first-in-class broad-spectrum antiviral agent | 25108173   |

**Remdesivir.** Remdesivir is a broad-spectrum antiviral medication. It is one of the most promising antiviral currently being investigated for COVID-19, and has been used in various clinical trials initiated in US, China, and other countries.
The compound is mainly distributed in three topics (Fig. 4), and the relevant documents are presented in Table 4. The antiviral effects of remdesivir for the treatment of Ebola virus disease have been demonstrated in cell culture and in non-human primates. The drug inhibits Ebola virus RNA synthesis by delaying chain termination (Tid 18). Remdesivir potently inhibits human and zoonotic coronaviruses in vitro and in a SARS-CoV mouse model (Tid 7). Resistance can be overcome with increased, nontoxic concentrations of remdesivir, further supporting the development of remdesivir as a broad-spectrum therapeutic to protect against contemporary and emerging coronaviruses. In the cases that inhibits other coronaviruses, remdesivir is highly efficacious against porcine deltacoronavirus (Tid 17). These results identify remdesivir as a broad-spectrum antviral against various viruses, including both contemporary human and highly divergent zoonotic coronavirus and potentially with the ability to fight future emerging coronavirus.

![Fig. 4. The entity-topic distribution of remdesivir](image)

**Table 4.** Relevant documents of remdesivir

| Tid | Year | Relevant doc                                                                 | PMID        |
|-----|------|-------------------------------------------------------------------------------|-------------|
| 18  | 2019 | Mechanism of Inhibition of Ebola Virus RNA-Dependent RNA Polymerase by Remdesivir | 30987343    |
| 7   | 2018 | Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral Polymerase and the Proofreading Exoribonuclease | 29511076    |
| 17  | 2019 | Broad spectrum antiviral remdesivir inhibits human endemic and zoonotic deltacoronaviruses with a highly divergent RNA dependent RNA polymerase | 31233808    |
4 Conclusion

Revealing the multipurpose use of potential treatment for COVID-19 is crucial for scholars to get fast understanding of the related compounds. In this study, we proposed that the applications of the treatments can be discovered through the entity-topic distributions indicating semantic shifts. The semantic shifts are detected from two perspectives, i.e., the overall and the over-time topic distributions. For each entity, the per-entity topic distribution and the notable peaks in the over-time trends are interpreted by examining the relevant documents in the corresponding topics, which presents the multipurpose use of the entity with different contexts in the topic-level.

This study has the following limitations: (1) Due to the collecting scale of the data set, the multipurpose use of entities primarily focuses on COVID-19 and related historical coronavirus. Their use for non-coronavirus related diseases is less considered. (2) The results are discussed based on the relevant documents. We preferentially considered the most relevant ones, and some of the entity-related but less relevant documents may be left out.

Acknowledgements. This research is supported by the MOE (Ministry of Education of China) Project of Humanities and Social Sciences [grant number 18YJC870002].

References

1. Mullard, A.: Flooded by the torrent: the COVID-19 drug pipeline. Lancet 395(10232), 1245–1246 (2020)
2. Wang, L. L., Lo, K., Chandrasekhar, Y., et al.: CORD-19: the Covid-19 open research dataset. ArXiv (2020)
3. Wei, C.H., Allot, A., Leaman, R., Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 47(W1), W587–W593 (2019)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
5. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (2012)
6. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50 (2010)
7. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015. pp. 399–408. Association for Computing Machinery, Shanghai (2015)
8. Wang, S., Schlobach, S., Klein, M.: Concept drift and how to identify it. Web Semant. Sci. Serv. Agents World Wide Web. 9(3), 247–265 (2011)
9. Hoffman, M., Bach, F.R., Blei, D.M.: online learning for latent dirichlet allocation. In: Advances in Neural Information Processing Systems, pp. 856–864 (2010)
10. Schneider, T.D.: Information Theory Primer with an Appendix on Logarithms. National Cancer Institute (2007)
11. Rossignol, J.F.: Nitazoxanide: a first-in-class broad-spectrum antiviral agent. Antiviral Res. 110, 94–103 (2014)
12. Xu, J., Shi, P.Y., Li, H., Zhou, J.: Broad spectrum antiviral agent niclosamide and its therapeutic potential. ACS Infect. Diseases 6(5), 909–915 (2020)