A HYBRID APPROACH USING PHRASES AND RULES FOR HINDI TO ENGLISH MACHINE TRANSLATION

Susmita Gupta and Niladri Chatterjee

Department of Mathematics, IIT Delhi, Delhi, India

ABSTRACT

The present work focuses on developing a hybrid approach for developing a machine translation (MT) scheme for automatic translation of Hindi sentences to English. Development of machine translation (MT) systems for Indian languages to English almost invariably suffers from the limited availability of linguistic resources. As a consequence, statistical, rule-based or example-based approaches have not been found to generate good quality translation for all types of sentences. Moreover, purely statistical or example based schemes do not leverage the semantic association (i.e. association of the noun case-endings (karakas)) with each other and also with the verb phrase to generate the English translation.

The scheme proposed in this work is based on the syntactic and semantic roles of the phrases of an input Hindi Sentence. The proposed solution is a hybrid scheme involving phrase-based, rule-based and statistical translation approaches. In our approach we try to identify the difficulties associated with the translation of different types of input sentences, viz. simple declarative, negative and interrogative sentences and propose a translation approach for them.

KEYWORDS

Hybrid MT, Rule-based MT, Example-based MT, Statistical MT, Noun case-ending.

1. INTRODUCTION

Past works on machine translation for Indian languages had mostly focused on translating English sentences to Indian languages. Not much work is found in the reverse direction i.e. from Indian languages to English. The primary reason for this may be attributed to the lack of resources for processing Indian languages. The situation becomes even more difficult because of the lack of existing parallel corpus essential for different translation paradigms, such as, example based machine translation (EBMT), statistical machine translation (SMT), among others. Furthermore, only a handful of parallel corpora are available for Hindi to English as one may find in the OPUS[1] corpus. Consequently, the quality of translation as obtained from various translators available online is not quite up to an acceptable level even for rather simple sentences. Table 1 shows some examples of translation results using Google translator which is one of the most popular online translator available. The table contains the translation of the Hindi sentence to English and the BLEU score representing the quality of the translation. The BLEU metric scores a translation on a scale of 0 to 1, but is frequently displayed as a percentage value. The closer the BLEU score is to 100%, that means the translation is more correlated to a human translation. Technically, the BLEU metric measures how many words overlap in a given translation when compared to a reference translation, giving higher scores to sequential words. It can be seen in Table 1 that the BLEU score obtained for Hindi to English Translator of Google translation are quite low, which can be interpreted also by the quality of translated results to a linguist who has the knowledge of both the languages. Note that the type of sentences been chosen for translation are of a special type (*multi-karaka*). This means that there are multiple occurrences of noun-case endings

[1] http://opus.lingfil.uu.se/

DOI: 10.5121/ijnlc.2017.6305

(*karakas*) in each of the Hindi sentences. The details of noun-case endings (*karakas*) are explained later in this section.

To address the challenges of Hindi to English machine translation and to build a good quality translation system, the inherent differences of the sentence structure between these two languages have to be carefully observed. The structures of English and Hindi sentences vary a lot. Development of a translation system needs considering and remembering these differences for producing a correct translation. The most important of them are mentioned below:

1. *Word Order*: English is classified as an SVO (Subject-Verb-Object) language and Hindi as an SOV (Subject-Object-Verb) language. An SVO language indicates, the position preceding a verb marks the subject and the object follows the verb. On the other hand, Hindi is a relatively free word-order language. For example, the following English sentence "Ram(S) eats (V) fruits (O)" is written as राम फ़ल खाता है (*raam*(S) *phal*(O*) khaataa hai*(V)). However, reordering of the constituent words often does not change the semantics of the sentence. For illustration,

   फ़ल खाता है राम (*phal khaataa hai raam*)
   राम खाता है फ़ल (*raam phal khaataa hai*)

   are also correct Hindi sentences conveying the same sense as the previously mentioned one [8] However, an MT system often fails to capture this equivalence. For illustration consider the Google translation of the two sentences

   चावल खाता है राम (*chaawal khaataa hai raam*),
   राम चावल खाता है (*raam chaawal khaataa hai*).

   Although both convey the same sense which is 'Ram eats rice', Google translates them as: 'Rice account is Ram' and 'Ram eats rice', respectively.

2. *Order of main verb and auxiliary verbs*: The relative order of the main verb and the auxiliary verb in a sentence are reversed in Hindi and English. The difference between Hindi and English verb groups is the order of root verb and its auxiliaries. In Hindi the main root verb is followed by its suffixes, as in जा रहा है (*jaa rahaa hai* ~ going) where *jaa rahaa hai* is the present continuous form of the verb *jaanaa* ~ to go. Here *rahaa hai* is the auxiliary verb that conveys the tense, gender and case information of the verb. However, in English the main verb follows the auxiliary verbs as in "is going".

Table 1. Translation of Hindi to English and vice versa

Hindi Sentence (H)	English Sentence (E)	Google Translations (dated 16th April 2017)	BLEU Score	Comments
राम को कविता की सुन्दर किताब पढ़नी है *(Raam ko Kavita ki sundar kitaab padhni hai)*	Ram wants to read Kavita's beautiful book	Ram has to read a beautiful book of poetry	16.2	"Kavita" is not identified as proper noun by the translator
राम को रोचक की सुन्दर पोशाक पसंद है (*Raam ko Rochak ki sundar poshaak pasand hai*)	Ram likes the beautiful dress of Rochak.	Ram likes to have interesting clothes of interesting	15.4	"Rochak" is not identified as proper noun by the translator

सीता दौड़ते हुए अपने घर गयी *(Sita daudte hue apne ghar gayii)*	Sita ran to her house	Sita ran to her house while running	12.3	Introduction of two verbs "ran" and "running", while only one is needed.
यह पोशाक जो आपको सूरज ने दी थी वह आप पर बहुत खिल रही है (*yah poshaak jo aapko Suraj ne dii thee wah aap par bohut khil rahii hai*)	This dress which Suraj gave you quite suits you	This dress which you gave to the sun is very blooming on you	9.8	"Suraj" is not identified as proper noun by the translator. "Blooming" does not make sense in the context of dress
कविता को नेहा के घर से सीता के घर तक कौन छोडने जायेगा (*Kavita ko Neha ke ghar se Sita ke ghar tak kaun chhorne jayega*)	Who will drop Kavita from Neha's house to Sita's house	Who will leave the poem from the house of Sita to the house of Sita	6.2	"Neha" and "Kavita"; the translator ignores both the proper nouns

3. *Presence of subject in a sentence*: The subject position in English cannot be empty and this leads to a forceful introduction of dummy 'it' or existential 'there' in certain sentences to fill the subject position. Hindi sentence construction does not need any such syntactic requirement. For illustration, consider the sentence "It is raining" ~ बारिश हो रही है (*baarish ho rahii hai*). The translation does not contain any word corresponding to 'It'.

4. *Preposition vs. Postposition:* In English, prepositions come prior to the words to which they relate. In Hindi, such words occur after the nouns they govern. For example, "on the door" translates to दरवाज़े पर (*darvaaze par*). Here *darvaaze* implies 'door', and *par* corresponds to the preposition 'on'. As they occur after the noun, they are often referred to as postpositions.

5. *Order of Verb and Adverb in a sentence:* As in English, in Hindi too adverbs modify verbs and adjectives. However, in English the verbs generally precede the adverbs. Unlike English, in Hindi this order is reverse. For example: "you speak fast" is represented in Hindi as तुम जल्दी बोलते हो (*Tum jaldi bolte ho*), where, *jaldi* implies 'fast', and *bolte ho* is the present indefinite form for second person for the verb 'speak'.

Based on the output from Table 1 and the differences of the two languages, Hindi and English, the basic problems of translations between Hindi to English sentences may be identified as follows:

1. Semantic information is not preserved in some of the translations. This is because based on the context of the sentence, a Hindi verb can be translated in various ways.
   For example: In the Hindi sentence
   फूल खिल रहा है (*phool khil rahaa hai*)
   [Flower is blooming],
   खिल (*khil)* [blooming] should be translated as "blooming". But in the sentence

   यह पोशाक आप पर खिल रही है (*yah poshaak aap par khil rahii hai*)
   [This dress suits you],
   खिल (*khil)* [blooming] should be translated as "suits".

2. Syntactically some translations are incorrect. The reason for this can be because the grammatical differences mentioned above is not considered in the statistical and the dictionary based schemes of translation.

3. In some cases, a few source language words are not translated at all in the target language. For example, the Google translator does not translate "suiting" at all and places the word as is in the translation.

The problems become even moreworse for the case when the source language (Hindi) has multiple noun case-endings (*karakas*) scenarios. For example, consider the following two Hindi sentences, which are structurally same but semantically different.

1. राम ने मेरे लिए सोने की घड़ी ली (*raam ne mere liye sone ki ghari lii*) [Ram took a gold clock for me]
2. राम ने मेरे लिए श्याम की घड़ी ली (*raam ne mere liye shyaam ki ghari li*) [Ram took Shyam's watch for me]

In the first sentence, the sense that the noun case-ending की(*kii)* conveys is that Ram took a watch "made of" gold, whereas in the second sentence की(*kii*) means that the watch "belongs to" Shyam. But for existing translators (eg Google) the statements fail to preserve the semantic information of the sentences as follows:

1. Ram took a gold clock for me
2. Ram took a look at Shyam's clock for me

which is clearly not preserving the semantics for the second sentence.

It is observed that the translation quality of statistical techniques [17] suffers because of the inherent property of the scheme, which is to generate the translation based on the probability computation from existing parallel corpora. Statistical schemes typically do not consider the syntactic or semantic information of the source and target languages. It is clear from the outputs given in Table 1 that the translations are neither faithful to the source sentence semantics nor fluent in the target language. More recently, syntax has been incorporated in SMT in both the source and the target, e.g. Syntax Augmented Machine Translation (SAMT) [27], Hierarchical Phrase-Based SMT (HPBSMT) [10]. Also, different phrase-based SMT schemes are increasingly looking forward towards incorporating semantic information into the translation. For example, Combinatory Categorial Grammar (CCG) augmented hierarchical phrase-based machine translation [1], CCG supertags in factored statistical machine translation [7].

Rule-based schemes also do not fit well in the purview for translation of Indian languages to English. This is because of the huge differences in the sentence structure of the two languages which leads to the need of a large set of translation rules to be developed. The sentence structure of Hindi as against English is free word order, which makes the task of translation rule creation a huge human effort.

The translations in example-based [24] schemes depend solely on the examples in the available corpus. A strong parallel corpus covering variety of sentence types is not widely available for Indian languages. Also, divergence between the two languages results in the lesser availability of useful translation examples even if similar structure input sentences appear in the corpus. Divergence [14] occurs when structurally similar sentences of the source language do not translate into sentences that are similar in structures in the target language. For illustration, consider the translations of the following English sentences:

It is running ~ यह दौड़ रहा है (*yah daud rahaa hai*)

It is raining ~ बारिश हो रही है (*baarish ho rahii hai*)

The above drawbacks motivated us towards developing a hybrid translation scheme that can take the advantages of each of the above schemes, and generate improved Hindi to English translation system.

The paper is organized as follows. Section 2 describes some of the previous works in the direction of translation involving Indian languages. This includes the work done in Indian

language translations for the case of multiple noun case-endings (*karakas*). The proposed translation approach is explained in Section 3, which is followed by Section 4 explaining the experiments conducted, and shares the translation results. Finally, Section 5 concludes the paper, by proposing few areas of future work.

2. PREVIOUS WORK

Over the last two decades' machine translation using Indian languages made some significant progress. Perhaps the most notable among them is the AnglaBharati system [21] which uses rules developed based on structural patterns to translate English sentences into different Indian languages. However, several other systems have been developed prior to it, following different translation paradigms. The most notable among them are mentioned below.

Anusaaraka [5], a translation system for mutual translation between different Indian languages is developed using the "direct translation" approach maintaining the grammatical structure of the source language. Hence the output is often syntactically incorrect when the source and target language grammars differ. The Mantra machine translation system [2] was developed using the transfer based approach. Here Tree Adjoining Grammar based taggers and parsers are used for transferring English text structure to Hindi to achieve translation. Anubharati [15] followed a hybrid approach where example based scheme was aided by statistical analysis corpora and linguistic rules. Another EBMT system Anubaad [3] also followed EBMT approach for translating English news sentences into Hindi. More recently many other systems have been made available for translating between various Indian languages. Some of them are: Punjabi to Hindi MT system [16], Hindi to Punjabi MT [13], UNL-Based systems for translating from English to Hindi, Bengali and Marathi [9], EBMT systems for translating between Hindi, Kannada and Tamil [2]. Barring the Hinglish translation system [22] none of the above translation schemes attempted to translate Hindi to English. The existing schemes focus mostly in translating English to Indian language, or between pairs of Indian languages.

Very recently in WMT14 [12] shared task, attempt has been done to develop statistical systems for Hindi to English, and English to Hindi translation. The core components of the translation systems are Phrase Based (Hindi- English) and factor based (English to Hindi). The work focuses on the usage of number, case, and Tree Adjoining Grammar information for translation. There is also a pre and post processing step involved to adjust the translation output for structural divergence between Hindi and English sentences. Also, there are a very limited set of rules which play the role in translating Hindi to English. Another translation system was developed by [26] in WMT2014 where apart from various other language pairs, Hindi - English translation has also been considered. This is based on string to tree system, where some improvements were done for out of vocabulary word translation through transliteration.

Apart from the WMT shared task translation systems [18] presented a hybrid machine translation architecture guided by syntax. In this work, the authors develop a hybrid translation system guided by the rule based machine translation engine and, before transference, few partial candidate translations given by the statistical translation schemes. These partial candidate translations are used to enrich the tree-based representation. The final hybrid translation is created by choosing the most probable combination among the available subsets with a statistical decoder.

[1] https://mantra-rajbhasha.rb-aai.in/

In the works described above, the syntactical and semantic information is not utilized completely to generate the translations. None of the translation approaches provide a complete set of translation rules guided by syntactical information. Further, there has not been any special focus on translating negative, and interrogative sentences which clearly has a different sentence structure, and therefore different translation rules apply for them as against simple positive sentences.

The present work focuses on Hindi to English translation. In our approach we focus on the structural/syntactic representation and semantic information to develop a strong translation paradigm. The structural information is captured by considering the syntactic phrases as units for translation. The semantic information obtained from the कारक (*karaka*) (case-ending) is used to generate the translation output. The details are explained in Section 3.

The evaluation of the translation output is done using the BLEU score. Although there are various metrics currently available for evaluating machine translation [28], the current work uses BLEU, as it is the most widely used evaluation metric till date. [29]

3. PROPOSED APPROACH

The translation approach followed in this work is a hybrid of phrase-based, rule-based and statistical approaches of machine translation. The strengths of each of the following approaches are used, to generate a hybrid scheme for machine translation.

3.1. *Phrase-based Approach*

This approach translates a given sentence phrase wise, and then recombines the individual translated phrases to generate the complete translation of the whole sentence. In the present work the phrases considered are the syntactic phrases derived from the LTRC Hindi Shallow Parser. [3]

3.2 Rule-based Approach

The translation of each phrase is recombined based on certain rules, designed specifically to solve the given sentence structure. The basis of these rules is derived from

- The translation pattern followed for Hindi to English with respect to various कारक (*karaka)* (noun case-ending) present in the Hindi Sentence.
- Type of the sentence (simple declarative, negative or interrogative)

Once these कारक (*karaka)* (noun case-ending) and the sentence type are identified properly, the translation of a Hindi Sentence can be achieved using some transfer rules explained in the Section 6.

3.3 Statistical Approach

Each phrase is translated using statistical techniques. The two well-known statistical translation techniques used for the current work are:

[3] http://ltrc.iiit.ac.in/analyzer/Hindi/

3.3.1 MOSES

Moses[4] is a statistical machine translation system. MOSES allows the language models for any language pair to be trained using a collection of parallel corpus. A parallel corpus consists of a set of translated texts of the two languages under consideration called the source and the target language. Word alignment of the parallel data is established using GIZA++[5]. Using these inputs a trained target language model is generated using a toolkit called SRILM[6] (SRI Language Modelling Toolkit), which ensures the fluency of the output. Using the language model generated by SRILM, any search algorithm can be used for finding the highest probability translation from among the number of choices available. One such efficient search algorithm used widely is the beam search[7].

3.3.2 Google translator

Google Translator is a utility provided by Google to translate written text from one language to another. This is one of the most popular and well known translators available currently.

The overall translation flow proposed is shown in Figure 1. Each step of the flow is explained as follows:

a) Parsing and Sentence Type identification of the Hindi sentence

The first step for translation of any Hindi sentence through the proposed scheme is parsing it. With this the syntactic phrases and the sentence type are identified.

The parsing is done using the LTRC shallow parser:

1. To identify syntactic phrases from the given Hindi sentence.
2. To identify the sentence type; positive, negative, or interrogative.

For example, consider the sentence

> पृथ्वी हर १२ महीने में सूरज का एक चक्कर लगाती है (*prithvii har 12 mahine mein suraj kaa ek chakkar lagaatii hai*)
>
> ~ The earth rotates around the sun once in every 12 months.

The parse output for the sentence is shown in Table 2. In Table 2, NP signifies the Noun Phrase, and VGF signifies the finite verb phrase.

In the given example five syntactic phrases are identified which are

1. पृथ्वी (*prithvii*) ~ (Earth)
2. हर १२ महीने में (*har 12 mahine mein*) (in every 12 months)
3. सूरज का (*suraj kaa*) (of the sun)
4. एक चक्कर (*ek chakkar*) (one round)
5. लगाती है (*lagaatii hai*) (takes)

The NEG (Negative) and WQ (Question Word) POS tags are used to signify negative and interrogative sentences.

Example of the parse output of a negative Hindi sentence

[4] http://www.statmt.org/moses/
[5] https://code.google.com/p/giza-pp/
[6] http://www.speech.sri.com/cgi-bin/run-distill?papers/icslp2002-srilm.ps.gz
[7] http://en.wikipedia.org/wiki/Beam_search

राम खाना नहीं खाता है (*Raam khaanaa nahii khaataa hai*) ~ Ram does not eat food

and Interrogative Hindi sentence

राम ने क्या खाया (*Raam ne kyaa khaayaa*) ~ What did Ram eat

is given in Table 3 and Table 4, respectively.

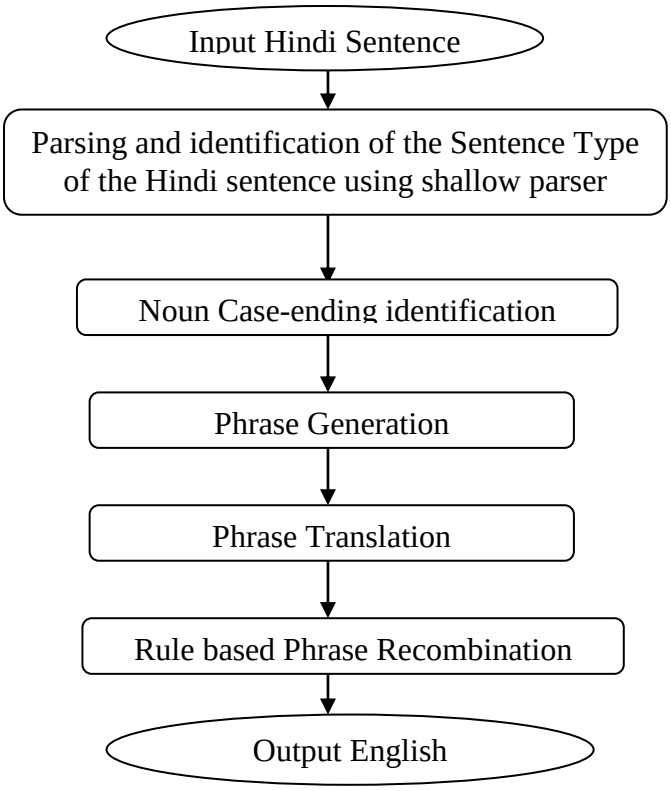

Figure 1 Overall Translation Flow

Table 2 Parse output from the LTRC Shallow Parser

```
<Sentence id="1">
1    ((     NP                <fs af='पृथ्वी,n,f,sg,3,d,0,0' head="पृथ्वी">
     1.1    पृथ्वी (prithvi)   NN      <fs af='पृथ्वी,n,f,sg,3,d,0,0' name="पृथ्वी"> ))

2    ((     NP                <fs af='महीना,n,m,pl,3,d,0_में,0' vpos="vib3_4" head="महीने">
     2.1    हर (har)   QF     <fs af='हर,adj,any,any,,any,,'>
     2.2    12         QC     <fs af='12,num,,,,,,'>
     2.3    महीने (mahine)    NN      <fs af='महीना,n,m,pl,3,d,0,0' name="महीने"> ))

3    ((     NP                <fs af='सूरज,n,m,sg,3,d,0_का,0' vpos="vib1_2" head="सूरज">
     3.1    सूरज (suraj)              NN      <fs af='सूरज,n,m,sg,3,d,0,0' name="सूरज">))

4    ((     NP                <fs af='चक्कर,n,m,sg,3,d,0,0' head="चक्कर">
     4.1    एक (ek)           QC      <fs af='एक,adj,any,any,,any,,'>
     4.2    चक्कर (chakkar)   NN      <fs af='चक्कर,n,m,sg,3,d,0,0' name="चक्कर">))

5    ((     VGF       <fs af='लगा,v,f,sg,2,,,ता_है,wA' vpos="tam1_2" head="लगाती">
     5.1    लगाती (lagaati) VM  <fs af='लगा,v,f,sg,any,,,ता,wA' name="लगाती">))

</Sentence>
```

Table 3 Parse output for Negative Hindi Sentence

<Sentence>
1 ((NP
1.1 राम NN <fs af='राम,n,m,sg,3,d,०,0'>|<fs af='राम,n,m,pl,3,d,०,0'>|<fs af='राम,n,m,sg,3,o,०,0'>))
2 ((VGNN
2.1 खाना VM <fs af='खाना,n,m,sg,3,d,०,0'>|<fs af='खा,v,m,sg,any,,ना,nA'>))
3 ((VGF
3.1 नहीं NEG <fs af='नहीं,n,f,sg,3,d,०,0'>|<fs af='नहीं,n,f,pl,3,d,०,0'>|<fs af='नहीं,n,f,sg,3,o,०,0'>|<fsaf='नहीं,n,f,pl,3,o,०,0'>|<fs af='नहीं,adv,,,,,,'>|<fs af='नहीं,avy,,,,,,'>|<fsaf='नह,v,f,pl,any,,या,yA'>
3.2 खाता VM <fs af='खाता,n,m,sg,3,d,०,0'>|<fs af='खा,v,m,sg,any,,ता,wA'>
3.3 है VAUX <fs af='है,v,any,sg,2,,है,hE'>|<fs af='है,v,any,sg,3,,है,hE'>))
</Sentence>

Table 4 Parse output for Interrogative Hindi Sentence

<Sentence id="1">
1 ((NP
1.1 राम NN <fs af='राम,n,m,sg,3,d,०,0'>|<fs af='राम,n,m,pl,3,d,०,0'>|<fs af='राम,n,m,sg,3,o,०,0'>
1.2 ने PSP <fs af='ने,psp,,,,,,'>
))
2 ((NP
2.1 क्या WQ <fs af='क्या,adv,,,,,,'>|<fs af='क्या,avy,,,,,,'>
))
3 ((VGF
3.1 खाया VM <fs f='खा,v,m,sg,any,,या,yA'>
))
</Sentence>

b) Noun Case-ending identification

Once the individual syntactic phrases are identified, the *karaka* (noun-case ending) information associated to each Noun Phrase is to be identified. This is done using the syntactic cues for identifying each *karaka*, defined in *AnnCorra: TreeBanks for Indian Languages Guidelines for Annotating Hindi TreeBank* [4]. For example, the syntactic cues for identifying *kartaa* as mentioned in [4] are as follows:
(a) *Kartaa* is normally in nominative case which is realized as Ø in Hindi.
(b) By default verb in active voice agrees with the *karta* in number, gender, and person.

Similarly, for each of the कारक (*karaka*) (noun case-ending) syntactic cues are laid down using which they are uniquely identified from a given Hindi sentence.

For the example sentence above, the *karaka* information for each Noun Phrase is as follows:

- पृथ्वी — *kartaa* (doer of an actions)
- हर १२ महीने में — *adhikaran* (location)
- सूरज का — *sambandh* (relation)
- एक चक्कर — *karma* (locus of the result of the action)
- लगाती है — Verb Phrase

c) Phrase generation

The final phrases for translation are generated by the following two rules:

1. Merge the *sambandh* (relation) noun case ending with the Noun Phrase (NP) following it. This is done to preserve the inter noun phrase associations after the translation which are depicted by the relation case ending words *kaa/ke/kii/raa/re/rii.*

2. For example, in the above phrases, सूरज का (*suraj kaa*) is the *sambandh* (relation) noun case ending. Hence, it will merge with the noun phrase following it. Hence, the final phrases to be considered by the proposed scheme are:

- पृथ्वी (*prithvii*)
- हर १२ महीने में (*har 12 mahine mein*)
- सूरज का एक चक्कर (*suraj kaa ek chakkar*)
- लगाती है (*lagaatii hai*)

3. If the sentence to be translates is identified as a negative sentence, then the negative (NEG) word is added to the verb phrase in the *Phrase Generation* step. This is done because the negation is associated with the Verb Phrase of the sentence.

For example, the sentence has the NEG word नहीं (*nahii)*

राम खाना नहीं खाता है (*raam khaanaa nahii khaataa hai*). Thus it will have the phrases as follows

- राम (*raam*)
- खाना (*khaanaa*)
- नहीं खाता है (*nahii khaataa hai*)

d) Phrase translation

The next step of the proposed scheme is to translate each phrase individually through statistical techniques. As mentioned in the previous sections, two well-known statistical schemes are used for the current work.

- *MOSES decoder, with GIZA++ and SRILM*

If a good training corpus is available for generating the language model of the target language, and for generating a good word alignment model, then MOSES decoder can be used easily for translation.

For the current work, the MOSES system is trained and a language model is developed using parallel corpus UMC002. The corpus is published by Charles University in Prague, UFAL. It contains nearly 45000 parallel sentences. Apart from that, there are parallel tokens for both the languages. These tokens are obtained from Wikipedia named entries for the years 2008 and 2009.

The translation of individual phrases using MOSES is as follows:

Table 5 Phrase translation with MOSES

पृथ्वी	Earth
हर १२ महीने में	in every 12 months
सूरज का एक चक्कर	sun's one round
लगाती है	does

- *Google translator*

In the absence of a good training corpus, Google translator /any other statistical translator can be used to translate the individual phrases. The translation of individual phrases using Google is as follows:

Table 6 Phrase translation with Google

पृथ्वी	Earth
हर १२ महीने में	every 12 months
सूरज का एक चक्कर	a round of sun
लगाती है	Applies

It has been observed that the negative verb phrase is not translated by Google translator properly. For example consider few negative verb phrases and the corresponding translations by Google translator

Table 7 Negative Phrase translation by google translator

नहीं गया	Not
नहीं जाता है	Is not
नहीं कर सकता है	You can not
नहीं जा रहा है	Not being
नहीं जाएगा	Will not

Note that the main verb जा is not at all translated in English to "go" in the above phrases. Because of this drawback of the translators, for translating the negated verb phrases, the main verb is individually translated and appended with a negation. For example for translating

नहीं जा रहा है , the translation of जा रहा है i.e. "is going" and a "not" is added before the main verb "going" .

While appending the negation (not) before the translated main verb, the grammatical rules for English language needs to be considered. The current scheme only considers the above rule for generating the translation of negated verb phrases. The scheme can be extended further to cater other verb types.

e) Rule based Phrase Recombination

The recombination of the individual translated phrases must be done based on the ordering rules derived after studying a variety of Hindi sentences, and the corresponding English translations from the UMC parallel corpus. The sentences considered for generating the recombination rules are based on Hindi sentences having different noun case endings and in different order. Some of them are as follows:

Example:

1. राम ने श्याम से सीता के लिए एक किताब माँगी (*Ram ne Shyam se Sita ke liye ek kitaab maangii*)
   Ram sought a book from Shyam for Sita
2. रवि को सतीश से बात करनी है (*Ravi ko Satish se baat karnii hai*)
   Ravi has to talk to Satish
3. राकेश ने श्याम की पढ़ाई में बहुत सहायता की है (*Rakesh ne Shyaam kii padhaii mein bohut sahayta kii hai)*
   Rakesh has helped a lot in Shyam's studies
4. मेरे चाचाजी हर साल नागपुर से संतरे भेजते है (*mere chachaji har saal Nagpur se santre bhejte hai*)
   Every year my uncle sends oranges from Nagpur
5. बहुत ज़ोर से बारिश हो रही है (*bohut zor se bearish ho rahi hai*)
   It is raining heavily
6. हमने उस उम्मीदवार को अपना मत दिया (*humne uss umeedwar ko apnaa mat diyaa*)
   We gave our vote to that candidate
7. तुम हमेशा शिकायत करते हो (*tum hameshaa shikayat karte ho*)
   You always complaint
8. हम सब उसके मज़ाक पर हँसे (*hum sab uske mazaak par hanse*)
   We laughed on his joke

Interrogative sentences

1. राम क्यों श्याम से नाराज़ है ? (*Raam kyu Shyaam se naraaz hai)*
   Why is Ram angry on Shyam?
2. क्या तुम मेरी बात सुनते हो ? (*kya tum meri baat sunte ho*)
   Do you hear my words ?
3. इस पक्षी का नाम क्या है ? (*iss pakshii kaa naam kyaa hai*)
   What is the name of this bird ?
4. खाना खाने का कमरा कहाँ है ? (*khaanaa khaane kaa kamraa kahaa hai*)
   Where is the dining room ?
5. हमारे सामान का क्या हुआ ? (*hamare samaan kaa kyaa hua*)
   What happened to our stuff?

Negative

1. राम खाना नहीं खाता है (*raam khaanaa nahii khataa hai*)
   Ram does not each food.
2. मैं सीता के घर नहीं जाऊंगा (*main Sita ke ghar nahi jaunga*)
   I will not go to Sita's house.

The following are the recombination rules that were generated for the various types of sentences.

Table 8 Recombination rules for Simple Positive sentences

Hindi Sentence Structure based on Noun Case Ending	**Corresponding Translated English Sentence Structure**
Kartaa(doer) ne *Karan*(instrument) se *Sampradaan*(beneficiary) ke liye *karma* (locus of the result of the action) ko VERB	*Kartaa*(doer) VERB *karma* (locus of the result of the action) from *Karan*(instrument) for *Sampradaan*(beneficiary)
Kartaa(doer) ko *Karan*(instrument) se *karma* (locus of the result of the action) VERB	*Kartaa*(doer) VERB *karma* (locus of the result of the action) to *Karan*(instrument)
Kartaa(doer) ne ((*sambandh*(relation) kii) *adhikaran*(location)) mein *karma* (locus of the result of the action) VERB	*Kartaa*(doer) VERB *karma* (locus of the result of the action) in *adhikaran*(location)+ (*sambandh*(relation)

Kartaa(doer) (*sambandh*(relation)) *apaadaan*(source) se *karma* (locus of the result of the action) VERB	*Kartaa*(doer) VERB *karma* (locus of the result of the action) from *apaadaan*(source) + (*sambandh*(relation))
Kartaa(doer) ne *Sampradaan*(beneficiary) ko *Karan*(instrument) VERB	*Kartaa*(doer) VERB *Karan*(instrument) to *Sampradaan*(beneficiary)
Kartaa(doer) ADVERB *karma* (locus of the result of the action) VERB	*Kartaa*(doer) ADVERB+VERB *karma* (locus of the result of the action)
Kartaa(doer) ((*sambandh*(relation) ke *adhikaran*(location)) par VERB	*Kartaa*(doer) VERB at *adhikaran*(location) + *sambandh*(relation)

Table 9 Recombination rules for Interrogative sentences

Hindi Noun Case Ending	**Corresponding English Sentence Structure**
Kartaa(doer) (kyu/kyaa/kahaa/..) *karma* (locus of the result of the action) se VERB	Why/What/Where is *Kartaa*(doer) verb on *karma* (locus of the result of the action)
Kya *Kartaa*(doer) *karma* (locus of the result of the action)ko VERB	Do *Kartaa*(doer) VERB *karma* (locus of the result of the action)
sambandh(relation)*Kartaa*(doer) kyaa VERB	What VERB *sambandh*(relation)+ *Kartaa*(doer)

Table 10 Recombination rules for Negative sentences

Hindi Noun Case Ending	**Corresponding English Sentence Structure**
Kartaa(doer) *karma* (locus of the result of the action) NEG VERB	*Kartaa*(doer) NEG VERB *karma* (locus of the result of the action)
Kartaa(doer) *sambandh*(relation) kaa/ke/kii *karma* (locus of the result of the action) ko NEG VERB	*Kartaa*(doer) NEG VERB to *sambandh*(relation) + *karma* (locus of the result of the action)

The phrases are to be recombined using filler words like *for*, *in* etc. The filler words are decided based on the Noun phrases' case ending/*karaka* type. There can be more than one filler word for the same *karaka* also. For example "at", "on", "in" are the common filler words for the *adhikaran* (location) case ending. For example:

Ram is sitting **at** his friend's place.
Ram is sitting **on** the table.
Ram is sitting **in** the room.

This results in more than one translation of each sentence, based on the filler word used in each of the translations. Based on the English language model obtained from SRILM, the top 5 translation are picked from the output list.

As per the recombination order described, and after choosing the best translation based on the English language model obtained from SRILM, the translation for the example Hindi sentence

पृथ्वी हर १२ महीने में सूरज का एक चक्कर लगाती है
(prithvii har 12 mahine mein suraj kaa ek chakkar lagaatii hai)

using MOSES and Google is as follows:

MOSES: Earth does sun's one round in every 12 months.
Google: Earth applies a round of sun in every 12 months.

For interrogative sentence containing *kyaa* (what), *kyuu* (why), *kisliye* (why) the recombination rule, 9a will be applicable. For interrogative sentence containing *kaun* (who), recombination rule 9b will be applicable. For example, consider the following sentence:

राम अपने घर से सीता के घर क्यों जा रहा है
(raam apne ghar se sita ke ghar kyuu jaa rahaa hai)

The phrases identified are

1. राम *(raam)*
2. अपने घर से *(apne ghar se)*
3. सीता के घर *(sita ke ghar)*
4. क्यों *(kyuu)*
5. जा रहा है *(jaa rahaa hai)*

The translation of each phrase, when recombined using the recombination rule of 9a (as *kyuu*/why rule is to be considered) results in the following translation

MOSES : Why is Ram going to Sita's house from his house.
Google : Why is Ram going to house of Sita from own house.

The scheme can be further extended if other rules are found to be suited for the phrases and thereafter generate the best translation based on the language model of the target language.

4. EXPERIMENTAL RESULTS

Experiments are conducted on a set of 5000 sentences from the UMC parallel corpus of nearly 45000 parallel sentences. From the corpus of 45000 sentences, 40000 sentences were analysed to create the recombination rules explained in Section 3, and remaining 5000 sentences were used for translation evaluation purpose. The individual phrases are translated using two translators: MOSES decoder and Google translator.

Following are the outputs of a few sentences using the proposed translation scheme, and using MOSES and Google, for translation the individual phrases.

1. राम ने श्याम के लिए सीता से एक किताब मांगी ।

MOSES: Ram borrowed book from Shyam for Sita
Google: Ram sought a book from Shyam for Sita

2. मैं तुम्हारे सुन्दर घर में जाउंगा ।

MOSES: I go in your home.
Google: I will go in your beautiful home.

3. राम को रवि की सुन्दर किताब पढ़नी है

MOSES: Ram wants to read beautiful book of Ravi.
Google: Ram has to read the beautiful book of Ravi.

4. मैंने दीवाली के दिन अपने घर में एक छोटी सी पूजा का आयोजन किया

MOSES: I organised a small worship in Diwali in house.
Google: I held a small ritual in Diwali in my house.

The BLEU score is used for the evaluation of the results. The Table 11 depicts the average BLEU score of the top 5 translations produced for 5000 sentences by the proposed scheme using the two translators for the phrases, MOSES and Google, and the corresponding comparison with the state of the art MOSES translator, Google translator and the translation proposed by [12]. The average BLEU score of the best translation using the proposed scheme with MOSES is 26.5 and with Google is 27.9. The proposed scheme is producing top-5 output options (which are having BLEU score of ~26.5 to 14.6 using MOSES and 27.9 to 18.5 using Google). The user can see all the options available and thereafter choose the best as per his judgement.

Table 11 BLEU score for 1000 test sentences

Top n-th translation	Proposed Scheme Using MOSES	Proposed Scheme Using Google	MOSES	Google	IIT Bombay translator [12]
1	26.5	27.9	23.3	24.5	25.7
2	25.4	25.6			
3	22.3	23.2			
4	19.7	19.2			
5	14.6	18.5			

The average BLEU score of the best translation obtained from the proposed scheme (26.5 and 27.9) outperforms when compared with classical MOSES (23.3), Google(24.5), and the IIT Bombay translator (25.7). The reason can be that, in the current work the relative ordering of the translated phrases is done using the recombination rules proposed in Table 8,9 and 10. These rules preserve the semantic information of the sentence and hence the quality of translation improves; which is not the case for the other translators.

5. CONCLUSION AND FUTURE WORK

The scheme provides a starting point for a Hybrid approach for translating Hindi sentences to English, with an attempt to preserve the semantic information of the source language sentence. For the current work the sentences in the problem domain are simple Hindi declarative sentences, negative sentences, and Content-question interrogative sentences. Since the semantic association between the syntactic phrases is the basis of the work, the scheme works well for these sentences. The scheme must be verified for complex, compound and exclamatory sentences. Since these sentences have different structure, the recombination rules will vary depending on the input sentence structure.

In the proposed scheme, the filler words (e.g. “for”, “from”, “in”) between the noun case endings’ translations are pre-defined as given in the recombination rule defined in Section 3. Consideration for other filler words which are applicable for each noun case ending can be done. Also, there can be other considerations, apart from noun case-endings (karakas) when dealing with sentences other than simple declarative sentences.

The scheme relies heavily on the quality of the chunker for identifying the Hindi phrases. As a future study, there can be a comparison between the translation qualities of the output for other chunkers apart for the Shallow Parser. The Indian languages which are like Hindi in structure can also utilize from this approach.

REFERENCES

[1] H. Almaghout, J.Jiang, and A.Way. CCG augmented hierarchical phrase-based machine translation In Proceedings of the 7th International Workshop on Spoken Language Translation, Paris, France, pp.211--218. 2010.

[2] P. Balajapally, P. Pydimarri, M. Ganapathiraju, N. Balakrishnan and R. Reedy, 2006. "Multilingual book reader: Transliteration, word-to-word translation and full-text translation" Proceeding of the 13th Biennial Conference and Exhibition Conference of Victorian Association for Library Automation Melbourne, Australia, pp: 1-12.

[3] S. Bandyopadhyay, (2004) "ANUBAAD - The Translator from English to Indian Languages", in proceedings of the VIIth State Science and Technology Congress. Calcutta. India. pp. 43-51.

[4] Akshar Bharati, Dipti Misra Sharma, Samar Husain, Lakshmi Bai, Rafiya Begam, Rajeev Sangal, AnnCorra : TreeBanks for Indian Languages Guidelines for Annotating Hindi TreeBank, Language Technologies Research Center IIIT, Hyderabad, India (version – 2.5) 17/09/2012

[5] Akshar Bharti, Vineet Chaitanya, Amba P. Kulkarni & Rajiv Sangal, (1997) "ANUSAARAKA: Machine Translation in stages', Vivek, a quarterly in Artificial Intelligence, Vol. 10, No. 3, NCST Mumbai, pp. 22-25

[6] Akshar Bharati, R. Moona, P. Reddy, B. Sankar, D.M. Sharma & R. Sangal, (2003) "Machine Translation: The Shakti Approach", Pre-Conference Tutorial, ICON-2003

[7] A. Birch, M. Osborne, and P. Koehn.CCG supertags in factored statistical machine translation. In proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic; pp.9-16, ACL 2007.

[8] N. Chatterjee, A. Johnson and M. Krishna. Some Improvements over the BLEU Metric for Measuring Translation Quality for Hindi. Proc. ICCTA, IEEE Computer Society, 2007, pp. 485 – 490

[9] S. Dave, J. Parikh, and P. Bhattacharyya. Interlingua Based English Hindi Machine Translation and Language Divergence. Journal of Machine Translation, 17, September 2002.

[10] C. David. A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting of the ACL, pages 263–270, Ann Arbor, MI, 2005.

[11] V. Dayal. Locality in WH Qantification: Questions and Relative Clause in Hindi. Kluwer Academic Publishers.1996.

[12] P. Dungarwal, R. Chatterjee, A. Mishra, A. Kunchukuttan, R. Shah, and P. Bhattacharyya. 2014. The iit bombay hind- english translation system at wmt 2014. ACL 2014, page 90.

[13] Vishal Goyal, Gurpreet Singh Lehal, 2011, "Hindi to Punjabi Machine Translation System" Proceedings of the ACL-HLT 2011 System Demonstrations, pages 1–6, Portland, Oregon, USA, 21 June 2011

[14] D. Gupta and N. Chatterjee Divergence in English to Hindi Translation: Some Studies. International Journal of Translation, Vol 15. No. 2, pp 5 – 24, 2003.

[15] R. Jain, R.M.K. Sinha, and A. Jain. Anubharti: Using hybrid example based approach for machine translation. In Symposium on Translation Support Systems, SYSTRANS, Kanpur, India, February, 2001

[16] G. S. Josan and G. S. Lehal, 2008, "A Punjabi to Hindi Machine Translation System" Coling 2008: Companion volume: Posters and Demonstrations, Manchester, UK, pp. 157-160

[17] P. Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In R.E. Frederking and K. Taylor, editors, Proceedings of the American Machine Translation Association, volume 3265 of Lecture Notes in Computer Science, pages 115–124. Springer, 2004.

[18] G. Labaka, C. Espa na-Bonet, L. Màrquez, K. Sarasola: A hybrid machine translation architecture guided by Syntax. In Machine Translation, 28 (2014), pp. 1–35

[19] Ananthakrishnan R, Kavitha M, Jayprasad J Hegde, Chandra Shekhar, Ritesh Shah, Sawani Bade & Sasikumar M., (2006) "MaTra: A Practical Approach to Fully- Automatic Indicative English Hindi Machine Translation", In the proceedings of MSPIL-06.

[20] R. Sinha, and A. Jain, Angla Hindi: An English to Hindi Machine-Aided Translation System, MT Summit IX, New Orleans, USA, 23-27 Sept. 2003

[21] R. Sinha and A Thakur. On Translation of Interrogative sentences from Hindi to English, International Conference on Machine Learning; Models, Technologies and Applications, MLMTA, Las Vegas, Nevada, USA, 2006.

[22] R. Mahesh K. Sinha & Anil Thakur, (2005) "Machine Translation of Bi-lingual Hindi-English (Hinglish) Text", *in proceedings of 10th Machine Translation Summit* organized by Asia-Pacific Association for Machine Translation (AAMT), Phuket, Thailand

[23] G. Singh and D. Lobiyal. A Computational Grammar For Hindi Verb Phrase, Proc. International Conference on Expert Systems for Development, 1994, 28-31 Mar 1994, pp: 244-249, DOI: 10.1109/ICESD.1994.302273.

[24] H. Somers. Review article: Example-based machine translation. Machine Translation, 14(2):113–157, 1999. McGregor, R. S. 1995. Outline of Hindi Grammar [3rd edition]. Oxford: Oxford University Press.

[25] H. Uchida and M. Zhu. Interlingua for Multilingual Machine Translation. In Proceedings of the Machine Translation Summit IV, Kobe, Japan, July 20-22, 1993.

[26] P. Williams, R. Sennrich, M. Nadejde, M. Huck, E. Hasler, and P. Koehn, "Edinburgh's Syntax-Based Systems at WMT 2014," in Proc. of the Workshop on Statistical Machine Translation (WMT), Baltimore, MD, USA, June 2014, pp. 207–214.

[27] A. Zollmann, A. Venugopal, S. Vogel and A. Waibel. The CMU-UKA Syntax Augmented Machine Translation System for IWSLT-06. In Proc. of International Workshop on Spoken Language Translation (IWSLT-06) -- system papers, Kyoto, Japan. 2006.

[28] https://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Automatic_evaluation

[29] A. Kalyani, H. Kamud, S. Pal Singh, and A. Kumar. Assessing the quality of mt systems for hindi to english translation. In International Journal of Computer Applications, volume 89, 2014.

AUTHORS

Susmita Gupta is a doctorate student in the in the Department of Mathematics, IIT Delhi. Her primary research areas include Machine translation, typically Example Based Machine Translation, and Hybrid approaches for translation, primarily from Indian languages to English. She had done her M.Tech in Computer Application from IIT Delhi, and B.Tech in Computer Science from Institute of Engineering and Technology, Lucknow.

Dr. Niladri Chatterjee is a Professor of Statistics and Computer Science in the Department of Mathematics, IIT Delhi. His primary research areas are:Natural Language Processing, Semantic Web, Statistical Modeling, His association with IIT Delhi is closed to 15 years. Prior to that he had worked in the Dept. of Computer science, University College London, and at Indian Statistical Institute, Calcutta. In 2010 he has been a Visiting Scientist in Dipartimento di Informatica, University of Pisa, Italy. He has over 70 publications international and national journals and conferences. He has been the Organizing Chair of "CICLING – 2012" in March 2012.