Think Globally, Apply Locally: Using Distributional Characteristics for Hindi Named Entity Identification

Shalini Gupta Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay
Mumbai, India.
{shalini, pb}@cse.iitb.ac.in

Abstract

In this paper, we present a novel approach for Hindi Named Entity Identification (NEI) in a large corpus. The key idea is to harness the global distributional characteristics of the words in the corpus. We show that combining the global distributional characteristics along with the local context information improves the NEI performance over statistical baseline systems that employ only local context. The improvement is very significant (about 10%) in scenarios where the test and train corpus belong to different genres. We also propose a novel measure for NEI based on term informativeness and show that it is competitive with the best measure and better than other well known information measures.

1 Introduction

NER is the task of identifying and classifying words in a document into predefined classes like person, location, organization, *etc.* It has many applications in Natural Language Processing (NLP) NER can be divided into two sub-tasks, Named Entity Identification (NEI) and Named Entity Classification (NEC). In this paper, we focus on the first step, *i.e.*, Named Entity Identification. NEI is useful in applications where a list of Named Entities (NEs) is required. Machine Translation needs identification of named entities, so that they can be transliterated.

For Indian languages, it is tough to identify named entities because of the lack of capitalization. Many approaches based on MEMM (Saha et al., 2008b), CRFs (Li and McCallum, 2003) and hybrid models have been tried for Hindi Named Entity Recognition. These approaches use only the local context for tagging the text. Many applications need entity identification in large corpora. When such a large corpus is to be tagged, one can use the global distributional characteristics of the words to identify the named entities. The state-of-the-art methods do not take advantage of these characteristics. Also, the performance of these systems degrades when the training and test corpus are from different domain or different genre. We present here our approach-*Combined Local and Global Information for Named Entity Identification* (CLGIN) which combines the global characteristics with the local context for Hindi Named Entity Identification. The approach comprises of two steps: (i) *Named Entity Identification using Global Information* (NGI) which uses the global distributional characteristics along with the language cues to identify NEs and (ii) Combining the tagging from step 1 with the MEMM based statistical system. We consider the MEMM based statistical system (S-MEMM) as the Baseline. Results show that the CLGIN approach outperforms the baseline S-MEMM system by a margin of about 10% when the training and test corpus belong to different genre and by a margin of about 2% when both, training and test corpus are similar. NGI also outperforms the baseline, in the former case, when training and test corpus are from different genre. Our contributions in this paper are:

- Developing an approach of harnessing the global characteristics of the corpus for Hindi Named Entity Identification using information measures, distributional similarity, lexicon, term co-occurrence and language cues
- Demonstrating that combining the global characteristics with the local contexts improves the accuracy; and with a very significant amount when the train and test corpus are not from same domain or similar genre
- Demonstrating that the system using only the

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 116–125,
Uppsala, Sweden, 16 July 2010.

global characteristics is also quite comparable with the existing systems and performs better than them, when train and test corpus are unrelated

- Introducing a new scoring function, which is quite competitive with the best measure and better than other well known information measures

Approach	Description
S-MEMM (Baseline)	MEMM based statistical system without inserting global information
NGI	Uses global distributional characteristics along with language information for NE Identification
CLGIN	Combines the global characteristics derived using NGI with S-MEMM

Table 1: Summary of Approaches

2 Related Work

There is a plethora of work on NER for English ranging from supervised approaches like HMMs(Bikel et al., 1999), Maximum Entropy (Borthwick, 1999) (Borthwick et al., 1998), CRF (Lafferty et al., 2001) and SVMs to unsupervised (Alfonseca and Manandhar, 2002), (Volker, 2005) and semi-supervised approaches (Li and Mccallum, 2005). However, these approaches do not perform well for Indian languages mainly due to lack of capitalization and unavailability of good gazetteer lists. The best F Score reported for Hindi NER using these approaches on a standard corpus (IJCNLP) is 65.13% ((Saha et al., 2008a)). Higher accuracies have been reported (81%) (Saha et al., 2008b), albeit, on a non-standard corpus using rules and comprehensive gazetteers.

Current state-of-the-art systems (Li and McCallum, 2003) (Saha et al., 2008b) use various language independent and language specific features, like, context word information, POS tags, suffix and prefix information, gazetteer lists, common preceding and following words, *etc*. The performance of these systems is significantly hampered when the test corpus is not similar to the training corpus. Few studies (Guo et al., 2009), (Poibeau and Kosseim, 2001) have been performed towards genre/domain adaptation. But this still remains an open area. Moreover, no work has been done towards this for Indian languages.

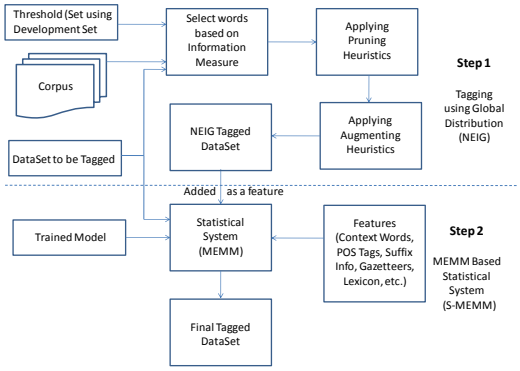

Figure 1: Block diagram of CLGIN Approach

One shortcoming of current approaches is that they do not leverage on global distributional characteristics of words (e.g., Information Content, Term Co-occurrence statistics, etc.) when a large corpus needs NEI. Rennie and Jaakkola (2005) introduced a new information measure and used it for NE detection. They used this approach only on uncapitalized and ungrammatical English text, like blogs where spellings and POS tags are not correct. Some semi-supervised approaches (Collins and Singer, 1999), (Riloff and Jones, 1999), (Paşca, 2007) have also used large available corpora to generate context patterns for named entities or for generating gazetteer lists and entity expansion using seed entities. Klementiev and Roth (2006) use cooccurrence of sets of terms within documents to boost the certainty (in a cross-lingual setting) that the terms in question were really transliterations of each other.

In this paper, we contend that using such global distributional characteristics improves the performance of Hindi NEI when applied to a large corpus. Further, we show that the performance of such systems which use global distribution characteristics is better than current state-of-the-art systems when the training and test corpus are not similar (different domain/genre) thereby being more suitable for domain adaptation.

3 MEMM based Statistical System (S-MEMM)

We implemented the Maximum Entropy Markov Model based system(Saha et al., 2008b) for NE Identification. We use this system as our Baseline and compare our approaches NGI and CLGIN with this baseline. We used various language dependent and independent features. An important

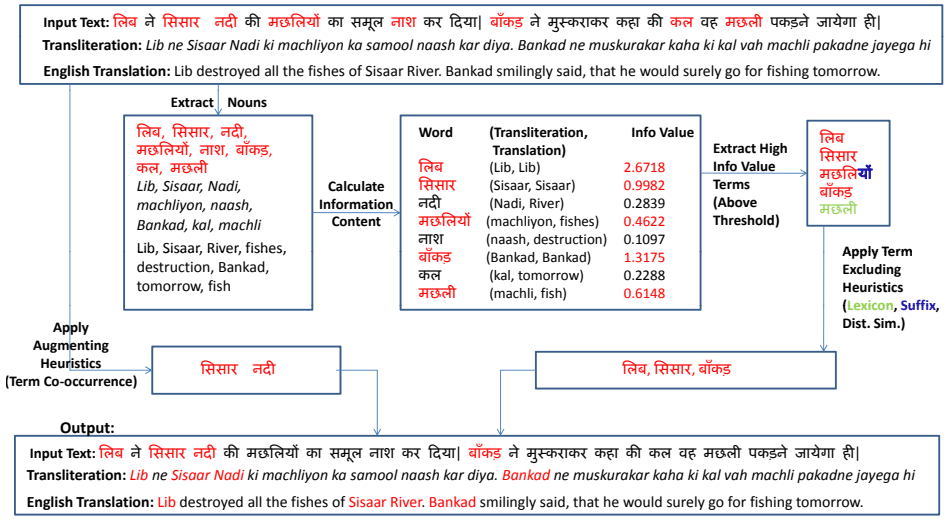

Figure 2: An Example explaining the NGI approach

modification was the use of **lexicon** along with traditionally used gazetteers. Gazetteers just improve the recall whereas including the lexicon improves the precision. The state-of-art Hindi NER systems do not use lexicon of general words but we found that using lexicons significantly improves the performance. Unlike English, NEs in Hindi are not capitalized and hence it becomes important to know, if a word is a common word or not.

Features used in S-MEMM were:

- Context Words: Preceding and succeeding two words of the current word
- Word suffix and prefix: Fixed length (size: 2) suffix information was used. Besides, suffix list of common location suffixes was created
- First word and last word information
- Previous NE Tag information
- Digit information
- Gazetteer Lists: Person and Location names, Frequent words after and before person, organization and location names, list of common initials, stopwords, *etc*.
- POS Tag Information
- Lexicons: If the stemmed word was present in the lexicon, this feature was true.

4 Our Approach-CLGIN

In this section, we describe our approach, CLGIN in detail. It combines the global information from the corpus with the local context. Figure 1 gives the block diagram of the system while tagging a corpus and Figure 2 explains the approach using an example. This approach involves two steps. Step 1 of CLGIN is NGI which creates a list of probable NEs (both uni-word and multi-word) from the given corpus and uses it to tag the whole corpus. Sections 4.1 and 4.2 explain this step in detail. Later, in step 2, it combines the tagging obtained from step 1, as a feature in the MEMM based statistical system. Output thus obtained from the MEMM system is the final output of the CLGIN approach. The creation of list in step 1, involves the following steps

- A list of all words which appeared as a noun at least once in the the corpus is extracted.
- List is ordered on the basis of the information content derived using the whole corpus. Words above the threshold (set during training using the development set) are selected as NEs.
- Heuristics are applied for pruning and augmenting the list.
- Multi-word NEs derived using term co-occurrence statistics along with language characteristics are added to the NE list.

The above process generates a list of NEs (uni-word and multi-word). In the second step, we provide this tagging to the S-MEMM along with other set of features described in Section 3

During training, the cutoff threshold is set for selecting NEs (in bullet 2) above. Also the tagging obtained from the step 1 is added as a feature to

S-MEMM and a model is trained during the training phase. The following sections describe this approach in detail.

4.1 Information Measures/Scoring Functions

Various measures have been introduced for determining the information content of the words. These include, IDF (Inverse Document Frequency) (Jones, 1972) , Residual IDF (Church and Gale, 1995), x^I- measure (Bookstein and Swanson, 1974), Gain (Papineni, 2001), *etc*. We introduced our own information measure, RF (Ratio of Frequencies).

4.1.1 RF (Ratio of Frequencies)

NEs are highly relevant words in a document (Clifton et al., 2002) and are expected to have high information content (Rennie and Jaakkola, 2005). It has been found that words that appear frequently in a set of documents and not so frequently in the rest of the documents are important with respect to that set of documents where they are frequent.

We expected the NEs to be concentrated in few documents. We defined a new criteria which measures the ratio of the total number of times the word appears in the corpus to the number of documents containing a word.

$RF(w) = \frac{cf(w)}{df(w)}$

where cf(w) is the total frequency of a word in the whole corpus and df(w) is the document frequency. This measure is different from the TF-IDF measure in terms of the term frequency. TF-IDF considers the frequency of the word in the document. RF considers it over the whole corpus.

We use the scoring function (information measure) to score all the words. During training, we fix a threshold using the development set. During testing, we pick words above the threshold as NEs. We then apply heuristics to augment this list as well as to exclude terms from the generated list.

4.2 Heuristics for Pruning and Augmenting NE List

Distributional Similarity: The underlying idea of Distributional Similarity is that a word is characterized by the company it keeps (Firth, 1957). Two words are said to be distributionally similar if they appear in similar contexts. From the previous step (Sect. 4.1), we get a list of words having high score. Say, top t, words were selected. In this step, we take t more words and then cluster together these words. The purpose at this phase is primarily to remove the false positives and to introduce more words which are expected to be NEs. For each distinct word, w in the corpus, we create a vector of the size of the number of distinct words in the corpus. Each term in the vector represents the frequency with which it appears in the context (context window: size 3) of word, w. It was observed that the NEs were clustered in some clusters and general words in other clusters. We tag a cluster as a NE cluster if most of the words in the cluster are good words. We define a word as good if it has high information content. If the sum of the ranks of 50% of the top ranked word is low, we tag the cluster as NE and add the words in that set as NEs. Also, if most of the words in the cluster have higher rank i.e. lower information content, we remove it from the NE set.

This heuristic is used for both **augmenting** the list as well to **exclude** terms from the list.

Lexicon: We used this as a list for **excluding** terms. Terms present in the lexicon have a high chance of not being NEs. When used alone, the lexicon is not very effective (explained in Section 5.2). But, when used with other approaches, it helps in improving the precision of the system significantly. State-of-art Hindi NER systems use lists of gazetteers for Person names, location names, organization names, *etc*. (Sangal et al., 2008), but lexicon of general words has not been used. Unlike English, for Indian languages, it is important to know, if a word is a general word or not. Lexicons as opposed to gazetteers are generic and can be applied to any domain. Unlike gazetteers, the words would be quite common and would appear in any text irrespective of the domain.

Suffixes: NEs in Hindi are open class words and appear as free morphemes. Unlike nouns, NEs, usually do not take any suffixes (attached to them). However, there are few exceptions like, लाल किले के बाहर (*laal kile ke baahar*, (outside Red Fort)) or when NEs are used as common nouns, देश को गांधियों की जरूरत है (*desh ko gandhiyon ki zaroorat hai*, The country needs Gandhis.) *etc*. We remove words appearing with common suffixes like एं (*ein*), ओं (*on*), येंगे (*yenge*), *etc*. from the NE list.

Term Co-occurrence: We use the term co-occurrence statistics to detect multi-word NEs. A word may be a NE in some context but not in another. E.g. महात्मा (*mahatma* "saint") when ap-

pearing with गांधी (*Gandhi* "Gandhi") is a NE, but may not be, otherwise. To identify such multiwords NEs, we use this heuristic. Such words can be identified using Term Co-occurrence. We use the given set of documents to find all word pairs. We then calculate Pointwise Mutual Information (PMI) (Church and Hanks, 1990) for each of these word pairs and order the pairs in descending order of their PMI values. Most of the word pairs belong to the following categories:

- Adjective Noun combination (Adjectives followed by noun): This was the most frequent combination. E.g. भीनी गंध (*bheeni gandh* "sweet smell")
- Noun Verb combination: दिल धड़कना (*dil dhadakna*, "heart beating")
- Adverb verb combination: खिलखिलाकर हंसना (*khilkhilakar hansna*, "merrily laugh")
- Cardinal/Ordinal Noun Combination: थोड़ी देर (*thodi der*, "some time")
- Named Entities
- Hindi Idioms: उल्लू सीधा (*ullu seedha*)
- Noun Noun Combination: ख्याती अर्जित (*khyati arjit*, "earn fame")
- Hindi Multiwords: जोश खरोश (*josh kharosh*)

We need to extract NEs from these word pairs. The first four combinations can be easily excluded because of the presence of a verb, cardinals and adjectives. Sometimes both words in the NEs appear as nouns. So, we cannot reject the Noun Noun combination. We handle rest of the cases by looking at the neighbours (context) of the word pairs.

We noticed three important things here:

- Multiwords which are followed (alteast once) by में (*mein*), से (*se*), ने (*ne*), के (*ke*), को (*ko*) (Hindi Case Markers) are usually NEs. We did not include की (*ki*) in the list because many words in the noun-noun combination are frequently followed by *ki* in the sense of किया/ करना (*kiya/karna*, "do/did") e.g. ख्याती अर्जित की (*khyati arjit ki*, "earned fame"), परीक्षा उत्तीर्ण की (*pariksha uttirand ki*, "cleared the exam"), *etc*.
- There were word pairs which were followed by a single word most of the time. E.g ईस्ट इंडिया (*East India*, "East India") was followed by कंपनी (*Company*, "Company") in almost all the cases. When *Company* appears alone, it may not be a NE, but when it appears with *East India*, it appears as a NE. Other examples of such word pairs were: खाँ इब्नू (*Khan Ibnu*, "Khan Ibnu") followed by अलीसम (*Alisam*, "Alisam")
- There were word pairs which were followed by uncommon words were not common words but were different words each time, it appeared. i.e. Most of the words following the word pair were not part of lexicon. गवर्नर जनरल (*governor general*, "Governor General") followed by [दलहौसी, बहदुर, सोलबरी, मैटकाफ़, लौर्ड ((*dalhousie, bahadur, solbari, metkaf, lord*), "Dalhousie, Bahadur, Solbari, Metkaf, Lord")] Such words are multi word NEs.

Corpus	No. of Tagged Documents	No. of Words	No. of NEs	Source Genre
Gyaan Nidhi	1570	569K	21K	Essay, Biography, History and Story

Table 2: Corpus Statistics

4.3 Step 2: Combining NGI with S-MEMM

The tagging obtained as the result of the step 1 (NGI), is given as input to the MEMM based statistical system (S-MEMM). This feature is introduced as a binary feature *OldTag=NE*. If a word is tagged as NE in the previous step, this feature is turned on, otherwise *OldTag=O* is turned on.

5 Experiments and Results

We have used *Gyaan Nidhi* Corpus for evaluation which is a collection of various books in Hindi. It contains about 75000 documents. The details of the corpus are given in Table 2. Names of persons, locations, organizations, books, plays, *etc*. were tagged as *NE* and other general words were tagged as *O* (others). The tagged documents are publicly made available at `http://www.cfilt.iitb.ac.in/ner.tar.gz`.

We use the following metrics for evaluation: Precision, Recall and F-Score. Precision is the ratio of the number of words correctly tagged as NEs to the total number of words tagged as NEs. Recall is the ratio of the number of words correctly tagged as NEs to the total number of NEs present in the data set. F Score is defined as $(F = 2 * P * R/(P + R))$

5.1 Comparison of Information Measures

We compare the performance of the various term informativeness measures for NEI which are Residual IDF[1], IDF [2], Gain[3] and x' measure [4] and the measure defined in Section 4.1.1. Table 3 shows the results averaged after five-fold cross validation. The graphs in the Figure 3 to Figure 7 show the distribution of words (nouns) over the range of values of each information measure.

Scoring Function	Prec.	Recall	F Score
Residual IDF	**0.476**	**0.537**	**0.504**
IDF	0.321	0.488	0.387
x-dash Measure	0.125	0.969	0.217
RF (Our Measure)	**0.624**	**0.396**	**0.484**
Gain	0.12	0.887	0.211

Table 3: Comparison of performance of various information measures

The best results were obtained using Residual IDF followed by Ratio of Frequencies (RF).

Method	Prec	Recall	F Score
S -MEMM (Baseline)	**0.871**	**0.762**	**0.812**
Res. IDF	0.476	0.537	0.504
Res. IDF + Dist Sim (DS)	0.588	0.522	0.553
Res. IDF + Lexicon (Lex)	0.586	0.569	0.572
Res. IDF + DS + Suffix	0.611	0.524	0.563
Res. IDF + Lex + Suffix	0.752	0.576	0.65
Res. IDF + Lex + Suffix + Term Cooccur (NGI)	**0.757**	**0.62**	**0.68**
CLGIN	**0.879**	**0.784**	**0.829**

Table 4: Performance of various Approaches (Here, train and test are similar)

5.2 NGI and CLGIN Approaches (Training and Test Set from Similar Genre)

Table 4 compares the results of S-MEMM, NGI approach and CLGIN. Besides, it also shows the step wise improvement of NGI approach. The final F-Score achieved using NGI approach was 68%. The F-Score of the Baseline system implemented using the MaxEnt package[1] from the OpenNLP community was 81.2%.

Using the lexicon alone gives an F-Score of only 11% (Precision: 5.97 Recall: 59.7 F-Score: 10.8562). But, when used with Residual IDF, the performance of the overall system improves significantly to about 57%. Note that, the use of lexicon resulted in an increase in precision (0.5860) which was accompanied by improvement in recall (0.5693) also. The cutoff thresholds in both cases (Rows 2 and 4 of Table 4) were different. Suffix information improved the systems performance to 65%. As words were removed, more words from the initial ordered list (ordered on the basis of score/information content) were added. Hence, there was a small improvement in recall, too. Improvement by distributional similarity was eclipsed after the pruning by lexicon and suffix information. But, in the absence of lexicon; distributional similarity and suffix information can be used as the pruning heuristics. Adding the multiword NEs to the list as explained in the section 4.2 using term co-occurrence statistics, improved the accuracy significantly by 3%. Word pairs were arranged in the decreasing order of their PMI values and a list was created. We found that 50% of the NE word pairs in the whole tagged corpus lied in the top 1% of this word pairs list and about 70% of NE word pairs were covered in just top 2% of the list.

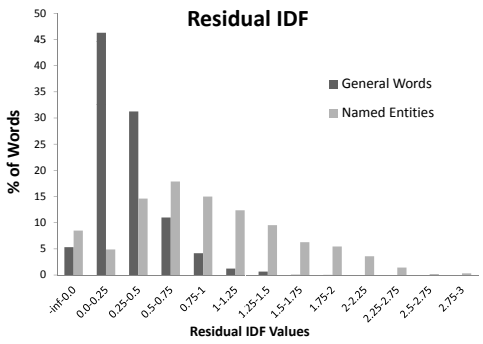

Figure 3: Distribution of Residual IDF values over the nouns in the corpus

CLGIN which combines the global information obtained through NGI with the Baseline S-MEMM system gives an improvement of about 2%. After including this feature, the F-Score increased to 82.8%.

5.3 Performance Comparison of Baseline, NGI and CLGIN (Training and Test Data from different genre)

In the above experiments, documents were randomly placed into different splits. Gyaan Nidhi is a collection of various books on several top-

[1]Observed IDF - Expected IDF
[2]IDF = -$log\frac{df(w)}{D}$
[3]Gain = $\frac{d_w}{D}(\frac{d_w}{D} - 1 - log\frac{d_w}{D})$
[4]$x'(w) = df(w) - cf(w)$
[1]http://maxent.sourceforge.net/index.html

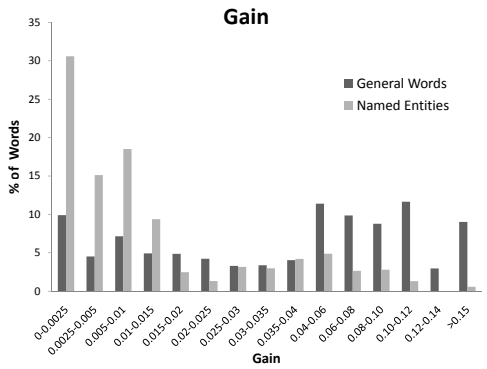

Figure 4: Distribution of Gain values over the nouns in the corpus

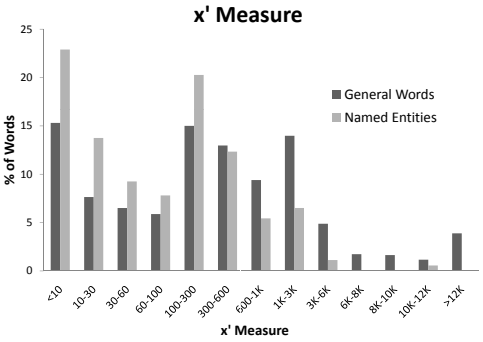

Figure 7: Distribution of x^I measure values over the nouns in the corpus

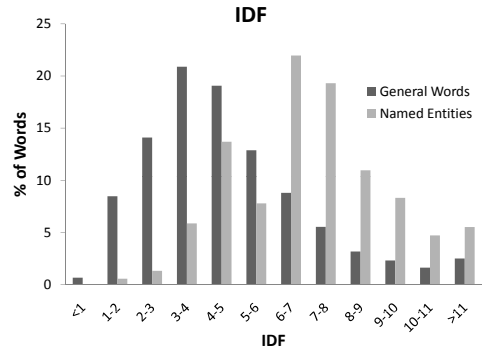

Figure 5: Distribution of IDF values over the nouns in the corpus

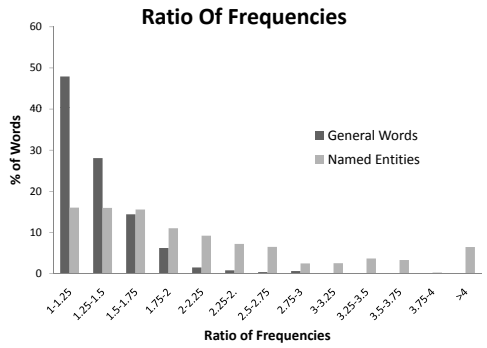

Figure 6: Distribution of Ratio of Frequencies(RF) values over the nouns in the corpus

ics. Random picking resulted into the mixing of the documents, with each split containing documents from all books. But, in this experiment, we divided documents into two groups such that documents from few books (genre: Story and History) were placed into one group and rest into another group (Genre: Biography and Essay). Table 5 compares the NGI and CLGIN approaches with S-MEMM and shows that the **CLGIN results are significantly better than the Baseline System**, when the training and test sets belong to different genre. The results were obtained after 2-fold cross validation.

Method	Prec.	Recall	F Score
S-MEMM	0.842	0.479	0.610
NGI	0.744	0.609	0.67
CLGIN	**0.867**	**0.622**	**0.723**

Table 5: Performance of various Approaches (Here, train and test are from different genre)

Similar improvements were seen when the sets were divided into (Story and Biography) and (Essay and History) (The proportions of train and test sets in this division were uneven). The F Score of NGI system was 0.6576 and S-MEMM was 0.4766. The F Score of the combined system (CLGIN) was 0.6524.

6 Discussion and Error Analysis

6.1 RF and other information measures

As can be seen from the graphs in Figures 3 to 7, Residual IDF best separates the NEs from the general words. The measure introduced by us, Ratio of Frequencies is also a good measure, although not as good as Residual IDF but performs better than other measures. The words having RF value greater than 2.5 can be picked up as NEs, giving a high recall and precision. It is evident that IDF is better than both, Gain and x^I measure, as most of the general words have low IDF and NEs lie in the high IDF zone. But, the general words and NEs are not very clearly separated. As the number of nouns is about 7-8 times the number of NEs, the

words having high IDF cannot be picked up. This would result in a low precision, as a large number of non-NEs would get mixed with the general words. Gain and x^I measure do not demarcate the NEs from the general words clearly. We observed that they are not good scoring functions for NEs.

Information Gain doesn't consider the frequency of the terms within the document itself. It only takes into account the document frequency for each word. x^I measure considers the frequency within document but it is highly biased towards high frequency words and hence doesn't perform well. Hence, common words like समय (*samay*, "time"), घर (*ghar*, "home"), etc. have higher scores compared to NEs like भारत(*bharat*, "India"), कलकत्ता (*kalkatta*, "Calcutta"), *etc.* Our measure on the other hand, overcomes this drawback, by considering the ratio. We could have combined the measures, instead of using only the best measure "Residual IDF", but the performance of "Gain", "IDF" and "x'-measure" was not good. Also, results of "RF" and "Residual IDF" were quite similar. Hence, we did not see any gain in combining the measures.

6.2 S-MEMM, NGI and CLGIN

The results in Section 5 show that adding the global information with the local context helps improve the tagging accuracy especially when the train and test data are from different genre. Several times, the local context is not sufficient to determine the word as a NE. For example, when the NEs are not followed by post positions or case markers, it becomes difficult for S-MEMM to identify NEs, *e.g.*, टैगोर एक अपवाद हैं, (*tagore ek apvaad hain*,"Tagore is an exception") or when the NEs are separated by commas, e.g. सुकुमारी दत्त, चुन्नीलाल.. (*Sukumari Dutt, Chunnilal ...* "Sukumari Dutt, Chunnilal .."). In such cases, because of the frequency statistics, the NGI approach is able to detect the words टैगोर (*Tagore*, "Tagore"), दत्त (*Dutt*, "Dutt"), *etc.* as NEs and frequently the CLGIN approach is able to detect such words as NEs.

The false positives in NEIG are words which are not present in the lexicon (uncommon words, words absent due to spelling variations *e.g.* सांप/साँप (*sanp* "snake")) but have high informativeness. Using the context words of these words is a possible way of eliminating these false positives. Many of the organization names having common words (मंडल (*mandal*, "board")) and person names (like प्रकाश (*prakash*,"light")) are present in the lexicon are not tagged by NEIG. Some errors were introduced because of the removal of morphed words. NEs like गुल्बानों, टोपे (*Gulbano, Tope*) were excluded.

Many of the errors using CLGIN are because of the presence of the words in the lexicon. This effect also gets passed on to the neighbouring words. But, the precision of CLGIN is significantly high compared to NGI because CLGIN uses context, as well.

The statistical system (S-MEMM) provides the context and the global system(NGI) provides a strong indication that the word is a NE and the performance of the combined approach(CLGIN) improves significantly.

7 Conclusion and Future Work

We presented an novel approach for Hindi NEI which combines the global distributional characteristics with local context. Results show that the proposed approach improves performance of NEI significantly, especially, when the train and test corpus belong to different genres. We also proposed a new measure for NEI which is based on term informativeness. The proposed measure performs quite competitively with the best known information measure in literature.

Future direction of the work will be to study the distributional characteristics of individual tags and move towards classification of identified entities. We also plan to extend the above approach to other Indian languages and other domains. We also expect further improvements in accuracy by replacing the MEMM model by CRF. Currently, we use a tagged corpus as development set to tune the cut-off threshold in NGI. To overcome this dependence and to make the approach unsupervised, a way out can be to find an approximation to the ratio of the number of nouns which are NEs to the number of nouns and then use this to decide the cut-off threshold.

Acknowledgments

We would like to acknowledge the efforts of Mr. Prabhakar Pandey and Mr. Devendra Kairwan for tagging the data with NE tags.

References

Enrique Alfonseca and Suresh Manandhar. 2002. An Unsupervised Method For General Named Entity Recognition and Automated Concept Discovery. In *Proceedings of the 1 st International Conference on General WordNet*.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An Algorithm that Learns What's In A Name.

A. Bookstein and D. R. Swanson. 1974. Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, 25:312–318.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Nyu: Description of the MENE Named Entity System as used in MUC-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7*.

Andrew Eliot Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York, NY, USA. Adviser-Grishman, Ralph.

Kenneth Church and William Gale. 1995. Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. In *Third Workshop on Very Large Corpora*, pages 121–130.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography.

Chris Clifton, Robert Cooley, and Jason Rennie. 2002. Topcat: Data mining for Topic Identification in a Text Corpus.

Michael Collins and Yoram Singer. 1999. Unsupervised Models for Named Entity Classification. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

J.R. Firth. 1957. A Synopsis of Linguistic Theory 1930-1955. In *In Studies in Linguistic Analysis*, pages 1–32.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009. Domain Adaptation with Latent Semantic Association for Named Entity Recognition. In *NAACL '09*, pages 281–289, Morristown, NJ, USA. Association for Computational Linguistics.

Karen Sprck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28:11–21.

Alexandre Klementiev and Dan Roth. 2006. Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 82–88, Morristown, NJ, USA. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wei Li and Andrew McCallum. 2003. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):290–294.

Wei Li and Andrew Mccallum. 2005. Semi-supervised Sequence Modeling with Syntactic Topic Models. In *AAAI-05, The Twentieth National Conference on Artificial Intelligence*.

Marius Paşca. 2007. Organizing and Searching the World Wide Web of facts – Step Two: Harnessing the Wisdom of the Crowds. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 101–110, New York, NY, USA. ACM.

Kishore Papineni. 2001. Why Inverse Document Frequency? In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Thierry Poibeau and Leila Kosseim. 2001. Proper Name Extraction from Non-Journalistic Texts. In *In Computational Linguistics in the Netherlands*, pages 144–157.

Jason D. M. Rennie and Tommi Jaakkola. 2005. Using Term Informativeness for Named Entity Detection. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–360, New York, NY, USA. ACM.

Ellen Riloff and Rosie Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, pages 474–479, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, and Pabitra Mitra. 2008a. A Hybrid Named Entity Recognition System for South and South East Asian Languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*,

pages 17–24, Hyderabad, India, January. Asian Federation of Natural Language Processing.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2008b. A Hybrid Feature Set Based Maximum Entropy Hindi Named Entity Recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, Kharagpur, India.

Rajeev Sangal, Dipti Sharma, and Anil Singh, editors. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. Asian Federation of Natural Language Processing, Hyderabad, India, January.

Johanna Volker. 2005. Towards Large-Scale, Open-Domain and Ontology-Based Named Entity Classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05*, pages 166–172. INCOMA Ltd.