ISSN 1870-4069

Context Pattern Based Agricultural Named Entity Recognition

Payal Biswas, Aditi Sharan, Ashish Kumar

Jawaharlal Nehru University, New Delhi, India
payal.biswas138@gmail.com

Abstract. Named entity recognition (NER) play a vital role in various application of Natural Language processing. Although a significant work has been done in general and biomedical domain NER, but agriculture domain has been ignored for a long time. Agriculture entity includes name of crops, crop diseases, fertilizers etc. Due to the inapplicability of conventional features which has been used for identifying general named entities, recognizing and extracting the agricultural entities become a rigorous and challenging task. As NER in agriculture domain has not been yet explored a lot, thus building up a NER system for agriculture domain is very recent and vital work. This paper proposes a novel context-based approach to develop a NER system for agriculture domain. The proposed approach employs the context pattern for extracting the required entity of interest. The experiment is carried out in two different genres 1) Word Context Pattern 2) POS context pattern. In word context pattern, merely the co-occurring word tokens corresponding to the required entity is considered. While in Part of Speech (POS) rather than considering the co-occurring word tokens, their POS structure is plied. We have proposed seven part of speech patterns which are most likely to comprise all the instances of required entity of interest. The remarkable point is that the proposed POS patterns have not only device the known agricultural entities but have also extracted out 55 hidden entities from the data set. To boost up the performance of the NER system semantic similarity module has also been exercised. The proposed approach attains an accuracy of 70.45 % and recall of 91.3% which is appreciable as the preparatory work.

Keywords: named entity recognition, agriculture NER, word context pattern, POS context pattern, semantic similarity, agriculture entity.

1 Introduction

Named Entity Recognition (NER) has grown as an important area of research in past two decades. It is a fundamental and key component in the field of text mining and Natural Language Processing (NLP). Lisa F. Rau [23] presented the first research paper in this area in 1991 at Seventh IEEE Conference on Artificial Intelligence Application [23] and then after in 1996, after the MUC-6 (Message Understanding Conference-6), it has been accelerated and never been declined since then [9]. In the taxonomy of computational linguistics, it falls under the domain of information extraction. Named Entity recognition can be defined as identifying the references to specific entities like names, including person, organization and location names and numeric expressions including time, date, money, and percent expression [20]. It involves processing structured as well as unstructured text document and extracts these information units

pp. 383–399; rec. 2019-09-15; acc. 2019-10-15

from the text. As named entities are the basic building block of content of the document, extracting these phrases from the document play an important role towards more intelligent information extraction and management. In the preparatory stage the task of NER involves extracting the name of place person and organization etc. in the text. Later on, these categories have been sub-categorize into more fine-grained classes like name of city, state [8, 14], name of film and scientist [7] and so on. The task of NER has also disseminated for different languages like Turkish [28], Arabic [1] and many more. Researchers working over the field of NLP and text mining have also glance their eye over different genre and domain like tweets data [25], clinical data [13], and biomedical data [32]. However, NER for agriculture domain has been ignored for a long time.

As we know India is principally an agriculture-based country and it is the backbone of Indian economy. It contributes a significant figure, approx. 13.7 % * to the Gross Domestic Product (GDP). For last six decades, various government and non-government organizations such as IARI, ICAR, IASRI and FAO are working in the field of agriculture. As a result of this ample amount of digital information become available in the internet. Agriculture domain is very much similar to the biomedical domain, but due to some unfavorable circumstances or unavailable resources no significant work has been done in this field. However, if someone want to do any kind of Natural Language Processing like Information retrieval, Machine translation, document summarization, Semantic web search, Question answering System or any other in agriculture domain one has to find out the agricultural named entities from the document.

Although there are various Named Entity Recognizers available now a day such as Stanford's Named Entity Recognizer, Python NLTK Named Entity Recognizer, Learning Based Java1 (LBJ) [26] 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. Similarly, various NER systems like ABNER, BNER, BANNER has been developed for biomedical domain but they are specially trained for recognizing biomedical entities hence they do not perform well for recognizing agricultural entities.

Hence if we want to develop any kind of NLP application for agriculture domain, we have to extract the basic entities particularly the agricultural named entities like plants name, plant diseases and fertilizer's name etc. from the text. A propaedeutic step has been taken by Biswas *et al.* [2] in this field. However, they have suggested solely a framework called AGNER, for building up a NER system for agriculture domain. Although there are various challenges comprising in building up a NER system for agricultural domain, in this paper we have taken an attempt to develop an agricultural NER using context pattern and semantic similarity measure.

2 Challenges with Agriculture NER

Since the research over NER lends its wings, a considerable amount of work has been done in the field of NER. In general, there are two main approaches for developing an NER system: Rule Based Approach and Machine learning approach. However, the rule-based approaches give good results, but developing such rules are time consuming, importable and require expert's knowledge. Nowadays machine learning techniques for developing the NER systems are in trend. Various machine learning approaches such as HMM [6, 35] SVM [30], CRF, [19, 29] maximum entropy [27] has been employed for framing the NER systems both for open domain NER as well as domain specific NER like biomedical [34, 12, 16] and tweeter [17,25]. These machine learning approaches are applied over some specific features. For an instance orthographic feature, word shape feature, prefix and suffix feature [29] and many others. Orthographic feature describes how a token is structured like ALLCAPS, INITCAP, alpha numeric, word containing roman digit etc. Word shape feature expounds that same category of tokens may share similar shape and prefixes and suffixes are some special sub words with which same class of tokens starts or ends. These features play an evincive role in exercising varied machine learning algorithms. However, the most hitching thing is that these features are not applicable with the agriculture entities. For example, the crop names like rice, wheat, oat, barley etc. neither follow any orthographic feature, nor word shape or prefix, suffix feature. Hence exploiting the machine learning algorithms for extracting the agricultural entities is not facile.

Even though agriculture domain seems similar to biomedical domain however their naming conventions are not same. Unlike agriculture entities, biomedical entities have some special characteristics such as long and descriptive naming sequences, conjunctive and disjunctive structure, abbreviations, cascaded construction etc. Moreover, in case of biomedical domain the presence of two annotated corpora GENIA and GENETAG [33] and the impendence of MEDLINE [31], scientific biomedical knowledgebase aids in fabricating the biomedical NER with satisfactory results. However as per our knowledge, very few works have been done in the field of NER in agriculture domain [2, 4, 18] and there is no such annotated agricultural corpora and full fleshed agricultural knowledge base.

3 Proposed Work

As it is discussed in the previous section that the various word level features like orthographic feature, word shape feature, prefix and suffix feature are not congruent for the agricultural entities hence we have to move for the context-based feature. In this work we have proposed an approach for developing an agricultural named entity recognizer, AGNER, which exploits the context of required named entities. The experiment is carried out in two phases: 1) using the context pattern of words 2) using the context pattern of POS.

3.1 Experiment 1: Word Context Pattern

The context pattern of words is devised in two steps. First step quest for the co-occurring word patterns and the next step check for the accuracy of their co-occurrence.

Co-Occurring Word Pattern Extraction. In order to extract the context patterns of words corresponding to the named entity terms, first of all the required entities are searched in the training dataset. In association with these entities different size context windows are extracted. Context windows are the set of tokens of fix size neighboring words in the text. Table 1 shows the context window of different size around the token W_0. In this work the maximum window size is taken as ±5.

Table 1. Context window of words.

Window	Description
W_0 $\mathbf{W_{+1}}$	A word right to the required token
$\mathbf{W_{-1}}$ W_0	A word left to the required token
$\mathbf{W_{-1}}W_0$ $\mathbf{W_{+1}}$	A word both left and right to the required token
W_0 $\mathbf{W_{+1}}$ $\mathbf{W_{+2}}$	Two words right to the required token
$\mathbf{W_{-2}}$ $\mathbf{W_{-1}}$ W_0	Two words left to the required token
$\mathbf{W_{-2}}$ $\mathbf{W_{-1}}$ W_0 $\mathbf{W_{+1}}$ $\mathbf{W_{+2}}$	Two words left and two words right to the required token
......$\mathbf{W_{-2}}$ $\mathbf{W_{-1}}$ W_0 $\mathbf{W_{+1}}$ $\mathbf{W_{+2}}$............	....

As a result of this multitudinous set of word phrases of different window size has been obtained. These word phrases are sorted on the basis of their frequency and few most frequent words patterns are extracted out.

Co-occurrence Evaluation. This step checks for the verity of co-occurrence of extracted terms with the required entity of interest. For this, first of all the terms which have been extracted in the word pattern extraction phase, as a co-occurred term with the required named entity, are retraced in the testing dataset. Now the co-occurrence of these entities has been cross checked with the other entities in the dataset. At last the entities which have been extracted from the probable position of named entity are examined for the required entity of interest. The algorithm, shown in Table 2, presents the implementation detail of the proposed word pattern approach.

Table 2. Algorithm for word context pattern.

Algorithm: Word Context Pattern

```
Input: $D_{TR}$ – Training Dataset, $D_{TS}$ – Testing Dataset, S – Seed Word
Output: Agriculture Entities E = Φ

  Word Pattern Extraction
E ← Φ, T←Φ;
for each $s \in S$ do
Find context window for $W_i$ in $D_{TR}$ where $i = \pm 5$
Extract co-occurred words, $CW_d(i) \leftarrow W_i$
Co-occurred words patterns $CW_dP \leftarrow$ Frequent $CW_d(i)$
Word Pattern $P_w = CW_dP - s$
Co-occurrence Evaluation
```

Algorithm: Word Context Pattern

```
7.  For each window $W_i$ in D_TS where $i = \pm 5$
8.  Extract n-gram, $N_i \leftarrow W_i$
9.  Search for n-gram $N$, where $N_i := P_w$
10. Extract token $T$ from $N$, where $T$ is in probable position in $N$.
11. Agriculture Entity, $E := E + T$
```

The flow diagram of the process, which is involved in the experiment 1, is shown in Fig 1.

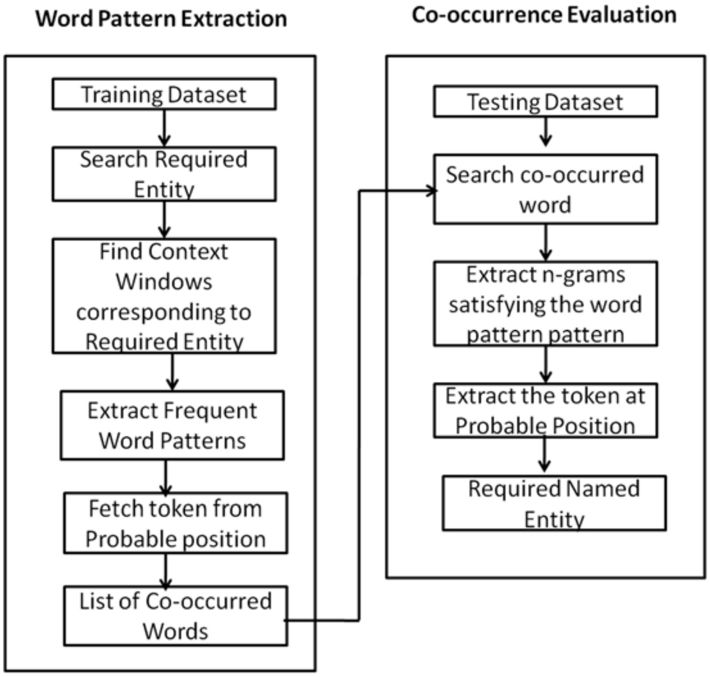
Word Pattern Extraction
Training Dataset
Search Required Entity
Find Context Windows corresponding to Required Entity
Extract Frequent Word Patterns
Fetch token from Probable position
List of Co-occurred Words
Co-occurrence Evaluation
Testing Dataset
Search co-occurred word
Extract n-grams satisfying the word pattern pattern
Extract the token at Probable Position
Required Named Entity

Fig.1. Flow diagram of word context pattern.

3.2 Experiment 2: Part of Speech Context Pattern

Unlike word pattern, in part of speech (POS) context pattern, rather than considering the absolute word, their part of speech structure is captured. This approach works in three steps: Pattern Extraction, Pattern Evaluation, and Pruning.

Pattern Extraction. The word phrases which have been devised from former approach are diverse in nature. Hence in order to achieve the homogeneity among the various word tokens of different size windows these word phrases have parsed for their POS structure [2]. Once after the obtained word phrases have been scanned for the POS structured, a finite set of unique POS patterns have been emerged. These patterns are then sorted on the basis of their frequency and few of the most frequent POS patterns are extracted. As a consequence, seven most frequent patterns have been derived. Table 3 quotes varied POS patterns, which has been instated from the context word window. In these patterns NN stands for noun phrase, IN is for preposition, JJ for adjunctive, and VB stands for verb phrase. The noun term written in bold and italic in the delineated patterns is the collocation of required entity of interest and that position is the most probable position having the named entity token. The process of pattern extraction process is shown in Fig 2.

Table 3. Part-of-Speech Context Patterns.

S no.	POS Pattern	Description	Example
1.	NN + IN + ***NN***	Any noun phrase followed by a preposition followed by required entity of interest.	…characteristics of rice.. …production of turnip… …yield of paprika….
2.	***NN*** + NN	Two consecutive nouns	…oats variety… …almond cultivar…
3.	JJ + ***NN***	Adjective followed by noun	…black pepper… …grind spelt…
4.	VB + ***NN***	Verb phrase followed by noun	…cultivated rice… …produced Walnuts…
5.	***NN*** + VB	Noun followed by verb	…bean planted… …sage harvested…
6.	IN + ***NN*** + IN	Noun phrase sandwich between two preposition	…of garlic in… …in hazelnut for… …with pistachios from …
7.	NN + IN + VB + ***NN***	Required noun phrase followed after a triplet of noun, preposition and verb.	…growth of cultivated oats… …sulphur at dried apricots… …advantage in producing ginger…

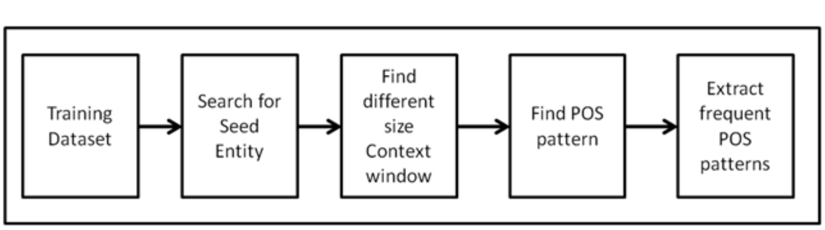
Training Dataset
Search for Seed Entity
Find different size Context window
Find POS pattern
Extract frequent POS patterns

Fig. 2. Flow diagram of POS pattern Extraction.

Pattern Evaluation. In the pattern evaluation phase, the exactitude of the patterns which has been derived in the pattern extraction phase are to be enumerated. To check the soundness of these patterns a separate subset of data called the testing dataset is taken. Then the whole dataset is parsed for the POS structure. Once the POS sequence of words in the dataset is obtained, we will look for the sub sequence matching the POS patterns structure. After getting the POS pattern sequence, the word sequence corresponding to that POS sequence are chunked out. In this chunk of word, the token laying in the most probable position of NE in the POS pattern sequence is then extracted. The steps involved in the pattern evaluation phrase is illustrated in the Fig. 3.

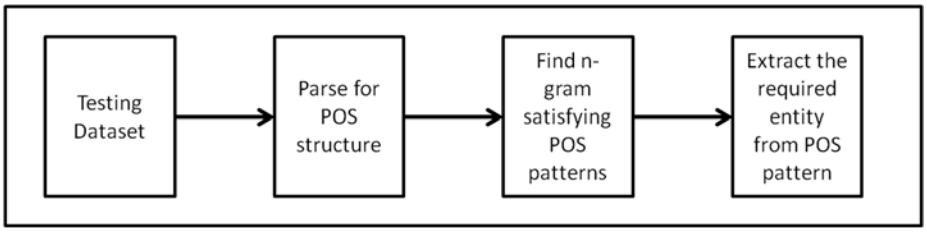
Testing Dataset
Parse for POS structure
Find n-gram satisfying POS patterns
Extract the required entity from POS pattern

Fig. 3. Flow diagram of POS pattern Evaluation.

Pruning. It is supposed that the list of tokens obtained as an output in the previous section contain the named entity terms. However along with the required entity of interest there is a faction of irrelevant terms. In order to improve the accuracy of the system, these irrelevant terms must be filtered out from the required relevant terms. There are various ways to drive this for an instance semantic similarity between different set of words, using difference between the information content from any knowledge base [24] like Word net, concept net or any domain specific ontology, exercise different clustering algorithm like k means and many other. We have used semantic similarity measure to alienate the extraneous words from the list of words. There are also multifarious techniques available for scaling the semantic similarity between set of words. Bollegala [5] and Haiyan [10] employed Web search engine to measure the semantic similarity between words using web search engine. Pekar and Staab [22] and Pedersen *et al.* [21] uses the relatedness of concepts in the thesaurus and distributional feature vector to compute the semantic similarity. While Li *et al.* [15] employed multiple information sources for doing the same. In this work we have plied the UMBC semantic similarity model proposed by Han *et al.* [11]. The online demonstration of the model is available at http://swoogle.umbc.edu/SimService/. It is based on align-and-penalize algorithm. The algorithm works over the concept that if two sentences or short text sequences are semantically equivalent then align them and the words that are poorly aligned have to give penalty. The alignment quality serves as a similarity measure. The Semantic Textual Similarity (STS) is computed using:

$$STS = T - P^{'} - P^{''}$$

where T is the term alignments score, $P^{'}$ is the penalty for bad term alignments and $P^{''}$ is the penalty for syntactic contradictions led by the alignments. However, $P^{''}$ had not been fully implemented, it is shown just for completeness. Alignments score T is define by the equation:

$$T = \frac{\sum_{t \in S_1} sim'\left(t, g\left(t\right)\right)}{2\left|S_1\right|} + \frac{\sum_{t \in S_2} sim'\left(t, g\left(t\right)\right)}{2\left|S_2\right|},$$

where *g(t)* is the aligning function:

$$g\left(t\right) = \arg\max sim'\left(t, t'\right).$$

$sim'\left(t, t'\right)$ is a wrapper function over the relation similarity model. The similarity between two words X and Y is computed by combining LSA similarity and Word Net relations:

$$sim_{\oplus}\left(x, y\right) = sim_{LSA}\left(x, y\right) + 0.5e^{-\alpha D\left(x, y\right)},$$

where D(x, y) is the minimal path distance between x and y in the word net taxonomy.

For penalizing the bad terms, a model had been developed that determine whether two terms belong to disjoint sets:

$$A_i = \left\{\left\langle t, g\left(t\right)\right\rangle \middle| t \in S_i \wedge sim'\left(t, g\left(t\right)\right) < 0.05\right\},$$

$$B_i = \left\{\left\langle t, g\left(t\right)\right\rangle \middle| t \in S_i \wedge \mathrm{t}\ \textit{is an antonym of}\ g\left(t\right)\right\}\ i \in \left\{1, 2\right\}.$$

P' is calculated using the formula:

$$\mathrm{P}_i^A = \frac{\sum_{\left\langle t, g(t)\right\rangle \in A_i}\left(sim'\left(t, g\left(t\right)\right) + w_f\left(t\right).w_p\left(t\right)\right)}{2\left|S_i\right|},$$

$$\mathrm{P}_i^B = \frac{\sum_{\left\langle t, g(t)\right\rangle \in B_i}\left(sim'\left(t, g\left(t\right)\right) + 0.5\right)}{2\left|S_i\right|},$$

$$P' = P_1^A + P_1^B + P_2^A + P_2^B.$$

The semantic similarity value returned by the model among two terms ranges from 0 and 1. The similarity value increases from 0 to 1. In this work we have taken the threshold, r as .4 for filtering out the irrelevant terms. Algorithm, shown in Table 4, presents the implementation detail of the proposed POS context pattern approach.

Table 4. Algorithm for POS Context Pattern.

Algorithm: Word Context Pattern.

```
Input: D_TR – Training Dataset, D_TS – Testing Dataset, S – Seed Word
Output: Agriculture Entities E = Φ

    POS Pattern Extraction
E ⟵ Φ, T ⟵ Φ;
for each $s \in S$ do
Find context window for $W_i$ in D_TR where $i = \pm 5$
```

Algorithm: Word Context Pattern.

```
4.  Extract co-occurred words, $CW_d(i) \leftarrow W_i$
5.  Get POS structure $S_{POS}$ by parsing the co-occurred word
        $S_{POS} \leftarrow Parse(\mathrm{CW}_d)$
6.  Part of speech patterns, $\mathrm{P}_{POS} \leftarrow Frequent(S_{POS})$
    Pattern Evaluation
7.  For each sentence $L_i$ in $\mathrm{D}_{\mathrm{TS}}$ do
8.  Get POS structure $S_{POS}$ by parsing each sentence $L$
        $S_{POS} \leftarrow Parse(\mathrm{L})$
9.  Search n-gram of part of speech þ, where $þ_{\mathrm{POS}} \in S_{POS}$ and $þ_{\mathrm{POS}} := P_{POS}$
10. Extract word token $T$ from the probable position in $þ_{\mathrm{POS}}$.
    Prunning
11. For each token $T_i \in T$ do
12. Semantic Similarity, $Sim \leftarrow Semantic\ Similarity(\mathrm{C}_i, \mathrm{T}_i)$
13. If $Sim \geq Threshold\ \Gamma$
14. Agriculture Entity, $E := E + T$
```

The flow diagram of the proposed model for POS context pattern is shown in Fig. 4.

4 Experiment and Result

4.1 Dataset Preparation

As NER in agriculture domain is an unenlightened field, we have not found any standard or benchmark dataset. Hereof in order to perform the experiments we have to make our own dataset. To prepare the dataset a renowned agricultural resource called AGRIS is employed. AGRIS is a global public domain database with more than 7.5 million structured bibliographical records on agricultural science and technology (https://en.wikipedia.org/wiki/AGRIS). More specifically it is an agricultural search engine which contains the references to the agricultural research articles, data, statistics, and multimedia material. For preparing the dataset we have extracted out the abstracts of the research articles available in the AGRIS database. In this paper five different class entities have been sorted out, which is to be tagged by the proposed NER system. These predefined classes are: Name of cereals, fruits, nuts, spices, and vegetables. For each of these classes we have used 15 varied elements or entities for extracting theabstract. As a result of this we have a set of 75 entities. These seventy-five entities are passed to the AGRIS data base and for each entity top ten abstracts have been extracted. Hence in total, for performing the experiment we have a dataset of 750 documents or abstracts containing 171,735 words. Out of these 15 entities in each class we have used five elements for learning and ten elements for testing. Thus, in total we have used 25 entities or 250 documents for learning and 50 entities or 500 documents for testing. Table 5 enlist the five classes and the entities contained in that classes in the testing dataset.

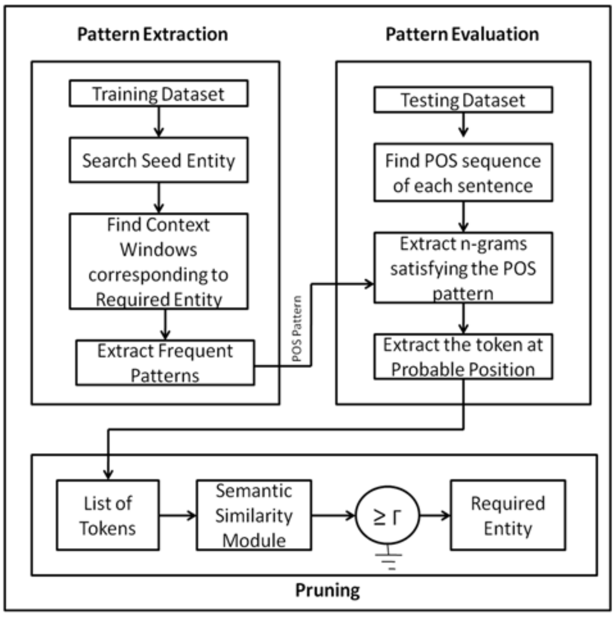
Pattern Extraction
Pattern Evaluation
Training Dataset
Search Seed Entity
Find Context Windows corresponding to Required Entity
Extract Frequent Patterns
POS Pattern
Testing Dataset
Find POS sequence of each sentence
Extract n-grams satisfying the POS pattern
Extract the token at Probable Position
List of Tokens
Semantic Similarity Module
≥ Γ
Required Entity
Pruning

Fig. 4. Flow diagram of POS Context Pattern.

Table 5. Agriculture class Vs Entities per class.

	Class Name				
	Cereal	**Fruit**	**Nut**	**Spice**	**Vegetable**
Entity Name	Oats (101)	Acorn (72)	Almond (70)	Balm (30)	Carob (69)
	Rice (323)	Apple (132)	Cashew (83)	Cinnamon (77)	Peanut (84)
	Rye (129)	Apricot (57)	Chestnut (86)	Clove (66)	Cassava (25)
	Wheat (325)	Blackberry (58)	Coconut (182)	Garlic (166)	Leeks (70)
	Barley (193)	Durian (108)	Hazelnut (64)	Ginger (90)	Okras (18)
	Bean (216)	Kiwi (48)	Pecan (26)	Mint (54)	Parsnips (18)
	Corn (265)	Mango (116)	Pistachios (68)	Paprika (62)	Rhubarb (52)
	Lentils (55)	Nectarine (47)	Walnut (72)	Pepper (98)	Tomato (67)
	Spelt (87)	Pear (231)	Raisin (27)	Sage (64)	Turnip (28)
		Plum (104)		Savory (38)	

4.2 Experiment and Result Analysis

In this paper the motto of the work is to extract the agricultural named entities from the dataset. Experiment 1 tries to find out these entities using the context pattern of words. For this at first the phrases or keywords are extracted which frequently co-occurred with the required named entities. Few of these extracted co-occurred word patterns have been enlisted below in Table 6.

To check the co-occurring accuracy of these word patterns with the agricultural entities, they are searched in the testing dataset and checked the token found in the probable position for named entity. However, after doing the experiment we found that the results obtained are not satisfactory. This is because of the reason that although the co-occurrence of these phrases is frequent with the agricultural entities but the probability of occurrence of these words patterns with the other entities is also very high.

Since the co-occurrence of word patterns are diverse in nature, no significant result has been obtained. Hence in order to devise more generalize co-occurring patterns we have move towards POS patterns.

Same as word patterns, for extracting the POS patterns also, n-gram of different size window is chunked out from the training data set. In this case rather than considering the word tokens the extracted n-gram phrases are parsed for the POS structure. For infesting the POS structure Stanford’s parser has been used. As a result of this a set of finite POS patterns have been obtained, out of which some frequent POS patterns have been sorted out (Table 3).

To check the accuracy of the procured patterns, these POS structures have been searched over the testing dataset. The n-gram of word which satisfies the POS structure is then extracted out and looks for the token in the probable position. At last the list of tokens obtained through the probable position is check for the required entity of interest or not. The number of relevant entities extracted out through each pattern in each class is shown in Table 7.

Table 6. Co-occurred Word Patterns.

S. No.	n- gram	Co-occurring Word Patterns
1.	yield of rice	yield of ____
2.	content in apple	content in ____
3.	production of cashew	production of ____
4.	quality of tomatoes	quality of ____
5.	wheat production	____ production
6.	barley grain	____ grain
7.	wheat flour	____ flour
8.	Almond cultivars	____ cultivars
9.	bean varieties	____ varieties
10.	mint plant	____ plant
11.	corn yield	____ yield
12.	common wheat	common ____
13.	red rice	red ___
14.	wild lentil	wild ____
15.	raw barley	raw ____
16.	fresh paprika	fresh ____
17.	apples produced	____ produced
18.	ginger grown	____ grown
19.	cultivated rice	cultivated ____
20.	polished barley	polished _____

Table 7. Relevant vs. Irrelevant entities extracted through POS patterns without semantic similarity measure.

POS Pattern	Cereal		Fruit		Nut		Spice		Vegetable	
	Rv	IRv	Rv	IRv	Rv	IRv	Rv	IRv	Rv	Irv
NN + **IN** + *NN*	385	1055	180	690	173	634	143	576	158	584
NN + **NN**	866	2506	382	1864	347	161	208	1421	166	1271
JJ + *NN*	272	1048	135	874	109	691	128	782	36	635
VB+ *NN*	154	501	70	403	71	391	15	75	40	310
NN + **VB**	225	1607	141	1266	116	986	140	1018	105	1105
IN + *NN* + **IN**	41	305	22	201	19	193	11	169	24	166
NN + **IN** + **VB** + *NN*	42	54	15	30	16	39	9	30	5	27

Here in Table 7, Rv referred to the relevant entities and IRv referred to the irrelevant entities extracted through patterns. Total number of entities extracted is the summation of all relevant and irrelevant entities extracted. The accuracy of the system can be measured using the formula:

$$\text{Precision}=\frac{Total\ no.of\ relevant\ entities\ extracted}{Total\ no.of\ entities\ extracted}.$$

Numerated value of precision for each of the patterns is enlisted in Table 8.

Table 8. Precision value of POS patterns without similarity measure.

POS Pattern	Cereal	Fruit	Nut	Spice	Vegetable
NN + **IN** + *NN*	26.73	20.68	21.43	19.88	21.29
NN + **NN**	25.68	17.0	68.30	12.76	11.55
JJ + *NN*	20.60	13.37	13.62	14.06	5.36
VB + *NN*	23.51	14.79	15.36	16.66	11.42
NN + **VB**	12.28	10.02	10.52	12.08	8.67
IN + *NN* + **IN**	11.85	9.86	8.96	6.11	12.63
NN + **IN** + **VB** + *NN*	43.75	33.33	29.09	23.07	15.62

Fig. 5 represents the accuracy or precision curve for each of the proposed POS patterns shown in Table 8 for each of the varied class of agricultural elements without semantic similarity. It can be observed through Table 8 and the precision curve plotted in Fig. 5 that the precision value for the proposed POS patterns is not appreciable. The precision is lousy because of the occurrence of the irrelevant entities therewith the relevant entities. Hence in order to improve the accuracy of the system the irrelevant entities must be filtered out.

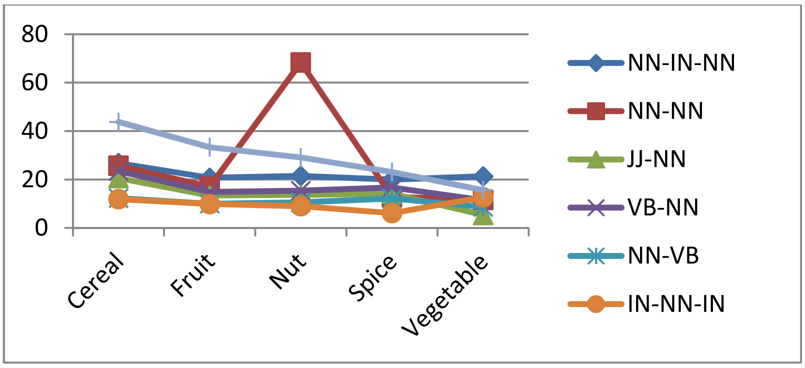

Fig. 5. Precision value of POS patterns (without semantic similarity measure).

To filter out the irrelevant entities from the list of extracted entities, semantic similarity measure has been used. We have employed the UMBC semantic similarity module to achieve this. The semantic similarity value of each term extracted out through patterns is measured with its class keyword that is cereal, fruit, nuts, spice, and vegetable. Terms for which the similarity measure is greater than or equal to .4 are kept as relevant named entities while the others are filtered out. The number of relevant and irrelevant entities obtained after the application of semantic similarity measure is enlisted in Table 9.

Table 9. Extracted Relevant vs. Irrelevant entities through POS patterns with semantic Similarity measure.

POS Pattern	Cereal		Fruit		Nut		Spice		Vegetable	
	Rv	Irv	Rv	Irv	Rv	Irv	Rv	Irv	Rv	Irv
NN + **IN** + *NN*	363	37	147	61	134	52	76	47	123	48
NN + **NN**	806	142	317	150	290	110	120	99	103	105
JJ + *NN*	161	33	115	40	45	20	81	19	20	21
VB + *NN*	147	34	58	21	64	18	32	20	32	20
NN + **VB**	220	64	105	98	92	60	68	51	97	44
IN + *NN* + **IN**	34	5	18	7	17	10	7	1	19	7
NN + **IN** + **VB** + *NN*	34	3	14	3	11	8	9	5	5	3

Table 10 represents the precision value of each of the proposed POS patterns derived from Table 9 for each of the varied class of agricultural elements with semantic similarity.

Table 10. Precision value of POS patterns with similarity measure.

POS Pattern	Cereal	Fruit	Nut	Spice	Vegetable
NN + **IN** + *NN*	90.75	70.67	72.04	61.78	71.92
NN + **NN**	85.02	67.88	72.5	54.79	49.51

JJ + *NN*	82.98	74.19	69.25	81.0	48.78
VB + *NN*	81.21	73.41	78.04	61.53	61.53
NN + **VB**	77.46	51.72	60.52	57.14	68.79
IN + *NN* + **IN**	87.17	72.0	62.96	87.5	73.07
NN + **IN** + **VB** + *NN*	91.89	82.35	57.89	64.28	62.5

The precision curve for the POS pattern with semantic similarity measure for each class of agriculture entity is illustrated in Fig 6.

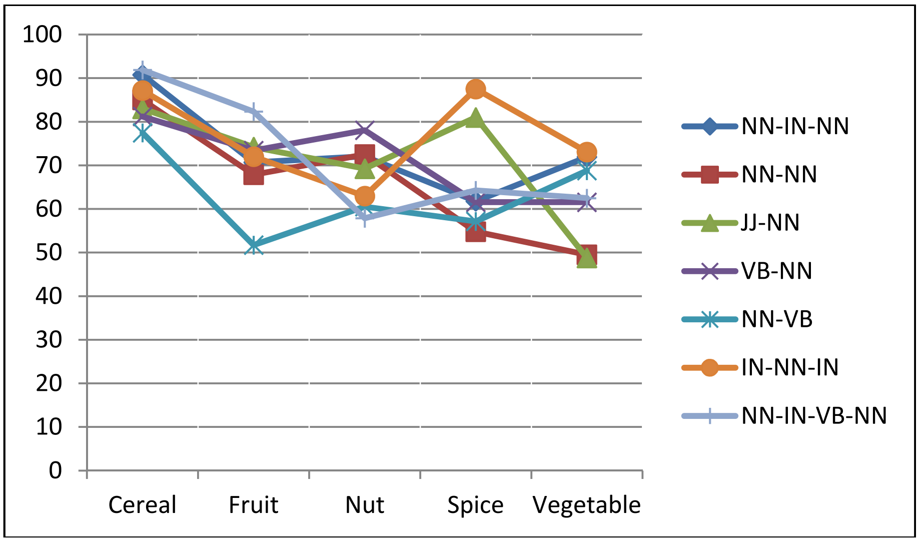

Fig. 6. Precision value of POS patterns (with semantic similarity measure).

It can look over through Table 8 and 10 and Fig. 5 and 6 that the application of semantic similarity has drastic effect on upgrading the performance of the system. The average precision of each pattern with and without similarity measure is framed in Table 11.

Table 11. Average precision for proposed POS patterns.

POS Pattern	Without Similarity Measure	With Similarity Measure
NN + **IN** + *NN*	22.00	73.43
NN + **NN**	27.05	65.94
JJ + *NN*	13.38	71.23
VB + *NN*	16.34	71.14
NN + **VB**	10.71	63.12
IN + *NN* + **IN**	9.88	76.54
NN + **IN** + **VB** + *NN*	28.97	71.78

The comparison of precision value of POS patterns for both that is without similarity measure and with similarity measure is shown in Fig. 7. As a result of application of semantic similarity measure the average accuracy of the system increases from 18.33% to 70.45%. The recall value of the system can be scaled using the formula:

$$\text{Recall}=\frac{Total\ no.of\ relevant\ entities\ extracted}{Total\ no.of\ relevant\ entities\ in\ the\ dataset}.$$

However, it cannot be calculated separately for each of the patterns. This is because of the overlapping nature of few of the POS patterns. Hence the recall value is accounted for the entire system rather than for varied separate patterns. In order to figure out this the entire data set is scanned for all the agriculture entity present in the data considering both hidden and known (fifty seed entities) agricultural entities. The recall is computed on the basis of out of all the relevant agricultural entities available in the data set, how many are extracted through patterns. It is found that there are 115 unique agricultural entities present in the data set and out of these 105 relevant agricultural named entities are extracted through patterns. Hence the recall value is obtained as 91.3%. The surpassing achievement is that the proposed patterns has extracted out fifty-three hidden agricultural named entities from the dataset in addition of fifty known seed entities.

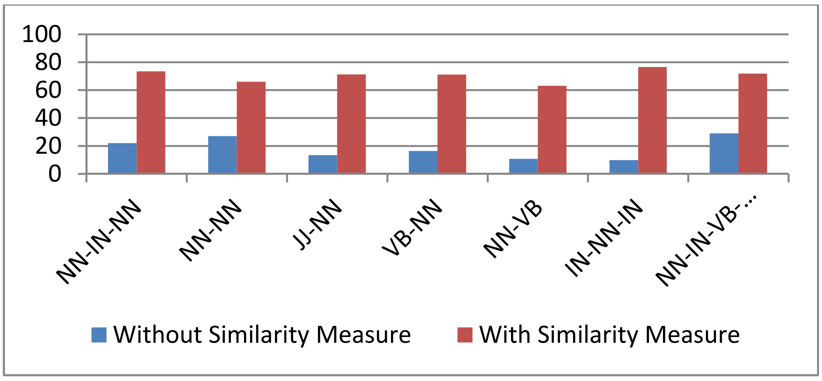

Fig. 7. Precision without similarity measure Vs Precision with similarity measure.

5 Conclusion

This paper focuses on developing an NER system for agriculture domain. For achieving this context pattern-based method has been used. The context has found using the co-occurrence pattern in different size window. Two type of context has been considered, word context pattern and POS context pattern. The output of the word context pattern is diverse in nature hence no significant result has obtained. Thus, the next experiment focuses on POS context pattern, which provide more generalize results. This approach receives good recall, but due to the impedance of bulk of irrelevant entities with the relevant ones, the accuracy of the system is very poor. Hence in order to filter out the irrelevant entities and improve the accuracy of the system semantic similarity measure is used. The experiment is deficient in a point that the semantic measure is relied over the word net database. Hence only those entities will be entertained for similarity measure which has enlisted in the word net database. Although the experimental results are not very surprising, but being a preparatory work in the field of agricultural NER, this can be an impellent for future work in this blooming topic of NER in agriculture domain.

References

1. Al-Jumaily, H., Martínez, P., Martínez-Fernández, J.L., Van der Goot, E.: A real time Named Entity Recognition system for Arabic text mining. Language Resources and Evaluation 46(4), 543–563 (2012)
2. Biswas, P., Sharan, A., Kumar, A.: AGNER: Entity tagger in agriculture domain. In: Proceeding of 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1134–1138 (2015)
3. Biswas, P., Sharan, A., Kumar, R.: Question Classification using syntactic and rule based approach. In: Advances in Computing, Communications and Informatics (ICACCI, 2014) International Conference on. IEEE, pp. 1033–1038 (2014)
4. Biswas P., Sharan A., Verma S.: Named entity recognition for agriculture domain using word net. Int J Comput Math Sci. 5(10), 29–36 (2016)
5. Bollegala D., Matsuo Y., Ishizuka M.: A web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge and Data Engineering 23(7), 977–90 (2010)
6. Collier, N., Nobata, C., Tsujii, J.: Extracting the names of genes and gene products with a hidden Markov model. In: Proceedings of COLING, pp. 201–7 (2000)
7. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence 165(1), 91–134 (2005)
8. Fleischman, M.: Automated subcategorization of named entities. In: ACL (Companion Volume), pp. 25–30 (2001)
9. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: 16th International Conference on Computational Linguistics, Vol. 96, pp. 466–477 (1996)
10. Haiyan, C.: Measuring semantic similarity between words using web search engines. Computer Science 42(2), 261-–267 (2015)
11. Han, L., Kashyap, A., Finin, T., Mayfield, J., Weese, J.: UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In: 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Vol. 1, pp. 44–52 (2013)
12. Kazama, J., Makino, T., Ohta, Y., Tsujii, J.: Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the workshop on natural language processing in the biomedical domain, pp. 1–8 (2002)
13. Lamurias, A., Ferreira, J., Couto, F.M.: Chemical named entity recognition: Improving recall using a comprehensive list of lexical features. In: 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014), pp. 253–260. Springer International Publishing (2014)
14. Lee, S., Lee, G.: Heuristic methods for reducing errors of geographic named entities learned by bootstrapping. In: International Joint Conference on Natural Language Processing, pp. 658–669. Springer, Berlin, Heidelberg (2005)
15. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering 15(4), 871–82 (2003)
16. Lin, Y.F., Tsai, T.H., Chou, W.C., Wu, K.P., Sung, T.Y., Hsu, W.L.: A maximum entropy approach to biomedical named entity recognition. In: Proceedings of the 4th International Conference on Data Mining in Bioinformatics. pp. 56-61. Springer-Verlag (2004)
17. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Human

Language Technologies, Vol. 1, pp. 359–367. Association for Computational Linguistics (2011)

18. Malarkodi, C.S., Lex, E., Devi, S.L.: Named Entity Recognition for the Agricultural Domain. Research in Computing Science 117, 121–132 (2016)
19. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Vol. 4, pp. 188–191. Association for Computational Linguistics (2003).
20. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), pp. 3–26 (2007)
21. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet: Similarity: measuring the relatedness of concepts. In: Demonstration papers at hlt-naacl 2004, pp. 38–41. Association for Computational Linguistics (2004)
22. Pekar, V., Steffen, S.: Word classification based on combined measures of distributional and semantic similarity. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Vol. 2, pp. 147–150. Association for Computational Linguistics (2003)
23. Rau, L.F.: Extracting company names from text. In: 7th IEEE Conference on Artificial Intelligence Application, Vol. 1, pp. 29–32 (1991)
24. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
25. Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534. Association for Computational Linguistics (2011)
26. Rizzolo, N., Roth, D.: Learning Based Java for Rapid Development of NLP Systems. In: LREC Vol. 5, pp. 313–323 (2010)
27. Saha, S.K., Sarkar, S., Mitra, P.: Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of biomedical informatics 42(5), 905–911 (2009)
28. Seker, G.A., Eryigit, G.: Initial Explorations on using CRFs for Turkish Named Entity Recognition. In: COLING, pp. 2459–2474 (2012)
29. Sun, C., Guan, Y., Wang, X., Lin, L.: Rich features based Conditional Random Fields for biological named entities recognition. Computers in Biology and Medicine 37(9), 1327–1333 (2007)
30. Takeuchi, K., Collier, N.: Use of support vector machines in extended named entity recognition. In: Proceedings of the sixth conference on natural language learning (CONLL), pp. 119–25 (2002)
31. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in biomedical text. Bioinformatics 18(8), 1124–1132 (2002)
32. Tohidi, H., Ibrahim, H., Murad, M.A.A.: Improving named entity recognition accuracy for gene and protein in biomedical text literature. International journal of data mining and bioinformatics 10(3), 239–268 (2014)
33. Tsai, R.T.H., Wu, S.H., Chou, W.C., Lin, Y.C., He, D., Hsiang, J., Sung, T.Y., Hsu, W.L.: Various criteria in the evaluation of biomedical named entity recognition. BMC bioinformatics 7(1), p. 92 (2006)
34. Zhang, J., Shen, D., Zhou, G., Su, J., Tan, C.L.: Enhancing HMM-based biomedical named entity recognition by studying special phenomena. Journal of biomedical informatics 37(6), 411–422 (2004)
35. Zhou, G.D., Su, J.: Named Entity Recognition using an HMM-based Chunk Tagger. In: Proceedings of the 40th annual meeting of the association for computational linguistics (ACL), pp. 473–480 (2002)