Evaluation of Punjabi Named Entity Recognition using Context Word Feature

Amandeep Kaur
University College of Engineering
Punjabi University, Patiala

Gurpreet Singh Josan
Department of Computer Science
Punjabi University, Patiala

ABSTRACT
Named Entity Recognition is the task of identifying and classifying Named Entities in the given text. In this paper evaluation of Named Entity Recognition in Punjabi language has been performed using context word feature. Words preceding and succeeding the target word are very helpful in determining its category. In this work context word feature of word window 7, 5 and 3 have been used. Experiments have been performed using different training and test sets. In this evaluation a Named Entity Tagset of 14 tags namely PERSON, ORGANIZATION, LOCATION, FACILITY, EVENT, RELATIONSHIP, TIME, DATE, DESIGNATION, TITLE-PERSON, NUMBER, MEASURE, ABBREVIATION and ARTIFACT has been used. It has been observed that word window 7 and 5 have given better results as compared to word window 3. Although F-scores and Precision values of word window 7 are slightly higher than that of word window 5 but recall of word window 7 was found to be lower than that word window 5.

General Terms
Natural Language Processing, Information Extraction, Named Entity Recognition

Keywords
Named Entities, Named Entity Recognition, Punjabi Language, Context word feature

1. INTRODUCTION
Named Entities (NE) are phrases that represent person, organization, location, number, time, measure etc. in a given text. Named Entity Recognition (NER) is a computational linguistics task in which every word in a document is classified as falling into some predefined categories: person, location, organization, date, time, percentage, monetary value and "none-of-the-above" [1]. NER involves two tasks: identification of Named Entities (NEs) and classification of NEs into different types, such as organization, location, person names etc.

According to [9], NER can be modeled as a sequence labeling task. Given an input sequence of words $W_1^n = w_1w_2w_3....w_n$, the NER task is to construct a label sequence $L_1^n = l_1l_2l_3....l_n$, where label l_i either belongs to the set of predefined classes for named entities or is none (representing words which are not proper nouns). For ex: Gandhi<PERSON> was<NONE> born<NONE> in<NONE> 1869<DATE>. Moreover, NEs are special words which are not defined under the grammatical rules of a language.

NER task was first introduced at DARPA sponsored Sixth Message Understanding Conference-6 (MUC-6) as one of the important sub-task of Information Extraction (IE) [6]. After that proper identification and classification of NEs has attracted the attention of Natural Language Processing (NLP) researchers. Information Retrieval and Extraction (IREX) program [14], Automatic Content Extraction (ACE)[1] program, Conference on Natural Language Learning 2002 and 2003(CoNLL 2002 and 2003)[13][12] have large contribution in emergence of NER.

NER plays an important role in various Natural Language Processing (NLP) applications like Information Extraction (IE), Question Answering (QA), Machine Translation (MT), Information Retrieval (IR), Text Summarization etc.

Although a lot of work has been done in English and Foreign languages with high accuracy but the research regarding NER for Indian languages (ILs) is still far behind. Features like Capitalization, which is helpful in other languages, are of no use in Indian languages making NER task more difficult and challenging in ILs.

The Workshop on Named Entity Recognition for South and South East Asian Languages (NERSSEAL) initiated NER research for ILs as a shared task in which five South Asian languages, Hindi, Bengali, Telugu, Oriya and Urdu were covered. The tagset defined for this task consists of 12 tags namely PERSON, ORGANIZATION, LOCATION, DESIGNATION, ABBREVIATION, TITLE-PERSON, BRAND, TITLE-OBJECT, TIME, MEASURE, NUMBER, TERMS. [17].

Like other Indian languages, Punjabi has a very old and rich literary history and is also going towards fast technological developments. A number of articles are available in the digital form for Punjabi Language. The research on digital form of Punjabi takes pace but still lacking various resources. No availability of NER system in Punjabi language is one of the major hurdles in research activities and thus the motive to pursue research in this direction. The NER research work in Punjabi language was initiated with 12 tags using Conditional Random Fields approach as presented in [8].

In this paper various experiments performed using word window 7, 5 and 3 as context word features with different training and test sets have been discussed. The paper is further organized as follows: Section II describes various NER approaches. In Section III related work is discussed. Section IV describes word regarding NER in Punjabi Language. Finally, in Section V various experiments and their results have been discussed.

2. NER APPROACHES
According to [10], the methods used for automatic identification and classification of named entities, are classified into following categories:

[1] http://www.itl.nist.gov/iaui/894.01/tests/ace/2000 /doc/ace-tides00/sld007.html

- Hand-made Rule-based NER
- Machine Learning-based NER
- Hybrid NER.

Hand-made Rule-based NER focuses on extracting names using human-made rules set. Most of the earlier studies were based on hand-crafted rules. These approaches are relying on manually coded rules and manually compiled corpora.

In Machine Learning-based NER, the systems look for patterns and relationships into text to make a model using statistical models and machine learning algorithms. The systems identify and classify nouns into particular classes such as persons, locations, times, etc base on this model, using machine learning algorithms. There are two types of machine learning model that are use for NER:

Supervised machine learning model – It involves using a program that can learn to classify a given set of labeled examples that are made up of the same number of features. The supervised learning approach requires preparing labeled training data to construct a statistical model, but it cannot achieve a good performance without a large amount of training data, because of data sparseness problem. In recent years several statistical methods based on supervised learning method were proposed which includes Hidden Markov Model (HMM), Conditional Random Fields (CRF), Maximum Entropy (ME), Support Vector Machines (SVM) and Decision Trees (DT).

Unsupervised machine learning model- It is another type of machine learning model, where an unsupervised model learns without any feedback. In unsupervised learning, the goal of the program is to build representations from data. These representations can then be used for data compression, classifying, decision making, and other purposes. Unlike the rule based method, these types of approaches can be easily port to different domain or languages.

In Hybrid NER system, the approach is to combine rule based and machine learning-based methods, and make new methods using strongest points from each method.

3. RELATED WORK

The Third International Joint Conference on Natural Language Processing (IJCNLP – 08) workshop on NER for South and South East Asian Languages (NERSSEAL), held in 2008 at IIIT Hyderabad, was a major attempt in the direction of NER for Indian languages with focus on Hindi, Bengali, Oriya, Telugu and Urdu languages. 12 research papers were selected for this workshop and out of which four were in shared task track. All these research papers were based on 12 tags NE tagset defined for NERSSEAL [17].

Sujan Kumar Saha in [11] described a hybrid system based on Maximum Entropy model, language specific rules and gazetteers. Language specific rules and context patterns were prepared for Hindi and Bengali only. For the other languages the system used the Maximum Entropy model. The system was able to recognize 12 classes of NEs with 65.13% f-value in Hindi, 65.96% f-value in Bengali and 44.65%, 18.74%, and 35.47% f-value in Oriya, Telugu and Urdu respectively.

The work regarding Telugu language is mentioned in [16]. They used Conditional Random Fields (CRF) approach for recognizing named entities using various language dependent and language independent features. The corpus was tagged using the IOB tagging scheme. They report precision, recall and F-Score of 64.07%, 34.57% and 44.91% respectively.

The work in [3] reported about the development of a NER system for Bengali language using Support Vector Machine (SVM) approach. For this experiment an overall average Recall, Precision and F-Score are claimed to be 94.3%, 89.4% and 91.8%, respectively.

In [4] author reported the development of CRF based NER system. This system used both language independent and language dependent features. Evaluation results have demonstrated the highest F-Score of 59.39% for Bengali, 33.12% for Hindi, 28.71% for Oriya and 35.52% for Telugu.

The work in [5] is also based upon Conditional Random Field approach. In this work authors have combined machine learning techniques with language specific heuristics. They have reported Lexical F-Score of 40.63, 50.06, 39.04, 40.94, and 43.46 for Bengali, Hindi, Oriya, Telugu and Urdu respectively.

Authors presented NER work for Telugu language in [18]. They developed a CRF based NER system for Telugu and tested it on several data sets for identifying mainly person, place and organization names. F-score between 80% and 97% was obtained in various experiments.

[2] discussed a three-stage approach of named-entity detection for Bangla language. The stages are based on the use of NE dictionary, rules for NE and left-right co-occurrence statistics. They only tried to identify the NEs, not classify them. For this task, they were able to achieve an overall F-measure of 89.51%

4. NER IN PUNJABI LANGUAGE

Punjabi or Panjabi (ਪੰਜਾਬੀ in Gurmukhi script, پنجابی in Shahmukhi script, Pañjābī in transliteration) is an Indo-Aryan language spoken by inhabitants of the historical Punjab region (in Pakistan and India). Speakers mainly include adherents of the religions of Sikhism, Hinduism and Islam.

Punjabi is the official language of the Indian state of Punjab and also one of the official languages of Delhi. According to the Ethnologue 2005[2] estimate there are 88 million native speakers of the Punjabi language and ranked 20th among the languages spoken in world.

It has been observed that vast amount of information about Punjabi language is available online, but this information is not present in a proper format which could be used to benefit the local users. Punjabi, like other Indian languages, is still lacking in the availability of resources like annotated corpora, name dictionaries, good morphological analyzers, POS taggers etc. in the required measure. Web sources for various gazetteer lists are available in English, but such lists are not available in Punjabi. This leads to little attention for Punjabi Language in Natural Language Processing (NLP) tasks especially in the area of NER.

NER is an important NLP task which has not been deeply explored for Punjabi Language and thus the motivation for this research. A small initiative in this direction was taken in 2009 as discussed in [8]. A small annotated corpus and gazetteers were manually created. Corpus was also manually tagged using coarse grained Parts of Speech (POS) tagset of 9 tags. Further extending this work, it was decided to create more exhaustive and improved resources for developing a Punjabi NER system.

[2] http://en.wikipedia.org/wiki/Ethnologue

Although various NER approaches have been identified and reported by authors for developing NER systems but Machine Learning techniques have been found to be very useful in this area. The efficiency of machine learning techniques is highly dependent on large corpora annotated with named entities.

So, it was decided to prepare an NE tagged annotated corpus for Punjabi language first, as none was available in the literature. In corpus annotation task two important challenges were faced: deciding NE Tagset (annotation categories) and annotation guidelines. It was soon realized that without clearly defined annotation guidelines and tagset, it is very difficult for the annotators to perform manual and/or automatic annotation. Even if the proper annotation guidelines and tagset are provided before annotation task, still there is high probability of amendments. As the annotation task progresses, more ambiguous and confusing cases are identified that do not fit the given tagset and annotation guidelines thus leads to inconsistency in the corpus. In [7], various tagset design issues and problematic cases were discussed and accordingly proposed additional tags to be used for NER task in Punjabi language. In order to develop the tagset, Extended Named Entity hierarchy provided in [15], CoNLL 2002 and 2003 tagset and 12 tags of NERSSEAL were referred.

For developing the annotated corpus, two online Punjabi newspapers Ajitweekly.com[3] and Ajitjalandhar.com[4] were used. Ajitjalandhar was found to be more useful as it also provides web archive as well as it is the most popular newspaper in Punjab region of India. E-news articles corresponding to Sports, Business, Entertainment, Health, Religion, and General news from State, National and International fields were collected.

4.1 Named Entity Tagset

In the Workshop on Named Entity Recognition for South and South East Asian Languages (NERSSEAL) the tagset defined consisted of 12 tags namely PERSON, ORGANIZATION, LOCATION, DESIGNATION, ABBREVIATION, BRAND, TITLE-PERSON, TITLE-OBJECT, TIME, MEASURE, NUMBER, TERMS. [17]

In [8], for evaluation of NER in Punjabi language, NERSSEAL tagset of 12 tags was used. During this work various ambiguous and problematic cases were realized. Various tagset design issues and author's recommendations about additional tags were presented and finally a tagset of 14 tags have been proposed in [7] which will be used in the current research work. Although Extended Named Entity hierarchy provides more than 100 tags but we have opted a limited tagset. NER research in Punjabi Language is in the initial stage. Keeping in view the resource scarcity issue, we propose an NE tagset of 14 tags namely PERSON, ORGANIZATION, LOCATION, FACILITY, EVENT, RELATIONSHIP, TIME, DATE, DESIGNATION, NUMBER, TITLE-PERSON, MEASURE, ABBREVIATION and ARTIFACT.

4.2 Corpus Annotation

The corpus has been annotated using the tagset mentioned in Table I. For each Named entity a tag has been specified. For instance, for annotating single person name in the data, the tag NEP is used that denotes Named Entity Person.

In order to tag a multiple entity which consists of more than one word, IOB tagging scheme is used. For instance, for annotating First name of a person, the tag B-NEP (Beginning of Named Entity Person) is used and for annotating Last name, the tag I-NEP (Intermediate of Named Entity Person) is used.

Table1. Proposed Named Entity Tagset

S.No.	Name	Tag	Examples
1	Person	NEP	ਰਣਜੀਤ [Ranjit]
2	Location	NEL	ਪੰਜਾਬ [Punjab]
3	Organization	NEO	ਕਾਂਗਰਸ [Congress]
4	Facility	NFAC	ਅਪੋਲੋ ਹਸਪਤਾਲ [Apollo Hospital]
5	Event	NEVE	ਲੋਕ ਸਭਾ ਚੋਣਾਂ [Lok Sabha Elections], ਓਲੰਪਿਕ ਖੇਡਾਂ [Olympics]
6	Relationship	NREL	ਭਰਾ [Brother], ਭੈਣ[Sister], ਮਾਮਾ [Mother's brother]
7	Time	NETI	ਦਸ ਸਾਲ [10 Years], ਦਸ ਸਾਲਾਂ [10 Years], ਦਸਵਾਂ ਸਾਲ [10th Year] , 2 ਵਜੇ [2 O'clock]
8	Date	NEDA	ਸਾਲ 2008 [Year 2008], ਐਤਵਾਰ [Sunday], 11 ਜੂਨ 2013 [11 June 2013]
9	Designation	NED	ਮੰਤਰੀ [Minister], ਕਪਤਾਨ [Captain], ਨਿਰਦੇਸ਼ਕ [Director], ਅਭਿਨੇਤਾ [Actor]
10	Title-Person	NETP	ਸ਼੍ਰੀ [Mr.], ਸੰਤ [Saint]
11	Number	NEN	ਇੱਕ [One], ਪੰਜਵਾਂ [Fifth], ਢਾਈ [2½], ੭ [Seven in Gurmukhi Script]
12	Measure	NEM	ਦੁਗੱਣਾ [2 times], 10 ਪ੍ਰਤੀਸ਼ਤ [10 %]

[3] http://www.ajitweekly.com
[4] http://www.ajitjalandhar.com

13	Abbreviation	NEA	ਆਈ.ਪੀ.ਐਲ [IPL], ਬੀ.ਜੇ.ਪੀ [BJP]
14	Artifact	NART	ਕ੍ਰਿਕਟ [Cricket as a Sport], ਪੰਜਾਬੀ [Punjabi language], ਇਤਿਹਾਸ [Subject of History]
15	Other (Not an NE)	O	

Moreover the corpus is annotated using Nested Named Entities which are also called embedded entities. We have used 3 levels of nesting in the data. While annotating corpus a number of nesting combinations were identified. Corpus has been annotated with useful combinations in the form of Nested Tags. Few of these combinations are discussed in Table 2.

Table2. Proposed Named Entity Tagset

S.No.	**Nested NE**	**Nested NE Tag**	**Example**
1	Number within Date	B-NEDA+NEN+O / I-NEDA+NEN+O	30 ਸਤੰਬਰ [30th September]
2	Person within Location	B-NEL+B-NEP+O/ B-NEL+NEP+O/ I-NEL+B-NEP+O/ I-NEL+I-NEP+O/ I-NEL+NEP+O	ਅੰਬੇਡਕਰ ਨਗਰ [Ambedkar Town]
3	Number within Measure	B-NEM+B-NEN+O/ B-NEM+NEN+O/ I-NEM+I-NEN+O/ I-NEM+NEN+O	11000 ਰੁਪਏ [Rs 11000]
4	Person within Organization	B-NEO+NEP+O/ B-NEO+B-NEP+O/ I-NEO+B-NEP+O/ I-NEO+I-NEP+O/ I-NEO+NEP+O	ਮਾਨ ਗਰੁੱਪ [Maan Group]
5	Number within Time	B-NETI+NEN+O/ I-NETI+NEN+O	ਪੰਜ ਮਹੀਨੇ [5 months]
6	Number within Event	B-NEVE+NEN+O/ I-NEVE+NEN+O	33ਵੀਂ ਸਬ - ਜੂਨੀਅਰ ਚੈਂਪੀਅਨਸ਼ਿਪ [33rd Sub -Junior Championship]
7	Person within Event	B-NEVE+NEP+O/ I-NEVE+NEP+O/	ਰੌਬਿਨ ਲੀਗ ਅਭਿਆਸ ਲੜੀ [Robin League Practice Series]
8	Location within Organization	B-NEO+B-NEL+O/ B-NEO+NEL+O/ I-NEO+NEL+O/ I-NEO+B-NEL+O/ I-NEO+I-NEL+O	ਹਰਿਆਣਾ ਮਾਰਕੀਟਿੰਗ ਬੋਰਡ [Haryana Marketing Board]
9	Location within Facility	B-NFAC+B-NEL+O/ B-NFAC+NEL+O/ I-NFAC+B-NEL+O/ I-NFAC+I-NEL+O/ I-NFAC+NEL+O	ਦਿੱਲੀ ਬਾਜ਼ਾਰ [Delhi Market]
10	Person within Facility	B-NFAC+B-NEP+O/ B-NFAC+NEP+O/ I-NFAC+B-NEP+O/ I-NFAC+I-NEP+O/ I-NFAC+NEP+O	ਵਾਲਮੀਕ ਮੰਦਿਰ [Valmik Temple]
11	Number within Time within Event	B-NEVE+B-ETI+NEN/ I-NEVE+B-NETI+NEN	ਦੋ ਰੋਜ਼ਾ ਕਬੱਡੀ ਟੂਰਨਾਮੈਂਟ [2 day's Kabaddi Tournament]

5. EXPERIMENTS AND RESULTS

The training corpus used in this work consists of 2, 00,000 words out of which 58000 are NEs. For training and testing the NER system we have used the C++ based OpenNLP CRF++ package[5], which is an open source implementation of Conditional Random Fields (CRFs) for segmenting /labeling sequential data. The system was trained on various training sets and experiments were conducted to evaluate the system on test data sets. The standard evaluation metrics, Precision, Recall and F-Score were used.

Experiments were performed on test data of 10,000 words that varies with the training sets. Initially the system was trained on first 20,000 words from the training corpus. Then this system was tested and evaluated on test data of next 10,000 words from the training corpus. The system was trained on 10 data sets with 20,000, 40,000, 60,000, 80,000, 10,000, 120,000, 140,000, 160,000 and 180,000 words. In this way, we kept on increasing the training data set size and evaluated the system on test data sets different from training data sets. In this work, experiments were performed using context word feature in which previous and next words of a particular word are considered. Word window sizes of 7(previous 3 words, current word and next 3 words), 5(previous 2 words, current word and next 2 words) and 3(previous word, current word and next word) were used. The evaluation results with

[5] http://crfpp.sourceforge.net

different Training sets and word window sizes are discussed in tables given below:

Table3. Training Size= 20,000 words, Sentences = 672

S.No.	Word Window Size	Precision	Recall	F-Score
1	7(w-3,w-2,w-1, w0,w1,w2,w3)	76.80	49.43	60.15
2	5(w-2,w-1, w0,w1,w2)	73.62	51.55	60.65
3	3 (w-1,w0,w1)	51.51	48.99	50.02

w-3 : Preceding 3rd word from current word
w-2: Preceding 2nd word from current word
w-1: Preceding word from current word
w0: Current word
w1: Succeeding word from current word
w2: Succeeding 2nd word from current word
w3: Succeeding 3rd word from current word

Table4. Training Size= 40K words, Sentences = 1382

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1, w0,w1,w2,w3)	79.44	51.58	62.77
2	5(w-2,w-1, w0,w1,w2)	75.85	53.99	63.08
3	3 (w-1,w0,w1)	57.50	50.12	53.56

Table5. Training Size= 60K words, Sentences = 2092

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	81.22	64.76	72.06
2	5(w-2,w-1 ,w0,w1,w2)	79.28	66.37	72.25
3	3 (w-1,w0,w1)	65.21	59.99	62.49

Table6. Training Size= 80K words, Sentences = 2820

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	79.10	70.10	74.33
2	5(w-2,w-1 ,w0,w1,w2)	75.06	71.49	73.23
3	3 (w-1,w0,w1)	59.70	66.42	62.88

Table7. Training Size= 100K words, Sentences = 3468

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	87.24	78.27	82.51
2	5(w-2,w-1,w0,w1,w2)	86.59	79.20	82.73
3	3 (w-1,w0,w1)	76.57	71.91	74.17

Table8. Training Size= 120K words, Sentences = 4192

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	88.77	78.54	83.34
2	5(w-2,w-1 ,w0,w1,w2)	87.85	79.70	83.23
3	3 (w-1,w0,w1)	77.50	71.09	74.16

Table9. Training Size= 140K words, Sentences = 4872

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	87.64	72.82	79.54
2	5(w-2,w-1 ,w0,w1,w2)	85.36	73.63	79.06
3	3 (w-1,w0,w1)	75.46	68.49	71.80

Table10. Training Size= 160K words, Sentences = 5460

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	91.38	74.94	82.35
2	5(w-2,w-1 ,w0,w1,w2)	90.59	75.94	82.62
3	3 (w-1,w0,w1)	79.77	71.65	75.49

Table11. Training Size= 180K words, Sentences = 6185

S.No.	Word Window	Precision	Recall	F-Score
1	7(w-3,w-2,w-1 ,w0,w1,w2,w3)	86.01	68.06	75.99
2	5(w-2,w-1 ,w0,w1,w2)	83.42	68.81	75.42
3	3 (w-1,w0,w1)	71.53	64.49	67.83

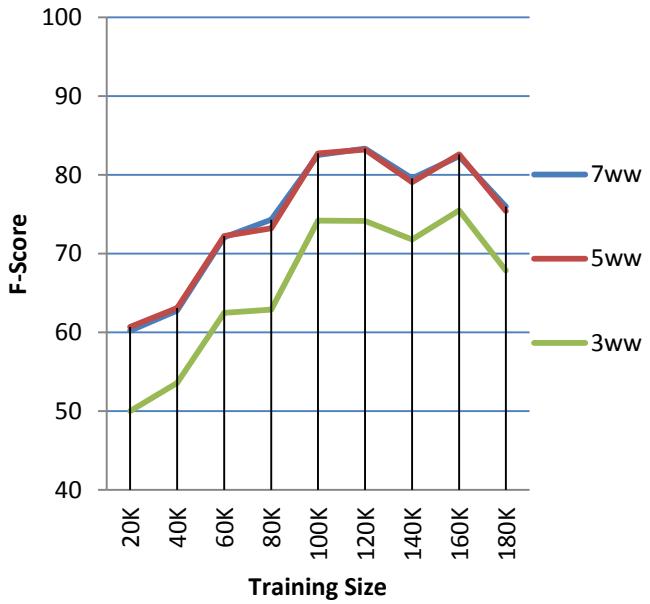

Figure 1: F-score value for different Training Sets

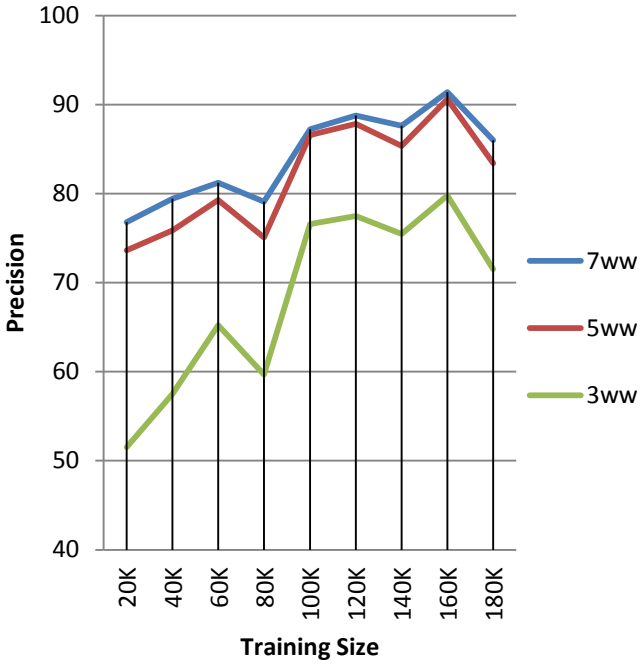

Figure 2: Precision value for different Training Sets

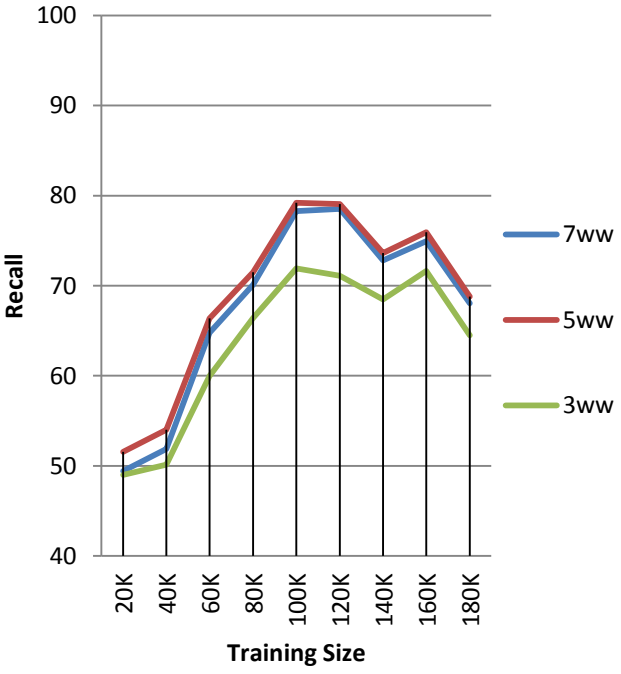

Figure 3: Recall value for different Training Sets

It has been observed that word windows 5 and 7 have given better results as compared to word window 3. F-Score values of context word windows 5 and 7 are quite similar. However, Precision value of word window 7 is higher that of word window 5 but Recall value of word window 7 is slightly lower than that of word window 5 for different training sets. As word window 5 gives slightly higher recall so it will be preferred for future work along with other additional features.

6. REFERENCES

[1] Borthwick, A., 1999. Maximum Entropy Approach to Named Entity Recognition. Ph.D. dissertation, Comput. Sci. Dept., New York Univ., New York, USA.

[2] Chaudhuri, B. B. and Bhattacharya, S., 2008. An Experiment on Automatic Detection of Named Entities in Bangla. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 75-82.

[3] Ekbal, A. and Bandyopadhyay, S., 2008. Bengali Named Entity Recognition using Support Vector Machine. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 51–58.

[4] Ekbal, A., Haque, R., Das, A., Poka V. and Bandyopadhyay, S., 2008. Language Independent Named Entity Recognition in Indian Languages. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 33–40.

[5] Gali, K., Surana, H., Vaidya, A., Shishtla, P. and Sharma, D. M., 2008. Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 25-32.

[6] Grishman, R. and Sundheim B., 1996. Message Understanding Conference - 6: A Brief History. In the Proceedings of the 16th International Conference on Computational Linguistics (COLING). 466 – 471.

[7] Kaur, A. and Josan, G., 2014. Improved Named Entity Tagset for Punjabi Language. In the Proceedings of 2014 RAECS.

[8] Kaur, A., Josan, G. and Kaur, J., 2009. Named Entity Recognition For Punjabi: A Conditional Random Field Approach. In Proceedings of ICON-2009: 7th International Conference on Natural Language Processing. 277-282.

[9] Lafferty, J.D., McCallum, A. and Pereira, F.C.N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of International Conference on Machine Learning. 282-289

[10] Mansouri, A., Suriani Affendey, L. and Mamat, A., 2008. Named Entity Recognition Approaches. International Journal of Computer Science and Network Security. 339-344.

[11] Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S. and Mitra, P., 2008. A Hybrid Approach for Named Entity Recognition in Indian Language. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 17-24.

[12] Sang, E. F. T. K. and Meulder, F. D., 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of 7th Conference on Natural Language Learning CoNLL-2003.

[13] Sang, E. F. T. K., 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In Proceedings of 6th Workshop on Computational Language Learning, CoNLL-2002.

[14] Sekine, S. and Ishara, H., 2000. IREX: IR & IE evaluation project in Japanese. In Proceedings of the 2nd International Conference on Language Resources and Evaluation.

[15] Sekine, S., Sudo, K. and Nobata, C., 2002. Extended Named Entity Hierarchy. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002.

[16] Shishtla, P. M., Gali, K., Pingali P. and Varma, V., 2008. Experiments in Telugu NER: A Conditional Random Field Approach. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 105-110.

[17] Singh, A. K., 2008. Named Entity Recognition for South and South East Asian Languages: Taking Stock. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 5–16.

[18] Srikanth, P. and Murthy, K. N., 2008. Named Entity Recognition for Telugu. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. 41-50.