**Imperial Journal of Interdisciplinary Research (IJIR)**
**Vol-2, Issue-6, 2016**
**ISSN: 2454-1362,** http://www.onlinejournal.in

# Extraction of Agricultural Elements using Unsupervised Learning.

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

[1,2,3,4,5,] Department of Information Technology, Sinhgad Academy of Engineering, Kondhwa B.K., Pune -48

***Abstract****: Recognition of crop name, disease and cure efficiently and accurately from different articles is one of the foremost challenges for computers. The proposed a system extracts and analyze data from agricultural corpuses. The subjective nature of such articles makes it difficult to extract significant key events from such data. The system aims to extract crop name, disease and cure in order to form a logical event that can be stored in a database to be used by researchers and agricultural practitioners. The idea is integrated for Agriculture field, but it can be applied to all fields like medical, science, crime, education, etc. Extracting data from an agricultural corpus requires significantly large database of resources which would be important to both researchers and agricultural practitioners. Articles published in newspapers and magazines are mostly subjective writing hence it tends to be difficult to analyze, objectify and classify relevant information from them. An unsupervised approach would extract data like crop name, disease, cure etc. without the concerning writing style of individual authors.*

## Introduction

Nowadays internet can serve us information related to any topic. The information provided by internet is vast. The main difficulty faced by the users is to fetch relevant information. Usually users have to go through several links provided by the search engine to get the required information.

In this project we aim to develop a system which would extract crop name, diseases and cure from given agricultural corpuses thus easing up the examining burden of going through them repeatedly. This system can be used by agricultural practitioners, researchers, students and others related to this field. Agricultural articles will be given to the system as input. The system will then extract the crop name, disease and cure from the articles and classify them into respective classes.

### 1. Literature Survey

The system mentioned in reference [1] aims at training the data and finding the correlated entities. The system fetches name of victim, place of crime and name of the police station where the act of crime was reported from the crime related articles.

The system mentioned in Reference [2] aims at extracting relations for a specific data type from the articles over internet. In order to achieve this a system called DIPRE (Dual Iterative Pattern Relation Extractor) is proposed. This system works on the duality between the rations and patterns.

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

Reference [4] cites a semi supervised relationship extractor. EM Algorithm is used to train the system.

### 2. Proposed System

Usually users have to go through different articles present over the world wide web and then analyze. This process becomes time taking. So we propose a system which aims to build software that can extract entities from data provided through the characteristics that we give through some seed examples. This system is unsupervised which means that that it has to identify named entities itself based only on seed examples and features provided in the data. Named entities, both domain specific (e.g., genes, enzymes, cells, proteins, organs, diseases) as well as generic (e.g., names of persons, organizations, locations, dates, email addresses) are

important content-carrying units within most documents.

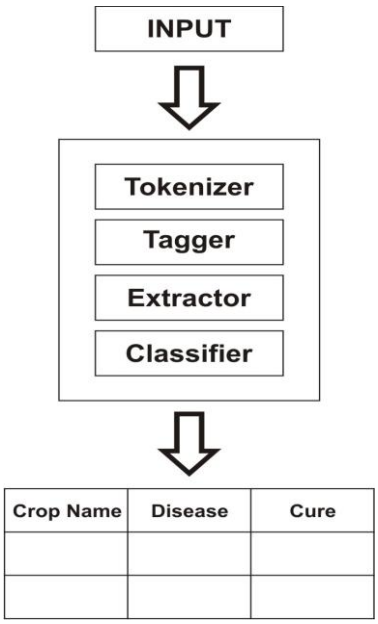


**Figure. 1 Proposed systems Architecture**

Recognizing crop name, disease and cure from the articles efficiently is one of the major challenges for computers. So we introduce 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. This process consists three steps: 1) Tokenization 2) Tagging 3) Extraction and classification.

### 3. EM Algorithm

Expectation Maximization (EM) algorithm [4] is an iterative method to find maximum likelihood of statistical models where equations cannot be solved directly. We propose to use EM algorithm for this relationship extraction problem.

The pair of named entity types (E1E2) and sentence features (F) are modeled as observed variables, whereas relations are modeled as hidden variables.

Let the number of instances in the dataset D be N and ith instance vectors are as ($x_i$ ,$F_i$, $z_i$)
where-

**Observed variable**, xi = ($x_{i1}$,$x_{i2}$……$x_{ik}$)
where K is the number of distinct pairs of entity types and $x_{ik}$ = 1 if $k^{th}$ pair of the
entity types is present in the $i^{th}$ instance and
0 otherwise.

**Observed variable**, $F_i$ = ($F_{i1}$,$F_{i2}$……$F_{iL}$)
where L is the number of distinct binary features characterizing the sentence and the entities in it and $F_{il}$ = 1
if the $l^{th}$ feature is present in the $i^{th}$ instance
and 0 otherwise.

**Hidden variable**, $z_i$ = ($z_{i1}$,$z_{i2}$,…..$z_{iM}$)
where M is the number of possible relations including a null relation and $z_{ij}$= 1 if the $j^{th}$ relation is present in the $i^{th}$ instance and 0 otherwise.

Let's assume that the instances in D are distributed identically and independently with the underlying Parameters Θ , the data likelihood can be expressed as,

$$L(D;\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[Pr(\vec{x_i},\vec{F_i},z_{ij}=1)\right]^{z_{ij}} \quad (1)$$

$$L(D;\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[Pr(\vec{F_i})Pr(z_{ij}=1|\vec{F_i}) \prod_{k=1}^{K} Pr(x_{ik}=1|z_{ij}=1,\vec{F_i})^{x_{ik}}\right]^{z_{ij}} \quad (2)$$

$Pr(z_{ij} = 1|\ F_i) = Pr(j^{th}\text{relation}|\ F_i)$ can be
modeled by a log-linear model using feature functions based on $F_i$. Similarly,
$Pr(x_{ik} = 1|z_{ij} = 1; F_i) = Pr(k^{th}$ E1E2_pair| $j^{th}$ relation; Fi) can be modeled by M log-linear models(one for each relation) using the same feature functions.
The feature functions (for the $i^{th}$ instance) are defined as the combinations of the features and the class labels as follows:
$f_{jl}(i; c) = F_{il}$, if j = c &
$f_{jl}(i; c) = 0$, if j 6= c
Using above definition for the feature functions,

$$Pr(Rel = j|\vec{F_i}) = \frac{\exp(\sum_{j'=1}^{M}\sum_{l=1}^{L}\lambda_{j'l}f_{j'l}(i,j))}{\sum_{j''=1}^{M}\exp(\sum_{j'=1}^{M}\sum_{l=1}^{L}\lambda_{j'l}f_{j'l}(i,j''))} \quad (3)$$

$$\text{and } Pr(x_{ik}=1|z_{ij}=1,\vec{F_i}) = \frac{\exp(\sum_{k'=1}^{K}\sum_{l=1}^{L}\alpha_{jk'l}f_{k'l}(i,k))}{\sum_{k''=1}^{K}\exp(\sum_{k'=1}^{K}\sum_{l=1}^{L}\alpha_{jk'l}f_{k'l}(i,k''))} \quad (4)$$

Therefore, the final expression for the data log likelihood is:

LL(D;Θ) =

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}\log(Pr(\vec{F_i})) + \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}\left(\sum_{j'=1}^{M}\sum_{l=1}^{L}\lambda_{j'l}f_{j'l}(i,j) - \log\sum_{j''=1}^{M}\exp(\sum_{j'=1}^{M}\sum_{l=1}^{L}\lambda_{j'l}f_{j'l}(i,j''))\right) + \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}\left(\sum_{k=1}^{K} x_{ik}\left(\sum_{k'=1}^{K}\sum_{l=1}^{L}\alpha_{jk'l}f_{k'l}(i,k) - \log\sum_{k''=1}^{K}\exp(\sum_{k'=1}^{K}\sum_{l=1}^{L}\alpha_{jk'l}f_{k'l}(i,k''))\right)\right) \quad (5)$$

Values of the hidden variables ($z_i$'s) and the parameters

(Θ : λ's and α's) can be estimated by the

EM algorithm. The parameter values are initialized in some way and the following EM steps are repeated till the log-likelihood is combined.

- **E Step:**

$$E(z_{ij}) = Pr(z_{ij} = 1|\vec{x_i}, \vec{F_i}) \quad (6)$$

$$E(z_{ij}) = \frac{Pr(z_{ij} = 1, \vec{x_i}, \vec{F_i})}{Pr(\vec{x_i}, \vec{F_i})} \quad (7)$$

$$E(z_{ij}) = \frac{Pr(z_{ij} = 1, \vec{x_i}, \vec{F_i})}{\sum_{j'=1}^{M} Pr(z_{ij'} = 1, \vec{x_i}, \vec{F_i})}, \forall_{i,j} \quad (8)$$

E(zij) can be calculated using the equations 2, 3 and 4 and using the current values of the parameters λ's and α's.

- **M Step:**

In this step, the data log-likelihood LL(D; Θ) is maximized using the current values of the hidden variables $z_i$'s.

### 4. Implementation Methodology

**Input:** Agricultural corpuses are the input to the system.

**Steps:**
Following steps are taken to generate output:

#### a) Tokenization:

Tokenization is the process of breaking a sentence into words called tokens. This list of tokens becomes input for next step.
For example:

Potato suffered from PotatoBlight

Tokenized Output is,

| | | | |
|---|---|---|---|
| Potato | suffered | from | PotatoBlight |

#### b) Tagging:

Part of speech tagging (POS tagging), also known as grammatical tagging is the process to mark a word in a text (corpus/article) corresponding to a particular part of speech, based on its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

The output of the tokenization process is input to tagging, so continuing the example

Tagged Output is,

Potato_NNP suffered_VBD from_IN PotatoBlight_NN

#### c) Extraction

The task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents is called Information extraction (IE). IE deals with parsing human language texts using techniques of natural language processing (NLP).

In simplified language, Extraction means automated or human-assisted acquisition of relations between concepts of textual or other data. In this step the machine finds and understands the limited relevant part of text. Then the machine gathers the information from many pieces of text.

Subtasks included in the process of extraction are:

**Named entity extraction:**
Named entity extraction is the process of extracting named entities of interest from an article/corpus. Recognition of known entity names such as PERSON, PLACE, ORGANIZATION, DISEASE, CROP etc. is done using named entity extraction. For example, consider the sentence "Potato suffered from PotatoBlight", the extracted named entities are Potato_CROP and PotatoBlight_DISEASE.

**Relationship extraction**
Relation Extraction is an important task in Information Extraction, after the extraction of name entities is done, next step is relation extraction which goes one step further and extracts the entities along with the relations between them. For example, in the sentence "Potato suffered from PotatoBlight" along with the named entities i.e. Potato_CROP and

PotatoBlight_DISEASE, relation extraction system is expected to identify that these two entities are related and recognize the relation type as "Affects". Table 1 [4] shows different types of relations.

**Table 1: Different types of relations**

| Type | Description |
|---|---|
| *Role* | Indicates the role a person plays at an organization. e.g. member, founder, citizen-of etc. |
| *At* | Represents location relationships like based-in, residence or located-at. |
| *Part* | Indicates part-whole relationships like part-of, subsidiary etc. |
| *Near* | Identifies relative locations. |
| *Social* | Represents various social or professional relationships between two persons like mother, sister, spouse, secretary etc. |
| *Affects* | This is specific to the agriculture domain and represents the relationship between DISEASE and CROP named entities |
| *Null* | We consider this additional relation which indicates that the two entities are not related. |

**d) Classification:**

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. The machine classifies the words into proper categories. We use EM algorithm, which uses few relation labeled seed examples and a large number of unlabeled examples to give the classified output.

**Output:**
Named entities i.e. Crop name, disease and cure are the output of the system.

## 5. Result

About 76% of the untagged articles/corpuses based on agricultural corpus are correctly tagged by our system. As we are using generic algorithm that can be applied over to a wide variety of data by simply manipulating the seed tuples, it can have a lot of applications in different fields as it allows computer to understand and use large amount of data. This it can be used on data sets relating to politics, sports or general crime to be used to predict unary and binary relations among entities occurring in such articles.

## 6. Conclusion and Future Work

In this paper, we have therefore proposed a system which uses unsupervised machine learning algorithm i.e.., EM Algorithm. Crop name, disease and cure can be classified into respective classes from the agricultural corpuses.

Our system is currently focused on agriculture field. In the future, the system can be trained to work in different arenas. Different languages can be integrated further into the system. The efficiency of the algorithm can be increased by using collaborative methods.

## 7. References

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

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

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

[4] Semi-supervised Relation Extraction using EM Algorithm. Sachin Pawar, Pushpak Bhattacharyya [Computer Science and Engineering, Indian Institute of Technology, Bombay]
Girish Keshav Palshikar [Systems Research Lab
Tata Consultancy Services Ltd., Pune]