Named Entity Recognition for Punjabi Language Text Summarization

Vishal Gupta
Assistant Professor,
UIET, Panjab University
Chandigarh, India.

Gurpreet Singh Lehal
Professor,
Department of Computer Science,
Punjabi University Patiala, India.

ABSTRACT
Named Entity Recognition (NER) is used to locate and classify atomic elements in text into predetermined classes such as the names of persons, organizations, locations, concepts etc. NER is used in many applications like text summarization, text classification, question answering and machine translation systems etc. For English a lot of work has already done in field of NER, where capitalization is a major clue for rules, whereas Indian Languages do not have such feature. This makes the task difficult for Indian languages. This paper explains the Named Entity Recognition System for Punjabi language text summarization. A Condition based approach has been used for developing NER system for Punjabi language. Various rules have been developed like prefix rule, suffix rule, propername rule, middlename rule and lastname rule. For implementing NER, various resources in Punjabi, have been developed like a list of prefix names, a list of suffix names, a list of proper names, middle names and last names. The Precision, Recall and F-Score for condition based NER approach are 89.32%, 83.4% and 86.25% respectively.

General Terms
Natural Language Processing, Text Mining, Text Summarization, Named Entity Recognition

Keywords
Punjabi language Named Entity Recognition (NER), condition based approach, Punjabi proper nouns identification and Named Entities (NE) etc.

1. INTRODUCTION
Named entities (NE) are those words containing the names of persons, locations, organization and concept etc. Named entity recognition (NER) is widely used in Natural Language Processing. The task of Named entity recognition is to categorize all proper nouns in a document into predefined classes like person, organization, location, etc. NER has many applications in NLP like machine translation, question-answering systems, indexing for information retrieval, data classification and automatic summarization. It is a two step process consisting of the identification of proper nouns and their classification. Identification is concerned with marking the presence of a word/phrase as NE in the given sentences and classification is for denoting role of the identified NE.

We have implemented a condition based method for Punjabi language named entity recognition for text summarization. Punjabi Text summarization system gives importance to those sentences containing named entities. Although a lot of work has been done in English and other foreign languages like Spanish, Chinese etc with high accuracy but regarding research in Indian languages is at initial stage only.

2. LITERATURE REVIEW
There are mainly two approaches to NER, linguistic approaches and machine learning approaches. Rule based models are used by linguistic approaches. Large amount of training data is used by machine learning approaches to acquire high-level language knowledge. Maximum Entropy Model (MaxEnt) , Decision Tree [1], Support Vector Machines [2] and Conditional Random Fields (CRFs) [3] are the various machine learning approaches used for NER tasks. Gazetteer lists may be used in both the approaches to build the NER system. A rule-based NER system has been developed by Grishman, 1995 [4] by using specialized name dictionaries including names of all countries, names of major cities, names of companies, common first names etc. For identifying the named entities in text, a set of rules or patterns is defined in rule based approaches. Gaizauskas et. al, 1996 [5] has developed another rule based NER system, which make use of several gazetteers like person name, organization name, location names, person names, human titles etc. MaxEnt based ML system [6] has been developed by Borthwick, 1999. 8 dictionaries have been used by this system. Using supervised and unsupervised learning, a lot of work has been done on NER for English language. Because of the capitalization of names in English, it is easier to identify NE. Labeled training data is not required for unsupervised learning approaches i.e. training requires few seed lists and large un annotated corpora. The goal of unsupervised learning is to build representations from data. For data compression, classifying, decision making and other purposes, these representations are then be used. By the use of unlabelled examples of data, Collins et. al, [7] has discussed an unsupervised model for named entity classification. Unsupervised named entity classification models [8] have been proposed by Kim et. al, 2002. A program that can learn to classify a given set of labeled examples that are made up of the same number of features involves using supervised learning. Preparing labeled training data to construct a statistical model is required in supervised learning approach.

Hidden Markov Model is a generative model. HMM is double stochastic process. First process generates the sequence of states Second stochastic process in HMM is responsible for generating the sequence of observations from the sequence of states. As its basic theory is elegant and easy to understand, it is advantageous.

The maximum entropy framework estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Maximum Entropy Markov

Models is a conditional probabilistic sequence model [9]. The basic idea behind this model is maximum entropy which states that the least biased model which considers all know facts is the one which maximizes entropy.

The states are associated with output labels. The benefit of this model is that it has resolved the problem of multiple feature representation and long term dependency issue faced by HMM. As compared to HMM, it has generally increased precision and recall. Label bias problem is the disadvantage of this model. It is biased towards states with lower outgoing transitions because the probability transition leaving any given state must sum to one. All observations are ignored for a state with single outgoing state transition. State transition can be changed for handling label bias problem.

Conditional Random Fields (CRFs) are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite -state machine. Based on values assigned to other assigned input nodes, this model is used to calculate the conditional probability of values on assigned output nodes. CRF model is very much useful in named entity recognition.

Support Vector Machines SVMs [10] are well known for their good generalization performance, and have been applied to many pattern recognition problems. Training and testing data are involved in classification task, which consist of some data instances. One class label and several features are contained in each instance in training set. In this classification model each name takes a fix label and its member of one fix class. In basic form, a SVM learns to find a linear hyperlane that separate both positive and negative examples with maximal margin. This learning bias has proved to have good properties in terms of generalization bounds for the induced classifiers.

A hybrid named entity recognition system is combination of more than one approaches of NER. A hybrid system has been introduced by Sirihari et. al.,2000 [11] which is combination of HMM, MaxEnt, and handcrafted grammatical rules.

Regarding the Indian languages, Punjabi language NER has been developed by Kaur et al., 2009 [12] using Conditional Random Field approach and reported precision, recall and F-score values of 88.05%, 74.85% and 80.92% respectively. The work regarding Telugu language [13] is mentioned in (Shishtla et al., 2008). The evaluation has reported precision, recall and F-score of 64.07%, 34.57% and 44.91% respectively. Ekbal et al., 2008 [14] reports about the development of a NER system for Bengali language. Recall, Precision and F-score are claimed to be 94.3%, 89.4% and 91.8% respectively. The work of (Gali et al., 2008) [15] has reported Lexical F-score of 40.63%, 50.06%, 39.04%, 40.94% and 43.46% for Bengali, Hindi, Oriya, Telugu and Urdu respectively. In (Krishnarao et al., 2007) a comparative study [16] of Conditional Random Field and Support Vector Machines for recognizing named entities in Hindi language is done. They have claimed F-score (lexical) of 47% and 37% for CRF and SVM respectively. Here are the characteristics and some problems faced by Hindi, Punjabi and other Indian languages"

- No capitalization
- Non-availability of large gazetteer
- Lack of standardization and spelling
- Number of frequently used words (common nouns) which can also be used as names are very large. "Also the frequency with which they can be used as common noun as against person name is more or less unpredictable."
- Lack of labeled data
- Scarcity of resources and tools
- Free word order language

3. METHODOLOGY

The Punjabi Named Entity Recognition System uses various gazetteer lists like prefix list, suffix list, middle name list, last name list and proper name lists for checking whether the a given word is proper name or not. After doing aanalysis of Punjabi corpus of Ajit Punjabi newspaper, various gazetteer lists have been developed. This corpus contains around 11.29 million words and 200003 unique words.

3.1 Punjabi Names Prefix List

The Prefix list contains various prefixes of names for checking whether next word is a proper name or not like ਸ. sa., ਸ਼੍ਰੀ. srī., ਡਾ: ḍā:, ਪ੍ਰਿ. pri. and ਸ਼੍ਰੀਮਤੀ srīmtī etc. We have manually analyzed these unique words and identified 14 prefixes and a list is developed by creating a frequency list from the Punjabi corpus. In the Punjabi corpus of 11.29 million words, the frequency count of these prefix words is 17127, which covers 0.15% of the corpus. Some of the most commonly occurring prefix words are displayed in Table1.

Table 1. Punjabi language Prefix words list

ਸ. (sa.)	ਸ਼੍ਰੀ. (srī.)	ਡਾ: (ḍā)
ਪ੍ਰਿ. (pri)	ਸ਼੍ਰੀਮਤੀ (srīmtī)	ਪ੍ਰੋ. (prō.)
ਇੰ. (iṃ.)	ਸ਼੍ਰੀਮਾਨ (srīmān)	ਪ੍ਰੌ. (prau)

3.2 Punjabi Names Suffix List

The Suffix list contains various suffixes of names for checking whether the current word is a proper name or not, like ਪੁਰ pur, ਗੜ੍ਹ gaḍah, ਪੁਰਾ purā, ਪੁਰੀ purī and ਜੀਤ jīt etc. We have manually analyzed unique words of Punjabi corpus and identified 50 suffixes and a list is developed by creating a frequency list from corpus. In the Punjabi corpus of 11.29 million words, the frequency count of suffix words is 225306, which covers 1.99% of the corpus. Some of the most commonly occurring suffix words are displayed in Table2.

Table 2. Punjabi language suffix list

ਪੁਰਾ (purā)	ਪੁਰੀ (purī)	ਜੀਤ (jīt)
ਮੀਤ (mīt)	ਜੋਤ (jōt)	ਦੀਪ (Dīp)

ਪੁਰ (pur)	ਬੀਰ (bīr)	ਗੜ੍ਹ (gaḍ੍ah)

3.3 Punjabi Middle Names List

The Punjabi middle name list contains various middle names of persons for checking whether that word is proper name or not, like ਕੁਮਾਰ kumār, ਕੁਮਾਰੀ kumārī and ਕੌਰ kaur etc. After manually analyzing unique words of Punjabi corpus, we have identified mainly 8 middle names and a list is developed by creating a frequency list from corpus. In the Punjabi corpus, the frequency count of middle name words is 97907, which covers 0.8672% of the corpus. Some of the most commonly occurring middle names are displayed in Table3.

Table 3. Punjabi language middle names list

ਕੁਮਾਰ (Kumār)	ਲਾਲ (lāl)
ਕੌਰ (kaur)	ਸਿੰਘ (siṅgh)
ਕੁਮਾਰੀ (kumārī)	ਕੁਮਾਰੀ (kumārī)

3.4 Punjabi Last Names List

The Punjabi last name list contains various last names of persons for checking whether that word is proper name or not, like ਖੁਰਾਨਾ khurānā, ਗੋਇਲ gōil, ਅੱਗਰਵਾਲ aggravāl, ਗੁਲਾਟੀ gulāṭ̣ī and ਕੱਕੜ kakkaṛ, etc. After manually analyzing unique words of Punjabi corpus, we have identified 310 last names and a list is developed by creating a frequency list from corpus. In the Punjabi corpus, the frequency count of last name words is 69268, which covers 0.6135% of the corpus. Some of the most commonly occurring last names are displayed in Table 4.

Table 4. Punjabi language last names list

ਅੱਗਰਵਾਲ (aggravāl)	ਖੁਰਾਨਾ (khurānā)	ਗੋਇਲ (gōil)
ਗੁਲਾਟੀ (gulāṭ̣ī)	ਕੱਕੜ (kakkaṛ)	ਰਾਵਤ (rāvat)
ਭੰਡਾਰੀ (bhaṇḍ̣ārī)	ਸ਼ੁਕਲਾ (shuklā)	ਮਹਿਰਾ (mehra)

3.5 Punjabi Proper Names List

Proper names are very much important in deciding a sentence's importance. Those sentences containing proper names are important. Some of Punjabi language proper nouns are given below in Table 5.

Table 5. Punjabi language proper names list

ਬਾਦਲ (bādal)	ਪਟਿਆਲਾ (paṭ̣iālā)	ਸੁਰਜੀਤ (surjīt)
ਜਲੰਧਰ (jalndhar)	ਭਾਜਪਾ (bhājapā)	ਨੰਗਲ (naṅgal)
ਮਨਪ੍ਰੀਤ (manprīt)	ਜਸਵੀਰ (jasvīr)	ਦੁਬਈ (dubaī)

3.6 Algorithm of condition based Named Entity Recognition for Punjabi Language

Initially Set the NER scores of each sentence as 0. For every word in each sentence follow following steps:

- Step 1 : If current Punjabi word is in Prefix list then increment the NER score of current sentence by 1 and prefix flag is set to true.
- Step 2 : Else If current Punjabi word Ends with any of words in suffix list then
  If prefix flag is false then increment the NER score of current sentence by 1 and suffix flag is set to true.
  Else set prefix flag to false.
- Step 3 : Else If current Punjabi word is in middle name list then middle name flag is set to true.
  If prefix flag, proper name flag and suffix flag are false then increment the NER score of current sentence by 1.
  Else set prefix flag, proper name flag and suffix flag to false.
- Step 4 :Else If current Punjabi word is in last name list then
  If prefix flag, proper name flag, suffix flag and middle name flag are false then increment the NER score of current sentence by 1.
  Else set prefix flag, proper name flag, suffix flag and middle name flag to false.
- Step 5 :Else If current Punjabi word is in proper name list then proper name flag is set true
  If prefix flag, suffix flag and middle name flag are false then increment the NER score of current sentence by 1.
  Else set prefix flag, suffix flag and middle name flag to false
  End If
- Step 6 :If middle name flag is true and Next Punjabi word is not in last name list or suffix list then middle name flag is set to false.
- Step 7 :If Suffix name flag is true and Next Punjabi word is not in last name list or middle name list then suffix flag is set to false.
- Step 8 :If prefix flag is true and Next Punjabi word is not in last name list or middle name list or suffix list or proper name list then prefix flag is set to false.
- Step 9 :If proper name flag is true and Next Punjabi word is not in last name list or middle name list then prefix flag is set to false.
- Step 7: Set all the flags to false at end of a sentence.

Input Text:

ਜ਼ਿਲ੍ਹਾ ਬਰਨਾਲਾ ਦੇ ਡਿਪਟੀ ਕਮਿਸ਼ਨਰ ਸ: ਅਰਸ਼ਦੀਪ ਸਿੰਘ ਥਿੰਦ ਨੇ ਜ਼ਿਲ੍ਹੇ ਅੰਦਰ ਘਰੇਲੂ ਗੈਸ ਦੀ ਸਮੱਸਿਆ ਨੂੰ ਮੁੱਖ ਰਖਦਿਆਂ ਗੈਸ ਵਿਭਾਗ

ਨਾਲ ਸਬੰਧਿਤ ਅਧਿਕਾਰੀਆਂ ਤੇ ਸਮੂਹ ਗੈਸ ਏਜੰਸੀਆਂ ਦੇ ਮਾਲਕਾਂ ਨਾਲ ਅਹਿਮ ਮੀਟਿੰਗ ਕੀਤੀ।

Barnālā dē ḍipṭī kamishanar sa: arashdīp siṅgh thind nē zilhē andar gharēlū gais dī samssiā nūṃ mukkh rakhdiāṃ gais vibhāg nāl sabndhit adhikārīāṃ tē samūh gais ējṃsīāṃ dē mālkāṃ nāl ahim mīṭiṅg kītī.

Output of NER:
ਬਰਨਾਲਾ barnālā and ਸ: ਅਰਸ਼ਦੀਪ ਸਿੰਘ ਥਿੰਦ sa: arashdīp siṅgh thind with NER Score=2.

Prefix rule (Rule1) has been used 5%, Suffix rule (Rule2) has been used 20%, Middle name rule (Rule3) has been used 16%, Last name rule (Rule4) has been used 11% and proper names rule (Rule5) has been used 48%. We can interpret that, the prefix rule is occasionally used as very less number of Punjabi names start with prefixes. The proper names rule (Rule5) has been largely used (48% usage) as in most of cases, Punjabi names appear as proper names without prefix, suffix, middle name and last name.

4. IMPLEMENTATION AND RESULTS

Punjabi Language NER system for Text Summarization has been implemented in VB.NET at front end and MS Access at back end. Regarding condition based NER system, an in depth analysis of output has been done over 50 Punjabi news documents as input. It is producing Precision=89.32%, Recall=83.4% and F-score=86.25%. An in depth error analysis of condition based system has been done over 50 news documents and it is giving 13.75% errors. Prefix rule is producing no errors, Suffix rule is producing 1% errors like ਅਫਸਰ aphsar and ਕਰਿਆਣਾ kariāṇā both are not present in Punjabi noun dictionary and lie under suffix rule which makes them as proper names which is not true. Middle name rule is producing 0.25% errors like a name ਕੌਰ ਸਿੰਘ Kaur siṅgh, as in this name, both middle names come together as a single name, but both lie under middle name rule and make NER score to 2 in this case. Last name rule is producing 10% errors like in case of a name ਕਰਤਾਰ ਸਿੰਘ ਜੰਗੀਆਣਾ kratār siṅgh jaṅgīāṇā and ਬੰਤਾ ਸਿੰਘ ਬੰਟੀ bantā siṅgh baṇṭī both are having last names not in last names list but ਜੰਗੀਆਣਾ jaṅgīāṇā and ਬੰਟੀ baṇṭī are in proper names list which will wrongly increment their NER scores to 2 for each case. Proper names rule is producing 0.25% errors, for example ਨਿਹਾਲ nihāl some times is used as proper name and some times in other context, but as it is lying in proper names list so NER system will always treat it as a proper name. Rest 2.25% errors are due to those proper names who do not lie under any of rules like ਗਰੀਣ ਐਵਿਨਿਊ grīṇ aiviniū or due to those proper names like ਬਹਾਦਰ bhādr which is some times used as proper name and some times used as noun.

5. CONCLUSIONS

In this paper, we have discussed the condition based named entity recognition approach for Punjabi text summarization system. Most of the lexical resources used in NER such as prefix list, Suffix list, Middle name list, Last name list and Punjabi proper name list had to be developed from scratch as no work had been done in that direction.
In condition based approach, five rules have been implemented like prefix rule, Suffix rule, Middle name rule, last name rule and proper name rule. For Punjabi language it is first time that these resources have developed and these may be helpful for future NLP applications in Punjabi.

6. REFERENCES

[1] Hideki Isozaki. 2001 Japanese named entity recognition based on a simple rule generator and decision tree learning" in the proceedings of the Association for Computational Linguistics, (pp 306-313). India.

[2] Takeuchi K. and Collier N. 2002 Use of Support Vector Machines in extended named entity Recognition. In the proceedings of the sixth Conference on Natural Language Learning (CoNLL-02), Taipei, Taiwan, China.

[3] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In the proceedings of International Conference on Machine Learning (pp 282-289). Williams College,

[4] R. Grishman 1995 The NYU system for MUC-6 or Where's the Syntax. In the proceedings of Sixth Message Understanding Conference (MUC-6) (pp167-195). Fairfax, Virginia.

[5] Wakao T., Gaizauskas R. and Wilks Y. (1996). Evaluation of an algorithm for the Recognition and Classification of Proper Names. In the proceedings of COLING-96.

[6] Andrew Borthwick. 1999 Maximum Entropy Approach to Named Entity Recognition, doctoral dissertation, New York University.

[7] Collins, Michael and Y. Singer 1999 Unsupervised models for Named Entity Classification. In the proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[8] J. Kim, I. Kang, K. Choi. 2002 Unsupervised Named Entity Classification Models and their Ensembles. In the proceedings of the 19th International Conference on Computational Linguistics.

[9] Darvinder Kaur and Vishal Gupta. 2010 A survey of Named Entity Recognition in English and other Indian Languages. In Proceedings of (IJCSI) International Journal of Computer Science Issues Vol. 7 Issue 6 (pp 239-245).

[10] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat 2008 Named Entity Recognition Approaches. In Proceedings of IJCSNS International Journal of Computer Science and Network Security VOL.8 No.2 pp 339-344

[11] Srihari R., Niu C. and Li W. 2000 A Hybrid Approach for Named Entity and Sub-Type Tagging. In the proceedings of the sixth Conference on Applied Natural Language Processing.

[12] Amandeep Kaur, Gurpreet Singh Josan and Jagroop Kaur. 2009 Named Entity Recognition for Punjabi: A Conditional Random Field Approach. In Proceedings of 7th international conference on Natural Language ProcessingICON-09. Macmillan Publishers, India.

[13] Praneeth M Shishtla, Karthik, Prasad Pingali and Vasudeva Verma 2008Experiments in Telgu NER: A Conditional Random Field Approach. In Proceedings of the IJCNLP-08

workshop on NER for South and South , East Asian Languages (pp105-110). Hyderabad, India.

[14] Asif Ekbal, Sivaji Bandyopadhyay.2008 Bengali Named Entity Recognition using Support Vector Machine. In the Proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages (pp 51-58). Hyderabad, India.

[15] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma.2008 Aggregating Macine Learning and Rule Based Heuristics for NER. In the Proceedings of the IJCNLP-08 worksop on NER for South and South East Asian Languages (pp 25-32). Hyderabad, India.

[16] Awaghad Ashish Krishnarao 2009 A Comparison of Performance of Sequential Learning Algorithm on task of NER for Indian Languages. In the Proceedings of the 9th International Conference on Computer Science (pp 123-132). Baton Rouge, LA, USA.