Semantic Search for Biomedical Texts using Predicate-Argument Structure

Mohammed Alliheedi1 and Robert E. Mercer2
1Department of Computer Science, Al Baha University, Prince Mohammad Bin Saud, Al Bahah 65527, Saudi Arabia
2Department of Computer Science, The University of Western Ontario, 1151 Richmond St., London, Ontario, Canada

Keywords: Biomedical Ontology, Semantic Search, Semantic Roles, Knowledge Representation, Formal Ontology.

Abstract: In this position paper we argue that using semantic roles in addition to using biologically-oriented ontologies and databases (or knowledge bases) will further enhance the generation of RDF triples that can be collected from biomedical text. RDF triples have been used to enhance semantic search beyond the simple use of linguistically oriented additions such as synonyms. We wish to focus on drug-virus interactions.

1 INTRODUCTION

In this position paper we argue that using semantic roles in addition to using linguistically-oriented and biologically-oriented ontologies and databases (or knowledge bases) will further enhance the generation of RDF triples that can be collected from biomedical text. RDF triples can then be used to enhance semantic search beyond the simple use of linguistically oriented additions such as synonyms (Kohlschein et al., 2018). Linguistically-oriented ontologies such as WordNet (Miller, 1995) can provide synonyms, hyponyms, hypernyms, and meronyms. VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) and BioFrameNet (Dolbey et al., 2006), an important biologically-oriented extension to FrameNet (FrameNet, 2020), provides means to connect verbs with semantic roles. BioFrameNet also provides connections to biological ontologies, such as, GO (Ashburner et al., 2000) and Entrez Gene (Entrez Gene, 2020). Recently, interest to extend VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) into the biological domain has reemerged with BioVerbNet (Chiu et al., 2019), which follows on the earlier work of Lippincott et al. (Lippincott et al., 2013), and a separate nascent investigation of biochemistry experimental method procedure verbs (Alliheedi et al., 2019b; Alliheedi and Mercer, 2019; Alliheedi et al., 2019a; Alliheedi et al., 2019c).

We wish to focus on drug-virus interactions and are particularly interested in searching the COVID-19 Open Research Dataset (Wang et al., 2020) for drug-virus interactions and how the drugs interact with the viruses. While these just mentioned linguistic ontologies provide important information about the verbs in the form of semantic roles that the verbs take as linguistic arguments, what is missing are the relations among biological entities. These relationships can be captured by machine accessible RDF triples, subject-verb-object triples that use the standard nomenclature provided by various terminology ontologies. Triples in the biology domain are collected by tools such as those found in Bio2RDF.org (Bio2RDF, 2020). This type of machine-accessible knowledge is strongly argued for in the FAIR guidelines (Wilkinson et al., 2016). Our immediate objective is to use the semantic role information found in the linguistic ontologies to generate the RDF triples found in scientific publications and the text sections found in many of the biological knowledge bases. Our long term goal is to use these RDF triples in a semantic search engine targeted on the COVID-19 Open Research Dataset.

2 ONTOLOGIES AND KNOWLEDGE SOURCES

In this section, we give an overview about existing biologically oriented ontologies in the literature. Table 1 shows some of well known ontologies that have been developed over the last decades. These ontologies include the Gene Ontology (GO) (Ashburner et al., 2000), the Foundational Model of Anatomy (FMA) (Rosse and Mejino Jr, 2003), the Relation Ontology (RO) (Smith et al., 2005), the
Table 1: List of developed ontologies in the literature.

<table>
<thead>
<tr>
<th>Ontology name</th>
<th>Domain</th>
<th>Developed by</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Gene Ontology (GO)</td>
<td>describes knowledge of the gene domain.</td>
<td>(Ashburner et al., 2000; Gene Ontology Consortium, 2011)</td>
</tr>
<tr>
<td>The Foundational Model of Anatomy Ontology (FMA)</td>
<td>is an ontology about human anatomy</td>
<td>(Rosse and Mejino Jr, 2003)</td>
</tr>
<tr>
<td>The Relation Ontology (RO)</td>
<td>consists of relations intended to be used across various ontologies in OBO Foundry</td>
<td>(Smith et al., 2005)</td>
</tr>
<tr>
<td>The Ontology of Scientific Experiments (EXPO)</td>
<td>defines the general knowledge about scientific experimental aspects (e.g., methodology and design)</td>
<td>(Soldatova and King, 2006)</td>
</tr>
<tr>
<td>The Ontology for Chemical Entities of Biological Interest (ChEBI)</td>
<td>a database of molecular entities focusing on small molecules</td>
<td>(Degtyarenko et al., 2007)</td>
</tr>
<tr>
<td>The Molecular Methods Database (MolMeth)</td>
<td>a resource consists of scientific protocol ontologies</td>
<td>(Klingström et al., 2013)</td>
</tr>
<tr>
<td>The Ontology of Scientific Experiments (EXACT)</td>
<td>describes scientific protocols and experiments</td>
<td>(Soldatova et al., 2013)</td>
</tr>
<tr>
<td>Semanticscience Integrated Ontology</td>
<td>describes scientific protocols and experiments</td>
<td>(Dumontier et al., 2014)</td>
</tr>
<tr>
<td>The Ontology for Biomedical Investigations (OBI)</td>
<td>a resource for annotating biomedical investigations</td>
<td>(Bandrowski et al., 2016)</td>
</tr>
</tbody>
</table>

The Gene Ontology (GO) describes knowledge of the gene domain. GO was built using various resources such as those in (FlyBase Consortium, 2003; Blake et al., 2000; Ringwald et al., 2000; Ball et al., 2000). Similarly to GO, ChEBI (Degtyarenko et al., 2007) was created using data from several resources such as IntEnz (Fleischmann et al., 2004), KEGG COMPOUND (Kanehisa et al., 2006), and the Chemical Ontology. While GO and ChEBI focus on Gene and molecular entities, OBI focuses on the other hand, focuses on the annotations of biomedical investigations and provides standard tools to represent study design, protocols and instrumentation used, the data generated, and the types of analysis performed on the data. Several ontologies (Courtot et al., 2008; Brinkman et al., 2010; Zheng et al., 2013; Soldatova et al., 2013; Dumontier et al., 2014) were developed and based on the OBI ontology.

EXPO provides detailed descriptions of various aspects of scientific experiments and their relationships (Soldatova and King, 2006). Descriptions of experimental processes are provided by OBI, and three real-world applications are discussed in (Brinkman et al., 2010). Some of the relations in these applications (e.g., inputs, outputs) add to the ontologies that describe biological entities. The beta cell genomics application ontology (BCGO) (Zheng et al., 2013) also uses OBI, but it tends to be a more descriptive ontology than some of the others that use OBI, but some of the relations in RO, the relation ontology (Smith et al., 2005), that are used (e.g., produces, translate to) do have an ordering sense.

EXACT (Soldatova et al., 2013) and the Semanticscience Integrated Ontology (Dumontier et al., 2014) describe scientific protocols and experiments.

MolMeth is a database which contains scientific protocol ontologies that conform to a set of laboratory protocol standards (Klingström et al., 2013).

Other ontologies describe general concepts that...
are useful to a biochemistry procedure-oriented ontology include: Ontologies consist of process such as (Lenat et al., 1985) and (Schlenoff et al., 2000), ontology for units of measure (Rijgersberg et al., 2013), classification of scenarios and plans (CLASP) (Devanbu and Litman, 1996), and materials ontology (Ashino, 2010). Foundational theories such as process calculus and regular grammar are essential for the formalization of procedure-oriented ontologies.

We consulted three knowledge sources in our discussion in Section 3. MedChemExpress (MCE)
2, a company, offers a wide range of high quality research chemicals and biochemicals. Its website provided synonym information for some biological elements. GeneCards
3 is a database of information on all known and predicted human genes. UniProt
4 is a database of protein sequence and functional information.

3 SEMANTIC ROLES

Semantic roles are defined as “the underlying relationships that participants have with a verb in a clause”
5. Minsky defined frame as “a data-structure representing a stereotyped situation” (Minsky, 1974), with frames having a header, slots and slot fillers. Fillmore (Fillmore, 1976) introduced the notion of frame semantics as a theory of meaning. A semantic frame is defined by Fillmore as “any coherent individuatatable perception, memory, experience, action or object”. In other words, these are coherently world events or experiences. In this work, we are interested in developing frame semantics at the verb level so that our headers are verbs and our slots are semantic roles, filled by the words which represent these roles. For example, to understand the word “buy”, one would access the knowledge contained in the commercial transaction frame which includes words such as the person who buys the goods (buyer), the goods that are being sold (goods), the person who sells the goods (seller), and the currency that the buyer and seller agree on (money). Motivated by Fillmore’s theory of frame semantics, FrameNet (Baker et al., 1998) was developed to create an online lexical resource for English. This framework includes more than 170,000 manually annotated sentences and 10,000 words. The computational linguistic community has been attracted to the concept of frame semantics and developed computational resources using this concept, such as VerbNet (Kipper-Schuler, 2005), an on-line verb lexicon for English and PropBank (Palmer et al., 2005), an annotated corpus with basic semantic propositions.

Since we are focusing on drug-virus interaction verbs with the associated semantic roles, we are particularly interested in analyzing the COVID-19 Open Research Dataset (Wang et al., 2020). Verbs evoke semantic roles in writing. Semantic roles provide salient pieces of information about experimental steps. Semantic roles are crucial aspects of identifying relevant information in the biomedical texts. This relevant information is essential to generate the RDF triples (i.e., subject, predicate, and object) which have been used in the literature (Hu et al., 2017). We can use the work by (Alliheedi and Mercer, 2019) to label each sentence with the appropriate semantic roles (see Table 2). We have used an annotation scheme for identifying the structured representation of knowledge in a set of sentences of biochemistry articles (Alliheedi and Mercer, 2019). The list of semantic roles include: Theme, Patient, Agent, Location, Goal. These semantic roles identify the arguments of both verbs (e.g., administered) and nominalised verbs (e.g., inhibitor is a nominalized form of the verb inhibit). Furthermore, the use of ontology that describes drug-virus interactions is salient because it provides the relations among various molecules.

Table 2: List of Semantic Roles developed in the work of (Alliheedi and Mercer, 2019).

<table>
<thead>
<tr>
<th>Semantic role</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td>Generally a human or an animate subject.</td>
</tr>
<tr>
<td>Patient</td>
<td>Participants that have undergone a process.</td>
</tr>
<tr>
<td>Predicate</td>
<td>A word that initiates the frame.</td>
</tr>
<tr>
<td>Theme</td>
<td>Participants in a location or undergoing a change of location.</td>
</tr>
<tr>
<td>Goal</td>
<td>Identifies a thing toward which an action is directed or a place to which something moves.</td>
</tr>
<tr>
<td>Factitive</td>
<td>A referent that results from the action or state identified by a verb.</td>
</tr>
<tr>
<td>Location</td>
<td>The physical place where the experiments took place.</td>
</tr>
<tr>
<td>Instrument</td>
<td>An object or force that comes in contact with an object and causes some change in them.</td>
</tr>
</tbody>
</table>

2https://www.medchemexpress.com
3https://www.genecards.org
4https://www.uniprot.org
5https://glossary.sil.org/term/semantic-role
4 EXAMPLE

We use as an example, a paper that discusses the antiviral drug sofosbuvir (Kirby et al., 2015). The title and abstract of this paper states that sofosbuvir is an inhibitor of the virus. This produces a very high-level RDF triple. We would like to have triples that describe this inhibition in more detail. What is the biochemistry that causes this inhibition? Reading further, we encounter some information about how this inhibition takes place. We first find mention of two metabolites of sofosbuvir, GS-461203 and GS-331007. But the abstract is not that informative regarding why these metabolites are mentioned. In the full text of the paper we find “GS-461203, the pharmacologically active nucleoside analog triphosphate metabolite of sofosbuvir, is incorporated into HCV RNA by NS5B polymerase, where it acts as a chain terminator.” So, this provides us with the information that we need for the GS-461203 metabolite. We remain interested in following up with GS-331007. We will need to look at other information sources. Figure 1 provides a visual presentation of the search that we now describe. We cannot find mention of GS-331007 in some of the knowledge bases that we search. However, we find in MedChemExpress (MCE) that PSI-6206 is a synonym. Now searching for PSI-6206, it is found in GeneCards under the POLR2A Gene. We note information about PSI-6206 and its importance as an inhibitor of the Hepatitis virus. But no other information is forthcoming. So, we then turned to UniProt. While looking in UniProt under POLR2A gene under Homo Sapien we find “(Microbial infection) Acts as an RNA-dependent RNA polymerase when associated with small delta antigen of Hepatitis delta virus, acting both as a replicate and transcriptase for the viral RNA circular genome.” and a link to a publication (Chang et al., 2008). In this publication we find important information about how the hepatitis delta virus uses host RNA polymerase which is the protein encoded by the POLR2A gene. “Previous studies have indicated that the replication of the RNA genome of hepatitis delta virus (HDV) involves redirection of RNA polymerase II (Pol II), a host enzyme that normally uses DNA as a template. … Taken together, we have demonstrated that with a low concentration of amanitin that only inhibited Pol II transcription and did not affect host Pol I or Pol III transcription there was a significant inhibition for HDV genomic and antigenomic RNAs. Thus, we believe that Pol II is required for the transcription of both genomic and antigenomic HDV RNAs.” (Chang et al., 2008).

---

6https://www.medchemexpress.com/PSI-6206.html
7https://www.genecards.org/Search/Keyword?queryString=PSI-6206
8https://www.uniprot.org
9https://www.uniprot.org/uniprot/P24928

Figure 1: Steps of the example to find information about GS-331007, a metabolite of sofosbuvir.
We now have a number of sources that have provided factual information and textual information. The factual information (e.g., sofosbuvir is a drug) can be readily made into an RDF triple. Knowledge about drugs (drugs ending in *ir* are anti-viral drugs) can be used to further refine the triple. The textual information can be analyzed using the semantic roles that we have discussed in Section 3. The semantic roles can be used to form RDF triples (agent is the subject, predicate is the verb, and theme and instrument are the objects). As an example, we can label "GS-461203, the pharmacologically active nucleoside analog triphosphate metabolite of sofosbuvir, is incorporated into HCV RNA by NS5B polymerase..." using the semantic roles in Table 2 as follows:

GS-461203/Patient; is incorporated/Predicate; into HCV RNA/Theme; by NS5B polymerase/Agent.

This example produces a number of RDF triples:

**Triples**
- sofosbuvir is-a drug
- sofosbuvir is-a anti-viral drug
- hepatitis delta is-a virus
- sofosbuvir inhibitis Hepatitis
- GS-461203 is-a metabolite of sofosbuvir
- hepatitis C virus is-a-synonym-of HCV
- HCV NS5B polymerase incorporates GS-461203 into HCV RNA
- GS-461203 acts-as chain terminator of HCV RNA synthesis
- GS-331007 is-a metabolite of sofosbuvir
- hepatitis delta virus is-a-synonym-of HDV
- P*OLR*2A gene encodes RNA polymerase
- RNA polymerase is-essential-for cell function
- RNA polymerase acts-as an RNA-dependent RNA polymerase
- RNA polymerase associates-with small delta antigen of Hepatitis delta virus
- RNA polymerase acts-as replicate for the viral RNA circular genome
- RNA polymerase acts-as transcriptase for the viral RNA circular genome
- RNA polymerase II is-a RNA polymerase
- RNA polymerase II is-required-for transcription of genomic HDV RNA
- RNA polymerase II is-required-for transcription of antigenomic HDV RNA

5 **SEMANTIC SEARCH**

Standard information retrieval has relied on the vector space model (Salton et al., 1975). The vector space model for representing documents in high-dimensional vector space has been validated by decades of research and development. This model represents documents and queries as vectors (based in various ways, such as the almost universally used TF-IDF (Salton et al., 1975), on the words occurring in the document and the query) and then using some measure (typically, the cosine difference between the query and each document vector) to retrieve relevant documents. Although very popular these types of representation of document semantics based solely on first order document-term statistics, such as TF-IDF, are limited in their expressiveness and search recall.

Semantic search (or semantic information retrieval) started in the early days of Artificial Intelligence (Raphael, 1964). There have been two main directions to improve the vector space model: 1) improve the vector representation of the query and the document in order to improve the relationship between the query and the relevant document vectors, and 2) provide meta-knowledge, typically in the form of synonyms, hypernyms, and hypernyms, but sometimes with the more distantly related meronyms, homonyms, or semantic fields, in general. Some of the methods that have been developed are directly applicable to semantic search.

The typical techniques used are to include synonyms, hypernyms, and hypernyms from the WordNet ontology (Fernández et al., 2011). Although we will investigate this knowledge source, we will also use the biomedical ontologies (Bandrowski et al., 2016; Courtot et al., 2008; Brinkman et al., 2010; Zheng et al., 2013; Soldatova et al., 2013; Dumontier et al., 2014; Soldatova and King, 2006; Zheng et al., 2013) to build our own ontologies (Liang et al., 2017; Seitter et al., 2016) because the specialized biomedical domains have specialized meanings for many common words and many technical terminologies that require specialized ontologies. In addition to these linguistically motivated semantic search extensions, we will also use RDF triples that have been derived from biomedical texts and information sources using our suggested application of semantic roles.

As an example of how a query with additional semantic information is used in a semantic search: **Query**: How does sofosbuvir help infected-patients with hepatitis C virus?

A search request is processed using the following procedure: The user search query is analyzed with linguistic tools and is matched against the word ontologies and the RDF data source, retrieving synonyms and RDF descriptions that semantically match each entity in the query with semantic information encoded in the RDF triples. A full text search query of various databases, based on the data returned by the previ-
ous step, is generated. All resulting information is returned to the user.

For this example query, the search would return papers and other information that talks about how sofosbuvir works to inhibit hepatitis viruses, such as the drug acting to prevent hepatitis C RNA synthesis.

6 CONCLUSION AND FUTURE WORK

The long-term goal that has motivated this position paper is to provide a semantic search system to query the COVID-19 Open Research Dataset (Wang et al., 2020). We have argued that RDF triples can assist this semantic search. Moving toward this goal, our immediate objective is to use the previously presented semantic roles to assist the automatic generation of the RDF triples found in scientific publications and the text sections found in many of the biological knowledge bases. This step together with information derived from ontologies and knowledge bases are necessary for building a semantic search system that is capable of extracting the relevant information from biomedical texts that goes beyond simple keyword searches.

Our proposed idea is the first step towards developing an automated semantic search. We aim to refine the methods discussed in this paper and to develop a semantic search system. We intend to focus on the drug-virus interactions (e.g., how different antiviral drugs interact with COVID-19). A few drug interaction ontologies exist\(^\text{10}\). A knowledge base with virus-drug interactions exist\(^\text{11}\). Having the RDF triples derived from text and this knowledge base would allow us to move toward a drug-virus ontology.

REFERENCES


---

\(^{10}\)https://bioportal.bioontology.org

\(^{11}\)https://go.drugbank.com


