Chapter 6

Experiments, Results and Discussions

In this chapter, we will talk about the experiments that were conducted to evaluate the English-Hindi MT system, Anusāraka. This system uses the notion of *pada* described in chapter 3, the dependency schema described in chapter 4 and transfer grammar rules described in chapter 5. We will also talk about the results of the Anusāraka system with comparison to Google Translate. We have used Google's web interface[1] for English-Hindi language pair. Finally, we will do the error analysis and present the conclusions.

6.1 Description of the Data and Evaluation

For our experiments, we have randomly picked 500 Full-text English sentences from Corpus of Contemporary American English (COCA) [44] as input where the shortest sentence was 3 word long and the longest sentence was 62 word long. The average sentence length was 22 word long.

The 500 sentences were then divided into 20 sets where each set contained 25 sentences. All these input sets were then translated into TL Hindi, using the MT systems, Anusāraka and Google Translate. These English sentences were then paired with their Hindi translations.

6.2 Evaluation of Translation Quality

The evaluation of the translation quality is a method which allows to assess and compare the output quality of the given translation system/s. MT output is evaluated for fluency and comprehensibility. This makes the evaluation process an important part of any MT system development. Apart from good or bad translations, a sentence can have many different translations. In such a scenario, either one should have multiple (all possible) reference translations or should opt for subjective evaluation by bilingual human judges. Therefore, evaluation of the translation quality of an MT system itself becomes a difficult problem.

[1] Google Translate's URL: https://translate.google.com

6.2.1 Evaluation Methods

Following two ways are used for MT evaluation:

1. Manual Evaluation: Manual evaluation or also called subjective evaluation, is one of the traditional methods of evaluating the quality of an MT system. In this method, the output of an MT system is judged sentence by sentence by bilingual human evaluators who know the source language as well as the target language. Since, humans judge the system output quality in this method, it produces high quality results and is considered to be an accurate and the most reliable method. But there are disadvantages to this method. It is time consuming, expensive and non-repeatable.

   Since evaluation is a laborious task, finding out the right number of bilingual or mono-lingual human judges itself becomes a challenging problem. More than one human evaluators evaluate every sentence to avoid biases. Their ratings are then averaged to a single score.

   In subjective evaluation itself, two approaches are followed: (a) Open Approach and (b) Blind Approach.

   (a) Open Approach: In this approach, the evaluators are provided with both the source sentence and its translation produced by the MT system under evaluation. The evaluators do a subjective analysis whether the given translated sentence is a fluent target language sentence and conveys all the information contained in the source language sentence and rate them accordingly.

   (b) Blind Approach: In this approach, the source language sentences are hidden from the evaluators, they have access only to the translated MT output. The evaluators check whether the given sentences sound fluent and natural in the given target language and rate them accordingly.

   There are some risks involved in this approach. A sentence may appear fluent and natural in the target language but may not have the crucial information that was there in the source language. For instance, a source sentence may have negation marker in it that might be missing in the target translation. The translated sentence would sound fluent and natural even without the information of the utmost importance to perform or not to perform an action.

2. Automatic Evaluation: Apart from human evaluation, automatic metrics such as BLEU [103], NIST [49], METEOR [6], WER [123] and LEPOR [1] etc. are widely used for MT evaluation. These metrics use reference translation to judge the quality of the translation of a machine. Though automatic metrics do not always match with manual evaluation results, they are widely used for translation evaluation. Automatic metrics play a very important role in monitoring the daily changes and weeding out the bad ideas during the development cycles of an MT system.

They are quick, inexpensive, re-usable, language independent and very good at fluency evaluation.

Despite the fact that evaluation by human evaluators is very expensive, non-repeatable and time consuming, human judgment is still considered to be the most reliable evaluation method in the field [30, 31, 32].

As said before, the focus in Anusāraka is towards accessibility. The Anusāraka system is not a mere machine translation system. It is a system that focuses more on accessing the information present in the source language rather than on the fluency. In Anusāraka, the emphasis is on comprehensibility rather than on the fluency of the target language sentence. Therefore, we decided to do the evaluations by the human judges. We choose the manual cum open evaluation method, so that we get a clear picture of how much information was transferred into the target language.

6.2.2 Evaluating Translation Quality in Terms of Comprehensibility

The translated sentences were uploaded on the graphical user interface *http://shakti.iiit.ac.in/MTESS/* for translation quality assessment. This interface is designed by IIIT-Hyderabad (International Institute of Information Technology, Hyderabad). It evaluates the comprehensibility of MT systems' output. Human evaluators rate the output of the given MT systems in terms of comprehensibility.

The sentences were then assigned to the human evaluators. The evaluators knew both the source language English and the target language Hindi but they were not professional translators. One human evaluator evaluated two different sets and each set was evaluated by three evaluators. In total, 30 evaluators rated the translation quality of both the systems on a scale of 0-4 described in Table 6.1. As per Bharati et al. (2004), this scale assesses translation quality in terms of comprehensibility [19].

Scale	Description
0	Unacceptable (does not make sense)
1	Unacceptable (major errors in translation, comprehensibility seriously effected)
2	Acceptable (some errors in translation but comprehensible)
3	Acceptable (No major errors in translation, fully comprehensible)
4	Acceptable (Perfect translations)

Table 6.1: MT comprehensibility evaluation rating scale

Note that the identity of any system was not disclosed to the human evaluators for unbiased ratings. We found that on this scale Google Translate obtained 44.2% comprehensibility score while Anusāraka obtained 51.8% comprehensibility score shown in Table 6.2.

MT System	Comprehensibility Score
Anusāraka	51.8%
Google Translate	44.2%

Table 6.2: Overall system comprehensibility results

6.3 Experiments Done for Testing Reordering

Out of the 20 sets described in Section 6.1, we selected 4 sets for English-Hindi word ordering quality assessment. The evaluators were asked to rank these sets on the basis of word order quality without giving much emphasis on translation quality, correct lexical substitutions. The evaluators rated the sentences '0' if the word order was unacceptable (but still the sentence might be comprehensible in some cases) and '1' if the order was acceptable and fluent.

The reordering rules for the Anusāraka system were tested for both gold as well as automatic constituency parse based input for reordering accuracy. The same rating approach was followed for Google Translate also. Table 6.3 shows the reordering results.

MT System	Word Ordering Score
Anusāraka with gold parse	67%
Anusāraka with automatic parse	65.5%
Google Translate	42%

Table 6.3: Reordering accuracy results evaluated by common people

The reordering rules were also tested on gold and automatic constituency parse based input by the developers. The gold constituency parse data was manually created by the developers. For creating the gold constituency parse data, we ran the input sentences on Stanford and then manually corrected the parses output. The reordering scale for developers was between 0-2, where '0' rating means 'unacceptable', '1' means 'non-fluent but acceptable' and '2' means 'fluent/acceptable' as shown in Table 6.4.

Scale	Description
0	unacceptable
1	non-fluent but acceptable
2	fluent/acceptable

Table 6.4: Reordering evaluation criteria for developers

The results for word ordering accuracy on the scale given in Table 6.4, are shown in Figure 6.1 and 6.2 through pie charts.

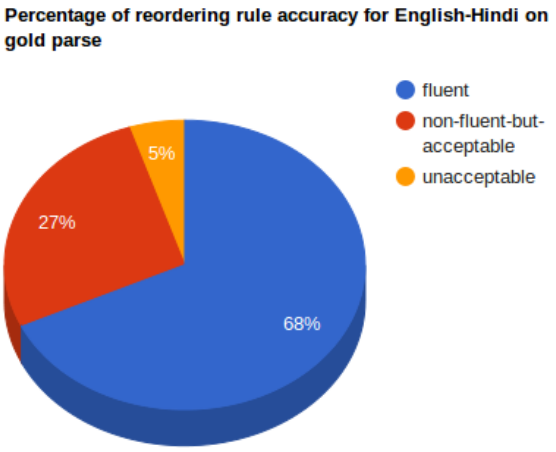

Figure 6.1: Percentage of reordering accuracy for English-Hindi with gold parse

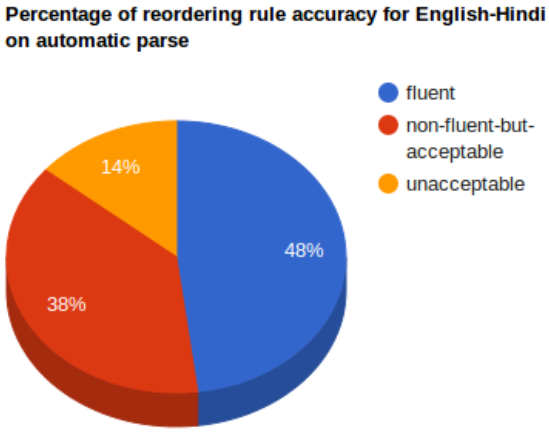

Figure 6.2: Percentage of reordering accuracy for English-Hindi with automatic parse

As expected, the results of the gold parse based word ordering are slightly better than the automatic parse based word ordering in case of evaluation done by the developers as well as by the common peoples. As mentioned in Chapter 5, the results confirm that a rule based reordering system performs better than a statistically trained system.

Since our system basically reorders phrases/*padas* and just because of incorrect reordering of one phrase in around a 50 word long sentence, penalizing the entire sentence does not sound justified. Therefore, we also report phrase reordering accuracy rather than sentence reordering accuracy for Anusāraka. Table 6.5 gives an overview of the corpus size taken for phrase ordering evaluation. And, Table 6.6 shows the results of the evaluation done for number of phrase reordering accuracy.

#Sentences	#Words	#Phrases
100	2151	1637

Table 6.5: Test Corpus

	#Phrases ordered correctly	#Phrases ordered incorrectly
Gold parse	97.5%	2.5%
Automatic parse	93%	7%

Table 6.6: Reordering results based on gold and automatic parse for Anusāraka

Since we do not have any control on Google Translate, we cannot give it constituency parsed input for translation, therefore, we cannot compare Anusāraka's constituency parse based results with Google Translate.

Using our tool, Karan et al. (2014) have reported 21.84 BLEU (bilingual evaluation understudy[2]) score improvement over the baseline 20.04. After that the system has been improved significantly.

6.4 Error Analysis

In this section, we will analyse various reasons behind low ratings. This includes both: 1. Incorrect Reordering and 2. Miscellaneous Problems:

1. **Reordering Problem:** In most of the cases, the adverbs like *now, mainly, likely*, etc. were misplaced as shown in 84.

   (84) Saleh is **now** in the U.S. for further medical treatment.
   Saleh ho.PR,3,SG **aba** meṃ U.S. ke liye āge cikitsā
   'Saleh **aba** āge kī cikitsā ke liye U.S. meṃ hai'
   'Saleh āge kī cikitsā ke liye **aba** U.S. meṃ hai'
   '*Saleh āge kī cikitsā ke liye U.S. meṃ **aba** hai'

2. **Miscellaneous Problems:** These problems can be further sub-classified into the following three classes: (a) WSD Problems, (b) Target Language Word Insertions and (c) Information Sharing or Repeating.

[2]BLEU is an algorithm for automatic evaluation of the quality of machine translation output [103].

(a) **WSD Problem:** Some cases rated as '0' were actually ordered correctly but the evaluators rated them low due to incorrect target language word substitutions. For instance, in 85, *reported* is translated as *sūcanā dī*, but, in system generated output, it was translated as *pāyā*. Otherwise, there was no mistake in placement of the source language words according to the target language word order.

(85) Hubbard reported from Cairo
Hubbard.NOM sūcanā de.PT se Cairo
'Hubbard ne Cairo se sūcanā dī.'

(b) **Target Language Word Insertions:** In order to arrive at a fluent target language word order, relative pronouns, subordinating conjunctions etc., which are not there in source language have to be inserted in the target language. Sometimes insertions like *ki* (that), *isaliye* (hence), etc. were made but they were not placed correctly in target language as shown in 86.

(86) Since chalk first touched slate, schoolchildren
jaba se cāka.NOM.SG pahalī bāra chū.PT,SG saleta.ACC skūla kā baccā.NOM,PL
have wanted to know: What is on the test?
AUX cāha.PT *INF_MRKR* jāna: kyā ho.PR,3,SG meṃ parīkṣā?
'?Jaba se cāka ne pahalī bāra saleta ko chūā, skūla ke baccoṃ ne jānanā cāhā parīkṣā meṃ kyā hai?'

'Jaba se cāka ne pahalī bāra saleta ko chūā, **taba se** skūla ke baccoṃ ne jānanā cāhā **ki** parīkṣā meṃ kyā hai?'

In 86, insertion of subordinating conjunctions: *taba se* (thenceforth) and *ki* (that) is required for a fluent and comprehensible sentence in Hindi.

(c) **Information Sharing or Repeating:** As discussed in chapter 5, sometimes items linked with coordination in English have to be repeated in Hindi. Or, the items that are repeated in English are shared through coordination in Hindi sometimes.

Often, a shared verb translates into a conjunct verb (noun/adjective + verbalizer) in Hindi. In a conjunct verb, the noun or adjective part is termed *kriyāmūla*. In such cases, only the verbalizer/helping verb [9] gets repeated not the entire "noun/adjective + verbalizer" sequence. The verbalizer then behaves as a verbal base to which the verbal inflections are attached and, shares the *kriyāmūla* as shown in bold in 87.

(87) ... he **can** and **will help** fight the country's
... vaha.MASC,3.SG ***BE ABLE TO*** aura ***FT*** madada kara ladanā deṡa kī
active Al-Qaida branch.
sakriya Al-Qaida ṡākhā
'... vaha deṡa meṃ sakriya Al-Qaida ṡākhā se ladane meṃ **madada kara sakatā hai** aura **karegā**'

Such cases need to be handled more accurately during translation.

6.5 Adapting The Tools for English-Indian Language Machine Translation

The focus of the research carried out in this thesis was on English-Hindi language pair. But we claim that this research is generic in nature and can be applied to any English-Indian language pair. This is because most of the Indian languages have many common features and the transparency of the tools built during this research allows to cover diverse cases easily.

In order to verify this claim, we took English-Telugu language pair as an example and adapted these tools to develop English-Telugu machine translation system.

For testing the English-Telugu MT system, we took the same 500 English sentences described in Section 6.1 and the same rating scale that was used for English-Hindi described in Table 6.4 was used for English-Telugu MT system.

To our surprise, the tools built for English-Hindi produced very good results for English-Telugu MT system without any changes according to the target language Telugu.

The major bottleneck in case of English-Telugu was WSD. There are no WSD rules at present and the size of English-Telugu dictionary is also very small. Otherwise, at the level of grammar, English-Hindi system works well for English-Telugu pair as well. For instance, let us take the reordering tool, we found that same set of rules work very well barring a few cases where Telugu diverges from Hindi. Therefore, the rules meant for English-Hindi did not perform very well in those cases. However, this is not a major task. In these cases, we can easily change the rules.

We found that around 67% sentences were rated as fluently ordered sentences. For example, let us see some cases that depict the transparency and ease to cover the major differences between Hindi and Telugu word order.

- Negation Marking: In Hindi negation marker '*nahīṃ*' precedes the verb whereas it follows the verb in Telugu as shown in 88 and 89 respectively.

  (88) Hari did **not** **read** the book.
  Hari *PAST* **nahīṃ**.NEG_MRKR **paḍha**.V pustaka
  'Hari ne pustaka **nahīṃ paḍhī**.'

  (89) Hari did **not** **read** the book.
  Hari *PAST* **ledu**.NEG_MRKR **chaduva**.V pustakamu
  'Hari pustakamu **chaduva ledu**.'

  Since, Anusāraka also produces a 'debug file' for every translation module, the developers can easily identity what action was performed by which rule and make changes accordingly. The above mentioned divergence is handled by a simple change in the rule that puts the negation marker before the verb (*tiṅanta pada*). For Telugu, the negation marker will be put after the verb.

- Subject Object Verb Order: Similar to Hindi, Telugu also follows SOV (Subject Object Verb) order. But if the object is a clausal/sentential object, Hindi follows SVO order whereas Telugu maintains SOV order. See example 90 and 91 for Hindi and Telugu respectively.

  (90) **She** **said** that Hari did not read the book.
  vaha.FEM.3.SG kaha.PAST ki Hari *PAST* nahīṃ paḍha pustaka
  '**Usane kahā** ki Hari ne pustaka nahīṃ paḍhī.'

  (91) **She said** that Hari did not read the book.
  āme cheppu.PAST ani Hari *PAST* ledu chaduva pustakamu
  '**Āme** Hari pustakamu chaduva ledu ani **cheppindi**.'

  In order to handle this difference, in the verb phrase (VP) reversal rule, the condition that restricts VP reversal for Hindi would be commented for English-Telugu system.

Thus, the research carried out in this thesis in generic enough in nature to develop any English-Indian language MT tool.

This experiment was conducted to test the extensibility of this approach to other Indian languages and more work needs to be done to further establish it. This was just an experiment to give it some direction and to see that it does work for English to other Indian languages.

6.6 Conclusion

In this chapter, we presented the experiments done for overall comprehensibility test of the translations produced by the English-Hindi Anusāraka system and the Google Translate. The translations generated by both the systems were evaluated by 30 human evaluators on a 0-4 scale to assess the translation quality in terms of comprehensibility. We found that the Anusāraka system obtained better results than the SMT system Google Translate which shows that the approach is accurate enough to enhance the translation quality.

We have also evaluated the output of both the systems for word ordering and found that Anusāraka system which is based on the hand crafted reordering rules designed using the insights from Pāṇinian Grammar outperforms the SMT system Google Translate.

We have also claimed that the approach presented in this thesis is generic in nature and can be applied to any English-Indian language machine translation system. We have shown an English-Telugu machine translation sample to prove this claim.