Abstract

The Hindi-Dogri Machine Translation System is based on the direct approach of Machine Translation (MT). It translates Hindi text into Dogri. This paper presents in brief the development of the system and details of testing and the evaluation of the results of this system.

Keywords: Accuracy Testing, Hindi-Dogri MTS, Intelligibility Testing, Quantitative Testing

1. Machine Translation

Machine Translation is the automatic translation of one language called the source language into another language called the target language. It can be applied both to text and speech. The idea of automated translation was conceived in 1947. In July 1949, Weaver marked the beginning of machine translation by bringing the idea of MT to general notice and within a few years research began at the University of Washington (Seattle), at the University of California at Los Angeles and at the Massachusetts Institute of Technology (MIT). Presently multilingual machine translation systems are being developed. Newer MT techniques are being used to yield much better results.

In India, research on machine translation began somewhere in mid 80’s. Currently, many machine translation systems are available for the translation of various Indian languages. Most of these systems are for English-Hindi or Indian language to other Indian language besides Hindi. Majority of the Indian machine translation systems have been developed for Hindi-English language pair. Many research centres are now focusing on the translation of foreign languages to Indian languages and the development of multilingual translation systems.

2. Translation Scenario in Jammu and Kashmir

The state of Jammu and Kashmir has a diversity of languages. The languages which are majorly spoken in the state are: Urdu, Dogri and Kashmiri. The translation of Hindi and English text into Urdu is available online. Google has its translation tool where English as well as Hindi can be converted to Urdu. Kashmir University is also working on the development of machine translation tools for Kashmiri.

No work of machine translation has been done on Dogri; though the Department of Information Technology (DIT), India has got some work done on software localization in Dogri. Some of the software tools that are available in Dogri are: Open Office, Firefox, Thunderbird, Pidgin messenger, Sunbird calendar, scribus page layout application etc. The present system is the first machine translation system developed for translation of Hindi text to Dogri.

3. Introduction to Dogri

Dogri is written using Devanāgarī script and has thirty eight segmental and five supra segmental phonemes.

*Author for correspondence
Segmental phonemes have been divided into two broad groups i.e. vowels and consonants. It has ten vowel phonemes and twenty eight consonant phonemes⁵. Some peculiarities of the language are:

- Dogri has the same basic consonants as Devanāgarī; but घ (gh), ङ (jh), ठ (dh), ड (dh), ढ (bh) exist only in Dogri orthography. Phonetically, they are used for tonal क (k), च (c), ट (t), ठ (t) at initial stage of a word e.g. घर (ghar/house), झड़ना (jharanā/ shed), ठाक़न (dhakkān/ lid), धमाकी (dhamaḵi/ threat), बहरी (bharī/ heavy) etc.
- In Dogri छ and ज phonemes are also used in the initial position of a word e.g. छूरा (ṇūṭā/ thumb), ज्ञाण (ṇāṇā/ infant), जार (ṇārām/ eleven); which is not in the case of Hindi.
- The use of chandrabindu ԋ and visarga ԋ is not prevalent in Dogri.
- ओ, ए, ऐ, झ, ञ are only used for transliteration of some Sanskrit words e.g. उष्ण (uṣṇa/ saint), ज्ञान (jnāna/ knowledge), कृष्ण (krṣṇa).
- Tone is a major supra-segmental feature of Dogri. There are three distinctive tones in Dogri namely: level tone, low or low-rising tone, high or high falling tone.
- In Dogri orthography apostrophe comma serves double purpose to:
  ➢ Single apostrophe comma is very frequent in the language and is referred as the ‘SURCHINHA’ in Dogri. It is used to express high falling tone after short vowel. For example: जिया (ji'yā /like), खळ्ला (kha'lla /below) etc.
  ➢ Its second purpose is to indicate syncopated forms, For Example: जिसले (This word is a combination of (jīṣ + बेल्लै) meaning ‘in the evening.’ The first apostrophe (between the shiro rekha/line) indicates high falling tone and the other (above the rekha/line) shows the syncopated form.
- Halant ह/ /h/ is also used to represent the tone e.g. ओह (oh / that),पीह्ना (pihna/to grind).
- In Dogri, an avagraha sign $, is used to indicate extra-long vowels e.g. लगा (lagā/began) लगाई (lagā' /affection)
- Numerals in Hindi and Dogri are written as below:
  Hindi: ० १ २ ३ ४ ५ ६ ७ ८ ९
  Dogri: 0 1 2 3 4 5 6 7 8 9⁶

4. Hindi-Dogri Machine Translation System

The Hindi-Dogri machine translation system has been developed using the direct approach of machine translation. It has been implemented in ASP.Net and the databases are in MS-Access with Hindi and Dogri text in Unicode format. This Machine Translation system accepts Hindi text as input and outputs Dogri text.

5. System Architecture

The Hindi-Dogri machine translation system consists of the following phases; each phase has its sub activities. This section discusses these activities in brief.

5.1 Pre-processing Phase

In this phase, the source text undergoes preprocessing activities. These activities may vary in different systems depending upon the requirement of preprocessing. The pre-processing phase in this system involves three activities: Text Normalization, proper noun identification and identification of collocations⁴.

5.1.1 Text Normalization

It is the first activity done in the pre-processing phase. In Hindi, there are many ways of spelling the same word. In this phase a standard word for all the variant words is chosen. This system has a collection of 415 words in the text normalization database.

5.1.2 Proper Noun Identification

In this sub phase, proper nouns such as names of persons, institutions, rivers, banks etc are identified in the source text, so that they are not translated. This sub phase is very important for good accuracy of the system.

5.1.3 Identifying Collocations

After checking for proper nouns, the source text undergoes this sub phase for identifying collocations. Collocations are combinations of words that cannot be translated word to word. If these words are translated, the word sense is changed. These words have a different meaning as an independent word and convey a different meaning by association. In this sub phase, the source text is checked for the presence of collocations, if found there meaning is taken
from the collocation database and the word is not sent further for lexicon look up. Presently this system has a collection of 1823 collocations and can further be extended.

5.2 Tokenization
The output of the preprocessing phase goes to the tokenizer, which extracts individual words from the sentence for further processing to generate its equivalent in the target language. The tokens are extracted from the text using space as a delimiter.

5.3 Target Text Generation
The tokens generated by the tokenizer are then forwarded to the translation engine for further processing. This phase consists of the following sub phases:

5.3.1 Word to Word Translation
The words in the source string that have not matched with title, collocation or surname database are translated word by word by referring the ‘dict’ database which is a Hindi-Dogri bilingual dictionary. Presently, this database has a collection of 18512 unannotated words.

5.3.2 Ambiguity Resolution
After lexicon look up, the source text undergoes ambiguity resolution. This system resolves ambiguity arising due to the morphemes और/aur and से/se using n-gram approach. This approach has demonstrated good performance in this system. Among 107 occurrences of the postposition से/se in the test data 91 were translated correctly by this system. Also, the total occurrence of the morpheme और/aur in the test data has been calculated to be 73; out of which 72 have been correctly translated. Therefore, it can be said that the n-gram approach has given considerable results.

5.3.3 Inflectional Analysis
A rule based approach has been followed for inflectional analysis. The rules for inflectional analysis have been derived from the morphological analyzer developed by IIIT, Hyderabad. This morphological analyzer works for Linux platform. It was converted to work on Windows platform and then inflection rules were extracted from it for Hindi language by Punjabi University, Patiala. These rules have been used for writing the equivalent rules for Dogri inflections.

5.3.4 Handling Special Cases of Dogri
Finally, before the output generation the text is checked for the following peculiarities of Dogri.

5.3.4.1 Handling Kar (कर)
The morpheme ‘Kar/ कर’ is used very frequently in Hindi and it takes different forms in Dogri. Therefore, it is important to handle it. In Hindi language, the word Kar (कर) takes the following two forms.
- As a morpheme it means ‘do’.
- As conjugative participle in compound verb.

In Dogri, कर (kar) is replaced with:
- जाइयै (jāiyai) for vowel ending words e.g. रखकर/rakh kar (रक्खयै/rakkhiyai).
- इयै for consonant ending e.g. जाकर/ jākar (जाइयै/ jāiyai).

5.4 Handling Kar (कर) Preceding Echo Words
Whenever Kar (कर) appears after echo words, it takes the following form in Dogri. An example to illustrate this is given below: The Hindi phrase भाग-भाग कर/ bhāg-bhāg kar will be translated in Dogri as नरसी-निससयै/nassī-nassiyai.

5.4.1 Handling Words Preceding the Morpheme लगा/लगी/ lage/
lage/

In Hindi sentences, the word लगा,/lagā is used to give the sense of continuity. When translating a sentence that contains लगा/लगी, lage/ in Dogri, आ/ā is removed from the words preceding lagā/lage, lage/lage.

5.4.2 Handling Words Preceding the Morpheme रहा/rahā; रही/rahi; रहे/rahe

The Hindi morpheme रहा/rahā takes the following forms in Dogri:
- If the word preceding रहा/rahā is vowel ending then आ/ā is added to it and रहा/rahā is changed to दा/dā.
- Else if it is consonant ending आ/ā is added to it and रहा/rahā is changed to दा.
- If the word is रही/rahī, then as in the previous case, आ/ā or आ/ā is appended to the preceding word and रही/rahī is changed to दी/dī.
• If the word is रहे, then like the other two cases आ (ा) or आ is appended to the preceding word, but रहे takes two forms depending on the word next to it. The two cases are described below:
  - If रहे/rahe is followed by हो/ho then, रहे हो is replaced with ने.
  - If रहे/rahe is followed by हैं/haim then, रहे is replaced with दे.

5.5 Output Generation

After going through all the above mentioned phases, the output of the system is generated. The following section discusses the testing of the system.

6. Evaluation Approaches

The evaluation of a MTS is an important task as it analysis the performance of the system. The variability in language makes evaluation a difficult task. i.e. a sentence in source language may have more than one correct translation in the target language. Several researchers have worked on evaluation techniques of Machine Translation systems and many measures and methods have been developed for this purpose. Broadly the evaluation of any machine translation system can be done in two ways namely: human evaluation and automatic evaluation.

Human evaluation of automatic translation is done by humans manually. The aim of evaluation is to judge the automatic translation relative to human translation. Automatic Evaluation is based on some metric for judging the translation with the best human judgments. In this type of evaluation, an algorithm is used for evaluating texts. Some methods for automatic evaluation are: BLUE, NIST and METEOR. Automatic evaluation techniques for evaluation of machine translation systems are more popular now-a-days.

6.1 Approach Followed for the Evaluation of Hindi-Dogri MTS

The present system is the first system developed for the language pair undertaken. There were no prior lexical resources available that could be used for this work. Therefore the system could not be tested using the automatic evaluation techniques, due to non-availability of Hindi-Dogri parallel corpus. This system has been tested using the human evaluation technique. This technique has been earlier used for evaluating the Hindi to Punjabi MTS and also for evaluating Punjabi-Hindi MTS.

The following section gives the details of the testing procedure.

7. System Testing

The system testing procedure comprises of selection, implementation and evaluation. Selection involves the choice of the test data, evaluators for testing, tests to be performed and the scoring methodology. The chosen tests are then implemented on the test data. Finally, the system is evaluated by calculating its efficiency.

7.1 Selection

7.1.1 Selection of Test Data

The test data has been selected using random sampling technique. Input sentences have been selected randomly from online resources such as news (sports, politics, world, regional, entertainment, travel etc.), articles (published in online newspapers), and literature (short stories by Sai Baba, Prem Chand and Yashwant Jain etc.) Sentences of simple as well as complex constructs have been incorporated in the test data set. A total of 120 documents (including news, articles and stories) were evaluated. These 120 documents were divided into half (i.e. 60 each) one set each for Intelligibility and accuracy testing. The following Table 1 shows the test data set:

| Test Data Set | Daily News | Articles | Stories | Total |
|---------------|------------|----------|---------|-------|
| Documents     | 60         | 40       | 20      | 120   |
| Sentences     | 127        | 117      | 245     | 489   |
| Words         | 2,393      | 1,589    | 3,277   | 7,259 |
Testing and Results of Hindi-Dogri Machine Translation System

include error analysis i.e. Sentence Error Rate (SER) and Word Error Rate (WER).

7.2 Subjective Tests

7.2.1 Intelligibility Tests

Intelligibility is defined as a measure of how understandable a sentence is. It is measured without the reference to the original sentence. This test checks for grammatical errors, mis-translations, and untranslated words.

7.2.2 Accuracy Tests

Accuracy Test or Fidelity measure is a measure of how much information is retained in the translated sentence as compared to the original. A highly intelligible output sentence may not be a correct translation of the source sentence. In this test it is checked whether the meaning of the source sentence is preserved in the translated sentence. This property is called Accuracy or Fidelity

7.3 Quantitative Tests

7.3.1 Word Error Analysis

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 sentence of reference. In this test the entire test data is checked for word errors.

7.3.2 Sentence Error Analysis

Sentence Error Rate (SER) is defined as percentage of sentences, whose translations have not matched in an exact manner with those of reference.

7.4 Scoring Methodology

The scoring methodology for intelligibility and Accuracy test has been adopted described by Van Slype Georges. The same methodology has been used for Hindi-Punjabi MTS and Punjabi-Hindi MTS. Table 2 and Table 3 show the scale for intelligibility test and accuracy test respectively.

8. Test Implementation

8.1 Intelligibility Test

In this test, the evaluators were given sheets containing only the target text. The evaluators did not have any clue about the source text. They judged each sentence on the basis of its comprehensibility. The intelligibility test sheet consisted of 30 news, 20 articles and 10 stories. These sheets were evaluated by 100 people. Average ratings for the sentences of the individual translations were then summed up to get the average scores. Percentage of intelligent sentences was calculated by counting the number of intelligent sentences. The scoring scale used for evaluation is given in Table 2 and the intelligibility test sheets are given in the appendix A.

8.2 Accuracy Test

In this test, the evaluators were provided with source text along with translated text for evaluation. They judged the target text with reference to the source text. The accuracy test sheets also consisted of 30 news, 20 articles and 10 stories. They were also evaluated by 100 people using the scale given in Table 3 and the test sheet given in Appendix B. Average ratings for the sentences of the individual translations were again then summed up to get the average scores and percentages calculated as mentioned above.

Table 2. Shows the four point scale used for Intelligibility testing

| Score | Significance |
|-------|--------------|
| 3     | The sentence is perfectly clear and intelligible. It is grammatically correct. |
| 2     | The sentence is generally clear and intelligible. Despite some inaccuracies, the information is conveyed. |
| 1     | The general idea is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choice. |
| 0     | The sentence is unintelligible. The meaning of the sentence is not understandable. |

Table 3. Shows the four point scale used for Accuracy testing

| Score | Significance |
|-------|--------------|
| 3     | Completely faithful |
| 2     | Fairly faithful: more than 50% of the original information passes in the translation. |
| 1     | Barely faithful: less than 50% of the original information passes in the translation. |
| 0     | Completely unfaithful. Doesn’t make sense. |
9. Results

9.1 Intelligibility Testing
The results of the responses given by the evaluators are given below (Figure 1):

- 70.78% sentences got the score 3 i.e. they were perfectly clear and intelligible.
- 27.76% sentences got the score 2 i.e. they were generally clear and intelligible.
- 1.43% sentences got the score 1 i.e. they were hard to understand.
- 0.02% sentence got the score 0 i.e. they were not understandable.

9.1.1 Analysis
According to the score given by the respondents, 98.54% sentences are intelligible. These sentences are those which have score 2 or above. One reason for such a high intelligibility score is the use of the same script by the language pair. Therefore there are hardly any sentences which are not understandable. Thus, we can say that the direct approach can translate Hindi text to Dogri Text with good accuracy.

9.2 Results of Accuracy Testing
The results drawn from the responses of the respondents are (Figure 2):

- 69.90% sentences achieve a match between 50 to 99%.
- 28.81% of remaining sentences were marked with less than 50% match against the correct sentences.
- Only 1.27% sentences are those which are found unfaithful.
- 0.017% sentences got Score 0.

9.2.1 Analysis
According to the score given by the respondents, 98.71% sentences are accurate. These sentences are those which have score 2 or above. The results of the accuracy test are more than satisfactory.

9.3 Word Error Rate (WER)
All errors in translated text were identified and their frequencies were noted. After robust analysis, Word Error rate in the test data was found to be 2.011%. Following Table 4 shows the percentage type of errors out of the total errors found:

The following Figure 3 shows the percentage of different type of errors

9.3.1 Analysis
From the table given above, it is observed that errors belong to the following categories:

- Untranslated words which imply that the size of the bilingual dictionary should be increased.

Table 4. Shows the WER analysis of the test data

| Error Category               | No. of Erroneous Words | Total Words | Percentage |
|------------------------------|------------------------|-------------|------------|
| Wrongly translated Words     | 61                     | 7259        | .84%       |
| Untranslated Words           | 66                     | 7259        | 0.90%      |
| Addition or Deletion Required| 19                     | 7259        | 0.26%      |
| Total Errors                 | 146                    | 7259        | 2.011%     |
Testing and Results of Hindi-Dogri Machine Translation System

9.4 Sentence Error Rate (SER)
The Sentence error rate is calculated as 16%. Following Table 5 shows the sentence errors for various test sets:
The following Figure 4 compares the wrong sentences with the total sentences in each category.

9.4.1 Analysis
The formation of wrong sentence is due to the following:

9.4.1.1 Gender Disagreement
During translation, sometimes correct gender of a word is not reflected in the translated language and it causes gender disagreement with verb/postposition in the target language. For example:

- Input sentence: उसे एक उपाय सूझा।/ use eka upāya sūjhā./He got an idea.
- Incorrect Output: उससी इक जुगती सूझी।/ ussī ika jugatī sūjhī.
- Correct Output: उससी इक जुगती सूझी।/ ussī ika jugatī sūjhī.

9.4.1.2 दा/da Postposition Agreement before Verb Phrase
In some cases, the translated text shows disagreement of postposition दा with all the Verb phrases in the sentence e.g.:

- Input sentence: वर्धमान महावीर का जन्मदिन आज है।/ vardhamāna mahāvīra kā janmadina āja hail /
- Incorrect Output: वर्धमान महावीर दा बरसगंध अन्ज ई।/ vardhamāna mahāvīra da barasagamdha āja hail
- Correct Output: वर्धमान महावीर दा बरसगंध अन्ज ई।/ vardhamāna mahāvīra da barasagamdha āja hail

The overall accuracy of Hindi to Dogri Machine Translation system is found to be 98.54% on the basis of intelligibility test and 98.71% on the basis of accuracy test with an error rate of 2.011% and SER calculated as 16%.

10. Comparison with Other Existing Systems
The results of the present system have been compared with Punjabi-to-Hindi MTS, Hindi-to-Punjabi MTS. These systems are also based on the direct approach of
machine translation. The Following Table 6 shows the comparison among various existing systems:

## 11. Conclusion

The Hindi-Dogri machine translation system is based on direct approach of translation. The overall accuracy of Hindi to Dogri Machine Translation system is found to be 98.54% on the basis of intelligibility test and 98.71% on the basis of accuracy test with an error rate of 2.011% and SER calculated as 16%. Therefore it can be said that the system output needs approx. 2% editing by a human editor. Evaluation of this system strengthens the fact that direct approach of machine translation is suitable for closely related languages.

## 12. Future Scope

This system is the beginning of NLP research on Dogri. In Future, the lexical resources developed for this system can be enhanced. The Hindi-Dogri dictionary can be annotated to handle issues like gender disagreement. The word sense disambiguation module can also be enhanced by including polysemous words. This system can be used to develop parallel corpus, which can be used for automatic evaluation of the system. Parallel corpus is also required for newer research techniques like statistical machine translation.

## 13. References

1. Arnold, et al. Machine Translation: An Introductory Guide. London: NCC Blackwell. 1994. ISBN: 1855542-17x.
2. Sudip Naskar, et al. Use of machine Translation in India: Current Status. Phuket, Thailand: Proceedings of MT SUMMIT X. 2005 Sept 13-15. p. 465–70.
3. Available from: http://www.kashmirizabaan.com.
4. Dubey P, Devanand. Machine translation System for Hindi-Dogri language pair. Katra: Proceedings of IEEE conference ICMIRA. 2013 Dec 21–23.
5. Dubey P, Pathania S, Devanand. Comparative study of Hindi and Dogri languages with regard to Machine Translation. Language In India. 2011 Oct 10; 11. ISSN 1930-2940.
6. Goyal V. Development of a Hindi to Punjabi Machine Translation System [PhD Thesis]. Patiala: Department of Computer Science, Punjabi University. 2010.
7. Goyal V, et al. Evaluation of Hindi to Punjabi Machine Translation System. IJCSI International Journal of Computer Science Issues. 2009; 4(1).
8. Josan GS, et al. Evaluation of direct machine translation system for Hindi to Punjabi. Available from: learnpunjabi.org/pdf/directtrans.pdf.
9. Slype V. Critical study of methods for evaluating the quality of machine translation. 1979 Nov 30. Available from: http://www.issco.unige.ch/en/research/projects/isle/van-slype.pdf
10. Somers HL. Current research in Machine Translation. Machine Translation. 1992-1993; 7(4):231–46.