Hindi है ‘to be’ in Bangla – Issues in Machine Translation

by

Soma Paul

in

ICON

Report No: IIIT/TR/2009/223

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2009
Hindi $hE$ ‘to be’ in Bangla – Issues in Machine Translation

Soma Paul
International Institute of Information Technology
Hyderabad, India
soma@iiit.ac.in

Abstract

The main verb $hE$ ‘to be’ has been used in Hindi in the equative as well as existential sense. It also occurs as a verbalizer in complex predicate. Bangla, on the other hand, is either verbless or has a copula $hOoyaa$ in equational sentence which is distinct from the existential verb $aach$. We observe that the drop of copula in Bangla has a close association with the appearance of classifier on the subject in those contexts. Hindi verb $hE$ ‘to be’ in negative contexts is translated into two different negative verbs. All these factors make the task of automatic translation of Hindi verb $hE$ ‘to be’ to Bangla very challenging. This paper intends to examine syntactic and semantic clues that will determine the translation of $hE$ to the appropriate form ($verbless$ or $hOoyaa$ or $aach$) in a given context in Bangla. A set of transfer rules has also been formulated as part of the transfer grammar module of Setu MT system. The rules transfer Hindi structure with the main verb $hE$ ‘to be’ to the rightful structure in Bangla.

1.0 Introduction

The main verb $hE$ ‘to be’ has both copulative and non-copulative usages in Hindi. The task of translation of these two kinds of $hE$ ‘to be’ in a genetically related language Bangla encounters the following issues:

a. The copula $hE$ ‘to be’ is generally absent in simple present tense in Bangla.

b. However, there are contexts in which the copula has to be mandiatorily present in Bangla even in simple present tense.

c. Bangla uses two different verbs for translating the Hindi verb $hE$ ‘to be’: $hOoyaa$ ‘to be’ in equational sense and $aach$ ‘to be’ for non-copulative use such as existence.

---

1 $hE$, the 2nd person singular non-honorific form (which is also the form for 3rd sg), is taken as a citation form for ‘to be’ in this paper.

2 The verb $hE$ ‘to be’ is also used as an auxiliary in Hindi. In this paper, we are not concerned about the auxiliary use of $hE$ ‘to be’. The phrase “Hindi verb $hE$” is to be understood as “Hindi main verb $hE$” in this paper.

3 They are, however, present in other tenses and also in non-finite form.

4 Most Dravidian languages make a clear cut distinction between existential and equational sentences. In these languages equational sentences are either verbless or have a copula which is distinct from the existential verb. Lakshmi Bai (1986) comments that an examination of modern IA languages with regard to the ‘be’ verb leads to a clear cut division of them into two
There is a close association between the dropping of copula and the usage of classifier in Bangla while Hindi does not have any classifier. These issues are very significant for building a transfer based machine translation system which works in three steps: a) Source language analysis; b) Structural transference and lexical transference; c) Generation of target language words. This paper presents architecture of the transfer grammar module of ‘Hindi to Bangla MT system’. The module is comprised of a set of rules each of which defines the context for mapping Hindi verb hE ‘to be’ to the rightful construction in Bangla.

The next section discusses the semantics of Hindi main verb hE ‘to be’. In section 3, we examine how Hindi verb hE ‘to be’ is translated differently in equational and existential sentences in Bangla. The ‘be’ verb in the context of negation is also discussed in this section as their treatment is different in the two languages under consideration. In section 4, we deal with the correlation between the dropping of copula and appearance of the classifier in Bangla, another factor which needs to be handled through transfer grammar module. This paper does a preliminary study of the aforementioned correlation. Section 5 presents the transfer grammar for translating Hindi verb hE ‘to be’ into Bangla.

2.0 Copulative and non-copulative use of hE ‘to be’ in Hindi

Copula has been defined as a link between the subject of a sentence and a predicate (a subject complement or an adverbial). In Hindi, the copula verb hE ‘to be’ conveys the following equative functions:

a. Identity
1. mE raam huM ‘I am Ram’
   I Ram am

2. kolkattaa se mumbai ki dUrI ‘The distance between Kolkata and Mumbai is 2000 Km’
   Kolkata from Mumbai of distance
   2000 km is

a. Class membership
3. raam DkTər hE ‘Ram is a doctor’
   Ram doctor is

b. Property of the subject
4. raam ecxhaa ləRkaa hE ‘Ram is a good boy’
   Ram good boy is

Copula constructions have been the focus of many studies previously—for example, the many language-specific papers in the series of volumes edited by Verhaar (1967-72), and the more recent semantically-based studies such as Declerck (1988) and Hengeveld (1992), as well as Stassen’s (1997) typological study of intransitive predication, including (some) copula constructions. Lakshmi Bai’s (1986) work gives a detailed typological study of copula and existential verb in Indo-Aryan and Dravidian languages.

Hindi verb hE ‘to be’ has non-copulative usage as well. When the existence of an entity is

d. There is a close association between the dropping of copula and the usage of classifier in Bangla while Hindi does not have any classifier.

geographical groups. On the one hand, we have the Eastern and Southern IA languages such as Nepali, Assamese, Bangla, Odiya, Konkani and Sinhalese which agree with Dravidian languages in maintaining a distinction between existential and equational sentences. On the other hand, we have the western block of languages such as Hindi, Gujarati, Sindhi, Punjabi and Kashmiri which lacks such a distinction.
affirmed by the sentence the verb \( hE \) ‘to be’ is used (see in (6) and (7)). The existence of an abstract relation can also be affirmed as shown in (5).

5. mere saath uskaa ek səmbəndh hE
   I-gen with he-gen one relation is
   ‘He has a relation with me’

6. is kəmre meM do dərwaaže hEM
   this room in two doors is
   ‘There are two doors in this room’

7. mertaa lɔRkɑa səb nepaal meM hE
   My son now Nepal in be
   ‘My son is now in Nepal’

The sentence in (7) expresses the boy’s existence in a location. A very similar construction shown in (8), on the other hand, conveys the sense of “happening of an event in the given location” and not that of “existence”:

8. is bər kɑa film festivəl kolkata meM hE
   This time gen film festival Kolkata in is
   ‘The film festival of this year is in Kolkata’

Thus the Hindi verb \( hE \) ‘to be’ can occur both in a stative sentence as well as in a context that conveys happening of an event. Bangla, however, uses two different strategies to express the semantics of ‘existence’ and ‘happening’. We will discuss this in detail in the next section.

In addition to this, \( hE \) ‘to be’ can also occur in complex predicate as exemplified below:

9. film fesTivəl aj se Suru hE
   Film festival today from begin is
   ‘Film festival begins from today’

Here \( Suru \) ‘begin’ combines with the verbalizer \( hE \) ‘to be’ to form the complex predicate \( Suru hE \) ‘begins’.

For most of the aforementioned cases including the complex predicate construction the Hindi verb \( hE \) ‘to be’ in simple present tense can remain unrealized in Bangla. Nevertheless, a simple rule such as “Drop Hindi main verb \( hE \) in simple present tense in Bangla” does not capture the behavior of the language as a whole. It is because there are contexts when the equivalent form of \( hE \) ‘to be’ has to be mandatorily present in simple present tense even in Bangla. The next section examines those contexts.

3.0 Correspondence of Hindi verb \( hE \) ‘to be’ in Bangla

The non-copulative \( hE \) is translated as \( aach \) ‘to be’ in Bangla in the sense of existence as illustrated below:

10. aamaar Saaathe or ækTaa SƏmbəndh oochrome aache
    I-gen with his one relation is
    ‘He has a relation with me’

11. ei ghOr-e duTo dərjaa aache
    This room-loc two door is
    ‘There are two windows in this room’

12. aamaar chele ækhon nepaale aache
    I-gen son now Nepal is
    ‘My son is now in Nepal’

Interestingly, the presence of verb is optional in (11) and (12). The sentences are grammatical even when the verb \( aach \) ‘to be’ is dropped. This is not, however, the case with the sentence in (10).

---

5 Complex predicate is a very productive construction in Indo-Aryan languages which is composed of a noun or an adjective and a verb.
The equative verb *hE* ‘to be’ of Hindi is “generally” dropped in Bangla as exemplified below:

13. *aami raam* ‘I am Ram’
   I        Ram

14. *raam Daaktaar* ‘Ram is a doctor’
   Ram      doctor

15. *raam bhaalo chele* ‘Ram is a good boy’
   Ram      good    boy

We will discuss here two contexts where the Hindi verb *hE* ‘to be’ in equational sense is not dropped in Bangla even in simple present tense.

First case: When the equative sentence occurs within a conditional type of clause, the copula cannot be dropped in Bangla (compare the sentence in (16) and (17):

16. *himaaləyə yədi bhaarət ka mukuT hE to*
   Himalya if   India gen crown is then
   *himaacəl hE us mukuT teM loʊaat*
   Himachal is that crown in placed
   *səbse durləbh rətnə*
   most rare jewel
   ‘If Himalaya is the crown of India, then Himachal is a rare jewel in that crown.

17. *jodi himaałOy bhaaroṭer mukuT hOy*
   If’ Himalya India-gen crown is
   *tObe himaachOl sei mukuTer ək durlObh rOtno*
   then Himachal that crown-gen one rare jewel
   Second case: When the subject refers to a class (*jati*) and the complement talks about a general property of that class the verb *hOoyaa* ‘to be’ appears in Bangla as illustrated below:

18. *paahaaR-er lok-eraa pəriSrəmi hOy*
   Hill-gen people-pl hard working is
   ‘People from hills are hard working’

19. *maaldaa-r aam miSTi hOy*
   Malda-gen mango sweet is
   ‘Mangoes from Malda are sweet’

For these cases, Hindi uses the habitual aspect form of the verb *hE* ‘to be’ as shown below:

20. *paahaaR ke log pəriSrəmI hoteM hEM*
   Hill gen people hard working be-hab be-3 pl

21. *maaldaa kaa aam miThaa hotaa hE*
   Malda    gen mango sweet    be-3 pr hab

It is noticed that whenever the sentence conveys the semantics of “happening of an event” in a general term, the verb corresponding to Hindi *hE* ‘to be’ surfaces up in Bangla as illustrated below:

22. *bOrSaakaal-e briSTi hOy*
   Monsoon-loc rain    is
   ‘It rains during monsoon’

23. *proti bOchor aamaadeer baaRi-te durgaa pujaa hOy*
   every year    our    house-loc Durga Puja    is
   ‘There happens Durga puja every year in our house’

Hindi in such case uses habitual aspect form of the verb *hE* ‘to be’ (similar to that in (20)):

24. *hər saal həmaare ghər meM dUrga pujaa hotI hE*
   Every year our house in    Durga puja   be-3 pr hab

However when the sense of generality is replaced by a sense of specificity, the verb is dropped in Bangla in the context of happening of an event. For example, when we remove *proti bachor* ‘every
year’ in (23) and instead use the phrase *ei bOchor* ‘this year’ (see in (25)) which brings in a sense of specificity the verb is not required any more:

25. *ei bOchor durgaa pujo aamaader baRi-te*
   This year Durga puja our house-loc
   ‘This year Durga puja is in our house’

In Hindi we use the simple present form of the verb in such case:

26. *is saal hōmaare ghor meM dUrgaa puja hE*
   This year our house-loc Durga puja is

Thus we have one more sense of “be” verb in this section besides “equative” and “existence” and that is “happening of an event”.

At this point I would like to revisit the discussion regarding the ‘existential’ and ‘happening’ reading of the verb “to be” in the context of negation. We have stated earlier that the verb in existential sense as well as in the sense of happening can be dropped in Bangla. The following sentence is legitimate and so is the sentence in (25)

27. *ækhon aamaar chele nepaal-e*
   now my son Nepal-loc

We have no way to determine from the syntax that the sentence in (25) expresses the sense of “happening of an event” and that in (27) conveys existence. The distinction becomes, however, significant for Bangla when the “be” verb is used in negative context. It is because Bangla uses two different negative verbs in these two contexts. Hindi does not make any distinction though. For example, the verbs in negative contexts in the following two Hindi sentences that express “happening of an event” and “existence” respectively is translated into Bangla using two different verbs in (30) and (31):

**Hindi**
28. *is saal durgaa pujaa hōmaare ghor*
   This year Durga puja our house
   *meM nāhī hE, hōmaare taaui ke ghor meM hE*
   in not is, our uncle gen house in is
   ‘Durga puja is not there in our house this year, but it will happen in my uncle’s house’

29. *meraa loDKaa ōbhlı nepaal meM nāhī hE*
   My son now Nepal in not is
   ‘My son is not in Nepal now’

**Bangla:**
30. *ei bOchor durgaa pujo aamaader*
   This year Durga puja our house-loc
   *baaRite nay, aamaar jEThaa-r baRi-te*
   house-loc is not, our uncle-gen house-loc

31. *aamaar chele ækhon nepaal-e nei*
   my son now Nepal-loc is not

Thus in the sense of happening *na ‘not to be’* is used in Bangla and *nei ‘is not’* is used to negate existence.

As far as complex predicate is concerned, Hindi verbalizer is generally dropped in Bangla as shown below:

32. **H:** *film fesTivEl aj se Suru hE*
   Film festival today from begin is
   ‘Film festival begins from today’

**B:** *film fesTivEl aj theke Suru ø*

Let me make a review of the observations that we have made till now in this section:
A Hindi copula in equative sense is dropped in Bangla unless
   i). The subject refers to a class and a generic property of that class is referred to
   ii) The copula is used in conditional clause
In case of (i) and (ii), the verb hOoyaa ‘to be’ is surfaced up

B. Hindi hE ‘to be’ used in the sense of happening is dropped in Bangla unless
   i). it is stated as a general or usual fact
In case of (i), the verb hOoyaa ‘to be’ is used

C. Hindi hE ‘to be’ in existential sense is translated as aach ‘to be’ in Bangla; however it might be optional most of the times.

D. Hindi nəhI hE ‘is not’ in existential sense is translated as nei ‘is not’ in Bangla and as the variant of nay ‘is not’ otherwise.

E. Hindi verbalizer hE ‘to be’ in a complex predicate is dropped in Bangla.

The next section discusses a phenomenon typical of Bangla, namely the use of classifier in the context of copula drop. Hindi does not have classifier.

4.0 Association of copula drop and appearance of classifier in Bangla

We have seen in the previous section that the equative copula can be dropped in Bangla. I agree with Bhattacharya’s (2005) observation that there is an intricate relationship between the absence of copula and the presence of the classifier. He explains that the presence of the classifier triggers the deletion of the copula in Bangla because they perform the same function.

Let us study the following sentences:

33. H: raamu əcchaa lɔRkaa hE
   B: raamu bhaalo chele ø
   Ramu good boy is ‘Ramu is a good boy’

34. H: yəho lɔRkaa əcchaa hE
    B: ei chele-Ta/*chele bhalo ø
    (this) boy good is ‘(This) boy is good’

35. H: kaam əsaan hE
    B: kaaj-Ta SOhoj ø
    work easy is ‘The work is easy’

In (33), the subject is definite entity; in (34) the demonstrative adds a sense of specificity. The sentence in (35) also signifies a particular task and therefore a Bangla speaker adds a classifier to kaaj ‘work’. On the contrary a classifier is not required in the following sentence:

36. taar kaache jekono kaaj sOhoj
    His near any work easy is ‘Any work is easy to him’

A Hindi translation of this sentence is the following:

37. uske lie koi bhl kaam əsaan hE
    his for any emp work easy is

The explicit use of jekono ‘any’ brings in the sense of indefiniteness and therefore a classifier is not required in such cases even in the absence of the copula. Thus the logic that works is the following: If not otherwise mentioned, what English, Hindi-Urdu does by putting something in between two nouns (a copula), Bangla does it by making the subject more definite/specific.
5.0 Transfer Rules for Hindi to Bangla translation

A set of transfer rules is proposed in this section following the strategies of the Transfer Grammar Module of Setu MT system. The architecture of Setu is designed for developing bidirectional MT systems for nine Indian languages; Bangla and Hindi being one of the paired languages.

Transfer component of a machine translation system maps the structure of the source language (SL) to the structure of target language (TL). Let me briefly put here cases we have observed so far in this paper. The objective is to examine for which cases we require transfer rules. The rest of the cases are treated at other levels of which we will discuss here in a nutshell.

<table>
<thead>
<tr>
<th>NO</th>
<th>Hindi</th>
<th>TAM</th>
<th>Sense</th>
<th>Operation on Bangla</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>हई ‘to be’</td>
<td>Present tense</td>
<td>Equative sense</td>
<td>i. Drop copula ii. Add classifier to the subject if it is singular and definite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>हई ‘to be’</td>
<td>Present tense</td>
<td>Happening</td>
<td>i. Drop Copula</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td>हई ‘to be’</td>
<td>Present tense</td>
<td>Existential</td>
<td>i. Translate to आच ‘be’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d.</td>
<td>होता हई</td>
<td>Present tense</td>
<td>General fact</td>
<td>i. Translate to हओआ ‘be’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Habitual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e.</td>
<td>नौहल हई</td>
<td>Present tense</td>
<td>Negation of existence</td>
<td>i. Translate to नौ ‘is not’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f.</td>
<td>नौहल हई</td>
<td>Present tense</td>
<td>Negation of an event or happening</td>
<td>i. Translate to variant of ना ‘to be’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g.</td>
<td>Verbalizer हई ‘to be’</td>
<td>Present tense</td>
<td>Complex predicate</td>
<td>i. Drop the verbalizer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Declarative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table1: Hindi verb हई ‘be’ in affirmative and negative context and its equivalent in Bangla
We can treat the main verb है ‘to be’ in Hindi as polysemous. A WSD\(^6\) tool is required to identify the sense of equation and existence for this verb in affirmative and negative context. Once that is done, the Lexical substitution module can appropriately substitute the word for Bangla. Similarly the TAM substitution tool can substitute hotaa है ‘be+pr hab’ to है ‘is’ in Bangla. Thus the morphosyntactic clue of Hindi has made the task of determining when to retain the equative verb है ‘to be’ in Bangla simpler.

Except for cases in (a), (b) and (g), the rest of the cases are therefore handled by modules such as WSD engine and TAM substitution module. In this section we will design rules for the first two cases as given in table 1. Before I describe the rules the input structure of the SL is discussed in some detail:

### 5.1 Source Language Input

The input structure of SL can have grammatical information of various levels. They can be

- a. POS tagged data with morphological information encoded
- b. Chunked data or shallow parsed data with (a)
- c. Fully parsed data with (b)
- d. Semantic features tagged data with (b) or (c)

We have used the following specifications on the input data:

- Morphological features, b) POS, c) Chunks,

\(^6\) One WSD module has been incorporated in the Setu MT pipeline on the Hindi language analysis side which fetches the correct sense id of a word from the Hindi wordnet

d) Karaka\(^7\) relations (simple parse)

The Chunk Labels used in this system is fewer in number. For the present purpose the three relevant chunks are NP (noun group), JJP (adjective group) and VG (verb group). With every word its POS and a feature structure is associated. The feature structure has the following feature and their value:

- root = “…”
- num = “sg/pl”
- gend = “masc/fem/neut”
- per = “1/2/3”
- tam = “tense/aspect/modality”
- cm = “case marking”
- Drel = “k1/k1s/k2/…”

The value of drels is karaka labels. Presently we have used the information of two karaka relations for writing transfer grammar rules. They are karta (k1) and karta samanadhikarna (k1s). This meaning of the latter tag is that the two nouns having the relation k1 and k1s with respect to a verb have the identical referent. This generally happens when the verb is copula. We have also used a special syntactico-semantic relation called POF (see AnnCorra) for the task. The noun or adjective component of the complex predicate is associated with the verbalizer by POF relation.

The input structure of the sentence “loRkaa ecchaa है” is shown below\(^8\):

\(^7\) As mentioned in AnnCorra (2007) karaka, according to Patanjali, is the one which does something (karotiiti kaarakam). In other words, ‘karakas’ are the roles of various participants in an action. For a noun to hold a karaka relation with a verb, it is important that they (noun and verb) have a direct relation. Panini has spelled out six karakas.

\(^8\) This is a simplified version of the format maintained in writing rules for the transfer grammar engine of Setu MT system. This is done for the sake of easy readability.
The percentage sign (%) separates the word from its POS tag. Every word is bounded within its chunk boundary and the chunk label is given at the start of the boundary. The word, its POS and the feature structure information is given within double parenthesis. The feature structure is given within rectangular bracket. In this sentence raam is in K1 relation and əcchaa is in k1s relation with the copula verb hE.

5.2 Transfer Rules

Each transfer rule is to be seen as a transformation rule that has two parts: a LHS (left hand side) and a RHS (right hand side). The LHS is the SL input structure and the RHS is the TL output structure. We have used the following two operations for the present task:

a. Deletion: A Chunk is deleted
b. Insertion: A classifier is inserted

The rules are as follows:

1. [NP~1 ((NN<drel=k1:1, cm=0>))] [NP*] [NP ((NN< drel=k1s:1 cm=0>))] [NP*] [VG ((VM<root="hE", tam="simple pr", stype=declarative name=1>),)] [NP~1] [NP~2] [NP*] [NP~1] [NP*] [NP~1] [NP*] [NP~1] [NP*]

Rule (1) states that when two nouns are related to the verb hE ‘to be’ with k1 and k1s relations on the LHS, delete the copula on RHS and insert a classifier on k1 marked noun. The VG is removed on RHS to show deletion. The classifier is represented as a feature ‘cl’ with value “Ti”. Rule (2) is identical to rule (1) with the difference that k1s is marked on a JJ within a JJP chunk. The feature structure of the main verb in Rule (1) and (2) contains a feature ‘stype’ whose value declares the sentence type. The value of stype is declared as “declarative”. Thus this set of rules ensures that the copula drop is operative on declarative sentence only. Rule (3) and (4) state that if there exists a noun or an adjective having a POF relation with the following verbalizer, drop the verbalizer. The strategies discussed till now in this section will enable us to handle all cases stated in table 1 except the one in (b). When the source language has the verb hE ‘to be’ and the sentence conveys a sense of happening, the verb is dropped in that context in Bangla. However, it is not an easy task to encode the notion of happening on the input data given the resources we have at our hand. I have not therefore formulated any separate rule for this case in this paper, even though the context for dropping the copula has been correctly identified.
6.0 Conclusion

This paper studies the Hindi verb है ‘to be’ and its corresponding equivalent in Bangla. Bangla is either verbless or have a copula হোয়া (hOoyaa) in equational sentence which is distinct from the existential verb অচ (aach). Three senses of है (hE) is identified in this paper. They are “equational”, “existential” and “happening of an event”. The paper discusses how Hindi है “to be” is translated to Bangla equivalents in all these senses. Finally a set of transfer rule is proposed that encodes the context for dropping the copula and the verbalizer হোয়া (hOoyaa) “to be” in Bangla.

References


