ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.617-621, August 2020

Location Identification Using Stanford NLP

S. Vishnu Manoj
Department of Computer Science and Application
Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
Vishnumanoj71@gmail.com

Abstract—Even since the 40's the scope of natural language processing has been primal dismay in computer science and Artificial Intelligence. It aspires to include the next strive forward in Artificial Knowledge which can perform both computers and Individual work with better malleability and apprehension. It incorporates various methods like machine translation, speech recognition, online learning, auto tutor etc. Researchers recalled it as a potential bridge that can amalgamate human spoken language and computer which uses programming language and binary codes. Since it is an impossible task to prepare a computer to recognize human natural language, further techniques and enhancements will foster the demanding yet rewarding and innovative computational trends. This paper confers a restrained domain metaphysical model for agriculture cultivation question answering system. The question answering system has been noted as a significant tactic for knowledge engineering research. Ontologies facilitate the computers in deciphering the restrain domain concept of semantics. Thus forming a significant technique for the question-answering system. This paper inculcates introduction ontology and the definition of a domain ontology for agriculture cultivation. The paper also focused on presenting the restricted domain ontology models and concept level vector space model of information retrieval.

Keywords—NLP – Natural Language Processing, Semantic, Stanford, Geocoding

I. INTRODUCTION

In the era of big data, we are mainly dealing with the manipulation of data. The date is the inevitable part we can't neglect it. Our existing system will give more importance to the data and information that's why we are implementing modern technology to make the information more important. Our paper is focusing on the ontology and extraction of information identifying places that are found with the help of Stanford NLP and Google Map API[1]. Ontology is interpreted as a collection of the word relating to a certain domain, The Horticulture domain is huge. There will be a lot of documents in a particular domain which contain a lot of information. In our electronic world, there will be no proper system for answering the query until the data are available only in books. Our farmers are badly affected by not get the proper information from any official government websites[2].

Accruing the vital information is the ultimate task for that we are using Natural Language Processing and Google map API's. The data are extracting from web pages using scrapping tools and process that information split into a different sentence. From each, it will process tokenized and get POS Tagging(Part Of Speech) from it. The POS Tagging results will help as the get the entities in each sentence and find out the relationship hidden in each sentence[3].

This work is used for Extracting semantic relations between entities in text. It will able to answer the question at underlying Domain.

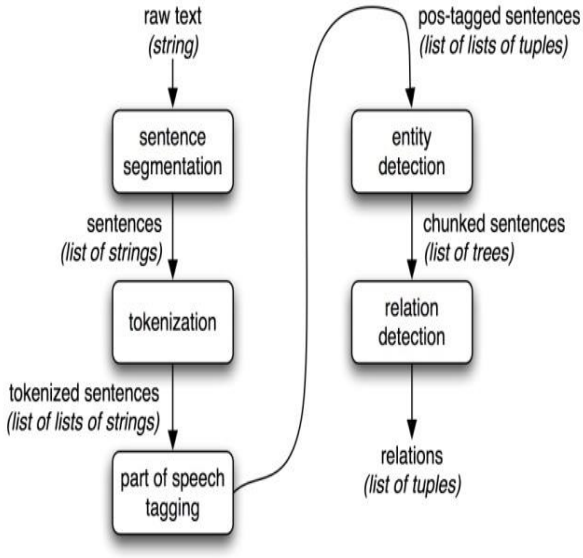


Fig. 1.Information Extraction

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.617-621, August 2020

II. LITERATURE SURVEY

Natural language processing technology provides a possible solution for agriculture and soil looking to learn actionable insights from unstructured data. Relevant Data can be fetched using Information Extraction Technology. It includes Named entity recognition[4], Co-referent noun phrases, Semantic roles, Relation extraction. Named entity recognition on the bases of detection and classification of expressions, which refers to specific person, place etc. Co-referent noun phrases extraction uses different expressions for the same person or object. NLP deals with supervised and unsupervised data.

The fig. 1 will show the step included in Information Extraction. The data that is extracted will process, the initial step of processing will be giving raw text to the existing system, It will perform sentence tokenization deals with the raw text will be split to different string it will be very helpful of extracting key phrases from the source document[5]. The Strings are split into small token called tokenization. using this tokenization text POS tagging will be performed. POS Tagging will read the text from in some languages and it will assign for each tokenized strings such as noun, verb, adjective, noun phrase

III. Proposed Methodology

The risk associated with natural language processing is when a user demands a piece of information especially using keys to retrieve textual information. It often needs natural language processing in retrieving textual information as they facilitate definitions for document content and raise user's query. NLP aims to analyze the statements and provide the users with the best results that satisfy their information demands. In response to a user's need, a textual knowledge retrieval system carries out the following task:- · Indexing the set of details - The NLP technique is implemented to provide the results for the user queries and to generate an index containing document descriptions[6]. Documents are defined through a set of terms which are best at its content. When a user raises a query, the system analyses it and transforms it to provide the user's information needs just like the way document content is defined The system analyses and compares the description of each document with that of the query and presents the results to the user with those documents in which descriptions are closer to the query description.

A. Objectives

1) place recognition from the Chronicle
2) The anxiety of NLP over a Standford system

The IE system works on a different scale based on our need it will perform its task. We pointed out there is no efficient system that will satisfy the farmer need. According to India culture, those who are poor mainly come in agriculture so education level is pathetic the information or knowledge it can't able to understand that is the big problem facing todayp[7]. To avoid this issue, Our System will be provided with the information they required. In our current situation, vital information is Agriculture entities we want to omit location details from the document.

- machine-learning
- intelligence
- Natural Language Processing

N These are the techniques that inverted index for information retrieval. The inverted index is the list of keywords and the links to corresponding documents. Frequently some preprocessing steps are taken before creating an inverted index

IV. PREPROCESSING

A. Tokenizing

A sequence character inside a document is a collection of tokens. tokenizing is the task of pull apart it's sentencing into trivial pieces, called tokens, and maybe tokenizing will eject certain characters such as punctuations, commas. The tricky task esoteric inside it is Splitting on whitespaces. it occurs most commonly with names, for example, Marine Drive, Sultan Bathery. A most common example used in San Francisco, Los Angeles. In this case, San Francisco is tokenized into (San, Franciso) but also with words that are sometimes written as a single word and sometimes space-separated (such as white space vs. whitespace). The whole meaning will be changed[8].

B. Stopping stop-words

Many times, some extremely common terms which are usually termed as insignificant and trivial, thereby of very little value are excluded from the vocabulary entirely. Such words are called stop words. The common strategy for arbitrating a stop list is to sort the term by collection frequency.[collection Frequency refers to the total number of times a word appears in the document], and then determining the frequently appearing terms. However, it must be noted that excluding stop words isn't always feasible[9].

C. Normalisation

Normalization is the process of changing a list of words into a uniform canonical sequence. It is a process which improves text matching. As an example, convert all words to lower case to make the searching process easier. Another instance is after breaking up documents (and also our query) into tokens if words in the query just match with tokens in the token list of the document that is considered to be an easy case. Though there are various scenarios where sequences of two-character are not quite the same but would like to have a match. For this case in point, if the tokens USA and U.S.A these two pairs mapped onto the term USA, moreover both the record text and queries, then searches for one term which will retrieve documents that contain either. But in another instance, the place named Karthikappally [a village in Alappuzha district] will normalize and detected as a person Karthi. These may also happen.

D. Stemming and Lemmatization.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.617-621, August 2020

Stemming usually refers to the process of reducing the derived word to their root form. Lemmatization is the process by which the inflexion ending of the word is removed, thus reducing it to the base form or the root word. The root word is referred to as 'Lemma'. The documents, for grammatical purposes, prefer to use a different form of the same word, Like organise, organises, organising. In addition to that, the families of related words with similar meanings come into play. Many times it would call for the search for one of these words to return documents that include another word in the set.

E. PosTagging

It is the process of reading text in some language and assigning parts of speech to each word such as noun, verb, adjective, etc using certain NLP software. The post-tagging process provides word chunks as output. It has also found a variety of uses in doing NLP tasks. Postags are extremely used to provide a word within the scope of a sentence, phrase, or document. Suppose if a sentence is given, there can be multiple interpretations possible that yields various kinds of parts of speech tags for different words. Whenever a sentence is given, firstly we have to know what specific meaning it conveys. Postagging resolves certain ambivalence when assigning a syntax category to each word present in a text. Pos tags like 'noun plural' has generally found its application in the computational field. It is also an NLP technique used for information retrieval[10].

The figure shows the steps taking place in Pos-tagging. There is a sentence given and Pos-tagging takes the responsibility of reading text in a certain language and assigns specific parts of speech in the form of tokens to each word.

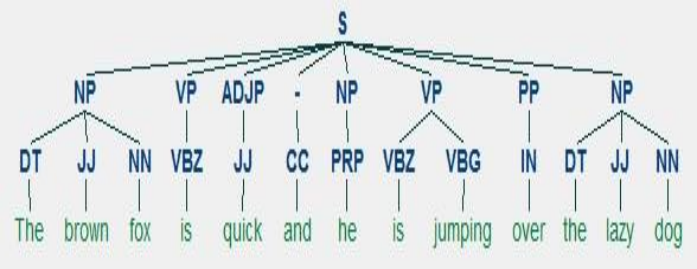


Fig. 2.POS-TAGGING

There are two steps involved and they are:-

1) A word or text tokenization

2) Apply Pos-tag to again tokenize text.

Each token abbreviation has got specific meaning . for eg:- NP stands for 'noun plural', VP for 'verb plural', etc. Hence Pos-tagger for each word of the sentence gives grammatical information.

V. METHODOLOGY

Web Crawling is a process in which the crawler traverses through the site and extract the relevant information. The needful information is retrieved from the web and the data retrieved is given to our system which is then taken for Pre-processing. The steps followed in Pre-processing are as follows:- Tokenization is the process of chopping up the given text into units. This task is done by locating word boundaries. Tokenization is also known as word segmentation because it breaks up a string into pieces called tokens. When Pre-processing is applied to the data, the output will be of tagged data(s). This tagged data(s) consists of NP,NNS ,NN ,IN ,NNP, DT, VBG etc. We are mainly focusing on the retrieval of location from the scraped document. So we use the Noun phrase for this purpose and it includes nouns. So it is impossible to predict whether it is location or some other nouns directly. So To avoid this problem, we reached a substitute method. There are practical complications related to data while searching for an insignificant location. eg: Vallikavu is a coastal region near karunagappally in Kollam, Kerala There will be practical difficulty to spot vallikavu as a place. The NLP tool strictly maintains the grammatical rules. The preposition 'at' is also frequently used in place phrases. One use is for exact addresses (addresses with a house or building number).In our English Language, When we speak about place and time, 'in','on','at' is the three main words that are commonly used. These common words show the relationship between the location. Using these words, NLP tools predict that the forthcoming word will be a place. But in the current instance, we are depending on NLP for the NOUN PHRASE extraction (NNP) that should be a place, person or a thing and it is not possible while we are going for the hierarchical fetching. For dealing with its Location Hierarchy, we can make use of GOOGLE MAP API's. It will embed the map surface to the corresponding project. This feature will get location hierarchy of particulate place but it is very difficult with NLP. so for achieving this particular task, we make use of STANFORD NLP Tools. Named Entity Recognition using Stanford is an alternative to NLTK classifier. This tagger is essentially seen because it is customary in named entity recognition, however, it uses a complicated applied mathematics learning rule for additional computationally pricey than the choice provided by NLTK. A big advantage of the Stanford NER tagger is that it provides us with some completely different models for withdrawal method named entities. We can use any of the following:

1) 3 category model for recognizing locations, persons, and organizations
2) 4 category model for recognizing locations, persons, organizations, and miscellaneous entities
3) 7 category model for recognizing locations, persons, organizations, times, money, percents, and dates

These three models are used as an approach by Stanford NER to fetch the location. Geocoding is a term that is used for this process Identifying the entities in the unstructured information or text is only making a fraction of our activity. By the guidance of ontology, models afford the linguistic content in

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.617-621, August 2020

the form of antique knowledge. Ontology picturized how one entity is related to other entities. Ontology is a set of concepts or Ideas that are categories in a domain that shows it's properties and their relationship between them. In the research set that tagged our need is inevitable. This concept the concepts of semantics being summarised, which has been one significant technique for question answering system. The main trouble that our project is dealing with there is no material about soil irrespective to the corresponding location in Kerala .so we demand to produce our ontology to run as per our demand. It is not at all an easy task for learning that information. for our requirements, the details that are prepared by SOIL SURVEY OF INDIAN is taken to the proposal. Our system only sees regions in Kerala. The role of soil in the existence of mankind is well known. Its role in crop production and productivity is often underestimated. A thorough understanding of the soil facilitates its better management which is possible only through a resource inventory. Soil Survey is a not establishment scientific tool for the generation of soil database based on our need. The soil survey organization has amassed valuable information on the soil of the state and established over 350 soil series. The Bench Mark of soil is selected among the existing and established soils that represent a typical range of characters. Using this helps to focus investigative efforts on key soils that have the greatest potential to apply new technologies across the stream. Adopting the map developed by Cartographic and Geomatics Laboratory we generate ontology of soil in Kerala. These above map help to create an ontology for our system protege tool is free to open source ontology editor and knowledge storing or management system. It very helps in semantics web development. Protege help to edit ontology using OWL Language(Web Ontology Language).ontology maybe use for knowledge knowing to the system as a part of Artificial Intelligence.It splashes the relation between the class(Entities).Classes, data properties, Object properties and individuals are the main concepts in the ontology. Class in ontology are set or collections of objects which are abstract. In object, properties make clear how the classes can narrate to each other based on their instance. data properties are typed literals. grouped level component of ontology is called individuals. To visualize the ontology graphically, we need to install the Protege OWL plugin. In our soil ontology "Location" is the class.14 districts as it's subclass and each subclass have instance it will be a major place with various soils. An instance of each subclass has soil as its data property. The main use of ontology because of graphical user interface makes the development much easier. SPARQL is a query language that use to retrieve the information from ontology.

Fig. 3. ontology graph

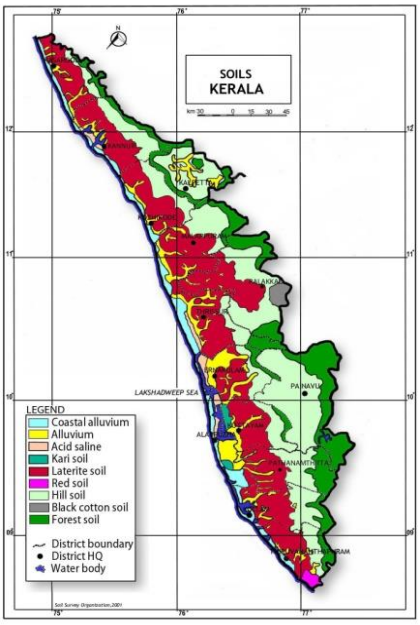
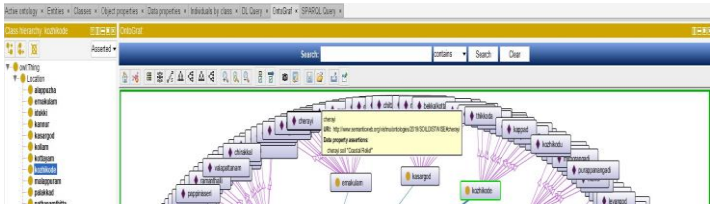


Fig. 4. Soil of Kerala

VI. CONCLUSION

This ontology will help to predict the soil that will present in various location of Kerala. NLP is a power full tool for predication and sentimental analysis with the help of language processing the world of research and artificial intelligence will . Get a hand on something that relates on



geographical location and its properties we can't neglect the importance of GOOGLE Map AP I'S

ACKNOWLEDGEMENT

The authors would like to thank the Department of Computer Science and Applications, Amrita Vishwa Vidyapeetham, Amritapuri Campus for maintaining a positive attitude throughout the semester and providing continual encouragement. While doing my work has given as countless suggestions, and was extremely proactive all times. Our sincere thanks to Dr S.N. Jyothi, Principal, Amrita School of Engineering, Mrs Ani R, Vice-Chairperson, CSA dept. head, for their support throughout the project.

We also extend our sincere thanks to Mrs Raji R, Mr Krishna S, Project co-ordinator, for providing all necessary facilities to carry out our project successfully. We have deep gratitude towards our faculties for giving us technical and moral support from time to time

VII. REFERENCES

[1] Akbik, Alan, and Jügen Broß. "Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns." In *www workshop*, vol. 48. 2009.

[2] Giuliano, Claudio, Alberto Lavelli, and Lorenza Romano. "Exploiting shallow linguistic information for relation extraction from biomedical literature." In *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

[3] Giuliano, Claudio, Alberto Lavelli, and Lorenza Romano. "Relation extraction and the influence of automatic named-entity recognition." *ACM Transactions on Speech and Language Processing (TSLP)* 5, no. 1 (2007): 1-26.

[4] Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." In *Coling 1992 volume 2: The 15th international conference on computational linguistics*. 1992.

[5] Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003-1011. 2009.

[6] Rusu, Delia, Lorand Dali, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. "Triplet extraction from sentences." In *Proceedings of the 10th International Multiconference" Information Society-IS*, pp. 8-12. 2007.

[7] Montes-y-Gómez, Manuel, Aurelio López-López, and Alexander Gelbukh. "Information retrieval with conceptual graph matching." In *International Conference on Database and Expert Systems Applications*, pp. 312-321. Springer, Berlin, Heidelberg, 2000.

[8] Augenstein, Isabelle, Sebastian Padó, and Sebastian Rudolph. "Lodifier: Generating linked data from unstructured text." In *Extended Semantic Web Conference*, pp. 210-224. Springer, Berlin, Heidelberg, 2012.

[9] Veena, G., Deepa Gupta, S. Lakshmi, and Jeenu T. Jacob. "Named Entity Recognition in Text Documents Using a Modified Conditional Random Field." In *Recent Findings in Intelligent Computing Techniques*, pp. 31-41. Springer, Singapore, 2018.

[10] Cp, Prathibhamol, Sankaran Narayanan, Vamsi Sai Kommuri, Sethu Subramanian, Kamal Bijlani, Savinay Nagendra, and Nikhil Podila. "ICACCI-02 (A): Artificial Intelligence and Machine Learning/Data Engineering/Biocomputing (Regular Papers)."