Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Extraction of Agricultural Elements Using Unsupervised Learning

[1]Nitika Sinha, [2]Aakash Rathod, [3]Pranay Gupta, [4]Pradnya Lanke, [5]Pankaja Alappanavar

[1, 2, 3 4] Sinhgad Academy of Engineering, Maharashtra, India

[5]Asst. Professor, Sinhgad Academy of Engineering, Maharashtra, India

Abstract- Efficient and accurate recognition of crop name, disease and cure from different articles is one of the major challenges for computers. We propose a system to extract and analyze data from agricultural corpuses. Due to the subjective nature of such articles it becomes difficult to extract relevant key events from such data. The aim is to extract crop name, disease and cure so as to form a coherent event that can be stored in a database to be used by researchers and agricultural practitioners. The idea is incorporated for Agriculture field, but it is applicable for all fields like medical, education, etc. Extracting data from an agricultural corpus would create a significantly large database of resources which would be valuable to both researchers and agricultural practitioners. Magazines and newspapers that publish articles tend to be subjective in their writing hence it becomes difficult to analyze, objectify and classify relevant information from them. An unsupervised approach would clearly extract data like crop name, disease, cure etc without taking to account each authors individual writing style.

Keywords- corpus, unsupervised machine learning, extractor, pattern extraction, EM algorithm.

I. INTRODUCTION

Nowadays internet is becoming the main source of information. But the information is dispersed. It would be useful if it is condensed and reusable. Most of the data present on internet is in textual format and users have to manually read to get the information they require but the proposed system reduces the human effort as it extracts the required information in the form of relation tuples which can be directly fed into further applications. In this project, we introduce a system which takes the agricultural corpuses as input and returns crop name, disease and cure in a tabular form as output. The domains involved here are Machine Learning and Natural Language Processing.

A. Unsupervised Machine Learning

Machine learning is a subfield of artificial intelligence that was derived from the study of pattern recognition and computational learning. Machine learning helps to develop programs that enable the system to learn and make predictions on data. Such algorithms operate by building a model from sample inputs to make data-driven predictions or decisions, instead of following strictly static program instructions.

In machine learning, the main concern of unsupervised learning is to find the hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or predefined instructions to obtain a proper solution. This differentiates unsupervised learning from supervised learning. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.

B. Pattern Recognition

Pattern recognition is a branch of machine learning which focuses on the recognition of patterns in data, although in some cases it is considered to be nearly same as machine learning. The training of pattern recognition systems is done from labelled “training” data (supervised learning), but when labelled data is unavailable other algorithms can be used to detect the unknown patterns (unsupervised learning). An object can be represented with a pattern. Depending on the nature of the patterns, pattern recognition can be classified into two types:

- Recognition of concrete items.
- Recognition of abstract items.

When a person gets a pattern, he deduces a logical conclusion and maps this conclusion with some general concepts or clues from the past experience. Pattern recognition problems may be logically divided into two categories:

- Capability of human beings to recognise patterns.
- Design and development of the systems which are capable of performing the task of recognition can be done using various theories and techniques of pattern recognition.

C. Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language, spoken or written. NLP is a component of artificial intelligence (AI). Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Modern NLP algorithms are based on machine learning, especially statistical machine learning. Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take as input a large set of features that are generated from the input data.

D. Extraction and Classification

Supervised Machine Learning Techniques attract the most attention of NER researchers because of their advantages such as high performance and domain neutrality (ease of adapting to different domains). Major supervised ML techniques that have been used for NER are: Support Vector Machines, Conditional Random Fields, Hidden Markov Models and Maximum Entropy. Most of the researchers in domain specific NER are focused on the biology and medicine domains describes a method for extracting parts of objects from wholes in a large corpus using hand-crafted patterns. The FARTUS system uses a series of hand crafted finite-state transducers to perform the NER task.

An algorithm to identify new extraction rules and A classifier based on point-wise mutual information for validating the extracted entities. We have extracted articles pertaining to agricultural corpuses using open source software called crawler4j that extracts related articles based on the user's input URL and target source. We then used these articles and manually parse them to recognize the patterns that are most commonly observed with the crop name and the disease. We eventually hope to extract more entities in future but for now we are limiting our scope to crop name, disease and its cure. We then used Stanford NLP toolkits for parsing and POS tagging.

II. LITERATURE SURVEY

A. EM Algorithm

Expectation-Maximization (EM) is an iterative technique used to find maximum likelihood estimates of parameters in probabilistic models, where the model rely on unobserved, also called latent variables. EM is a cyclic technique between performing the expectation (E) step, computing an expectation of the likelihood by counting the latest variables as observed data, and maximization (M) step, computing the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E steps. The parameters found on the M step are so used to start another E step, and the process iterates until some criterion is satisfied. EM is frequently used for data clustering like for example in Gaussian mixtures.

In the Expectation step, find the expected values of the latent variables.

In Maximization step, firstly plug in the expected values of the latent variables in the log-likelihood of the augmented data. Then maximize the log-likelihood in order to reevaluate the parameters.

B. Crime Analysis Using Self Learning

In Reference [1] EM algorithm is used to train the data and find correlated entities. This is in order to extract details from a crime related article. This system fetches the victim, place of crime and the police station where this act of crime was reported.

C. Extracting Relations From Internet

The system in Reference [2] focuses on extracting relations for a particular data type from the articles over internet. For this purpose a system called DIPRE (Dual Iterative Pattern Relation Extractor) is proposed. This system works on the duality between the rations and patterns.

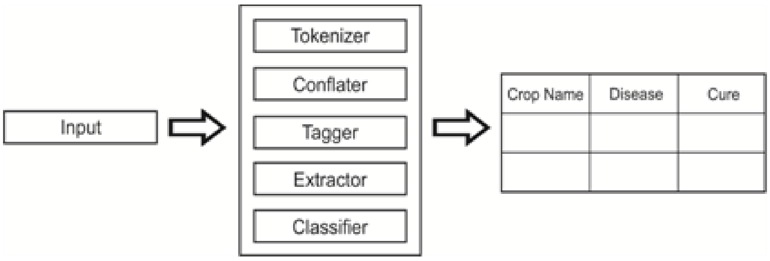
D. Seed: A Framework For Extracting Social Events From Press News

In Reference [3] the proposed system SEED (Social Entertainment Event Detection) aims to discover social events from the press news. The system extracts DATE, LOCATION, PLACE and ARTIST from the news articles. The method is divided into two steps. First step is to recognize four classes from the press news. NER is used for this purpose. Next step is to extract ternary relationship between the entities. For this purpose Relation Extractor (RE) is used. Provided enough resources true social events can be discovered.

III. PROPOSED SYSTEM

In the proposed system we aim to build software that can extract entities from data provided through the features that we give through some seed examples. This system is unsupervised which implies that that it has to recognize named entities itself based only on seed examples and features provided in the data. Named entities, both generic (e.g., names of persons, locations, organizations, dates, email addresses) as well as domain specific (e.g., genes, enzymes, proteins, cells, organs, diseases) are important content-carrying units within most documents. The task of named entity recognition (NER) is identifying all the occurrences of a given named entity type (e.g., ORG) in the given document.

NER is a complicated task due to various factors such as the variable nature of documents, semantics based and language dependent aspects of the task. Consequently, a large number of approaches have been designed for performing NER. One simple approach to NER is to use a gazette or list; e.g., a gazette for the named entity DEGREE would contain names of all known educational degree, such as Bachelor of Arts, B.A., Ph.D. etc. The gazette-based approach results into fast and more precise named entity recognition, since one simply looks for occurrences of any entries in the gazette though occasionally one needs some post-processing to differentiate between an occurrence of London as PERSON(e.g., Jack London) or as CITY. But the accuracy (recall) of the gazette-based approach is critically dependent on the completeness of the gazette since there may be thousands (or even millions) of examples for the named entity. One possible solution is to design the gazette (for a particular named entity) automatically from the large unlabeled corpus of text documents accessible either inside the organization or over the Internet. The problem then is how to do this automatically for sentences spanning different documents or even sentences within the same document but in separate sentences.



System Architecture

Efficient and accurate recognition of crop name, disease and cure from different articles is one of the major challenges for computers. We are introducing a system for this problem which takes the articles as input and returns crop name, disease and cure as output using Machine Learning and Natural language Processing techniques. The whole process consists of three steps:

1) Tokenization
2) Tagging
3) Extraction and classification.

This extracted output will be helpful for research works and advancement in agricultural field.

IV. CONCLUSIONS

Thus the system reduces human effort as it extracts the useful information from textual data in form of ordered pairs which can be directly used for further applications. The future scope is to make a generalized algorithm that can be applied over a wide variety of applications by just changing the inputs and to extend the system for different languages.

REFERENCES

[1] Extracting Patterns and Relations from the World Wide Web. S. Brin[Stanford University]

[2] Crime Analysis using Self Learning. Pranav Ruke, Stephy Mathew, Meghna Mohanty, Pankaja Alappanavar, Gandhali Gurjar[Sinhgad Academy of Engineering]

[3] SEED: A framework for extracting social events from press news. Salvatore Orlando, Francesco Pizzolon, Gabriele Tolomoi.