Development of Morphological Analyzer for Hindi

Mayuri Rastogi
Department of Computer Science
Amity University
Lucknow, India

Pooja Khanna
Faculty in Department of Computer Science
Amity University
Lucknow, India

ABSTRACT
One of the important phase of Natural language Processing is Morphological Analysis that helps in work of machine translation. Effective implementation of morphological analyzer can be seen in language which is rich in morphemes. Hindi being an inflected language has capability of generating hundreds of words from the root word. It is morphologically rich language, due to wide variety of words available in Hindi. This paper is primarily concerned with the design of a morphological analyzer for Hindi language. The input to this analyzer will be a Hindi word or sentence and after doing the proper analysis it will return the root word along with its feature as output. The features will have categories like part of speech, gender, number, and person. Two approaches will be followed by analyzer-rule based and corpus based. Till now none of the developed morphological analyzers have worked for both inflectional and derivational morphemes.

Keywords
Morpheme inflected, root word, corpus.

1. INTRODUCTION
Basically there are five stages of NLP namely-morphological analysis, syntactic analysis, semantic analysis and pragmatic analysis. Morphological analysis is concerned with analyzing individual words into their components. There are basically two classes of morphology-

Class 1- inflectional morphology

Class 2-derivational morphology

In inflectional morphology the new word formed is of same class as that of previous word i.e. if previous word is of class noun then after adding affixes to it will remain a noun. This can be explained by a example as- घड़ी (Noun) becomes घडियाँ (Noun) on adding ियाँ as suffix

In derivational morphology, the new word formed will be of different class than the previous word or let's say root word. For example- मूर्ख (fool) (Adj) becomes मूर्खता (foolishness) (Noun) on adding ता as suffix.

The objective of this paper is to present a idea about a morphological analyzer for Hindi language that works on both classes of morphemes. Also this paper aim to present a effective morphological analyzer as till now most of the morphological analyzers have been developed either for inflectional or for derivational morpheme. In this paper we will follow two approaches for development of my morphological analyzer.

Approach 1- corpus based

Approach 2- rule based used for derivational morphology.

The first approach follows a database that stores the root word along with some of the features of root word like-gender (masculine or feminine), category (noun, pronoun, verb, adverb, adjective, particles, connectives and interjections) and number (singular or plural).

The second approach takes into consideration careful analysis of Hindi word and then formation of rules which put the word in certain category specified earlier. Before applying any of the above approaches careful study of inflected and derivational word of Hindi language was done so that the developed analyzer works for both kinds of words.

In the below sections, other related works, structure of Hindi sentences/words and approach followed is explained.

2. RELATED WORK
There is no morphological analyzer that works for both inflectional and derivational morphology to the best of our knowledge.

However, morphological analyzer and generator for inflectional morphology is discussed in paper by Vishal Goyal,Gurpreet S. Lehal[1] in which they focused on the time taken to search for a word in database to know the grammatical information of word. But this method was limited for languages that have finite or less number of inflections.

Niraj Aswani and Robert Gaizauskas[2] presented a rule-based morphological analyzer. The approach uses suffix replacement rule for finding out the root word from its inflected form. The major drawback in their work was that not all rules produced correct results.

Deepak kumar, Manjeet Singh and Seema Shukla[3] presented paper with idea to use FST(Finite State Transducer) for developing morphological analyzer for Hindi language. They discussed two methodologies for development of morphological analyzer- lexicon generator and generation of morphological processor.

Nikhil Kamparthi, Abhilash Inumella and Dipti MisraSharma[4] presented an algorithm to upgrade Hindi inflectional analyzer to derivational. The algorithm used the principle of Porter's stemmer and Krovetz stemmer. The approach followed included study of Hindi derivatives, defining derivational rules, using Wikipedia for confirming genuine data n devising algorithm for derivational analysis.

The paper on "The effect of dictionary and lexicon to morphological analysis" by Mohd Yunus Sharum[5] highlights that use of lexicon in morphological analyzer improves the correctness and performance of the analyzer (in terms of quality of output).

The lightweight stemmer for Hindi language developed by Ramanathan and Rao[6] was based on hand-crafted rules. After careful observation of Hindi words for noun, verbs, adjectives and adverbs they produced list of suffixes for each category.

3. STRUCTURE OF HINDI WORDS

To develop morphological analyzer for Hindi language it is important to know about the actual structure of Hindi words and sentences-how they are formed, their special characteristics etc. Hindi is quite similar to other Indo-Aryan languages in terms of linguistic characteristics. There are ten vowels in Hindi language. The Hindi syllable contains a vowel as it nucleus, followed or preceded by consonants. Words usually have two or three syllable.

Hindi morphological structure of Hindi consists of various word classes that include description about their derivational and inflection forms. In the forthcoming sections, details about the word classes are given which are referred from book "Modern Hindi Grammar" by Omkar K.Koul [11]

3.1 Nouns

Nouns in Hindi are inflected for gender, number and case.

3.1.1 Gender

There are three types of nouns:

Type I have masculine nouns ending with आ/a: /

Types II have all other masculine nouns.

Types III have all other feminine nouns.

Generally आ/a/ ending masculine nouns have feminine forms ending in ई/i: /

Masculine			Feminine		
लड़का	Ladka	boy	लड़की	Ladki	girl
चरखा	Charkha	wheel	चरखी	Charkha	Reel

ई/i: / ending animate masculine nouns have their feminine forms ending in -अन/-an/

Masculine			Feminine		
तेली	Teeli	Oilrman	तेलन	Teelan	oilwoman
माली	Maali	Gardener	मालन	Maalan	Gardener

Nouns ending in आ/a:/ form their feminine by replacing आ/a:/ with –इया/-iya:/

Masculine			Feminine		
गुड्डा	Gudda	Boy toy	गुडिया	Guddiya	Girl toy
चूहा	Chuha	Rat	चुहिया	Chuhiya	Rat

Generally –आ/a: / ending nouns are masculine are replaced by - ई/-i: / to form feminine

Masculine	Feminine
राजा	Raja	King	रानी	Rani	Queen
बिल्ला	Billa	Cat	बिल्ली	Billi	Cat

Add suffix नी/ni: / to the masculine nouns to form the feminine

Masculine			Feminine		
डाक्टर	Docter	Docter	डाक्टरनी	Doctorni	docter

Add suffix ई/-i: / to the masculine noun to form feminine

Masculine			Feminine		
देव	dev	God	देवी	devi	goddess
नट	nat	Acrobat	नटी	nati	Showgirl

3.1.2 Number

Singular and plural are the two types of number.

The आ/a: ending masculine noun (including pronoun and adjective) with some exceptions change into plural ending with ए/e/

Singular		Plural	
चीता	Leopard	चीते	Leopards
छाता	Umbrella	छाते	Umbrellas
*पिता	Father	पिता	Fathers
*नेता	Leader	नेता	Leaders

* remain same in their plural form

All other consonants and other vowel-ending nouns do not change their plural forms.

मच्छर	Mosquito
गिलहरी	Squirrel

To form feminine plurals add the suffix एं/ĕ/ to the consonant-ending singular forms

Singular		Plural	
पुस्तक	Book	पुस्तकें	Books

Feminine nouns ending with ई on adding इयाँ becomes plural

Singular		Plural	
नारी	Woman	नारियॉ	Women
सखी	Friend	सखियॉ	Friends

In this last vowel of the stem is removed.

3.1.3 Noun Derivation

Nouns in Hindi are derived from nouns, adjectives and verbs by using suffixes.

3.1.3.1 Nouns from Nouns:

Suffixes are दार–da:r ,गर-gar and दान –da:n

थाना	दार	थानेदार	Inspector
सौदा	गर	सौदागर	Merchant
खजाना	ची	खजानची	Cashier
रोशन	दान	रोशनदान	Window
गुसल	खाना	गुसलखाना	Bathroom

3.1.3.2 Nouns from Adjectives:

Suffixes for this purpose are ई –i, ता –t:a ,pan, आई-a:I, इयत-iyat, आस-a:s

खुश	Happy	खुशी	Happiness
बुरा	Bad	बुराई	Badness
एक	One	एकता	Unity
बालक	Boy	बालकपन	Childness
आदमी	Man	आदमियत	Humanly
खट्टा	Sour	खटास	Sourness

3.1.3.3 Nouns from Verbs

To derive nouns from verbs suffixes used are अस-as, अन-an, ई-ee, वत-vat, ना-na

पढ़	Read	पढ़ना	Reading
धड़क	Throb	धडकन	Throbbing
लड़	Quarrel	लड़ाई	Dispute
बना	Make	बनावट	Shape

3.2. Adjectives-

In Hindi these are classified as inflected and uninflected.

3.2.1 Inflected Adjectives:

Masculine		Feminine
Singular	Plural	Singular/plural
भूख	भूखे	भूखी
गोरा	गोरे	गोरी
मोटा	मोटे	मोटी

Uninflected Adjectives:

सुन्दर लड़का/लड़की	Sundar ladka/ladki	beautiful boy/girl
सफेद आदमी/औरत	Safed aadmi/aurat	white man/woman

3.2.2 Derivation of Adjectives:

Adjectives are derived from nouns by adding suffixes आ-aa, ई-I, उ-u, ईला-ila , लू-lu , इक-ik, जनक-janak, दाई-daai, मय-may, वन-van, आना-ana, नाक-nau, ईन-inn, मंद-mand, दार-dar.

Noun		Adjective	
सच	truth	सच्चा	truthful
बाजार	market	बाजारू	common
रस	juice	रसीला	juicy
दया	mercy	दयालु	Kind
दिन	day	दैनक	daily
बल	strength	बलवान	strong
विदेश	foreign	विदेशी	foreigner

3.3 Verbs

There are two types of verbs: main verb and auxiliary verb. Main verb is classified as simple, conjunct and compound verb.

Verbal construction is classified as:

3.3.1 Intransitive verb

3.3.2 Transitive verb

3.3.3 Ditransitve verb

3.3.4 Causative verb

3.3.5 Dative verb

3.3.6 Conjuct verb

3.3.7 Compound verb

Here we have considered intransitive, transitive and ditransitive verb in our database.

3.3.1 Intransitive verbs

They are like आ-aa, जा-ja, उठ-uth, बैठ-baith, they do not take direct object and are not marked by any preposition in present or future tense.

Eg.-

अमित घर जाएगा

Amit ghar jae:ga

Amit will go home

3.3.2 Transitive verbs

They are derived from intransitive verbs by certain vocalic changes to the verb roots.

Eg.-

Intransitive		Transitive	
मर		मार	
पिस	be ground	पीस	Grind
घूम	go round	घुमा	turn around

In some cases beside vocal changes consonantal changes also take place.

Eg-

Intransitive		Transitive	
फाड़	Be torn	फाड़ना	Tear
बेच	Be sold	बेचना	Sell

3.3.2 Ditransitive verb

They are like देना dena(to give), सुना (to tell), बेचना becna(to sell). They take three arguments-subject, object and indirect object. Indirect objects are in dative while rest follow transitive pattern.

Eg-

राम ने रीता को पुस्तक दी

Ram gave book to Rita.

4. APPROACH FOLLOWED

The methodology used by me for development of morphological analyzer for Hindi can be explained as-

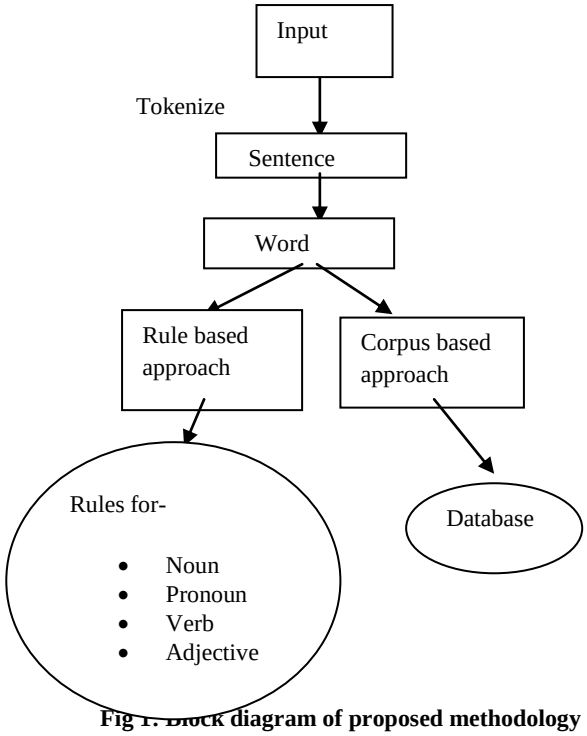

Fig 1: Block diagram of proposed methodology

Thus, from above the methodology followed is quite clear.

The proposed morphological analyzer works for both- Hindi words and sentences.

If the input to analyzer is a sentence, the sentence is tokenized into words. The word is then searched in database or the maintained corpa.

Database is maintained as-

INPUT {Word_id(Primary key), name, W_Word Category, Gender, Number}

Database stores the root word along with some features like category, gender etc. also it is desirable to store some common or frequently used common noun in database for better accuracy however it is a tedious job. With space nowadays of no major issue, searching time is primarily of concern. Efficient searching technique need to be implied for effective use of maintained database or dictionary. The prime focus will be on nouns except proper noun.

If the tokenized word is not found in database then rules are applied and the root word along with its features is displayed

Rules here are not learnt automatically.

5. CONCLUSION AND RESULT ANALYSIS

The methodology illustrated above can be analyzed with the examples of Hindi word-

ILLUSTRATION I:

INPUT- निडरता

CHECK if the input is a word or sentence. Here it is a word, so go and search for match in database. If match found then find the features of word given as input (rules) and display morphemes of the input word along with the features of the word.

OUTPUT-

निडरता = नि(prefix) + डर(Category) + Feminine(Gender) + Any(Number)+ ता (Suffix)

ILLUSTRATION II:

INPUT- "मनु धीरे चलता है"

CHECK if the input is word or sentence. Here input is a word, so sentence is tokenized into individual tokens or words and then matching of each word with the databade takes place. The first token मनु was serached in database, match found and display features of मनु. Move to next token which is धीरे and repeat above procedure. Similarly do for third and fourth token which are चलता and है respectively.

OUTPUT-

मनु = मनु (rootword) + Noun(Category) + Mascuilne(Gender) + Singular(Number)

धीरे = धीरे (rootword) + Adj(Category) + Any(Gender) + Singular(Number)

चलता = चलता (rootword)+ Verb(Category) + Mascuilne(Gender) + Singular(Number)

है= indeclinable

ILLUSTRATION III:

INPUT- सरलो

CHECK if the input is word. Here it is a word indeed but no match is found in the database as there is no word like सरलो in Hindi langugae however सरल is a word so invalid word is given out as output.

OUTPUT-

Invalid word.

5.1 Conclusion

The paper presents a method for developing Hindi morphological analyzer that works on inflectional and derivational morphemes by using rule based and dictionary or corpus based approach. The commonly used words along with their features are stored in corpus. Time and accuracy of result is preferred over memory space. But memory space is not a problem these days, neither in terms of cost nor in terms of storage requirements- so this method can be used to get good results. Moreover, this method has an advantage over other suffix trimming or using Porter's stemmer and Krovetz stemmer as it gives root word along with the features rather than just giving root word as the result. The approach discussed in this paper has capability of working on either type of morphology.

6. ACKNOWLEDGEMENTS

This research paper is made possible through the help and support from parents, teachers and friends. I dedicate my acknowledgement of gratitude towards my guide 'Ms. Pooja Khanna' for her kind support and advice on every step. My thanks to the experts who have contributed towards development of this paper and to my parents and family, without whose financial and moral support, this work would not have been materialized.

7. REFERENCES

[1] Vishal Goyal, Gurpreet Singh Lehal,"Hindi morphological Analyzer and generator", First International Conference on Emerging Trends in Engineering and Techonology, USA,pp.1156-1159, 2008.

[2] NirajAswani, Robert Gaizauskas, "Developing Morphological Analyzers for South Asian Languages: Experimenting with the Hindi and Gujrati Languages", Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valleta, Malta pp.811-815, May, 2010.

[3] Deepak Kumar, Manjeer Singh, Seema Shukla, "FST based Morphological Analyzer for Hindi Language", Internatinal Journal of Computer Science Issues(IJCSI), Vol.9, pp. 349-353, July, 2012.

[4] Nikhil Kanuparthi, Abhilash Inumella, Dipti Misra Sharma, "Hindi Derivational Morphological Analyzer", Proceedings of the twelfth meeting of the Special Interest Group on Computational Morphology & Phonology, Canada, pp.10-16, June, 2012.

[5] Mohd Yunus Sharum, "The Effects of Dictionary and Lexicon to Morphological Analysis", International Conference on Computer and Information Application(ICCIA),pp 9-12,2010.

[6] A. Ramanathan and D.Rao, "A Lightweight Stemmer for Hindi", ACM Transactions on Asian Language Information Processing(TALIP),Vol.2, pp. 130-142,2003.

[7] Gill Mandeep Singh, Lehal gurpreet Singh, Joshi S.S.,"A full form Lexicon Based Morphological Analysis and Generation tool", Punjabi International Journal of Cybermatics and Informatics, Hyderabad, India. October 2007, pp 38.

[8] Omkar N. Koul, "Modern Hindi Grammar", Dunwoody Press, USA, 2008.

[9] LTRC, IIIT Hyderabad http://ltrc.iiit.ac.in

[10] Teena Bajaj, Prateek Bhatia, "Semi supervised learning Approach of Hindi Morphology", All India Conferences on Advances in Communication Computers, Control and Knowledge Management (AICACCC-KM), Bahadurgarh ,Feb,2008.