Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 1337–1345

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques

Veena Gangadharan[a], Deepa Gupta[b]

[a] *Dept of Computer Science and Applications,Amrita Vishwa Vidyapeetham ,Amritapuri,India*
[b] *Dept of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham,India*

Abstract

Named Entity Recognition (NER) is one of the fundamental process in Natural Language Processing applications. In this paper, we propose an Agriculture Named Entity Recognition using Topic Modelling techniques (AERTM Algorithm). In the agriculture domain, we have identified Names of Crops, Soil Types, Names of Pathogen, Crop Diseases and Fertilizers as the key entities. Our work presents a hybrid approach using the agriculture vocabulary AGROVOC and the AERTM algorithm. We used AGROVOC for identifying crop names. But it failed to identify Soil Types, Crop Diseases and Fertilizers. Hence, for those entities we propose a Latent Dirichlet Allocation (LDA) based topic modelling algorithm. These named entities can be used for creating a knowledge base which can be further used mainly in Relation Extraction systems, forums supported by various Government distinguished repositories, etc. Because of the absence of benchmark agriculture data, we tested our model using 3000 sentences extracted from reputed agriculture sites. Human evaluation of the method confirms that our approach gives an accuracy of 80%.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).

Keywords: Agriculture; Named Entity Recognition; AGROVOC; Crops; Soils; Fertilizer; Diseases; LDA; Topic Modeling.

1. Introduction

NLP Applications mainly consists of Question-Answering and Relation Extraction Systems. In an open domain, named entities are Person Name, Place Name, Organization and Date. In Question-Answering systems, these entities are the answers to most of the factoid type questions. The major focus of Relation Extraction systems is to extract relations present between these named entities. NER is a challenging problem in NLP and large amount of work has been done in this area. There exist different methods for NER such as Dictionary based, Rule based and Corpus based. Dictionary (Terminology) based NER uses machine readable dictionary such as Wordnet for searching the

E-mail address: veenag@am.amrita.edu,g_deepa@blr.amrita.edu

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).
10.1016/j.procs.2020.04.143

definition of word present in the sentence. Since it is a lexical database it is difficult to search noun phrases or multi-words. Manually created rules in the form of regular expressions are used in the case of Rule based NER systems. This method is time consuming since it uses human experts for rule creation. These rules are particularly restricted to sentences in a particular domain. Corpus-based NER methods are based on domain specific annotated corpora and entities are predicted using ML algorithms. The applicability of this trained model in other domain is extremely difficult. In this paper, we present an unsupervised approach for identifying the named entities that aims to bridge this gap. For our study we have taken documents related to agriculture in Kerala State.

In Indian economy, agriculture plays a very important role since most of the Indian population depends on agriculture and agro-industries. For various information such as crops and soil, soil management, irrigation, use of fertilizers and farming practices, crop diseases and fertilizers farmers depend on agriculture department. The forums provided by the agriculture sites are the primary access point to the farmers. There exist a lots of documents proposed by state and central government to provide useful information to the farmers. In most of the forums, these queries are managed manually and it is time consuming. Such forums are very effective if it could be able to automatically extract information from the documents. These e-services need to extract proper features from unstructured documents to answer the farmers queries. Majority of these services depend on annotated corpus or knowledge bases. Problem of information extraction for domains such as agriculture is particularly challenging due to non-availability of any tagged corpus. Our proposed model automatically extract the related words present in a document and bridges this gap. This work presents a hybrid approach based on Dictionary and Machine Learning algorithm for named entity Recognition in agriculture documents. Our main focus in this work is to extract named entities with out using an annotated corpus. Unsupervised machine learning algorithms are characterized by their ability to extract semantically meaningful features from raw data. Our model identified domain specific entities like Crops, Soil Types, Names of Pathogen, Crop Diseases and Fertilizers. Because of the absence of benchmark agriculture data, we extract useful information from agriculture related government sites.

This paper is organized as follows. Section 2 presents a review of NER systems in agriculture documents. In Section 3 proposed algorithm is discussed. Results and Evaluations are presented in section 4. Section 5 concludes the paper with possible future works.

2. Related Past Works

There are number of models proposed in open domain [1] [2] [3] for Named Entity Recognition and these models are unable to detect agriculture related phrases. There are limited number of tools and models proposed in agriculture domain for Entity tagging. Most of the methods are particularly for identifying crop names. AGROVOC [4] was developed in the 1980s as a multilingual structured thesaurus for subject fields in agriculture, forestry, fisheries, food and related domains. It is published by Food and Agriculture Organization (FAO) and edited by a community of experts. Using this vocabulary search we can identify most of the crop names in Indian agriculture. But this vocabulary is unable to detect some phrases of crop names, (Ash Gourd,Yellow Cucumber,Chow Chow, Elephant-ear, Eddoe,etc.) names of soil, (Red Soil, Coastal Alluvium, Acid Saline soils, Kari Soils,etc) most of the names of crop diseases (Sheath Blight, Bacterial Blight, Rice Blast, etc.) in Indian agriculture. Chatterjee N, Kaushik N proposed RENT [5]. They used domain specific regular expression to extract terms present in the documents. This approach uses human experts for writing regular expressions. Automatic NE gazetteers using a variant of Multiword Expression Distance (MED) is used in [6]. There are three main NE tags in their system, namely crop, disease and chemical treatment. The Gazetteers were generated automatically for each NE type using similarity of new terms with candidate phrases. In the work [7], it is proposed to include AGNER,a NER system in Agriculture domain. They used AGROVOC hierarchy for domain specific knowledge. Each term is labelled with fine grain label and another coarse grain label. This method suffer the disadvantages of AGROVOC. The work presented in [8] propose NER methods using Wordnet. Wordnet is a dictionary of English words where semantic relationships of words are expressed by graph structure. Many phrases in Agriculture domain are not included in the Wordnet. Machine learning algorithms for NER is proposed in [9]. Conditional Random fields are used for entity classification. They have created a Named Entity tag set consisting of 19 fine grained tags. CRF is computationally complex and makes it very difficult to re-train the model when newer data becomes available.

The major focus of our work is to label the agriculture terms and phrases using an unsupervised topic modelling algorithm called LDA. One advantage of our AERTM algorithm is the lack of hand-engineered features. Despite using the domain specific regular expressions, our model is competitive with existing methods.

3. Proposed Method

For agriculture NER, we propose a hybrid approach that takes the advantages of AGROVOC and machine learning based algorithm called LDA. AGROVOC is agriculture vocabulary consists of over 36,000 concepts available in 33 languages. Using this agriculture vocabulary we could be able to check whether a particular phrase is crop name or not. It failed to recognize some of the local crop names of Kerala, so an additional agriculture Gazetter is maintained in our work. To tag a particular phrase as CROP we used AGROVOC and the Gazetter and for the other entities we have used the AERTM Algorithm. This algorithm output the clusters of related words automatically.

There are two phases in our AERTM algorithm, the *Training Phase* and *Testing Phase*. In the first phase the LDA model is trained using the *Agriculture Dataset or the trainset*.

Figure 1 shows proposed system architecture and the AERTM algorithm is explained in Algorithm 1. The complexity of AERTM algorithm is O(n2).

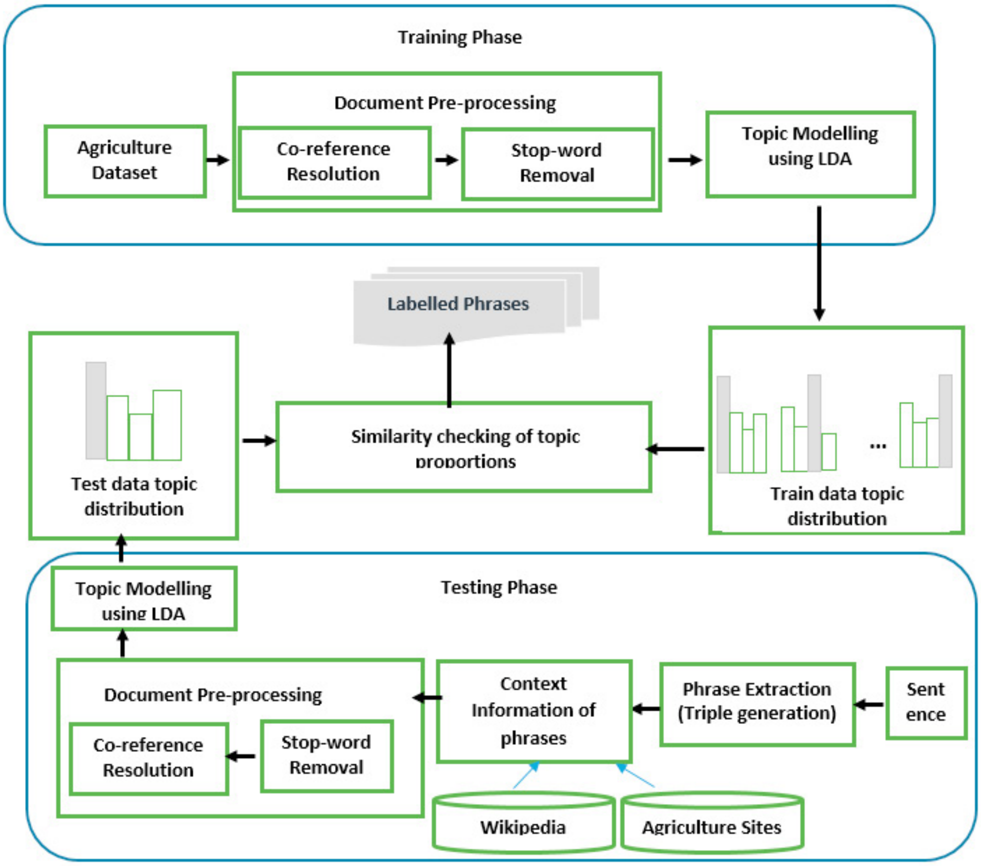


Fig. 1: Schematic Diagram of Agriculture Entity Recognition using Topic Modelling

Whenever a new phrase came for tagging, first we check whether it is a CROP in the AGROVOC and in the Gazetter. If it is not there context information is extracted from recognized government agriculture sites. After preprocessing its is stored as *testset*. In the second phase the LDA model is tested with *testset*. Similarity of *testset* and *traintset*is computed and based on that phrase is labelled.

The following subsections describe the details of our proposed method.

3.1. Training data preparation

In this work we created our own data set pertaining to agriculture topics DISEASE, SOIL AND FERTILISER by crawling the agriculture related websites like

`https://farmer.gov.in/StateAgriDepartments.aspx,`

`http://keralaagriculture.gov.in/`

`http://www.kissankerala.net/home.jsp`

`http://krishisewa.com/articles/disease-management/444-diseases-rice.html`

This *trainset* is used by the AERTM algorithm for learning the hidden parameters.

3.2. Document Pre-processing

This step is common to both training and testing the data sets. Document pre-processing including sentence segmentation and Corefernce resolution. Coreference resolution is the task of finding related expressions of an entity in a text. All the pronouns in the text document are replaced by its related entity by Corefernce resolution. There has been extensive methods for Corefernce resolution systems [10] [11] [12]. This work used the open source Standford Corefernce resolution.

3.3. Topic Modelling using Latent Dirichlet Allocation

This phase is considered as a muli-label classification problem. Here the class labels are DISEASE, SOIL and FERTILIZER. We have used Latent Dirichlet Allocation (LDA) [13] [14] [15] technique for Topic Modelling. LDA map the given document to a set of topics and the words in the document is captured by these imaginary topics. In our work, the model is trained using unlabelled agriculture data set. In LDA each document is considered to have a set of various topics and these topics are produced from a mixture of words. For identifying named entities in the document, a words topic distribution provides useful information. In probabilistic Graphic Models dependencies among variable names are defined as equation 1:

$$p(\beta, \theta, z, w \mid \alpha, \eta) = \prod_{i=1}^{K} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha)(\prod_{n=1}^{N} p(z_d, n|\theta_d) p(w_{d,n} |\beta_1 : k, z_{d,n})) \tag{1}$$

α and η are parameters of the Dirichlet prior on the per-document topic and per-topic word distributions. K represents the total number of topics in the Corpus and for a particular topic β_i represents the word distribution. D is the total number of documents in the collection. The topic distribution for a particular document d is denoted by θ_d. N is the total number of words in the document d. The observed word in the document is denoted by $w_{d,n}$ and per-word topic assignment is denoted by $z_{d,n}$ where n represents the current word. The only observed parameter in this model is $w_{d,n}$. From this observed word w , we compute the posterior distribution of other parameters. The purpose of our model is to compute what is the contribution of each topic to create the document. Topic with highest distribution is more significant to the dataset. Our algorithm outputs K topics from the dataset and we choose the major three topics related with the topics DISEASE, FERTILIZER and SOIL. When a noun phrase with all relevant sentences came for entity tagging the AERTM algorithm calculates its probability distribution with respect to existing data sets.

3.4. Test data Preparation

This module accepts unlabelled sentences and its context information is extracted from web. First the input sentence is parsed and its major phrases are identified. These extracted phrases are considered as candidates for generating the context information. Figure 2 shows the dependency tree generated for a given input sentence. Using the Triple generation [16] we identified all the phrases present in that sentence. In the above example the noun phrases extracted are *sheath blight* and *Rhizoctonia solani.*

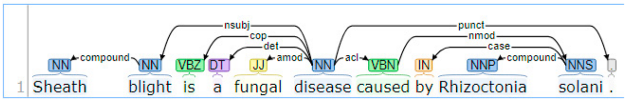


Fig. 2: Dependency Tree generated for the given input *Sheath blight is a fungal disease caused by Rhizoctonia solani.*

The major sentences pertaining to the above phrases are extracted from agriculture sites and named as *testset*. This module is executed in two steps:

- Step1: It searched for the phrases in AGROVOC. If the entries are not found in AGROVOC step 2 is executed else the Phrase is labelled as CROP.
- Step2: It identified all the sentences which have is-a/is-type-of relations with phrase. We selected top 10 sentences from the result and labelled as testset. Table 1 shows the testset.

Table 1: Sample testset for the phrase *Sheath blight*

1.Sheath blight is a fungal disease caused by Rhizoctonia solani

2. sheath blight is one of the most economically significant rice diseases

3.sheath blight is a major soil borne disease causing economic losses

4.Sheath blight of paddy is one of the most widely spreading diseases of paddy

5.Sheath blight disease usually appears in the later growth stages of the plant.

Algorithm 1 Entity Recognition using Topic Modelling techniques -The AERTM Algorithm

```
Result: Labelled Noun phrases
Let doc represents the input;
 Let trainset represents the agriculture data set extracted from agricultre web sites.
 Let $S_i$, $i$=1 to $N$ be the set of sentences in doc where $N$ is the total number of sentences in doc;
 Step1:Apply co-reference resolution on doc;
 Step2:Remove stop words present in doc;
 while i is less than N do
     Step3:Identify the noun phrases present in si and store in a list named nplist; $nplist_i$, $i$=1 to $M$ where $M$ represents
       number of phrases in sentence;
     Step4: Apply LDA on trainset and store the topic proportions in $tp_i$ $i$=1 to $T$ where $T$ represents number of
       training set;
     while nplist is not empty do
         if $nplist_i$ is in AGROVOC then
             Label the phrase as CROP
         else
             Step5: Extract sentences with is-a/is-type of relations with nplist(i)
               and store in a new doc testset;
               Step6: Apply LDA on testset and store the topic proportions in q;
               Step7:Compute similarity of testset topic proportions with trainset using KLdivergence($tp_i$,q);
               Step8:Based on similarity in step7, phrases are labelled with Label of Topics.
         end
     end
end
```

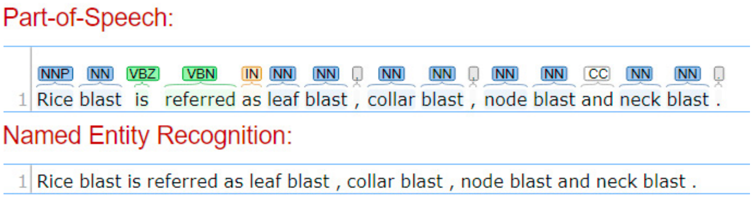
3.4.1. Similarity checking of topic distributions

The KullbackLeibler divergence (KL divergence) [17] [18] [19] is used for identifying the similarity between topic distributions. The KL divergence is a measure of how one probability distribution is different from a second, reference probability distribution. We want to evaluate how similar the *testset* distribution to thetrainset distribution. Let p and q are two topic distributions of documents *d1* (Trainset) and *d2*(Testset). Then the similarity of p and q is calculated in equation 2 as.

$$D_{KL}(p\,||q) = \sum_{i=1}^{N} p(x_i).log\frac{p(x_i)}{q(x_i)} \tag{2}$$

4. Results and Evaluation

For our experiments, we train an LDA model using 3000 sentences extracted from different recognized agriculture sites. Figure 3 shows result from Standford NER for the sentence ***Rice blast*** *is referred as* ***leaf blast, collar blast, node blast*** *and* ***neck blast.***This open domain NER system is unable to detect entities like ***Rice blast, leaf blast, collar blast, node blast and neck blast.*** Figure 4 shows output from our AERTM Algorithm. This algorithm output labelled phrases as well as keywords.



Part-of-Speech:

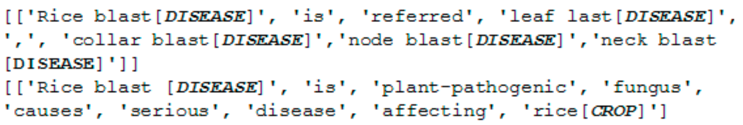
NNP NN VBZ VBN IN NN NN . NN NN . NN NN CC NN NN .

1 Rice blast is referred as leaf blast , collar blast , node blast and neck blast .

Named Entity Recognition:

1 Rice blast is referred as leaf blast , collar blast , node blast and neck blast .

Fig. 3: Result of Standford NER



```
[['Rice blast[DISEASE]', 'is', 'referred', 'leaf last[DISEASE]',
',', 'collar blast[DISEASE]','node blast[DISEASE]','neck blast
[DISEASE]']]
[['Rice blast [DISEASE]', 'is', 'plant-pathogenic', 'fungus',
'causes', 'serious', 'disease', 'affecting', 'rice[CROP]']
```

Fig. 4: Result of AERTM algorithm

Topic Coherence is used to evaluate the topic models [20]. It is defined as the average of the pairwise word-similarity scores of the words in the topic. Each topic in the model consists of words, and the topic coherence is applied to the top N words. A good model will generate topics with high topic coherence scores. Figure 5 shows topic coherence of our LDA model with varying number of topics. In our model its shows good coherence when the number of topics is 3.

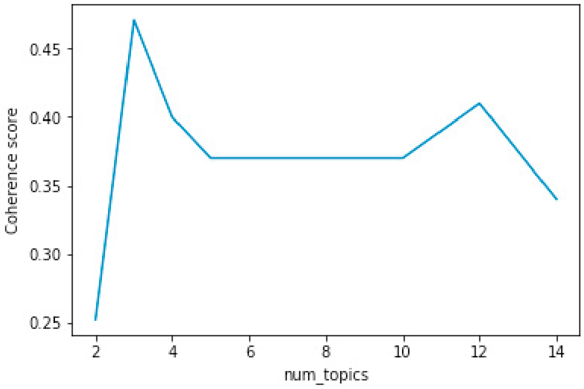


Fig. 5: Topic Coherence

To evaluate Topic models there are many approaches such as Perplexity. It is statistical measure of how well a sample is predicted using the probabilistic model. Perplexity of our model is -6.25. Best model is the one with the lowest perplexity.

Since LDA is an unsupervised technique, there exists no prior information on the number of topics in our corpus. We used LDA visualization tool pyLDAvis and assessed topic models. Figure 6 shows visualization of topics generated by the pyLDAvis inter active chart. The produced topics and the associated keywords are shown in the figure. Each topic is represented by a bubble. The more prevalent topic is shown by larger bubble. In the left hand side of the plot, the big non-overlapping bubbles represents the prominent topics in our document and the topic per word distribution is shown on the right hand side of the plot. The words and bars on the right-hand side depend on the selection of bubbles on the left- hand side. Figure 7 shows selected topic and corresponding keywords presented in Figure 6.

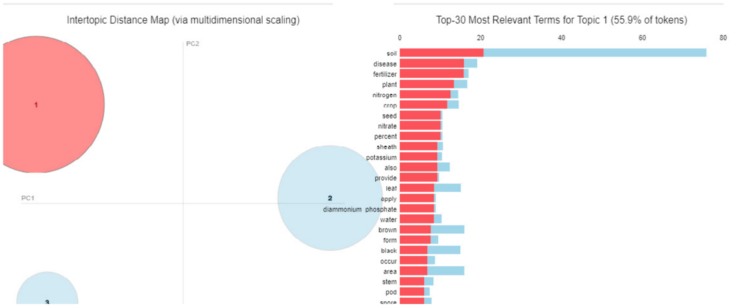


Fig. 6: Intertopic Distance Map with number of topics 3

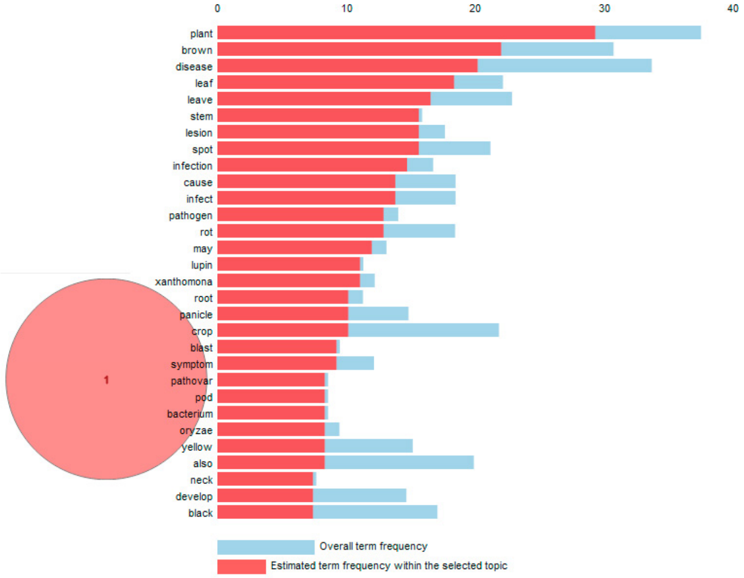
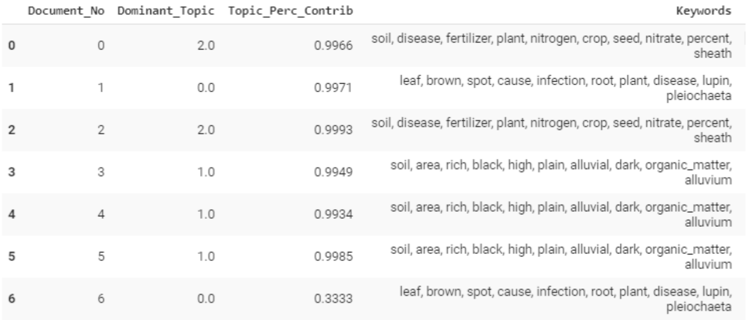


Fig. 7: Topic-Keywords

Figure 8 shows list of dominant topics and keywords present in that topic. It has the Dominant topic, contribution column and the keywords in the document.



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords
0	0	2.0	0.9966	soil, disease, fertilizer, plant, nitrogen, crop, seed, nitrate, percent, sheath
1	1	0.0	0.9971	leaf, brown, spot, cause, infection, root, plant, disease, lupin, pleiochaeta
2	2	2.0	0.9993	soil, disease, fertilizer, plant, nitrogen, crop, seed, nitrate, percent, sheath
3	3	1.0	0.9949	soil, area, rich, black, high, plain, alluvial, dark, organic_matter, alluvium
4	4	1.0	0.9934	soil, area, rich, black, high, plain, alluvial, dark, organic_matter, alluvium
5	5	1.0	0.9985	soil, area, rich, black, high, plain, alluvial, dark, organic_matter, alluvium
6	6	0.0	0.3333	leaf, brown, spot, cause, infection, root, plant, disease, lupin, pleiochaeta

Fig. 8: List of dominant topics and Keywords

Per topic word distribution is shown in Figure 9. The X-axis shows the most relevant keywords in a particular Topic and Y-axis shows the number of count of keywords. From this figure the most prevalent keywords can be extracted.

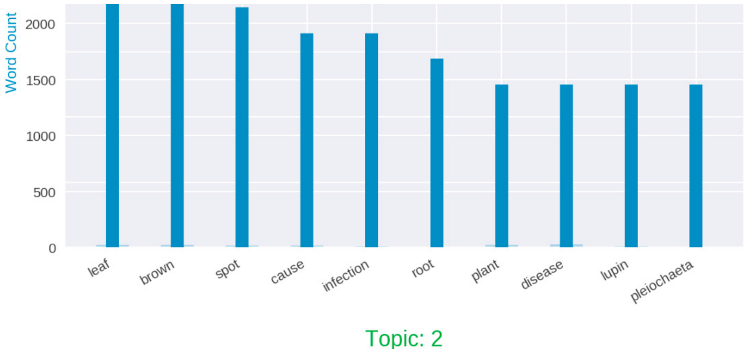


Fig. 9: Topic-Word Distribution

Table 2 shows some of the phrases identified by our algorithm.

Table 2: Entities identified

Disease	Fertilizer	soil
['node', 'blast']	['diammonium', 'phosphate']	['yellow', 'clayey', 'soil']
['bacterial', 'leaf', 'streak']	['monoammonium', 'phosphate']	['dark', 'topsoil']
['sheath', 'blight']	['nitrogen']	['carbonate', 'planosols', 'soil']
['sheath', 'brown', 'rot']	['phosphorus']	['alluvial', 'deposit', 'gleysols', 'soil']
['brown', 'leaf', 'spot']	['triple', 'super', 'phosphate']	['dark', 'brown', 'topsoil']
['root', 'rot']	['ammonia', 'component']	['saline', 'soil']
['dark', 'brown', 'spot']	['phosphorus', 'fertilizer']	['red', 'soil']

In the first level of the AERTM algorithm all the disease related words like names of bacteria, fungus are grouped under a single cluster.

In order to differentiate a phrase as PATHOGEN or DISEASE, a second level of search in another dictionary of Pathogens is conducted. This Gazetter for Pathogens is created by extracting 2500 bacterial names published in the *International Journal of Systematic Bacteriology.*Some of the entries in the Gazetter created for PATHOGEN are shown in Table 3.

Table 3: Gazetter for PATHOGEN

Name of Bacteria
['actinoplanes', 'kalakoutskii']
['actinoplanes', 'brasiliensis']
['actinoplanes', 'deccanensis']
['actinoplanes', 'beretta']
['actinoplanes', 'philippinensis]'

5. Conclusion and Future Works

In this work our attention is placed particularly for tagging agriculture terms and phrases using an unsupervised approach called LDA and AGROVOC. The algorithm is tested using 3000 sentences. We identified Names of Crops, Soil Types, Crop Diseases, Names of Pathogen and Fertilizers present in a document specific to Agriculture. The tagged phrases can be used for creating a knowledge base in agriculture domain, which can be further used mainly in Relation Extraction Systems, Question-Answering in Agriculture Forums supported by various government distinguished repositories, etc. Human evaluation of the method confirms that our approach gives an accuracy of 80%. One of the first future lines of work regarding the proposed method is to explore the topic clusters and identity the relation exists between the named entities.

References

[1] Trond Grenager Jenny Rose Finkel and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

[2] Soman K.P Premjith B Gopalakrishnan, A. A deep learning-based named entity recognition in biomedical domain. Lecture Notes in Electrical Engineering, Springer Verlag, Volume 545, p.517-5262019 (2018).

[3] Fousiya Anand Kumar Soman K.P Prasad, Gowri. Named entity recognition for malayalam language: A crf based approach. Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on, IEEE, Chennai.

[4] http://aims.fao.org. URL http://aims.fao.org/vest-registry/vocabularies/agrovoc.

[5] Kaushik N Chatterjee N. Rent: regular expression and nlp-based term extraction scheme for agricultural domain. Proceedings of the international conference on data engineering and communication technology 2017.

[6] Sachin P. Girish K.P Patil, A. Named entity extraction using information distance. International Joint Conference on Natural Language.

[7] Ashish Kumar Payai Biswas, Aditi Sharan. Agner: Entity tagger in agriculture domain. Second International Conference on Computing for Sustainable Global Development 2015.

[8] Ashish Kumar Payai Biswas, Aditi Sharan. Named entity recognition for agriculture domain using word net. International Journal of Computer and Mathematical Sciences.

[9] Sobha Lalitha Devi1 Malarkodi C, Elisabeth Lex. Named entity recognition for the agricultural domain. Research in Computing Science 2016.

[10] Deepa Gupta Anna Neethu Daniel S. Roshny Veena, G. A learning method for coreference resolution using semantic role labeling features. International Conference on Advances in Computing, Communications and Informatics 2017.

[11] Sruthy Krishnan Veena, G. A concept based graph model for document representation using coreference resolution. Springer Intelligent Systems Technologies and Applications 2016.

[12] Kevin Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. URL AssociationforComputationalLinguistics(ACL)2016.

[13] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003.

[14] Intuitive guide to latent dirichlet allocation. URL https://towardsdatascience.com/.

[15] Introduction to latent dirichlet allocation. URL https://blog.echen.me.

[16] Lekha N.K Veena, G. A concept based clustering model for document similarity. International Conference on Data Science and Engineering, ICDSE 2014.

[17] www.countbayesie.com. Kullback-leibler divergence explained. URL https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained.

[18] A simple introduction to kullback-leibler divergence through python code(2018). URL https://bigdatascientistblog.wordpress.com.

[19] From wikipedia:. URL https://en.wikipedia.org/wiki/Cross_entropy.

[20] Evaluation of topic modeling: Topic coherence. URL https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/.