ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 9, March 2013

Hindi Morphological Analysis and Inflection Generator for English to Hindi Translation

Pawan Deep Singh, Archana Kore, Rekha Sugandhi, Gaurav Arya, Sneha Jadhav

Abstract -For incorporating syntactic and morphological information for English to Hindi statistical machine translation simple and computationally less exhaustive concepts are: (i) applying simple transformation rules on the English parse tree and (ii) by using a simple suffix separation, distributed morphological analysis and inflection addition. With only a small amount of bilingual training data and limited tools for Hindi a reasonable performance and substantial improvements are achievable. This paper primarily presents an analysis of nominal inflection in Hindi within the framework of Distributed Morphology. In this paper we will discuss in detail the inflectional categories, inflectional classes, morphological processes operating at syntax, the distribution of vocabulary items and readjustment rules for Hindi nouns and analyze the experimental results we obtained from the system developed and as well as accuracy of the same.

Index- **Morphological Analyzer, Inflection, parts of speech, machine translation.**

I. INTRODUCTION

Hindi is a partially free order language (i.e. the order of the words in a Hindi sentence is not fixed but the order of words in a group/phrase is fixed. Hindi allows greater freedom in word-order, identifying constituents through case marking. Hindi has a relatively richer system of morphology. Morphology is the field of the linguistics that studies the internal structure of the words and their forms in different uses and constructions. It can be mainly divided into two branches – derivational morphology and inflectional morphology. Morphological Analysis and generation are essential steps in any NLP Application. Morphological analysis means taking a word as input and identifying their stems and affixes. Morphological Analysis provides information about a word's semantics and the syntactic role it plays in a sentence. Morphological Analysis is essential for Hindi as it has a rich system of inflectional morphology as like other Indo-Aryan family languages. Morphological Analyzer and generator is a tool for analyzing the given word and generator for generating word given the stem and its features (like affixes). If we had an exhaustive lexicon which listed all the word forms of all the roots, and along with each word form it listed its feature values then clearly we do not need a morphological analyzer. Then we need only to look a given word in the lexicon and retrieve its feature values. But this method has several problems. First, it is extremely wasteful of memory space. Every form of the word is listed which contribute to the large number of entries in such a lexicon. Even when two roots follow the same rule, the present system stores the same information redundantly. Second, it does not show relationship among different roots that have similar word forms. Thus, it fails to represent a linguistic generalization. Third, some languages have a rich and productive morphology like Hindi. The number of word forms might well be infinite in such a case. Clearly, this will not work with such category of languages. As an example, the masculine noun "horse" is written as "घोड़ा" in the direct singular while its oblique singular is "घोड़े" and as a rule it is used in conjunction with post-positions to designate other complements, as in ""“घोड़े को" (dative singular). As for plural forms, the direct case is written as "घोड़े" or the obliquecase as "घोड़_". Hindi adjectives may be either inflected or uninflected. Uninflected adjectives remain unchanged before all nouns and under all circumstances, the same as with English adjectives (e.g., "सुंदर" (beautiful)). All inflected adjectives usually end in '-आ' (e.g., "काला" (black)) and their inflection depends on the gender and case of the noun they alter (e.g., as for the masculine noun "काला घोड़ा" (black horse), "काले घोड़े" (black horses) or with the feminine noun in "काली बिल्ली" (black cat), "काली बिल्लियाँ "(black cats)) [Kellogg 1938]..

II. MORPHOLOGICAL INFORMATION

The inflectional morphology of Hindi does not permit agglutination. This helps keep the number of inflectional morphological rules manageable.However,the derivational suffixes are agglutinative, leading to an explosion in the number of root word-forms in the inflectional root lexicon. As an example, assume that the following sentence pair is part of the bilingual training corpus:

English: Players should just play.
Hindi: f[kykfM;ksa dks dsoy [ksyuk pkfg,A
khilaadiyom ko kevala khelanaa caahie
Hindi (suffix addition is as):
f[kykM ÷;ksa dks dsoy [ksy uk pkfg,A
khilaada iyom ko kevala khela naa caahie

Without using morphology, the system is constrained to the choice of : f[kykfM;ksa (khilaadiyom) for the word players. Also, the general relationship between the oblique case (indicated by the suffix ÷;k (iyom)) and the case marker dks (ko) is not learnt, but only the specific relationship between: f[kykfM;ksa (khilaadiyom) and dks (ko). This indicates the necessity of using morphological information for languages such as Hindi.

Noun Case Marking: In English major elements like subject, object can be usually identified by their positions in sentence, but in Hindi they can be placed anywhere without changing the meaning of the sentence ie. Hindi is very free formed language. For example following two sentences give the same meaning.

1. राम ने गीता को देखा.

2. गीता ने राम को देखा.

So to identify the Ram as a subject and Geeta as a object we required Case Markers. Case Markers can be classified and analyzed are shown in following table. In the absence of case marker the case is called as "Nominative".

Case	Marker	Example
Nominative	Void	He went to market. वह बाजार के लिए चला गया.
Ergative	ने	The boy bought a Pen. लड़के ने एक कलम खरीदा.
Dative	को	He gave the pen to the boy. उसने लड़के को कलम दिया.
Accusative	को	Ram beat him. राम ने उस को मारा.
Ablative	से	The book is written by the boy. लड़के से पुस्तक लिखा गया.
Locative	मे	The student is in the classroom. छात्र कक्षा में है.
Instrumental	से	The girl wrote with the pen. लड़की ने कलम से लिखा.

The division of case markers can be done on the basis of their functions how they are handled. Thus division of case marker is based on Morphological, Structural, and Semantic.

1. *Morphological based division*

The nouns can appear in three forms in Hindi Nominative, Oblique, and Vocative. The default form of Hindi noun is Nominative. If nouns are followed by postpositions then this is case is Oblique. Vocative nouns used to address people(s).The analysis of noun forms are shown in following table.

Nominative	Oblique	Vocative
Boy - लड़का	Boy - लड़के	Boy - लड़के
Girl - लड़की	Girl - लड़की	Girl - लड़की
Boys - लड़के	Boys - लड़कों	Boys- लड़को
Girls-लड़किया	Girls-लड़कियों	Girls-लड़कियो

2. *Structural based Division*

Nouns can have functions like subject, object (direct or indirect) etc. based on the structure of sentence. Analysis of this is given in following table.

Nouns Function	Marker
Subject, Object	Void / से
Subject	ने
Object, Subject indirect object,	को

3. *Semantic based division*

Case markers are assigned as per the gender and the number. This affect the meaning of sentence. Analysis of this is shown in following table.

Gender and Number	Marker
Masculine-Singular	का
Masculine-plural	के
Feminine	की

Postpostions

Pure postpositions are not controlled by verbal predicates so the sentence is complete in meaning with or without them. Case Markers that follow the noun can be handled at syntactic level.

Postpostions	Marker
In	मे
On	पर
Upto	तक
For	के लिए

APPROACH

In our system we are tokenizing words using nltk, then a Stanford parser is called that performs tagging of words and generating relations between them which it stores into outputparser.txt. From the outputparser.txt we are generating tags_output.txt, through which these relations we are storing in mysql database table named relation_tab.An eng lemma of words is obtained by calling function lemmatize onto wordnetlemmatizer object. Using the info obtained till now i.e. eng_lemma, postags, wntag entries are made into dynatab for each of word in eng sentence. After it particularities are handled i.e. if a word is proper noun then its gender is guessed using naive bays classifier and if it a verb then its tense is obtained from tagging.

- Then Word Sense Disambiguation is applied to obtain the right sense word after which inflection addition remains the most crucial step for obtaining a grammatical correct form of translated Hindi sentence.

- While generating inflection we are focusing onto pos_tags of words and accordingly applying inflection onto it..

- proper nouns(NNP) and adjective(JJ) are not inflected and thus hin_lemma remains as it is in hin_inflected.

- for plural noun NNS, we have found out end of word and depending on gender a new end is attached..This new_add is obtained from InflectNounNum where the gender and word_end extracts a particular entry.

- the quality of inflection improves if proper helping verbs are inserted in-between, thus a table of helping verbs is maintained which depending on the tense, gender (masculine, feminine) and multiplicity i.e. singular or plural give the proper helping verbs that are added to already unicoded words .

- if it's a continuous tense then depending on previous word, it is guessing tense and depending on degree ,gender and multiplicity it is finding out the helping verb.

III. MORPHOLOGICAL ANALYSIS AND INFLECTION GENERATOR DATABASE SCHEMA

Following schema describes the tables that we are reffering while performing inflection-

1. DynaTab{id(primarykey), eng_lemma, postag, wntag, gender, tense, mult, nsub, deg_v, hinlemma, hin_gender, hin_deg, hin_inflected}: It stores all information related to each of word in sentence to be translated i.e. root word,part of speech tag, wntag, gender, tense, multilplicity, Inflected word etc.

2. EngLex{pk_engDB, EngWord, POS}: It lists out different english words that we will need while translating an input sentence and their Part of speech tag.

3.FutureTenseVerbs{TenseForm,NarrativePoint,MasSing,MasPlu,FemSing,FemPlu}: It stores future tense verbs, its narrative point, gender and multiplicity.

4.HinLex{fk_HinDB,HindiWord,Gender,NarrativePoint};

5.InflectNounNum {gender,word_end, newadd}: It stores the new end to be attached to a word depending on its current end and gender which will improve the inflection .

6. PastTenseVerbs{gaya, liya, diya}; This table list out different verbs in past tense to which while inflecting the helping verbs gaya,liya,diya are attached. Thus whenever a past tense verb encounters then depending on the coloumn in which it is placed into PastTenseVerbs table a helping verb is attached.

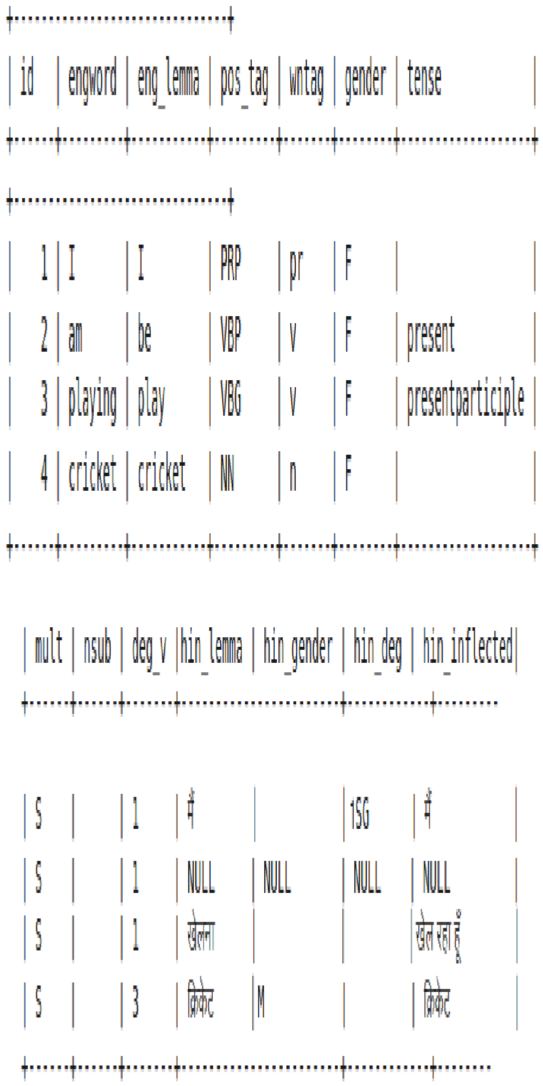
7.PresentTenseVerbs(TenseForm,NarrativePoint,MasSing, MasPlu, FemSing, FemPlu): It enlist different present tense verbs and their tense form, masculinity or feminity and multiplicty.

8. Relation_tab {relation, ind1,ind2}:This table is generated depending on stanford dependencies. Relations between different words in sentence are extracted from output file of Stanford parser and stored into relation_tab.

Following are the examples to explain in detail how these tables are used to generating proper inflection.

1. I am plating cricket

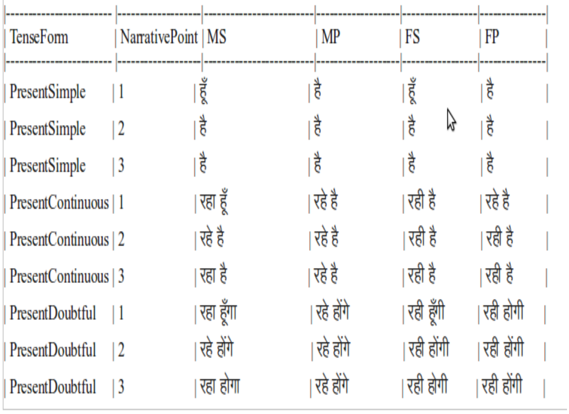
Words in sentence are tokenized using nltk.then tagging is obtained using Stanford parser which also generate relation between these words which is used in generating Dyna_tab entries. the dynatab entries for all words are-



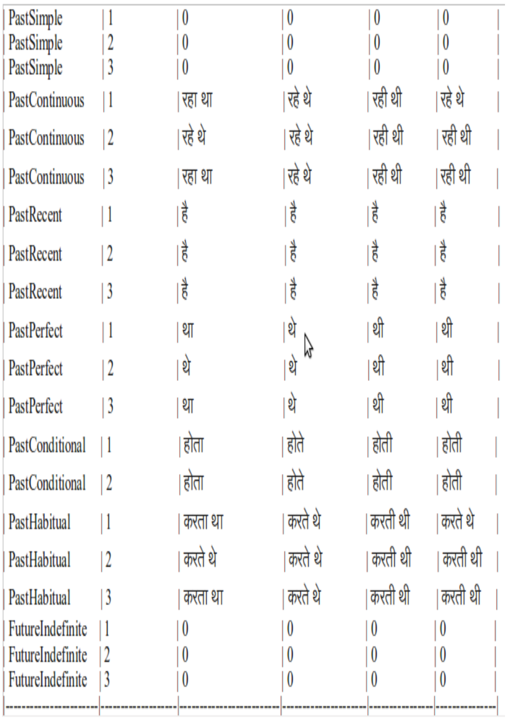
id	engword	eng_lemma	pos_tag	wntag	gender	tense
1	I	I	PRP	pr	F	
2	am	be	VBP	v	F	present
3	playing	play	VBG	v	F	presentparticiple
4	cricket	cricket	NN	n	F	

mult	nsub	deg_v	hin_lemma	hin_gender	hin_deg	hin_inflected
S		1	मैं		1SG	मैं
S		1	NULL	NULL	NULL	NULL
S		1	खेलना			खेल रहा हूँ
S		3	क्रिकेट	M		क्रिकेट

By word-to-word translation playing gets translated to khelalna but to get proper meaning helping verbs are need to be added. Depending on Tense Form, Narrative Point, masculinity or femininity, singularity or plurality helping verbs are found. In current sentence as playing is present continuous the helping verb that is added is 'raha hoo'.and after rearranging in proper SVO format finally generating result-मैं क्रिकेट खेल रहा हूँ| The helping_verbs table that we are referring is as follows



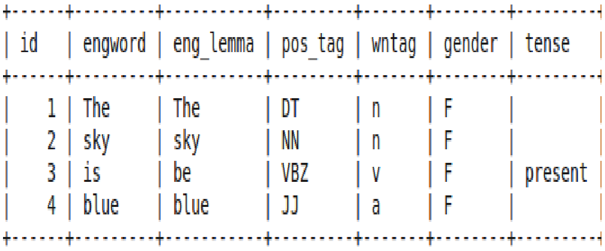
TenseForm	NarrativePoint	MS	MP	FS	FP
PresentSimple	1	हूँ	है	हूँ	है
PresentSimple	2	है	है	है	है
PresentSimple	3	है	है	है	है
PresentContinuous	1	रहा हूँ	रहे है	रही है	रहे है
PresentContinuous	2	रहे है	रहे है	रही है	रही है
PresentContinuous	3	रहा है	रहे है	रही है	रही है
PresentDoubtful	1	रहा हूँगा	रहे होंगे	रही हूँगी	रही होगी
PresentDoubtful	2	रहे होंगे	रहे होंगे	रही होंगी	रही होंगी
PresentDoubtful	3	रहा होगा	रहे होंगे	रही होगी	रही होंगी



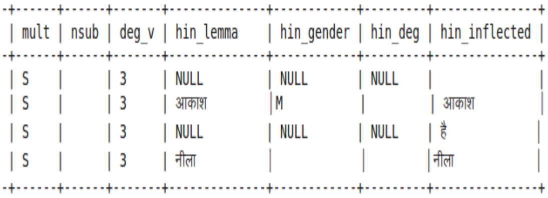
PastSimple	1	0	0	0	0
PastSimple	2	0	0	0	0
PastSimple	3	0	0	0	0
PastContinuous	1	रहा था	रहे थे	रही थी	रहे थे
PastContinuous	2	रहे थे	रहे थे	रही थी	रही थी
PastContinuous	3	रहा था	रहे थे	रही थी	रही थी
PastRecent	1	है	है	है	है
PastRecent	2	है	है	है	है
PastRecent	3	है	है	है	है
PastPerfect	1	था	थे	थी	थी
PastPerfect	2	थे	थे	थी	थी
PastPerfect	3	था	थे	थी	थी
PastConditional	1	होता	होते	होती	होती
PastConditional	2	होता	होते	होती	होती
PastHabitual	1	करता था	करते थे	करती थी	करते थे
PastHabitual	2	करते थे	करते थे	करती थी	करती थी
PastHabitual	3	करता था	करते थे	करती थी	करती थी
FutureIndefinite	1	0	0	0	0
FutureIndefinite	2	0	0	0	0
FutureIndefinite	3	0	0	0	0

2. The sky is blue.

After tokenizing and tagging the dynatab entries are as follow-



id	engword	eng_lemma	pos_tag	wntag	gender	tense
1	The	The	DT	n	F	
2	sky	sky	NN	n	F	
3	is	be	VBZ	v	F	present
4	blue	blue	JJ	a	F	



mult	nsub	deg_v	hin_lemma	hin_gender	hin_deg	hin_inflected
S		3	NULL	NULL	NULL	
S		3	आकाश	M		आकाश
S		3	NULL	NULL	NULL	है
S		3	नीला			नीला

As the sentence is in simple present tense and sky is singular, masculine the helping verb that is added is "hai".As blue is adjective and sky is singular noun they are getting translated as it is generating a final result- आकाश नीला है ।

IV. CONCLUSION

The Hindi morphological analyzer and generator discussed in this paper stores all the commonly used word forms for all Hindi root words in its database. In this paper, we have presented a comprehensive analysis of Hindi morphology and inflection and its implementation in a DM-based Morphological Analyzer. The system was able to completely analyze in most cases accurately. The system failures were driven primarily by external factors. The linguistic analysis that was presented in the earlier part of the paper is economical in descriptive terms. The system takes into consideration most of the tenses and verbs and hence aim to produce accurate results. It can be further expanded to add inflection on adjectives which it at present does not inflect. Also suffix trimming approach to get possible root can result indifferent results. Hence combining this inflection generator along with a Word Sense Disambiguation (WSD) engine to stem the input words and to use their morphological information for sense disambiguation can present a strong case for NLP tools for English to Hindi translation based on well-reasoned and well-argued linguistic analyses.

REFERENCES

[1] Vishal Goyal, Gurpreet Singh Lehal. 2008. Hindi Morphological Analyzer and Generator, pp. 1156–1159.IEEE Computer Society Press, California, USA.

[2] Ananthakrishnan Ramanathan, Hansraj Choudhary Avishek Ghosh. Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT

[3] Ananthakrishnan, R., Bhattacharyya, P., Hegde J.,J., Shah, R. M., and Sasikumar, M., Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation, Proceedings of IJCNLP, 2008.

[4] Niraj Aswani, Robert Gaizauskas. 2010. Developing Morphological Analyzers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages. In Proceedings of LREC.

[5] Bharati, Akshar, Rajeev Sangal and S.M. Bendre (1998b). Some Observations on Corpora of Some Indian Languages. Knowledge Based Computing Systems, Tata McGraw-Hill

[6] Avramidis, E., and Koehn, P., Enriching Morphologically. Poor Languages for Statistical Machine Translation, Proceedings of ACL-08: HLT, 2008.

[7] Bhuvaneshwari C Melinamath, Shubhagini D. 2011. A robust Morphological analyzer to capture Kannada noun Morphology, VOL 13. IPCSIT.

[8] Shapiro, M. 2000. A Primer of Modern Standard Hindi Grammar. Delhi: Motilal Banarsidass Publications.