Final Report: ”Accelerated Discovery: Microbe Test Case”

Maya Ramanath[*]
IIT-Delhi

1 Introduction

In this paper, we report on our efforts to construct a knowledge-base of interesting information, extracted from a corpus of bio-chemical research papers. A knowledge-base (or KB) consists of *triples* of the form ⟨`s p o`⟩ where `s` and `o` are entities (subject and object, respectively) and `p` is a relationship connecting the two (also known as the predicate). Table 1 shows several examples of the kinds entities and triples we would like to extract.

Subject	**Predicate**	**Object**
`pseudomonas`	`found_in`	`sea`
`pseudomonas`	`acts_on`	`acetone`
`hydrocarbon_degradation`	`carried_on`	`acetone`
`pseudomonas`	`produce`	`carboxylase`
`hydrocarbon_degradation`	`occurs_in`	`sea`

Table 1: Example triples extracted from a corpus of bio-chemical research papers

Such a KB can help in a number of different kinds of applications, including search and logical inferencing. For example, suppose we wanted to find all microbes found in a specific environment, such as `sand`. We could simply query the KB using a structured query language (mainly, SPARQL is used to query KBs represented in the RDF format as shown in Table 1) to retrieve all such microbes.

1.1 KB construction

The input available to us were: i) a set of 100 research papers from the bio-chemical domain provided by the Exxon-Mobil SMEs, and, ii) a set of terms, such as names of microbes, substrates, etc. which were considered to be of interest. We initially started with the aim of extracting and organizing information from the research paper corpus in to a KB. Our approach was to, first, construct a dictionary of interesting entities. Second, using this list of entities, to extract relationships among them from the corpus. We briefly describe the steps here and elaborate further in the rest of this paper.

1.1.1 Document Pre-processing

The pdf documents provided to us consists the unstructured (textual) content from which we extract triples and construct our KB. In order to this, we first converted the pdf documents into text. We ignored images, tables, etc. and focused only the textual content.

1.1.2 Constructing a dictionary of entities

We had available to us, the set of terms deemed interesting by the Exxon-Mobil SMEs. We classified these terms into types, such as `microbe`, `substrate`, `environment`, etc. When annotating the documents with these entities, we found that this list was not comprehensive. We describe in Section 2 our attempts to create a more exhaustive dictionary using the corpus available to us.

[*] Joint work with students Shubham Singla and Manas Joshi

1.1.3 Extracting relationships

Once the dictionary of entities was created, we then annotated the entire corpus with the entities and implemented several different approaches to extract the relationships between them. We describe these methods in Section 3.

2 Entity extraction

2.1 Types of interesting entities

We are interested in the following types of entities:

- **Microbes**: Microbes or microorganisms are the small life forms that can not be seen by naked eye surrounding us. For example, bacteria like Cyanobacteria, Escherichia coli, Fungi like Penicillium, Rhizopus stolonifer (a bread mold), etc.
- **Substrates**: A substrate is a chemical substance on which a reaction or a process is carried on. For example Benzene, Ethanol, gasoline etc.
- **Environment**: Environment here refers to the surroundings or conditions in which a process takes place or a microbe lives in. For example groundwater, hydrocarbon-contaminated soil etc.
- **Reaction/Process**: It refers to a process in which a substance undergoes a physical or chemical change. For example - fermentation, oxidation, cracking(of oil).
- **Enzyme**: Enzymes are biological catalysts. They accelerate a chemical reaction. For example - monooxygenase, carboxylase etc.
- **Nutrient**: Nutrients are substances used by organisms to maintain life and growth. For example - iron, NaCl etc.
- **Property**: These are used to describe a microbe. For example - anaerobic, rod-shaped etc.

The list we received from Exxon-Mobile SMEs consisted of four different kinds of entities. There were 23 entities which represented ”property” (abiotic, contaminated, prokaryotic), 17 represented surrounding conditions or ”environment” (wastewater, beach, soil), 10 represented a ”chemical reaction” (degradation, denaturing, fertilization), 6 represented a ”substance” (diesel, oil, hydrocarbon). We used these entities as given to annotate our document set. However, on closer examination of these annotations, we noticed that fine-grained entities were not being annotated (indeed, they were not in our expert-prepared entity set at all). For example, on annotating the entity `soil`, we found that the word ”soil” usually came with a qualifier, and in fact, there were different kinds of soils. Examples include, `hydrocarbon-free_soil`, `methanogenic_soil`. Similarly for annotated entity `degradation`, we wanted to extract kind of degradation, such as, `aerobic_degradation`, `benzene_degradation`, etc. Fine-grained entity annotation leads to more meaningful relation extraction, as we will show later. Even without the goal of relation extraction, having a dictionary of precise and fine-grained entities helps in query formulation.

2.2 Automatic entity extraction

Our key observation was that, the entities that we annotated in our documents typically came with adjectives that further described the entities (such as ”anoxic” groundwater, ”alluvial” sand). Therefore, our idea was to *automatically* extract these adjectives and acquire fine-grained entities. We proceeded as follows.

We used Stanford CoreNLP[1] to assign a part of speech tag to each word in the corpus. We then formulated the rules to extract the fine-grained entities[2]. These rules are enumerated in Table 2.2. With these rules, we were able to extract around 2015 entities. We sampled 200 entities and calculated the accuracy (correctness) of these entities. We achieve and accuracy of 81%.

[1] https://stanfordnlp.github.io/CoreNLP/index.html

[2] Note that we use the terminology of Penn Treebank. For example, ”JJ” to indicate adjectives, ”NN” to indicate noun phrases, etc.

#	Rule	Example
1	E' = JJ JJ . . . E	aromatic(JJ) hydrocarbon-contaminated(JJ) sediments(E)
2	E' = NN S E	alkyl(NN) substituted(VBN) aromatic(JJ) XXX(E)
3	E' = NN JJ E	microbial(NN) hydrocarbon(JJ) biodegradation(E)
4	E' = NN VB E	hydrocarbon(NN) contaminated(VBN) antarctic(E)
5	E' = E(JJ) . . . NN	aromatic(E) hydrocarbon-degrading(NN) bacteria(NN)
6	E' = JJ CC JJ E	aliphatic(JJ) and(CC) aromatic(JJ) hydrocarbons(E)

Table 2: Rules to extract fine-grained entities. E' refers to the new entity formed from the old, seed entity E. JJ is an adjective, NN refers to a noun and should be understood as representing all its variations (NNS, NNP, etc.), VB is a verb and should be understood as representing all its variations (VBG, VBN, etc.), CC is a conjunction

2.3 Entity acquisition from web-resources

Despite finding fine-grained entities, our final entity set was still restricted by our initial entity set, that Exxon-Mobil's SMEs provided us with. But for several of the entity types, we observed that there are a number of web-resources that can provide us with additional entities. Therefore, we scraped these websites for lists of entities. The list is as follows.

- BRENDA (http://www.brenda-enzymes.info) - A set of 7974 enzymes
- CyanoBase (http://genome.microbedb.jp/cyanobase/) - A set of 376 microorganisms
- MicrobesOnline (http://www.microbesonline.org/) - A set of 3707 genomes
- metaMicrobesOnline (http://meta.microbesonline.org/) - A set of 1429 metagenomes
- PubMLST (https://pubmlst.org/) - A set of 99 bacteria
- ChEBI (http://www.ebi.ac.uk/chebi/) - A set of 395852 compounds
- GenomeNet(http://www.genome.jp/) - A set of 12034 compounds and 10827 reactions

The entities obtained from these resources were matched with our corpus and entities having a significant term frequency were put into our database.

With this entity dictionary in place, we then annotated our corpus with these entities and extracted relationships among them. We describe the techniques we used to extract these relations in the next section.

3 Relation extraction

We were interested in the following kinds of relationships:

- **FoundIn** - This type of relation exists between a microbe and environment – that is, which environment is a microbe found in. For example: Pseudomonas is found in sea.
- **Of** - This type of relation exists between a microbe and its property. It tells us about a particular attribute of the microbe. eg, heterotrophic bacterial population of culturable gasoline degraders, here heterotrophic bacterial population is being treated as entity type microbe which are culturable gasoline degraders and it is their property.
- **ActOn** - This type of relation exists between microbes and substrates. It tells the type of substances on which a microbe acts on and results in a chemical reaction. eg, Pseudomonas acts on acetone, here Pseudomonas is a microbe acting on acetone which is a substrate.
- **CarriedOn** - This type of relation exists between microbe and reaction/process. It states the type of reaction in which a microbe participates. This type of relation also exists between a process/reaction and substrate. It identifies the type of reaction carried on a substrate to produce end product. eg, hydrocarbon degradation carried on acetone by Pseudomonas, here reaction of hydrocarbon degradation is carried on acetone which is a substrate by a microbe, Pseudomonas.

Relation	Logistic Regression			Random Forests			Neural Networks		
	P	R	F-score	P	R	F-score	P	R	F-score
FoundIn	0.311	0.164	0.215	0.892	0.193	0.317	0.839	0.152	0.257
Of	0.392	0.18	0.247	0.629	0.351	0.45	0.612	0.712	0.658
ActsOn	0.205	0.865	0.331	0.609	0.736	0.66	0.559	0.865	0.679
CarriedOn	0.469	0.949	0.628	0.813	0.622	0.70	0.765	0.827	0.795
OccursIn	0.403	0.92	0.561	0.855	0.908	0.88	.851	0.958	0.902

Table 3: Main results for relation extraction. **P** is precision, **R** is recall

- **Produce** - This type of relation exists between microbe and enzyme. It identifies the type of enzyme being produced by a microbe after a chemical reaction. eg, Pseudomonas produces carboxylase, here Pseudomonas which is a microbe produces carboxylase, an enzyme.
- **OccursIn** - This type of relation exists between reaction/process and environment. It tells us about the possible places where a reaction can take place. Indirectly, it tells us about the environment in which a microbe involved in that reaction can be found. eg, Hydrocarbon degradation occurs in sea, here reaction of degradation of hydrocarbons takes place in environment which is found near to sea.
- **Inhibit_Catalyze** - This type of relation exists between enzymes and process/reaction. The role of an enzyme is to help in either inhibition or catalysis of a reaction or a process which is predicted by this relation. eg, Oxygenase helps in inhibition of photosynthesis, here oxygenase is an enzyme which inhibits the process of photosynthesis. Unfortunately, we were unable to acquire a sufficient number of examples.

3.1 Methodology for relationship extraction

3.1.1 Creating Data Set

Creation of Gold Standard

We asked students from the Biochemical Engineering department to help us with annotations of both entities and relationships. Three students spent several hours to provide annotations on approximately 40 pdf files provided by the Exxon-Mobil SMEs and annotated a total of around 3000 relations. We used the 'brat' software to enable easy annotation[3].

Training and testing

We used 70% of the annotated relations as our training set and the remaining 30% as the test set. Both positive and negative examples were part of the training and 2-fold cross validation was performed.

3.1.2 Extraction algorithms

We used the following feature set for each algorithm we experimented with:

1. All the words in between as feature
2. All the words in between as feature along with the type of entity
3. All the words in between as feature along with the entity type and number of words in between

We experimented with logistic regression, random forests and neural networks. The results are summarized in Table 3.

[3] http://brat.nlplab.org

4 Summary and Future Work

In this report, we have briefly described our attempts to construct a KB consisting of interesting entities and relationships among these entities in the bio-chemical engineering domain. We did this in two parts. First, we constructed dictionary of fine-grained entities, and second, we trained several models to make use of these entities and other features to learn how to extract interesting relations. We were able to achieve a reasonable precision and recall for most of the relationships.

As future work, there is still much to be done. First, we need a much larger training set and expert annotators to ensure that the annotations are correct. Second, we need experts to make a list of interesting entities and relationships that is more comprehensive than the one we have now. Finally, we need to use these models on a larger scale to acquire a bigger set of entities and relations.

A second major direction of work is in making use of the facts in the knowledge base to infer unknown facts. For example, one of the goals would be to figure out if a certain microbe is aerobic or anaerobic. Given a set of logical rules to determine whether a microbe is aerobic (or not), we can perform logical inferencing on the collected set of facts.