Using Bootstrapping to Extract Information from Scientific Literature to Populate a Natural Products Ontology

Patrick Ivan S. Altea, Dan Eduard E. Dagdag, Eric Jerome G. Embestro, Nixon Power S. Ong, Nathalie Rose Lim-Cheng*
De La Salle University-Manila
2401 Taft Avenue, 1004 Manila, Philippines
**Corresponding Author: nathalie.lim@dlsu.edu.ph*

Abstract: Medicinal plants play a vital role in developing drugs and treating illnesses to preserve the health and well-being of the people. Scientists all over the world have been conducting studies and research on natural products to discover and seek for new medications. To aid the researchers in conducting studies, there are existing online natural products databases to browse or search for chemical information and properties. However, these existing online natural products databases lack data about the natural products that are exclusively found in the Philippines; the inclusion of these data into existing databases cannot be done directly into the database. Creating a database that includes natural products in the Philippines could help boost and produce more research on organic medicines in the country. This paper discusses the development of an ontology population system that extracts data from scientific publications containing information about natural products collected from the Philippines. It focuses on identifying natural products' corresponding active chemical compounds that are said to be effective for treating certain medical conditions. In this paper, we present our findings on why, from a high of 72.63% precision (among other metrics), the results dropped to 31.34% after feeding more data into the automatic extraction system using bootstrapping.

Key Words: Ontology population; natural products; information extraction

1. INTRODUCTION

Health and well-being became a global trend in the world in the recent decade. There is a rise in the demand for the Philippine's natural products since there was an increased amount of exportation of these products, like malunggay (Moringa oleifera) and lagundi (Vitex negundo). Our own Department of Science and Technology, through PCHRD, funds research on discovery of health products from natural products. It was found that Moringa oleifera is not just beneficial for lactation and colorectal cancer. Recent researches are also exploring the potential of this natural product against SARS-CoV-2 (COVID-19).

Scientists identify chemical compounds found in natural products, extract these for testing to determine its efficacy towards curtailing the spread of or even curing certain diseases. Chemical compounds found in one natural product may also exist in others. Thus, being able to find alternative sources would be beneficial for the manufacturing of medicines. The search, identification, and analysis can be facilitated via repositories of information on these natural products to allow scientists to determine what compounds does the natural product have, which compounds have been studied previously for which particular medical condition, and so on.

There exist natural product databases such as Super Natural II (Banerjee et al., 2014),and YaTCM (Li et al., 2018). Both are open access, web-based databases. Super Natural II contains information of over 300,000 molecules with its corresponding structures in 2D and in 3D along with its toxicity class. This version can also search for similar compounds by allowing template-based search and provides information on the pathways that are associated with synthesis degradation of the natural products and their mechanism of action. YaTCM, on the other hand, contains over 47,000 natural compounds. Information includes prescription, herbs, ingredients, definite or putative protein targets, pathways, and diseases. The application provides corresponding common names and images of plant-based products and specific plant parts used for treating certain medical conditions. However, this natural products database application only contains natural products that are found in China.

Though both are open access, these do not allow the user to add or modify entries in its database. Since most natural products from the Philippines are not included in these databases, scientists could not perform research to include/compare Philippine natural products with those from other countries.

In this paper, we present our database that can contain information about Philippine natural products. The database is designed as an ontology which is a set of concepts on a particular subject area that shows their properties and relations. Maynard et al. (2008) also stated that ontologies are beneficial for healthcare and bio-medicine due to the fact that such concepts evolve and update rapidly.

Since, there are new research on natural products being published regularly. It is tedious to manually read through the scientific publications to manually populate the ontology. To facilitate this process of populating the ontology, we used natural language processing (NLP) techniques, specifically using bootstrapping. Bootstrapping is a technique to extract information from unstructured texts of a subject domain. It utilizes the relationships and patterns between two entities found in the document. The objective of bootstrapping is to obtain more patterns as more documents are processed to extract other possible entities concerning the natural products domain.

Section 2 discusses our ontology to model the natural products. Section 3 presents the modules involved in the extraction process using bootstrapping. Section 4 shows the performance of the automatic extraction. Lastly, in Section 5, we present our conclusion and future work.

2. OUR PHILIPPINE NATURAL PRODUCTS ONTOLOGY

The designed ontology of the natural products was extended from the existing ontology of Philippine medicinal plants by Lim-Cheng et al. (2014) to prevent the reduplication of curating similar or intersecting information which had already been collected and validated. Figure 1 shows the augmented medicinal plants ontology that include chemical compounds found in parts of these natural products. From four (4) entities in medicinal plant ontology, there are now fourteen (14) entities or concepts in our ontology, namely: MedicinalPlant, Species, SpeciesPlantPart, Location, Genus, Family, PlantPart, Preparation, BodyPart, Illness, Compound, CompoundClass, BiologicalActivity, and CellLine.

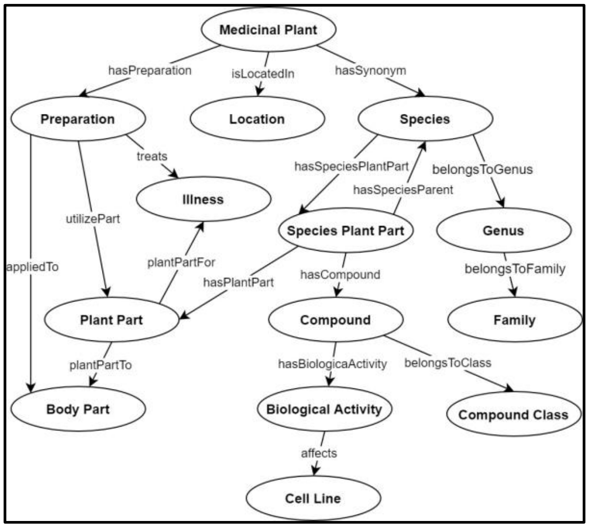


Fig. 1. Natural Products Ontology

From six (6) relationships in the original plant ontology, we now have fifteen (15) relationships. These are listed below in the format of <relationship>(<entity1>, <entity2>)

(1) hasSynonym(MedicinalPlant, Species)
(2) isLocatedIn(MedicinalPlant, Location)
(3) hasPreparation(MedicinalPlant, Preparation)
(4) belongsToGenus(Species, Genus)
(5) hasSpeciesPlantPart(Species, SpeciesPlantPart)
(6) hasSpeciesParent(SpeciesPlantPart, Species)
(7) hasPlantPart(SpeciesPlantPart, PlantPart)
(8) hasCompound(SpeciesPlantPart, Compound)
(9) belongsToFamily(Genus, Family)
(10) treats(Preparation, Illness)
(11) appliedTo(Preparation, BodyPart)

(12) utilizePart(Preparation, PlantPart)
(13) belongsToClass(Compund, CompoundClass)
(14) hasBiologicalActivity(Compound, BiologicalActivity)
(15) hasCellLine(BiologicalActivity, CellLine)

Details of the entities and the relationships can be seen in (Altea, et.al, 2020). Figure 2 further illustrates the ontology design, via display of the entries stored in the corresponding concepts of the plant Batino (scientific name "Alstonia macrophylla").

This ontology design was validated by faculty members from the Chemistry Department and Biology Department of De La Salle University. They are doing research either on extracting chemical compounds of natural products and determining their efficacy on medical conditions or on using available data about properties of natural products to perform analysis and predictions on their possible applicability to other medical conditions.

3. SYSTEM DESIGN

The system accepts the pdf document, converts it to text only as part of the Preprocessing Module, then feeds these together with seed tuples and seed patterns to the Bootstrapping Module to produce more tuples and patterns. The tuples eventually are presented in the User Interface for validation by the expert. Once validated, these entries are then included into the ontology. Refer to Fig. 3 for the architectural diagram of the system.

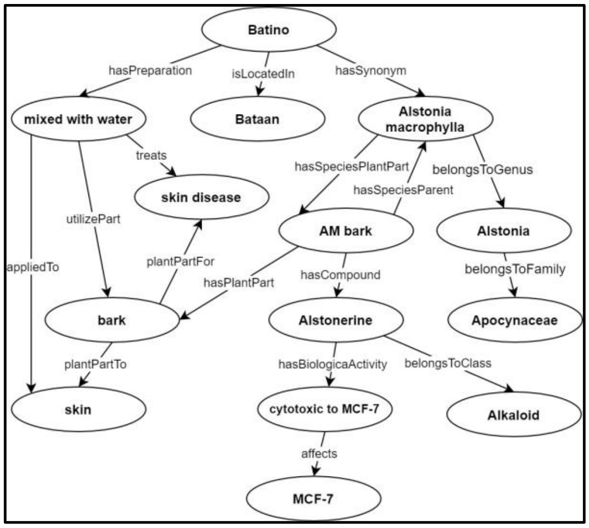


Fig. 2. Sample Natural Products Ontology

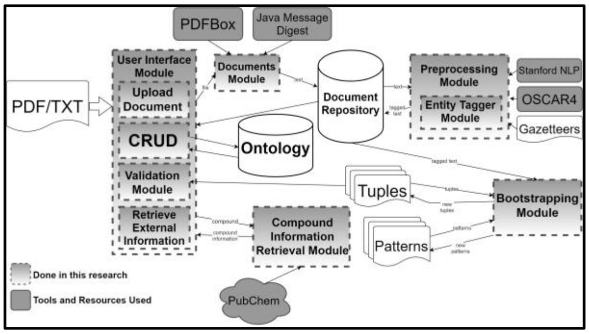
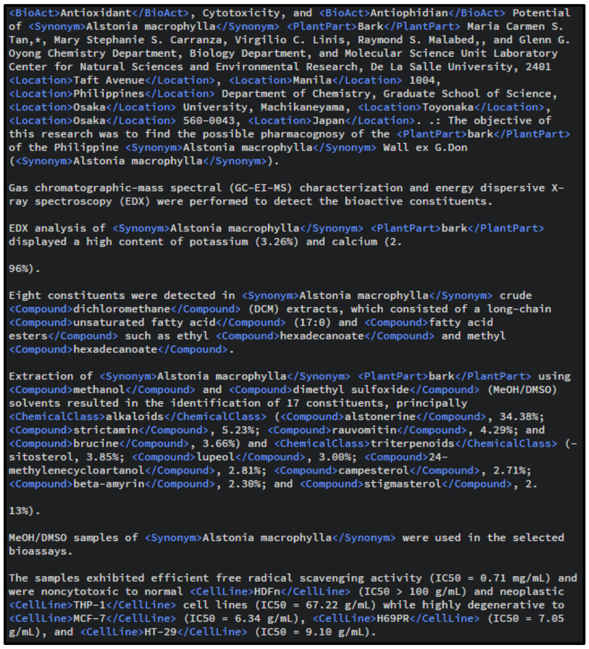


Fig. 3. System Architecture



```
<BioAct>Antioxidant</BioAct>, Cytotoxicity, and <BioAct>Antiophidian</BioAct> Potential
of <Synonym>Alstonia macrophylla</Synonym> <PlantPart>Bark</PlantPart> Maria Carmen S.
Tan,*, Mary Stephanie S. Carranza, Virgilio C. Linis, Raymond S. Malabed,, and Glenn G.
Oyong Chemistry Department, Biology Department, and Molecular Science Unit Laboratory
Center for Natural Sciences and Environmental Research, De La Salle University, 2401
<Location>Taft Avenue</Location>, <Location>Manila</Location> 1004,
<Location>Philippines</Location> Department of Chemistry, Graduate School of Science,
<Location>Osaka</Location> University, Machikaneyama, <Location>Toyonaka</Location>,
<Location>Osaka</Location> 560-0043, <Location>Japan</Location>. .: The objective of
this research was to find the possible pharmacognosy of the <PlantPart>bark</PlantPart>
of the Philippine <Synonym>Alstonia macrophylla</Synonym> Wall ex G.Don
(<Synonym>Alstonia macrophylla</Synonym>).

Gas chromatographic-mass spectral (GC-EI-MS) characterization and energy dispersive X-
ray spectroscopy (EDX) were performed to detect the bioactive constituents.

EDX analysis of <Synonym>Alstonia macrophylla</Synonym> <PlantPart>bark</PlantPart>
displayed a high content of potassium (3.26%) and calcium (2.

96%).

Eight constituents were detected in <Synonym>Alstonia macrophylla</Synonym> crude
<Compound>dichloromethane</Compound> (DCM) extracts, which consisted of a long-chain
<Compound>unsaturated fatty acid</Compound> (17:0) and <Compound>fatty acid
esters</Compound> such as ethyl <Compound>hexadecanoate</Compound> and methyl
<Compound>hexadecanoate</Compound>.

Extraction of <Synonym>Alstonia macrophylla</Synonym> <PlantPart>bark</PlantPart> using
<Compound>methanol</Compound> and <Compound>dimethyl sulfoxide</Compound> (MeOH/DMSO)
solvents resulted in the identification of 17 constituents, principally
<ChemicalClass>alkaloids</ChemicalClass> (<Compound>alstonerine</Compound>, 34.38%;
<Compound>strictamin</Compound>, 5.23%; <Compound>rauvomitin</Compound>, 4.29%; and
<Compound>brucine</Compound>, 3.66%) and <ChemicalClass>triterpenoids</ChemicalClass> (-
sitosterol, 3.85%; <Compound>lupeol</Compound>, 3.00%; <Compound>24-
methylenecycloartanol</Compound>, 2.81%; <Compound>campesterol</Compound>, 2.71%;
<Compound>beta-amyrin</Compound>, 2.30%; and <Compound>stigmasterol</Compound>, 2.

13%).

MeOH/DMSO samples of <Synonym>Alstonia macrophylla</Synonym> were used in the selected
bioassays.

The samples exhibited efficient free radical scavenging activity (IC50 = 0.71 mg/mL) and
were noncytotoxic to normal <CellLine>HDFn</CellLine> (IC50 > 100 g/mL) and neoplastic
<CellLine>THP-1</CellLine> cell lines (IC50 = 67.22 g/mL) while highly degenerative to
<CellLine>MCF-7</CellLine> (IC50 = 6.34 g/mL), <CellLine>H69PR</CellLine> (IC50 = 7.05
g/mL), and <CellLine>HT-29</CellLine> (IC50 = 9.10 g/mL).
```

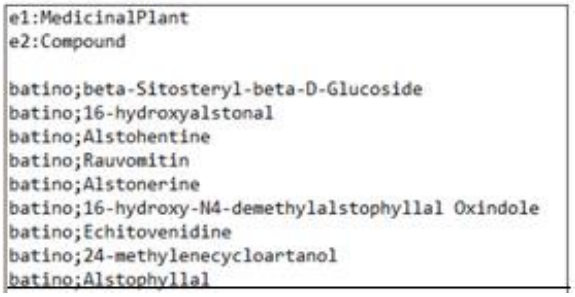
Fig. 4. Sample Output of the Preprocessing Module

3.1 Preprocessing Module

Each article to be used for extraction is first converted to a text file; Doing so removes the images, figures, and tables from the document since information is only extracted from the texts written by the authors. The converter used is from the reusable library of Apache' PDFBox Java API. Once the articles have been converted to a text file, the papers go into the cleaning phase where unnecessary characters in Unicode and new lines are removed. All Greek symbols are also converted into their English word counterparts. Whitespaces or indentations are also reconfigured to remove sentence or line breaks. Using the cleaned text, all sentences are now split per line using RegEx. Each sentence is separated by having a new line ("\n"). After the entire article has been cleaned and prepared, all information is now tagged by their corresponding classification. All entities to be tagged are based on the classes from the designed natural products ontology. Entities were tagged through the use of lexicons and gazetteers by looking up an entity from the lists. Techniques such as name resolution, name expansion, and coreferencing were also implemented to obtain accurate information from the papers. The sample output of this module is shown in Figure 4.

3.2 Bootstrapping

Once the uploaded documents have been tagged from the Documents Module, seed patterns are searched from the Tagged Documents with the help of POS tagger, Wordnet, and seed data. The seed data is a set of text files, where each text file contains the initial contents manually identified from scientific papers. Fig. 5 shows an excerpt of the contents of the MedicinalPlant-Compound seed.



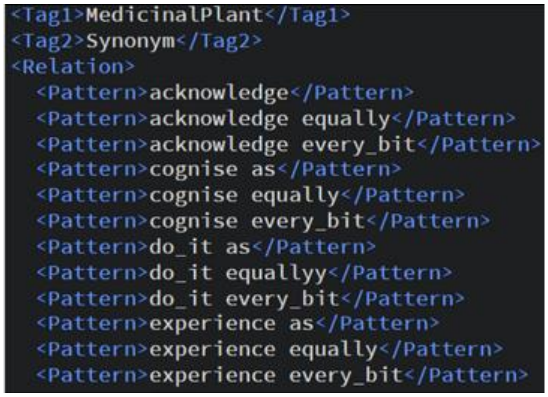
```
e1:MedicinalPlant
e2:Compound

batino;beta-Sitosteryl-beta-D-Glucoside
batino;16-hydroxyalstonal
batino;Alstohentine
batino;Rauvomitin
batino;Alstonerine
batino;16-hydroxy-N4-demethylalstophyllal Oxindole
batino;Echitovenidine
batino;24-methylenecycloartanol
batino;Alstophyllal
```

Fig. 5. Medicinal Plant – Compound Seed Entries

It searches the tagged documents sentence by sentence and with the help of the seed data it stores the sentence to a TreeSet if it contains the seeds inside the Seed Data. After the sentences has been stored, it retrieves the words between the two matching entities. The retrieved words are tagged by the POS tagger to identify the part-of-speech each word had used. Once the tagged documents have been POS tagged, it is trimmed to follow the format of V/VP/VW*P based from Fader et al. (2011).

After the tagged sentence has been trimmed, Wordnet generates all possible seed relations between the found seeds and store it as seed output. Fig. 6 has the excerpt of the sample of generated seed relations.

Once the seed output has been generated, the bootstrapping now uses the generated seed outputs to locate seed relations that do not exist in the seed data by identifying whether the phrases between the seed relations contains the seed pattern. The found seed relations from the tagged documents are saved to a different XML file so that it can be used by the validation page from the User Interface Module.



```
<Tag1>MedicinalPlant</Tag1>
<Tag2>Synonym</Tag2>
<Relation>
  <Pattern>acknowledge</Pattern>
  <Pattern>acknowledge equally</Pattern>
  <Pattern>acknowledge every_bit</Pattern>
  <Pattern>cognise as</Pattern>
  <Pattern>cognise equally</Pattern>
  <Pattern>cognise every_bit</Pattern>
  <Pattern>do_it as</Pattern>
  <Pattern>do_it equallyy</Pattern>
  <Pattern>do_it every_bit</Pattern>
  <Pattern>experience as</Pattern>
  <Pattern>experience equally</Pattern>
  <Pattern>experience every_bit</Pattern>
```

Fig. 6 Generated Seed Relations

4. RESULTS AND FINDINGS

The main goal of the bootstrapping technique is to extract more information without need for manually encoding many seed information. To test how well the bootstrapping technique would perform, we need to run the ontology population system multiple times, adding more documents per iteration and compare the results.

Due to constraints in manual annotating articles, we only had twenty-five (25) scientific articles manually annotated to be used as the reference data. Two of the annotated articles were validated by two of the expert consultants (same ones who gave input on the ontology design). The experts also validated the annotation guidelines (i.e., the documented process on how to identify entities that have to be extracted). This validated guideline was used as basis in manually annotating the rest of the 23 documents (meaning annotations from these were no longer validated). It should be noted that all 25 articles were published by experts on the field of natural products affiliated to De La Salle University.

As there are only 25 annotated articles and we wanted to perform three runs to compare the results, we started with nine (9) articles for the first run and eight (8) additional articles for each succeeding run, thus having 17 articles for the second run, and 25 articles used for the third run.

For the first run, we choose the articles where there were the most manually annotated entities and where these articles are written by different authors and published in different publication venues. Also considered here was that these articles would contain data to manually populate the text files for our seed data. In the succeeding runs, the eight articles (for each run) is chosen randomly from the remaining sixteen (16).

Part of the bootstrapping process was to generate the patterns, as discussed in Section 3.2. The nine documents in the first run were also used to generate initial seed patterns for the seed output.

Table 1. Test Results

Metric	Run 1	Run 2	Run 3
Accuracy	40.77	32.80	25.74
Precision	72.63	42.51	31.34
Recall	48.17	58.96	58.99
F-Measure	57.93	49.40	40.94
No. of Documents	9	17	25

In the second run, the additional eight (8) articles were fed to the ontology population system, using the resulting data (the seed data, seed pattern, and seed output) of the previous run/s. The same process is done for the third run. As can be seen in the results of Table 1, from 72.63% precision and 40.77% accuracy in the first run, it has dropped to 31.24% and 25.74% respectively in the second run.

One of the causes was the inclusion of additional information (background information or for comparison) in the document, not related to the natural product being discussed. An example is where the compounds of Alstonia scholaris were found in the documents of Alstonia macrophylla. Those non-related entity relationships are still automatically added by the bootstrapping process of the system since the bootstrapping process only considers the patterns between the two entities disregarding whether or not the entities are actually the ones being mainly discussed in the paper; which is why the number of false positives for the Alstonia macrophylla paper has obtained a high count decreasing the score for accuracy and precision.

Another cause of the problem is on the coreferencing. An example is on the article on Pleurotus eryngii, where the plant name was not coreferenced, properly when the abbreviated plant name (such as P. Eryngii) was used in the text.

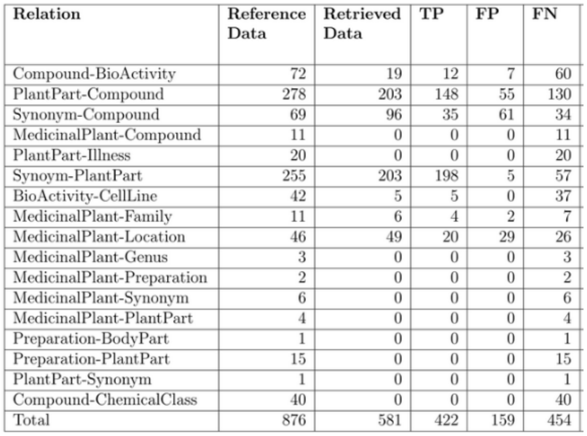
Analyzing the implementation of the bootstrapping, the main concerns are (1) very small number of seed data for some entity relationships, (2) some seed data are not directly translated from the entity relationships indicated in the ontology design, and (3) over generation of seed patterns.

Table 2 lists the results of the first run showing the number of extracted data per entity relation, while Table 3 contains that of the second. TP refers to the true positives which are the number of correctly retrieved entries, FP are the number of incorrectly retrieved entries, and FN are the number of entries that were not retrieved.

Table 2 shows that entity relationships like MedicinalPlant-Synonym and MedicinalPlant-PlantPart have less than ten seed data. Recall that in the first run, the reference data in also the content of the seed data. This small number also means that there are also at most that many number of initial seed patterns for this relationship, which causes other entries (even for succeeding runs where there are more expected data to be found for the said relationships, shown in Table 3) to not be found.

On the concern that seed data do not directly translate from the relationships in the ontology design. However, entities for some relationships do not appear in the same sentence or in the article. For example, Preparation-Illness (i.e., treats relationship) do not appear (in the same sentence) in the articles used in the study. What appears instead are PlantPart-Illness.

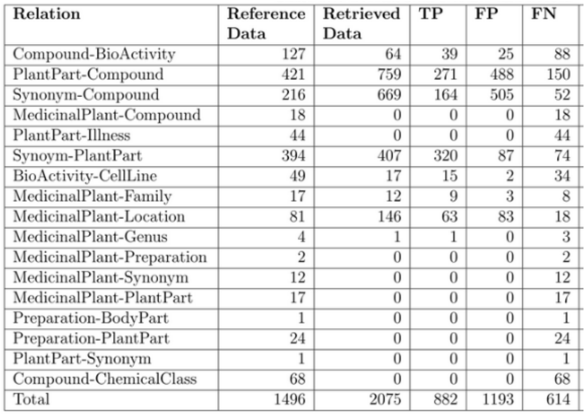
Table 2. First Run Results



Relation	Reference Data	Retrieved Data	TP	FP	FN
Compound-BioActivity	72	19	12	7	60
PlantPart-Compound	278	203	148	55	130
Synonym-Compound	69	96	35	61	34
MedicinalPlant-Compound	11	0	0	0	11
PlantPart-Illness	20	0	0	0	20
Synoym-PlantPart	255	203	198	5	57
BioActivity-CellLine	42	5	5	0	37
MedicinalPlant-Family	11	6	4	2	7
MedicinalPlant-Location	46	49	20	29	26
MedicinalPlant-Genus	3	0	0	0	3
MedicinalPlant-Preparation	2	0	0	0	2
MedicinalPlant-Synonym	6	0	0	0	6
MedicinalPlant-PlantPart	4	0	0	0	4
Preparation-BodyPart	1	0	0	0	1
Preparation-PlantPart	15	0	0	0	15
PlantPart-Synonym	1	0	0	0	1
Compound-ChemicalClass	40	0	0	0	40
Total	876	581	422	159	454

Lastly, on the over generation, from Figure 6 we see that from an original pattern “is_known_as” as in <MedicinalPlant> “is_known_as” <Synonym> that appeared in a sentence in an article, the seed pattern would over generate to include “acknowledge” and “cognize” as synonyms of “known”, and “equally” as synonym of “as”. This also does not stop there, since synonyms of “acknowledge” and “cognize” are also retrieved, which resulted to “do_it” and “experience”, among others. Then, combinations of the synonyms are re-formed (producing, for example, "experience equally") to be part of the seed patterns.

Table 3. Second Run Results



Relation	Reference Data	Retrieved Data	TP	FP	FN
Compound-BioActivity	127	64	39	25	88
PlantPart-Compound	421	759	271	488	150
Synonym-Compound	216	669	164	505	52
MedicinalPlant-Compound	18	0	0	0	18
PlantPart-Illness	44	0	0	0	44
Synoym-PlantPart	394	407	320	87	74
BioActivity-CellLine	49	17	15	2	34
MedicinalPlant-Family	17	12	9	3	8
MedicinalPlant-Location	81	146	63	83	18
MedicinalPlant-Genus	4	1	1	0	3
MedicinalPlant-Preparation	2	0	0	0	2
MedicinalPlant-Synonym	12	0	0	0	12
MedicinalPlant-PlantPart	17	0	0	0	17
Preparation-BodyPart	1	0	0	0	1
Preparation-PlantPart	24	0	0	0	24
PlantPart-Synonym	1	0	0	0	1
Compound-ChemicalClass	68	0	0	0	68
Total	1496	2075	882	1193	614

5. CONCLUSIONS

Currently, the ontology population system can extract significant information from the scientific literature through the utilization of a bootstrapping algorithm to find possible patterns as to how entities are going to be extracted from the papers. Although the number of false positives are high from the extraction process, the reasonable amount of true positives justify that almost all of the information is obtained to populate the natural products ontology. Additionally, all entries extracted will first be validated before populating the ontology to ensure that the data on the ontology are all valid and correct. Although the extraction procedure is not perfect, the bootstrapping algorithm proves to be a notable approach in dealing with information extraction on unstructured texts such as scientific literature, especially since the number of papers on natural products published all over the world are increasing.

It should be noted too that the tests here involve only twenty-five (25) articles. From the findings, we see several improvement opportunities from the preprocessing to the bootstrapping process. In the preprocessing, for example, we need to improve the coreferencing. In the bootstrapping process, we see that using WordNet to retrieve more synonyms to generate more patterns produce too many patterns that are not useful or relevant to the domain. We are looking into limiting the amount of synonyms that will appear as patterns for the bootstrapping process.

Once we have improved on these, among others, we will test on more articles. Validation of these extracted results would be key. Once we have collected a sufficient number of validated data, we plan to also explore machine learning as an approach and compare results of using bootstrapping with machine learning, not just based on the metrics of precision and recall, but also on the amount of minimum data requirement for these approaches to produce acceptable results. To date, we have already improved on the front-end search features developed from the API from those published in (Altea, et. Al, 2020) to facilitate the validation process.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Angelyn Lao, Dr. Raymond Malabed, Mr. Vincent Antonio Ng, Dr. Mariquit De Los Reyes, and Dr. Maria Carmen Tan for their evaluation and feedback on the ontology design.

7. REFERENCES

Altea, P. I., Dagdag, D. E., Embestro, E. J., Ong, N. P. & Lim-Cheng, N. R. (2020). An Ontology of Philippine Natural Products and the API to Retrieve Data from the Ontology. In DLSU Research Congress 2020, De La Salle University, Manila, Philippines.

Banerjee, P., Erehman, J., Gohlke, B., Wilhelm, T., Preissner, R., & Dunkel, M. (2014, 10). Super natural ii–a database of natural products. Nucleic acids research. doi: 10.1093/nar/gku886

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 1535{1545).

Li, B., Ma, C., Zhao, X., Hu, Z., Du, T., Xu, X., Lin, J. (2018, 11). Yatcm: Yet another traditional chinese medicine database for drug discovery. Computational and Structural Biotechnology Journal, 16. doi: 10.1016/j.csbj.2018.11.002

Lim-Cheng, N. R., Co, J. R., Gaudiel, C. H. S., Umadac, D. F., & Victor, N. L. (2014). Semi-Automatic Population of Ontology of Philippine Medicinal Plants from On-line Text. In DLSU Research Congress, De La Salle University, Manila, Philippines.

Maynard, D., Li, Y., & Peters, W. (2008). NLP techniques for term extraction and ontology population