ISSN 2007-9737

A Framework that Uses the Web for Named Entity Class Identification: Case Study for Indian Classical Music Forums

Joe Cheri Ross, Aditya Joshi, Pushpak Bhattacharyya

Dept. of Computer Science & Engg., Indian Institute of Technology Bombay, Mumbai, India

{joe, adityaj, pb}@cse.iitb.ac.in

Abstract. Identification of named entity(NE) class (semantic class) is crucial for NLP problems like coreference resolution where semantic compatibility between the entity mentions is imperative to coreference decision. Short and noisy text containing the entity makes it challenging to extract the NE class of the entity through the context. We introduce a framework for named entity class identification for a given entity, using the web when the entity boundaries are known. The proposed framework will be beneficial for specialized domains where data and class label challenges exist. We demonstrate the benefit of our framework through a case study of Indian classical music forums. Apart from person and location included in standard semantic classes, here we also consider *raga*[1], song, instrument and music concept. Our baseline approach follows a heuristic based method making use of Freebase, a structured web repository. The search engine based approaches acquire context from the web for an entity and perform named entity class identification. This approach shows improvement compared to baseline performance and it is further improved with the hierarchical classification introduced. In summary, our framework is a first-of-its-kind validation of viability of the web for NE class identification.

Keywords. Named Entity Recognition, Named Entity Class Identification, Music Data

1 Introduction

Named entity class (semantic class) identification aims to classify a named entity into one out many semantic classes. In generic domain data, these semantic classes may be person, location, organization, geo-political entity (GPE) etc. This must be distinguished from named entity recognition which involves determination of named entity boundary followed by identification of its class. The input for named entity class identification is a string indicating a named entity, while the output is one among many semantic classes. In other words, our formulation of named entity class identification for a specific domain assumes that the entity boundaries are given. Information extraction tasks like coreference resolution and question answering need the named entity class to be resolved automatically for entities of whose the boundaries are manually annotated or predicted. Certain domains, however, may have specific named entity classes in addition to or excluding some of these classes. For instance for information extraction from biomedical text, domain specific classes like protein, DNA, RNA , cell are introduced [7]. In case of such specific domains, there may be several challenges: (i) There may be insufficient or no annotated data available for training a named entity recognition system (ii) The text may be noisy making it difficult to consider context, or (iii) the named entity classes include domain-specific classes.

In this paper, we present a **framework for named class identification** for a specific domain when the entity boundaries are given. Our framework consists of three steps, and leverages on web as a knowledge repository, in order to perform the target classification. The utility of

[1] Ragas are melodic modes in Indian classical music.

doi: 10.13053/CyS-20-3-2464

our framework lies during setting up a named entity class identification system for new, specific domains. In such cases, information extraction task may require to have specific named entity classes for proper distinction between the entities. Also, in many cases, the context of the named entity whose class needs to be determined, may not be available. It is in such cases that the the capability of search engines and other online knowledge bases to retrieve relevant information for a named entity, is beneficial. motivates the idea of using search engines for gathering documents for identifying the named entity class of an entity. Google has more than 30 trillion web pages indexed [13] making it a rich source of information for any domain.

To demonstrate the utility of our framework, we consider the **domain of Indian classical music forums**. We conduct this study using entities from an online forum on Indian classical music, Rasikas.org [10]. Considering the nature and domain of the text here it is hard to utilize the context of the entities for class identification. We present three approaches to use the web: (a) a baseline, rule-based approach that uses a structured web repository, (b) a supervised approach that uses search engine results and topic models, and (c) a supervised approach that improves upon (b) with task-specific hierarchy of classifiers.

The rest of the paper is organized as follows. Section 2 describes related work on named entity recognition (NER) in general and research directions making use of web resources. Section 3 describes the proposed framework and section 4 briefs on domain specific aspects of Indian classical music forums taken for case study. Section 5 describes the baseline method and the methods using search engine. Section 6 explains the experiments and results on selected entities and Section 7 summarizes our conclusions.

2 Related Work

Existing approaches for named entity recognition (NER) combine entity boundary identification and named entity class identification. There exists quite a large number of supervised learning approaches for NER. Most of these approaches rely on an annotated dataset from similar domain for training the system. To the best of our knowledge the proposed framework is first-of-its-kind on named entity class identification when the entity boundaries are known. SVM based NER discussed in [5] classifies every word in a sentence through features related to the word and the preceding and succeeding words. This system is trained with CRL (Communication Research Laboratory) data prepared for IREX (Information Retrieval and Extraction Exercise, [12]). MUC-6 and MUC-7 dataset served for training in the HMM based approach in [16] which used word features, semantic features and gazetteer based features. CRF based method in [8] used CoNLL-2003 English shared task data for training. All of the above approaches classify the named entities into standard set of named entity classes.

Web is used as a resource in some of the researches for NER and NER related tasks. The unsupervised approach for an NER related task in [4] describes a web based approach to bootstrap for identifying more candidates in particular classes given some seed candidates as input. [14] proposed an approach to perform named entity recognition on entire web through a supervised approach. A bootstrap based method is employed to generate training data from the web for this supervised approach. The unsupervised approach discussed in [9] generates a large gazetteer list from web and this is then used during disambiguating and classifying entities in a given document using simple heuristics, taking context of each entity into account. Similar approaches use web resources like Wikipedia for building an extensive gazetteer list for NER [11]. [6] proposed a method to gather training data from web with the learning examples for each class. The major distinction of our approach with the existing approaches is the utilization of web instead of the context of an entity while finding the named entity class of the entity.

3 Our Framework

Implementation of named entity class identification for new domains can be challenging. We present a generalized framework that uses the web as a resource in order to perform named entity class identification.

Figure 1 presents our three-step framework to set up a system for named entity class identification. The first step is to **understand the domain** of operation. This includes studying the challenges of the domain in terms of availability of datasets, and then determining the class labels. These class labels must be derived from the domain of operation. The second step is to **devise a web-based mechanism** to harness information from the web. Alternatives to do this are: using structured web-based knowledge repositories or using search engines and other retrieval mechanisms in order to extract relevant content. In this paper, we compare two approaches to do this. The goal of this step is to determine the context of an entity whose named entity class must be determined. The third step is to **set up the classification mechanism** to perform the task. Like any typical classification task, this classification may be rule-based or supervised, and may use a combination of other approaches (ensembles, hierarchies, etc.) In this paper, we compare two approaches to perform this classification.

The three steps above indicate how such a system can be set up. In the rest of the paper, we show how our framework can be used for named entity class identification for Indian classical music entities.

4 Understanding the Domain: Indian Classical Music Forums

Rasikas.org is one prominent online forums having discussions on various topics pertaining to Carnatic music. Carnatic music is the south Indian system of Indian classical music. The main topics of discussion in the forum includes raga [1], *tala* (rhythm), *vidwans & vidushis* (musicians), *vaggeyakaras* (composers), *kutcheri* (concert) reviews & recordings, album reviews, etc. A sample forum post is as shown here

```
Sri Ragam is the asampoorna mela
equivalent of K Priya acc to MD's school.
Thyagaraja gave life to K.Priya with
his excellent compos, where as MD never
touched this raga.  In Sri ragam we have
plenty of compos by the trinity incl the
famous Endaro Sri Ranjani is a lovely
janya of K Priya with plenty of compos by
both T & MD.
```

Table 1. Named entity classes and examples

Class	Examples
Person	*Sri Tyagaraja Swami, M. S. Subbulakshmi*
Raga	*Mayamalavagowla, Surabhi*
Song	*dEvAdi dAva sadAshiva, Isha paahimaam*
Instrument	*Veena, Mridangam*
Concept	*Arohana, Janya*

To perform coreference resolution on this dataset, we require that the entities to be classified into domain specific named entity classes viz. person, raga, song, music instrument (hereafter ‘instrument’), music concept (hereafter ‘concept’). Table 1 shows instances of each class from the dataset.

Each forum post is a short discourse text comprising 4-5 sentences average. Forum post have noisy content in the form of a few grammatical errors, less structuring and spelling discrepancies. Spelling discrepancies are found more with named entities where the entities are spelled variably in different posts. For example *’Muthuswami Dikshitar’, ’Dikshithar’* and *’diksitar’* refer to the same person.

The context of occurrences of certain entities have nothing much to tell about the class of the entity. Also, the context can be very similar for classes like song, raga, concept. In the following examples, it is difficult to infer *balahamsa* as a raga and *mysOre vAsudEvacharya* as a person.

```
w.r.t balahamsa, IMO it seems to be
characterised (nowadays at least) mainly
by variations of one prayoha - r/mgs ?
```

```
I heard a recording of the mysOre
vAsudEvacharya krithi "mahAtmulE
teliyalEru" sung by SK Vaagesh (probably
from an AIR program) at a friend's place
few years back.
```

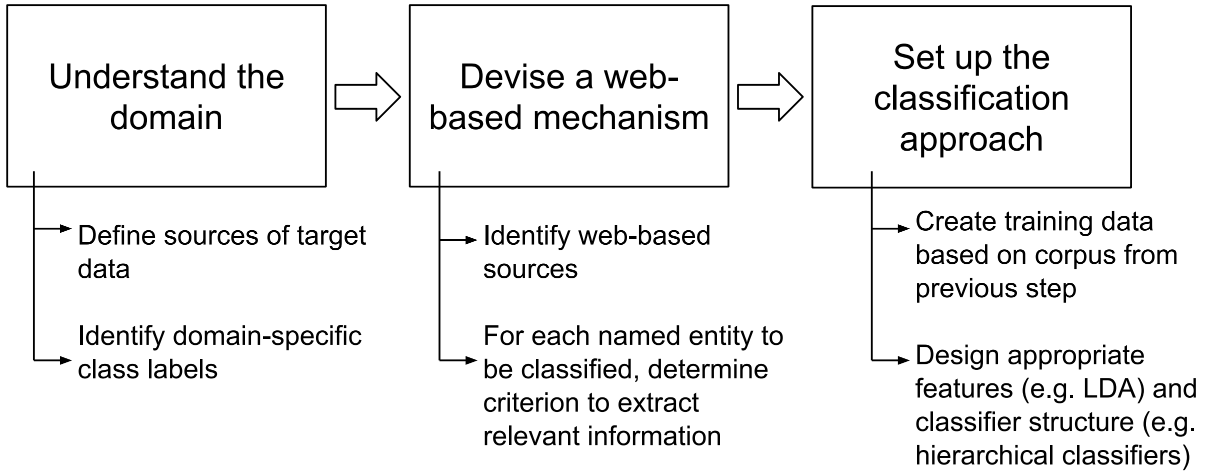
Understand the domain
Define sources of target data
Identify domain-specific class labels
Devise a web-based mechanism
Identify web-based sources
For each named entity to be classified, determine criterion to extract relevant information
Set up the classification approach
Create training data based on corpus from previous step
Design appropriate features (e.g. LDA) and classifier structure (e.g. hierarchical classifiers)

Fig. 1. Our framework to set up a named entity class identification that uses the web, for a new domain

Also, the extensive usage of Indian terms in text makes it harder to infer class from the context. There are instances where an entity appears alone as a separate sentence. This usually happens with composition names, followed by description in the subsequent sentences.

5 Devising a Web-Based Mechanism & Setting Up Classification Mechanism

Web serves as a general knowledge repository, that can be effectively harnessed for the task at hand. This particularly holds true in case of specific domains such as ours, where general-purpose knowledge repositories may not contain the requisite information.

We present three approaches for named entity class idenfitication.

5.1 Baseline: Heuristic-Based Approach That Uses Freebase

Freebase is a vast repository of world knowledge extracted from popular wikis and stored as a database of structured knowledge [3]. It is rich with information from specific domains like Indian classical music. This motivates the first heuristic-based approach for identifying the semantic class with the help of certain information fields in Freebase database.

Algorithm 1 NE class identification through Freebase (approach 1)

```
procedure NE_CLASSID_FREEBASE($entity$)    ▷ NE class of entity
    Get Google suggestions for $entity$
    $sel\_suggestion \leftarrow \operatorname*{argmax}_{gs \in suggestions} sim(entity, gs)$    ▷ gs with highest similarity with entity
    Search $sel\_suggestion$ in Freebase to identify the type of $entity$
    If the Freebase entity does not have type information is_a pattern is searched in Freebase description to know the type return NE class
```

This approach is described in Algorithm 1. We try to minimize spelling discrepancies using *Google Suggest* and among the suggestions, we consider the suggestion with the highest similarity with input entity string for the subsequent steps. The robustness of *Google Suggest* in handling spelling discrepancies is put to use here to obtain better search results. Jaro-Winkler distance [15], a type of string edit distance is employed to get the similarity, capable of giving more importance to the initial part of the entity words. The differences at the start of the string are more significant than the ones towards the end. The entity strings from the forum are less likely to have spelling discrepancies at the initial part.

The selected string from *Google Suggest* suggestions is searched in the Freebase. The type (/common/topic/notable types) of the selected entity obtained from Freebase is taken as the semantic class of the input entity. This type is mapped to one of named entity classes defined for this domain, except a few which are considered as 'other'. If type is not present for an entity, we search for 'is a' pattern in the description available with Freebase. Mostly raga entities are identified through this pattern in the description.

5.2 Supervised Classification Based on Web Search

This method relies on documents returned by the search engine for identifying the semantic class of an entity. Given the fact that Bing/Google has a large number of indexed pages, the chances of getting relevant pages for an entity even from a narrow domain is quite high. A classifier is pre-trained for classifying documents returned by the search engine to the relevant NE class. Algorithm 2 describes the procedure for training the model and classification of an entity string.

Algorithm 2 NE class identification through web search (approach 2)

```
procedure NE_CLASSID_WEB_TRAIN
    Get documents for each NE class
    Get word clusters using LDA-based topic models from all the documents
    Train the bag of words classifier with the document set for each NE class
return Learned model

procedure NE_CLASSID_WEB_TEST($entity$, $learned\_model$)
    Get top k web search results for $entity$ string
    Get the web content of the top k results
    Combine the retrieved content into single document and then classify with the $learned\_model$
return NE class
```

The classifier is trained with handpicked documents for each NE class. The documents for person, raga and instrument classes come mostly from Wikipedia[2], whereas song and music concept related documents are from other sources. The classifier uses bag-of-words model for document classification. We use probabilistic models based on LDA [2] to discover clusters of words called topics. These topics represent themes underlying in the dataset. To avoid named entities getting into the bag of words, all the proper nouns in the text are masked before applying LDA.

The NE identification procedure gets the top k web search results for the searched entity string. The main web content of these k results are extracted[3]. The content extracted from these websites are merged to form a single document. This document is classified as one of the NE classes with the pre-trained model (output of NE_classid_web_train).

5.3 Hierarchical Hybrid Classification Based on Web Search

As the third approach, we consider a hierarchical classification approach. In this case, we segregate classification of concepts and songs using a rule-based method. In this approach, the learned supervised classifier will classify only the entities which are not classified by the rule-based classifiers for song and concept. The method is depicted in Figure 2. An input entity string is given to song classification module to identify the entity as song or not. A few different heuristics are tried for song classification. One method checks for if the majority of the web search results are links to music websites. The exhaustive list of 143 music websites is used to check for if a returned link is a music website or not. A simplified version of this method is tried to check if the first link returned by Google search is a music website or not.

The entities which are classified as not song by the song classifier are passed on to the concept classifier. The concept classifier follows a gazetteer based approach with a gazetteer

[2] https://en.wikipedia.org

[3] Python library "Boiler Pipe" is used for main html content extraction.

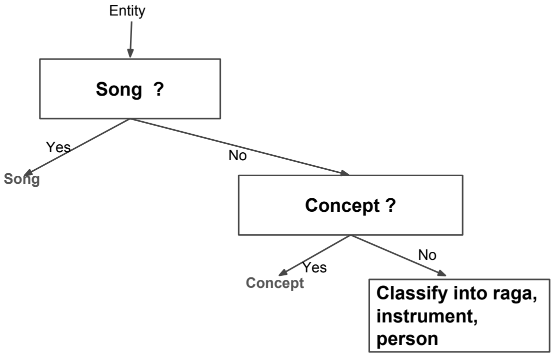

Fig. 2. Separation of concept and song classification from the rest using hierarchical classification.

covering most of the concepts in Carnatic music. The entities having a Jaro-Winkler distance based similarity above a defined threshold value are classified as concepts. The entities which are not classified as concepts are passed on to the document classifier for getting classified as one among person, raga and instrument classes.

6 Experiments and Results

6.1 Experiment Setup

We consider 5-class classification for our experiments. Our classes are: person, raga, song, instrument, and concept. Since there is only a few occurrences of location class instances, we do not consider location as a label.

6.2 Comparison of Methods

Table 2 shows the performance of the baseline heuristic-based method. Out of 619 test entities, this method assigns no semantic class to 254 entities. The reported result takes into account only the entities classified by the method. Considering the classified entities, the overall precision is 0.77, recall 0.43 and F-score 0.55.

In addition, the confusion matrix for this method is shown in Table 3. We see that even among the entities for which NE class is predicted, the mis-classification is high. Concept instances are getting mis-classified always since a meaningful type could not be seen in Freebase corresponding to any concept instance. The 'other' class mentioned in the confusion matrix includes the instances which are classified to types (ex. film) which cannot be mapped to the defined classes.

Table 2. Results of Freebase based identification

Class	Precision	Recall	F1	Support
Concept	1.00	0.01	0.02	87
Instrument	0.83	1.00	0.91	5
Person	0.72	0.77	0.74	118
Raga	0.81	0.39	0.52	124
Song	0.19	0.42	0.26	31
avg / total	**0.77**	**0.43**	**0.55**	**365**

Table 3. Confusion matrix: Freebase based identification; c: concept, i: instrument, p: person, r: raga, s: song

	c	i	p	r	s	o
c	1	36	17	11	22	0
i	0	0	0	0	0	5
p	0	21	91	0	5	1
r	0	32	15	48	29	0
s	0	14	4	0	13	0

Table 4. Training documents

Class	#Documents	#Words
Person	150	121777
Raga	141	53969
Song	102	84722
Instrument	121	51747
Concept	50	55440

The supervised approach described in Algorithm 2 depends on a pre-trained model for classifying an entity. Table 4 describes the training documents selected for each class to train the pre-trained model. Documents for person, raga, instrument and concept are mostly taken from Wikipedia whereas documents for song class are handpicked from other websites. While searching the word 'Carnatic' (a sub-genre of Indian classical music

Table 5. Results of web search based identification

	Experiment A			Experiment B			
Class	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
Concept	0.75	0.05	0.09	0.75	0.05	0.09	182
Instrument	0.86	0.86	0.86	0.86	0.86	0.86	7
Person	0.64	0.92	0.75	0.71	0.90	0.80	145
Raga	0.52	0.94	0.67	0.54	0.92	0.68	211
Song	0.62	0.11	0.18	0.86	0.65	0.74	74
avg / total	**0.63**	**0.57**	**0.60**	**0.68**	**0.63**	**0.65**	**619**

in case of our dataset) is appended to the search string for improved disambiguation. For these experiments, the top-5 (k=5) web search results are taken for an entity searched. Two popular search engines *Google* and *Bing* are used for searching the entity string. The results of the method is given in Experiment A of Table 5.

The drop in accuracy is majorly due to the confusions with concept and song classes. The confusion matrix in Table 6 shows that the music concept and song classes are getting heavily confused with person and raga classes. The accuracy for person, raga and instrument classes is high compared to concept and song classes. Large overlap of the words in webpages related to music concept and raga classes is the primary reason for the confusions between them. The songs are getting mostly classified as person and raga. The web content available on searching a song name are media links or pages having lyrics, notation and other information related to the song which is not helpful in classification to song class. The content of these websites have raga and singer information as well tending to classify them as raga or person.

Experiment B in Table 5 shows the results of the third approach that uses hierarchical classification. Though the concept and song instances which are missed by the respective classifiers lead to low accuracy, the results are better compared to approach 2. Confusion matrix in 7 shows that, though the mis-classifications of concept class instances remain almost the same, the improvement in song classification leads to better overall accuracy.

Table 6. Confusion matrix: web search based identification (Experiment A); c: concept, i: instrument, p: person, r: raga, s: song

	c	i	p	r	s
c	9	1	34	136	2
i	0	6	1	0	0
p	0	0	133	11	1
r	1	0	10	198	2
s	2	0	31	33	8

Table 7. Confusion matrix: web search based identification (Experiment B); c: concept, i: instrument, p: person, r: raga, s: song

	c	i	p	r	s
c	9	1	31	141	0
i	0	6	1	0	0
p	1	0	131	10	3
r	1	0	10	195	5
s	1	0	11	14	48

6.3 Error Analysis

In this section, we analyze the errors of our method that uses hierarchical classification.

Person: Few person instances are confused with raga. For example singer ‘Ilayaraja’ when searched after appending word ‘carnatic’ returns irrelevant pages having only a few information about this composer. This may be happening because Ilayaraja has more contributions to Indian popular music compared to Carnatic. In the case of singer 'Rajalakshmi', pages having her songs get retrieved with a fair occurrences of the term 'raga'.

Raga: Many raga names are confused with person and song. The raga names having ambiguity with person names or other entities are likely to get classified as person. 'Snehapriya', 'K Priya' (short form of Karaharapriya), 'Ranjani' which are likely to be confused with Indian person names are classified as person. The search results of certain raga names return mostly links to music websites causing the song classification to classify them as song.

Song: Song names not meeting the song classification criteria tend to get categorized as one of the other classes. Song names like 'Bhairavi krithi', 'Thyagaraja krithis' having a raga name or a person name as a part of it are not classified as song. Song names for which search engine return websites with lyrics are also not classified as song.

Concept: The gazetteer based approach fails to identify many concepts which are combination of other concepts as in 'madhyama sruti', 'shuddha rishabam', 'raga alapan'. There also exists many Indian terms related to music but not music concepts like 'shishya', 'bhakti rasa', 'Kelvi gnanam' marked as concepts in the ground truth. These terms are not classified as concepts. Also, the absence of many concepts in the gazetteer is a reason for poor performance.

7 Conclusion & Future Work

This work deals with named entity class identification in novel domains. Such domains may be challenging due to lack of data, or presence of specific class labels. We presented a framework to perform this task of identification. Our framework helps to setup a platform for named entity class identification of entities making use of web resources, ignoring the context of the entities. Our methods utilizing the popular search engines to procure context are compared against the baseline approach with Freebase. The domain specificities pertaining to named entity classes are major determinants in designing the hierarchical classification model. From our case study with Indian classical music forums, it is evident that a system design driven by domain understanding is helpful. Compared to baseline approach based on Freebase, search engine based approach yields better accuracy. The segregation of certain classes through hierarchical classification further improved the accuracy.

The method which extracts web content needs improvement to diligently filter the text to contain meaningful content related to the searched entity. Usage of nuanced LDA-based topic models will help us to identify better word clusters, in the future.

References

1. **Bhagyalekshmy, S. (1990).** *Ragas in Carnatic music*. South Asia Books.
2. **Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003).** Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022.
3. **Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008).** Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, pp. 1247–1250.
4. **Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005).** Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, Vol. 165, No. 1, pp. 91–134.
5. **Isozaki, H. & Kazawa, H. (2002).** Efficient support vector classifiers for named entity recognition. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 1–7.
6. **Karaa, W. B. A. (2011).** Named entity recognition using web document corpus. *arXiv preprint arXiv:1102.5728*.
7. **Kazama, J., Makino, T., Ohta, Y., & Tsujii, J. (2002).** Tuning support vector machines for biomedical named entity recognition. *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, Association for Computational Linguistics, pp. 1–8.
8. **McCallum, A. & Li, W. (2003).** Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*,

Association for Computational Linguistics, pp. 188–191.

9. **Nadeau, D., Turney, P., & Matwin, S. (2006).** Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity.
10. **rasikas (2005).** Rasikas.org.
11. **Ratinov, L. & Roth, D. (2009).** Design challenges and misconceptions in named entity recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 147–155.
12. **Sekine, S. & Eriguchi, Y. (2000).** Japanese named entity extraction evaluation: analysis of results. *Proceedings of the 18th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, pp. 1106–1110.
13. **statisticbrain (2016).** statisticbrain.com.
14. **Whitelaw, C., Kehlenbeck, A., Petrovic, N., & Ungar, L. (2008).** Web-scale named entity recognition. *Proceedings of the 17th ACM conference on Information and knowledge management*, ACM, pp. 123–132.
15. **Winkler, W. E. (1999).** The state of record linkage and current research problems. *Statistical Research Division, US Census Bureau*, Citeseer.
16. **Zhou, G. & Su, J. (2002).** Named entity recognition using an hmm-based chunk tagger. *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 473–480.

Joe Cheri Ross is a PhD Student in the Department of Computer Science and Engineering. His primary area of research is music information retrieval. His current focus is on extracting information from music related text using natural language processing methods.

Aditya Joshi is a PhD student at IITB-Monash Research Academy, a joint PhD program between Indian Institute of Technology Bombay, India and Monash University, Australia. His primary area of research is sentiment analysis.

Pushpak Bhattacharyya is Vijay and Seeta Vashee Chair Professor at Indian Institute of Technology Bombay, and also the Director of Indian Institute of Technology Patna. With a research experience of over 25 years, he has conducted innovative research in several disciplines of NLP. He has also authored a book titled ‘Machine Translation’.

Article received on 15/01/2016; accepted on 17/03/2016.
Corresponding author is Joe Cheri Ross.