Machine Translation system for Standard Punjabi to Malwai Dialect

Harjeet Singh
Research Scholar
Punjab Technical University
Jalandhar
zrjeet@gmail.com

Vishal Goyal
Dept. of Computer Science,
Punjabi University
Patiala
vishal.pup@gmail.com

Ravinder Khanna
Dean (R&D)
MM University,Sadopur
Ambala
ravikh_2006@yahoo.com

Abstract

A lot of work is done on the standard varieties of Indian languages in the field of MT but dialectal variety of a language is still an unexplored area. Machine Translation between a Standard language and its dialect is easy as standard variety and its dialect are closely related to each other. When the Language pair is closely related then due to the common grammar and vocabulary, it become easy to develop a MT system.

In Punjab generally Malwai, Majhi, Doabi and Powadi dialect are used for the oral communication and no work is still done on any dialect in the field of Machine Translation. This Paper discusses the various phases in the development of Machine Translation system for Standard Punjabi - Malwai dialect pair.

Keywords: Machine Translation, Malwai dialect,

1. Introduction

Machine Translation (MT) is a process in which computer software is used to translate a text from one natural language called the source language(SL) into the another language called the target language(TL). There are number of approaches of Machine Translation and which approach to be used depends upon the resources available and language pair involved. Some approaches of MT are:

1.1 **Rule Based Machine Translation** is based on collection of grammar rules of source and target languages, a bilingual or multilingual lexicon. The rules are generally written with linguistic knowledge collected from linguistics. When the text is inputted in this system then on the basis of morphological, syntactic and semantic analysis of both the source and target languages, the output sentence is generated. Rule based approaches are categorized as Direct, Transfer and Interlingua approaches.

1.2 **Corpus Based Machine Translation** is based on the huge amount of raw data in the form of bilingual parallel corpus. This data contains text and their corresponding translations. This approach is further classified as Statistical based MT and Example based MT.

1.3 **Hybrid Machine Translation** combines the benefits of more than one approach for better results. Presently, the hybrid approach of statistical MT and Rule based MT is used. The hybrid approach can be used in many ways. Sometimes, Rule are used in the first step of translation and statistical information is used to improve the results later on. In other cases, the rules are used for pre-processing the source data and post-processing the target test generated by Statistical MT system.

2. Approach Used for Standard Punjabi- Malwai dialect MT system:

The comparative study of Standard Punjabi and Malwai dialect has been done to know the difference between the pair. It is concluded from the study that Standard Punjabi and Malwai dialect are closely related. Both are similar with regard to script, word order, lexicon and grammar but the difference lies between inflection of words. The literature survey on Machine Translation of Closely related languages suggests that Direct approach of Machine Translation is suitable for our system.

Some system which are developed using this Direct approach are: GAT system for English-Russian , MARK-II system for Russian-English , LOGOS system for English to Vietnamese, SYSTRAN system for English to Russian, CULT system for Chinese to English, ALPS system for English to French, German and Portuguese, RUSLAN system for Czech to Russian, CESILKO system for Czeck to Slovak, Tatar system for Turkish to Crimean, Punjabi to Hindi Machine Translation system.

The Direct approach of Machine Translation is based on word to word translation. This approach requires bilingual dictionary of the language pair. There was no machine readable dictionary available for the Standard Punjabi - Malwai dialect but the dictionaries are available which explains the meaning of the Malwai words. So the maximum time of the research has been depleted for the preparation of machine readable dictionary for Standard Punjabi - Malwai dialect and to gather the linguistic data.

3. Phases in Machine Translation of Standard Punjabi to Malwai Dialect:

The Machine Translation of Standard Punjabi to Malwai dialect has the following phases:

- ✓ Preprocessing of the source Text(Standard Punjabi),
- ✓ Grammatical Translation Rules
- ✓ Lexicon lookup
- ✓ translation of the text to the target(Malwai dialect) language,

The various phases are discussed as follows:

3.1 **Pre-processing of the source text**: It involves identification of collocations and proper nouns in the source text. These are discussed as follows

3.1.1 **Identifying Collocations**: Collocations means two or more words that cannot be translated word to word. If these are translated the word sense is changed. The correct Malwai dialect translation is stored in the collocation table of the database.

For example the collocation ਉੱਤਰ ਪ੍ਰਦੇਸ਼ (uttar pradesh) . If this collocation will be translated word by word, it will be translated to ਜਵਾਬ ਪ੍ਰਦੇਸ਼ but it will be translated to ਉੱਤਰ ਪ੍ਰਦੇਸ਼.

3.1.2 **Identifying Proper Noun:** After the identification of collocation phase in the inputted text, the system extracts proper nouns like personal, days of month, days of week, country names, city names, bank names, organization names, ocean names, river names, university names etc. from the inputted text. If these words are translated word to word their meaning is changed. So for this purpose two databases namely titles and surnames are referenced.

In this phase the system checks for abbreviation sign(.) in the inputted text and the word next to the sign is treated as name so therefore not translated.

For example ਡਾ. ਆਨੰਦ / Dr. Anand. The word ਆਨੰਦ means ਮਜਾ in Malwai dialect. So Anand will not be translated and will be kept same in the translated text as it is in the source text.

If the name is not preceded by any title then the name is identified using the surname database. When a surname in source string in the inputted text is matched with the item in the surname database, the word before that is supposed to be a name and not translated.

3.2 **Grammatical Translation Rules**: The words that depends upon the previous and the following words must be translated before the root in order not to lose the context information. The rules are defined in this phase to handle such words.

3.3 **lexicon lookup :** The output of the pre-processing phase is sent to the Tokenizer also known as lexical analyzer that divides the given sentence based on the spaces between them into units called tokens. The tokens are inputted to the translation engine for the generation of the target text. The translation engine consist of bilingual dictionary.

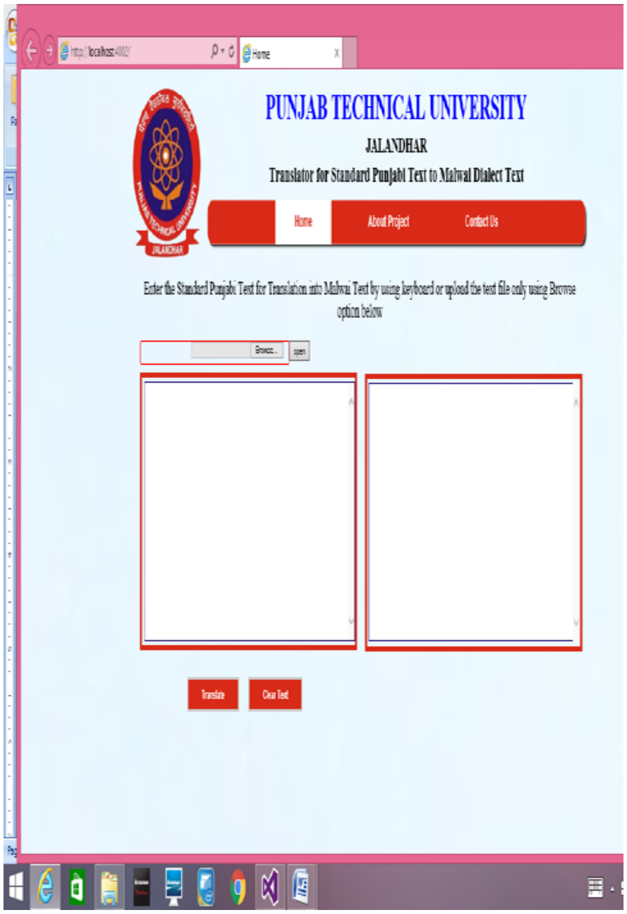
If **t**he token provided by the Tokenizer is not title or surname then those are translated by looking up in the dict database which is Standard Punjabi - Malwai dialect dictionary. The dict database consist of around 8000 entries. If no entry is found in the dict databases then it is passed on without translation.

3.4 **Output Generation:** After all the phases, the final output in the target language(Malwai dialect) is generated.

4. Implementations

This machine translation system has been implemented in ASP.Net and the databases are in MS-Access with Standard Punjabi and Malwai text in Unicode format. This Machine Translation system accepts Standard text as input and provides Malwai text as output in Unicode. The formal testing of the system i.e. Intelligibility and Accuracy testing has not been done yet.

The sample run of the system is as follows:


PUNJAB TECHNICAL UNIVERSITY

JALANDHAR

Translator for Standard Punjabi Text to Malwai Dialect Text

Enter the Standard Punjabi Text for Translation into Malwai Text by using keyboard or upload the text file only using Browse option below

5. Conclusion:

The Paper discusses in brief about the various phases in the machine translation of Standard Punjabi to Malwai dialect. This is the first system to be developed for translating Standard Punjabi to Malwai dialect. This tool can be used to develop machine translation system for other dialects of Standard Punjabi also.

References

1.http://en.wikipedia.org/wiki/Machine_translation [Accessed on 18-01-2013.]

2. V.Goyal , G.S. Lehal, “Web Based Hindi to Punjabi Machine Translation System”, *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 2, pp. 148-151[2010].

3.G.S. Josan, G. S.Lehal, “A Punjabi to Hindi Machine Translation System”, in the *proceedings of the 22nd International Conference on Computational Linguistics, (COLING ’08)* pp. 157-160[2008].

4. R Harshawardan, “Rule based machine translation system for English to Malayalam Language”. A Master thesis dissertation Amrita Vishwa Vidyapeetham Coimbatore[2011].

5. Vishal Goyal, " Development of a Hindi to Punjabi Machine Translation System" Phd Thesis, Department of Computer Science, Punjabi University, Patiala. Available at: *www. academia.edu/482683/LANGUAGE_IN_INDIA.*

6. G.S. Joshan, " A Punjabi to Hindi Machine Translation System", Phd Thesis,

Department of Computer Science, Punjabi University, Patiala[2011].

7. Kamaljeet Kaur, "A Punjabi to English Machine Translation System for legal Documents", Phd Thesis, Department of Computer Science, Punjabi University, Patiala.

8. S. Dwivedi and P.P. Sukhdeva“ Machine Translation System in Indian Perspective”. *Journal of Computer Science 6(10). ISSN 1549-3636*. pp 1111-1116[2010].

9. Yves Scherrer, Owen Rambow “Word-based dialect identification with georeferenced rules”[2010]. *Extracted from http://delivery.acm.org/10.1145/1880000/1870770 / p 1151 scherrer.pdf?ip=117.205.57.154&acc=OPEN&CFID=266881214&CFTOKEN=27883754&__acm__=1359391097_d48fb85ebd91aaaea1e15b2af9481ecd*

10. Kemal Altintas et.al.," A Machine Translation System Between a Pair of Closely Related Languages". extracted from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.1634&rep=rep1&type=pdf

11. Jan HAJIC et.al., "Machine Translation of very close languages" extracted from http://www.aclweb.org/anthology/A00-1002 accessed on 22-01-2013.

12. Petr Homola et.al.," Improving Machine Translation Between Closely Related Romance Languages" extracted from http://mt-archive.info/EAMT-2008-Homola.pdf accessed on 22-01-2013.

13. Kevin P. Scannell," Machine translation for closely related language pairs" extracted from http://borel.slu.edu/pub/ga2gd.pdf accessed on 22-01-2013.

14. Jan Hajic et.al., " CESILKO – an MT system for closely related languages" extracted from http://mt-archive.info/ACL-2000-Hajic.pdf accessed on 23-01-2013.

15. Vladislav Kubon et.al," A Comparison of MT Methods for Closely Related Languages: a Case Study on Czech – Slovak Language Pair" extracted from http://alt.qcri.org/LT4CloseLang/pdf/LT4CloseLang12.pdf accessed on 23-01-2013.