Presented at the DLSU Research Congress 2014
De La Salle University, Manila, Philippines
March 6-8, 2014

Semi-Automatic Population of Ontology of Philippine Medicinal Plants from On-line Text

Nathalie Rose Lim-Cheng, Junn Richmond C. Co, Christa Hannah S. Gaudiel, Darah F. Umadac, and Nadine L. Victor
De La Salle University-Manila
** Nathalie Rose Lim-Cheng: nats.lim@delasalle.ph*

Abstract: Ontologies are formal representations of knowledge organized as concepts of a domain with their relationships defined. These are used in various fields including biomedical informatics. However, to create a functional ontology, it must be populated with instances of the defined concepts. Since it is a tedious task with constantly updating data, especially in the field of healthcare, information extraction (IE) can be employed to fulfill this task semi-automatically. In this paper, semi-automatic ontology population through information extraction from online articles will be discussed. The design of the ontology is based on available information that can be consistently extracted from the available on-line text. These information focus on the medicinal properties of the plant (i.e., what illness can it be applied to, which body part does it affect/cure, the preparation instruction, and the plant part to be used). Lastly, we present some results from initial usability assessment and from comparison with Gold standard. The test shows we can get an 85.71% accuracy.

Key Words: Ontology population, medicinal plants, information extraction

1. INTRODUCTION

In 2009, the World Health Organization (WHO) estimated that 80% of the world population use herbal medicines as part of their primary health care. In the United States, the use of herbal medicines increased in the last 20 years due to the expensive cost of prescription medications. Similarly, the use of herbal medicine in Southeast Asian countries such as Thailand, Malaysia and Philippines is also prevalent. In the Philippines, medicinal plants are used especially by those who have limited access to pharmaceutical medicines and those who cannot afford expensive medications.

Due to the acceptance of medicinal plant use, Thailand and Malaysia both developed ontologies used for health care systems. Unfortunately, in the Philippines, it has yet to be instigated. Since there are several online sources on Philippine medicinal plants, they can be used as sources of textual

information to populate a medicinal plant ontology.

Ontologies are formal representations of knowledge organized as concepts of a domain with their relationships defined. These are used in various fields including biomedical informatics. However, to create a functional ontology, it must be populated with instances of the defined concepts. Since it is a tedious task with constantly updating data, especially in the field of healthcare, information extraction (IE) can be employed to fulfill this task semi-automatically.

This paper presents part of our research on populating an ontology of Philippine medicinal plants, with the intention of harnessing this information for a recommendation system and question-answering system, among others, to provide information on what available alternative remedies can be utilized by the local community. The paper is organized as follows: Section 2 presents existing ontology population systems. Section 3 discusses the Philippine medicinal plant ontology and the extraction process. Section 4 presents the usability test on accessing the contents of the ontology and validation test by comparing the system result with that of the Gold standard. Lastly, we present our conclusion in Section 5.

2. EXISTING ONTOLOGY POPULATION SYSTEMS

OntoLearn is an ontology population system that uses text mining and machine learning techniques (Navigli, et al, 2004). It starts with an existing generic ontology (like WordNet) and a set of documents in a given domain. It then produces a domain extended and trimmed version of the initial ontology. Another ontology population system, Ontosophie, semi-automatically populates ontology with instances from unstructured text. This system is based on supervised learning, where it learns extraction rules from annotated text and applies those rules on new articles to populate the ontology (Celjuska and Vargas-Vera, 2004).

OntoLearn learns concepts by undergoing three phases, namely, (1) Terminology Extraction, (2) Semantic interpretation of multi-word expressions (MWEs), and (3) Extending and trimming the initial ontology. In terminology extraction, a list of domain MWEs is extracted from a set of documents using natural language processing and statistical techniques. In semantic interpretation of MWEs, the semantic interpretation is based on a principle, compositional interpretation, and on a novel algorithm called structural semantic interconnections (SSI). Compositional interpretation signifies that the meaning of a multi-word expression can be derived compositionally from its components. Lastly, in extending and trimming of the initial ontology, the terms are organized in a sub-tree and appended under the appropriate node of the initial ontology. This is done after the terms have been semantically interpreted (Navigli, et al, 2004).

Ontosophie undergoes three phases to populate an existing ontology as well, namely (1) Annotation, (2) Learning, (3) Extraction and Ontology Population. In the first phase, a set of plain text or HTML documents are annotated with XML tags and assigned to one of the predefined classes in the ontology. After document annotation, the learning phase may begin. In the learning phase, Ontosophie uses shallow parsing to recognize syntactic constructs without generating a complete parse tree for each sentence using the Marmot NLP system. After preprocessing with Marmot, the documents are then used to generate extraction rules using Crystal. Crystal is a conceptual dictionary induction system, which derives extraction rules from a training corpus. It also provides coverage and error values to provide a feel of the generated rules' precision or confidence. Lastly, in the population and extraction phase, the system extracts appropriate entities from an article and feeds these entities to the ontology. The document is preprocessed using the Marmot NLP system prior to the extraction (Celjuska and Vargas-Vera, 2004).

3. PHILIPPINE MEDICINAL PLANT ONTOLOGY

3.1 Ontology Description

Six concepts have been defined, all of which are represented as classes in OWL, as well as the relations among the classes referred to as properties, relevant to the medicinal plants domain. In the

ontology, a medicinal plant is associated with an illness, a body part, a preparation, a plant part and a place. Figure 1 illustrates the medicinal plants' information and their relationships in the ontology.

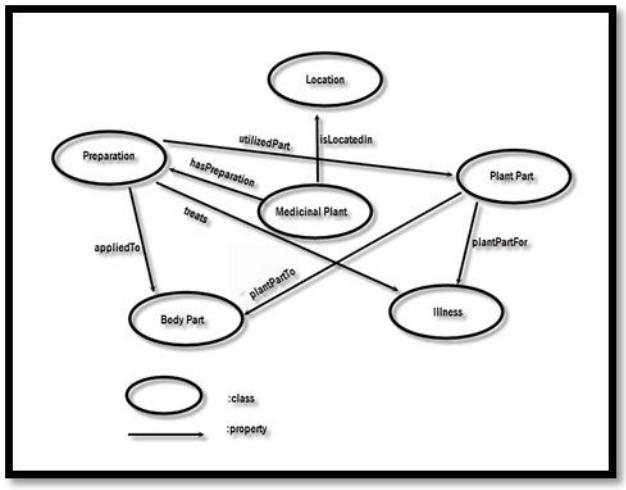

Fig. 1. Medicinal plants ontology

Below are the partially defined classes of the ontology:

1. The MedicinalPlant class represents a set of medicinal plant instances in the ontology.

2. The Illness class represents a set of illnesses treated by the medicinal plants.

3. The BodyPart class represents the set of body parts affected by the medicinal plants and illnesses.

4. The Location class represents a set of provinces and countries where the medicinal plants are cultivated / available.

5. The PlantPart class represents the plant parts utilized in the medicinal plants.

6. The Preparation class represents the preparation for the medicinal plant given an illness.

Below are the properties of the MedicinalPlant class, the Preparation class, and the PlantPart class:

1. MedicinalPlant class

(a) The hasPreparation property defines what preparations are to be done with the medicinal plant.

(b) The isLocatedIn property defines where a medicinal plant can be found.

2. Preparation class

(a) The utilizedPart property defines what parts of the medicinal plant can be used to treat illnesses.

(b) The appliedTo property defines where the medicinal plant should be applied to cure a specific illness.

(c) The treats property defines what illnesses can be cured by the medicinal plant.

3. PlantPart class

(a) The plantPartTo property defines where the plant part should be applied to.

(b) The plantPartFor property defines what illness will the plant part be used for.

3.2 Components

To populate the ontology, data needs to be extracted from different medicinal plant articles. To accomplish this, online articles has to go through the back-end and front end processing. The back-end section comprises of Article Retrieval Module, Preprocessing Module, Semantic Annotation Module, Coreference Module, Template and Validation Module, while the front end section comprises of Save Module, Add/Edit Module, and Search Module. Refer to Figure 2 for the architecture of the system. Further discussion on each component is discussed below.

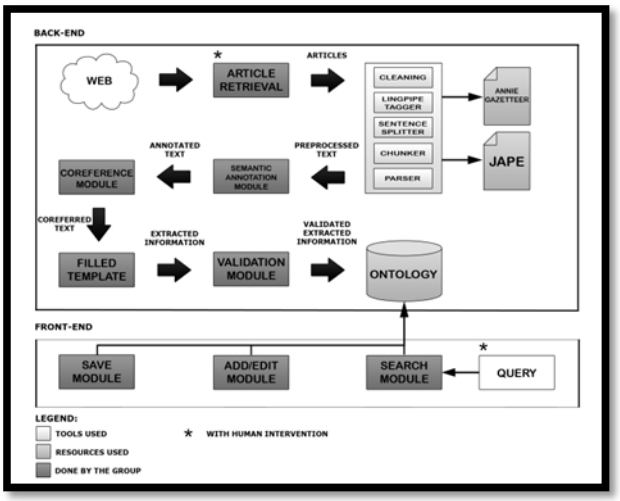

Fig. 2. Ontology Population System Architecture

3.2.1 Back-end Components

1. Article Retrieval Module - this module retrieves the relevant articles from a website (e.g. www.stuartxchange.org). The relevance of the article is determined using keyword searching (e.g., treats). The retrieval of documents is done using the Crawler4J web crawler. As a result, it outputs a text file of the retrieved article.

2. Preprocessing Module - this module processes the text articles from the article retrieval module. It consists of the cleaning, tokenizing, part-of-speech (POS) tagging, sentence splitting, noun phrase chunker, and parser and named entity recognition (NER). The cleaning module removes unnecessary spaces, texts and symbols from the input article. The cleaned article is then passed to the ANNIE English Tokenizer to determine the article contents' token types. Once the token types are identified, the tokens are then tagged with corresponding Penn Tree Bank Part-of-Speech tags (e.g. noun, adjective and verb) using a LingPipe tokenizer. The tokenizer uses a tagger model trained using the GENIA Corpus. These tagged tokens are passed to the noun phrase chunker and the parser, then to the NER, where the ANNIE Gazetteer and JAPE Rules are used to annotate entities according to their entity types (e.g. Location, PlantName).

3. Semantic Annotation Module - this module determines the relationships among the different entities using the proponents' handcrafted JAPE rules. The rules are constructed based on sentence patterns.

4. Coreference Resolution Module - this module uses Anaphora resolution to solve the referencing problems which are normally found in pronouns and nouns.

5. Template - this contains the extracted information from the input article. It contains attributes such as Medicinal Plant Name, Synonyms, Local Name, Location, Contraindication, Preparation, Plant Part (e.g. leaves, stem and roots), Illness, and Body Part. The template is then presented to the user for validation.

6. Validation Module - this module allows the user to validate the contents that are specified in the template. This allows the user to determine which information will be added to the ontology.

3.2.2 Front-end Components

1. Save Module - this module saves the contents shown in the validation screen to allow validation at another time.

2. Add/Edit Module - this module allows the user to manually add or edit the information from the added medicinal plant instance.

3. Search Module - this module allows the user to look for a specific medicinal plant in the ontology. This is done by allowing the user to provide a related keyword to the search field category (e.g. medicinal plant name, illness, body part, effect, preparation, plant part and place).

4. RESULTS

The ontology population system was tested by senior B.S. Biology students from De La Salle University and three professors in terms of usability and validity.

In testing the usability, the students and the experts evaluated the front-end features of the system based on a given set of criteria, with ranking of 5 for Strongly Agree down to 1 for Strongly Disagree. Seen in Table 1 are the results of the usability testing.

Table 1. System Usability Evaluation Results

Criteria	Average Result
User Interface Organization and Ease of Use in terms of Sequence of Steps	4.16
Information Extracted and Stored in Ontology are Informative	4.34
Features in the Front-end System are Easy to use	4.28
Data is presented well	4.24
Data extracted from articles are complete.	4.22
Validation process prior to storing in ontology is effective	4.12

What is notable in the results is that the system got an average rating of 4.34 in terms of the informativeness of the data provided by the system. The data being the information extracted from textual articles and stored as instances in the ontology.

The evaluators recommended improvements to the user interface for better organization and emphasis on the main functionalities, possibly via providing confirmation messages for all user tasks; making the system available over the internet; automatic removal of incorrect extracted information; assuring the credibility of the sources; and adopting the proper format for presenting scientific names. These recommendations were incorporated into the updated system.

In testing the validity of the system, only the three professors serving as experts were asked to do the evaluation. They were asked to annotate 12 medicinal plant articles; 6 from stuartxchange.org and 6 from other websites (i.e., philippineherbalmedicine.org, healthmad.com, herbalmedicineph.blogspot.com, cebu-philippines.net, buzzle.com). It should be noted that manual reviews of articles from stuartxchange.com were done early in the research and the insights learned from this study served as the basis for constructing the extraction rules. However, the articles used for the Gold standard were not used in the early study. Furthermore, articles from other websites were also used in the study to determine if the extraction rules can be applied to other medicinal plant articles.

To ensure consistency, if there are differences in the experts' manual annotations, they were again consulted for clarification. The final consolidated manual annotations now serve as the Gold standard. This is then compared to the information extracted by the system. Tables 2 and 3 show the results of comparing the Gold standard with the output of the system.

Tables 2a and 3a show actual count of the type of information that can be extracted from the articles from stuartxchange.org and from websites, respectively. The column on the Gold standard refers to the count as computed from the annotations of the experts. The Retrieved column is the total number of that information extracted from the articles. TP refers to true positive, meaning the number of correctly extracted information. FP refers to false positive, meaning these are data the system extracted but are incorrect (as determined by the Gold standard). Lastly, FN refers to false negative which means the system was not able to extract the information.

Based on the results, the results of the articles from stuartxchange.org had an average precision of 70.85% and average recall of 61.44%.

On the other hand, the system had an average precision of 47.33% and average recall of 42.01% from articles from other websites.

Table 2a. Validation Results on stuartxchange.org

Type of Information	Gold	Retrieved	tp	fp	fn
Plant Name	6	7	6	1	0
Synonyms	17	26	15	11	2
Local Names	52	41	28	13	24
Location	31	26	18	8	13
Contra-indication	35	6	6	0	29
Preparation	37	43	19	24	18

Table 2b. Summary results on stuartxchange.org

Type of Information	Precision	Recall	F-measure	Accuracy
Plant Name	85.71	100.00	92.31	85.71
Synonyms	57.69	88.24	69.77	53.57
Local Names	68.29	53.85	60.22	43.08
Location	69.23	58.06	63.16	46.15
Contra-indication	100.00	17.14	29.27	17.14
Preparation	44.19	51.35	47.50	31.15

Table 3a. Validation Results on other websites

Type of Information	Gold	Retrieved	tp	fp	fn
Plant Name	29	78	12	66	17
Synonyms	16	13	10	3	6
Local Names	42	25	13	12	29
Location	13	12	9	3	4
Contra-indication	0	7	0	7	0
Preparation	50	37	24	13	26

Table 3b. Summary results on other websites

Type of Information	Precision	Recall	F-Measure	Accuracy
Plant Name	15.38	41.38	22.43	12.63
Synonyms	76.92	62.50	68.97	52.63
Local Names	52.00	30.95	38.81	24.07
Location	75.00	69.23	72.00	56.25
Contra-indication	0.00	0.00	0.00	0.00
Preparation	64.68	48.00	55.17	38.10

5. CONCLUSION AND FUTURE WORK

Currently, the information extraction rules works better on articles from stuartxchange.org. However, it can also work on other articles if more extraction rules will be defined and incorporated into the system to cover other sentence construction patterns. Though the data in the ontology is far from perfect or complete, according to the experts, they agree that the information extracted by the system is relevant for future use.

The ontology can also be expanded to accommodate other information related to the medicinal plants. These concepts may include edible plants, vitamins, and minerals. Adding these concepts to the ontology can be useful in determining, for example, which plants or plant parts may contain a particular vitamin or mineral for treating certain illnesses. Recommendation systems can use these information for natural diet plans, alternative medicine information and the like.

6. ACKNOWLEDGEMENT

We would like to express our gratitude to the consultants and evaluators of the system. Special

thanks go to Dr. Esperanza Maribel G. Agoo and Ms. Chona Camille E. Vince Cruz, both from the Biology Department of the College of Science, and Mr. Neil Arvin Bretana, a Bioinformatics professor from the Software Technology Department of the College of Computer Studies for contributing their expertise and feedback to further improve the system.

7. REFERENCES

Aziz, Z., & Tey, N. P. (2009). Herbal medicines: prevalence and predictors of use among malaysian adults. Complement Ther Med..

Celjuska, D. & Vargas-Vera, M. (2004) Ontosophie: A Semi-Automatic System for Ontology Population from Text. In Proceedings International Conference on Natural Language Processing ICON 2004., Hyderabad, India.

Herbal medicine. (2009, September). Retrieved from http://www.umm.edu/altmed/articles/herbal-medicine-000351.htm

Navigli, R., Velardi, P., Cucchiarelli, A., & Neri, F. (2004). Quantitative and qualitative evaluation of the ontolearn ontology learning system. In Proceedingsof the 20th international conference on computational linguistics. Stroudsburg, PA, USA : Association for Computational Linguistics. Disponible sur http://dx.doi.org/10.3115/1220355.1220505

Satyapan, N., Patarakitvanit, S., Temboonkiet, S., Vudhironarit, T., & Tankanitlert, J. (2010). J med assoc thai. Journal of the medical association of thailand.