CHAPTER 2
THE MATCHING PROBLEM

2.1 INTRODUCTION
The solution of matching problem may involve the parameters and resources used. The
parameters may be dependent or independent. A function operates on these parameters to
give the solution. The resources can be categorized into two types i.e. internal and external.
Internal resources may be data from database, any other dependent modules, memory, CPU
and so on. The external resources are the special resources developed for the domain i.e.
dictionary, on-line input or processing module and so on. In our research work, matching
function takes the resources and operates on parameters to give result.

In this chapter, we have defined the problem statement along with proposed system
architecture. Each approach has been defined with the function needed to match along with
necessary parameters and resources. We have elaborated the scope of the research in term of
business and applications, also principle findings of the research work have been mentioned.

2.2 PROBLEM STATEMENT AND PROPOSED DESIGN

2.2.1 PROBLEM STATEMENT

There have been different formalizations of matching and its results. We provide and used a
general definition as follows:

The matching operation determines a pair of matching by using an operator. There are many
parameters which can extend the definition of matching process, namely i) use and
interpretation of input strings for the languages ii) the matching parameters P e.g. writing
style, phonemic form, threshold value and so on and iii) external resources like phonetic
rules, data from DBMS system, bi-lingual dictionary, synset, ontology and so on. The entire
process can be defined as follows:
The matching process can be viewed as a function $f$ which matches a pair of strings $S_1$ and $S_2$ with input $I$, a set of parameters $P$ and a set of resources $R$. It returns a result as yes or no or information as IR from a system [54].

$$\text{Result} = f(S_1, S_2, I, P, R)$$

This can be schematically represented as in figure 2.1.

In phonetic matching process, we can match single string entered by a user to a system’s own data maintained for a particular domain or we can match two strings. Instead of using text-to-speech conversion, we can consider input as writing style for a string which we called as phonetic equivalent. The matching process will give result of matching in the form of IR. In another approach, two strings can be inputted to match and results can be obtained as yes or no. In this case, resources may be phonetic rules for the languages.

In semantic matching process, the system will match semantically by taking either one string or a pair. The result will be in the form of IR after matching. The resources $R$ would be language dictionary, ontology and synset-based dictionary.
In order to solve phonetic and semantic issues related to Hindi and Marathi languages, we need an efficient system or approach. The main research problem can be divided into two sub problems as follows:

1. If a user doesn’t know the exact keyword/s to be entered in a query then he may use different writing style for a keyword. There may be various difficulties arise like system will respond with wrong result or the system will not accept the query terms or it may fail. Main motivation of the research is to acquire the relevant information from information system irrespective of typographical or spelling variations of the words in a query. Irrespective of languages, the phonetic matching issues have to be handled which lead to information retrieval for any domain.

Our research addresses the phonetic matching issues by applying phonetic rules for each language instead of using TTPs.

2. If user’s entered keywords are different than the system’s keywords. Then the issues related to semantically matching with keywords searching and retrieval is to be handled. Once these issues are being handled efficiently then the system can be extended to an inference system. The performance of the system should be calculated, which may depend on many factors such as query terms, its corresponding synsets, searching within ontology, traversing the ontology and display. Main motivation of the research is to acquire the relevant information from information system irrespective of use of synonyms in a query for Hindi and Marathi.

Our research addresses these issues and proposed a system with highest performance in terms of precision and recall.

**2.2.2 PROPOSED DESIGN**

The proposed designs for phonetic matching are shown in figure 2.2 and figure 2.3. They involve the modules such as query, parser, phonetic form (for approach II) and matcher. The
matcher is a function which takes the input parameters and resources needed to match. After matching, the result may be in the form of yes or no or as information retrieval.

2.2.2.1 Phonetic Matching Approaches

![Figure 2.2: Phonetic Matching Approach-I Using Mapping Methodology](image)

![Figure 2.3: Phonetic Matching Approach-II using Phonetic Rules of Hindi and Marathi](image)

2.2.2.2 Semantic Matching Approach

The proposed design for semantic matching is as shown in figure 2.4. This involved the modules such as query, parser, stemmer, ontology builder and matcher. The matcher is a function which takes the input parameters and resources needed to match. After matching, the result will be in the form of yes or no or as inference which can be called as information retrieval.
Inference as result

Figure 2.4: Semantic Matching Approach Using Ontology

This approach will form and process query two times. First, to build ontology from database that represents the knowledge for the domain from query terms and second when a user acquires an inference from a domain from query terms. In both approaches, the matcher will match the strings with necessary parameters and resources to acquire the required result.

2.3 SCOPE OF THE RESEARCH

In multilingual environment, we need a system which will provide the way of handling the multiple languages i.e. their scripts, writing styles, translations or transliterations, phonetic issues, processing, providing information as knowledge, use of that information etc. There are numerous applications as an example where we maintained the data in multiple languages. Some of them are government records, land records, real estate records, government schemes,
railway reservation charts, PhD thesis etc. If this information is not available in multiple languages, lot of difficulties may arise for people especially in rural areas to understand the literature, where percentage of literacy is very less. Also, some people may make mistakes in making the pronunciation of certain words or writing styles may be different. Some people may use alternative words for the same word whose meaning may be same. In this scenario, you may need a system which will handle syntactic as well as semantic issues of Indian language.

By understanding the need of a system and by looking at the goal and objective as vision, each project objective is also linked to the business vision. Business objectives can be stated as business measures such as increasing profits by 20%. In our case study, either at shopping mall or at grocery shop, if such a system is available then the profit surely may increase by more than 20% approximately, since all categories of people will use the system. They will acquire the service much faster than expectation.

2.4 PRINCIPAL FINDINGS

Our main objectives have been successfully completed by designing, implementing, testing with real time scenarios. The main objective of our system is to solve various syntactic and semantic issues related to Indian languages and to implement the information retrieval technique for both.

The principal findings of our research work are as below –

a) We have proposed a system for handling the phonetic matching through writing style, in which strings are transformed into various forms as combinations of consonants, vowels and/or modifiers. This solves the problem of writing the strings and information retrieval for the same.

b) Further, we extended our research for phonetic matching by transforming the entered strings into its phonetic form by applying language’s rules instead of using IPA. This makes faster and accurate processing of strings. A string pair with variations in vowel/s at the end of the string makes difference in matching. A string pair with
addition or deletion of vowel/s in the middle of the string makes no difference in matching.

c) We proposed the semantic issues in two stages; first, we transformed the entire system into ontological form according to our proposed ontology building approach. We computed the performance of number of users per sub-domain in terms of ontology terms, time to build, precision and recall.

d) In the second stage, the system is being used as inference system by considering the semantic issues from ontology. We compute the semantic similarity by using synsets for Hindi and Marathi languages instead of using WordNet for each language. Finally, we got better performance in terms of precision and recall values. In this way the IR issues have been handled.

2.5 SUMMARY

In this chapter, we introduced the multilingual environment for Hindi and Marathi languages. We focused on the language issues along with phonetic and semantic matching issues. The goal is to provide IR in multilingual environment for Hindi and Marathi users, where this goal will be satisfied by phonetic and semantic matching approaches which give rise to applying the approach for any domain.