ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

Analysis of Phonetic Matching Approaches for Indic Languages

Sandeep Chaware[1], Srikantha Rao[2]

[1]Research Scholar, MPSTME, NMIMS University, Mumbai, INDIA
[2]PhD Supervisor, MPSTME, NMIMS University, Mumbai, INDIA
smchaware[1]@gmail.com, dr_s_rao[2]@yahoo.com

ABSTRACT- ***Phonetic matching plays an important role in multilingual information retrieval, where data is manipulated in multiple languages. User needs information in their local language which may be different from the language where data has been maintained. In such an environment, we need a system which matches the strings phonetically irrespective of errors either exactly or approximately. There are many errors or variations can be considered but here we had considered typographical errors, spelling errors as differ in vowel and matching of compound words. There are many approaches has been proposed like soundex, q-gram, phoenix etc., but they may produce an ambiguity in matching or may not be applicable to Indian languages. In this paper, we proposed approaches which match the strings either in Hindi or Marathi accurately. We evaluated the three approaches namely Soundex, Q-gram and Indic-Phonetic by generating cases like length-of-string (LOS), differ in vowel and compound words for Hindi and Marathi. We found that Indic-Phonetic approach is an efficient and accurate as compared to other two approaches.***

Keywords- **Soundex, Q-gram, Indic-phonetic, threshold, phonetic matching.**

I. INTRODUCTION

Phonetic matching is needed when many diversified people come together. They either speak with different pronunciation styles or write many languages in various writing styles, but their meaning is same. It deals with the similarity of two or more strings by pronunciation regardless of their actual spelling. A typical example is Just-Dial service, where a telephone operator is given a name for finding out the information. The operator guesses the possible spelling of it and search from the database for approximate or exact result (or spelling may be provided which may be incorrect).

Phonetic matching plays an important role in information retrieval in multilingual environment. Information retrieval needs an exact match for a given string. Phonetic matching can be defined as a process of identifying a set of strings those is most likely to be similar in sound to a given keyword. The strings can be spelled using different writing styles but they can be matched phonetically. All the strings represent the same keyword only way of writing is different. Since in rural areas, the word may be spelled or pronounce either wrongly or differently. We can retrieve the data using phonetic matching. There is no need of exact string matches.

There are many approaches had been proposed in order to find the phonetic matching of strings like Soundex, Q-gram, edit distance, Caverphone, Phonix etc. Each approach generates a code for the strings and matches them through edit distances. If the edit distance is within threshold then those strings are more close to each other [1]. But, these approaches have been not found suitable for all the strings. There are some limitations like first, generation of code follows long procedure, and second, they are generating the same code for mismatch strings or generate different codes for match strings.

In this paper, we proposed approaches for Indic languages such as Hindi and Marathi which provides a simple and efficient way of matching the strings. Our scheme will work on text encoding technique by using phonetic rules of the languages.

The paper organization is as follows. Section I gives the introduction to the phonetic matching. Section II is helpful to understand the existing phonetic matching approaches. Section III gives the details of the proposed phonetic matching approaches for Hindi and Marathi. Section IV shows the results and performance of the proposed approaches and at last section V concludes the paper and followed by the references.

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

II. PHONETIC MATCHING APPROACHES: A SURVEY

In this section, we describe the various approaches for phonetic matching. Existing approaches were developed to match the string approximately or exactly, where in our approach we can match the strings by combining both. These approaches used for phonetic matching.

Code:	0	1	2	3	4	5	6
Letters:	अ, इ, ओ, उ, व, ह	ब, प	च, ग, ज, क, स	ड, ट, त	ल, ळ	म, न	र

Table 1: Soundex Codes for Hindi as per Algorithm

A. Soundex Approach

In this approach, each string is converted into a code which consists of a first letter of a string and three numbers. The numbers are assigned to each letter as per guidelines described by the algorithm. Zeros are added at the end if necessary to produce four-character code. Additional letters are discarded [3]. The pitfalls of this algorithm are, first, it produces the same code for phonetically two different strings or producing different codes even if they are same. Second, this approach works for only English not for Indian languages since no codes are being assigned for some of the letters of the alphabet [1, 2].

Consider the two strings in English as, '**sandy'** and '**sandhya'** for phonetic matching. After applying the rules and algorithm, we are getting the same code as '**s530'.** So, by looking at the code we have to say that both strings are phonetically matching, but both the strings are phonetically different.

B. Q-Gram Approach

This method measures the string distances based on q-gram counts, where q-gram of a string s is any substring of s of some fixed length q. A simple such measure is to count number of q-grams for the two strings, with a higher count yielding a stronger match. But, this algorithm is not exactly phonetic since they do not operate based on comparison of the phonetic characteristics of words. Since phonetically similar words often have similar spellings this technique can provide favorable results, it also successfully matches misspelled or otherwise muted words even if they are rendered phonetically disparate [1].

For example, for the strings, '**sandy'** and '**sandhya'**, with **q = 2**. With q-gram algorithm, we are forming 4-grams but only 3-grams are being matched and for 2 grams there are no grams to match. We have to find out similarity of strings as phonetically same which depends on number of matching q-grams.

III. PROPOSED MATCHING APPROACHES FOR HINDI AND MARATHI

A. Soundex Approach

If we consider the soundex algorithm and coding scheme for Hindi or Marathi, we can assign the code for each letter as per the table 1. If we match the strings with this algorithm then we will get different codes for the same string or we cannot generate the code. From this methodology, the code can be generated by taking halant for each consonants concatenated together with vowel. This will change the code and it takes long time to parse the string and assign the code to it. We found out that there is no code for many letters such as ण, झ, छ, घ for Hindi language.

If we consider the same rules and algorithm for Hindi as shown in table 1, the codes for 'सँडी' and 'संध्या' strings are same as 'स530'. By using soundex method, we are getting the same code for both the strings entered in Hindi. As codes are same, we have to say that both strings match phonetically but they are not matching. There is an ambiguity to match.

B. Q-Gram Approach

If we apply the same rules in order to form q-grams for Hindi or Marathi strings then the entire string will change or no q-grams to match. One problem with this method is that if parsing of a string goes wrong then it affects the generated code.

Consider the same strings as, 'सँडी' and 'संध्या'. If we apply Q-gram algorithm for Hindi, we will get only two q-grams

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

(where q=2) that matches but two q-grams does not matches. There is also an ambiguity for phonetically match.

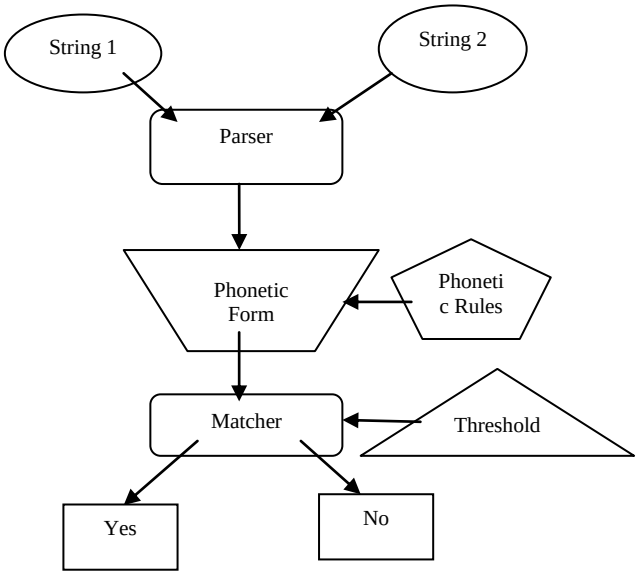
C. *Indic-Phonetic Approach*

In this approach, we are providing the user interface to provide the keywords or strings in local languages like Hindi or Marathi for phonetically matching. Each keyword is entered by using Indic IMEs for Hindi and Marathi, which are available on bhashaindia.com web site [4]. Each entered string will be parsed to get exact combination of vowels, consonants and/or modifiers as phonetic string according to phonetic rules for each language mentioned below. We are assigning the codes for each phonetic string in order to match. The difference between these codes are compared with the threshold value, if it is less or equal to threshold value then these two strings are phonetically matched else not.

1) Methodology:

The main objectives of the proposed system are, to convert the entered strings into its equivalent phonetic forms by applying phonetic rules for each language and to compare the generated code to match. In this approach, we are forming the rules as per pronunciation of letters in Hindi and Marathi languages. Each string will be interpreted and converted to its phonetic form using these rules. We are forming the codes for each string as per Unicode table [5] which will be common for both the languages as the script is same for both Hindi and Marathi. These codes are compared with threshold as 5% fixed value. If they are within a range then we can say that they are phonetically matched, otherwise not. Figure 1 shows the overall approach for proposed Indic-Phonetic approach for Hindi and Marathi.

Fig. 1: Proposed Indic-phonetic Matching Approach



2) Phonetic Rules:

Some phonetic rules have been formulated according to the occurrences in pronunciation of letters in Hindi and Marathi languages words [6]. These rules will give the exact letter to be considered for the pronunciation of letters before and after each letter. Each letter from the parsed string is being checked against the rule of the language. The rules of pronunciation have been applied in order to acquire the correct form of each letter as vowel or consonant or modifier.

3) Threshold Calculations:

Now, for each letter from a phonetic string the Unicode has been assigned from the Unicode mapping table [5]. Finally, we acquired the total Unicode for the string as summation of all Unicode values of each letter of the string. We had used the difference between the codes as a threshold value for matching. If the result crosses the threshold value then the entered two strings are not phonetically matched else they matched.

Example:

Consider two strings as ‘संतोष’ and ‘संथोष’.

The phonetic forms of the strings are as ‘स्अन्अत्ओष’ and ‘स्अन्अथ्ओष’.

After parsing the strings, we acquire its corresponding forms as:

स्अन्अत्ओष = स ् अ न ् अ त ् ओ ष

स्अन्अथ्ओष = स ् अ न ् अ थ ् ओ ष

Then we assigned the code to each string as:

स ् अ न ् अ त ् ओ ष = 23487

स ् अ न ् अ थ ् ओ ष = 23488

By considering 5% threshold to match, the difference is calculated as:

(23488 - 23487)/23488) * 100 = 0.0042%

So, the threshold value as edit distance is 0.0042% between the two strings.

4) Evaluation Approaches:

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

We evaluated the three phonetic matching approaches, namely soundex, Q-gram and our proposed Indic-Phonetic with various cases by taking some parameters like length-of-string (LOS), strings differ in vowel and strings with compound words. We made a large database consisting of 50 pairs of string for both Hindi and Marathi. Some pairs are of length of 2 letters, some are three, some are of three or/and four letters and some pairs are differ in vowel/s, differ in consonant/s and compound words [7]. We have taken five cases as per the parameters, which are mentioned below:

Case I (LOS = 2): In this case, we evaluated the three approaches by considering a string pair of TWO letters to match. For example, 'टेढ़ा and पेढ़ा' string pair in Hindi or Marathi has been evaluated by all three matching approaches. This string pair is differs in consonant also in the end but looks like same phonetically.

Case II (LOS = 3): In this case, we evaluated the three approaches by considering a string pair of THREE letters to match. For example, 'पाटन and पाठन' string pair in Hindi or Marathi has been evaluated by all three matching approaches.

Case III (LOS = 3 or 4): In this case, we evaluated the three approaches by considering a string pair of THREE OR FOUR letters to match. For example, 'परिवार and परिवाद' or 'करता and करतार' string pair in Hindi or Marathi has been evaluated by all three matching approaches.

Case IV (Differ in vowel): In this case, we evaluated the three approaches by considering a string pair as change in vowel in a string to match. For example, 'पापड़ and पापड़ी' string pair in Hindi or Marathi has been evaluated by all three matching approaches. This pair differs in vowel at the end.

Case V (compound words): In this case, we evaluated the three approaches by considering a string pair as compound string to match. For example, we evaluated the string pairs like 'पाप-पुण्य and 'पापपुण्य' or 'त्रिभुवन' and 'त्रि भुवन 'in Hindi or Marathi by all three matching approaches.

IV. RESULTS

The table 2 shows the results of matching of all 3 approaches according to length-of-string (LOS), differ in vowel/s and matching of compound words for Hindi. We can conclude that soundex approach has maximum number of non-matching pairs since the string pairs are not starting with same letter. Q-gram and Indic-Phonetic approaches give almost same number of matching string pairs since Q-gram approach depends on number of q-grams that matches.

Table 2: Result of Matching for Hindi according to LOS

Hindi	LOS	Matching	Non-matching
Soundex	LOS = 2	5	10
	LOS = 3	9	21
	LOS = 3 or 4	3	2
	Compound Word	15	0
	Differ in vowel/s	10	0
Q-Gram	LOS = 2	10	5
	LOS = 3	19	11
	LOS = 3 or 4	5	0
	Compound Word	9	6
	Differ in vowel/s	7	3
Indic-Phonetic	LOS = 2	11	4
	LOS = 3	23	7
	LOS = 3 or 4	4	1
	Compound Word	4	11
	Differ in vowel/s	4	6

The table 3 shows the results of matching of all 3 approaches according to length-of-string (LOS), differ in vowel/s and matching of compound words for Marathi. We can conclude that soundex approach has maximum number of non-matching since the string pairs are not starting with same letter. Q-gram and Indic-Phonetic approaches give almost same number of matching string pairs since Q-gram approach depends on number of q-grams that matches.

Table 3: Result of Matching for Marathi according to LOS

Marathi	LOS	Matching	Non-matching
Soundex	LOS = 2	3	30
	LOS = 3	2	12
	LOS = 3 or 4	2	1
	Compound Word	15	0
	Differ in vowel/s	10	0
Q-Gram	LOS = 2	16	17
	LOS = 3	5	9
	LOS = 3 or 4	2	1
	Compound Word	12	3
	Differ in vowel/s	4	6
Indic-Phonetic	LOS = 2	13	20
	LOS = 3	4	10
	LOS = 3 or 4	2	1

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

	Compound Word	5	10
	Differ in vowel/s	7	3

A. Analysis of Results

The overall analysis of results of the approaches is shown from figure 2 to figure 10 for Hindi and Marathi. Figure 2 and figure 3 shows the overall analysis of matching of the string pairs where Indic-Phonetic approach gives highest matching and soundex approach gives least matching performance for Hindi. Soundex approach has least performance where as Q-gram and Indic-Phonetic gives almost same performance for Marathi.

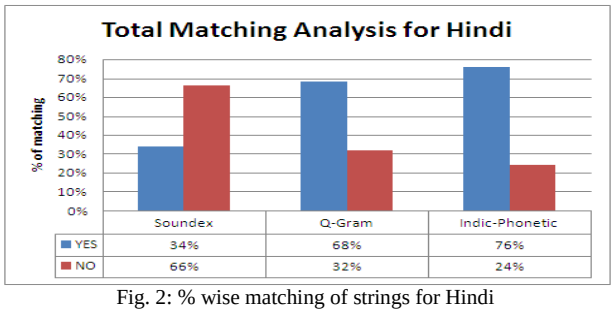


Fig. 2: % wise matching of strings for Hindi

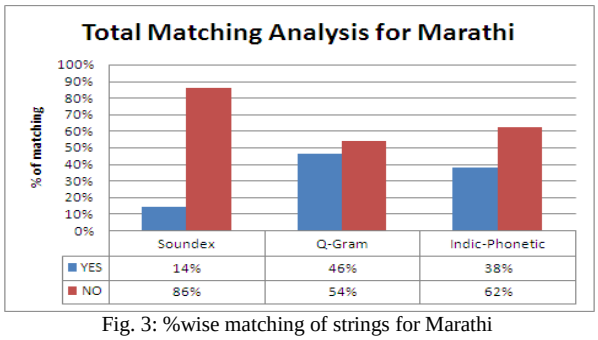


Fig. 3: %wise matching of strings for Marathi

Figure 4 shows the performance analysis of string pairs for Hindi where length-of-string (LOS) is the parameters. The string pairs with LOS = 3 has poor performance by soundex and Q-gram whereas Indic-Phonetic approach gives best performance for any LOS. Figure 5 shows the performance string pairs of Marathi where soundex gives very poor performance for any length of the string. The string pairs with LOS = 3 or 4, the Indic-Phonetic approach matches 75% whereas other two matches only 25%. Soundex performance is very poor as compared to other two approaches. String pair with LOS = 3, IP performed well as compared to Q-gram approach.

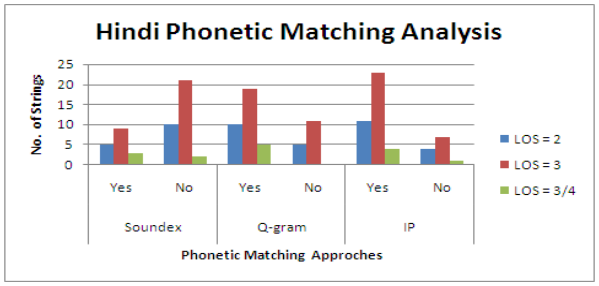


Fig. 4: Phonetic Matching Results Analysis for Hindi According to LOS

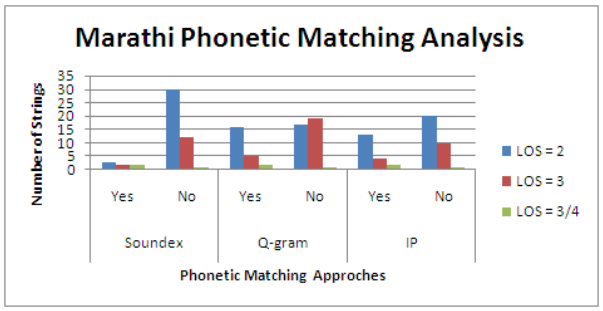


Fig. 5: Phonetic Matching Results Analysis for Marathi According to LOS

Figure 6 shows the overall matching of the string pairs by comparing the matching and non-matching options of all the three approaches. 75% of Marathi string pairs are not matching for combined approach and Q-gram approach gives negative performance to Hindi string pairs. Indic-Phonetic performed moderate and accurate.

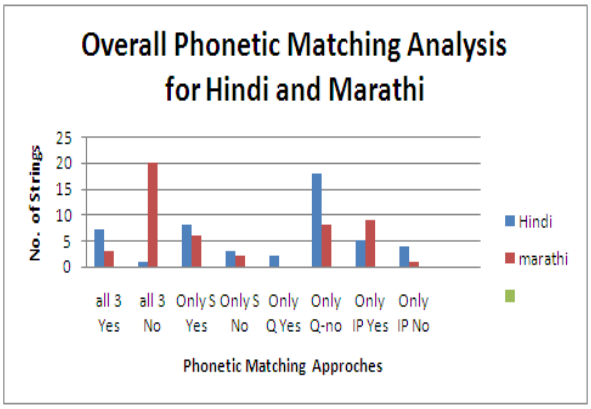


Fig. 6: Overall Phonetic Matching Results Analysis for Hindi and Marathi According to Matching Approaches

Figure 7 and figure 8 shows the performance of matching of the strings which are differs in vowel/s. In this approach, the soundex approach matches every string since each string starts with same letter in a pair. Q-gram gives 50-50% performance and Indic-Phonetic approach gives accurate performance since vowel/s has modified the string and change the code of the string.

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

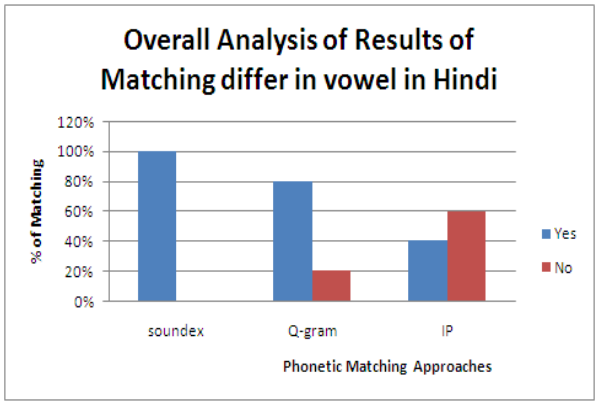


Fig. 7: Overall Analysis of Results of Matching differ in vowel in Hindi

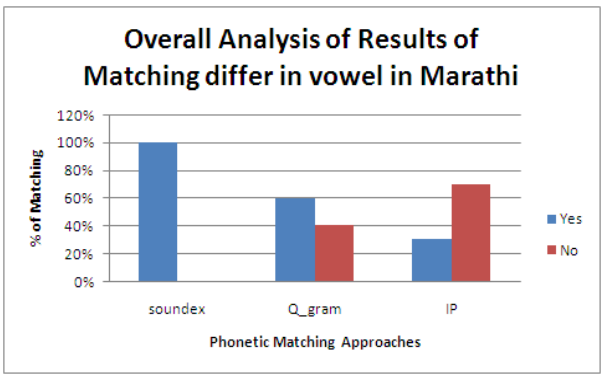


Fig. 8: Overall Analysis of Results of Matching differ in vowel in Marathi

Figure 9 and figure 10 shows the performance of the matching of compound words in Hindi and Marathi. In this case, soundex approach has matched all the strings, Q-gram approach has been given 50% matching performance and Indic-Phonetic approach matched accurately with 25% matching.

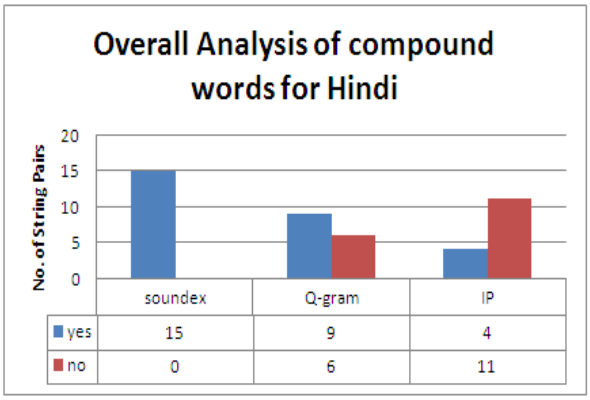


Fig. 9: Overall Analysis of Results of Matching of compound words in Hindi

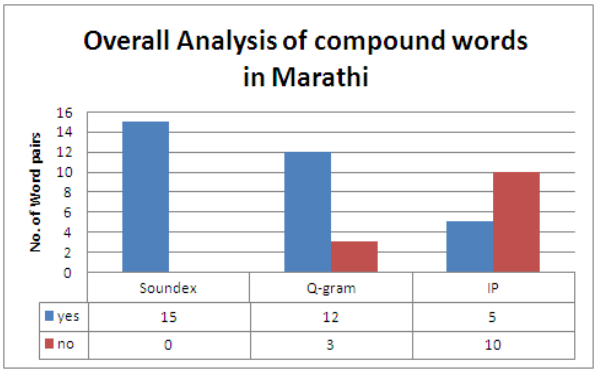


Fig. 10: Overall Analysis of Results of Matching of compound words in Marathi

V. CONCLUSIONS

There are many features of phonetic for Indian languages. In this paper, we proposed a rule-based approach for phonetic matching. We tried to include most of them by forming the rules in order to match the strings for Hindi and Marathi languages. In this approach, we explored and compared some of the phonetic matching approaches by taking cases like LOS, differ in vowel/s, and compound words. We have found out that these approaches are lagging in accommodating all the letters from an alphabet for both Hindi and Marathi languages. Our proposed approach has performed best and accurate for all the cases. Advantages of our approach are that it gives the user and developer a simple, easy and efficient way of phonetic matching.

REFERENCES

[1] Justin Zobel and Philip Dart. 'Phonetic String Matching: Lessons from Information Retrieval'. In Proceedings of the 19th SIGIR (1996) on Research and Development in Information Retrieval, pages 166-172, 1996.
[2] Alexander Beider and Stephen P. Morse, **'**Phonetic Matching: A Better Soundex**'**. Association of Professional Genealogists Quarterly, March 2010.
[3] The Soundex: http://en.wikipedia.org/wiki/Soundex
[4] Microsoft Bhasha Empowering Indic Languages Computing: http://www.bhashaindia.com
[5] Technology Development for Indian Languages: http://tdil.mit.gov.in/unicodeapril03.pdf
[6] Sandeep Chaware and Srikatha Rao, 'Rule-based Phonetic Matching Approach for Hindi and Marathi' at Computer Science & Engineering: An International Journal (CSEIJ), Vol. 1, No. 3, Aug. 2011, pp 13-24.
[7] Peter Christen, A Comparison of Personal Name Matching: Technique and Practical Issues, Joint Science Technical Report Series, ANU, 2006.

ISSN 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2012

AUTHOR'S BIOGRAPHY

[1]Sandeep Chaware is a research scholar at MPSTME, NMIMS University, Mumbai. His area of research is phonetic and semantic matching approaches for Hindi and Marathi.

[2]Srikantha Rao is a PhD supervisor at MPSTME, NMIMS University, Mumbai. He is guiding the students in various research areas like natural language processing, parallel computing etc.