Translation of *TO infinitives* in Anusaaraka Platform: an English Hindi MT system

Akshar Bharati, Sukhada, Soma Paul
Language Technology Research Center
International Institute of Information Technology
Hyderabad, India
sukhada@research.iiit.ac.in
soma@iiit.ac.in

Abstract

In this paper, we study the *infinitive TO* constructions of English which can be variedly translated into Hindi. We observe that there can be different equivalents of *infinitive TO* into Hindi. Factors such as numerous semantic variants in translated equivalents and the syntactic complexity of corresponding English expressions of *infinitive TO* cause great difficulties in the English-Hindi translation. We systematically analyze and describe the phenomenon and propose translation rules for the conversion of English *infinitive TO* into Hindi. The rules have been implemented in the Anusaaraka Platform, an open source English-Hindi Machine Translation tool. The problem has been treated as translation disambiguation of the *infinitive TO*. We examine contexts of *infinitive TO* when it occurs as a dependent of various kinds of main verbs and attempt to discover clues for different translations into Hindi. We achieved a translation accuracy of over 80%. The experiments show that Anusaaraka gives significant improvement in translation quality of *infinitive TO* over Google Translator and Anuvad MT systems.

1 Introduction

We study the *infinitive TO* constructions of English which can be variedly translated into Hindi. The translation of *infinitive TO* in "TO verb" constructions is *-nā* form of the verb which is a *kṛdanta* (participial) form in Hindi, as illustrated in (1) and (2).

(1) I want **to go**.
maiṃ cāha-tā[1] hūṃ **jā-nā**
maiṃ **jā-nā** cāha-tā hūṃ

(2) I prefer **to be** in the woods alone.
maiṃ pasanda_karatā_hūṃ **raha-nā** jaṃgala_meṃ akelā
maiṃ akelā jaṃgala meṃ **rahanā** pasanda karatā hūṃ

However we observe that there can be different equivalents of *infinitive TO* into Hindi. Factors such as numerous semantic variants in translated equivalents and the syntactic complexity of corresponding English expressions of *infinitive TO* cause great difficulties in English-Hindi translation. Therefore an English-Hindi translation software such as Google Translator[2] gives non-satisfactory translations and another MT system Anuvad[3] gives poor translations of *infinitive TO*.

We systematically analyze and describe the problem and propose translation rules for the conversion of English *infinitive TO* into Hindi. The rules have been implemented in Anusaaraka Platform[4], an open source English-Hindi Machine Translation tool. The problem has been treated as word translation disambiguation (WTD) of the *infinitive TO*.

This paper examines the behavior of the main verb on which the *infinitive TO* is a dependent and attempts to discover clues for translation variations in Hindi. We discover some interesting clues for translation as discussed below:

1. Structural Clue: The raising, exceptional case marking (ECM), control verbs license *infinitive TO* as their dependents. Translation of infinitives in the context of these type of verbs

[1]Following conventions for writing gloss are followed in this paper: '-' between morph boundary; '_' between word boundary in case of local word grouping. TO has been consistently glossed as *-nā*. In actual translation layer (3rd layer) the translation of *infinitive TO* construction has been given.

[2]http://translate.google.com

[3]http://nlp.cdacmumbai.in:8081/anuvad/

[4]http://anusaaraka.iiit.ac.in/

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 186–195, Goa, India. December 2014.

systematically vary. English also uses periphrastic compounds with the verb 'get' in the causative sense when we want "to convince someone or trick someone into doing something" (Section 5 discusses all these verb types in detail).

2. Translation Clue: The translation of main verbs determines the translation of its TO infinitival dependent. This presents an important case that shows how the translation of the target language determines the information flow in that language. For example, the English verb *want* in *I want to go home* can have three translation equivalents: (i) *cāhanā*, (ii) *icchā karanā* and (iii) *icchā rakhanā*. If *want* is translated as *cāhanā* in Hindi then *infinitive TO* translates into *-nā*, with other two translations it translates into *-ne kā*[5], as shown in (3).

   (3) I want **to go** home.
   maiṃ ghara **jā-nā** cāhatā hūṃ
   maiṃ ghara **jāne kī** icchā rakhatā hūṃ
   maiṃ ghara **jāne kī** icchā karatā hūṃ

3. Verb specific semantic Clue: The *-nā* form in Hindi takes different postpositions such as *-ne kā, -ne meṃ, -ne se* and so on. We consider that such variation is typically dependent on the semantics of source language verbs which might be sometimes difficult to formalize in terms of rules or conditions.

This paper explores the possibilities of identifying contexts that will help us predict the translation of *infinitive TO*. For the above mentioned cases we have created rules for translation disambiguation. We understand that there are cases where it is difficult to determine a specific rule for disambiguation because either we do not discover the context or it is difficult to translate the contextual clue into a rule that can be implemented. These cases can be handled through case-based reasoning where a deterministic rule is not available. Thus the WTD module proposed in this paper follows a hybrid approach. We have made an attempt to find out structural and semantic clues in the source language that can help us to predict translation variations. We have generated an output on the basis of the rules we framed. We manaually evaluated 100 such test sentences and achieved 80% accuracy. In comparision with Google Translator and Anuvad, we achieved significant improvement in translation.

The paper is organized as follows. Section 2 presents a brief review of word translation disambiguation. Section 3 gives an overview of the Anusaaraka system which has been used as a translation platform for implementing the translation rules. Section 4 briefly presents insights from Sanskrit grammar for the interpretation of *infinitive TO*. The insights have motivated the design of our rules. Section 5 illustrates different contexts where *TO* construction occur and also presents the translation equivalents in Hindi. Section 6 deliberates on our approach in handling *infinitive TO* Finally Section 7 presents the results.

2 Related Work

Earlier WSD based approaches like the one used in (Chan et al., 2007) assumed that different senses of a word in a source language may have different translations in the target language, depending on the particular meaning of the word in context. Hence, the assumption is that in resolving sense ambiguity on the source side, a WSD system will be able to help an MT system to determine the correct translation of an ambiguous word. However, in the context of translation, word sense disambiguation amounts to selecting the correct target translation which is termed as word translation disambiguation (WTD). This aims to select the best translation(s) given a source word in a context and from a set of target candidates.

In the current predominant paradigm for data driven phrase based statistical machine translation, the task of WTD is not explicitly addressed. Instead the influence of context on word translation probabilities is implicitly encoded in the model both in the phrasal translation pairs learned from parallel texts and stored in the phrase translation table and in the target language model (Bungum and Gambäck, 2011). The assumption is that both phrase table and n-gram language model in a way capture collocation and local dependencies and thus helps to disambiguate a possible translation candidate. (Chan et al., 2007) have made an effort to integrate a state-of-the-art WSD system into a state-of-the-art hierarchical phrase-based MT system, Hiero. They show that integrating a WSD

[5] In case the following word is a feminine, the postposition *kā* takes feminine gender and becomes *kī*.

system improves the performance of a state-of-the-art statistical MT system on an actual translation task. For their WSD classifier they select a window of three words (w_{-1}, w, w_{+1}), where w is the word to be disambiguated. One potential problem of such approach is that the amount of context taken into account is rather small. It is clear that WTD often depends on cues from a wider textual context, for instance, elsewhere in the same sentence, paragraph of the document as a whole. This is beyond the scope of most phrase-based MT approaches which work with relatively small phrases.

(Li and Li, 2004) propose a bilingual bootstrapping (BB) approach to disambiguate words to be translated. This approach does not require parallel corpora. Instead they make use of a small amount of classified data and a large amount of unclassified data in both the source and the target language in translation. It repeatedly constructs classifiers by classifying data in each of the languages and by exchanging information regarding the classified data between the two languages (Li and Li, 2004).

(Bharati et al., 2005) have made an attempt to disambiguate English *infinitive TO* from the MT perspective. They have devised rules for translating *infinitive TO* in Hindi. They analyze the phenomena which are discussed in Pāṇini's Aṣṭādhyaī for Sanskrit language. They missed the cases where a verb along with the dependent "TO VERB" translates into one verb unit in target language, such as causativization (see Section 5.4) and the cases where the "infinitive TO" marks subjunctive mood in TL as shown in "Rule 6" in Section 6.2.

3 Anusaaraka as an MT platform

Anusaaraka, a machine translation cum language accessor system, is a unique approach to develop machine translation system based on the insights of information dynamics from Paninian Grammar Formalism. The major goals of the system as stated in (Chaudhury et al., 2010) are the following:

- Reduce the language barrier by felicitating access from one language to others.
- Demonstrate the practical usability of the Indian traditional grammatical system in the modern context.

He	seems	to	be	intelligent.
वह{पु.}	प्रतीत~होना~{@s}	@to{->को[की~ओर]/ना}	होना{0}	बुद्धिमान^सुबोध[-].
वह	प्रतीत_हो{ता_है}	-	-	बुद्धिमान.
वह	प्रतीत होता है	-	-	बुद्धिमान.
वह	प्रतीत होता है	-	-	बुद्धिमान.

Figure 1: Anusaaraka interface showing output for the sentence *He seems to be intelligent.*

- Provide a free and open source machine translation platform for Indian language.

The Anusaaraka system prefers faithful representation of information to naturalness of translation because it aims at no loss of information. In order to achieve that it has designed a special graphical interface as shown in Fig. 1:

The layered output represented by this interface provides an access to all the stages of translation making the whole process transparent. For instance the output in Fig. 1 shows that the infinitival verb group *to be* can be translated as *honā* in isolation as it is clear in the initial layer. But it is dropped in the final Hindi output as shown in the final layers. Thus Anusaaraka provides a "Robust Fall Back Mechanism" which ensures a safety net by providing a "padasutra layer[6] ", which is a word to word translation represented in special formulatic form, representing various senses of the source language word. Users get opportunity to select one of the senses and continue reading the source text with better comprehension.

One of the unique ideas of Anusaaraka system is to utilize human intervention from the earlier stage of development of the system. It talks about a need for sharing the load between man and machine. Machines are equipped with large memory storage, they can "remember" large quantities of information. Humans are good at interpretation.

4 Insights from Sanskrit Grammar

Most of the *infinitive TO* verb constructions in English correspond to the *kṛt* (non-finite) suffix, *tumun* (*tum*) in Sanskrit. According to Sanskrit grammar, a word ending in a *kṛt* affix, where the *kṛt* affix ends in the letter *m*, is designated as an *avyaya* (indeclinable) (A. 1-1-39). Patanjali

[6]The concept of padasutra assumes that polysemous words have a "core meaning" and other meanings are natural extension of that meaning. In Anusaaraka, an attempt is made to relate all these meanings and show their relationship by means of a formula. This formula is termed as padasutra.

says the meaning of the affix *tumun* is *bhāva* (action)[7] (Patanjali, 1975).

Another law *avyayakṛto bhaāve* says that the *kṛt* affixes which are also *avyaya*s denote *bhāva* (action). In English and Hindi *bhāva* is denoted by *to* and *-nā* affixes respectively. Ex.
Eng: 'to go', 'to read', 'to eat', 'to dance', 'to be', 'to feed' etc.
Hnd: *'jānā'*, *'padḥanā'*, *'khānā'*, *'nācanā'*, *'honā'*, *'khilānā'* etc.
The 'infinitive TO' forms of a verb in English seem to be indeclinable as these forms do not take any affixes further. In Hindi, though the affix denotes *bhāva*, it is not indeclinable. Hence the words ending in the affix *-nā* can take zero or some postposition like *'kā', 'ke liye'* etc. So, the 'to' in Hindi is translated as '*-nā* *' where '*' denotes zero or a postposition like *kā, se, meṃ, ke liye* etc.

5 Contexts of *infinitive TO* and their translation equivalents in Hindi

We have focused on the following constructions where *infinitive TO* occurs: the context of *raising*, *control* (subject control and object control) and ECM verbs in English. The examples of each case are illustrated below with their Hindi equivalents. We attempt to identify contexts that might account for the translation variations of these constructions into Hindi. Both raising and control verbs take an infinitival complement with 'TO', however they differ in what they take as their subject.

5.1 Raising verbs

Raising verbs are those verbs whose subject is not its logical subject. We notice that the *infinitive TO* is represented in Hindi in two different ways depending on what the infinitive verb is. If the infinitive verb is any verb other than copula, it occurs in its participial form as exemplified below in (4) and (5):

(4) The girl appeared to enjoy the film.
laṛakī{fem} laga{3,pt} **ānanda_uṭhānā**{fem} philma
laṛakī philma kā **ānanda uṭhātī huī** lagī

(5) The boy seems to know everything.
laṛakā laga{3,pr} **jāna-nā**{masc} saba-kucha
laṛakā saba-kucha **jānatā huā** lagatā hai

[7] *tumarthaścha kaḥ? bhāvaḥ.* What is the meaning of the words that end in *tumun* affix? It is *bhāvaḥ* (action) (3.4.26.2)

It is interesting to note that the Hindi equivalent expression corresponding to the *infinitive TO* (as ānanda_uṭhānā in (4) agrees with the subject (laṛakī) of the sentence. We consider this non-finite form to be a *kṛdanta viśeṣaṇa* (adjectival participial) of the subject *laṛakī*. Given this observation, we propose to make a dependency representation of the above case as shown in Fig. 2:

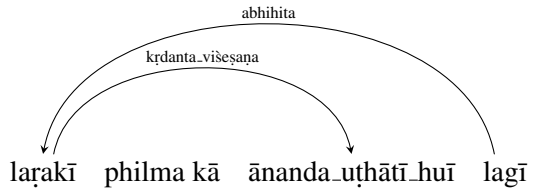

Figure 2: Dependency tree of example (4)

The tree in Fig. 2 represents information better than the one in Fig. 3, which does not account for the feminine marking on *kṛdanta viśeṣaṇa*:

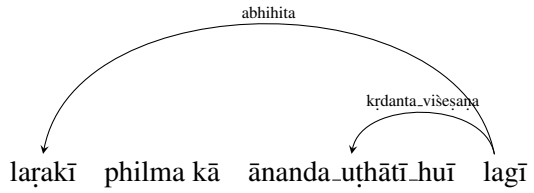

Figure 3: Dependency tree of example (4)

The analysis represented by Fig. 2 correctly predicts the translation equivalent in Hindi and thus can be used as a clue for determining the Hindi equivalents of the English raising verbs 'seem' and 'appear'.

When the *infinitive TO* takes the verb 'be', we note that the infinitives are consistently dropped in Hindi as shown below:

(6) The car proved to be expensive.
gāṛī nikala{3,pt} **ho-nā** mahangī
gāṛī mahangī nikalī

(7) Ram turned out to be a smart guy.
rāma nikala{3,pt} **ho-nā** eka buddhimāna laṛakā
rāma eka buddhimāna laṛakā nikalā

(8) Higher floors tend **to be** hotter.
jyādā uṃcī manjila{pl} jā{3,pr} **ho-nā** garama{comp_degree}
jyādā uṃcī manjileṃ jyādā garama hotī haiṃ

(9) The boy seems **to be** intelligent.
laṛakā laga{3,pr} **ho-nā** buddhimāna
laṛakā buddhimāna lagatā hai

The syntactic analysis of these sentences are same as the one given in Fig. 4. For example, the translation equivalent of the sentences from (6)-(9) will have the following dependency analysis:

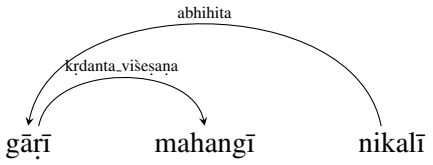

Figure 4: Dependency tree of example (6)

Aspectual and modal verbs of English have also been treated as raising verbs (Taylor, 2006). The verbs with *infinitive TO* are consistently translated into *-nā* form in Hindi as shown below:

(10) Mohan began **to feel** useless.
mohana ṣuru kara{3,pt} **mahasusa karanā** bekāra
mohana ne bekāra **mahasūsa karanā** ṣuru kiyā

(11) She will continue **to do** the work.
vaha jārī rakha{3,ft,fem} **kara-nā** kāma
vaha kāma **karanā** jārī rakhegī

(12) This ought **to be** a very good moment for him.
yaha cāhiye **ho-nā** eka bahuta acchā kṣaṇa liye usake
yaha usake liye eka bahuta acchā kṣaṇa **honā** cāhiye

5.2 Control verbs

Control verbs are the verbs which share one of its arguments with that of the *infinitive TO* argument. When the subject is shared, those verbs are called subject control verbs. We note that the translation of *infinitive TO* in the context of subject control verb is always into *-nā kṛdanta* form. However, different postpositions can occur with the kṛdanta form depending on the semantics of the main verb of which the *infinitive TO* is an object:

(13) He forgot **to tell** you something.
vaha bhūla jā{3,pt} **batā-nā** āpako kucha
vaha āpako kucha **batānā** bhūla gayā

(14) I hate **to say** this.
maiṃ nāpasanda kara{3,pr} **kaha-nā** yaha
maiṃ yaha kahanā nāpasanda karatā hūṃ

(15) He is presently attempting **to do** the translation work.
vaha rahā hai abhī prayāsa kara **kara-nā** anuvāda kārya
vaha abhī anuvāda kārya **karane kā** prayāsa kara rahā hai

(16) He decided **to take** a nap on the sofa.
usane phaisalā kara{3,pt} **le-nā** jhapakī para sophā
usane sophe para jhapkī **lene kā** phaisalā kiyā

(17) He managed **to get** home on Sunday
vaha kāmayāba raha{3,pt} **ā-nā** ghara para ravivāra
vaha ravivāra ko ghara **āne meṃ** kāmayāba rahā

(18) They failed **to make** remarkable discoveries.
ve asaphala raha{3,pt,pl} **kara-nā** ullekhanīya khoja
ve ullekhanīya khoja **karane meṃ** asaphala rahe

The Hindi correspondence of the *infinitive TO* in (13) and (14) is kṛdanta form *-nā*; this form occurs in its ṣaṣṭhī (6th case maker) variant (-ne[8] kā) in (15) and (16) and saptamī (7th case marker) variant (-ne meṃ) in (17) and (18).

When the infinitive TO is not an argument of the subject control verbs, it conveys a sense of "purpose". In Hindi the postposition *ke liye* expresses the semantics of purpose.

(19) She moved **to stand** behind Fiona.
vaha kadama badhā{3,pt} **khadā ho-nā** pīche Phionā
usane Phionā ke pīche **khadā hone ke liye** kadama baṛhāye

(20) Dad is negotiating **to sell** his property.
pitā bātacīta kara{3,pr_cont} **beca-nā** vaha{gen,fem} sampatti
pitā usakī sampatti **becane ke liye** bātacīta kara rahe haiṃ

[8]"-ne" is the oblique form of the suffix *-nā* which appears when it is followed by postpositions.

(21) The staff bribed police **to get** information on politicians.
karmacāri{pl} rishvata_de{3,pt} pulisa_ko **prāpta_kara-nā** sūcanā para rājanītijñoṃ
karmacāriyoṃ ne rājanītijñoṃ para sūcanā **prāpta karane ke liye** pulisa ko riṡvata dī

In case of object control verb the object of the main verb and the subject of the embedded *infinitive TO* verb are co-indexed. We note that the Hindi equivalent of *infinitive TO* in the context of object control verb is mainly *-ne ke liye* as exemplified below:

(22) We ask students **to write** something about themselves.
hama kaha{3,pr} vidyārthiyoṃ **likha-nā** kucha bāre meṃ khuda
hama vidyārthiyoṃ se khuda ke bāre meṃ kucha **likha-ne_ke_liye** kahate_haiṃ

(23) New rules push members to share more information about themselves.
naye niyama{pl} bādhya_kara{3,pr,pl} sadasya-pl **sāṃjhā_kara-nā** aura adhika jānakārī bāre meṃ khuda
naye niyama sadasyoṃ ko khuda ke bāre meṃ aura adhika jānakārī **sāṃjhā karane ke liye** bādhya karate haiṃ

We understand from the aforementioned discussion that *infinitive TO* is translated into kṛdanta form in Hindi. It appears that the selection of postpositions in different contexts depends on the semantics of the control verb. Similar observation is made in (Bharati et al., 2005).

5.3 Exceptional Case Marking verbs

In English, there are verbs which assign accusative case to nouns which are not its argument but the argument of the embedded *infinitive TO* constructions. Such constructions are very differently translated in Hindi as shown below:

(24) I want the students **to go**.
maiṃ cāha{1,pr} vidyārthī **jā-nā**
meṃ cāhatā huṃ ki vidyārthī **jāyeṃ**

(25) We need volunteers **to serve** as medical assistants.
hama_ko jarurata_hai svayaṃsevaka **sevā_kara-nā** ke_rūpa_me auṣadhīy sahāyaka
hameṃ jarurata hai kī svayaṃsevaka auṣadhīya sahāyaka ke rūpa me **sevā kareṃ**.

In (24) and (25), the *infinitive TO* is translated as a clause with subjunctive form of the verb. However we notice that ECM verbs can be variedly translated in Hindi for which no immediate contextual clue is available.

5.4 Causative periphrastic compound

English causative construction is periphrastic in nature where the grammatical meaning is distributed among more than one words. One causative construction in English uses 'get' as exemplified below:

(26) They got me **to talk** to the police.
ve{nom} prāptā_kara{3,pt} maiṃ{acc} **bāta_kara-nā** se - pulisa
unhoṃne merī pulisa se **bāta karavāyī**

(27) I got the mechanic **to check** the brakes of my car.
meṃ prāpta_kara{3,pt} kārīgara **jāṃca_kara-nā** breka kā merī kāra
maiṃne kārīgara se merī kāra ke breka kī **jāṃca karavāyī**

This form of causative construction is used when we want to convince someone or trick someone into doing something. Such construction is systematically translated into causative form of the embedded verb with the drop of equivalent of 'get' in Hindi.

6 Our Approach to WTD

We have distributed the task of WTD in two parts in consonance with the observation made in (Kulkarni, 2003) in the context of design and development of Anusaaraka system:

1. A need to share load between man and machine.

2. Distinguish reliable knowledge from heuristics.

We often come across ambiguous cases where it is difficult to state the choice of a particular target translation for a word in terms of certain conditions from the context. This is so because the

information is not easily logically available in the context, but is rather distributed hence difficult to tap through certain conditions. Therefore, we propose to handle the WTD task of *infinitive TO* at two levels:

1. Rule based approach
2. Case based reasoning approach

1. Rule based approach: In order to handle logical type of cases, linguistic knowledge is represented in terms of rules. The discussion in Section 5 guides us to formulate rules and implement them. When number of rules increase, maintenance of rules becomes important in the sense that no rule should clash with any other rule and the syntactic format of the rules should be correct. The use of expert system CLIPS[9] for the rule writing makes the task simple. While making the rules, the developer is also requested to give at least one example English sentence with its translation for which the rule is written. Such an effort also helps in growing the parallel corpora.

2. Case based reasoning approach: We have identified cases where it is difficult to identify context which can be used as conditions in the rules. For example, the discussion in Section 5 illustrates that *-nā* kṛdanta form occurs with different postpositions such as *-ne kā, -ne meṃ, -ne ke liye* and so on while translating *infinitive TO* in the context of control verbs the semantics of individual verbs might give us clue for selecting the right postposition in a given case. But specifying that semantics in concrete fact is not easy. Also, we noted that the *infinitive TO* in the context of ECM verbs can be translated in various ways. For such cases, we have decided to adopt the case based reasoning option. We will develop translation copora for such cases and use machine learning technique for learning the correct translation automatically. However, further discussion on this approach is beyond the scope of this paper.

6.1 Data Preparation

We have taken the list of ECM, control and raising verbs from Treebank IIa Guidelines[10]. The guidelines have 31 ECM, 34 raising verbs, 99 subject control verbs, 52 object control verbs and 34 raising verbs. We extracted sentences for these verbs from COCA[11] (Zhou and McKinley, 2005). Then the sentences were simplified as and when required and were manually translated into Hindi. We observed the patterns of translation from these translated pairs of sentences.

6.2 Formulation of Rules

Rule 1. The 'to' in 'infinitive TO' constructions translates into *nā* in Hindi if it occurs as an infinitival predicate of the following verbs when they have a PRO embedded subject, (with an embedded subject they will follow "Rule 6"): apt, begin, choose, continue, end, fail, figure, forget, happen, hate, keep, learn, like, love, need, ought, prefer, prove, quit, remain, start, stop, tend, want and wind. Ex.

(28) a. Jennifer began **to take** precautions.
Jeniphar ṡuru_kara{3,pt} **barata-nā** sāvadhanī{fem}
Jeniphar ne sāvadhanī **baratanā** *ṡuru kiyā*

b. He chose **to go** into teaching.
vaha{masc} cuna{3,pt} **jā-nā** meṃ ṡikṣaṇa
usane ṡikṣaṇa meṃ **jā-nā** cunā

Rule 2. If the 'infinitive TO' constructions are arguments of the verbs 'appear' or 'seem' then 'to' translates into 'verb + -tā huā' in Hindi. Ex.

(29) a. It appears **to move**.
yaha laga{3,pr} **cala-nā**
yaha **calatā huā** lagatā hai

b. She appeared **to enjoy** it.
vaha{3,fem} laga{3,pt,fem}
ānanda_uṭhā-nā yaha{acc}
vaha isakā **ānanda uṭhātī huī** lagī

Rule 3. If "infinitive TO" verb is an argument of a verb that translates into a conjunct verb and the first part of the verb is a noun as in *phaisalā kara,*

[9] http://clipsrules.sourceforge.net/

[10] Treebank IIa is the annotation style used in the English Treebank being created as part of the OntoNotes Project (DARPA GALE). It is based on the original Penn Treebank II Style (Taylor, 2006). http://www-users.york.ac.uk/~lang22/TB2a_Guidelines.htm

[11] COCA (Corpus of Contemporary American English) is the largest freely-available corpus of English. It contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. It allows limit searches by frequency and compare the frequency of words, phrases, and grammatical constructions. http://corpus.byu.edu/coca/

niścaya kara, ānanda uṭhā, āśā kara, paravāha kara, lakṣya rakha, anumati de etc. in Hindi then it is translated as '-ne kā'. Ex.

(30) a. I decided **to go** ahead.
maiṃ phaisalā_kara 3,pt{**jā-nā** āge
maiṃne āge **jāne kā** phaisalā kiyā
b. We have opted **to take** the research.
hama{1,pl} phaisalā_le **le-nā** - śodhakārya
hamane śodhakārya **lene kā** phaisalā liyā hai

Exception to this rule:

(31) She declined **to comment**.
vaha{1,sg,fem} manā_kara **tippaṇī kara-nā**
usane **tippaṇī karane se** manā kiyā

Rule 4.: If the 'infinitive TO' constructions are 'to BE' where 'BE' occurs as a 'copula' verb then 'to BE' is dropped while translating it into Hindi. Ex.

(32) a. The car proved **to be** expensive.
- kāra sābita_ho{3,pl} *ho-nā* mahaṃgā{fem}
kāra mahaṃgī sābita huī
b. The number of inputs is assumed **to be** two.
- saṃkhyā kā{fem} inaputa hai māna{1,sg,passive} **ho-nā** do
inaputa kī saṃkhyā do mānī gayī hai

Rule 5.: English uses MAKE, HAVE and GET verbs for causativization, whereas Hindi uses *-ā* and *-vā* suffixes to the root to represent direct and indirect causation respectively (Ramchand, 2008). The pattern *GET + animate + to + Verb* marks causatives in English. For example in (33-a) the main verb 'got' and to-infinitive 'to paint' form a causative verb. Hence we group these verbs together and causativize them in Hindi.

(33) a. I **got** the boy **to paint** my house.
maiṃ{nom} prāpta_kara{pt} - ladakā **raṃga-nā** merā ghara
maine ladake se merā ghara **raṃgavāyā**
b. They **got** me **to talk** to the police.
ve{nom} prāpta_kara{pt} merī **bāta kara-nā** se - pulisa
unhoṃne pulisa se merī **bāta karavāī**

Rule 6. The 'TO infinitive dependent' of some verbs gets transferred into subjunctive clause in Hindi. Some verbs in this category are *command, demand, insist, order, recommend, suggest, want* and *wish*.

(34) I want him **to go**.
maiṃ usako* **jānā *cāhatā hūṃ*
maiṃ usakā* **jānā *cāhatā hūṃ*
maiṃ cāhatā hūṃ ki vaha **jāye**

Rule 7. By default the 'infinitive TO' constructions translate into 'verb + *-ne ke liye*' in Hindi. Ex.

(35) a. 7000 people turned out **to see** him.
7000 loga ā{3,pl,pt} - **dekha-nā** use
7000 loga use **dekhane ke liye** āye.
b. You were elected **to do** something.
āpa the cuna{2,pt} **kara-nā** kucha
āpa kucha **karane ke liye** cune gaye the

7 Results and Error Analysis

We randomly picked 100 sentences from COCA for testing the rules. We ran three translation systems Anusaaraka, Google and Anuvad on these 100 test sentences. Three evaluators evaluated the output of the systems for their accuracy. Accuracy was measured on a scale of 0-2; 0 being incomprehensible, 2 being comprehensible and 1 comprehensible with some effort. Generally when the output is not grammatical but the reader can comprehend the meaning from the output, the score 1 was given for such cases. Table 1 reports the results.

From Table 1, we observe that the performance of Anusaaraka is distinctly better than the two other systems.

	Anusaaraka	Google	Anuvad
Correct Translation	80	70	46
Incorrect Translation	20	30	54
Accuracy	**80%**	70%	46%

Table 1: Anusaaraka accuracy results compared with other MT systems.

We also compared the output of Anusaaraka with revised rules with the performance of the older version of Anusaaraka where the default

translation of TO infinitive was given as *-ne ke liye*. We observed a distinct improvement of the system when we implemented our rules as shown in Table 2:

	Without Formulated Rules	With Formulated Rules
Correct Translation	50	80
Incorrect Translation	50	20
Accuracy	50%	**80%**

Table 2: Anusaaraka accuracy results before and after application of the *to-infinitive* rules.

We categorized the verbs which the TO infinitive is a dependent of into different verb types and examined the performance of Anusaaraka for each type of verb class.

Verb Type	Total	Correct	Accuracy
Aspectual	12	9	75%
Causative	4	4	100%
ECM	13	12	92%
Object control	25	20	80%
Raising	9	4	44%
Subject control	37	31	83%

Table 3: Accuracy results for various type of verbs present in the test set.

We observe from Table 3 that the TO infinitive with *Raising* type of verbs have mostly been incorrectly translated. The errors in various types of verb translations can be classified as follows:

1. Parser Error: Sometimes, the 'TO' is tagged as preposition and the parser inadvertently considers the *infinitive TO* as preposition and as a consequence the whole parse goes wrong. For example, infinitive TO (in bold characters) has been wrongly projected as a prepositional phrase (PP) for the following sentence: *I am going* **to direct people** *to read your writings at our website*.

2. For rule 3, it is important that our conjunct verb list be exhaustive. If a conjunct verb is not identified while translation, this rule will not fire and the translation of *TO infinitive* will be incorrect. For example, in (36), the word 'advise' is translated as *sujhāva denā* in Hindi. Since we do not have that conjunct verb present in the list, hence, the *TO infinitive* "to pay" was translated as *dhyāna de-ne ke liye* while it should have been translated as *dhyāna de-ne kā*:

   (36) I have advised them **to pay** attention to their intuition.
   maiṃ sujhāva de{1,pt} unako dhyāna **de-nā** apane antarjñāna ko
   maiṃne unako apane antarjñāna kī ora dhyāna **dene kā** sujhāva diyā

3. Sometimes, a specific verb of a verb class has a very different behavior and therefore they cannot be handled with rules. For example the raising verb 'happen' with its dependent TO infinitive is translated into different constructions into Hindi such as:

   (37) a. He **happened to see** the article.
   vaha ho{pt} **dekh-nā** - lekha usakī lekha para **najara padī**
   b. I **happened to go** to the market one Saturday.
   meiṃ ho{pt} **jā-nā** ko - bājāra eka ṣanivāra
   merā eka ṣanivāra ko bājāra **jānā huā**
   c. I **happen to disagree** with my husband on a lot of issues.
   meiṃ **ho**{1,pr} **matabheda_ho-nā** ke__sātha merā pati para bahuta sāre viṣayoṃ para
   merā mere pati ke sātha bahuta sāre viṣayoṃ para **matabheda rahatā hai**

   We observe that the word 'happen' is not a straightforward case to translate into Hindi. At present, our system does not handle 'happen to V' constructions.

8 Conclusion

In this paper, we presented the design and implementation of a resource namely WTD rules for disambiguating English *infinitive TO* in the context of English-Hindi machine translation. The results are promising and show that with the use of

contextual knowledge, machine can produce satisfactory translation of English 'infinitive TO' in the context of raising, control, ECM and periphrastic causative constructions. Since availability of these constructions in parallel copora is not always possible, hence, we chose to utilize contextual translation and semantic clues for writing WTD rules. However, we also recognize cases where contextual clue is not available. Thus the method of WTD in this system respects the concept of sharing the work load between man and machine. As future work, we will create parallel corpora for such cases for case base reasoning.

Acknowledgments

The authors are grateful to Prof. Vineet Chaitanya, Dr. Dipti Misra Sharma and Dr. Aditi Mukherjee for having useful discussions on various aspacts of the subject. We also thank Banasthali Vidyapith students, especially Ayushi Agarwal, Shivani Pathak, Anshika Sharma and Prajya Jha for evaluating the test output.

References

Akshar Bharati, R Vaishnavi Rao, and AP Tirupati. 2005. WSD of To-Infinitive into Hindi: An Information Based Approach.

Erwin Marsi André Lynum Lars Bungum and Björn Gambäck. 2011. Word Translation Disambiguation without Parallel Texts. *LIHMT 2011*, page 66.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 33. Citeseer.

Sriram Chaudhury, Ankitha Rao, and Dipti M Sharma. 2010. Anusaaraka: An expert system based machine translation system. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, pages 1–6. IEEE.

Amba P Kulkarni. 2003. Design and Architecture of Anusaaraka-An Approach to Machine Translation. *Volume*, 1:Q4.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 30(1):1–22.

Patanjali. 1975. *Patanjali's Vyakarana mahabhasya : with English translation and notes*. Bhandarkar Oriental Research Institute.

Gillian Catriona Ramchand. 2008. *Verb meaning and the lexicon: A first phase syntax*, volume 116. Cambridge University Press.

Ann Taylor. 2006. Treebank 2a guidelines.

Zhinan Zhou and Philip K. McKinley. 2005. COCA: A Contract-Based Infrastructure for Collaborative Quality-of-Service Adaptation. Technical report, July.