

**NAMED ENTITY EXTRACTION
IN AGRICULTURE DOMAIN
TARGETING PLANT PRODUCT NAMES**

*Dissertation submitted to the Jawaharlal Nehru University
in partial fulfillment of the requirements
for the award of the degree of*

**MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND TECHNOLOGY**

**By
ASHISH KUMAR**

**Under the supervision of
Dr. ADITI SHARAN**



**SCHOOL OF COMPUTER AND SYSTEMS SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067, INDIA
JULY 2015**



जवाहरलाल नॅहरू विश्वविद्यालय

SCHOOL OF COMPUTER & SYSTEMS SCIENCES

JAWAHAR LAL NEHRU UNIVERSITY

NEW DELHI-110067, INDIA

DECLARATION

I hereby declare that the dissertation entitled “**Named Entity Extraction in Agriculture Domain Targeting Plant Product Names**”, submitted by me to the School of Computer and Systems Sciences, **Jawaharlal Nehru University, New Delhi**, in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science and Technology**, is a bona fide work carried out by me under the supervision of **Dr. Aditi Sharan**.

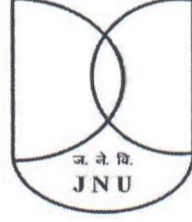
The matter embodied in this dissertation has not been submitted to any other university or institution for the award of any other degree or diploma.

Ashish Kumar

M.TECH-CSE

SC&SS, JNU,

New Delhi-110067



जवाहरलाल नॅहरू विश्वविद्यालय

SCHOOL OF COMPUTER & SYSTEMS SCIENCES
JAWAHAR LAL NEHRU UNIVERSITY
NEW DELHI-110067, INDIA

CERTIFICATE

This is to certify that the dissertation entitled “Named Entity Extraction in Agriculture Domain Targeting Plant Product Names”, submitted by Ashish Kumar to the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Science and Technology, is a bona fide work carried out by him under my supervision.

The matter embodied in this dissertation has not been submitted to any other university or institution for the award of any other degree or diploma.

Dr. Aditi Sharan

(Supervisor)

SC&SS, JNU, New Delhi

Dr. C.P. Katti

(Dean)

SC&SS, JNU, New Delhi
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi-110067

To My Loving Family and Friends...

Acknowledgement

I would like to gratefully acknowledge the enthusiastic supervision of Dr. Aditi Sharan during this work. This work wouldn't have been possible without her constant support, valuable suggestions and comments during my whole tenure of this dissertation work. I feel privileged to work under her for my master's dissertation. Apart from the academic guidance she has always been a great mentor of mine in encouraging me to be disciplined and well organized. I must surely say, she has given her best in providing me the infrastructure required, which led to the successful completion of my dissertation. I would take this opportunity to thank her once again for her esteemed support and I, from the bottom of my heart would like to wish her the best in all her future endeavors.

I wish to thank my colleagues Miss Deena and my senior Mr. Jagendra Singh, Mr. Mayank Saini, Mr. Rakesh Kumar, Mr. Chandra Shekhar Yadav and special thanks to Miss. Payal Biswas for creating a home like environment in our lab to keep the stress away. I would also like to thank my friend Mr. Devki Nandan and others for suggesting to remove errors in my dissertation. Thank you guys!!

Where would I be without my family? My parents deserve special mention for their inseparable support and prayers. My Father, Mr. Kishori Lal, in the first place is the person who put the fundamentals of my learning character, showing me the joy of intellectual pursuit ever since I was a child. My Mother, Mrs. Yogendri Devi, is the one who sincerely raised me with her caring and gently love. Ashwani and Abhishek thanks for being supportive and caring siblings.

Finally, I would like to thank the whole faculty members of our department for clarifying my doubts throughout this work. Last but not least, thanks are due to the JNU administration for creating such a secular and healthy environment amongst the students.

Abstract

With the advancement of high-speed network technology and the popularization of the Internet, the amount of accessible data is growing exponentially leading to information overload. With this growth of data, it is difficult to extract and manage useful information. Information extraction captures and outputs factual information contained within a document. It is recognised that a fundamental task in Information Extraction is Named Entity Recognition (NER), the goals of which are identifying named entities in unstructured documents, and classifying them into pre-defined semantic categories.

In the first phase of NER main aim was to extract named entity from unstructured text like newspapers and the entity classes were name (name of person, name of organization, name of location), time(date, year) and numbers(money, percentage, quantity). But later on these classes were extended to distance, city, country, river etc.

Initially researches were focused for general domain, but with the time researchers have shown their interest on different domains like Biomedical, Tourism and also on different languages like French, Greek, Italian and Hindi. Thus domain specific name entity extraction is emerging as fertile area of research.

A lot of data is available in agriculture domain which needs to be properly searched, processed and managed. It becomes a big challenge to handle such a large amount of data and extract useful information from this data. Extraction of agriculture entities become a prerequisite for many application such as question answering system, machine translation and entity tagger etc. However till now no entity extractor is available for extracting agricultural named entities. There can be large number of agricultural entities such as cereals and crops, pest and pesticides, fertilizers, plant diseases etc. Obviously it is not possible to target the extraction of all these entities. As a preliminary attempt, we have tried to develop an entity extractor for agriculture domain that will classify cereal and crop names from the agriculture corpus.

We have proposed a system that uses the context of words to extract the named entities. With the help of available agriculture corpus and few seed entities we were able to extract most of the new cereals and crops. Our approach uses context pattern induction based

domain specific approach, but this approach can be easily adaptable for other domain also.

Table of Contents

Declaration	i
Certificate	Error! Bookmark not defined.
Acknowledgement.....	iv
Abstract	v
Table of Contents	vii
List of Figures	ix
List of Tables.....	x
List of Abbreviations.....	xi
Chapter 1.....	1
Introduction	
1.1 Named Entity Extraction	1
1.1.1 Identifying Domain Specific Entities.....	2
1.1.2 Identifying Domain Specific Entities.....	2
1.2 Need of Named Entity in Agriculture Domain.....	3
1.2.1 Need	3
1.2.2 Background and Motivation	4
1.3 Issues in Identifying Named Entities in Agriculture Domain	6
Chapter 2.....	8
Related Work	
Chapter 3.....	12
Approaches and Features for Developing NER	
3.1 Approach for NER.....	12
3.2 Feature Selection	14
3.2.1 Word-level features.....	15
3.2.2 List lookup features.....	16
3.2.3 Document and corpus features	17
Chapter 4.....	19
Proposed Work	
4.1 Notion of Semantic Matrix (Motivation).....	20
4.2 Proposed Modification	22
4.3 Proposed Approach.....	22
4.3.1 An Overview of Approach 1 (SVB):	22
4.3.2 An overview of Approach 2 (NV-SVB):	23

4.3.3 Description of Proposed Approaches.....	23
Chapter 5.....	30
Experimental Results and Analysis	
5.1 Corpus/ Dataset Selection and Preparation	30
5.2 Steps in Experiments	32
5.3 Analysis of Result.....	36
Chapter 6.....	41
Conclusion	
References	42
List of Publication	47

List of Figures

Figure 1.1: Output of Stanford NER	5
Figure 1.2: Output of Python NLTK Named Entity Recognizer	5
Figure 3.1: Approaches for named entity extraction	13
Figure 5.1: Snapshot of website: “ http://www.agriculturalproductsindia.com/ ”	31

List of Tables

Table 3.1: Subcategories of word-level features.....	15
Table 3.2: List of list lookup features.....	16
Table 3.3: List of documents and corpus features.....	17
Table 4.1: List of variables used in algorithms	24
Table 5.1: Example of Sequence List.....	32
Table 5.2: Example of Word List.....	33
Table 5.3: Example of Co-occurrence Matrix.....	33
Table 5.4: Example of Distance Matrix	34
Table 5.5: Output entities corresponds to seeds Barley and Rye	35
Table 5.6: Example of Co-occurrence Sub Matrix	35
Table 5.7: Confusion Matrix	36
Table 5.8: Result of experiment 1 using random seed selection	38
Table 5.9: Result of experiment 1 using manually seed selection	38
Table 5.10: Result of experiment 2 using random seed selection	39
Table 5.11: Result of experiment 2 using manual seed selection	39

List of Abbreviations

ABNER	A Biomedical Named Entity Recognizer
AGROVOC	Agricultural Vocabulary
BANNER	Named Entity Recognition System, Primarily Intended For Biomedical Text
BNER	Biomedical Named Entity Recognition
CRF	Conditional Random Field
DT	Decision Tree
GDP	Gross Domestic Product
HMM	Hidden Markov Model
IE	Information Extraction
IR	Information Retrieval
LBJ	Learning Based Java1
MEMM	Maximum Entropy Markov Model
MUC	Message Understanding Conference
NE	Named Entity
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification
NLTK	Natural Language Toolkit
NV-SVB	Semantic Vector based only for noun and verbs
POS	Part of Speech
SVM	Support Vector Machine
SVB	Semantic Vector based
VSM	Vector Space Model
XML	EXtensible Markup Language

Chapter 1

INTRODUCTION

The ‘information explosion’ has created the exceptionally large amount of published information which is still growing at an astonishing rate. The problem of information management becomes more challenging with the growing amount of information. A key to this challenge rests on the technology of Information Extraction, which automates the transformation of un-structured textual data into structured representation. The structured information can be easily interpreted and manipulated by machines. Named Entity Recognition is considered to be a fundamental task of Information Extraction. The goal of Named Entity Recognition is to identify references of named entities in unstructured documents, and classify them into pre-defined semantic categories. Since natural languages are polysemous, so name references are ambiguous. Resolving ambiguity concerns recognizing the true referent entity of a name reference, essentially a further named entity ‘recognition’ step and often a compulsory process required by tasks built on top of NER.

1.1 Named Entity Extraction

The term “Named Entity” (NE) is often used in Information Extraction (IE) applications. The term “Named Entity” was first coined in the sixth message understanding conference (MUC-6) (Grishman & Sundheim, 1996). The goal was to extract named entities such as people, organization or location names from news articles. Over the past years, the task of Named Entity Extraction in the newswire domain has attracted considerable amount of research and a number of successful systems such as LBJ (Rizzolo & Roth, 2010) with accuracies of over 90% have been developed.(Raja, Subramani & Natarajan, 2014)

Information Extraction and Text Mining systems have many components in which Named Entity Recognition is the most important. The task of NER is to find all proper noun phrases

(and other easily recognizable phrases) from a text and categorize them into a small predefined set of semantic classes such as names, locations, dates, organizations, drugs, diseases, books etc. While applying text mining techniques like extraction of relations from the text, semantic hierarchies and building ontologies, etc. as a preprocessing step, Name Entity Extraction is essential (Fresko, Rosenfeld & Feldman, 2005).

1.1.1 Identifying General Named Entities

There are predefined categories of named entity, in addition to this various opinions are there, based on which categories should be considered as named entity. These opinions also define the size of named entity. There are some common conventions, based on which entities are marked according to XML format as described in Message Understanding Conference (MUC).

"ENAMEX" tags are used for names, "NUMEX" tags are used for numerical entities, and "TIMEX" tags are used for temporal entities. Consider the following example:

“Jim bought 300 shares of Acme Corp. in 2006”

```
<ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX
TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX
TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX
TYPE="DATE">2006</TIMEX>.
```

Basic categories generally agreed upon include the following:

Names (ENAMEX) : Person, Organization, Location

Times (TIMEX) : Date, Time

Numbers (NUMEX) : Money, Percent, Quantity

However, State/Province, Country, City, Distance, Speed, Age, Weight, River, etc. may also be considered as categories/subcategories (AFNER - Named Entity Recognition, 2015).

1.1.2 Identifying Domain Specific Entities

The above discussed entities are usually considered in case of open or general domain. Depending on the project requirements, categories chosen for a particular Named Entity Recognition (NER) project may vary. If in a particular field biological classification is

important, then the biological terms may need to be more refined such as protein , gene etc. Similarly, for geographical classification, a particular type of location should be classified to location entity. Thus there is a need to identify named entities from a document belonging to a specific domain. Our aim is to identify some specific entities from Agriculture domain.

1.2 Need of Named Entity in Agriculture Domain

1.2.1 Need

Named entity extraction in agriculture domain has various applications such as extracting a specific piece of information from gigantic source of agriculture text, monitoring the frequency or tracking the occurrence of an agriculture entity in the agriculture related document. Agricultural Named Entity recognizer also helps in locating the entities in the sentence to find out the answer of a particular agriculture related question in Question Answering System.

For example if the given question is “Which crop is suitable for black soil ? ” Through Question Processing Module we will able to predict that since *Crop* is the *Head Word* of the question the answer will be of type *Crop*. Now in the text if there is a statement that “Maize is grown in Black soil .” and if the Named Entity Recognizer is able to recognize that *Maize* is a name of crop then after pattern matching the system can easily predict that *Maize* is the actual answer for which the question is looking for.

Similarly in case of Information Retrieval (IR), suppose the query given by the user is “Characteristics of Basmati Rice” then for efficient information retrieval query expansion will take place. But since the term Basmati Rice is the name of proper entity, it should be taken as it is and hence should not be expanded. For this purpose, it should be known to the Query Expansion module that Basmati Rice is a Named Entity.

In brief, if we want to extract any kind of knowledge from agriculture related text or want to do any kind of processing such as Summarization, Information Retrieval, Machine Translation etc. the basic key step is recognizing and extracting the agriculture related entities. Hence I realized that there is a need to develop a “ Named Entity Recognizer for Agriculture Domain” which as per my knowledge is the first step in this direction.

1.2.2 Background and Motivation

General purpose entities are well defined. However entities pertaining to a specific domain such as: medical, agriculture, geography and tourism etc. depend on the nature of the domain. Domain specific NER improves the efficiency of text mining applications in that domain. Our focus will be on extracting named entities in agriculture domain.

India is principally an agriculture based country. It is the backbone of Indian economy which contributes a significant figure approx. 13.7 % * to the Gross Domestic Product (GDP). The size of available agricultural information in electronic form is very large and the information repository is still increasing day by day at a very high rate. With this enormous amount of textual information in Agricultural domain, there is a need for effective text mining and knowledge discovery system that can help people working in the field of agriculture to gather information and make use of the knowledge encoded in text documents in an efficient manner. The most fundamental and key component for data mining and knowledge acquisition is Named Entity Recognition (NER).

The Named Entity Recognizer (NER) available till date, are either for open domain or specific domain like biomedical. Thus NER for agriculture domain becomes an interesting research problem. There are various Named Entity Recognizers are available now a days such as Stanford's Named Entity Recognizer, Python NLTK Named Entity Recognizer, Learning Based Java1 (LBJ) (Rizzolo & Roth, 2010) and many others. But since they are open domain NER, they can only able to recognize the name of Place, Person and Organization but unable to tag the Agricultural related entities.

The snapshot of the output of Stanford NER for agricultural related text are given in Figure 1.1.

We can notice that it has tagged Rice as Pers that is person Name rather than tagging as crop and not able to recognize *Oryza Sativa* which is a variety of Rice.

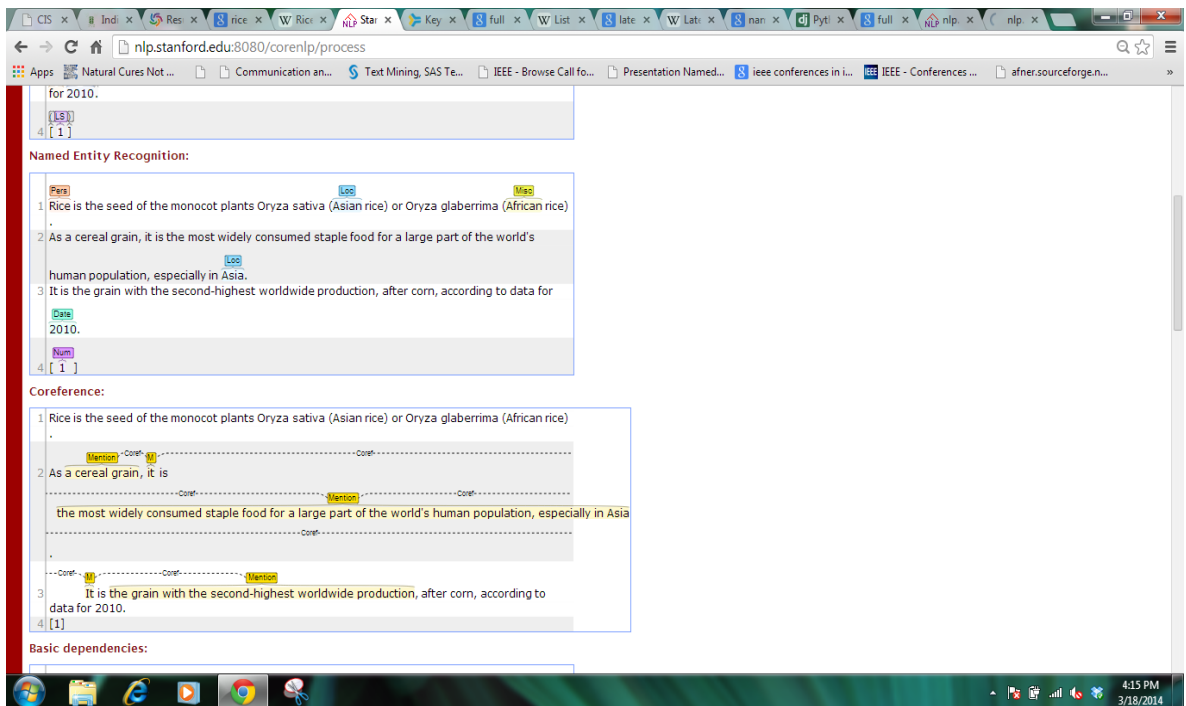


Figure 1.1: Output of Stanford NER

Similarly the output of Python NLTK Named Entity Recognizer which recognizes Rice as GPE that is Geo-political Entities and Oryza as person name rather than Rice name, is in Figure 1.2.

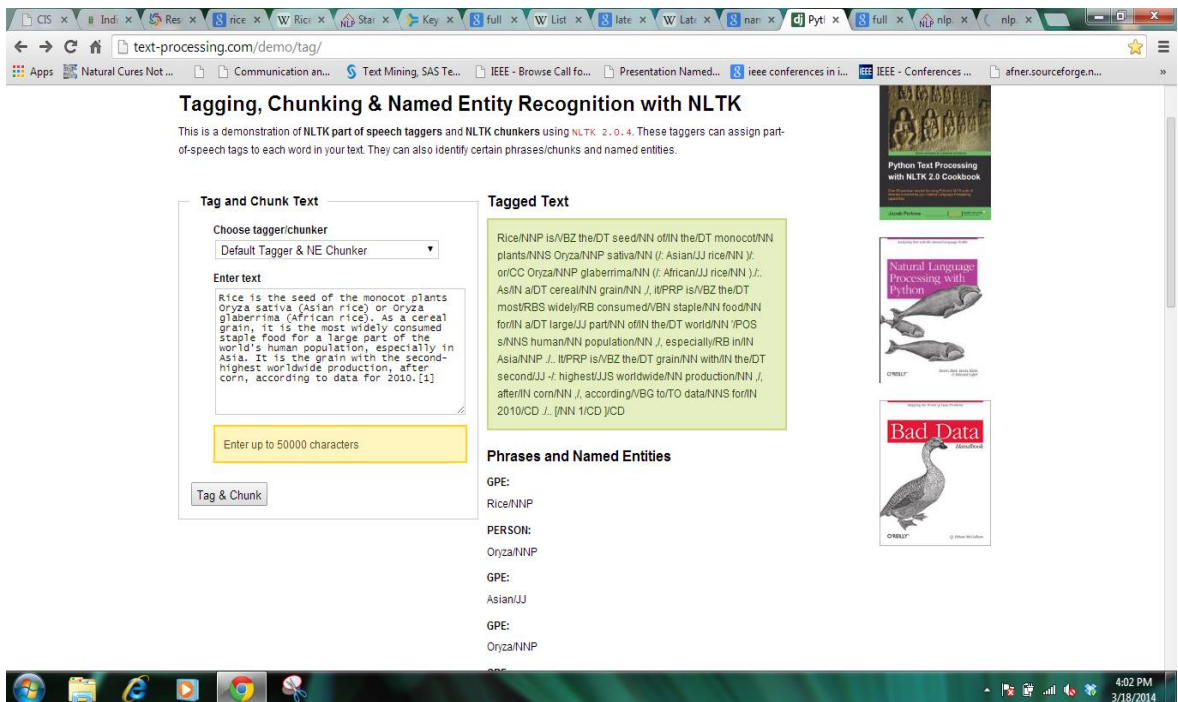


Figure 1.2: Output of Python NLTK Named Entity Recognizer

Ample amount of work in NER has been done with respect of open domain but very limited work has been done domain specific named entity extraction. Only Biomedical domain is the one which has been taken in account for domain specific NER. For last few years, after the availability of GENIA corpus, considerable amount of work has also been done in biomedical domain. Various Named Entity Recognizers for biomedical domain such as ABNER (A Biomedical Named Entity Recognizer), Biomedical Named Entity Recognition (BNER), BANNER (Named Entity Recognition System, Primarily Intended For Biomedical Text) and many others have been developed. No doubt they are working well, but since they had particularly developed for the biomedical domain they have been trained to recognize and tag the biomedical entities such as protein name, gene name , gene products etc. but unable to recognize the agricultural entities.

Agricultural domain is very much similar to the biomedical domain having its own specialized terminology and complex naming conventions but due to the unavailability of agricultural corpus and sufficient agricultural resources the task of NER in Agricultural domain has consistently lacked behind.

1.3 Issues in Identifying Named Entities in Agriculture Domain

As this is a preliminary attempt, we start with some very basic issues in identifying name entities in agriculture domain:

1: Identification of agricultural entities

How to justify, a particular term is agricultural entity or not is itself a challenge. That is defining a notion of what constitutes an agriculture entity itself is not clear . For example consumption, animal feed, etc.

2: Selecting a set of entities to be tagged in context of agriculture domain

There are different types of entities available in Agricultural domain but not all the entities are of our interest, so finding the interested entities to be tagged may be a challenging task.

3: Assigning proper tag which can be given generic or focused

The next step is to assign these terms with the agricultural related tags. Assigning the proper tag to agricultural term is also a challenging task. Because sometime we have to select

generic tag and some time we need to select focused tag. For example, watermelon can be tagged using two different tags, product or Fruit. Product is a generic tag while Fruit is a focused tag.

4: Using multiword entities and their boundary

What to do with the words which are composed of two or more tokens. Whether to take them as it is or treat the atomic terms as a single entity. For example, Pearl barley is composed using two words: Pearl and Barley. Pearl is a Fishery products and Barley is a Cereal whereas complete phrase pearl barley is Cereal products.

5: Unavailability of Benchmark Dataset

An ample amount of work has been done in the field of NER for open domain and biomedical domain. But agriculture domain has not been yet taken in consideration for NER perspective. Hence there is an unavailability of any authenticate or benchmark agricultural data set. Thus in order to do the work we have to start with creating our own dataset.

Considering all the above discussed problems, it is difficult to consider and identify all the named entities in Agriculture domain. Keeping this in mind the scope of our problem is confined to a specific set of entities, which can be objectively defined. There are many significant entities such as crop name, disease name, pest & pesticides and fertilizer and many more. All these entities plays a vital role for developing NER system for agriculture. However considering the focus of agriculture domain, crop names can be considered as set of entities with prime importance. Thus in this work we are focusing on extraction of crop names from agriculture related data.

The layout of this dissertation is as follows. Chapter 2 describes the related work in the field of named entity recognition. Chapter 3 gives a description of various approaches and features for developing NER. Chapter 4 contains the proposed work. Chapter 5 explains the experiments and evaluation of results. After that Chapter 6 concludes the work.

Chapter 2

RELATED WORK

Named Entity Recognition has grown as an important area of research in past two decades. Lisa F. Rau presented the first research paper in this area in 1991 at Seventh IEEE Conference on Artificial Intelligence Applications (Rau, 1991). The designed system can “extract and recognize names” incredibly the company’s name. Based on heuristics and handcrafted rules, the system can perform the recognition. The publication rate was relatively low during the period of 1991 to 1995. Then in 1996, after MUC-6, it has been accelerated and never been declined since then.

Besides NER in English, researchers are also pointing on the problems related to language independence and multilingualism e.g. CONLL-2003 uses German for study where MUC-6 conference and IREX conference uses Japanese for study. Abundant literature is available in Chinese (Wang, Li & Chang, 1992), (Chen & Lee, 1994) and (Yu, Bai & Wu, 1998). Similarly French is studied by (Petasis, Vichot, Wolinski, Paliouras, Karkaletsis & Spyropoulos, 2000) and (Poibeau, 2003), Greek by (Boutsis et al., 2000). and Italian by (Black, Rinaldi & Mowatt, 1995) and (Cucchiarelli & Velardi, 2001). Many other languages such as Korean (Whitelaw & Patrick, 2003), Polish (Piskorski, 2004), Romanian (Cucerzan & Yarowsky, 1999), Russian (Popov, Kirilov, Maynard & Manov, 2004), Danish (Bick, 2004), Hindi (May, Brunstein, Natarajan & Weischedel, 2003) have also received attention.

Researchers working in the field of NER have also focused over diverse genres and domains. (Maynard, Tablan, Ursu, Cunningham & Wilks, 2001) designed a system for emails, scientific texts and religious texts. (Minkov, Wang & Cohen, 2005) created a system specifically designed for email documents. Domain specific named entity extraction is emerging as a recent research area.

Initially NERC was used for finding the “proper names” particularly names of “persons”, “places” and “organizations” (Coates-Stephens, 1992; Thielen, 1995) but later on subcategories were created for each entity in a fine grained manner e.g. the entity type “location or place” can be separated into multiple subcategories such as country, state, city, etc. (Fleischman, 2001) and (Lee & Lee, 2005). Similarly, “person” can be fine-grained into sub-categories “doctor”, “engineer”, “politician”, etc. (Fleischman & Hovy, 2002).

Two new terms “timex” and “numex” were coined in 2003, regarding MUC for numeric quantities e.g. “date”, “time”, “money”, “percent”, etc. CONLL conference used “miscellaneous” type for including the proper names falling outside the classic types. For example “film”, “scientist” (Etzioni et al., 2005), “email address”, “phone number” (Witten, Bray, Mahoui & Teahan, 1999), (Maynard, Tablan, Ursu, Cunningham & Wilks 2001), “research area”, “project name” (Zhu, Uren & Motta, 2005), “book title” (Brin, 1999), “job title” (Cohen & Sarawagi, 2004), “brand” (Bick, 2004), etc.

All the previously discussed research works are particularly on open domains whereas if we proceed towards domain specific NER, the named entities will not remain same as name of Place, Person and Organization. The entities needed to be tuned according to the particular domain. For example in case of Biomedical Domain common entity names are gene name, gene product, protein name, etc. In last ten years, various researchers working in the field of NER are interested in Biomedical domain and an abundant amount of work has also been done e.g. (Settles, 2004), (Kazama, Makino, Ohta & Tsujii, 2001), (Lin et al., 2004), (Saha, Chatterji, Dandapat, Sarkar & Mitra, 2008). Among the entire available domain specific NER, Biomedical domain is little bit similar to Agriculture domain in some extent. Some of the prodigious works in Biomedical domain is discussed below:

(Lee, Hwang, Kim & Rim, 2004) presented a two-phase SVM based named entity recognizer, which consists of a two phases namely boundary identification phase and a semantic classification phase. It is used to resolve the multi-class problem and unbalanced class distribution problem by employing an ontology based hierarchical classification method, breaking the NE recognition task into two individual subtasks where for each subtask, they used appropriate SVM classifiers and relevant features. They were able to achieve 74.8 F-score for the boundary identification and 66.7 for semantic classification.

In biomedical domain, a named entity recognition system Power Bio NE has been presented by (Zhou, Zhang, Su, Shen & Tan, 2004). Various evidential features are proposed here for dealing with the special phenomena of naming conventions in the biomedical domain: word formation pattern; morphological pattern, such as prefix and suffix; part-of-speech; head noun trigger; special verb trigger and name alias feature. HMM-based named entity recognizer is used to integrate all the features effectively and efficiently. For resolving the data sparseness problem, they have proposed a k -Nearest Neighbour (k -NN) algorithm.

Unlike this (Seki & Mostafa, 2005) presented a hybrid approach without using any natural language processing tool (e.g. part of speech taggers, syntactic parsers, etc.) that completely relies on surface clues to reduce the overhead incurred during processing. For initial detection of names of protein, they used a set of simple heuristics and for locating complete protein names they use some probabilistic model.

Enormous work had been done in recognizing non-nested NEs, whereas nested NEs (one containing another) have been generally neglected for a long time. In recent years, Nested Named Entity has also become popular among various researchers as it represents important relations between various entities as well as it accounts for 16.7% of all named entities in GENIA corpus. Three techniques for modelling and recognizing nested entities were introduced by (Alex, Haddow & Grover 2007). The techniques introduced are namely layering, cascading, and joined label tagging. They also compared these techniques by means of a conventional sequence tagger. There is a difficulty with the Nested NER that the standard methods employed in conversion of NER to sequence tagging problem i.e. when each token has assigned a tag to indicate the beginning (B), inside (I), or outside (O) of an entity is not applied directly when token belongs to more than one entity. In this work, they have explored methods for reducing the nested NER problem to one or more BIO problems so that the existing NER tools can be easily applied.

The latest work in the field of NER has been done by (Tang, Cao, Wang, Chen & Xu, 2014), which exploits a large set of features for Named Entity Recognition. Systematically, they have investigated three different types of word representation (WR) features for BNER, namely clustering-based representation, distributional representation, and word embedding. They have improved the F -measure on the BioCreAtIvE II GM and JNLPBA corpora by combining all the three types of WR features. The increase in F -measure is 3.75% and 1.39%

for BioCreAtIvE II GM and JNLPBA corpora respectively while comparing with the systems using baseline features.

Chapter 3

APPROACHES AND FEATURES FOR DEVELOPING NER

Entity extraction can be considered as a classification problem in which words are assigned to one or more semantic classes. An entity extractor first identifies the significant entity words of a sentence and then classifies them into some predefined entity types. Entity Extraction is not a trivial task. Many approaches have been developed for extracting Named Entities. There are two important modules for developing NER Systems:

- Approach to be used
- Feature Selection

3.1 Approach for NER

Various approaches (Sasidhar, Yohan, Babu & Govardhan, 2011) which has been used for developing NER systems can broadly categorized into two classes:

A) Rule Based Approach

B) Machine learning based Approach.

Rule Based or Handcrafted Approach: Rule based approach includes Dictionary/Gazetteer Based Approach (List Lookup Approach), Linguistic Approach.

The task of identifying named entities using list seems easier. If we have a list of named entities, we can search this list and assign appropriate entity type. The advantages of List Lookup approach is that it is very simple, fast and language independent. However these approaches have following limitations –

- It only works for entities in the gazetteer.
- It is impossible to create a generalized list for all the possible NEs.
- It cannot resolve ambiguity and does not have learning capacity
- Efficient searching techniques are required for large set of data.

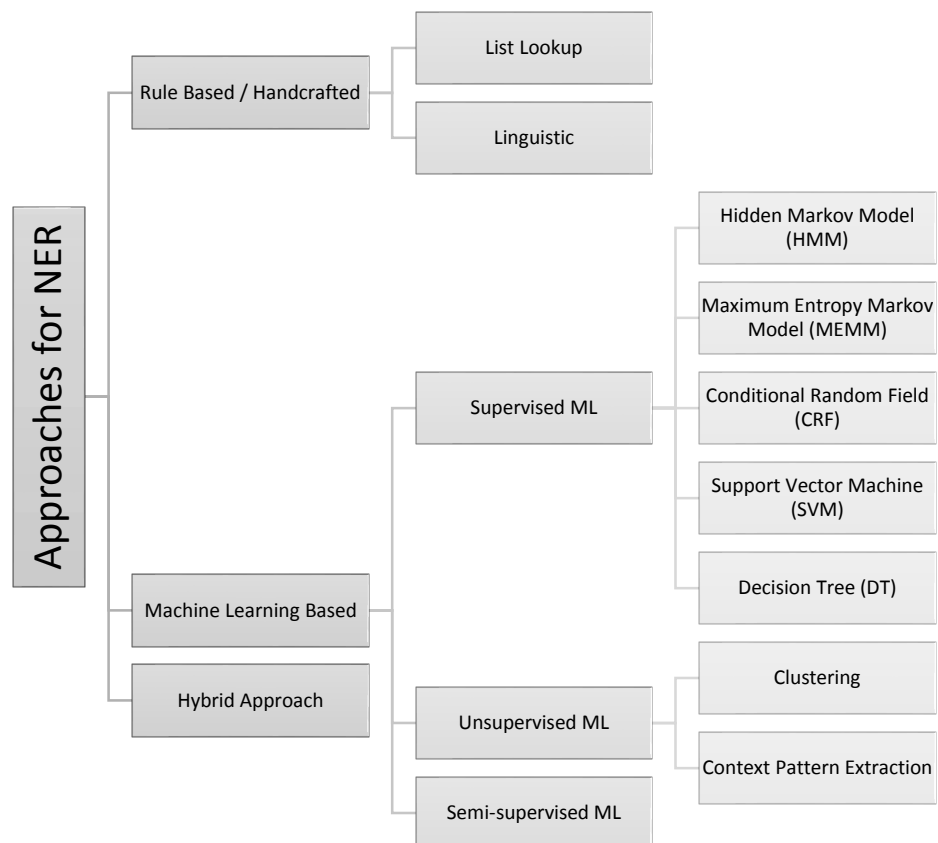


Figure 3.1: Approaches for named entity extraction

Linguistic Approach uses language rules to extract named entity. The rules are written manually by linguists. The advantage of this approach is that if rules used are rich and expressive then it gives very good results but this requires a lot of work by domain experts.

Following are the disadvantages of Linguistic Approaches -

- The development is generally time-consuming and sometimes it is hard to accommodate changes in the system.
- Rule based NER is domain specific i.e. NER made for a particular language or domain can be used only for that language or domain and not for other.

Rule Based or Handcrafted Approach usually accomplish better accuracy, but in order to prepare and maintain the extraction knowledge, it need huge amounts of skilled labor by domain experts and linguists.

Machine Learning Approach: Rule based approach has many limitations, because of it, recent research in NER is staring on machine learning techniques, which only rely upon a manually labeled training set of documents and learning algorithms. The learner is trained to learn the features of tagged entities and build a model using learnt features. Once the model is trained it can be used for extraction of new named entities. Some of the ML techniques that have been used for named entity recognition are: CRF (conditional Random Fields), MEM (Maximum Entropy Model), SVM (Support Vector Machine) and HMM (Hidden Markov Model).

Machine learning approach is easily adaptable and flexible in comparison to rule based approach. One can apply a machine learning approach of one domain to another domain with slight changes in the original approach. While a rule based approach cannot be applied in new domain without completely changing it. Machine based approach is much cheaper than rule based approach. However current machine-learning approaches capture important vindication behind NER problem much less effectively than human experts who handcraft the rules, even though machine-learning approaches always provide significant statistical information that is not achievable by human experts (Zhou & Su, 2002).

To apply any machine learning approach we need some important features. These features are used to train the machine learning based models so that it can produce good result.

Following section discuss in details some important features which can be used to develop the NER systems.

3.2 Feature Selection

Feature: To determine whether a particular word in the corpus represents an entity or not, we need some information associated with that word. This information is nothing but the feature of that word. Features are considered to be descriptors or characteristic attributes of words designed for algorithmic consumption.

Features can be easily represented in the form of a vector known as feature vector. Feature vector represents the information about the each feature presented in the word, this information can be stored as Boolean attribute, numeric attribute or nominal attribute. Suppose we are using 29 features then length of feature vector will be 29. A particular feature can be represented as:

+1 if it is presented in the word

-1 if it is not presented in the word

0 if we are not considering that feature of word (Baluja, Mittal & Sukthankar, 2000)

Text Based Features: Here, we present the most commonly used features for named entity recognition and classification. These features are organized along three different axes namely Word-level features, List lookup features and Document and corpus features as discussed below.

3.2.1 Word-level features

Word-level features represent the character configuration of the words. These features are used to describe word case, punctuation, numerical value and special characters.

Table 3.1: Subcategories of word-level features.

Features	Examples
Case	Starts with a capital letter Word is all uppercased The word is mixed case (e.g., ProSys, eBay)
Punctuation	Ends with period, has internal period (e.g., St., I.B.M.) Internal apostrophe, hyphen or ampersand (e.g., O'Connor)
Digit	Digit pattern Cardinal and Ordinal Roman number Word with digits (e.g., W3C, 3M)

Character	Possessive mark, first person pronoun Greek letters
Morphology	Prefix, suffix, singular version, stem Common ending
Part-of-speech	proper name, verb, noun, foreign word
Function	Alpha, non-alpha, n-gram lowercase, uppercase version pattern, summarized pattern token length, phrase length

3.2.2 List lookup features

Lists are the privileged features in NERC. The term “list” can also be known as “gazetteer”, “lexicon” or “dictionary”. The “is a” relation can be represented by using List inclusion (e.g., Paris is a city). It may appear obvious that if a word (Paris) is an element of a list of cities, then the probability of this word to be city, in a given text, is high. Because of polysemy property, the probabilities can almost never be 1 (e.g. the probability of “Fast” to represent a company is less because of the commonly used adjective “fast” which is more frequent).

Table 3.2: List of list lookup features.

Features	Examples
General list	General dictionary Stop words (function words) Capitalized nouns (e.g., January, Monday) Common abbreviations
List of entities	Organization, government, airline, educational First name, last name, celebrity Astral body, continent, country, state, city

List of entity cues	Typical words in organization Person title, name prefix, post-nominal letters Location typical word, cardinal point
----------------------------	---

3.2.3 Document and corpus features

Document features are defined over both document content and document structure. Large collection of documents (corpora) are also considered to be excellent source of features. In this section, we list features that go beyond the single word and multi-word expression and include meta-information about documents and corpus statistics. (Nadeau & Sekine, 2007)

Table 3.3: List of documents and corpus features.

Features	Examples
Multiple occurrences	Other entities in the context Uppercased and lowercased occurrences Anaphora, co-reference
Local syntax	Enumeration, apposition Position in sentence, in paragraph, and in document
Meta information	Uri, Email header, XML section, Bulleted/numbered lists, tables, figures
Corpus frequency	Word and phrase frequency Co-occurrences Multiword unit permanency

The features which we have discussed above in this section can be presented in any corpus. But generally it has been seen that most of the word level features are useful in open domain but when we move to specific domain word level feature are not so promising. In that case list lookup feature and document and corpus level can be useful. So Word level features are not helping in distinguishing because the entity names themselves are not very focused. Because of considering these limitation we have decided to use context level features

As we have seen we cannot use all features at a time. Most NER systems (or taggers) are harshly limited in the number of features they may think over, because the computational overhead of handling large numbers of features is expensive, and because the risk of overtraining expands with the number of features (Mayfield, McNamee & Piatko, 2003). Features are also language dependent like root information of the word can be used in morphologically rich languages because words can be recovered in various forms depending on its number, case, tense, etc. On the other hand capitalization information can be used generally in English language. So we can see that features are language dependent. Thus, the feature set must be finely superimposed to be effective.

Because of the unavailability of annotated corpus we have decided to use context level features such as Context words (frequent words for a particular class presented in a word window). Since we are working with agriculture corpus which is available in English language, so we can think POS (Part of Speech) as an feature for our problem. POS taggers can easily provide features for machine learning. The POS of the current word and the neighboring words may be useful feature for NER (Saha, Chatterji, Dandapat, Sarkar & Mitra, 2008).

Chapter 4

PROPOSED WORK

The Named Entity Recognizers (NER) available till date, are either for Open domain or for some specific domains like Biomedical. Open domain techniques do not cover domain specific entities. Therefore, separate NER are required for identifying entities in specific domains. A lot of textual data is available in Agriculture domain. However, it is unorganized, unstructured and hence not much usable. In order to utilize this information or if we want to do any kind of processing with this information, entity extraction may be a prerequisite for many applications such as: question answering system, machine translation and entity tagger etc. Different entities exist in Agriculture domain such as: plant names, cereals and crops, fertilizers, pest and pesticides, plant diseases, etc. and no entity extractor is available. Thus, NER for Agriculture Domain becomes an interesting research problem. As it is not possible to cover the possible entities to start with, we have targeted for extraction of cereals and crops names as these can be considered to be most basic and commonly used entity in Agriculture domain.

The objective of our work is to identify plant names from the textual data in agriculture domain. The entities of our interest are cereals and crops.

In our knowledge this is a primitive attempt to identify named entities in agriculture domain. No serious work has been done in this direction so we have to start from scratch.

We faced following problems in applying some well-known techniques for identifying named entities:

- Limitation of Word based features: We observed that the entities are of very general nature, which can get mixed up with common words. In other words, there are not very specific word based features that can be used to separate out plant names.

- The non-existence of a standard ontology/knowledgebase for identifying agriculture entities. Though AGROVOC (Multilingual Agricultural Thesaurus) is available as a thesaurus, however, it does not cover all the cereals and crops.
- Non availability of benchmark and labeled dataset.

Keeping in view that word based features cannot work well in our case, we switched over to the use of semantic vectors for identifying our entities of interest. Semantic vector captures the contextual information from the text. We observed that semantic vectors could play an important role in identifying named entities

Following are our specific objectives:

- 1: To utilize the semantic information for extracting entities
- 2: To propose a framework for developing a system for extracting cereal and crop names from agriculture corpus using semantic vector.
- 3: To implement the proposed framework

4.1 Notion of Semantic Matrix (Motivation)

Our work is inspired by the concept of distributional hypothesis that says: words that hang together in similar contexts lean towards having similar meanings (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990). It has been observed that similar type of entities shares similar context. So it can be assumed that this context helps us in extracting named entities.

Context Pattern: A pattern is an arrangement or sequence regularly found in comparable objects or events. Context pattern is a pattern enclosing certain context. The context of a particular word can be viewed as some preceding and some following words of the given word. While extracting the named entities, these context words can play significant role.

The context of a word is useful for extracting named entities but the difficulty is how to capture that context. There are many methods suggested for capturing the context but most of them have high computational complexity. Most famous method for context extraction is window method. In this method window represents a span of words, and it is passed over

the whole corpus. This window contains certain number of words and this number can be varied.

Semantic similarity allows to represent contextual information in the form of matrix and hence it reduces the complexity of processing the context-based information. The notion of semantic vector is based on the Vector Space Model (VSM) of semantic. So it is easy to apply the vector based model to identify entities.

As the semantic vector is created at backend, most of efforts done are the backend. The semantic information is captured in the form of a co-occurrence matrix. Due to this structure, the computational efficiency of algorithm increases.

Our description of co-occurrence matrix is based on the paper suggested by (Lund & Burgess, 1996).

Co-occurrence matrix : For a given document, word-by-word co-occurrence frequency matrix is generated by using a sliding window of fixed size: all the words present within the window are considered as co-occurring with each other. When the window is slided across the document, an aggregated co-occurrence matrix is produced in a definite vocabulary for all the words. The strength of association between two words is inversely proportional to their distance. This matrix is direction sensitive: The co-occurrence information preceding and following a word are write down separately by the row and column vectors. This order information appears to be very interesting (Chen & Lu, 2011).

It is obvious that many words do not occur together, so the matrix is very sparse (Konkol, Brychcín & Konopík, 2015).

Similarity between the words is often represented by similarity between context associated with the two words. In this case, we are focusing on the crops name and according to the concept of semantic similarity, context of two crops name will be more similar rather than any two arbitrary words. Context based similarity between two words can be calculated using the semantic vectors of the words.

Distances between word vectors were examined to determine whether or not similarity in word meaning corresponded to similarity in patterns of vector elements by LUND AND

BURGESS. They used Euclidean distance between vectors to find the similarity between these vectors.

As the distance vector is used, similarity is inverse of distance.

4.2 Proposed Modification

We found that approach suggested by (Lund & Burgess, 1996) is only to find out the words that share aspects of meaning. However, our main motivation is to find out the named entities in agriculture domain. Above approach as such is not very useful for us, therefore we suggest some modifications.

We suggest some changes in the construction of co-occurrence matrix and further we use cosine based distance measure instead of Euclidean distance that can produce better results. We propose two approaches as given below:

1. SVB (Semantic Vector based) approach
2. NV-SVB (Semantic Vector based on noun and verbs) approach

4.3 Proposed Approach

4.3.1 An Overview of Approach 1 (SVB):

Our proposed approach is based on semantic vector, In the modified approach we have changed the co-occurrence matrix in such a way that each row vector of a word represent the information about the preceding and succeeding words within a window of fixed size. We have further made a change while calculating the similarity between word pairs, instead of using Minkowski distance formula we have used cosine distance formula.

Cosine distance gives the angular cosine distance between vectors u and v as given in equation 1.

Cosine Distance = (1 - Cosine Similarity)Eq. 1

The angle between two vectors u and v is represented as Cosine Similarity and is expressed as given in equation 2.

$$\text{Cosine Similarity} = \frac{u \cdot v}{\|u\| \|v\|} = \frac{\sum_{i=1}^n u_i \times v_i}{\sqrt{\sum_{i=1}^n (u_i)^2} \times \sqrt{\sum_{i=1}^n (v_i)^2}} \dots\dots\dots \text{Eq. 2}$$

We start our work by generating co-occurrence matrix of all the terms in the corpus. Further we identify a set of seed entities that correspond to some popularly used plant names.

Once the co-occurrence matrix is built, we considered rows corresponding to seed entities. The words in these rows were sorted on the basis of their cosine similarity with the corresponding seed entity. The topmost words/the words above certain threshold of similarity were considered as newly discovered named entities.

4.3.2 An overview of Approach 2 (NV-SVB):

We have observed that not all words are important in the extraction of named entities, mostly named entities are noun and to extract these nouns, verbs associated with them play a significant role.

In this approach, we have focused on the noun and verbs present in the corpus. We have used a similar notion of co-occurrence matrix as in proposed work 1. However, we have considered only nouns and verbs in construction of co-occurrence matrix. In the newly proposed matrix rows of the matrix correspond to the nouns and columns correspond to the verbs.

For the given seed words, we have extracted the verbs that are co-occurring with the seeds using co-occurrence. These verbs help in extracting new entities.

4.3.3 Description of Proposed Approaches

As discussed earlier, our approaches rely on the notion of co-occurrence matrix. Further some seed entities are identified and the co-occurrence matrix has been used for extracting more entities. In this section, we present the algorithms developed for our proposed approaches.

Table 4.1 presents the list of important variables used in the algorithms along with their clarification.

Table 4.1: List of variables used in algorithms

Variable Name	Variable Type	Explanation
SL	Sequence List	List of all words and phrases along with POS tag in sequential order of their occurrence in the corpus.
WL	Word List	Dictionary sequence of each term in Sequence List with their frequency.
CM	Co-occurrence Matrix	For each word pair in Word List, CM_{ij} represents, total number of times j^{th} term is occurring in context of i^{th} term within a fixed window size.
CSM	Co-occurrence Sub Matrix	This is a sub matrix of co-occurrence matrix, it is constructed only for noun and verbs in Word List. Rows represent nouns and columns represent verbs
SDM	Sorted Distance Matrix	Each row corresponding to each term in CM represents a vector. Cosine based distance between two terms is measured by computing the one minus cosine of the angle between these two vectors. Distance Matrix represents pairwise cosine distance between the terms in CM. DM_{ij} stores the cosine based distance between two terms i and j . Each row of distance matrix is sorted in increasing order, giving a new matrix

		Sorted Distance Matrix (SDM), each entry of SDM corresponds to a pair of value (word index, distance).
SWL	Similarity Word List	We extracted some specific rows of SDM that correspond to the seed entities. For each seed entity we extracted the topmost similar words (below some threshold) from the corresponding rows using SDM. These words were stored in the Similarity Word List SWL of the corresponding seed word. Finally this SWL was used to extract new named entities
VL	Verb List	Verb List contains the verbs that are co-occurring with seed words.
EL	Entity List	This list contains the newly extracted entities as per our experimental result.

Following are the steps in our proposed algorithm 1:

1. Pre-process the data (corpus)
2. Select the seed entities
3. Construct the co-occurrence matrix
4. Construct the distance matrix
5. Construct the similarity word list
6. Extract the entities based on seed entities

We start with the pre-processing that involves sentence extraction and POS tagging. After POS tagging, the consecutive nouns are combined to extract noun phrases. Once the data is preprocessed we construct the co-occurrence matrix, the co-occurrence matrix consists of

window based approach where window is span of words. Each word is represented by co-occurrence vector.

After construction of co-occurrence matrix, we calculated the distance between each pair of words using cosine based distance.

As cosine is similarity measure, we used one minus cosine as a score for calculating cosine based distance between two words. This value was stored in the distance matrix. Using the distance matrix we have created a similarity list of words that contain the most similar words of given word (seed) based on some predefined threshold. In the final step, we extracted new entities based on the seed entities using the similarity word list. The detailed algorithms are discussed as given below.

1. Construction of co-occurrence matrix

```
Input: window size ws, word list WL, sequence list SL
Output: Co-occurrence matrix CM
1: Assign, Total rows = Total Column = no. of words in WL(say wwl)
2: Initialize  $CM(i,j) = 0$  (where  $i,j = 1$  to  $wwl$ )
3: for each word x in SL
    Get the index  $x_i$  of x in WL
    for all following words y of x within ws in SL
        Get the index  $y_i$  of y in WL
         $CM(x_i, y_i) = CM(x_i, y_i) + 1;$ 
4:  $CM = CM + \text{transpose}(CM)$ 
5: return co-occurrence matrix CM
```

2. Construction of sorted distance matrix

```
Input: co-occurrence matrix CM
Output: Sorted distance Matrix SDM
1: for each row i of CM
    Calculate the cosine-distance with each row j of CM
        Distance matrix  $DM(i,j) = \text{cosine-distance}(i,j)$ 
2: Sort each row of DM in ascending order
2: return the sorted distance matrix SDM
```

3. Construction of similarity word list

```
Input: sorted distance matrix SDM, threshold distance td, seed words
(s1, s2, s3...)
Output: similarity Word List SWL for each seed words and Entity List EL
1: Add seed words in an entity list EL
2: for each row i of SDM
    for each column j of SDM corresponding to (s1, s2, s3...)
        if  $SDM(i,j).distance < td$ 
            get the corresponding word w
            add  $SWL(i,j) = \text{word};$ 
    Find the top k similar words from SWL for i
    Add these k words in to the EL, if not present
3: return similarity Word Lists SWL, Entity List EL
```

Following are the steps in our proposed algorithm 2:

1. Pre-process the data (corpus)
2. Construct the co-occurrence matrix
3. Construct the co-occurrence sub matrix
(rows represent nouns and columns represent verbs)
4. Get most frequently co-occurring verbs corresponding to seed nouns
5. Extract the nouns (entities) using verbs from step 4.

Here also, we start with the pre-processing data and construction of co-occurrence matrix as in previous approach. After that we find out co-occurrence sub matrix from the original co-occurrence matrix by selecting only nouns in rows and verbs in columns. Then using some seed entities in our hand we extract the verbs from the co-occurrence sub matrix. Now these extracted verbs helps us in extracting new entities. The detailed algorithms are discussed in below:

1. Construction of co-occurrence sub-matrix

```
Input: co-occurrence matrix CM
Output: co-occurrence sub matrix CSM
1: for all rows corresponding to nouns from CM
2: for all columns corresponding to verbs from CM
3: stores the entries in CSM
4: return the co-occurrence sub matrix CSM
```


2. Get verb list VL of most frequent verbs corresponding to seed nouns

```
Input: co-occurrence sub matrix CSM, seed words (s1, s2, s3...)
Output: verb list VL
1: for each seed word x from (s1, s2, s3...)
    Find the verbs (v1, v2, v3...) corresponding to x from CSM
    Add these verbs to verb list VL
2: return the verb list VL
```

3. Entity extraction process using verb list VL

```
Input: co-occurrence sub matrix CSM, verb list VL
Output: Entity List EL
1: for each verb v from verb list (v1, v2, v3...)
    Find the corresponding nouns (e1, e2, e3...) of v from CSM
    Add these noun to output entity list EL
2: return the entity list EL
```

Chapter 5

EXPERIMENTAL RESULTS AND ANALYSIS

In this chapter, we present the implementation details, steps in the experiment, experimental results and their analysis.

Objective: *To extract the named entities in agriculture domain in particular names of cereals and crops using Semantic Vector Based approach.*

Experiments were performed on the two approaches as proposed in the previous chapter.

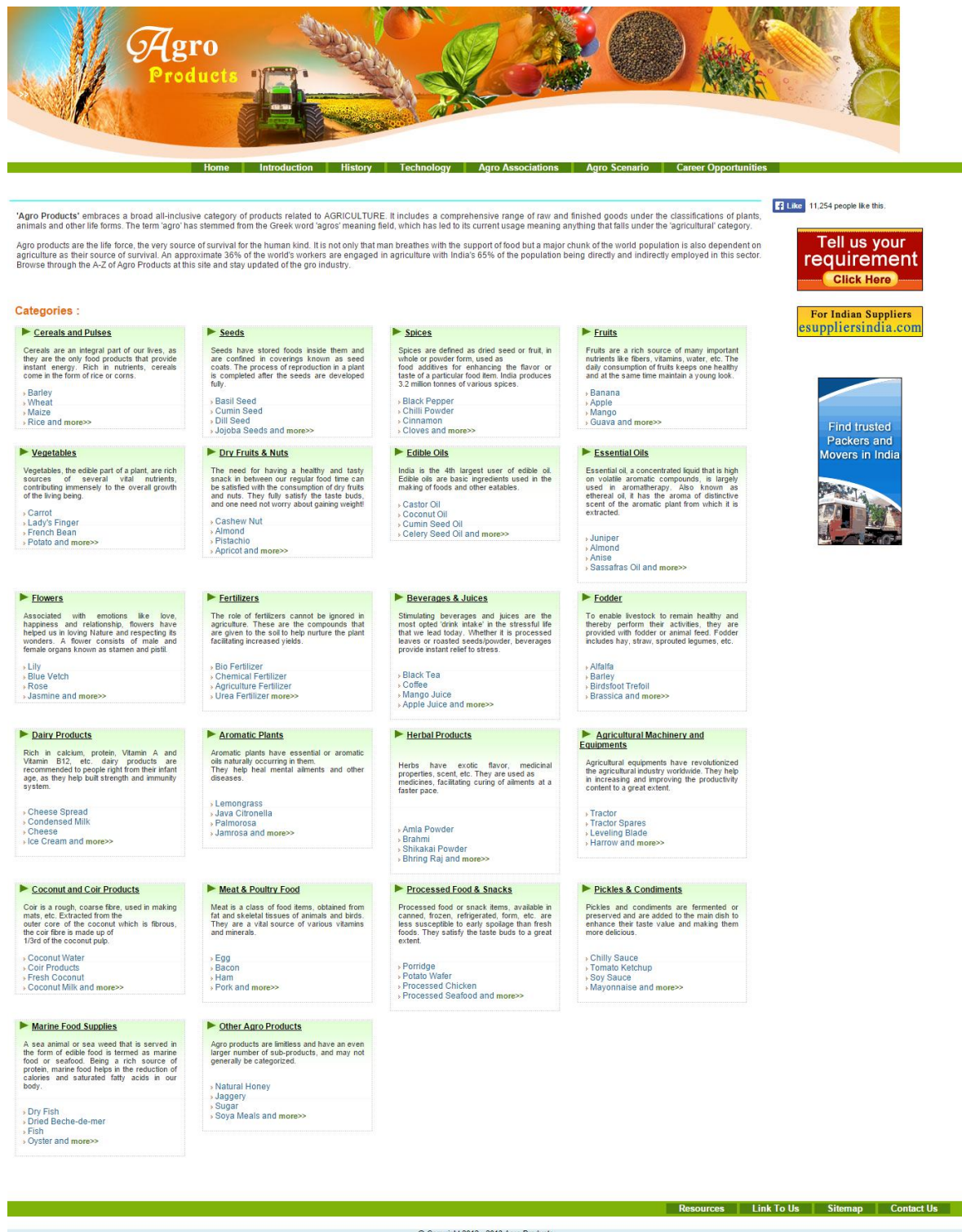
1: First part consists implementation of work done by (Lund & Burgess, 1996) with a slight modification in the original approach by changing context window and similarity measure formula.

2: Second approach is completely different which uses only nouns and verbs from the corpus to extract the named entities.

5.1 Corpus/ Dataset Selection and Preparation

In the absence of standard benchmark dataset for our problem, we have designed our own dataset. In order to perform experiments we have crawled the data from the web. Most of data have been crawled from (Agro Products, 2014). This website is dedicated to Agriculture Industry, it contains information about agriculture products, agriculture technology, careers in agriculture, industry scenario and association to agriculture in India.

A snapshot of webpage is shown in Figure 5.1



Agro Products

Home Introduction History Technology Agro Associations Agro Scenario Career Opportunities

11,254 people like this.

Tell us your requirement
Click Here

For Indian Suppliers
esuppliersindia.com

Find trusted Packers and Movers in India

Categories :

- Cereals and Pulses**
Cereals are an integral part of our lives, as they are the only food products that provide instant energy. Rich in nutrients, cereals come in the form of rice or corns.
 - Barley
 - Wheat
 - Maize
 - Rice and more>>
- Vegetables**
Vegetables, the edible part of a plant, are rich sources of several vital nutrients, contributing immensely to the overall growth of the living being.
 - Carrot
 - Lady's Finger
 - French Bean
 - Potato and more>>
- Flowers**
Associated with emotions like love, happiness and relationship, flowers have helped us in loving Nature and respecting its wonders. A flower consists of male and female organs known as stamen and pistil.
 - Lily
 - Blue Vetch
 - Rose
 - Jasmine and more>>
- Dairy Products**
Rich in calcium, protein, Vitamin A and Vitamin B12, etc., dairy products are recommended to people right from their infant age, as they help build strength and immunity system.
 - Cheese Spread
 - Condensed Milk
 - Cheese
 - Ice Cream and more>>
- Coconut and Coir Products**
Coir is a rough, coarse fibre, used in making mats, etc. Extracted from the outer core of the coconut which is fibrous, the coir fibre is made up of 1/3rd of the coconut pup.
 - Coconut Water
 - Coir Products
 - Fresh Coconut
 - Coconut Milk and more>>
- Marine Food Supplies**
A sea animal or sea weed that is served in the form of edible food is termed as marine food or seafood. Being a rich source of protein, marine food helps in the reduction of calories and saturated fatty acids in our body.
 - Dry Fish
 - Dried Beche-de-mer
 - Fish
 - Oyster and more>>
- Seeds**
Seeds have stored foods inside them and are confined in coverings known as seed coats. The process of reproduction in a plant is completed after the seeds are developed fully.
 - Basil Seed
 - Cumin Seed
 - Dill Seed
 - Jajoba Seeds and more>>
- Dry Fruits & Nuts**
The need for having a healthy and tasty snack in between our regular food time can be satisfied with the consumption of dry fruits and nuts. They fully satisfy the taste buds, and one need not worry about gaining weight!
 - Cashew Nut
 - Almond
 - Pistachio
 - Apricot and more>>
- Fertilizers**
The role of fertilizers cannot be ignored in agriculture. These are the compounds that are given to the soil to help nurture the plant facilitating increased yields.
 - Bio Fertilizer
 - Chemical Fertilizer
 - Agriculture Fertilizer
 - Urea Fertilizer more>>
- Aromatic Plants**
Aromatic plants have essential or aromatic oils naturally occurring in them. They help heal mental ailments and other diseases.
 - Lemongrass
 - Java Citronella
 - Falmarosa
 - Jamrosia and more>>
- Meat & Poultry Food**
Meat is a class of food items, obtained from fat and skeletal tissues of animals and birds. They are a vital source of various vitamins and minerals.
 - Egg
 - Bacon
 - Ham
 - Pork and more>>
- Other Agro Products**
Agro products are limitless and have an even larger number of sub-products, and may not generally be categorized.
 - Natural Honey
 - Jaggery
 - Sugar
 - Soya Meals and more>>
- Spices**
Spices are defined as dried seed or fruit, in whole or powder form, used as food additives for enhancing the flavor or taste of a particular food item. India produces 3.2 million tonnes of various spices.
 - Black Pepper
 - Chilli Powder
 - Cinnamon
 - Cloves and more>>
- Edible Oils**
India is the 4th largest user of edible oil. Edible oils are basic ingredients used in the making of foods and other eatables.
 - Castor Oil
 - Coconut Oil
 - Cumin Seed Oil
 - Calary Seed Oil and more>>
- Beverages & Juices**
Stimulating beverages and juices are the most opted 'drink intake' in the stressful life that we lead today. Whether it is processed leaves or roasted seeds/powder, beverages provide instant relief to stress.
 - Black Tea
 - Coffee
 - Mango Juice
 - Apple Juice and more>>
- Herbal Products**
Herbs have exotic flavor, medicinal properties, scent, etc. They are used as medicines, facilitating curing of ailments at a faster pace.
 - Amia Powder
 - Brahmi
 - Shikakai Powder
 - Bhiring Raj and more>>
- Processed Food & Snacks**
Processed food or snack items, available in canned, frozen, refrigerated, form, etc. are less susceptible to early spoilage than fresh foods. They satisfy the taste buds to a great extent.
 - Porridge
 - Potato Wafer
 - Processed Chicken
 - Processed Seafood and more>>
- Fruits**
Fruits are a rich source of many important nutrients like fibers, vitamins, water, etc. The daily consumption of fruits keeps one healthy and at the same time maintain a young look.
 - Banana
 - Apple
 - Mango
 - Guava and more>>
- Essential Oils**
Essential oil, a concentrated liquid that is high on volatile aromatic compounds, is largely used in aromatherapy. Also known as ethereal oil, it has the aroma of distinctive scent of the aromatic plant from which it is extracted.
 - Juniper
 - Almond
 - Anise
 - Sassafras Oil and more>>
- Fodder**
To enable livestock to remain healthy and thereby perform their activities, they are provided with fodder or animal feed. Fodder includes hay, straw, sprouted legumes, etc.
 - Alfalfa
 - Barley
 - Birdsfoot Trefoil
 - Bassica and more>>
- Agricultural Machinery and Equipments**
Agricultural equipments have revolutionized the agricultural industry worldwide. They help in increasing and improving the productivity content to a great extent.
 - Tractor
 - Tractor Spares
 - Leveling Blade
 - Harrow and more>>
- Pickles & Condiments**
Pickles and condiments are fermented or preserved and are added to the main dish to enhance their taste value and making them more delicious.
 - Chilly Sauce
 - Tomato Ketchup
 - Soy Sauce
 - Mayonnaise and more>>

Resources Link To Us Sitemap Contact Us

© Copyright 2012 - 2013 Agro Products

Figure 5.1: Snapshot of website: "http://www.agriculturalproductsindia.com/"

After crawling of web pages preprocessing were applied. Preprocessing step includes sentence extraction, phrase extraction and POS tagging.

There are 2206 sentences and 5137 distinct words.

We also have a list of cereals and crops which contains 324 cereals and crops names, from different online sources, which is used as benchmark for evaluating our result.

For POS (part of speech) tagging, Stanford POS tagger was used. The tagger accepts text as input, tokenizes the text and then assigns parts of speech, such as noun, verb, adjective, etc. to each token. It assigns the labels based on both its linguistic definition, as well as its context. The consecutive nouns were combined in order to produce the noun phrases.

5.2 Steps in Experiments

After pre-processing the text we created two lists

1: Sequence List (SL): that is basically the original text but in the list form with POS tag

2: Word List (WL): this is dictionary sequence of original sequence list, this list also contains the information about the frequency of each word/phrase.

For clarity we are presenting a small example. Let us consider a sample text:

“Barley was considered to be the first ever cereal crop to be domesticated”.....Ex. 1

Table 5.1: Example of Sequence List

Word	POS
'barley'	'NN'
'was'	'VBD'
'considered'	'VBN'
'to'	'TO'
'be'	'VB'
'the'	'DT'
'first'	'JJ'
'ever'	'RB'
'cereal crop'	'NP'
'to'	'TO'
'be'	'VB'
'domesticated'	'VBN'

Table 5.2: Example of Word List

Word	POS	Frequency
'barley'	'NN'	1
'be'	'VB'	2
'cereal crop'	'NP'	1
'considered'	'VBN'	1
'domesticated'	'VBN'	1
'ever'	'RB'	1
'first'	'JJ'	1
'the'	'DT'	1
'to'	'TO'	2
'was'	'VBD'	1

Construction of Co-occurrence Matrix (CM):

Co-occurrence matrix was constructed over the whole corpus. Co-occurrence matrix is a term to term matrix. Window size 'ws' was taken as 2.

To construct the co-occurrence matrix a window is scanned over the whole corpus. For each word in the window its entry is made in co-occurrence matrix.

After the construction of co-occurrence matrix by looking row or column of corresponding word we can get the information about the context of that word in the whole corpus. The co-occurrence matrix for example 1 is shown in Table 5.3

Table 5.3: Example of Co-occurrence Matrix

	barley	be	cereal crop	considered	domesticated	ever	first	the	to	was
barley	0	0	0	1	0	0	0	0	0	1
be	0	0	1	1	1	0	1	1	2	0
cereal crop	0	1	0	0	0	1	1	0	1	0
considered	1	1	0	0	0	0	0	0	1	1
domesticated	0	1	0	0	0	0	0	0	1	0
ever	0	0	1	0	0	0	1	1	1	0
first	0	1	1	0	0	1	0	1	0	0
the	0	1	0	0	0	1	1	0	1	0
to	0	2	1	1	1	1	0	1	0	1
was	1	0	0	1	0	0	0	0	1	0

Construction of Distance Matrix:

Distance matrix is constructed by considering each term row as a term vector and pairwise cosine similarity is calculated for each pair of terms. Distance matrix is a symmetric matrix. DM_{ij} represents similarity between i^{th} and j^{th} word. Table 5.4 shows distance matrix for example 1.

Table 5.4: Example of Distance Matrix

	barley	be	cereal crop	considered	domesticated	ever	first	the	to	was
barley	0	0.7643	1.0000	0.6464	1.0000	1.0000	1.0000	1.0000	0.5528	0.5918
be	0.7643	0	0.5000	0.6667	0.5286	0.1667	0.6667	0.5000	0.5784	0.4226
cereal crop	1.0000	0.5000	0	0.5000	0.2929	0.5000	0.5000	0	0.5257	0.7113
considered	0.6464	0.6667	0.5000	0	0.2929	0.7500	0.7500	0.5000	0.5257	0.4226
domesticated	1.0000	0.5286	0.2929	0.2929	0	0.6464	0.6464	0.2929	0.5528	0.5918
ever	1.0000	0.1667	0.5000	0.7500	0.6464	0	0.5000	0.5000	0.6838	0.7113
first	1.0000	0.6667	0.5000	0.7500	0.6464	0.5000	0	0.5000	0.2094	1.0000
the	1.0000	0.5000	0	0.5000	0.2929	0.5000	0.5000	0	0.5257	0.7113
to	0.5528	0.5784	0.5257	0.5257	0.5528	0.6838	0.2094	0.5257	0	0.8174
was	0.5918	0.4226	0.7113	0.4226	0.5918	0.7113	1.0000	0.7113	0.8174	0

In the distance matrix range is between [0, 1]. Here 0 represent most similar and 1 represent completely different.

Construction of Similarity Word List (SWL):

Using the distance matrix, we can get the list of most similar words for a particular word. So we have created a similarity word list for each seed using the distance matrix as described in table 4.1 in previous chapter. For this we considered only the words having distance value within some threshold. Threshold value was set to 0.5. So words with distance values between (0 to 0.5) can be considered as similar words. The words were sorted on the basis of their similarity values to get the similarity word list with the most similar word occurring at top position and least similar word occurring at end.

Our assumption is that a crop name will be similar to other crop name. If we have some input crop name as seed then we can easily extract new crop name using the above similarity word list. For this method we do not need many input seed words only few seed words can also provide good result.

Once we have similarity word list for all seed words in the corpus, we can easily extract the new crops.

By taking only two seed words ‘*barley*’ and ‘*rye*’, we were able to extracted new crops: malted barley, wheat, rice, rice bran, maize, vetch, sorghum, triticale, lentils, black gram.

Table 5.5: Output entities corresponds to seeds Barley and Rye

barley	malted barley	Wheat	Composi tion	rice	This	rice bran	maize	vetch	sorghum	triticale
rye	Wheat	Maize	the	rice	This	barley	rice bran	lentils	black gram	and

As an initial attempt we were able to extract new entities, not all the entities are covered. Moreover the list contains many noisy words also, therefore there is a large scope of improving the result. We have improved upon our experiment by focusing only on nouns and verbs. Our improved experiment is presented next.

(Experiment 2).

For the second experiment we created co-occurrence matrix only for nouns and verbs in the text.

Table 5.6 shows co-occurrence sub matrix for example 1:

Table 5.6: Example of Co-occurrence Sub Matrix

Noun\verb	'be'	'considered'	'domesticated'	'was'
'barley'	0	1	0	1
'cereal crop'	1	0	0	0

Once the co-occurrence matrix is created, we considered some cereals name as seeds words and using these seed words we extracted the common verbs that co-occur with the seeds.

Now we have common verbs in our hand, these verbs are quite useful in extraction of new cereals name. These verbs were sorted in order to their frequency and discarded the verbs having very low frequency (in our experiment we discarded verbs with frequency 1).

Now we extracted a list of nouns with the help of extracted verbs. This extracted list of nouns contains the entity of our interest.

5.3 Analysis of Result

Following are the evaluating parameters for analyzing our results: confusion matrix, precision, recall and f-measure.

Confusion Matrix: Confusion Matrix is a special table that allows the evaluation of algorithm's performance. It contain the information about actual and predicted classifications done by a classification system (Kohavi & Provost, 1998).

Table 5.7: Confusion Matrix

Model \ Actual	Correct	Not Correct
Selected	TP	FP
Not Selected	FN	TN

TP (True Positive) → correctly selected by model

FN (False negative) → not selected by model but actually correct

TN (True Negative) → not selected by model and not correct

FP (False Positive) → wrongly selected by model

Precision measures the accuracy of the result obtained.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \dots\dots\dots \text{Eq. 5.1}$$

Recall measures the coverage of the result obtained.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \dots\dots\dots \text{Eq. 5.2}$$

In many cases both recall and precision may not be high simultaneously in other words in most of the cases there has to be a tradeoff between precision and recall. Therefore recall and precision provides different views of evaluation. In order to judge the quality of result based on both precision and recall, F-measure can be used. F-measure is the harmonic mean of recall and precision.

$$\text{F-measure} = \left(\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right) \dots\dots\dots \text{Eq. 5.3}$$

We observed that the selection of seeds may affect the quality of result. Therefore, we performed our experiments by selecting the seeds randomly and by selecting the seeds manually which we are more common and are expected to provide better result. In each experiment we gradually increased the number of seeds to check the performance on different numbers of seeds.

Experiment 1:

This experiment is done using SVB approach. After the completion of this experiment we are selecting topmost m similar entities corresponding to given seed entities and we can get a list of newly extracted entities. The result of experiment using random seed selection are presented in Table 5.8. The parameter values considered are: number of seed entities (variable as shown in table 5.8), similarity threshold ($td \geq 0.5$, $m=10$).

Table 5.8: Result of experiment 1 using random seed selection

Number of Seeds	10	20	30	40	50
Precision	43.3962264	37.1134021	33.0769231	34.751773	28.1553398
Recall	24.2105263	37.8947368	45.2631579	51.5789474	61.0526316
F-measure	31.0810811	37.5	38.2222222	41.5254237	38.538206

It can be observed from Table 5.8 that on increasing number of seeds precision degrades but recall increases.

The result of experiment 1 corresponding to manual seed selection are presented in Table 5.9. The parameter values are same as for random seed selection.

Table 5.9: Result of experiment 1 using manually seed selection

Number of Seeds	10	20	30	40	50
Precision	62.962963	46.1538462	43.3333333	41.0714286	37.5
Recall	17.8947368	31.5789474	41.0526316	48.4210526	56.8421053
F-measure	27.8688525	37.5	42.1621622	44.4444444	45.1882845

Here also, on increasing number of seeds precision degrades but recall increases. It can be observed that precision is better for manual seed selection whereas random seed selection is

giving better recall, F-measure is varying for different number of seeds. As the data sample and number of entities are not very large, we cannot say whether random or manual selection works better. However we can comment that the method of seed selection and number of seeds are important factors in determining the quality of result.

Experiment 2:

This experiment is done using NV-SVB approach. The result of experiment using random seed selection are presented in Table 5.10. The parameter values considered are: number of seed entities (variable as shown in Table 5.10), similarity threshold ($td \geq 0.5$) and $m=10$.

Again the experiments were performed by selecting the seeds randomly and manually and number of seeds values were varied to see their effect on quality of result. The results using random seed and manual seed selection are presented in Table 5.10 and Table 5.11 respectively.

Table 5.10: Result of experiment 2 using random seed selection

Number of Seeds	10	20	30	40	50
Precision	9.20245399	8.46394984	8.54816825	7.85463072	8.14371257
Recall	63.1578947	56.8421053	66.3157895	70.5263158	71.5789474
F-measure	16.064257	14.73397	15.1442308	14.1350211	14.6236559

Table 5.11: Result of experiment 2 using manual seed selection

Number of Seeds	10	20	30	40	50
Precision	7.53424658	7.45614035	7.36728061	7.24174654	7.21102863
Recall	69.4736842	71.5789474	71.5789474	71.5789474	71.5789474
F-measure	13.5942327	13.5054618	13.3595285	13.1528046	13.1021195

In this experiment random selection and general seed selection does not have much impact on the result. Precision and recall values are almost similar for both types of seed selection.

Second experiment is good for very little number of seeds. Any increase in number of seeds does not make too much increase in performance.

As compared to first experiment this experiment covers more entities because recall is high but we achieved this performance at the cost of precision loss. In spite of low precision value in comparison to first experiment we cannot say that experiment 2 is less accurate than experiment 1. Firstly as the recall is higher therefore definitely this experiment is covering more entities in comparison to experiment 1. Secondly the decline in the precision is mainly because the length of list containing the resultant entity is very large in comparison to the list obtained in experiment 1. It may not be justifiable to compare the precision when the length of result is different. Obviously the list with longer length is expected to be less precise.

It can be emphasized that recall achieved is quite motivating and we are able to cover up to 75% entities of our interest.

Chapter 6

CONCLUSION

In this work we have tried to develop an entity extractor for extracting cereals and crops from agriculture text. Till now no entity extractor is available for extracting entities for agriculture domain, we think our work is a preliminary and important step in this direction. We propose a novel context representation beyond the previously dominant bag of words approach. We have applied the context based approach for developing the system. In general the context based approach are computationally intensive. However in our work context is captured in the form of matrix, therefore the approach has a good computational efficiency.

In absence of benchmark data we considered the textual corpus from agriculture related website (<http://www.agriculturalproductsindia.com/>) and entities from different sources were considered as benchmark data. Due to the unavailability of well fleshed agricultural resources we have not achieved surpassing accuracy but as a preliminary attempt the results are quite motivating and in future results can be further improve by using multilayered resources.

In future we can also automatize the system by using machine learning techniques to distinguish between the agriculture terms and non-agriculture terms. A possible future work includes an improvement in number of entities classes for tagging, we can also develop an entity tagger for other agriculture entities like pest and pesticides, crop diseases, fertilizers, etc.

On the other hand we can use ontologies in named entity recognition. By using ontologies we can use different level of ontology to tag a particular entity. Different level of entity tags will increase the understanding level of that entity.

References

- AFNER - Named Entity Recognition (2015, March 18). Retrieved from AFNER - Named Entity Recognition website: <http://afner.sourceforge.net/what.html>
- Agro Products (2014, December 29). Retrieved from Agro Products website: <http://www.agriculturalproductsindia.com/>
- Alex, B., Haddow, B., & Grover, C. (2007). Recognising Nested Named Entities in Biomedical Text. *In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 65-72.
- Baluja, S., Mittal, V. O., & Sukthankar, R. (2000). Applying Machine Learning for High-Performance Named-Entity Extraction. *Computational Intelligence*, 16(4), 586-595.
- Bick, E. (2004). A Named Entity Recognizer for Danish. *In Proceedings of 4th International Conference on Language Resources and Evaluation*, 305-308.
- Black, W. J., Rinaldi, F., & Mowatt, D. (1998). FACILE: Description of the NE System Used for MUC-7. *In Proceedings of the 7th Message Understanding Conference*, 1-10.
- Boutsis, S., Demiros, I., Giouli, V., Liakata, M., Papageorgiou, H., & Piperidis, S. (2000). A System for Recognition of Named Entities in Greek. *Natural Language Processing—NLP 2000*, Springer Berlin Heidelberg, 424-435.
- Brin, S. (1999). Extracting Patterns and Relations from the World Wide Web. *The World Wide Web and Databases*, Springer Berlin Heidelberg, 172-183.
- Chen, H. H., & Lee, J. C. (1996). Identification and Classification of Proper Nouns in Chinese Texts. *In Proceedings of the 16th Conference on Computational Linguistics-1*, 222-229.
- Chen, Z., & Lu, Y. (2011). A Word Co-occurrence Matrix Based Method for Relevance Feedback. *Journal of Computational Information Systems*, 7(1), 17-24.
- Coates-Stephens, S. (1992). The Analysis and Acquisition of Proper Names for Robust Text Understanding, Doctoral dissertation, *City University, London*.
- Cohen, W. W., & Sarawagi, S. (2004). Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. *In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 89-98.

-
- Cucchiarelli, A., & Velardi, P. (2001). Unsupervised Named Entity Recognition using Syntactic and Semantic Contextual Evidence. *Computational Linguistics*, 27(1), 123-131.
- Cucerzan, S., & Yarowsky, D. (1999, June). Language Independent Named Entity Recognition Combining Morphological And Contextual Evidence. *In Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC*, 90-99.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the Association for Information Science and Technology (JASIS)*, 41(6), 391-407.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., & Yates, A. (2005). Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial intelligence*, 165(1), 91-134.
- Fleischman, M. (2001). Automated Subcategorization of Named Entities. *In Proceedings of the Conference of the European Chapter of Association for Computational Linguistic*, 25-30.
- Fleischman, M., & Hovy, E. (2002). Fine Grained Classification of Named Entities. *In Proceedings of the 19th International Conference on Computational Linguistics- 1*, 1-7.
- Fresko, M., Rosenfeld, B., & Feldman, R. (2005). A Hybrid Approach to NER by MEMM and Manual Rules. *In Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 361-362.
- Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *In Proceedings of the 16th Conference on Computational Linguistics*, 1, 466-471.
- Kazama, J. I., Makino, T., Ohta, Y., & Tsujii, J. I. (2002). Tuning Support Vector Machines for Biomedical Named Entity Recognition. *In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, 3, 1-8.
- Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent Semantics in Named Entity Recognition. *Expert Systems with Applications*, 42(7), 3470-3479.
- Lee, K. J., Hwang, Y. S., Kim, S., & Rim, H. C. (2004). Biomedical Named Entity Recognition using Two-phase Model Based on SVMs. *Journal of Biomedical Informatics*, 37(6), 436-447.
- Lee, S., & Lee, G. G. (2005). Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping. *In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2005)*, 658-669.
-

-
- Lin, Y. F., Tsai, T. H., Chou, W. C., Wu, K. P., Sung, T. Y., & Hsu, W. L. (2004). A Maximum Entropy Approach to Biomedical Named Entity Recognition. *In Proceedings of 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2004)*, 56-61.
- Lund, K., & Burgess, C. (1996). Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.
- May, J., Brunstein, A., Natarajan, P., & Weischedel, R. (2003). Surprise! What's in a Cebuano or Hindi Name? *Transactions on Asian Language Information Processing (TALIP)*, ACM, 2(3), 169-180.
- Mayfield, J., McNamee, P., & Piatko, C. (2003). Named Entity Recognition using Hundreds of Thousands of Features. *In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 4, 184-187.
- Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. (2001). Named Entity Recognition from Diverse Text Types. *In Proceedings of Conference on Recent Advances in Natural Language Processing*, 257-274.
- Minkov, E., Wang, R. C., & Cohen, W. W. (2005). Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 443-450.
- Nadeau, D., & Sekine, S. (2007). A survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1), 3-26.
- Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 426-433.
- Piskorski, J. (2004). Extraction of Polish Named-Entities. *In Proceedings of 4th International Conference on Language Resources and Evaluation*, 1-4.
- Poibeau, T. (2003). The Multilingual Named Entity Recognition Framework. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 2, 155-158.
- Popov, B., Kirilov, A., Maynard, D., & Manov, D. (2004). Creation of Reusable Components and Language Resources for Named Entity Recognition in Russian. *In Proceedings of the International Conference on Language Resources and Evaluation*, 309-312.
- Provost, F. J., Fawcett, T., & Kohavi, R. (1998, July). The Case against Accuracy Estimation for Comparing Induction Algorithms. *In Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, 98, 445-453.

-
- Raja, K., Subramani, S., & Natarajan, J. (2014). A Hybrid Named Entity Tagger for Tagging Human Proteins/Genes. *International Journal of Data Mining and Bioinformatics*, 10(3), 315-328.
- Rau, L. F. (1991). Extracting Company Names from Text. *In Proceedings of Seventh IEEE Conference on Artificial Intelligence Applications*, 1, 29-32.
- Rizzolo, N., & Roth, D. (2010). Learning Based Java for Rapid Development of NLP Systems. *Language Resources and Evaluation*, 957-964.
- Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S., & Mitra, P. (2008). A Hybrid Approach for Named Entity Recognition in Indian Languages. *In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, 17-24.
- Saha, S. K., Sarkar, S., & Mitra, P. (2009). Feature Selection Techniques for Maximum Entropy based Biomedical Named Entity Recognition. *Journal of Biomedical Informatics*, 42(5), 905-911.
- Sasidhar, B., Yohan, P. M., Babu, A. V., & Govardhan, A. (2011). A Survey on Named Entity Recognition in Indian Languages with Particular Reference to Telugu. *International Journal of Computer Science*, 8(2), 438-443.
- Seki, K., & Mostafa, J. (2005). A Hybrid Approach to Protein Name Identification in Biomedical Texts. *Information Processing & Management*, 41(4), 723-743.
- Settles, B. (2004). Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. *In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 104-107.
- Tang, B., Cao, H., Wang, X., Chen, Q., & Xu, H. (2014). Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *Biomedical Research International*, 1-6.
- Thielen, C. (1995). An Approach to Proper Name Tagging for German. *In Proceedings of Conference of European Chapter of the Association for Computational Linguistics (SIGDAT)*, 1-7.
- Wang, L. J., Li, W. C., & Chang, C. H. (1992). Recognizing Unregistered Names for Mandarin Word Identification. *In Proceedings of the 14th Conference on Computational Linguistics*, 4, 1239-1243.
- Whitelaw, C., & Patrick, J. (2003). Evaluating Corpora for Named Entity Recognition using Character-level Features. *AI 2003: Advances in Artificial Intelligence*, Springer Berlin Heidelberg, 910-921.
- Witten, I. H., Bray, Z., Mahoui, M., & Teahan, W. J. (1999). Using Language Models for Generic Entity Extraction. *In Proceedings of the ICML Workshop on Text Mining*, 1-11.

- Yu, S., Bai, S., & Wu, P. (1998). Description of the Kent Ridge Digital Labs System used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference*, 7, 1-16.
- Zhou, G., & Su, J. (2002). Named Entity Recognition using an HMM-based Chunk Tagger. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 473-480.
- Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing Names in Biomedical Texts: A Machine Learning Approach. *Bioinformatics*, 20(7), 1178-1190.
- Zhu, J., Uren, V., & Motta, E. (2005). E-Spotter: Adaptive Named Entity Recognition for Web Browsing. *Professional Knowledge Management*, Springer Berlin Heidelberg, 518-529.

List of Publication

AGNER: Entity Tagger in Agriculture Domain. *In Proceedings of the 2nd International Conference on "Computing for Sustainable Global Development"*, 2, 1134- 1138.