Chapter 6
Results and Discussion

6.1 Introduction
This chapter presents the results and discussions on the work carried out in this thesis. This work, as mentioned earlier, has focused on the development of EnConverter and DeConverter for Punjabi language. The development process of these tools has been explained in detail in the previous chapters. Apart from the process that was followed, the outcomes of EnConverter and DeConverter have also been discussed. The important aspect of these tools that has been presented in this chapter is their testing for the perspective of a good machine translation system.

The testing of a MT system is an important aspect in order to establish the usefulness of the system. The evaluation of proposed system, consisting of Punjabi EnConverter and DeConverter, has been performed with the help of one thousand Punjabi sentences. These sentences have been selected in such a way that generation of all possible UNL relations and attributes can be tested. The results of adequacy, fluency tests and BLEU score are very encouraging. The quantitative test, namely, error analysis has also been performed by calculating Sentence Error Rate (SER) and Word Error Rate (WER) of the proposed UNL based MT system. The results of this experimentation are presented in subsequent sections.

6.2 Selection of sentences for test cases
An important aspect in MT system evaluation is to make appropriate selection of sentences for evaluating an MT system. According to Balkan (1998), there are three types of sources for evaluation of MT system, namely, test corpora (a collection of naturally occurring text), test suites (a collection of usually artificially constructed inputs, where each input is designed to probe a system’s treatment to a specific phenomenon) and test collections (a set of inputs associated with a corresponding set of expected outputs). Accordingly, one thousand sentences have been considered for evaluation of proposed UNL based MT system for Punjabi language. As mentioned, these sentences have been
considered in such a way that generation of all possible UNL relations and attributes can be tested.

6.3 Evaluation of Punjabi EnConverter

Evaluation of Punjabi EnConverter is performed by using test collections, i.e., a set of one thousand Punjabi sentences are created with their corresponding set of expected UNL expressions as output. For this purpose Spanish UNL Language Server (Spanish Language Center, 2004) and agricultural domain threads developed by IIT Bombay, India are considered as gold-standard EnConverters.

Spanish Language Server contains English sentences with their corresponding UNL expressions generated by the system. These English sentences were translated manually into equivalent Punjabi sentences and then inputted to the proposed Punjabi EnConverter system for their EnConversion to UNL. The UNL expression generated by proposed system is compared with the UNL expression generated by Spanish Language Server. The agricultural domain threads developed by IIT Bombay have Hindi language sentences with their equivalent UNL expressions. These Hindi sentences were manually translated into Punjabi language and then processed in Punjabi EnConverter for their equivalent UNL expressions. Similarly, the comparison has been performed for the UNL expressions generated by proposed system with the corresponding UNL expressions given at agricultural domain threads.

The two UNL expressions match with each other if the UNL relations, including associated UWs and UNL attributes present in the expressions are same (Jain and Damani, 2009). It has been seen that proposed system handles the resolution of UNL relations and generation of attributes for these sentences with a very reasonable accuracy. Results of these evaluations on the basis of each UNL relation are depicted graphically in Figure 6.1.
6.4 Evaluation of Punjabi DeConverter

Proposed Punjabi DeConverter is evaluated by inputting UNL expressions generated by Punjabi EnConverter to it and the output of Punjabi DeConverter is compared with input Punjabi sentence given to Punjabi EnConverter. The approach followed in this work for evaluating UNL based MT system for Punjabi language is given in Figure 6.2. The two Punjabi sentences (original sentence inputted to Punjabi EnConverter and the sentence generated by DeConverter) are compared and the system is evaluated based on different tests on these outputs.

6.5 Tests for the evaluation

There are number of tests available for evaluating MT systems. Slype (1979) describes subjective tests and quantitative tests for evaluation of an MT system. Subjective tests
like adequacy and fluency tests have been performed on the proposed system. BLEU score has been calculated to evaluate the quality of output. The quantitative test, namely, error analysis has also been performed by calculating Sentence Error Rate (SER) and Word Error Rate (WER) of the proposed UNL based MT system for Punjabi language.

6.5.1 Fluency test
Fluency refers to the degree to which the target is well formed according to the rules of target language grammar. A fluent segment is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of that language (LDC, 2005). There are number of discussions available in the literature regarding the choice for point of scale in the evaluation. It has been suggested that one can use 5-point scale for these tests (LDC, 2005). Singh et al. (2007) and Goyal (2010), however, suggested that 5-point scale is a too fine-grained scale and in this scale the distinction may result in evaluators worrying a lot about making an accurate call, and intuitive judgment may get affected. It also makes the evaluation even more subjective. Thus, a 4-point scale has been followed for evaluation of the proposed system as followed by Singh et al. (2007) and Goyal (2010). The details of 4-point scale for the fluency score is given in Table 6.1.

Table 6.1: 4-point scale of fluency score

<table>
<thead>
<tr>
<th>Score</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Perfect</td>
<td>Good grammar</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Easy-to-understand but flawed grammar</td>
</tr>
<tr>
<td>2</td>
<td>Acceptable</td>
<td>Broken, understandable with effort</td>
</tr>
<tr>
<td>1</td>
<td>Nonsense</td>
<td>Incomplete</td>
</tr>
</tbody>
</table>

6.5.2 Adequacy test
Adequacy refers to the degree to which the information present in original sentence is also communicated in translated sentence. Thus, for adequacy judgments, the gold-standard will serve as a proxy for the original source language text. For the evaluation of proposed Punjabi DeConverter, the sentence generated by the system is compared with the input Punjabi sentence of EnConverter, whose UNL expression is fed to the Punjabi
DeConverter for DeConversion. Thus, for the adequacy test the evaluator is also presented with the gold-standard translation, \textit{i.e.}, the original input Punjabi sentence to Punjabi EnConverter, with reference to which they can compare the generated output. Again, a 4-point scale has been followed for the adequacy test of the proposed system as followed by Singh \textit{et al.} (2007) and Goyal (2010). The details of 4-point scale for the adequacy score is given in Table 6.2.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>All</td>
<td>No loss of meaning</td>
</tr>
<tr>
<td>3</td>
<td>Most</td>
<td>Most of the meaning is conveyed</td>
</tr>
<tr>
<td>2</td>
<td>Some</td>
<td>Some of the meaning is conveyed</td>
</tr>
<tr>
<td>1</td>
<td>None</td>
<td>Hardly any meaning is conveyed</td>
</tr>
</tbody>
</table>

6.5.3 BLEU score
Bilingual Evaluation Understudy (BLEU) score is commonly used for automatically evaluating the quality of output produced by an MT system. BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Scores are calculated for individual translated segments, generally sentences, by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to estimate the translation’s overall quality. In BLEU score intelligibility or grammatical correctness are not taken into account. BLEU’s output is always a number between 0 and 1. This value indicates how similar the original and translated sentences are. In case BLEU score is closer to one, the two sentences are closer to each other.

During the processing of BLEU score the \textit{n}-grams of the translated sentence is compared with the \textit{n}-grams of the original sentences and system counts the number of matches between both the original and translated sentences. These matches are position independent. If we have higher matches, the quality of translation system is better. The metric calculates scores for individual sentences, and then averages these scores over the whole corpus in order to reach a final score (Goyal, 2010).
6.5.4 Tests based on quantitative metrics

Quantitative metrics are based on the analysis of errors made by the MT system. It performs error analysis to establish how seriously errors affect the translation output. The error analysis includes Word Error Rate (WER) and Sentence Error Rate (SER). Word Error Rate (WER) is defined as percentage of words which are to be inserted, deleted, or replaced in the translation in order to obtain the original sentence. Sentence Error Rate (SER) is defined as percentage of sentences, whose translations have not matched in an exact manner with those of reference (Goyal, 2010).

6.6 Experiments

It is also very important to choose appropriate evaluators for performing the tests on the proposed system. The proposed system has been evaluated by ten evaluators having a good knowledge of Punjabi and a basic knowledge of UNL system. As suggested in LDC (2005) and by Singh et al. (2007), the evaluators were asked to provide their intuitive reaction to the output and to work as quickly as comfortable. First of all, fluency judgments have been taken. The evaluators did not have any clue about the original source sentence for these judgments. After fluency judgments, the evaluators were asked to look at the original source sentences for adequacy judgments.

6.7 Results for fluency score

The fluency score of proposed system is 3.61 (on a 4-point scale). The response by evaluators is further analyzed and following are some important findings for the fluency test.

- 73.50% sentences got a score 4, i.e., these are perfect having no grammatical mistake.
- 15.50% sentences got a score 3, i.e., these are fair and easy to understand.
- 9.50% sentences got a score 2, i.e., these are acceptable and are understandable with effort.
- 1.50% sentences got a score 1, i.e., these are hard to understand.

Here, it is worth mentioning that proposed system generated 89.0% grammatically correct sentences. These sentences are those that have a score of 3 or above.
6.8 Results for adequacy score

The adequacy score of proposed system is 3.70 (on a 4-point scale). The response by the evaluators has again been further analyzed and following are some important findings for the adequacy test.

- 79.2% sentences got a score 4, *i.e.*, there is no loss of meaning during their translation.
- 12.8% sentences got a score 3, *i.e.*, most of the meaning of these source sentences is conveyed in the translated sentence.
- 7.0% sentences got a score 2, *i.e.*, some of the meaning of these source sentences is conveyed in the translated sentence.
- 1.0% sentences got a score 1, *i.e.*, hardly any meaning of these source sentences is conveyed in the translated sentence.

Here, it is worth mentioning that proposed system generated 92% sentences that are faithful to the original sentences. These sentences are those that have a score of 3 or above.

We have also observed that there is a strong correlation between fluency and adequacy scores. This relation between adequacy and fluency is explored further in Figure 6.3. It shows the distribution of adequacy scores for various values of fluency.

![Figure 6.3: Distribution of adequacy scores for various fluency scores](image-url)
6.9 Results for BLEU score

The quality of proposed system is also evaluated on the basis of BLEU score. The original sentences inputted to Punjabi EnConverter have been considered as reference sentences and corresponding sentences generated by Punjabi DeConverter have been considered as machine generated sentences. The proposed machine translation system is able to achieve a BLEU score of 0.72.

6.10 Error Analysis

The error analysis has also been performed for the proposed system after calculating fluency, adequacy and BLEU score. The error analysis has been performed for classified error list that includes the errors like addition of extra words, removal of words, wrong order of words and wrong choice of words in the generated output. All these errors in generated output have been identified and noted. After robust analysis, word error rate is found to be 5.43%. The results of word error rate are depicted in the graph given in Figure 6.4.

![Word Error Analysis](image)

Figure 6.4: Word Error Analysis of the proposed system

From the above figure, it can be concluded that majority of the errors are due to wrong word order. The order of words in the generated sentence is decided by the syntax planning phase of Punjabi DeConverter. This order is governed by a matrix based on priority of UNL relations sharing a common parent. Since, Punjabi is a free order language the change in the word order will not severely affect the accuracy of the system. The function word insertion rule base of Punjabi DeConverter is responsible for insertion
of extra words in the generated output. The morphology generation rules are responsible for removal of words and wrong choice of words in the generated output.

After word error rate analysis, the sentence error rate analysis has been performed by calculating the number of sentences, whose translations have not matched in an exact manner with those of reference sentences. The proposed system has a sentence error rate of 20.8%.

However, in case of Punjabi EnConverter the wrong parsing of input sentence by Punjabi shallow parser results into generation of incorrect UNL expression, which causes errors in the generated output. Sometimes, the non availability of semantic information of verbs in lexicon also causes incorrect firing of rules, which result into incorrect output.
Chapter Summary

The proposed system has been evaluated on the basis of subjective tests and quantitative metrics. From the analysis in this chapter, it has been concluded that proposed system generates 89.0% grammatically correct sentences; and generates 92.0% sentences that are faithful to the original sentences. The system could achieve a fluency score of 3.61 (on a 4-point scale) and adequacy score of 3.70 (on a 4-point scale). The proposed system is able to achieve a BLEU score of 0.72. The proposed system has a word error rate of 5.43% and sentence error rate of 20.8%. These scores of the proposed system can be improved further by improving the rule base and lexicon.