ISSN: 2229-6959 (ONLINE)
DOI: 10.21917/ijsc.2016.0188

ONTOLOGY EXTRACTION FOR AGRICULTURE DOMAIN IN MARATHI LANGUAGE USING NLP TECHNIQUES

Prachi Dalvi[1], Varsha Mandave[2], Madhu Gothkhindi[3], Ankita Patil[4], S. Kadam[5] and Soudamini Pawar[6]

Department of Computer Engineering, D Y Patil College of Engineering, India

E-mail: [1]prachidalvi409@rediffmail.com, [2]varsha.m9922@gmail.com, [3]madhu.gotkhindi@gmail.com, [4]patilrankita1994@gmail.com, [5]suvarna.kadam@gmail.com, [6]psoudamini@yahoo.co.in

Abstract

Ontology is defined as shared specification of conceptual vocabulary used for formulating knowledge-level theories about a domain of discourse. Dataset is created by manually collecting information about different diseases related to crops. Ontology modeling is used for knowledge representation of various domains. India is an agricultural based economic country. Majority of Indian population relies on farming but the technologies are sparsely used for the aid of farmers. Ontology based modeling for agricultural knowledge can change this scenario. The farmers can understand it easily in their native language. We proposed a system which will model and extract knowledge in Marathi language. In this paper, we review various existing agriculture ontology's along with some of Natural Language Processing (NLP) models. Model ontology for agriculture domain system aims to retrieve relevant answers to the farmer's query. We explored Rule-Based and Conditional Random Fields based models for Ontology extraction. The extraction methods and preprocessing phases of proposed system is discussed.

Keywords:

Ontology Modeling, Agriculture, NLP, Marathi, Domain Ontology

1. INTRODUCTION

More than 70 percent of population in India has agriculture as a mean of livelihood. The agriculture domain is very vast. Large number of data has been written in books and till today lots of electronic data is available. Farmers in India are badly affected by not being able to get vital information required to support their farming activities in a timely manner. Some of the required information can be found in government websites, agriculture department leaflets, and from radio and television programs. Due to its unstructured and varied format, and lack of targeted delivery methods, knowledge is not reaching the farmers. 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 few dialects of Marathi are Standard Marathi, Varhadi, Dangi and Ahirani.

There are over 68 million people of western and central India 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.

Marathi is spoken in the complete Maharashtra state which consists of 34 different districts. Marathi language is the most effective and common way of communication between farmers in Maharashtra. Most of the farmers are able to understand Marathi language in Maharashtra.

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 have proved their usefulness in different applications scenarios, such as natural language processing, semantic web, intelligent information integration, knowledge-based systems and digital libraries. Ontologies are developed to separate domain knowledge from operational knowledge. Reuse of domain knowledge and operational knowledge is possible using ontologies.

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 in relation to climate, culture, history, languages, and local plant varieties. Farmers in India belong from different states and different states have different languages. Language becomes a barrier as the farmers are 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 that is interpretations are obtained by matching patterns of words against the input utterance.

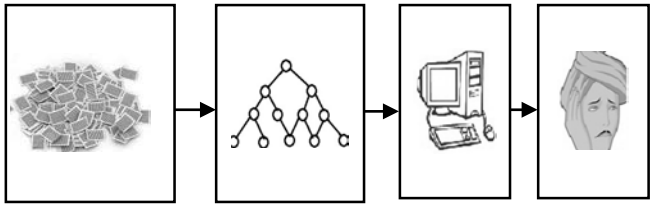

Fig.1. Architectural Overview

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 understandability which means the farmers can understand it easily in their native language.
- Describe and represent data in an explicit manner.
- Largely helpful for agriculture education system, farmers, agriculture domain experts and researchers.

The remainder of this paper is structured as follows: Section 2 describes Literature review on existing agriculture ontologies are discussed in text format. In section 3 Challenges are discussed. In section 4, proposed system and Ontology extraction methods are discussed. In section 5 results are discussed in tabular format along with ontology evaluation terminologies.

2. LITERATURE REVIEW

Juana Maria Ruiz-Martinez and four researchers [2] 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 knowledge acquisition techniques 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. AGROVOC is widely used in specialized libraries as well as digital libraries and repositories to index content. It is 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 [18].

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 [19].

2.1 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 the inter-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 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 [16]. Crop specific ontology's for rice crop were also built in IITK. 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.

2.1.1 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 [3].

2.1.2 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 [3] [4].

2.1.3 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 [5] [6].

2.1.4 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 [3].

2.1.5 World Agriculture Information Center (WAICENT):

WAICENT's is a multilingual knowledge management system. It is powered by FAO. With the help of WAICENT, FAO knowledge of agriculture is available to users around the world through internet [7].

2.1.6 Citrus Water and Nutrient Management System (CWMS):

The basic purpose behind this system is balance processes for Water and nutrient for citrus production and included 700 symbols and 500 equations. Block, soil cell, soil profile, soil layer, root distribution, irrigation system, and weather are included in this system [15].

2.1.7 OntoSim-Sugarcane:

OntoSim is an application and basic purpose of this system is to represents hydrology, nutrient cycling, plant growth, soil moisture, crop growth on organic soils and nutrient uptake in southern Florida sugarcane production and 195 equations and 247 symbols are included in this collection [16].

2.1.8 OntoCrop Ontology:

This is the ontology constructed for horticulture domain of agriculture and the author examines the usage in particular domain. This ontology is used,

- As a refining and classification tool facilitating indexing and searching process in a repository environment.
- As a domain model for rule knowledge base construction [17].

2.2 CHALLENGES

2.2.1 Data Collection:

Agriculture data in Marathi language is not available electronically. This was the main challenged we faced during data collection. The data was first collected in English language and then we converted it into Marathi using Google Translate.

2.2.2 Structured Data:

The another challenged we faced was structure data was not available for Agriculture in Marathi language. So we had to deal with unstructured data.

2.2.3 Standard Tool:

There is no standard ontology modeling tools, to model ontology in Marathi language. Hence the ontology is modeled manually.

2.2.4 Processing:

Applying pre-processing techniques on natural language (Marathi) was difficult. Parsing Marathi statements was difficult.

3. PROPOSED SYSTEM

The ontology extraction process is described in Fig.2. 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:

(a) Preprocessing

(b) Keyword identification

(c) Knowledge extraction.

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

3.1 PREPROCESSING

Preprocessing is an important task in Natural Language Processing (NLP). In the proposed system 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.

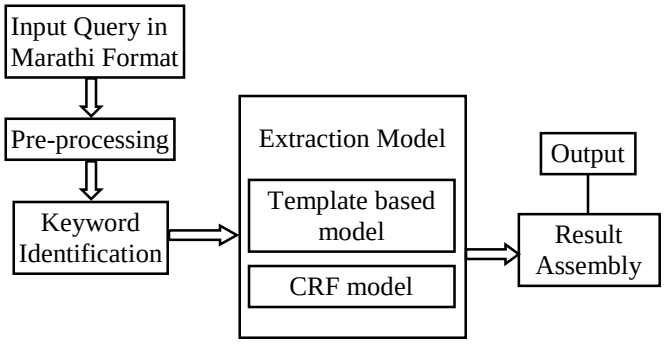

Fig.2. System Flow

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.

3.1.1 Part-of-Speech Tagging [12][8]:

Part-of-Speech (POS) tagging is a starting point for processing textual information. The words having similar syntactic behaviour are grouped into classes.

3.1.2 Tokenization:

In tokenization, a stream of text is break into words, phrases, symbols which are called as 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.

3.1.3 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.

3.1.4 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”.

3.1.5 Syntactically driven parsing:

The way that words can fit together to form higher level units such as phrases, clauses and sentences is called syntax. Syntax analyses are obtained by application of grammar that determines what sentences are legal in the language that is being parsed.

3.2 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.

Table.1. Questions Designed

Question ID	Question
1	मीरचि पीकावर मावा रोग आला आहे. ऊपाय सांगा
2	वांगे च्या फुलावर स्पायडरमाईटस आहे तर कोणते औषधे वापरु ?
3	संत्रा वर काली माशी पडली आहे. माशी घालवण्यासाठी काय करु ?
4	संत्रा फळातील रसशोषणारी किड साठी कोणत्या उपाययोजणा करायला हव्या ?
5	पपई च्या पानांवर रिंग स्पॉट व्हायरस मुळे पाने वाकडे झालेत, पाने सरळ आणि व्हायरस संपवण्यासाठी उपाय सांगा.
6	पपई वर काळे ठिपके पडले आहेत, कशाची फवारणी केल्यास ते जातील या बद्दल माहिती कळवावी.
7	लालकोळी ची मिरची वर लागण झाली आहे. कसा नष्ट करता येईल ?
8	डाळिंब च्या खोडावरील खवले किड नियंत्रित कशी करु ?
9	शेवगा झाडा च्या पान व शेंगा खाणारी अळि पडलि आहे. योग्य मार्गदर्शन करा.

3.3 EXTRACTION METHODS

3.3.1 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 words are identified so that they can extract. The first operation that the algorithm must naturally execute is the preprocessing. After preprocessing 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 कोणते, कोणत्या. The extracting strategy combined with regular expression and term searching within such vocabulary. The Table.1 shows the different questions asked by the farmers.

3.3.2 Conditional Random Fields (CRF's) Method:

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

We have used CRF for knowledge Extraction in our project. Features include word identity, transitions among class labels, starting features, ending features, word score features and features for handling words that are new or have only been observed in other states so far. For training and testing, we included the other category crops like mango, tomato, bhendi etc. Our CRF implementation consists of two steps: First, the CRF identifies relevant terms. These terms are marked as being a part of a relevant entity. If consecutive words are identified as belonging to one entity (e.g. for mango crop), they are deterministically designated one concept. Second, CRF can be used to classify the identified relevant entities. This is done by merging the consecutive words, identified as entities into one concept. This concept is represented as a concatenation of the consecutive entity words. Finally, CRF will provide answer for farmer query.

4. RESULTS AND DISCUSSIONS

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

Table.2. Ontology Evaluation

Ontology Evaluation Perspective	Metric	Measure
Correctness	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

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

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

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

Considering two class disease and non-disease, if system provides disease then value 1 is considered and if system does not show the results it will be considered as 2 and in non-disease class. We have evaluated the values provided by the system with agriculture expert to check the correctness.

Table.3. Results

Question ID	Keywords	Extracted Value
Q.1	मीरचि, मावा	इमिडाक्लोप्रिड, डायमेथोएट, फॉस्फोमिडॉन, या पैकि फवारणी घ्यावी.
Q.2	वांगे, स्पायडरमाईटस	इमिडाक्लोप्रिड, डायमेथोएट, फॉस्फोमिडॉन, थायमेथॉक्झाम या पैकि एकाची फवारणी घ्यावी.
Q.3	संत्रा	वरील पीक किंवा रोगा बद्दल माहिती उपलब्ध नाही.
Q.4	संत्रा,रसशोषणारी किड	गुळवेल, वसंतवेल या सारखे गवत काढुन टाकावे .फळांना बॅग ने झाकुन टाकावे . तसेच संध्याकाळी शेतात दाट धुर करावा
Q.5	पपई , रिंग स्पॉट व्हायरस	पपईच्या शेताच्या जवळपास वेलवर्गिय पिकांची लागवड करु नये. मावा किड नियंत्रणात ठेवावी.
Q.6	पपई , काळे ठिपके	डायथोन एम ४५, जिनेब, कॅपटन, कॉपर औक्सिक्लोराईड, कॉपर हायड्रॉक्साईड, यापैकी एकाची फवारणी घ्यावी.
Q.7	मिरची, लालकोळी	डायकोफॉल, सल्फर ची धुरळणी, अबामेक्टिन
Q.8	डाळिंब, खवले	फॉस्फोमिडॉन, मोनोक्रोटोफॉस, एन्डोसल्फान,
Q.9	शेवगा, शेंगा खाणारी अळि	वरील पीक किंवा रोगा बद्दल माहिती उपलब्ध नाही.

Table.4. Classification

Question ID	Classification	Expert
Q.1	1	1
Q.2	1	1
Q.3	2	1
Q.4	1	1
Q.5	1	1
Q.6	1	1
Q.7	1	1
Q.8	1	1
Q.9	2	1

Table.5. Precision, Recall and F-score

Question ID	Precision	Recall	F -score
Q.1	0.745	0.789	0.766
Q.2	0.712	0.755	0.732
Q.3	1	1	1
Q.4	0.587	0.595	0.590
Q.5	0.861	0.870	0.865
Q.6	0.746	0.798	0.771
Q.7	0.567	0.543	0.554
Q.8	0.674	0.631	0.651
Q.9	1	1	1

Overall performance of rule based system: In the below table Precision, Recall and F-score values of each question are shown. For question 3 and question 9 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. The Fig.3 and Fig.4 shows the screenshots of the system implementation.

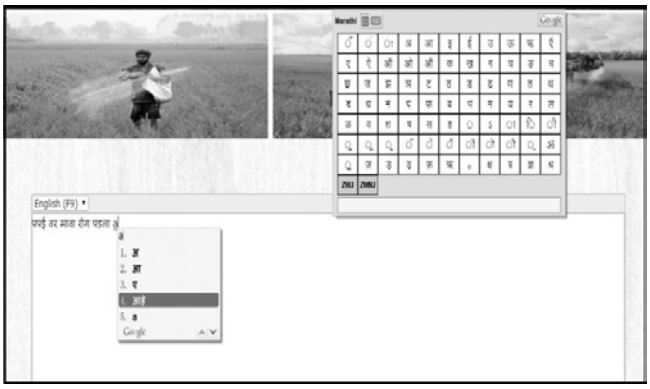

Fig.3: Enter query in Marathi

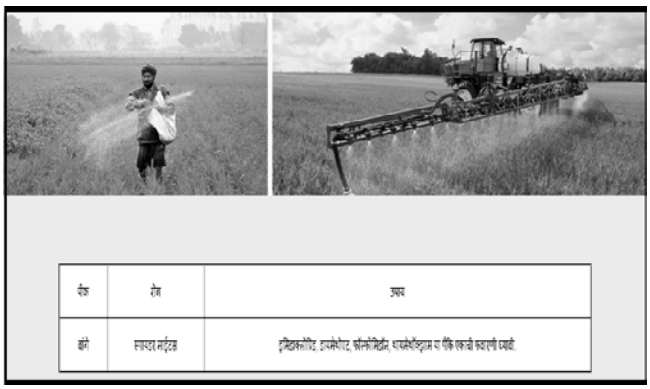

Fig.4. Answer of query

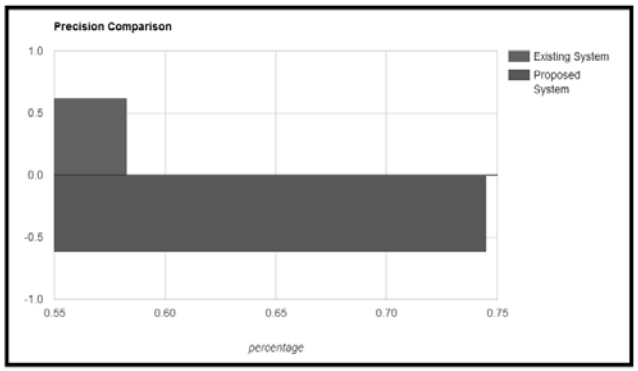

Fig.5. Precision graph for above query

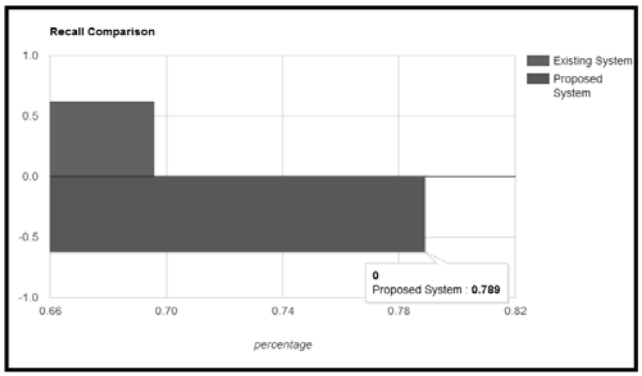

Fig.6. Recall graph for above query

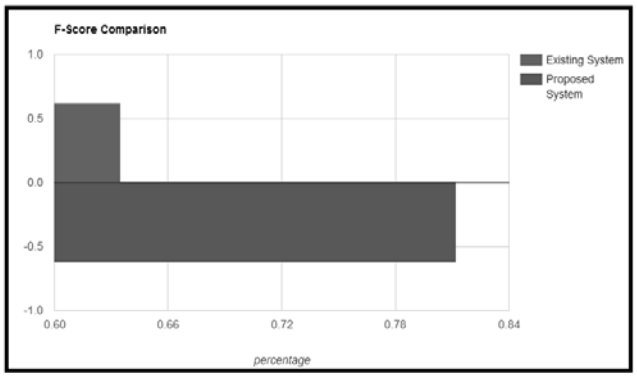

Fig.7. F-Score graph for above query

5. CONCLUSION

In this paper, we have shown ontology modelling for agriculture domain in Marathi language. According to the farmer's query, the relevant information will be extracted and solution will be provided. The average F-score we obtained for 9 questions is 76.98%. In future, the system can be expanded for more number of crops and can be made available in other Indian languages.

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] 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.

[4] 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.

[5] 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.

[6] Caterina Caracciolo et al., "The Agrovoc Linked Dataset", *Semantic Web*, Vol. 4, No. 3, pp. 341-348, 2013.

[7] 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.

[8] A. Mangstl, J.R. Judy and F.L.H. Ward, "The World Agricultural Information Centre (Waicent) Faos Information Gateway", *Proceedings of 1st European Conference for Information Technology in Agriculture*, pp. 189-198, 1997.

[9] Chris Manning and Hinrich Schutze, "*Foundations of Statistical Natural Language Processing*", MIT press, 1999.

[10] Daniel Jurafsky and James H. Martin, "*Speech and Language Processing*", 2nd Edition, Prentice hall, 2008.

[11] Leyla Zhuhadar, "A Synergistic Strategy for Combining Thesaurus based and Corpus-based Approaches in Building Ontology for Multilingual Search Engines", *Computers in Human Behavior*, Vol. 51, pp. 1107-1115, 2015.

[12] Hui Wang,Weide Zhang, Qiang Zeng, Zuofeng Li, Kaiyan Feng and Lei Liu, "Extracting Important Information from Chinese Operation Notes with Natural Language Processing Methods", *Journal of Biomedical Informatics*, Vol. 48, pp. 130-136, 2014.

[13] Alexandre Trilla, "Natural Language Processing Techniques in Text-to-Speech Synthesis and Automatic Speech Recognition", *Departament de Tecnologies Media*, pp. 1-5, 2009.

[14] Agropedia, Available at: *http://www.agropedia.iitk.ac.in*

[15] Howard Beck, Kelly Morgan, Yunchul Jung, Sabine Grunwald, Ho-Young Kwon and Jin Wu, "Ontology-based Simulation in Agricultural Systems Modeling", *Agricultural Systems*, Vol. 103, No. 7, pp. 463-477, 2010

[16] 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.

[17] Michael T. Maliappis, "Applying an Agricultural Ontology to Web-based Applications", *International Journal of*

Metadata, Semantics and Ontologies, Vol. 4, No. 1-2, pp. 133-140, 2009.

[18] 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.

[19] 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.