Generation of Bilingual Dictionaries using Structural Properties

Ajay Dubey and Vasudeva Varma

Search and Information Extraction Laboratory,
International Institute of Information Technology,
Hyderabad, AP, 500032, India

ajay.dubey@research.iiit.ac.in, vv@iiit.ac.in

Abstract. Building bilingual dictionaries from Wikipedia has been extensively studied in the area of computation linguistics. These dictionaries play a crucial role in Natural Language Processing(NLP) applications like Cross-Lingual Information Retrieval, Machine Translation and Named Entity Recognition. To build these dictionaries, most of the existing approaches use information present in Wikipedia titles, info-boxes and categories. Interestingly, not many use the structural properties of a document like sections, subsections, etc. In this work we exploit the structural properties of documents to build a bilingual English-Hindi dictionary. The main intuition behind this approach is that documents in different languages discussing the same topic are likely to have similar structural elements. Though we present our experiments only for Hindi, our approach is language independent and can be easily extended to other languages. The major contribution of our work is that the dictionary contains translation and transliteration of words which include Named Entities to a large extent. We evaluate our dictionary using manually computed precision. We generated a massive list of 72k tokens using our approach with 0.75 precision.

Keywords. Bilingual dictionary, comparable corpora, structural elements.

Generación de diccionarios bilingües usando las propiedades estructurales

Resumen. Compilación de diccionarios bilingües usando Wikipedia ha sido estudiada mucho en la lingüística computacional. Estos diccionarios juegan un papel crítico en tales aplicaciones del procesamiento de lenguaje natural (PLN) como recuperación de información inter-lingüística, traducción automática y reconocimiento de nombres. La mayoría de los enfoques existentes para la construcción de estos diccionarios usa la información presente en títulos de Wikipedia, info-boxes y categorías. Es interesante que pocos investigadores hayan usado las propiedades estructurales de documentos tales como secciones, sub-secciones, etc. Este trabajo utiliza las propiedades estructurales de documentos para construir un diccionario bilingüe inglés-hindi. La intuición principal en la cual se basa este enfoque es el hecho de que la discusión de un cierto tema en idiomas diferentes puede tener los elementes estructurales similares. Los experimentos se realizaron sólo para hindi, pero el enfoque no depende del idioma particular y puede ser extendida fácilmente a otros idiomas. La mayor aportación de este trabajo es la inclusión en el diccionario las palabras que son nombres traducidos y transliterados. El diccionario fue evaluado mediante la precisión calculada manualmente. Se generó una lista muy grande de 72K tokens usando el enfoque propuesto con la precisión de 0.75.

Palabras clave. Diccionario bilingüe, corpus comparable, elementos estructurales.

1 Introduction

Multilingual content over the internet is increasing at a rapid pace. Also, comparable corpora, i.e. documents in different languages talking about the same topic, are the main focus of the increasing internet. We believe that language should not become a hindrance in meeting a user's information need. Cross-Language Information Access(CLIA) systems play an important role to overcome this barrier up to a certain extent. Bilingual dictionaries are the backbone of most CLIA and NLP applications like Machine Translation, Cross-Language Information Retrieval and Named Entity Recognition. Language tools with high accuracy are available in abundance for resource rich languages. Unfortunately, not many tools with considerable accuracy exists for resource poor languages. The scarcity of such tools motivated us to strive for a language independent approach.

Currently, most of the approaches rely heavily on Wikipedia, we here propose an approach which can be applied to any bilingual comparable corpora. Interestingly, none of the published

works have looked into the basic structure of documents/articles. Now onwards the terms article and document will be used interchangeably. Article[1] is defined as a written work published in a print or electronic medium. Articles can be classified into academic papers, news articles, blogs, essays and speeches. All types of article are generally further divided into structures. These structures can be sections, subsections, introduction and different details; for example, an e-mail document structure would be addressee, sender, subject, greetings, signature and timestamps. Another example is the structure of an academic paper which includes abstract, introduction, related work, approach, experiments, results, conclusion, and references. Irrespective of language an article about a certain topic will have similar sections discussing the topic in different languages. Most important contribution of this paper is to find similar sections of documents across languages. These section headings are good candidates for being entries to a bilingual dictionary. Now onwards the terms section headings and subheadings will be used interchangeably. Searching for corresponding bilingual tokens in these sections is more logical rather than finding them in the entire article.

Most dictionaries do not contain Named Entities, which play an important role in Topic Modeling and many NLP applications specially in Machine Translation. In this paper we try to capture Named Entities using transliteration. The rest of the paper is organized as follows: Section 2 talks about the related work. Section 3 describes our approach. Dataset used for our experiment is explained in Section 4, followed by the results in Section 5. Finally we conclude our paper in Section 6.

2 Related Work

Earlier work on creation of bilingual dictionaries can be broadly classified into two types, manually and automatic. In the following section we discuss both of them briefly.

[1] Article http://en.wikipedia.org/wiki/Article_(publishing)

2.1 Manually Constructed Dictionaries

Bilingual dictionaries are necessary in understanding different languages. This necessity has been observed from the very ancient era. The earliest work of creation of bilingual dictionaries was found in the form of clay tablets from about the 2300s BC[2]. It consisted of words in Sumerian language and their equivalents in the Akkadian language. Even the earliest modern European dictionaries were bilingual in nature.

Noah Webster, an American lexicographer, published an expanded dictionary[3] in 1825 containing seventy thousand words in many European languages like Old-English, German, Greek, Latin, Italian, Spanish, French, and also in Hebrew, Arabic and Sanskrit. It took 25 years for an expert to compile such a dictionary. Dictionary creation is a costly, complex, and time-consuming task which requires a lot of human effort. Hence instead of using linguists crowd sourcing techniques are used to build such dictionaries where users are generally second language learners.

The EDICT project started in 1991 by Jim Breen aimed at Japanese-English dictionary, to provide assistance in reading Japanese text. Because of its inadequate structure mentioned by Breen et al. [3] the project channeled into JMdict. Since then, JMdict dictionary has been extended by a lot of people. It comprises a lot of tokens including large number of domain specific terms. Even with the help of such a huge community of people and over a period of a decade, the dictionary does not exhaustively contain local domain-specific and latest terms.

Shabdanjali[4], a Hindi-English dictionary is yet another example of manually constructed dictionary which started in May 1999. It was developed and is being continuously upgraded through a voluntary collaborative effort. Total number of entries in its latest version reached slightly above 25,000 words.

Manual dictionaries require a lot of effort, time and money and yet they are not exhaustive due to addition of new words everyday. Hence the

[2] Dictionary as Mentioned in WebCite http://www.webcitation.org/5kwbLyr75

[3] American Dictionaries in Wikipedia http://en.wikipedia.org/wiki/Dictionary

[4] http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Shabdanjali/dict-README.html

methods which generate dictionaries from parallel or comparable corpora automatically are important.

2.2 Automatically Constructed Dictionaries

Automatic Construction of dictionaries can be further classified into approaches that use parallel corpora and others that use comparable corpora. Parallel corpora contain translation of documents from one language to another. On the other hand comparable corpora contain documents in different languages having similar texts.

Statistical models proposed by Brown et al. [4] and Kay et al. [9] can be used for building dictionaries. These models require availability of huge bilingual corpora which are difficult to find for under-resourced languages. Melamed et al. [11] and Och et al. [14] also used parallel corpora to align text. Parallel corpora generate good results for high frequency terms but accuracy decreases tremendously for low frequency words. In parallel corpora large amount of text is added and omitted to make sense of a text clear and leads to poor accuracy which is verified by Fung et al. [8]. Also, such large parallel corpora are not available in many languages.

A lot of work has been done on generation of dictionaries using Wikipedia as a comparable corpora. Tyers et al. [19] used a seed list and inter language links present in the Wikipedia structure. They captured titles of documents by inter language links for the words present in their seed list of English language. Bilingual dictionary was extracted from Wikipedia by Maike Erdmann et al. [5] using inter language links, redirect pages and anchor text. They achieved good results but their approach required to assign weights manually to each category. In their later paper [6], they worked on removing noisy data and false entries from their dictionary by applying a classifier. Bharadwaj et al. [16] proposed an iterative approach to extract English-Hindi dictionary from Wikipedia. They iteratively mine text from Wikipedia titles, info-boxes, categories and first paragraphs of Wikipedia documents to build dictionary. Another approach based on Wikipedia was proposed by Rahimi et al. [15] by aligning Wikipedia titles only. The work divided multi-word title alignments to shorter aligned phrases, to build word association English-Persian dictionary. The focus of this approach was to enhance the performance of CLIA system instead of coverage.

Quite a lot of work is also done on finding similar/parallel sentences in comparable corpora. Adafre et al. [1] used a seed lexicon to find similar sentences in comparable corpora. This approach was later modified by Mohammadi et al. [12] using N-gram of sentences and various similarity measures. A binary classifier was built to determine whether two sentence of different languages are similar or not by Smith et al. [17]. They used multiple features like log-probability of alignment, number of aligned words, longest sequence of aligned words and number of words having multiple meaning. Since they also used orthographic features and edit distance measures, their approach is too specific for closely originated languages only. A language independent method of extraction of parallel sentences from Wikipedia was proposed by Bharadwaj et al. [2].

There are some graph theory based approaches too for creating bilingual dictionaries. University of Washington along with Google Seattle build Massive, Multilingual Dictionary using Probabilistic Inference popularly known as PAN-Dictionary [18]. This dictionary is a sense-disambiguated lexical translation resource. Wiktionary was used to build this dictionary which has 80 thousand senses. Laws et al. [10] generated a cross-lingual thesaurus using graph similarity measures and SimRank algorithm on two graphs for different languages.

There are approaches that use language specific resources as well. Fatiha et al. [7] used POS tagged based context vector approach for calculating similarity between two words of source and target language. This approach performs poorly for resource-poor languages as it is dependent on accuracy of POS Tagger.

3 Proposed Approach

We propose an approach to create a bilingual dictionary from comparable corpora and not limited to Wikipedia. The approach is language independent and can be applied to other languages as well. In this paper we will discuss generation of English-Hindi Bilingual Dictionary. The approach is focused on those languages which are phonetically rich like Indian languages. Our dictionary contains

transliteration and translation of words including Named Entities across languages.

3.1 Generating Named Entity Dictionary

Since Named Entities are not found in regular dictionaries but are very crucial in CLIA systems, we first try to capture Named Entities (NEs). We used Stanford Named Entity Recognizer to extract NEs in English language. We search phrases in Hindi language, which are transliteration of NEs extracted from English documents. Since Hindi is a phonetically rich language (i.e. every character in Hindi can be mapped to English based on its sound), Hindi text is transliterated to English. We build the transliteration module with the help of Nayan et al. [13] and Editex algorithm as mentioned in Zobel et al. [20]. The algorithm for generating NEs dictionary is presented as Algorithm 1.

3.2 Generation of the Title Dictionary

Comparable corpora contain documents in multiple language talking about the same topic but they may be written by completely different authors. Hence their content will not be exact translations but similar text. Titles of such documents are accurate candidates of dictionaries. With the help of the transliteration module, we applied a modified approach of Rahimi et al. [15] to capture word level association (mappings). Our algorithm is a two-pass algorithm over entire corpus title pairs presented as Algorithm 2.

3.3 Finding Similar Sections across Languages

In this section we find similar sections of documents across languages. To do this we try to find mappings of headings of sections across languages. We observed that not all documents have same subheadings. Documents talk about different topics like countries, actors, players, etc., hence section headings of corresponding documents are entirely different. Therefore we need to cluster documents to find similar sections and their corresponding subheadings in documents across languages. Clustering is done only on English documents. Here we make an assumption that clusters formed in English documents will be similar to clusters of Hindi documents.

This assumption is based on the intuition that documents across languages representing the same theme/topic might end up forming similar clusters. In this paper, clustering of documents is based on Wikipedia categories, meaning documents under the same category are assigned to the same cluster. In absence of such categories, documents can be clustered using subheadings and other features.

After forming the clusters we build a subheading mapping for every cluster across languages to find the similar word/keywords inside these sections. These subheading mappings are good candidates for dictionary entries. There were many challenges encountered while building this mapping. The content of Hindi document is less than that of English documents. Hence several sections in Hindi documents are either missing or clubbed up to represent multiple sections of English documents. Some domain specific documents contain extra information in Hindi documents as well. In some documents it was observed that sections of correspondent documents across languages have a different order. So these mappings are generated by Algorithm 3 using the following formula:

$$score_{(en_i - hi_j)} = log_2\{2 - \frac{|pos_{en_i} - pos_{hi_j}| + m}{2 * m}\} \quad (1)$$

where

- $pos_{en_i} \equiv$ position of en_i subheading in English document subheading list.
- $pos_{hi_j} \equiv$ position of hi_j subheading in Hindi document subheading list.
- $m \equiv$ maximum of lengths of subheading list in English document and Hindi document.

3.4 Similar Sentences and Co-occurring Words

Finding bilingual lexicons/sentences in text become easier as text tends to become rather parallel than comparable. Adafre et al. [1] found parallel sentences in documents using a bilingual lexicon. This approach can perform well only if applied to appropriate context. They tried to find parallel sentences across entire documents. In this paper we try to find them in similar sections. We split the content of sections into sentences. Since the content in Hindi document is less, we try to find the most similar English sentence by computing

Algorithm 1 : Generation of the English-Hindi NE mappings

```
for all do en-doc ∈ english-corpus
    en_NE_List ← Stanford-NER(en-doc)
    hi-doc ← corresponding hindi-document in hindi-corpus
    for all do hi-word ∈ hi-doc
        transliterated-word ← phonetic-Transliteration(hi-word)
        for all do en-NE-word ∈ en_NE_list
            if then Editex(transliterated-word, en-NE-word) < phonetic-threshold
                ne_Mapping_List ← ne_Mapping_List ∪ (en-NE-word, hi-word)
            end if
        end for
    end for
end for
```

Algorithm 2 : Generation of the Title Dictionary

```
for all do en-doc ∈ english-corpus
    en-title ← title of en-doc
    hi-doc ← corresponding hindi-document in hindi-corpus
    hi-title ← title of hi-doc
    title_Dictionary ← title_Dictionary ∪ (en-title, hi-title)
    if then en-title and hi-title contain multi words
        en_Words ← Split(en-title)
        hi_Words ← Split(hi-title)
        remove already known mappings from en_Words and hi_Words
        if then Only one pair of words are left in en_Words and hi_Words
            title_Dictionary ← title_Dictionary ∪ (en_Words, hi_Words)
        else
            for all do en-word ∈ en_Words
                for all do hi-word ∈ hi_Words
                    transliterated-word ← phonetic-Transliteration(hi-word)
                    if then Editex(transliterated-word, en-word) < phonetic-threshold
                        title_Dictionary ← title_Dictionary ∪ (en-word, hi-word)
                    end if
                end for
            end for
        end if
    end if
end for
```

Algorithm 3 : Generate Subheading mappings for every cluster

```
for all do en-doc ∈ cluster
    hi-doc ← corresponding hindi-document in hindi-corpus
    for all do (en-subheading, hi-subheading)
        if then (en-subheading, hi-subheading) present in dictionary
            subheading_Mapping ← subheading_Mapping ∪ (en-subheading, hi-subheading)
        end if
        scoreMap[(en-subheading, hi-subheading)] += relativePositionScore(en-subheading, hi-subheading)
    end for
end for
while s docoreMap is not empty
    (en-subheading, hi-subheading) ← maximumScoreEntry(scoreMap)
    subheading_Mapping ← subheading_Mapping ∪ (en-subheading, hi-subheading)
    remove all other entries from scoreMap continaing en-subheading or hi-subheading
end while
```

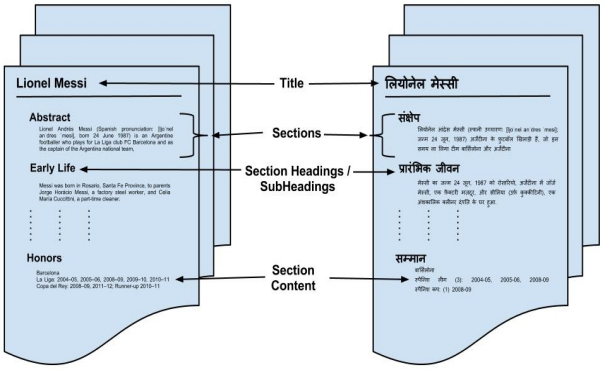

Fig. 1. Comparable documents gathered from Wikipedia

the number of similar words present in them. After getting these partially parallel sentences, we remove stop-words from them. From the remaining sentence we add all the co-occurring word pairs to a map. Iteratively, most frequent co-occurring word pair is added to our dictionary and the rest of word pairs containing either part of most co-occurring word pair are discarded. We also keep a tab on minimum co-occurrence of word pair to be added to the dictionary.

4 Dataset

For dataset, we use English and Hindi language Wikipedia dumps. We enlisted all the titles for which documents are present in both languages with the help of interwiki links. These documents were crawled, parsed, cleaned and stored as files in local disk. There were 21,384 pairs of documents in total. Each document contains title, section heading and content in that section as shown in Figure 1. Since we do not use any Wikipedia specific properties, our approach can be applied to other corpus as well which can be converted into this format.

5 Results

Evaluation of the dictionary formed by the proposed approach is done using the metrics of precision. Precision(P) is a parameter to judge accuracy. It is the ratio of total number of correctly(N) mapped word pairs to total(T) number of mappings in the dictionary:

$$P = \frac{N}{T} \tag{2}$$

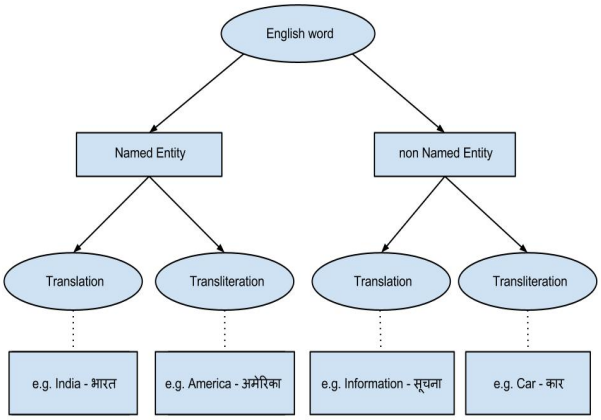

Fig. 2. Types of bilingual words

Our dictionary contains all four types of bilingual word pairs which are NE translations, NE transliterations, non-NE translations and non-NE transliterations, see Figure 2 for more details. In Table 1, we report manually evaluated precision of tokens collected from different phases of the proposed approach. A group of 3 native language speakers were assigned the task of evaluating samples of dictionary. The kappa score for inter annotator agreement between the annotators was found to be 0.76.

There are many reasons for not performing automatic evaluations. Firstly, Named Entities are not present in dictionaries, which are major part of our dictionary. Secondly, freely available dictionaries on web containing non-Named Entities present a lot of challenges like spelling variations, different morphological variations. Spelling variations of word “information” and “beautiful” is shown in Figure 3. Colour and color are also an example of different spelling variations.

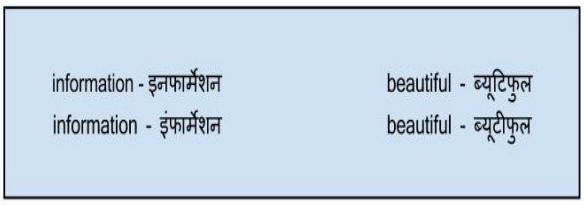

Fig. 3. Spelling variations

Table 1. Precision scores of various phases of dictionary

Phase (gathering)	Tokens	Precision
Named Entities	32500	0.74
Title Dictionaries	26335	0.89
Subheading Mapping	1362	0.86
Co-Occurring Words	10288	0.56
Overall	72220	0.75

6 Conclusion and Future Work

In this paper we used comparable corpora to generate a bilingual dictionary. To do this we first built a small dictionary using word association methods and transliteration. Using this small dictionary we found similar sections of documents across languages. Partially parallel sentences were extracted from these similar sections. Most co-occurring words in these extracted parallel sentences were added to the dictionary.

The overall construction process of dictionary from comparable corpora is automatic. Our focus has been low-resourced languages and requires very minimal information about languages for building the transliteration system only. Since our system is language independent, it can be extended to other languages as well. The approach can also be applied to other bilingual comparable corpora like multilingual news or magazine corpus and is not limited to Wikipedia only.

In future, we want to further improve our method, by understanding features for mapping words in parallel sentences. We are planning to conduct experiments on other languages as well. In addition to it, we observed that there is no proper mechanism to evaluate such dictionaries. As a manual evaluation of such a dictionary is a costly and tedious job, so in future we will try to come up with approaches to evaluate these dictionaries automatically with less human intervention.

References

1. **Adafre, S. & de Rijke, M.** (**2006**). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 62–69.
2. **Bharadwaj, R. & Varma, V.** (**2011**). Language independent identification of parallel sentences using wikipedia. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 11–12.
3. **Breen, J.** (**2004**). Jmdict: a japanese-multilingual dictionary. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Association for Computational Linguistics, 71–79.
4. **Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., & Roossin, P.** (**1990**). A statistical approach to machine translation. *Computational linguistics*, 16(2), 79–85.
5. **Erdmann, M., Nakayama, K., Hara, T., & Nishio, S.** (**2008**). An approach for extracting bilingual terminology from wikipedia. In *Database Systems for Advanced Applications*. Springer, 380–392.
6. **Erdmann, M., Nakayama, K., Hara, T., & Nishio, S.** (**2009**). Improving the extraction of bilingual terminology from wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(4), 31.
7. **Fatiha, S.** (**2011**). Extracting the multilingual terminology from a web-based encyclopedia. In *Research Challenges in Information Science (RCIS), 2011 Fifth International Conference on*. IEEE, 1–5.
8. **Fung, P. & McKeown, K.** (**1997**). A technical word-and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1), 53–87.
9. **Kay, M. & Röscheisen, M.** (**1993**). Text-translation alignment. *computational Linguistics*, 19(1), 121–142.
10. **Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., & Schütze, H.** (**2010**). A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 614–622.
11. **Melamed, I.** (**1997**). A word-to-word model of translational equivalence. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 490–497.
12. **Mohammadi, M. & GhasemAghaee, N.** (**2010**). Building bilingual parallel corpora based on wikipedia. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volume 2. IEEE, 264–268.
13. **Nayan, A., Rao, B., Singh, P., Sanyal, S., & Sanyal, R.** (**2008**). Named entity recognition for indian languages. *NER for South and South East Asian Languages*, 97.

14. **Och, F. & Ney, H.** (**2003**). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19–51.

15. **Rahimi, Z. & Shakery, A.** (**2010**). Creating a wikipedia-based persian-english word association dictionary. In *Telecommunications (IST), 2010 5th International Symposium on*. IEEE, 562–567.

16. **Rohit Bharadwaj, G., Tandon, N., & Varma, V.** (**2009**). An iterative approach to extract dictionaries from wikipedia for under-resourced languages. *Proc. ICON2010*.

17. **Smith, J., Quirk, C., & Toutanova, K.** (**2010**). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 403–411.

18. **Soderland, S., Etzioni, O., Weld, D., Skinner, M., Bilmes, J., et al.** (**2009**). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 262–270.

19. **Tyers, F. & Pienaar, J.** (**2008**). Extracting bilingual word pairs from wikipedia. *Collaboration: interoperability between people in the creation of language resources for less-resourced languages*, 19.

20. **Zobel, J. & Dart, P.** (**1996**). Phonetic string matching: Lessons from information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 166–172.

Ajay Dubey is currently a dual degree B.Tech and M.S Research student in the department of Computer Science and Engineering at International Institute of Information Technology, Hyderabad, India. Since July 2010, he has been working as a research scholar at Search and Information Extraction Lab at IIIT Hyderabad. His primary research interests are in Multilingual Lexical Resource Generation, Focused Crawling, Sentiment Analysis and Information Retrieval and Extraction.

Vasudeva Varma is a professor at International Institute of Information Technology, Hyderabad since 2002. He is heading Search and Information Extraction Lab and Software Engineering Research Lab at IIIT Hyderabad. His research interests include search (information retrieval), information extraction, information access, knowledge management, cloud computing and software engineering. He published a book on Software Architecture (Pearson Education) and over seventy technical papers in journals and conferences. In 2004, he obtained young scientist award and grant from Department of Science and Technology, Government of India, for his proposal on personalized search engines. In 2007, he was given Research Faculty Award by AOL Labs.

Article received on 06/12/2012; accepted on 11/01/2013.