ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Ontology Extraction for Agriculture Domain Using NLP Techniques and Speech Command

Krunali Patel[1], Bhawna Jain[2], Ashish Bankar[3], Aarya Sarda[4]

U.G. Student, Department of Computer Engineering, SSBT's COET Bambhori , Jalgaon, India

ABSTRACT: The shared specification of conceptual vocabulary used for formulating knowledge-level theories about a domain of discourse is known as Ontology. Data set is created by manually collecting information about different diseases related to crops, it's pesticides and weedicides. Ontology modelling is used for knowledge representation of various domains. Ontology extraction is a process in which important concepts related to a domain are extracted and relationships between them is formed. Ontology modelling is used for knowledge representation of various domains. Majority of Indian population relies on farming but the technologies are sparsely used for the aid and benefit of farmers. India is an agricultural based economic country. Ontology based modelling for agricultural knowledge can change this scenario. The farmers can understand it easily in their native languages like Marathi, Hindi or any other Indian languages. The Ontology Extraction system will model and extract knowledge in Marathi language. A review of various existing agriculture ontology along with some of Natural Language Processing (NLP) models is overviewed. The concept of NLP is useful for input processing and for human-computer interaction. Ontology model for agriculture domain system aims to retrieve appropriate answers to the farmers query and Rule-Based and Conditional Random Fields based models for Ontology extraction is explored. The extraction methods and pre-processing phases of proposed system is discussed.

KEYWORDS: Ontology, NLP, Pre-processing, Tokenization, Stemming.

I. INTRODUCTION

Agricultural information and its related domains are now widely available in the internet. This information is very useful especially to farmers for them to improve their production with respect to changing circumstances and conditions. This agriculture information is normally published on the internet in diverse formats such as Relational Databases, XML, RSS, webpages and others. Daily usage of searching information through search service providers also grow rapidly. Statistically, about 93% of internet traffic is occupied by information searching services. India being a diverse country and language changes after every 20 kilometers it becomes difficult to communicate. And as majority of Indian farmers are not educated, it becomes difficult for them to handle English language. So it is necessary to have a system which will have farmers to gain knowledge in their native language. Marathi is regional language of Maharashtra state. It uses modified version of Devanagari script and Marathi like Standard Marathi, Varhadi , Dangi and Ahirani. Over 68 million people of western speaks the Marathi language. Marathi is an Indo-Aryan language. It is written in Devanagari script similar to the National Language of India i.e. Hindi. Sanskrit language is written using the Devanagari script. In India, Marathi language has the largest number of native speakers.

The most widely quoted definition of ''ontology'' was given by Tom Gruber in 1993, who defines ontology as (Gruber, 1993) [1]: "An explicit specification of a conceptualization".

Ontologies in specific domains such as Health care have been developed on a large scale. In health care, the information regarding medical treatments is consistent worldwide. But in agricultural field, the information changes according to environmental conditions and geographic locations. Agricultural information has strong local characteristics related to climate, culture, history, languages, and also local plant varieties. Farmers in India belong from different states and different states have different languages. Language becomes a barrier as the farmers are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

unaware about other languages. Due to this, it is difficult to build a universal ontology that will provide answers to farmer's queries according to environmental conditions and in native language. The proposed system extracts the knowledge in native language i.e. Marathi. It will help the farmers speaking Marathi language to gain knowledge regarding crop diseases .Natural Language Processing (NLP) is a very active area of research and development in Computer Science. NLP applications are machine translation and automatic speech recognition. Natural language processing techniques are used to process input which is in the form of natural language i.e. human understandable. The idea behind the natural language processing is to interpret input as whole by combining the structure and meaning of words .The words that is interpretations are obtained by matching patterns of words against the input utterance.
The objective of this paper is to highlight the techniques or methods found during the phase of keyword identification, Extraction and constructing agricultural domain for Marathi language. The use of ontology for extraction purpose may provide substantial benefit to user in terms of:

- To help in terms of understand ability which means the farmers can understand it easily in their native language.
- Describes and represent the data in an explicit manner.
- Largely helpful for agriculture education system, farmers, agriculture domain experts and researchers.

II. LITERATURE REVIEW

Juana Maria Ruiz-Martinez [2] and four researchers had proposed Ontology learning from biomedical natural language documents using UMLS. They proposed a methodology for building biomedical ontologies from texts. This approach relies on natural language processing and also knowledge acquisition techniques.This is done to obtain the relevant concepts and relations to be included in OWL ontology. Caterina Caracciolo, Armando Stellato and five researchers [2] provides an overall description of the AGROVOC Linked Dataset and details its maintenance and publication process. AGROVOC is managed by FAO, and owned, maintained by an international community of experts and institutions active in the area of agriculture. It is widely used in specialized libraries as well as digital libraries and repositories to index content and also used as a specialized tagging resource for knowledge and content organization by FAO and other third-party stakeholders.

Gelian Song, Maohua Wang, Xiao Ying puts forward a kind of agriculture domain knowledge ontology representation method. Through the crop planting information expression and integration unity, transform the natural language description or unstructured information into formal, structured knowledge records. And use that knowledge to support agricultural problem solving and decision support effectively. Food Safety Semantic Retrieval System is an ontology-based semantic retrieval experimental system, includes all aspects of food safety knowledge in the field of International Journal of Applied Information Systems emergencies. This system provides the users to access the accumulation of the knowledge in the food safety domain [3].

According to Ling Cao et al, Agriculture Literature Retrieval System defined agriculture literature concepts captured from Encyclopedia of Chinese Agriculture and Catalogue of Ancient Chinese Agricultural Literatures. There are more than 10,000 keywords extracted from the research papers of Chinese agricultural history [4].

Existing Agriculture Ontologies

Agriculture is considered to be a very important sector in creation of raw food items. For economic growth of country agro based industries play a vital role. It is important that all the data regarding agriculture domain should be well organized and properly arranged, so that the farmer can easily retrieve theinter- related data. Ontology extraction techniques can be used for extracting relevant information.

There are many ontology's available online in agricultural domain which includes ontology's for different crops, types of corps, fisheries, animal husbandry, etc. Following are such examples: Agropedia platform is basically an agricultural Wikipedia, which is used for wide range of application in agriculture in India and developed by Indian Institute of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

India-Kanpur (IITK). This knowledge repository consists of universal meta-model and localized content for a variety of users with appropriate interfaces that supports information access in multiple languages [5]. Crop specific ontology's for rice crop were also built in Indian Institute of India-Kanpur. But in India researches are still working in Indian Institute of Technology-Bombay (IITB) in crop specific ontology for cotton crop and building ontology form text document.

Integrated Agriculture Information Framework (IAIF): Integrated Agriculture Information Framework (IAIF) is one of the useful solutions for ontology extraction. This IAIF technique makes knowledge extraction possible from various domain related repositories. Main functions of IAIF technique are combining, merge and aggregate the data in existing knowledge repositories. The three sub ontologies included in IAIF agriculture ontology are Domain ontology, Resource Ontology, Linking Ontology[6].

Scalable Service Oriented Agriculture Ontology for Precision Farming (ONTAgri): Scalable Service Oriented Agriculture Ontology for Precision Farming (ONTAgri) is proposed to use in agriculture domain and this domain consist of several farming practices such as irrigation fertilization and pesticides spraying[6][7].

AGROVOC: AGROVAC is a structured thesaurus created in 1980, by FAO and European Communities. It covers the fields of food, agriculture, forestry, fisheries, etc. It is a multilingual thesaurus[8][9].

Agricultural Ontology Service (AOS): AOS is designed for utilization of AGROVOC encyclopedia at its core. It also serves as a common set of core terms and relationships as well as the richer relationship which can be shared among knowledge organization system. The main purpose of AOS is to achieve interoperability among different agriculture systems[6].

World Agriculture Information Center (WAICENT): WAICENTs is a multilingual knowledge management system. It is powered by FAO. With the help of WAICENT, FAO i.e, Food Agriculture, knowledge of agriculture is available to users around the world through internet[10].

III. PROPOSED SYSTEM

The ontology extraction process is described in Fig.1. The input of the process is a query entered by the farmer in Marathi language and the output is answer related to query i.e. pesticides name.
Three main phases are necessary

- Pre-processing
- Keyword identification
- Knowledge extraction.

For each query, these phases extract the ontological entities contained in the current text.

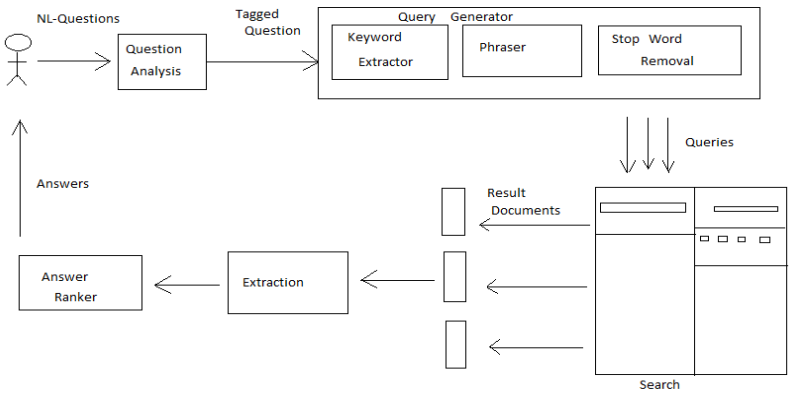

Fig. 1: System Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Pre-processing
Pre-processing is an important task in Natural Language Processing (NLP). The farmer will enter the query in Marathi language. So in the area of natural language processing, data preprocessing used for extracting interesting, non-trivial and knowledge from unstructured Marathi text query. Proposed system extract domain specific terms from the text corpus. The text corpus is processed using various techniques like morph analysis, POS tagging and stop word removal etc. To extract key phrases from the text corpus different lexical patterns are applied.
Relevance of the key term is calculated by counting the frequency of the term in text corpus.

(a) Part-of-Speech Tagging: Part-of-Speech (POS) tagging is a starting point for processing textual information. The words having similar syntactic behavior are grouped into classes.
(b) Tokenization: Tokenization is the process of breaking a stream of text into words, phrases, symbols called tokens. Exploration of the words in a sentence is done by tokenization. These tokens are given for parsing. Tokenization is used to identify the meaningful keywords.
(c) Stop Word Removal: Stop words are not useful for searching. Stop words are used to join words together in a sentence. They occur very frequently in text query, But these words are meaningless. Stop words like and, or, are, this etc are not used for classification of documents, so they must be removed.
(d) Stemming: The process of conflating the variant forms of a word into a common representation is called stemming. For example, the words: presentation, presented, presenting
are reduced to a common representation present.
(e) Syntactically driven parsing: The way that words can t together to form higher
level units such as phrases, clauses and sentences is called syntax. Syntax analysis is obtained by application of grammar that determines what sentences are legal in the language
that is being parsed.

Keyword Identification:
Identify keywords is one of the important task when working with text. Keyword identification is useful because they reduce the dimensionality of text to the most important features. First we locate the attributes by identifying related keywords. We picked one to three keywords for each question i.e, crop name and disease name. By identifying most useful keywords from farmer query related pesticides are extracted.

Extraction Methods
(a) Rule-based method: Rule based models help you to write the rules explicitly. A rule based system consists of a set of rules, a working memory for storing states, a schema for Matching the rules and a conflict resolution schema if more than one rule is applicable.
By using rule based method, we developed the rules needed to extract the main verb of sentence, along with its subject and objects. Proposed system relies on a training corpus of sentences, in which the correct words to be extracted are identified. The first operation that the algorithm must naturally execute is the pre-processing. After pre-processing that is stop words removal, tokenization and part-of-speech tagging, we apply different transformation rules. The training sentences are split into short segments containing at most one word to extract. According to questions different category of question adopted different rule. We also collected frequently mentioned terms into question-specific vocabulary, such as units. The extracting strategy combined with regular expression and term searching within such vocabulary.

(b) Conditional Random Fields (CRFs) method: Conditional random fields (CRFs) are a probabilistic framework for labelling and segmenting structured data. The idea is to define a conditional probability distribution over label sequences given at a particular observation sequence. CRF is an discriminative model. It does not assume the features that are independent.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

IV. RESULT AND DISCUSSION

Ontology evaluation basically depends on two aspects i.e. quality and correctness. Number of frameworks and methodologies are available for ontology evaluation.

Table.1. Ontology Evaluation

Ontology Evaluation Perspective	Metric	Measure
Correlation	Accuracy	Precision: total number correctly found over whole knowledge defined in ontology Recall: total correctly found over all knowledge that should be found
	Consistency	Count: Number of terms with inconsistent meaning
Quality	Efficiency	Size
	Clarity	Number of word senses

Standard metrics (precision, recall and F-score) will be used for measuring the performance. Let S be the size of the ground truth list (doctors annotations), D is the number of Correct , distinct values extracted by our system and N be the total number of values returned by the system.

$$\text{recall} = \frac{D}{S}$$

$$= \frac{\text{number of correct,distinct values returned by the system}}{\text{size of the ground truth list}} \quad (1)$$

$$\text{precision} = \frac{D}{N}$$

$$= \frac{\text{number of correct,distinct values returned by the system}}{\text{total number of results returned by the system}} \quad (2)$$

$$\text{F score} = \frac{2*\text{precison}*\text{recall}}{\text{precison}+\text{recall}} \quad (3)$$

Table.2. Test Cases for Input by Speech Command in English Language

Test case ID	Input	Output	Expected Output	Actual Output	Result
1.	Pests for sugarcane	All information about pest for sugarcane and it's images.	All information about pest for sugarcane and it's images.	All information about pest for sugarcane and it's images.	PASS
2.	Weeds for pomegranate	All information about weed of pomegranate and it's Images.	All information about weed of pomegranate and it's Images.	All information about weed of pomegranate and it's Images.	PASS
3.	Weads of cotton	All information about cotton and its weeds.	All information about weeds for cotton and it's images.	Result not found.	FAIL

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

In the above Table 2. Few Test cases are discussed where ID no. 1 & 2 shows the expected result while due to recognition of incorrect word the result is not found. The Standard metrics are being used for measuring the performance of the system as shown in the Table.3.

Table.3. Ontology Evaluation

Test case ID	Precision	Recall	F-Score
1.	0.745	0.789	0.766
2.	0.712	0.755	0.732
3.	1	1	1

In the above Table.3 Precision, Recall and F-score values of each question are shown. For Test case ID 3 there are no values as the information is not present in the ontology. The system will return that the crop name or disease name is not present.

V. CONCLUSION AND FUTURE WORK

Due to huge dependency upon agriculture field, proper growth must be important in this field and for this farmer plays vital role to improvement. For this farmer must be aware about all well detailed information about and things related to agriculture. Ontology's semantic feature is not limited to text only. It describe and represent the data in explicit manner. The aim was to help users in terms of understandability which means farmer can understand it easily in their native language. This technology is largely helpful for agriculture education system, farmers, agriculture, domain experts, researchers. The main benefit is that it increases accuracy as well as performance ensuring the relevant information. Ontology extraction provides satisfactory solution for their search queries. Able to assist users by reducing agriculture inputs and shortens the response time.

In the future, the system can be expanded for more number of crops and can be made available in other Indian languages. Also functionality like user can easily navigate to nearby pesticides shop from their location and weather information is also provided.

REFERENCES

1. Brijesh Bhatt and Pushpak Bhattacharya, "Domain Specific Ontology Extractor for Indian Languages", Proceedings of the 10th Workshop on Asian Language Resources, pp. 75-84, 2012.
2. Juana Maria Ruiz Martinez, et al., "Ontology Learning from Biomedical Natural Language Documents using UMLS", *Expert Systems with Applications*, Vol. 38, No. 10, pp. 12365-12378, 2011.
3. Yuehua Yang, Junping Du and Meiyu Liang, "Study on Food Safety Semantic Retrieval System based on Domain Ontology", Proceedings of IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 40-44, 2011.
4. Ling Cao and Lin He, "Domain Ontology-based Construction of Agriculture Literature Retrieval System", *Proceeding of 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1-3, 2008.
5. Ho-Young Kwon, Sabine Grunwald, Howard W. Beck, Yunchul Jung, Samira H Daroub, Timothy A. Lang and Kelly T. Morgan, "Ontology-based Simulation of Water Flow in Organic Soils applied to Florida Sugarcane", Agricultural Water Management, Vol. 97, No. 1, pp. 112-122, 2010.
6. Gelian Song et al., "Study on Precision Agriculture Knowledge Presentation with Ontology", Proceedings of Conference on Modelling, Identification and Control, Vol. 3, pp. 732-738, 2012
7. Rayner Alfred et al., "Ontology-Based Query Expansion for Supporting Information Retrieval in Agriculture", Proceedings of 8th International Conference on Knowledge Management in Organizations, pp. 299-311, 2014.
8. Aqeel-ur Rehman and Zubair A. Shaikh, "ONTAgri: Scalable Service Oriented Agriculture Ontology for Precision Farming", *Proceedings of International Conference on Agricultural and Biosystems Engineering*, pp. 1-2, 2011.
9. Caterina Caracciolo et al., "The Agrovoc Linked Dataset", Semantic Web, Vol. 4, No. 3, pp. 341-348, 2013.
10. Boris Lauser, Margherita Sini, Anita Liang, Johannes Keizer and Stephen Katz, "From Agrovoc to the Agricultural Ontology Service/Concept Server", Food and Agriculture Organization of the United Nations, pp. 1-10, 2006