Shree Harsh Attri
T.V. Prasad
G. Ramakrishna

HiPHET: HYBRID APPROACH FOR TRANSLATING CODE-MIXED LANGUAGE (HINGLISH) TO PURE LANGUAGES (HINDI AND ENGLISH)

Abstract *Bilingual code-mixed (hybrid) languages have become very popular in India as a result of the spread of technology in the form of television, the Internet, and social media. Due to this increase in the use of code-mixed languages in day-to-day communication, the need for maintaining the integrity of the Indian languages has arisen. As a result of this, a tool named Hinglish to Pure Hindi and English Translator was developed. The tool is capable of translating in three ways; namely, Hinglish into pure Hindi and English, pure Hindi into pure English, and vice versa. The tool has achieved an accuracy of 91% in giving Hindi sentences and of 84% in giving English sentences as output when the input sentences were in Hinglish. The tool has also been compared with another similar tools.*

Keywords code-mixed language, pure language, Hinglish, hybrid language, machine translation, HiPHET, rule based MT

Citation Computer Science 21(3) 2020: 371–391

1. Introduction

In the era of machine translation (MT) and NLP, research on the translation of one monolingual language (MLL) into another and vice versa has become very prominent today. Google Translate translates approx. 100 MLLs into other MLLs individually [18]. Table 1 gives a list of a few monolingual MT systems for non-Indian languages [11]. These systems were developed in countries like Japan, Russia, Sweden, Poland, Spain, etc. for the translation of one MLL into another.

Table 1
Some monolingual MT systems for non-Indian languages

S. No.	MT System	From – To	Year
1.	English Japanese MT system [25]	English to Japanese	1982
2.	RUSLAN [12]	Czech to Russian language	1987
3.	PONS [22]	Norwegian to Swedish	1995
4.	CESILKO [13] [37]	Czech to Slovak language	2000
5.	Bulgarian to Polish MT system [6, 24]	Bulgarian to Polish	2009
6.	APERTIUM [2, 9, 38]	Portuguese to Spanish and vice-versa	2006
7.	ga2gd [30]	Irish and Scottish Gaelic	2006

Table 2 presents a list of a few monolingual MT systems for Indian languages [11]. These MT systems were developed to translate English into Indian languages like Hindi, Telugu, Tamil, etc. as well as translations among Indian languages, like Hindi into Telugu, Punjabi, Marathi, Bengali, etc. and vice-versa.

Table 2
Some monolingual MT systems for Indian languages

S. No.	MT System	From – To	Year
1.	Anglabharti [34]	English to Indian languages	1991
2.	Anusaaraka [3]	One Indian language to another	1995
3.	Mantra (MAchiNe assisted TRAnslation tool) [26]	English to Hindi Language	1999
4.	Vaasaanubaada [36]	Bengali-Assamese News text	2002
5.	Anglabharti-II [33]	English to Hindi	2004
6.	Anubharti [33]	Hindi to English language	2004
7.	Shiva and Shakti [4, 21]	English to Hindi translation	2004

Nowadays, bilingualism is a unique feature in human communication that has been achieved by the mixing of at least two different languages in everyday speech so as to make communication easier [27]. This bilingualism is considered to be a derivative of code switching (CS)/code mixing (CM)/hybrid language (HL) by sociolinguistics such as Hinglish (Hindi+English) [7], Tenglish (Telugu+English), and

Tamlish (Tamil+English). Such harsh language has become more frequent on the Internet as well while writing messages on social media platforms like WhatsApp, Facebook, Twitter, etc. This phenomenon is very common in every multilingual state, like Cantonese-English in Hong Kong, Mandarin-English in Singapore and Malaysia [14], etc.

Due to the emergence of social media, code-switched text data has flooded the Internet. Researchers and data scientists have started to perform research and build tools to detect and translate such code-mixed messages and utterances into one MLL based on the dominant language in the code-mixed text and then translate it into another MLL, if required.

Research tool "Hinglish to Pure Hindi and English Translator (HiPHET)" was developed to translate Hinglish into the pure Hindi and English languages simultaneously; the experiments in this paper focused on the translation capabilities, considering all of the grammatical aspects like nouns, pronouns, verbs, adjectives, phrase word ordering, etc. of HiPHET vis-a-vis another indigenous MT system that claims to translate Hinglish into the pure Hindi and English Languages.

Various research institutes in India such as IIT Kanpur, CDAC Noida, TDIL, etc. are working on MT; they have developed various MT systems for Indian languages like Anusaaraka systems, Mantra systems, Anglabharti, etc. [8, 15]

At present, there are no tools other than HiPHET available for translating HL (Hinglish) into pure languages (PLs) (Hindi and English) simultaneously.

2. Issues in research work

India is a multi-lingual country where most of the languages are code-mixed today due to the influence of English. Hence, the need arose for a code-mixed translation tool for Hinglish, which culminated in HiPHET.

The development of MT systems is comprised of a deep alliance among linguists who form the language rules and computer programmers who code the linguistic rules. The grammar must be optimized to obtain the goal of accurate translation by the use of a bilingual corpus, competent parsing algorithms, and the rearrangement of tags to form a sentence based on the target language's sentence structure.

This research tool focuses on two-way solutions: code-mixed into pure monolingual translations, and a pure MLL into another pure MLL. The steps of the translation related to HL's [5, 15–17, 32] are as follows:

1. Designing and developing word corpus for code-switched language and MLLs.
2. Analyzing and developing grammatical rules for code-switched language and MLLs to translate into PLs.
3. Lexical analysis:
   a) identification of phrases in input sentence,
   b) language identification of each token,
   c) root word identification of each token.

4. Morphological analysis:
   a) identification of features of root words of both languages, like part of speech, tense, number (singular/plural), gender, etc.,
   b) identification of sentence form, whether sentence is general or interrogative.
5. Reverse morphology:
   a) transformation of source lexis into target lexis considering language aspects like part of speech, tense, number (singular/plural), gender, etc.,
   b) conversion to word phrase; e.g.,

$$going \to jA\ rahA,$$

   c) transforming word phrase into one word; for instance,

$$uDatA\ hai \to flies,$$

   d) dropping of verb followers (karanA, kiyA, etc.) from code-mixed phrase and translating into target language as shown in Table 3,
   e) addition of gender-specific postpositions between two nouns/pronouns; e.g,

$$rAma\ kA\ pen \to rAma\ kI\ kalam.$$

6. Sentence Formation
   a) arrangement of translated Hindi tokens as per Hindi sentence structure,
   b) arrangement of translated English tokens as per English sentence structure.

Table 3
Dropping of verb follower like karanA, kiyA, etc.

S. No.	Code-Mixed Verb Phrase	Hindi Translation	English Translation
1	drink karanA	pInA	drink
2	drink kiyA	piyA	drank

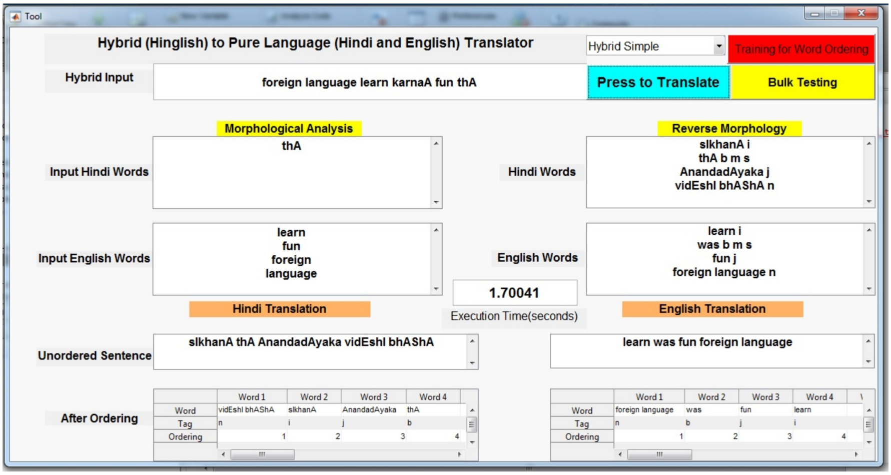
3. Methodology: hybrid language to pure language translation tool

3.1. Tool architecture

The issues identified in the previous section have been resolved by developing a tool named HiPHET, which consists of 32 implemented components. These components are related to the following:

1. translation of HL into pure Hindi and English languages,
2. pure Hindi to pure English translation,
3. pure English to pure Hindi translation.

A snapshot of **"Hinglish to Pure Hindi and English Translator (HiPHET)"** is presented in Figure 1. In this snapshot, the Hinglish input sentence *"foreign language learn karanA fun thA"* was analyzed morphologically and reverse-morphologically to give output words as shown in the snapshot.



Hybrid Simple

Hybrid Input: foreign language learn karnaA fun thA

Input Hindi Words:
thA

Hindi Words:
sIkhanA i
thA b m s
AnandadAyaka j
vidEshI bhAShA n

Input English Words:
learn
fun
foreign
language

English Words:
learn i
was b m s
fun j
foreign language n

1.70041
Execution Time(seconds)

Unordered Sentence: sIkhanA thA AnandadAyaka vidEshI bhAShA

learn was fun foreign language

After Ordering

	Word 1	Word 2	Word 3	Word 4
Word	vidEshI bhAShA	sIkhanA	AnandadAyaka	thA
Tag	n	i	j	b
Ordering	1	2	3	4

	Word 1	Word 2	Word 3	Word 4
Word	foreign language	was	fun	learn
Tag	n	b	j	i
Ordering	1	2	3	4

Figure 1. HiPHET Translation Output: Hinglish to pure Hindi and English languages

Each word in the Reverse Morphology section is followed by its PoS tag, its gender, its number (where relevant); furthermore, the translated outputs in pure Hindi and English are also presented with the time taken for execution. The last step "After Ordering" is to order the unordered sentence of both Hindi and English; this phase is called the word ordering of a sentence in both languages individually, which is also shown in the snapshot of HiPHET. The various PoS tags used in the dictionary of HiPHET are given in Table 4.

Table 4
Details of POS tags used by HiPHET

Entity	POS Tag
Noun	n
Pronoun	p
Regular Verb	v
Irregular Verb	i
Auxiliary Verb	b
Adjective	j
Adverb	d
Conjunction	c
Preposition	r
Phrases/Idioms	h
Numerals	g
Honorifics	o

A bilingual database of approximately 12,000 words has been created. This database consists of a dictionary with entities like nouns, pronouns, verbs, auxiliary verbs, adjectives, adverbs, conjunctions, prepositions, phrases, idioms, honorific words, numerals, etc. as well as a set of rules for word ordering. Experiments were conducted on HiPHET for a dataset created specifically for Hinglish as described in [17]. This dataset consisted of sentences in the four categories of very simple, simple, complex, and very complex sentences.

The tool is based on the hybrid parsing techniques presented in [15] and enhanced in this paper as depicted in Figure 2.

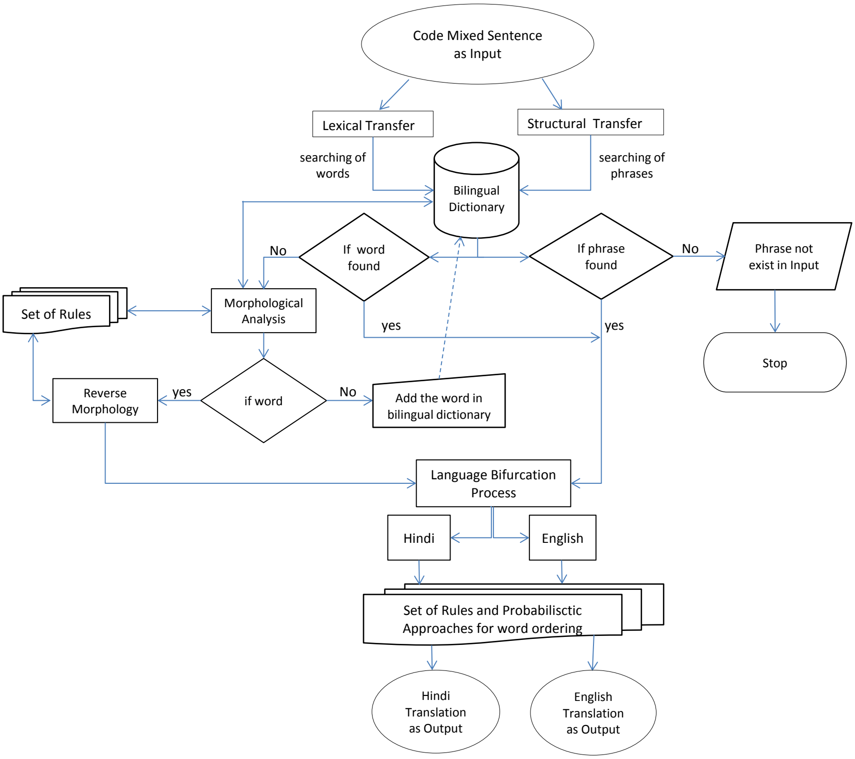

Code Mixed Sentence as Input
Lexical Transfer
Structural Transfer
searching of words
searching of phrases
Bilingual Dictionary
No
If word found
If phrase found
No
Phrase not exist in Input
Set of Rules
Morphological Analysis
yes
yes
Stop
Reverse Morphology
yes
if word
No
Add the word in bilingual dictionary
Language Bifurcation Process
Hindi
English
Set of Rules and Probabilisctic Approaches for word ordering
Hindi Translation as Output
English Translation as Output

Figure 2. Structure of generalized bilingual translation tool

The following steps of algorithm [15–17] demonstrate the working of the tool:

1. Input sentence (Hybrid/Hindi/English).
2. Call phrase/idiom identification.
3. If phrase/idiom present in sentence:
    - separate phrase/idiom from sentence;
    - tokenize remaining sentence; else, if phrase/idiom is absent, then;
    - tokenize whole sentence.
4. Tokenize sentence and store output in LA_Sentence.

5. For each token in LA_Sentence:
   - if token ∈ Hindi Language, then call Morphological_Analysis (Hindi Token) and store result in HM_Token; call Hindi reverse morphology, store result in HRM_Token;
   - add HRM_Token to HRM_Sentence; else, if token ∈ English Language, then call Morphological_Analysis (English Token) and store result in EM_Token; call English reverse morphology, store result in ERM_Token;
   - add ERM_Token to ERM_Sentence.
6. Call modules:
   - Hindi word ordering (HRM_Sentence),
   - English word ordering (ERM_Sentence).
7. Output:
   - translated Hindi sentence,
   - translated English sentence.

3.2. Techniques used

The novel focus of this research work was to translate code-mixed languages to PL's. Since a code-mixed language is highly complex, implementing an MT system becomes a difficult task. The code-mixed language required hybrid grammar rules to translate into pure grammar, and 1,693 rules have been formed and coded to date to translate into the pure Hindi and English languages simultaneously. Following the MT approaches, [1] were used in HiPHET:

1. Direct MT Approach.
2. Rule-Based MT Approach.
3. Hybrid MT Approach in word ordering.

The Direct MT method is dependent on dictionary entries so that a sentence in a source language gets translated into a target language word by word. Generally, such translations are done without morphological aspects and are most suitable in translating phrases/idioms that appear in a sentence as depicted in Table 5. It is also known as the dictionary-based approach.

Table 5
Bilingual dictionary phrase/idiom examples

English Phrase	Hindi Phrase	Phrase/Idiom
as a matter of fact	vAstava mEin	Phrase
as soon as possible	yathA shIghra	Phrase
by hook or by crook	kisI bhI taraha sE	Phrase
may I have your attention please	kripyA Apa dhyAna dIjIyE	Phrase
once in a blue moon	sAla mEin eka bAra	Phrase
red-handed	rangE hAthOn	Idiom

HiPHET deals with the bilingual grammar in a bidirectional manner. For example, consider the following Hinglish sentence with a phrase in English:

mujhE file send kara as soon as possible.

In this sentence, HiPHET identified the phrase *"as soon as possible"* as English and translated it in the direction of English to Hindi to the phrase *"yathA shIghra"*. Thus, the output of this sentence in HiPHET was:

Hindi: "mujhE yathA shIghra sanchikA bhEj"
English: "Send me the file as soon as possible"

Similarly, consider another Hinglish sentence with a phrase in Hindi:

"Student cheating karatA rangE hAthOn pakaDA gayA"

Here, HiPHET identified the idiom *"rangE hAthOn"* to belong to Hindi and translated it in the direction of Hindi to English to the idiom *"red-handed."* HiPHET gave the following output for this sentence:

Hindi:"vidhyArthI nakal kartA rangE hAthOn pakaDA gayA"
English:"The student got caught red-handed while cheating"

The rule-based approach is dependent on a bilingual dictionary and linguistic rules, which are formed by linguistic experts [28]. These rules lead to achieving a morphological analysis, a syntactic analysis, and a source for target sentence formation. These rules have been implemented in HiPHET as part of the algorithms mentioned above.

The combination of the example-based approach [11] and rule-based approach leading to a hybrid approach for word ordering was formed and implemented in HiPHET. The hybrid approach that HiPHET used was to create a bilingual example-based dictionary of PoS Rules for word ordering. Each example in the bilingual dictionary consisted of a PoS rule for a Hindi sentence and a PoS rule for the corresponding English sentence. For example, consider the Hinglish input sentence *"vaha eka boy hai"* (Tab. 6).

Table 6
Example of word ordering in HiPHET

Input	vaha eka boy hai (Hinglish)	
Parameter	Hindi	English
Unordered Translation	vaha eka hai laDkA	he a boy is
PoS tags structure for unordered sentence	**p a i n** for **p**ronoun **a**rticle aux**i**liary_verb **n**oun	**p a n i** for **p**ronoun **a**rticle **n**oun aux**i**liary_verb
Example of PoS rule pair for word ordering	**p a n i** for **p**ronoun **a**rticle **n**oun aux**i**liary_verb	**p i a n** for **p**ronoun aux**i**liary_verb **a**rticle **n**oun
Word-ordered output sentence	vaha eka laDkA hai	he is a boy

4. Experimental results

The tool was given Hinglish sentences as input, and the results included sentences in pure Hindi and pure English that were obtained from each of the Hinglish sentences. The results for the pure Hindi output are presented in Figure 3, and the results for the pure English output are depicted in Figure 4.

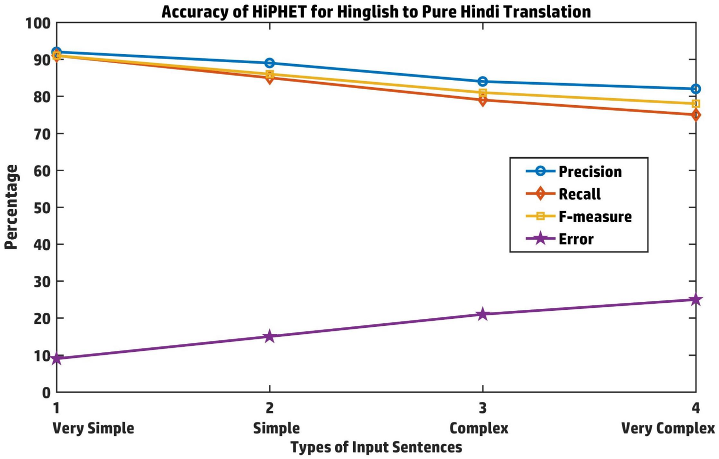


Figure 3. Results of experiments performed using HiPHET for Hinglish to pure Hindi translation

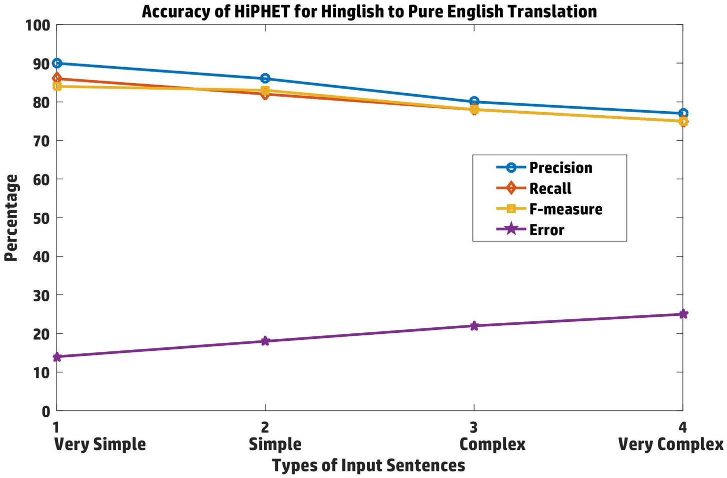


Figure 4. Results of Experiments Performed using HiPHET for Hinglish to Pure English Translation

The precision, recall, F-Measure, and error values for the output sentences were dependent on the type of input sentence, which included sentences in the very simple, simple, complex, and very complex categories. The accuracy of the pure Hindi output sentences was found to be higher as compared to the accuracy of the pure English output sentences [17]. As such, Hinglish is based on Hindi into which English words have been inserted while considering the syntactic and semantic structures of Hindi rather than English. Due to this, the MT of Hinglish into pure Hindi was found to be more accurate as compared to the MT of Hinglish to pure English sentences.

5. Comparison with other translation tools

HiPHET is compared with another tool called Hinglish MT (2005) [35], which was developed by incorporating the morphological analyzers of two MT systems (2004); namely, (AnglaBharti-II) for English-to-Hindi translation, and (AnuBharti-II) for Hindi-to-English translation. Hinglish MT translates Hinglish into Hindi first and then Hindi into English [10, 34].

Another famous translation tool (namely, Google Translate) is very accurate, but it only carries out monolingual translation. Although Google Translate is highly efficient, it cannot be compared with HiPHET, as HiPHET performs code-mixed bilingual translation. Google Translate uses neural machine translation, which requires a very large corpus for accuracy. The accuracy of Google Translate appears to be high, but it is not very good in the case of complex sentences. Besides this, Google Translate is not able to perform accurate translations of some blending words like "youngisthAna" or English slang words in sentences (e.g., "couch potato" is translated to "sofE AlU") in Hindi. Hindi slang words in sentences like "jhakAsa" are not translated at all.

However, HiPHET translates Hinglish into Hindi and English simultaneously. While doing so, HiPHET can also translate pure Hindi into pure English as well as pure English into pure Hindi.

These two tools are compared by considering the various parameters that are discussed below, and a comparison of these two tools is presented in Table 13. These parameters are as follows:

1. **Techniques Used**: refers to techniques used for developing tools.
2. **Domain Specific Translation**: refers to whether tool is generalized or specific to domain selected for creating data set of input sentences.
3. **Handling Phrases**: set of words gives different meaning from those of individual words when clubbed together (see Table 5).
4. **Handling Idioms**: formulaic expression (also called idiomatic phrase) that has symbolic meaning used to add color or poetry to conversation [19, 20] (see Table 5).

5. **Declensions**: Changes in word forms that indicate number, gender, grammatical cases, etc. [29]

   a) **noun declensions**:
   Case 1: while translating some English nouns into Hindi plural, the declinable adjective *"bahuta sE/bahuta sArE"* is prefixed to actual translated noun; for example,

   Apples → bahuta sE sEba

   Case 2: removing Hindi adjectives that are placed before noun to translate into Hindi plural noun and then translate into English plural noun; for example,

   bahuta sE hAthI → many Elephants

   In this example, the identification of a declinable adjective and its translation followed by adding a suffix to the translated English noun has been done.
   Case 3: Hinglish plural "schoolOn" is combination of English word "school" and Hindi plural suffix "On" and translated into Hindi and English as follows:

   Hindi → vidhyAlyOn
   English → schools

   b) **pronoun declensions**: pronouns "I or he/she" have declension form while translating into Hindi (as shown in Table 7).

   c) **adjective declensions**: Hindi language adjectives have declensions based on gender of noun (as depicted in Table 8).

Table 7
Pronoun declension example

Pronoun	Declension forms in Hindi	Hindi Example	English Example
I	mErE main	**mErE** pAsa eka kalama hai **main** jA rahA hUn	I have a pen I am going
he/she	vaha usakE	**vaha** laDakI hai **usakE** pAsa pustaka hai	She is a girl She has a book

Table 8
Adjective declension example

Adjective	Declension forms in Hindi	Hindi Example	English Example
wet	gIlA gIlI	laDakA **gIlA** hai laDakI **gIlI** hai	Boy is wet Girl is wet

6. **Indeclinable Words**: words that are not inflected due to number, gender, or any grammatical rules are known as indeclinable words [29] (see Table 9).

Table 9
Indeclinable word example

Indeclinable Word in		Example	
English	Hindi	Hindi	English
when	kaba	laDakA **kaba** AEgA laDakI **kaba** AEgI	when will boy come when will girl come

7. **Tenses**: Hinglish as such does not possess its own tenses but derives them from Hindi and English only.

8. **Hybrid Phrase**: phrase combining words from Hindi as well as English. For example, "lAThI charge, auspicious mantras, tatkAla reservation, swatch bhArat mission."

9. **Blending Word**: single word that is combination of parts from both Hindi and English words or from two English words (as shown in Table 10).

Table 10
Blended Word example

S.No.	Blended Word	Words Combination
1	Hinglish	Hindi+English
2	youngisthAna	young+hindusthAna
3	fantabulous	fantastic+fabulous
4	smog	smoke+fog

10. **Code-Switched Compounding**: concept in which two complete words (one from Hindi and another from English) are compounded together to form single word in Hinglish; for example, "railgADI" ("rail" is English and "gADI" is Hindi).

11. **Compound Words to One Word**: phenomenon where group of words forming compound word are translated to single word in Hindi and vice-versa; for example,

    raw brown sugar → khAnDa
    father-in-law → sasur

12. **One Word to Compound Words**: when single word in one language forms group of words when translated to another language; for example,

    generally → sAmAnya taura para

13. **Numerals**:

    a) numbers in digit; e.g., 108.

b) numbers in words; e.g.,

One hundred eight → eka sAu ATha

c) currency; e.g.,

rupee → rupaiyA
rupees → rupaya

14. **Addressing of Respect**: refers to certain salutations meant specifically to impart respect to others; e.g.,

Mr. → shrimAn
Mrs. → shrimati
Late → svargiya

15. **Verbs**: refers to regular and irregular verbs in English and Hindi and vice--versa; e.g.,

regular verbs	→	play – played – played
irregular verbs	→	eat – ate – eaten

16. **Auxiliary Verbs**: means helping verbs that denote tenses in sentence in both Hindi and English; e.g.,

is	→	hai	(singular, present tense)
are	→	hain	(plural, past tense)
was	→	thA	(singular, past tense)
were	→	thE	(plural, past tense)

17. **Adverbs**: word that describes action in verb, noun, adjective, etc.; for example,

Hinglish:	laDakI nE song **sweetly** gAyA
Hindi:	laDakI nE gIta **madhurtA sE** gAyA
English:	the girl sang the song **sweetly**

Here, the adverb "sweetly→madhurtA sE" is deriving the action of noun "song."

18. **Colloquial Words (Slang)**: words that are distorted forms of original words forming part of day-to-day communication (as depicted in Table 11).

19. **Abbreviations/Acronyms**: short forms of specific terminology; for example,

Full Name	**Abbreviations**
Indian **A**dministrative **S**ervice	IAS
Indian **I**nstitute of **T**echnology	IIT
University **G**rants **C**ommission	UGC

20. **Active Voice**: refers to voice of sentence in which subject acts upon its verb; e.g., see Table 12.

21. **Passive Voice**: refers to sentence in which subject is recipient of action of verb; e.g., see Table 12.

22. **Negative Sentences**: sentences that result in negative meaning (see Table 12).

23. **Interrogative Sentences**: sentence asking question (see Table 12).

24. **Punctuations**: marks like commas, inverted commas, periods, and other symbols used in Hindi or English language sentences while writing to clarify meaning.

25. **Polysemous Words**: refers to word that takes different meanings depending on following auxiliary verb in Hindi, for example:

    kala hai → tomorrow
    kala thA → yesterday

    The word "kala" in Hindi means "yesterday" if followed by Hindi past tense auxiliary verbs like "thA/thI/thE", and the same word means tomorrow if followed by Hindi present or future tense auxiliary verbs like "hai/hOgA," etc.

26. **Input Word Spelling Correction**: automatic correction of word input by user, for example.

Wrongly spelled word as Input		**Correct Spelling as Output**
Engilsh	→	English
Aple	→	Apple

Table 11

Example of Hindi and English slang words and their respective English and Hindi meanings

S.No.	Language	Words in Slang	Hindi Meaning	English Meaning
1	Hindi Slang	jhakAsa	bahuta achcHA	fantastic
2	Hindi Slang	suTTA	dhumrapAna DanDikA	cigarette
3	Hindi Slang	pakAU	atyant ubAU	extremely boring
4	English Slang	yuck	bahuta burA	disgusting
5	English Slang	sucks	amAnya	unacceptable
6	English Slang	couch potato	AlasI	lazy

Table 12

Examples of various types of sentences in Hinglish with translations into pure Hindi and English languages

Types of sentences as example	Hinglish	Pure English	Pure Hindi
Active Voice	vaha dinner eat karatA hai	He eats dinner	vaha rAtri bhOja khAtA hai
Passive Voice	Dinner usakE dvArA eat kiyA gayA	Dinner was eaten by him	rAtri bhOja usakE dvArA kiyA gayA
Negative Sentence	mErI sister school nahI gaI	My sister did not go to school	mErI bahan vidhAlya nahI gaI
Interrogative	kyA Apa another cup of tea like karEngE	Would you like another cup of tea	kyA Apa eka aura chAyE kA pyAlA pasanda karEngE

Table 13
Comparative Study of Translation Tools

S. No	Tool Names→ Parameters↓	Hinglish MT	HiPHET
1	Year	2005	2019
2	Developed by	Dr. R. Mahesh K. Sinha, Anil Thakur	Shree Harsh Attri, Dr. T. V. Prasad, Dr. G. Ramakrishna
3	Developed at	Indian Institute of Technology, Kanpur, UP, India	K.L. University, Vijayawada, AP, India
4	Techniques Used	Based on Techniques used for AnuBharti-II and AnglaBharti-II	Hybrid Technique (Rule Based, Direct Translation, Transfer Based and Example Based)
5	Translation: Domain-Specific	General	General
6	Hinglish to Hindi and English to Translation	Yes	Yes
7	Hindi to English Translation	Yes	Yes
8	English to Hindi Translation	No	Yes
9	Handling Phrases	INA	Yes
10	Handling Idioms	Yes	Yes
11	Declensions	INA	Yes
12	Indeclinable	INA	Yes
13	Tenses	Yes	Yes
14	Hybrid Phrases	No	No
15	Blending Words	No	Yes
16	Code-Switched Compounding	INA	Yes
17	Compound Words to One Word	INA	Yes
18	One Word to Compound Words	INA	Yes
19	Numerals	INA	Yes
19.1	Numbers in digits	INA	No
19.2	Numbers in words	INA	Yes
19.3	Currency	INA	Yes
20	Addressing of respect	No	No
21	Verbs	Yes	Yes
22	Auxiliary Verbs	Yes	Yes
23	Adverbs	Yes	Yes
24	Colloquial Words	No	Yes
25	Abbreviations/Acronyms	No	Partial
26	Active Voice	Yes	Yes
27	Passive Voice	INA	Partial
28	Negative Sentences	Yes	Yes

Table 13 (cont.)

S. No	Tool Names→ Parameters↓	Hinglish MT	HiPHET
29	Interrogative Sentences	Yes	Yes
30	Input Word Spelling Correction	No	No
31	Punctuations	Yes	No
32	Polysemous Words/Verbs	No	Yes
33	Merits	Produced satisfactory acceptable results in more than 90% of the cases	Has maximum accuracy of 91% for pure Hindi output and maximum accuracy of 84% for pure English output. Translates Hinglish into pure Hindi and English simultaneously, Translates Hindi into English and vice versa directly.
34	Demerits	Not capable of resolving meanings of polysemous verbs	Lack in accuracy in Word ordering for complex sentences
35	Remarks	System first translates Hinglish into Hindi and then from Hindi into English	Aims to translate Hinglish into pure Hindi and pure English Languages simultaneously. Tool allows adding more rules as per requirements. Handles various types of sentences like simple, very simple, compound, and complex.
LEGEND: INA = Information Not Available			

6. Discussion

Translation by another tool, Hinglish MT (2005), consists of a translation into an intermediate MLL and then into another MLL [10, 34, 35]. However, HiPHET translates directly from the bilingual source language into the target languages. HiPHET is domain-independent and can be used for any domain by modifying the dictionary. HiPHET is language-dependent and cannot be used for any languages other than Hinglish, Hindi, or English. Furthermore, although HiPHET is highly accurate in certain translation and word ordering, it is lacking in correctly ordering those words that form parts of phrases and idioms if they are present in a sentence.

HiPHET uses the process of grammar engineering by building linguistic models using the direct MT and rule-based techniques for PoS tagging, stemming, pre-processing, morphological analysis, reverse morphological analysis, and post-processing. Hybrid techniques with "example-based and rule-based" approaches are used by HiPHET for mixed-word translation, word ordering, and idiom/phrase translation. Together, the 32 major modules form a hybrid system that combines rule-based techniques with an example-based approach. HiPHET also reports the time taken to translate an input sentence.

There are many other tools available hat translate from one Indian language into another language. However, there is only one tool available (namely, Hinglish MT), which was developed in 2005 [35]. Like HiPHET, this tool is domain-independent but language-dependent. Hinglish MT (2005) produced satisfactory acceptable results in more than 90% of the cases for Hinglish inputs [31, 35]. This tool used a dataset without any categorization of sentences on the basis of complexity. However, the tool does not translate Hinglish sentences directly into pure Hindi and pure English sentences. The tool first translates Hinglish into Pure Hindi and then translates this intermediate Hindi version into pure English. Thus, the tool never directly translates Hinglish into pure English at all, whereas HiPHET has a maximum accuracy of 91% for pure Hindi output and a maximum accuracy of 84% for pure English output in the case of sentences without phrases/idioms.

Figures 3 and 4 summarize the results of the experiments performed using HiPHET. These results are based on the dataset described earlier in Section 4. It is observable that the errors increase with longer sentences that are being translated. This is evident due to the fact that longer sentences have more-complex forms, and HiPHET gives results that are on par with other Hinglish translation tools. Furthermore, the addition of more words to the dictionary was found to improve accuracy. Similarly, the inclusion of more rules in HiPHET also enhanced its accuracy. Thus, the number of errors was reduced as the outcome of including more words and rules.

A limitation of HiPHET is that a sentence that has multiple nouns or pronouns cannot be ordered correctly. This is due to the fact that the proper ordering of nouns requires knowledge of the semantics of a sentence. This problem will be taken up by the authors in later research work. Similarly, one limitation of Hinglish MT (2005) is that it is unable to resolve the meaning of polysemous verbs [31, 35]. HiPHET also has a limitation of translating proper nouns such as names of persons since there are as many names as there are people, and this is almost infinite. The matter becomes more complicated when the names of people or places have similarities with other words, raising ambiguity. Thus, a name such as "Bengali Babu" causes an ambiguity, as Bengali has two meanings; that is, the name of the Bengali language, and the first name of a person. Similarly, other names such as "Tamil Rajan, Punjab Singh", etc. present similar cases. Similarly, the translation of context dependent words such as mouse (whether the speaker is referring to a live mouse or a computer mouse) depends on the contextual information [23]. This is another limitation of HiPHET, since it does not accept any contextual information as input.

7. Conclusions

It can be concluded that HiPHET is the only tool that translates three ways; i.e., (a) Hybrid into pure Hindi and pure English; (b) pure Hindi into pure English; and (c) pure English into pure Hindi.

The accuracy of HiPHET is on par with Hinglish MT (as was compared in this paper). However, as discussed above, the tool falls short in accuracy in the case of ordering certain sentences that have a phrase/idiom as part of it. Furthermore, HiPHET has certain features that are absent in Hinglish MT. These include the translation of blending words, colloquial words, abbreviations, and polysemous words.

8. Future scope

The authors will work on the word-ordering algorithm described in this paper. This future work will also handle the problem of multiple nouns or pronouns in a sentence (as discussed above). The tool can be combined with speech-recognition and speech-synthesis tools to create a system that receives a spoken bilingual code-mixed sentence, translates it, and then speaks the translated sentence to the listener. This will be especially useful for visually impaired people. It can also be used by tourists who visit a country where code-mixed languages are extensively spoken. It will be useful in various national and international meetings as well as in legislative assemblies (e.g., Lok Sabha and Rajya Sabha proceedings). And last, similar tools can be developed for other code-mixed languages as well; for example, Tamil+English, Telugu+English, Marathi+English, etc., or any other mixed pairs of languages.

References

[1] Antony P.J.: Machine Translation Approaches and Survey for Indian Languages, *ROCLING/IJCLCLP*, vol. 18(1), pp. 47–78. 2013.

[2] Armentano-Oller C., Carrasco R.C., Corbí-Bellot A.M., et al.: Open-Source Portuguese–Spanish Machine Translation. In: Vieira R., Quaresma P., Nunes M..G.V., Mamede N.J., Oliveira C., Dias M.C. (eds.), *Computational Processing of the Portuguese Language. PROPOR 2006*, Lecture Notes in Computer Science, vol. 3960, pp. 50–59, Springer, Berlin, Heidelberg, 2006.

[3] Bharati A., Chaitanya V., Kulkarni A.P., Sangal R.: Anusaaraka: Machine Translation in Stages, *CoRR*, vol. cs.CL/0306130, 2003.

[4] Bharti A., Moona R., Reddy R., Sankar B., Sharma D.M., Sangal R.: Machine Translation: The Shakti Approach. In: *Tutorial at ICON-2003*, 2003.

[5] Chakrawarti K.R., Bansal P.: Approaches for Improving Hindi to English Machine Translation System, *Indian Journal of Science and Technology*, vol. 10(16), pp. 1–8, 2017, https://doi.org/10.17485/IJST/2017/V10I16/111895.

[6] Dimitrova L., Koseska V., Roszko D., Roszko R.: Bulgarian-Polish-Lithuanian Corpus: Current Development. In: *Proceedings of the Workshop Multilingual resources, technologies and evaluation for central and Eastern European languages*, pp. 1–8, Association for Computational Linguistics, 2009.

[7] Dixit P.: Hinglish as a Hybrid Language: An Analytical Study, *International Journal of Research and Analytical Reviews*, vol. 3(1), pp. 162–167, 2016.

[8] Dwivedi S.K., Sukhadeve P.P.: Machine Translation System in Indian Perspectives, *Journal of Computer Science*, vol. 6(10), pp. 1111–1116, 2010.

[9] Forcada M.L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J.A., Sánchez-Martínez F., Ramírez-Sánchez G., Tyers F.M.: Apertium: a free/open-source platform for rule-based machine translation, *Machine Translation*, vol. 25(2), pp. 127–144, 2011.

[10] Goyal P., Mittal M.R., Mukherjee A., Raina A., Sharma D., Shukla P., Vikram K.: A bilingual Parser for Hindi, English and code-switching structure. In: *Proceedings of the Workshop on Computational Linguistics for the Languages of South Asia, 10th Conference of the European Chapter*, pp. 15–22, 2003.

[11] Gupta D., Chatterjee N.: Identification of Divergence for English to Hindi EBMT. In: *Proceedings of MT Summit-IX*, pp. 141–148, 2003.

[12] Hajič J.: RUSLAN – An MT System Between Closely Related Languages. In: *3rd Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen, Denmarks*, 1987.

[13] Hajič J., Hric J., Kuboň V.: ČESÍLKO – An MT system for closely related languages. In: *ACL2000, Tutorial Abstracts and Demonstration Notes, ACL-Washington*, pp. 7–8, 2000.

[14] Hamed I., Elmahdy M., Abdennadher S.: Building a First Language Model for Code-switch Arabic-English, *Procedia Computer Science*, vol. 117, pp. 208–216, 2017, https://doi.org/10.1016/j.procs.2017.10.111.

[15] Harsh A.S., T.V. Prasad, Ramakrishna G.: Issues in parsing and POS tagging of hybrid language. In: *2012 IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom)*, pp. 20–24, 2012, https://doi.org/10.1109/CyberneticsCom.2012.6381609.

[16] Harsh A.S., T.V. Prasad, Ramakrishna G.: Identification and Translation of Noun in Bilingual Code Mixed Language into Pure Form, *International Journal of Applied Engineering Research*, vol. 10(9), pp. 21591–21604, 2015.

[17] Harsh A.S., T.V. Prasad, Ramakrishna G.: Translation of Code Mixed Language to Monolingual Languages using Rule Based Approach, *International Journal of Cloud Computing* (in press).

[18] Information about Google Translator available at: https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-translate-how-work-foreign-languages-interpreter-app-search-engine-a8406131.html.

[19] Information about Idiom available at: https://7esl.com/english-idioms/#What_is_an_Idiom.

[20] Information about Idiom available at: https://www.smart-words.org/quotes-sayings/idioms-meaning.html.

[21] Information about Shiva and Shakti available at: https://web.iiit.ac.in/∼ papi_reddy/test.pdf.

[22] Information on PONS available at: https://en.pons.com/translate/german-norwegian.

[23] Mall S., Jaiswal U.C.: Word sense disambiguation in Hindi applied to Hindi-English machine translation, *Computer Modelling & New Technologies*, vol. 21(2), pp. 58–68, 2017.

[24] Marinov S.: Structural Similarities in MT A Bulgarian-Polish case, 2003.

[25] Nagao M., Tsujii J.i., Yada K., Kakimoto T.: An English Japanese Machine Translation System of the Titles of Scientific and Engineering Papers. In: *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, 1982, https://www.aclweb.org/anthology/C82-1039.

[26] Naskar S., Bandyopadhyay S.: Use of Machine Translation in India: Current Status, 2005.

[27] Parshad R.D., Bhowmick S., Chand V., Kumarui N., Sinha N.: What is India Speaking? Exploring the "Hinglish" invasion, *Physica A: Statistical Mechanics and its Applications*, vol. 449, pp. 375–389, 2016.

[28] Poornima C., Dhanalakshmi V., Anand Kumar M., Soman K.P.: Rule based Sentence Simplification for English to Tamil Machine Translation System, *International Journal of Computer Applications*, vol. 25(8), pp. 38–42, 2011, https://doi.org/10.5120/3050-4147.

[29] Rao T.K.: *Telugu to Sanskrit Machine Translation System – An Hybrid Approach*, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India, Ph.D. thesis, 2017.

[30] Scannell K.: Machine translation for closely related language pairs. In: *Proceedings of the LREC 2006 Workshop on Strategies for Developing Machine Translation for Minority Languages*, pp. 103–107, 2006.

[31] Singh Lehal G., Goyal V.: Advances in Machine Translation Systems, *Language in India*, vol. 9, pp. 138–150, 2009.

[32] Singh K., Sen I., Kumaraguru P.: Language Identification and Named Entity Recognition in Hinglish Code Mixed Tweets. In: *Proceedings of ACL 2018, Student Research Workshop, Melbourne, Australia*, pp. 52–58, Association for Computational Linguistics, 2018, https://doi.org/10.18653/v1/P18-3008.

[33] Sinha R.M.K.: An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures. In: *Proceedings of International Symposium on MT, NLP and Translation Support System*, 2004.

[34] Sinha R.M.K., Jain A.: Angla Hindi: An English to Hindi Machine-Aided Translation System. In: *MT Summit IX*, New Orleans, USA, 2003.

[35] Sinha R.M.K., Thakur A.: Machine Translation of Bi-lingual Hindi-English (Hinglish) Text. In: *MT Summit X*, Phuket, Thailand, 2005.

[36] Vijayanand K., Choudhury S.I., Ratna P.: VAASAANUBAADA: automatic machine translation of bilingual Bengali-Assamese news texts. In: *Proceedings of Language Engineering Conference*, pp. 183–188, 2002, https://doi.org/10.1109/LEC.2002.1182307.

[37] Web enabled ČESÍLKO, available at: https://lindat.mff.cuni.cz/services/cesilko/about.php.

[38] Web enabled open source Apertium, available at: http://www.apertium.org.

Affiliations

Shree Harsh Attri
K.L. University, Department of CSE, Vijayawada (Andhra Pradesh), India; Department of CSE, JIMS Engineering Management Technical Campus (JEMTEC), Greater Noida (Uttar Pradesh), India; shreeharshattri@gmail.com
ORCID ID: https://orcid.org/0000-0001-6915-0048

T.V. Prasad
GIET Institutions, Rajahmundry (Andhra Pradesh), India, tvprasad2002@yahoo.com

G. Ramakrishna
K.L. University, Vijayawada (Andhra Pradesh), India, ramakrishna_10@yahoo.com

Received: 16.01.2020
Revised: 24.07.2020
Accepted: 05.08.2020