Hybrid Approach for Named Entity Recognition

Kanwalpreet Singh Bajwa
Student (Mtech, CSE)
Punjabi University Regional
Centre, Mohali

Amardeep Kaur
Assistant Professor (CSE)
Punjabi University Regional
Centre, Mohali

ABSTRACT
This paper proposes the Named Entity Recognition (NER) system for Punjabi language using a hybrid approach in which rule based approach and machine learning approach i.e. Hidden Markov Model (HMM) is combined. With no Dataset available, the Named Entities (NEs) were manually tagged which led us to the creation of training and testing dataset, under the linguistic supervision. Using hybrid approach, the proposed system is able to recognize Name of person, Location, Time, Date, Designation, Organization, Title-person, Event, Abbreviation, Facility, Number, Artifact, Relation and Measure. This paper presents two versions of NER for Punjabi language, the first version is designed with HMM only and the second version is designed hybrid approach in which HMM is used in combination with handcrafted rules. NER system with proposed hybrid approach is able to achieve the precision of 72.92%, Recall of 76.27%, F-measure of 74.56% with hybrid approach and Precision, Recall and F-measure of 47.57%, 48.98%, 48.27% respectively has been achieved by using HMM only. This paper has also compared proposed method with simple HMM and observed that proposed NER system performs better.

General Terms
Named Entity recognition, Natural Language Processing (NLP), Hybrid approach.

Keywords
Named Entity Recognition (NER), NLP, and Hidden Markov Model (HMM), Rule based approach.

1. INTRODUCTION
The term "Named Entity" now extensively used in Natural Language Processing was first introduced at the sixth Message Understanding Conference-6 (MUC-6) [1]. The focus of MUC-6 was on Information Extraction (IE) tasks, where ordered information concerning activities of companies and defense related information is extracted from unstructured text, such as newspaper articles. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)". Named Entities (NEs) have a unique status in Natural Language Processing (NLP) due to their distinguishing nature which other elements of language do not have, e.g. NEs indicate particular concepts and things in the world which are not listed in grammar or lexicons [2]. Named Entity Recognition is a sub problem of NLP. It is a computational linguistic task which figures out atomic elements in text and classifies them into predetermined classes such as name of place, organization, date, etc. Initial requirement for the development of NER system is corpus. Corpus is an unprocessed data in specific language for which NER tool is being developed. NER incorporates two tasks 1) Identification of Named Entities (NEs) and 2) Organizing NEs in different categories. For example, if "*Anil Ambani*" is a named entity in corpus, then it is required to identify the position of this entity and subsequent step would classify this entity into predefined category which is PERSON in this case.

Abounding work has been done in NER for various languages across the world, but research for Indian Languages is at pre-mature stage particularly Punjabi language which belongs to the class of Indo-Aryan Language. According to Ethnologue-3 2005 estimate there are 88 million native speakers of Punjabi language and ranked 20th among the languages spoken in world [3]. Unlike English language where capitalization is the major clue for recognizing Named Entities, Punjabi lacks this feature and in addition to it, there are so many other problems faced by researchers while designing NERC system for Punjabi language such as ambiguity in names, lack of standardization and spelling, non-availability of large gazetteer, scarcity of resources and tools etc.

Hidden Markov Model is a generative model which incorporates double stochastic process. First stochastic process generates the sequence of states where as second stochastic process is responsible for generating the sequence of observations from the sequence of states. In Vivek et al., 2013, author has introduced a HMM based NER for seven languages i.e. English, Bengali, Tamil, Hindi, Punjabi, Marathi and Telgu. Rule based is another approach which has been used for designing NER system [4, 5]. In rule based approach, rules are crafted for every entity and NEs are recognised using these rules. The proposed study has clubbed rule based and HMM to design a hybrid approach to develop NER for Punjabi language and achieved better results than [6].

The remaining sections of this paper is organised as follows: section 2 discusses about related work and section 3 describes the methodology of proposed study. An experiment and results are presented in section 4 and study is concluded in section 5.

2. RELATED WORK
NER system can be build using diverse techniques. The two major approaches to NER are: Rule based (Linguistic) approaches and Machine Learning (ML) based approaches. In Rule based approaches, language based rules or human written rules are applied to recognise named entities from unstructured text. Prior knowledge of grammar of language is required to develop NER tool for that language. These approaches are successfully implemented in [1][4][7][8] [9][10]. In Machine Learning approaches, large amount of NE annotated data is prepared which is then trained using various ML techniques such as Hidden Markov Model [11][12][13] [14][15], State Vector Machines [16][17], Maximum Entropy Model (Max Ent) [12][14][18][19], Robust risk minimization [20] and Conditional Random Fields (CRFs) [16][21].

Different ML techniques can also be combined with Rule based approaches to develop NER system. Rohini Srihari et al. has combined Hidden Markov Model (HMM), MaxEnt and handcrafted rules to build a NER system. Radu et al. has also combined Risk Minimization Classifier and HMM classifier to build NER system German and English language. NER systems can also make use of gazetteer lists to identifying names [4].

Considering the successful implementation of HMM in developing Named Entity Recognition for English language [11], Chinese language [13], as well as for Punjab language [22], and successful implementation of rule based approach [4], this study proposes NER system for Punjabi language using Hybrid approach.

3. METHODOLOGY

The proposed study has incorporated Bikel's HMM [15] in addition to some handcrafted rules to generate NER system for Punjabi text. These rules are mostly language dependent which highly optimize the output of the system. Architecture of the proposed NER system is shown in figure 1. The proposed architecture is divided into 4 stages 1) Data set preparation, 2) Data preprocessing, 3) HMM training, 4) Testing.

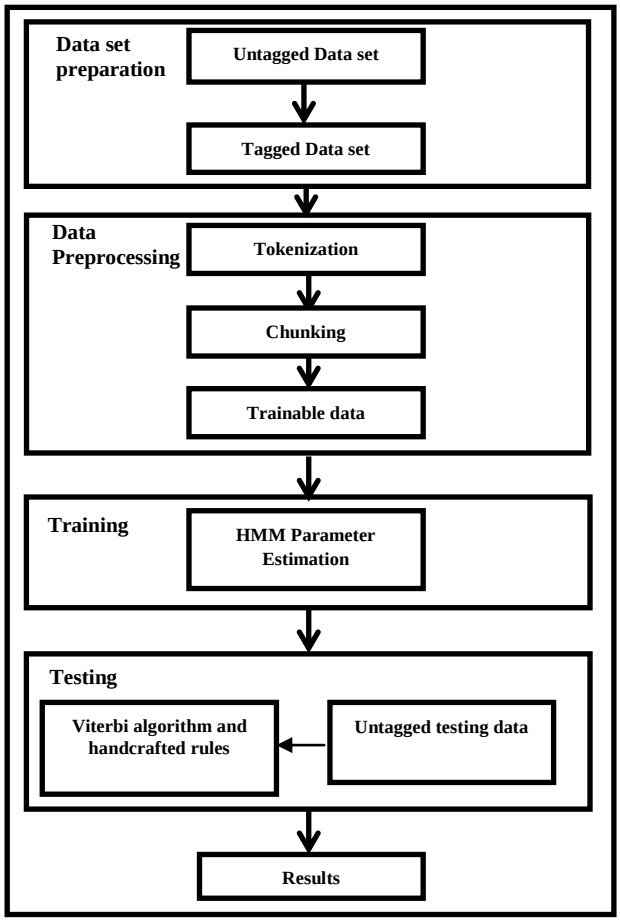

Figure 1: Architecture of proposed system

Table 1. NE classes and Tags used in proposed study [22]

Names	Tags	Examples
Person	NEP	ਰਣਜੀਤ [Ranjit]
Location	NEL	ਪੰਜਾਬ [Punjab]
Organisation	NEO	ਕਾਂਗਰਸ [Congress]
Facility	NFAC	ਅਪੋਲੋ ਹਸਪਤਾਲ [Apollo Hospital]
Event	NEVE	ਲੋਕ ਸਭਾ ਚੋਣਾਂ [Lok Sabha Elections]
Relation	NREL	ਭਰਾ [Brother] ਭੈਣ[Sister]
Time	NETI	ਦਸ ਸਾਲ [10 Years] ਦਸਵਾਂ ਸਾਲ [10th Year] 2 ਵਜੇ [2 O'clock]
Date	NEDA	ਸਾਲ 2008 [Year 2008] ਐਤਵਾਰ [Sunday] 11 ਜੂਨ 2013 [11 June 2013]
Designation	NED	ਮੰਤਰੀ [Minister] ਕਪਤਾਨ [Captain]
Title-Person	NETP	ਸ਼੍ਰੀ [Mr.] ਸੰਤ [Saint]
Number	NEN	ਇੱਕ [One] ਪੰਜਵਾਂ [Fifth]
Measure	NEM	ਦੁਗੱਣਾ [2 times] 10 ਪਰ੍ਤੀਸ਼ਤ [10 %]
Abbreviation	NEA	ਆਈ.ਪੀ.ਐਲ [IPL] ਬੀ.ਜੇ.ਪੀ [BJP]
Artifact	NART	ਕ੍ਰਿਕੇਟ [Cricket as a Sport] ਪੰਜਾਬੀ [Punjabi language]

3.1 Dataset

The Data set used in proposed study is prepared from Punjabi text available on online newspapers' sites such as ajitnews.com, parvasi.com, doabanews.com and is authenticated from linguistic before it is used in proposed study. It constitutes 9.5k words and 500 lines. The study has used tag set of 14 tags proposed in [22] for manually tagging NEs in Punjabi text and <OTHER> tag is used for words which are not NEs. 400 Tagged lines are used as training set and remaining 100 untagged lines are considered as testing data.

3.2 Data Preprocessing

In a proposed method, data set is tokenized before processing i.e. each line of data set is split into words and each word represents either NE or not a NE. But there are some NEs which consist of more than two words and for these entities

only tokenization is not sufficient. For example" ਭਾਰਤੀ ਫਿਲਮ ਉਦਯੋਗ "is a name of an organization constituting three words, if only tokenization is applied in pre-processing then this single NE will be divided into three entities which individually represent nothing so another preprocessing technique is applied to data set after tokenization which is known as chunking. Chunking groups all words which represent single NE. Taking above example of organisation name, chunker groups all three constituting words and considers it as single entity. After pre-processing techniques data set is fully prepared for training.

3.3 Training

The Hidden Markov Model (HMM) is a statistical model which is also known as a probabilistic generative model of sequence. Basically HMM is double stochastic process in which first stochastic process generates the sequence of states and second stochastic process is responsible for generating sequence of observations from sequence of states. States in case of proposed study are NE classes which are the tags of NEs to be recognised by proposed system.

The first stochastic process of HMM can be described by considering two assumptions. The first assumption is that the current state of HMM depends only on its previous state i.e. probability of transition to current state depends only on previous state. Suppose q_1,q_2,q_3,....,q_n are number of states that in case of proposed studies are tags then the transition probability of any state is calculated as:

$$P(q_t|\, q_1, q_2, \cdots, q_{t-1}, \lambda) = P(q_t|q_{t-1}, \lambda) \qquad (1)$$

Second assumption is that transition probabilities are time invariant i.e. $a_{ij} = P(q_t = j|q_{t-1} = i)$ where a_{ij} is the transition probability to state j from state i. Considering these assumptions transition matrix $\{a_{ij}\}$ is computed which is an N×N matrix where N denotes the number of state in the model which is 14 in proposed study. Start probability or initial state distribution (π) is a probability that the sentence start with particular tag or state. It is computed as:

$$\pi = \{\pi_i\} \text{ And } \pi_i = P(q_1 = i),\ 1 \le i \le N \qquad (2)$$

$$\text{Where } \pi_i \ge 0 \text{ and } \sum_{i=1}^{N} \pi_i = 1$$

Second Stochastic process which is responsible for generating sequence of observations assumes that observation o_t at any step t only depends on current state q_t i.e.

$$b_{ij} = P(o_t = j|q_t = i) \qquad (3)$$

Considering this assumption, an emission matrix (B) is defined as B = $\{b_{ij}\}$ which is of size N×M where M denotes number of signals/observations emitted from each state and $b_{ij} \ge 0$ where $\sum_{j=1}^{M} b_{ij} = 1$.

Computation of three parameters of HMM that is start probability, transition probability, emission probability constitutes training of proposed system and classification of words into NEs includes finding most likely sequence of NE classes/tags given the sequence of words and handcrafted rules constitutes testing of proposed system.

3.4 Testing

Testing of a proposed study has been designed in two phases. In First phase, proposed system is tested with viterbi decoder and its output is further optimized with handcrafted rules which constitutes second phase of testing.

3.4.1 Viterbi Decoder

Viterbi algorithm is used for classification which maximizes the conditional probability of state/tag sequence given the sequence of words $P(NE_{1:T}| < w >_{1:T}, \lambda)$ where T is a number of observations in testing data.

The probability of generating the first word of name class or tag is factored into two parts:$P(NE_t|NE_{t-1}).P(< w >_t |NE_t, NE_{t-1})$. The conditional probability which is to be maximized is shown in equation (4):

$$\mathrm{P}(NE_{1:T} \mid < w >_{1:T}, \lambda) = \frac{P(NE_{1:T}, < w >_{1:T}| \lambda)}{P(< w >_{1:T}| \lambda)} \qquad (4)$$

The optimal sequence of NE classes/tags which is obtained by maximizing above probability is computed as:

$$NE^*_{1:T} = _{NE_{1:T}}\ argmax\ P(NE_{1:T}, < w >_{1:T}| \lambda)$$

$$= arg_{NE_{1:T}} \max P(NE_{1:T}| < w >_{1:T}, \lambda) \qquad (5)$$

In Viterbi algorithm there is one auxiliary variable (δ_i) that is introduced to derive this recursive algorithm. This variable is defined as:

$$\delta_i = \max_{0 \le j \le N} \{P(< w >_t |NE_t = i, NE_{t-1} = j).$$

$$P(NE_t = i|NE_{t-1} = j, < w >_{t-1}, \lambda).\ \delta_{t-1}(j)\} \qquad (6)$$

There is a need to keep the track of arguments that maximizes the auxiliary variable at each time step in order to retrieve the optimal sequence of states. To keep this track one variable ($\Psi(i)$) is defined as:

$$\Psi(i) = arg \max_{0 \le j \le N} \{P(< w >_t |NE_t = i, NE_{t-1} = j).$$

$$P(NE_t = i|NE_{t-1} = j, < w >_{t-1}, \lambda).\ \delta_{t-1}(j)\} \qquad (7)$$

In a proposed study, the most likely sequence of states or optimized path generated by the viterbi algorithm are further optimized by applying handcrafted rules over each state in a path.

3.4.2 Handcrafted Rules

The proposed study has developed a number of rules specific for Punjabi to extract language dependent features. These rules have been developed for every Named Entity to be recognised by proposed system

3.4.2.1 Name of person rules

An entity is extracted as a name of a person if elements such as "ਜੀਤ", "ਜੋਤ", " ਕੌਰ", "ਸਿੰਘ", "ਪ੍ਰੀਤ" are found in that entity. If the suffix of observed entity contains "ਖਾਨ", "ਪਠਾਨ", "ਗਰੇਵਾਲ" or " ਗਿੱਲ" then also that entity is observed as the name of a person. In another proposed rule, if next word to the current observed word is "ਨੇ" and state for current word is not Designation (NED) then current word is extracted as name of person (NEP). In other rule, if tag/state of previous words to current observed word is Title (NETP) then current word is extracted as NEP i.e. name of person.

3.4.2.2 Designation rules

Two types of rules are proposed for designation class, first type recognises one-word designation where as second type recognises two-word designation entity. In rule of first type, if single word entity contains elements such as 'ਉਪ', 'ਮੁੱਖ',

'ਮੰਤਰੀ' or 'ਵਾਦੀ' then that entity is extracted as Designation entity i.e. present tag/state for this entity is replaced with state 'NED'. For two-word designations, there is a rule in which if first word is 'ਡਿਪਟੀ' then this word and next word is collectively extracted as Designations i.e. state NED is extracted for it. In another rule, if any of two words are 'ਐਡਵੋਕੇਟ', 'ਡਾਇਰੈਕਟਰ', 'ਜੱਜ' or 'ਵਕੀਲ' then this two word entity is extracted as a designation.

3.4.2.3 Location rules

For location entity if Punjabi word 'ਵਿਖੇ' or 'ਵਾਸੀ' or 'ਪਰਤਣ' is found in observation then previous word to observed word is extracted as the name of a location. When word 'ਪਿੰਡ' is found while scanning testing data then next word to it is extracted as location name. In another rule, if keyword 'ਪੁਰ' is found in any observed word then that word is recognised as location name.

3.4.2.4 Organisation rules

In proposed rule for Organisation class, when keywords 'ਪਾਰਟੀ', 'ਦਲ', 'ਸਭਾ', 'ਕਮਿਸ਼ਨ', 'ਮੰਚ', 'ਯੂਨੀਅਨ', 'ਐਸੋਸੀਏਸ਼ਨ', 'ਕਲੱਬ', 'ਕੰਪਨੀ' are observed then these keywords and word preceding it collectively extracted as organisation name i.e. state/tag 'NEO' is replaced with current state in sequence.

3.4.2.5 Facility Rules

If any observed entity ends with words such as 'ਯੂਨੀਵਰਸਿਟੀ'[university], 'ਹੋਟਲ'[hotel], 'ਸਟੇਡੀਅਮ'[stadium], 'ਬੈਂਕ'[bank], 'ਕਲੀਨਿਕ'[clinic], 'ਹਸਪਤਾਲ'[hospital], 'ਸਕੂਲ'[school], 'ਕਾਲਜ'[college],etc then observed entity is recognised as Facility.

3.4.2.6 Date and Time Rules

If keyword 'ਸਾਲ' or 'ਸੰਨ' is found followed by four [1-2][0-9][0-9][0-9] digits then it represents year and if two digits ≤ 31 preceded by name of month such as 'ਦਸੰਬਰ'[December], 'ਜਨਵਰੀ'[January], etc is found then in both cases entity is extracted as date. If two [1-9][0-9] or one [1-9] digit is found preceding word 'ਸਾਲ'[year] then these collectively extracted as time. In another rule if word ''ਚ' is preceded by four [1-2][0-9][0-9][0-9] digits then the entity is recognised as date entity.

3.4.2.7 Rules for Number and Measure

If any number of [0-9] digits constituting numeric word are not followed by keywords 'ਕਰੋੜ', 'ਫੀਸਦੀ', 'ਪੈਂਏ', 'ਲਿਟਰ', 'ਪ੍ਰਤੀਸ਼ਤ', etc then words are extracted as number entity but if these numerical words are followed by such keywords then they are collectively extracted as measure entity.

3.4.2.8 Rules for Title-person and Abbreviations

After analysing Punjabi text, study has proposed some rules and in first rule, if word 'ਸ਼੍ਰੀ' or 'ਸ.' or 'ਬਾਬਾ', etc is found before name of person i.e. state/tag of next word is NEP then that word is recognised as Title-person entity. In another rule if any observed word contains '-ਏ-' then that word is also recognised as Title-person entity. There is one rule for recognising Abbreviations, if more than two single letter of Punjabi language in Punjabi word preceded by '.' then that word is extracted as abbreviation entity.

3.4.2.9 Relation and Event Rules

The study has also proposed rule for relation entity such as, if word 'ਦੀ' is observed and tag/state of previous word is NEP then next word to observed word is recognised as relation entity. The rule for Event entity has also been proposed in this study. In rule for event, if keywords such as 'ਪੁਰਸਕਾਰ', 'ਜੁਬਲੀ', 'ਦਿਵਸ', 'ਟੂਰਨਾਮੈਂਟ', etc found in any observed entity then observed entity is recognised as event and in another rule, if word 'ਦੌਰਾਨ'' is observed then previous word to observed word is recognised as event.

3.5 Validation parameters

The proposed study has used three evaluation parameters to validate this study. These three parameters are Precision, recall and F-measure. Precision (P) measures the correct number of NE obtained by proposed study over total number NEs extracted. Recall (R) measures the correct number of NEs obtained over the total number of NEs present in testing text.

$$P = \frac{Correct\ number\ of\ NEs\ extracted}{Total\ NEs\ extracted} \quad (8)$$

$$R = \frac{Correct\ number\ of\ NEs\ extracted}{Total\ number\ of\ NEs\ present\ in\ text} \quad (9)$$

F-measure is obtained by harmonic mean of recall and precision i.e.

$$F = \frac{2 \times R \times P}{R + P} \quad (10)$$

These three parameters constitute results of proposed study.

4. EXPERIMENT AND RESULTS

The proposed study has developed two versions of NER system for Punjabi language and this experiment is performed on *Intel®Pentium(R) Dual CPU T3200 @ 2.00Ghz *2 Machine* 32 bit Linux operating system. First version is designed with HMM only and second version has combined HMM with handcrafted rules for designing proposed NER system. Both versions use same data set which is prepared under supervision of linguistic of Punjabi language, section 3.1 describes how data set is prepared. When data set is prepared then next step in proposed study is data preprocessing in which tokenization and chunking techniques are applied to dataset, these are explained in section 3.2. For training, both versions use HMM which results in computation of three parameters i.e. Start probability, Transition probability, and Emission probability. Till this step design of both versions is same and provided with same testing data but testing criteria for two versions are different. First version uses viterbi decoder for decoding HMM as explained in section 3.4.1 and the most likely state sequence given by viterbi is taken as only output and in results, Precision of 47.57%, Recall of 48.99% and F-measure of 48.27% has been obtained. Table 2 shows results delivered by first version with respect to individual NEs. These results are obtained by manually checking output of each line of testing data. The approach used in first version has also been studied in [6] for generating NER for Indian languages. F-measure of 54.55%, Precision of 54.97% and Recall of 54.13% has been claimed with NER for Punjabi language. The major reason behind the variance between our results and results claimed in [6] is dissimilarity in the data set used. The volume of data set

used in [6] is more than our created dataset. But second version is a new approach proposed for NER for Punjabi, the output given by viterbi decoder is filtered by handcrafted rules explained in section 3.4.1 i.e. firstly the testing data is passed through viterbi decoder and then the most likely sequence of states/tags given as its output is further optimized by analysing each state/tag in sequence with respect to rules defined for each state/tag this led to highly optimized state sequence as an output of proposed NER system. Till the time of conducting this study, no research has been reported using this hybrid approach. Our study has achieved the precision of 72.92%, Recall of 76.27% and F-measure of 74.56%, which is eminent in performance as compared to results delivered by NER system, designed with HMM approach proposed in [6]. Table 3 describes results of our proposed version of NER system with respect to individual NEs to be recognised by it.

Table 2. Results obtained using HMM

Named Entity (NE)	Precision	Recall	F-measure
Person	21.43%	21.42%	21.42%
Location	58.33%	56.94%	57.63%
Organisation	59.68%	55.91%	57.73%
Facility	35.29%	32.35%	33.75%
Event	50.06%	86.36%	63.38%
Relation	100%	100%	100%
Time	12.25%	12.25%	12.25%
Date	20%	20%	20%
Designation	49.58%	50%	49.79%
Title-Person	67.16%	62.74%	64.87%
Number	55.73%	57.81%	56.75%
Measure	30%	30%	30%
Abbreviation	37.50%	37.50%	37.50%
Artifact	69.04%	62.54%	65.63%
Total	**47.57%**	**48.98%**	**48.27%**

Table 3. Results obtained using HMM in combination with handcrafted rules

Named Entity (NE)	Precision	Recall	F-measure
Person	81.74%	85.34%	83.50%
Location	64.68%	72.83%	68.51%
Organisation	66.93%	68.66%	67.78%
Facility	86.36%	82.00%	84.14%
Event	54.03%	77.26%	63.59%
Relation	94.44%	100%	97.14%
Time	65%	65%	65%
Date	75%	75%	75%
Designation	59.80%	60.63%	60.21%
Title-Person	85%	100%	91.89%
Number	77.67%	73.75%	75.65%
Measure	90.62%	93.75%	92.16%
Abbreviation	65%	60%	62.40%
Artifact	54.61%	53.54%	54.07%
Total	**72.92%**	**76.27%**	**74.56%**

In proposed study, overall performances of two NER systems developed are also compared. Figure 2 shows the comparison of performances of two generated versions of NER on the basis of three evaluation parameters described in section 3.5.

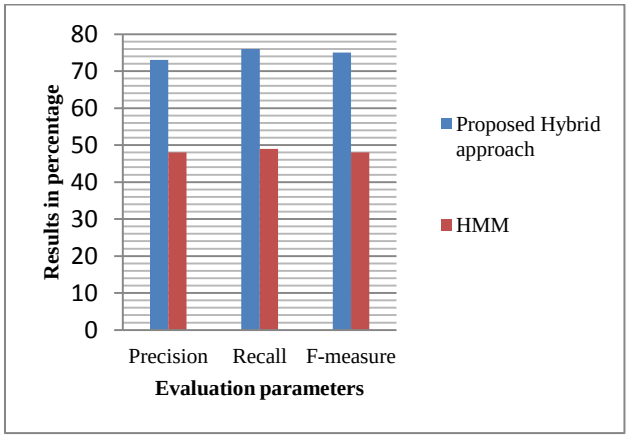

Figure 2: Performance comparison of NER system with two different approaches.

5. CONCLUSION

This study has proposed a new Hybrid HMM utilization for named entity recognition for Punjabi language. The proposed hybrid method is first of its kind for any Indic language and especially in case of Punjabi language. In proposed study, dataset of its own kind has been prepared by using 14 different NE classes proposed in [22]. The study has performed better in hybrid approach constituting supervised learning and handcrafted rules for Punjabi language as compare to using only supervised machine learning. In proposed study, when only HMM is used, results obtained are not satisfactory but when same HMM approach is combined with handcrafted rules then results obtained are pretty notable and this variance is shown in figure 2.

Future works include forming new rules to improve existing results. Tag-set also plays a vital role in improving efficiency of NER system because of ambiguities in Indic languages, so generating more tags and using them to generate NER systems can also constitute future work. As gazetteers for Punjabi language is not available, so preparing gazetteer and using these gazetteers with proposed study can also improve results.

6. REFERENCES

[1] Grishman, R., & Sundheim, B. (1996, August). Message Understanding Conference-6: A Brief History. In COLING Vol. 96 (pp. 466-471).

[2] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, *30*(1) (pp.3-26).

[3] A.Singh and J. Rani (2013).Maximum Entropy Approach based Named Entity Recognition in Punjabi Language. International journal of Computer Application (IJCA) vol.84,no.3(pp.1–5).

[4] Kamaldeep kaur and Vishal Gupta (2012, June). Name Entity Recognition for Punjabi Language. IRACST – International Journal of Computer Science and Information Technology & Security (IJCSITS),vol. 2, pp. 561-567.

[5] Srihari, R., Niu, C., & Li, W. (2000, April). A hybrid approach for named entity and sub-type tagging. In Proceedings of the sixth conference on Applied natural language processing (pp. 247-254). Association for Computational Linguistics.

[6] Gayen, Vivekananda, and Kamal Sarkar. 2014 An HMM Based Named Entity Recognition System for Indian Languages. The JU System at ICON 2013.

[7] Vishal gupta and Gurpreet Singh Lehal. (2014).Named Entity Recognition for Punjabi Language Text Summarization. (IJCA)International Journal of Computer Applications vol. 33, no. 3 (pp. 28–32).

[8] S. Chan, W. Lam, X. Yu. 2007. A cascaded approach to biomedical named entity recognition using a unified model. In the proceedings of Seventh IEEE International Conference on Data Mining ICDM 2007 (pp. 93-102). IEEE.

[9] McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. Corpus processing for lexical acquisition (pp. 21-39).

[10] Wakao, T., Gaizauskas, R., & Wilks, Y. (1996, August). Evaluation of an algorithm for the recognition and classification of proper names. In Proceedings of the 16th conference on Computational linguistics-Volume 1(pp. 418-423). Association for Computational Linguistics.

[11] Wang, Jing Liu, Zhijing1.2008 A novel arithmetic of named entity identification. In the Proceedings of 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008 vol. 4 (pp. 457-461). IEEE.

[12] D.Klein, J.Smarr, H.Nguyen, C.Manning. 2003.Named entity recognition with character-level models. In the Proceedings of the seventh conference on Natural language learning, HLTNAACL 2003, vol. 4, pp. 180-183.ACM.

[13] B. Todorovic,S. Rancic, I. Markovic, E. Mulalic, V. Ilic. 2008. Named Entity Recognition and Classification using Context Hidden Markov Model. In the proceedings of Neural Network Applications in Electrical Engineering, NEUREL 2008 no. 1 (pp. 43-46).IEEE.

[14] R. Florian, A. Ittycheriah, H. Jing, T. Zang (2003, May). Named entity recognition through classifier combination. In the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 168-171). Association for Computational Linguistics,

[15] Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997, March). Nymble: a high-performance learning name-finder. In Proceedings of the fifth conference on Applied natural language processing (pp. 194-201). Association for Computational Linguistics.

[16] A.Krishnarao, H. Gahlot, A. Srinet, D. Khushwaha. 2009 A Comparative Study of Named Entity Recognition for Hindi Using Sequential Learning Algorithms. In the proceedings of International Advance Computing Conference, IACC pp. 1164-1169.IEEE.

[17] Asahara, M., & Matsumoto, Y. (2003, May). Japanese named entity extraction with redundant morphological analysis. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 8-15). Association for Computational Linguistics.

[18] S. Chan, W. Lam, X. Yu. 2007. A cascaded approach to biomedical named entity recognition using a unified model. In the proceedings of Seventh IEEE International Conference on Data Mining ICDM 2007 (pp. 93-102). IEEE.

[19] Borthwick, Andrew, Sterling, J., Agichtein, E., Grishman, R. 1998. NYU: Description of the MENE Named Entity System as used in MUC-7. In the Proceeding Seventh Message Understanding Conference (MUC-7).

[20] Zhang, T., & Johnson, D. (2003, May). A robust risk minimization based named entity recognition system. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 204-207). Association for Computational Linguistics.

[21] McCallum, A., & Li, W. (2003, May). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 188-191). Association for Computational Linguistics.

[22] Aman deep Kaur & Gurpreet singh josan (2014, March). Improved Named Entity Tagset for Punjabi Language. In Engineering and Computational Sciences (RAECS), 2014 Recent Advances in (pp. 1-5). IEEE.