APPLYING SEMANTIC TECHNOLOGIES FOR KNOWLEDGE DISCOVERY FROM PATENTS – A CASE STUDY FOR THE COVID-19 PANDEMIC

Hidir Aras, Ph.D.
Project Leader TDM / IT Development and Applied Research
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

Posted August 3, 2020

CONTENT

INTRODUCTION

In the last decades, an increasing number of scientific publications and patents has been published worldwide, which constitute an essential resource for aiding current drug research and helping the development of innovative methods and applications in the fight against novel diseases. Although patents related to a new phenomenon like the COVID-19 pandemic will take time to be submitted and granted, previously published patents may already contain essential knowledge related to vaccines, antiviral drugs, medical treatment, specific information such as virus morphology, diseases pathology, and side effects and their interrelations. In many instances, relevant data is only disclosed in patent documents. Commonly, patent applications are first published approximately 18 months after they have been filed. This means that patents related to COVID-19 may not appear until mid- or end-2021. Nevertheless, patent applications dealing with SARS-CoV-2 and patents related to other events of coronavirus outbreaks, especially SARS-CoV-1 and MERS, can be particularly helpful as a resource for

gaining new insights during the development of new methods of treatment against COVID-19.

Finding and exploiting knowledge in patent texts is a difficult task, as it requires a high intellectual effort to sift through several hundreds or thousands of documents with up to 25 pages of technical descriptions, background knowledge or examples of an invention. Established patent search approaches with keyword queries are often only a first step within a sequence of elaborated methods.

In recent years, existing rule-based analysis techniques have been improved through the increasing use of text mining, machine learning and natural language processing. It has been shown that the inclusion of semantic information, be it implicit semantics (e.g. through word embedding) or explicit domain knowledge using expertly created ontologies, can help to efficiently explore data collections and to ask and answer complex natural language questions in an interactive way. This makes it possible both to collect factual information and to explore, justify and identify further (derived) more substantial information within the texts by writing complex (related, subsequent) semantic user queries that can easily be linked to the users' natural language question, which is associated with important aspects of one or more research questions.

Moreover, the results of such a semantic query aren't just a list of ranked patent documents but rather specific answers (e.g. bio-medical knowledge entities) extracted from a certain section or paragraph of the patent text linked to other highly relevant knowledge sources such as the Linked Open Data (LOD)[1] cloud or any other relevant semantic knowledge base.

Such new means for exploration and analysis of large amounts of data will help scientists to benefit from patent information and/or scientific articles in combination and exploit specific information or gain new evidence for answering essential questions for their own research.

For scientists in life sciences, urgent research questions for combating the COVID-19 pandemic are the development of vaccines and antibody-based approaches, or the identification and testing of active pharmaceutical ingredients. Any knowledge related to these questions hidden in certain sections, paragraphs, or sentences of the description text or the claims of a patents is of great importance. During the current COVID-19 pandemic, Kaggle (a data science community platform) posted a call[2] for Data Scientists to develop methods that help to answer a list of research questions related to COVID-19. As an important step, a research dataset with scholarly articles called CORD-19 (COVID-19 Open Research Dataset) has been made publicly available and is regularly updated by the White House and a coalition of leading research groups.

Motivated by all these developments and based on our many years of experience in patent search, retrieval and analysis, we have started the development of new and innovative methods for searching and analyzing collections of patent information and scientific texts employing enhanced methods for search and semantic analysis.

PART I: SEMANTIC SEARCH AND DEEP PATENT TEXT ANALYSIS

Scientists try to apply existing knowledge in order to gain new insights and to find answers for their complex research questions. Most often their search for answers is interdisciplinary, necessitating to combine knowledge from several relevant while striving for a specific solution. Hence, the informational requirements of scientists can become quite complex, i.e. their research questions or search queries, must be formulated in a sufficiently meaningful way to be comprehensible to other researchers. In this sense, keyword-based statistical methods are limited in various aspects, and semantic information must be taken into account in order to represent and seek for useful information for

satisfying specific information needs. While a search query can be formulated by asking a natural language question, users often formulate a query with one or more sentences describing the topic of interest.

As part of our ongoing work, we are developing a data processing and analysis pipeline that applies natural language processing and machine learning techniques for deep text analysis of patents and scientific articles. We have integrated the data analysis pipeline into a semantic search engine (Figure 1) based on patent-specific embedding models such as word or phrase embedding with word2vec, GloVe or BERT. The engine is based on a patent corpus created by extracting relevant documents from all full-text databases of STN.

The semantic search engine will be able to identify and extract the most important sentences and passages from patent texts in order to answer specific questions from the users. While one goal is to find related answers more quickly, it will further contribute to a more systematic examination of the patent result set by taking into account relevant annotations such as biomedical entities. In particular, automatically extracted and highlighted knowledge (added value) will help to gain important insights and uncover unused information in the textual description of a solution or invention.

We plan to release a first prototype of the patent search engine for COVID-19 soon and open it to the scientific community. For the COVID-19 corpus, typical example semantic queries are, for example, "vaccines and therapeutics for coronavirus", "drugs for treating coronavirus patients", "*viral inhibitors against coronavirus such as naproxen, clarithromycin, and minocycline*".

DATA PROCESSING AND SIMILARITY ANALYSIS

The patent dataset consists of patents related to the coronavirus and related aspects such as formulations, therapeutic use, etc. Therefore, 105,700 patent documents (approx. 36,000 patent families) were extracted from STN and evaluated on the basis of search queries defined by FIZ Karlsruhe experts. The text of the patent documents (title, abstract and claims) was processed using natural language processing and shallow syntactic analysis to break long sentences into smaller phrase chunks and normalize them to identify descriptive noun phrases. Based on the words and phrases of the patent text, an embedding model was created to capture the semantics of each word or phrase using deep learning.

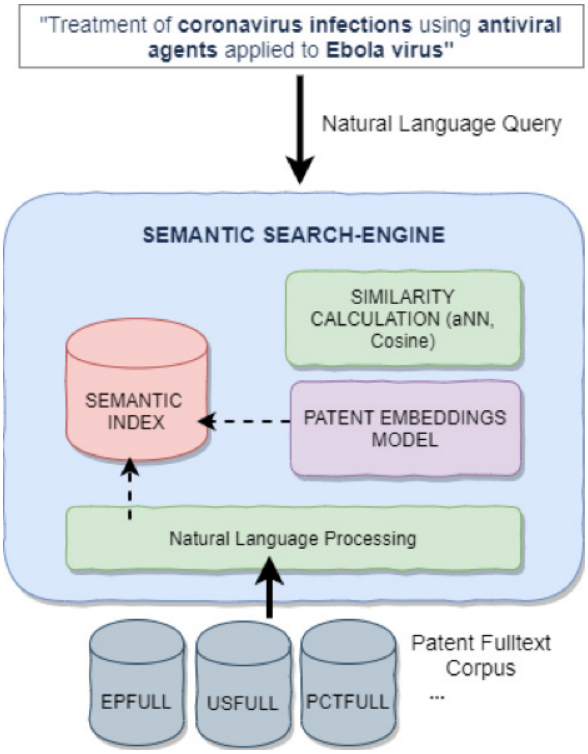

Figure 1. Semantic Search Engine for COVID-19.

The approximate similarity matching index is created by converting the documents in the patent corpus and the user query into their vector representation in an appropriate feature space employing the embeddings model. Hereby, a domain-specific patent embeddings model is utilized to represent each sentence/paragraph in the vector space.

Related patents are identified by measuring the semantic similarity between a user query and the sentences/paragraphs in the full text of a patent. An approximate nearest neighbor algorithm is

used to find the most relevant and semantically similar sentences/paragraphs in relation to a given query vector. The final ranked list of patent documents can be obtained by averaging the highest top similarity scores of the sentences/paragraphs found or taking the maximum of all hits in a winner-takes-all manner.

EXTRACTING BIO-MEDICAL ENTITIES FROM PATENT TEXT

We also extract the biomedical entities such as diseases, genes and chemical components for each document/record in the search results. A phrase matching pipeline is being developed to extract coronavirus-related entities including those related to COVID-19 for each patent document. The terms of the search query and extracted entities are highlighted and linked to the external knowledge base such as Unified Medical Language System UMLS[3] (based on the concept identity number, definitions and aliases). The development of the search engine is currently in progress and a first prototype is available for evaluation by experts.

A SEARCH EXAMPLE – FOR THE AVIGAN (FAVIPIRAVIR) CASE

In the following, an example of a semantic search query for identifying important patents related to antiviral agents for combating virus diseases is shown. In our scenario, scientists start their search by expressing their information need with one or more sentences such as:

"Treatment of coronavirus infections using antiviral agents applied to Ebola virus".

After query execution the search engine returns the following list of highly related and most relevant top ten patents.

RANK	PN	PUB DATE	FAMILY ID	APPLICANT	SCORE
1	US 20180021333	25.01.2018	52264442	INSERM	0.84
2	WO 2016061549	21.04.2016	50209372	SIRNAOMICS, INC.	0.824
3	WO 2017211843	14.12.2017	56675135	ABIVAX	0.818
4	KR 2004072720	18.08.2004	197442	PHARMACIA & UPJOHN COMPANY LLC	0.816
5	WO 2017202789	30.11.2017	56663582	INSERM	0.81
6	CN 111093627	01.05.2020	61522369	NAN	0.788
7	AU 2010306914	19.04.2012	900169	GEMMUS PHARMA, INC.	0.771
8	WO 2019182947	26.09.2019	64705915	PURDUE RESEARCH FOUNDATION	0.766
9	US 20130210915	15.08.2013	17349171	EIRIUM AB	0.766
10	KR 2017016975	14.02.2017	50552087	US HEALTH, UNIVERSITY OF KANSAS	0.754

The patent application listed at rank 5 **WO2017202789** with the title "METHODS AND PHARMACEUTICAL COMPOSITIONS FOR THE TREATMENT OF FILOVIRUS INFECTIONS", was published 2017 for combating the Ebola virus disease employing the Avigan antiviral drug together with another important patent application with the publication number **US20180021333**.

Although neither the brand name Avigan nor its drug name favipiravir are mentioned in the query the deep learning models allow to calculate the relatedness to coronavirus infections by considering inherent semantics from the patent corpus.

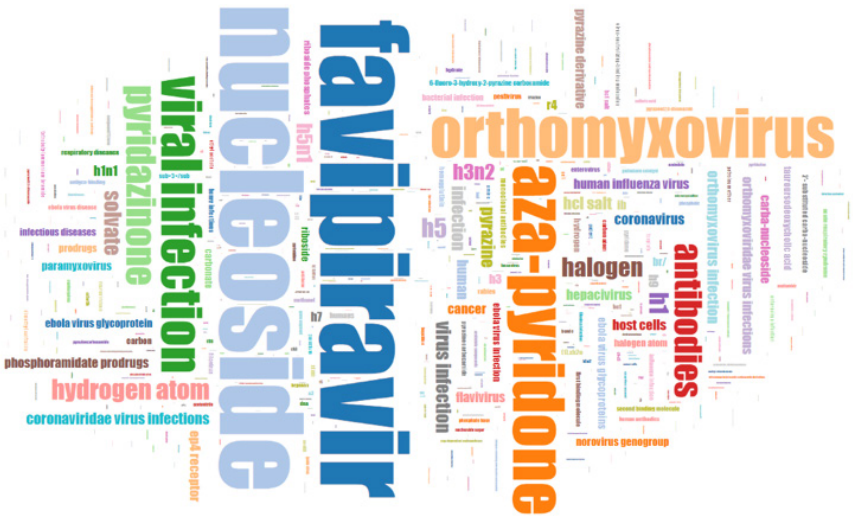

Figure 2. Bio-medical entities extracted from the Avigan patent corpus (size corresponds to occurrence frequency).

Looking at the entities listed in the patents (see overview visualization as term cloud in Figure 2) the following important *basic types and biomedical terms* are identified and classified automatically employing bio-medical named entity recognition on the fly:

- ORGANISM: human, ebola virus, filo virus
- DISEASE: filovirus infection, ebola virus infection
- CHEMICAL: 6-Fluoro-3-hydroxy-2-pyrazine-2-carboxamide
- DRUG: favipiravir, gemcitabine, obatoclax

Other important entities are extracted from the top ten to 20 most relevant results by analyzing more related patent applications (AU 2010306914, WO 2016172205 and possibly WO2019014247) and entity types:

- ORGANISM: h1n1, h3n2, h5n1 (**influenza**), cdv-3, cdv-11 (**canine distemper**), protozoan
- DISEASE: **ebola**, influenza, canine distemper, parasitic infection, viral infection, respiratory disease, **coronavirus disease**, pest
- CHEMICAL: **6-Fluoro-3-hydroxy-2-pyrazine-2-carboxamide**, 2',2'-difluoro-2'-deoxycytidine
- DRUG: **favipiravir, t-705, gemcitabine, obatoclax**
  - Properties: anti-retroviral, **antiviral, rna-polymerase-inhibitor, acute respiratory syndrome**
- GENE/PROTEIN: **ep4-receptor**, hemagglutinin (HA)

Moreover, analyzing entity-type relations employing standard data science methods allows to extract lightweight ontological dependencies between infections caused by various virus types and associated drugs/compounds used in therapeutic treatment.

EXTRACTED ONTOLOGY:

- **infection** ⇢ **virus** ⇢ ‹ebola, filovirus, hiv, hcmv, orthomyxovirus, hepatitis c, human influenza›
- **drug** ⇢ prodrug ⇢ ‹favipiravir, gemcitabine, obatoclax›

The patent application WO2017202789 from the above list was also reported in a comprehensive case study[4] by experts at FIZ Karlsruhe, who consulted chemical structure databases like Derwent Markush Database or Registry as well as value-added databases such as Derwent World Patents Index or CAplus in order to identify patents concerning favipiravir, a substance that is in discussion as a COVID-19 treatment. The Derwent title contains the term favipiravir, while the semantic search is able to exploit the semantic context of the search query in order to identify this important drug as well by just analyzing patent full-text, without an exhaustive search and consulting value-added databases.

"Treating a *filovirus infection* comprising Ebola virus disease, comprises administering combination of *favipiravir* and obatoclax, or a combination of *favipiravir* and *gemcitabine*".

Moreover, the patent analysis describes favipiravir as a *small molecule prodrug* which is metabolized in its *active form favipiravir-RTP*. Besides combating the ebola disease infections, this drug was also involved in treatment against *human influenza* (type A and B), *filovirus and hepatitis c*.

Digging deeper and investigating the extracted entity types via co-occurrence analysis (Figure 3) allows to explore knowledge about various *inhibitory effects* on different virus types, the applied *pharmaceutical synthesis* and the compound structures involved. For example, the "ep4 receptor" plays an important role in the treatment of respiratory viral diseases. Here, antiviral agents such as favipiravir are used together with an *ep4-receptor agonist*. Such and more specific knowledge entities and their interrelations can be discovered by means of the presented semantic analysis and exploration of the patent corpus. While at higher levels conceptual entities can be observed in the visualization, zooming to certain areas, e.g. to the human influenza virus, will reveal more details, such as the *rna polymerase inhibition* activity of pyrazine derivatives and their role in various viral diseases, such as the human influenza virus.

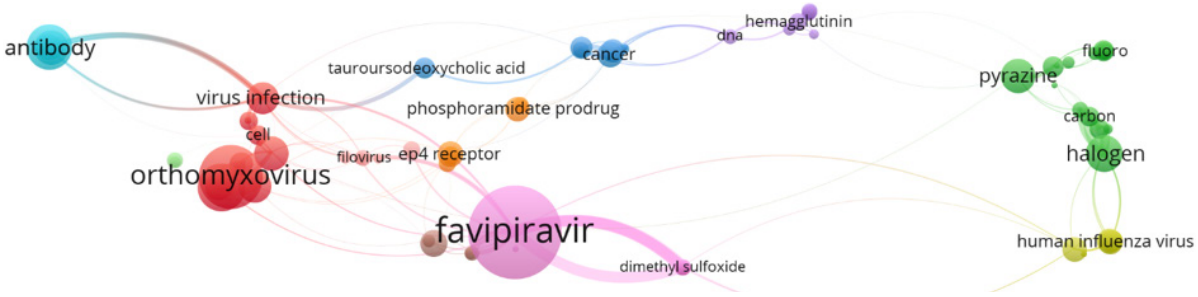

Figure 3. Co-occurrence Analysis of Favipiravir for the treatment of virus diseases[5].

SUMMARY AND OUTLOOK

A semantics-driven search engine makes it possible for users to explore patent information and gain insight into selected topics, methods and specific inventions by allowing users to formulate their information needs in a few sentences. For the analysis of results the user can focus on using widely applied data science methods such as similarity comparison, classification and clustering or topic analysis. Especially for scientists who are not trained to manually search and evaluate a significant amount of patent documents using classical patent retrieval approaches, the complexity and learning curve for the use of such analysis tools is reduced.

In addition, techniques for extracting specific entities, such as bio-medical named entities, will provide more efficient access to the results under consideration. Since these techniques not only identify important parts of the patent text by highlighting relevant terms, they also provide conceptual information about the identified types. This makes it easier to link such entities to a knowledge graph that is of interest for further work. For the application under investigation, bio-medical named entities from patents can be linked to the covidGraph[6] (a knowledge graph for COVID-19), which allows for the simultaneous investigation of multiple sources of knowledge such as scientific articles, drug and chemical information, etc.

In particular, rich information on the chemical structure from published datasets, such as the research dataset on antiviral drug candidates[7] published by CAS in March 2020, enables the study of antiviral strategies with small molecules and biologicals targeting complex molecular interactions involved in the infection and replication of coronaviruses, thus providing extremely important and valuable information for drug development and medical science. In addition to supporting the efforts for the development of new drugs based on active ingredients known to be effective against other RNA viruses, including SARS-CoV and MERS-CoV, scientists easily can obtain information about the chemical structures mentioned in the early patents, while taking into account the usage of different generic drug names, trade names or terminologies for these compounds. In Part II of this blog series we will explore new means for exploring patent information and scientific texts using Deep Learning and Knowledge Graphs for different application scenarios such as Patent Landscaping[8].

REFERENCES

1 https://data.epo.org/linked-data/
2 COVID-19 Open Research Challenge (Tasks) https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
3 https://www.nlm.nih.gov/research/umls/
4 https://www.fiz-karlsruhe.de/en/nachricht/fiz-special-corona-favipiravir-kandidat-fuer-ein-medikament-gegen-covid-19
5 Mark Neumann, Daniel King, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing Published in BioNLP@ACL 2019 Computer Science.
6 https://covidgraph.org/
7 https://www.cas.org/covid-19-antiviral-compounds-dataset
8 https://link.springer.com/article/10.1007/s10506-018-9222-4