DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY PATNA

REPORT OF INTERNSHIP

Submitted by :-

RAJESH KUMAR (15010105)

SANJAY BABU JAISWAL (15010109)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY MANIPUR

Supervisor :
Dr. Asif Ekbal
Associate Professor
Dept. of CSE
IIT PATNA

Debajyoty Banik
PhD Research Scholar
Dept. of CSE
IIT Patna

Abstract

We made a GUI which will do translation in 11x11 different indian languages like "English,Hindi,Bengali, Gujarati ,Tamil,Telugu,Marathi,Malayalam,Urdu,Konkani,Punjabi .

We present collection of 110 Statistical Machine Translation systems which is made from parallel corpora of 11 Indian Languages which belongs to Indo-Aryan and Dravidian families. We made analysis of relationship between translation accuracy and language families involved in the system. We think that results obtained from this analysis will provide guidelines for creating machine translation system for specific Indian language pairs. For our studies , we built phrase based systems and with some extensions. Across more than one languages , we show improvements on the baseline phrase based systems using these extensions : (1) source side reordering for English-Indian language translation, (2) transliteration of untranslated words for Indian language-Indian language translation.

Keywords -

SMT(Statistical Machine Translation)
Indian Language Machine Translation
Phrase Based Statistical Machine Translation

Declaration

I declare that this submission represents my idea in my own words and where others' idea or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/sources in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from proper permission has not been taken when needed.

Date: *(Signature)*

(RAJESH KUMAR)
(IIIT MANIPUR)

(SANJAY BABU JAISWAL)
(IIIT MANIPUR)

Acknowledgement

Dr. Asif Ekbal

Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Patna

Debajyoty Banik

PhD Research Scholar
Department of Computer Science and Engineering
Indian Institute of Technology Patna

- Rajesh Kumar
- Sanjay Babu Jaiswal

Contents

List of Tables

List of Figures

List of abbreviations

I

ILCI	Indian Language Corpora Initiative

I

IL	Indian Language

P

PBSMT	Phrase Based SMT

S

SMT	Statistical Machine Language

S

SOV	Subject Object Verb

Chapter 1

Introduction

A QUOTATION RELATED TO YOUR PROBLEM STATEMENT

"······"Reading a poem in translation," wrote Bialek, "is like kissing a woman through a veil"; and reading Greek poems, with a mixture of katharevousa and the demotic, is like kissing two women. Translation is a kind of transubstantiation; one poem becomes another. You choose your philosophy of translation just as you choose how to live: the free adaptation that sacrifices detail to meaning, the strict crib that sacrifices meaning to exactitude. The poet moves from life to language, the translator moves from language to life; both like the immigrant, try to identify the invisible, what's between the lines, the mysterious implications."

– Anne Michaels, Fugitive Pieces ······

1.1

In India we have different types of language families like Indo-Aryan ,Dravidian, Tibeto-Burman and Austro-Asiatic. After having this diversity in languages ,they also have some common features like (1) relatively free word order, with SOV being the common word order, (2) similar orthographic systems descended from the Brahmi script based on listening sense of phonetic principles, (3) vocabulary and grammatical tradition derived from Sanskrit, and (4) morphological richness. Due to this diversity of languages ,there is need of translation from one language to another language for government, business and social requirements. For this task we have Statistical Machine Translation (SMT) technology which provides the possibilty of covering a large number of translation pairs efficiently. For Indian languags ,the SMT approach has been used for some languages like English and Hindi ,along with other major languages like Urdu, Telugu, Bengali.

1.2

In this work, we build phrase-based SMT systems and their extensions for 110 language pairs using the Indian Language Corpora Initiative (ILCI) corpus. Our main aim is to achieve the following research directions: (1). Observe patterns between translation accuarcy and languages invloved. Do the patterns suggest that unique SMT system architecture be created for each language pair?

(2). Investigate ways of using shared charactericts of Indian languages to reduce the effort and resources required for building systems for other Indian language pairs.

(3). Find if learnings form improvements of SMT systems in one language pair can be easily transfered to other language pairs, making simultaneous progress of all Indian language SMT systems viable .

(4). Explore how far phrase based SMT systems for Indian languages can be improved through preprocessing and post-editing extensions.

(5). Identify the challenges for SMT involving all major Indian languages.

(6). Find out best principles to build SMT systems for specific language pairs.

Chapter 2

Existing System Study

2.1 Different SMT Models Explored for various language pairs

2.1. Baseline phrase based system (S1) : Phrase based SMT (PBSMT) systems have been developed for for many language pairs and are easily extensible to new language pairs sinnce they don't need linguistic resources. We study the performance of PBSMT for the relationship between translation accuracy and language families from the point of view of language divergences like word order and morphology and the effect of corpus size.

2.2. English -Indian Languages(IL) PBSMT with generic source side reordering rules(S2) : Preprocessing the training and test corpus by reordering the source side sentences to make them conform to target word order has been shown to be useful. The improvement occurs for two reasons: (1). The decoder's search space can consider candiadates with better word order. (2). The quality of the phrase table created is better since the alignment template method for phrase can match longer phrases. In addition , there are rules like prepositions becoming postpositions. This principle holds across all Indian languages , hence we hypothesize that the rules will benefit translation from English to any Indian language .

2.3. English-IL PBSMT with Hindi-tuned source side reodering rules(S3) : These rules are refinements of S2 with additional rules found through a focused analysis of word order divergence observed in the English -Hindi translation pair. These include rules for handling interrogative sentences, infinite clauses, adjectival and adverbial phrases.

2.4. IL-IL PBSMT with post-editing using translitration(S4): There are many words spoken in different geographical region which are untranslated. So after translation there are many words which do not get translated these are called named entites found in particular region.So these words get transliterated in later stage by post editing transliteration.

Chapter 3

System Analysis, Design & Implementation

3.1 Data set and Resources

To build the translation system we need a large data set for 11 languages to make 110 machine translation systems using corpus which contain 50000 parallel sentences. Under 11 languages we have 7 from the Indo-Aryan family which contains Hindi ,Urdu ,Konkani,Gujarati,Bengali,Marathi and Punjabi and 3 from Dravidian family Malayalam,Tamil,Telugu and last one from the West Germanic family as English. These sentences are collected from tourism and health field of these languages. These corpus is divided into three parts as training set ,testing set and tuning set among the 11languages parallely .We use the Moses system for training the phrase-based system with the grow-diag-final model.For extracting phases ,we used heuristic model and msd-bidirectional-fe model for lexicalized reordering.

3.2 Unicode Normalization of Indic Scripts

There are some problems with the text written in Indic sripts that they face multiple Unicode codepoints for representing the same script character.This problem happens because of need for compatibilty with other standards and use of some control characters for supplying rendering information. If we represent the same character many times then it creates data sparsity. So ,there is need of conversion of corpus to a canonical Unicode representation with the help of Indic Unicode Normalizer.

Chapter 4

Conclusion

We can separate the language based on translation accuracy into different language pairs.These separation leads to language pairs with high accuracy like Indo-Aryan languages and low accuracy with Dravidian languages.We have seen on increasing the training corpus size will show good translation accuracy. In the case of Indo-Aryan language ,increase in corpus size will increase the translation accuracy but in the Dravidian language ,the increase in corpus size will not affect the accuracy too much.In Indian language SMT there are some common properties of Indian languages which make translation easy. Like same script of Indian languages make transliteration between them easier. The factors which affect the translation quality are rich morphology of Indian languages and word order divergence between English and Indian languages. So morphology should be considered in SMT model to make translation easier in morphological rich Indian languages. Source side reordering and post-editing transliteration can give significant improvement over baseline phrase-based system. For example, source side reordering is used for English to Indian language translation.

Appendix A

User manual

A.1 Description of the system

My application contains Two Combo Box in which a list of languages is listed inlcuding " English , Hindi , Bengali , Gujarati , Tamil , Telugu , Marathi , Malayalam , Urdu , Konkani , Punjabi " and Two Text Area in which one is for taking input text or source language and second Text Area is for targeted Language or in which we want to translate .Three buttons including one for exit , reset and third one for performing Translation based on input selected through two Combo Boxes . Simulataneously thirteen check boxes is provided in bottom to perform simultaneous translation .

A.2 Step to install my implemented system

Change the path in decoding-bengali-english.sh

1. First You need to set the path according to your choice in the "decoding-bengali-english.sh" according to your files and folders where you copied the gui-model folder .

2.Change the path in .ini file according where you have placed the main gui$_{m}odelfolder. like /smt/mosesdecoder/scripts/tokenizertokenizer.perl-len < \ /gui-model/bengali-english/test.en.text > \ /gui-model/bengali-englishtest.tok.en$

Change the path in moses.ini

1.PhraseDictionaryMemory name=TranslationModel0 num features=4 path=/home/rajesh/set-eng-to-hindi/working/train/model/phrase-table.gz input-factor=0 output-factor=0 LexicalReordering name=LexicalReordering0 num features=6 type=wbe-msd-bidirectional-fe-allff input factor=0 output-factor=0
path=/home/rajesh/set-eng-to-hindi/working/train/model/reordering-table.wbe-msd-bidirectional-fe.gz
Distortion IRSTLM name=LM0 factor=0 path=/home/rajesh/set-eng-to-hindi/lmcorpora-

train.lm.arpa.binary order=3

A . Now you can either copy .exe file or .jar file .
B . You can install it by double clicking if you are using window .
C . If you are using linux/Ubuntu operating system then you can type java -jar filename.jar after going to copied path of application/jar file from terminal .
D . Note For this application you can use any version of jdk above then 1.4.0

Appendix B

Screenshot and Description of the Implemented System

B.1 Main GUI Interface

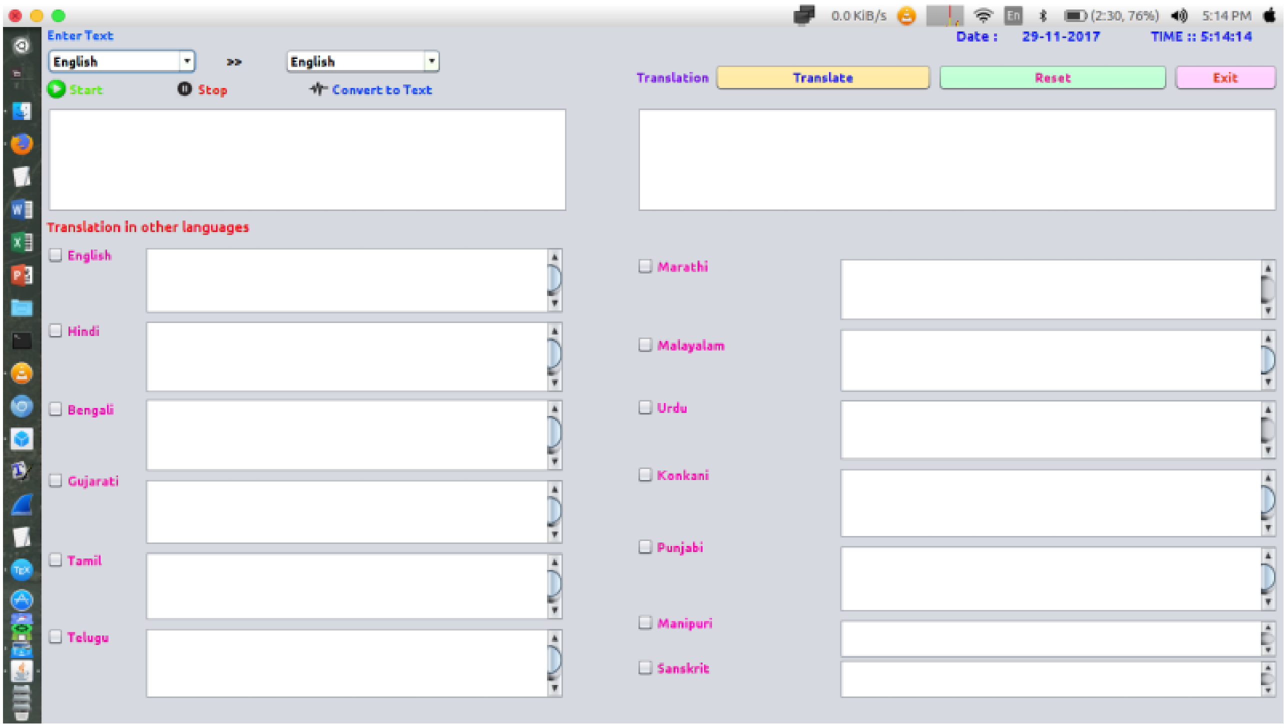
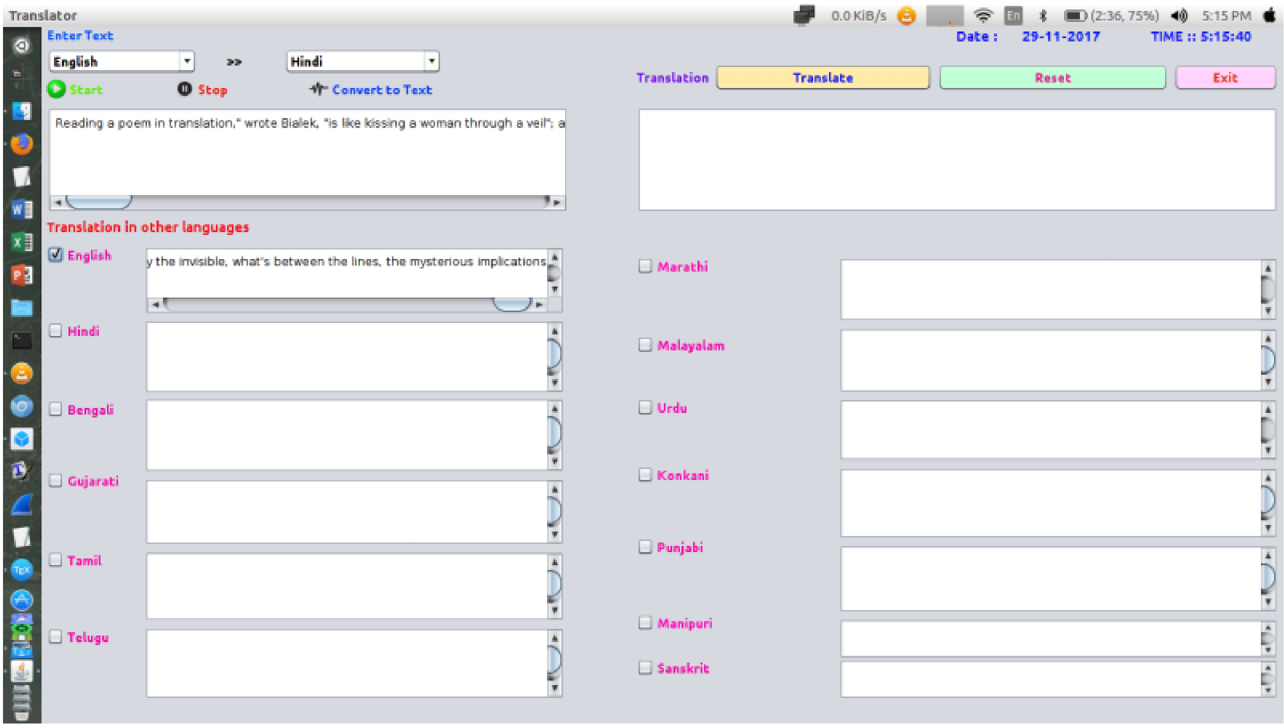


Figure B.1: Main GUI Interface

Description of Figure B.1This is the Basic GUI Interface Design in which 15 Text Area , 13 Check Box , 3 buttons , 2 Combo Box , 8 Labels are used . Check Boxes are used to choose the language in which we want to translate and corresponding Text Area is used to show Translated Language . Buttons with named as ”Translate” is used for Doing Translation , ”Reset” is used for Doing reset all Fields , Exit is used for Exiting or Terminating the application. ”Date” and ”TIME” Labels are used to show current date and time and remaining are used to guide user .

B.2 Basic Translation Testing Design



Date : 29-11-2017 TIME :: 5:15:40

Reading a poem in translation," wrote Bialek, "is like kissing a woman through a veil"; a

y the invisible, what's between the lines, the mysterious implications,

Figure B.2: English to English Testing

Description of Figure ??Testing for English to English.

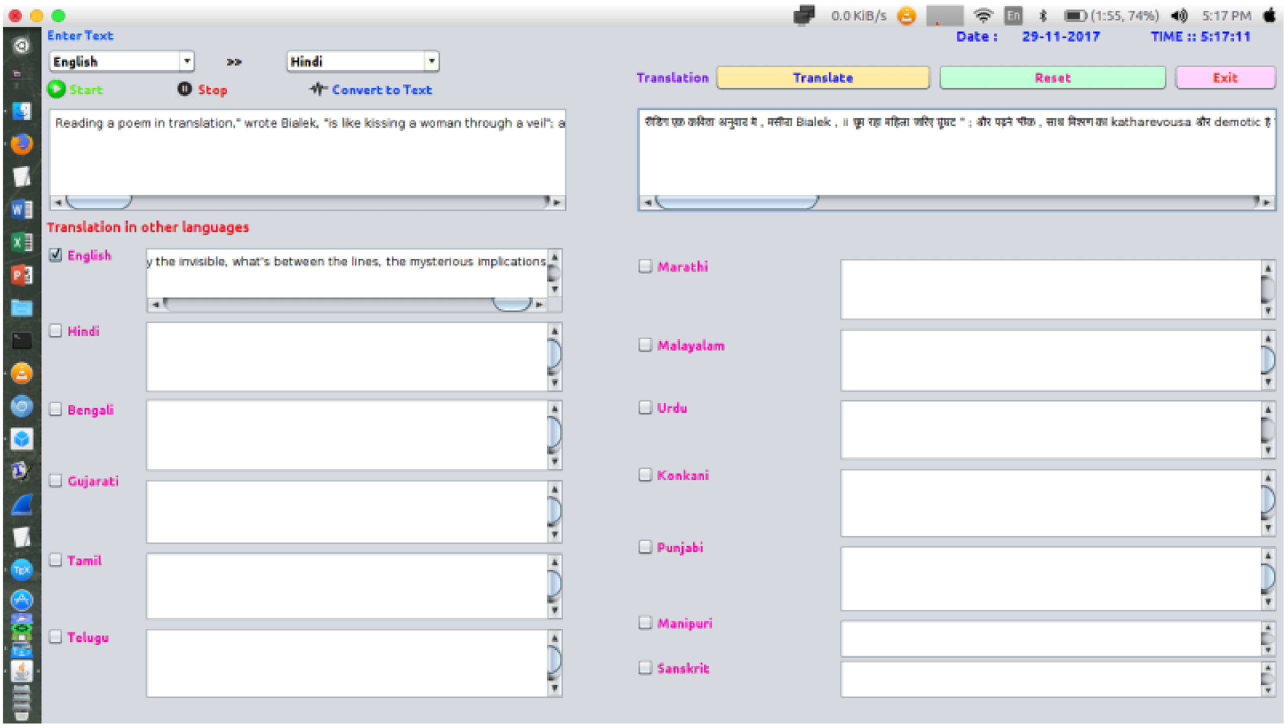
B.3 Basic Translation Model Including both Check Box and Text Area

Description of Figure ??Translation from English to Hindi using both Text Area and Check Box.

B.4 Translation Testing with Big Data

Description of Figure ??Translation Testing with Big Data and testing for time delay and results accuracy.

B.5 Translation Testing with Big Data including both Check box and Text Area



English >> Hindi

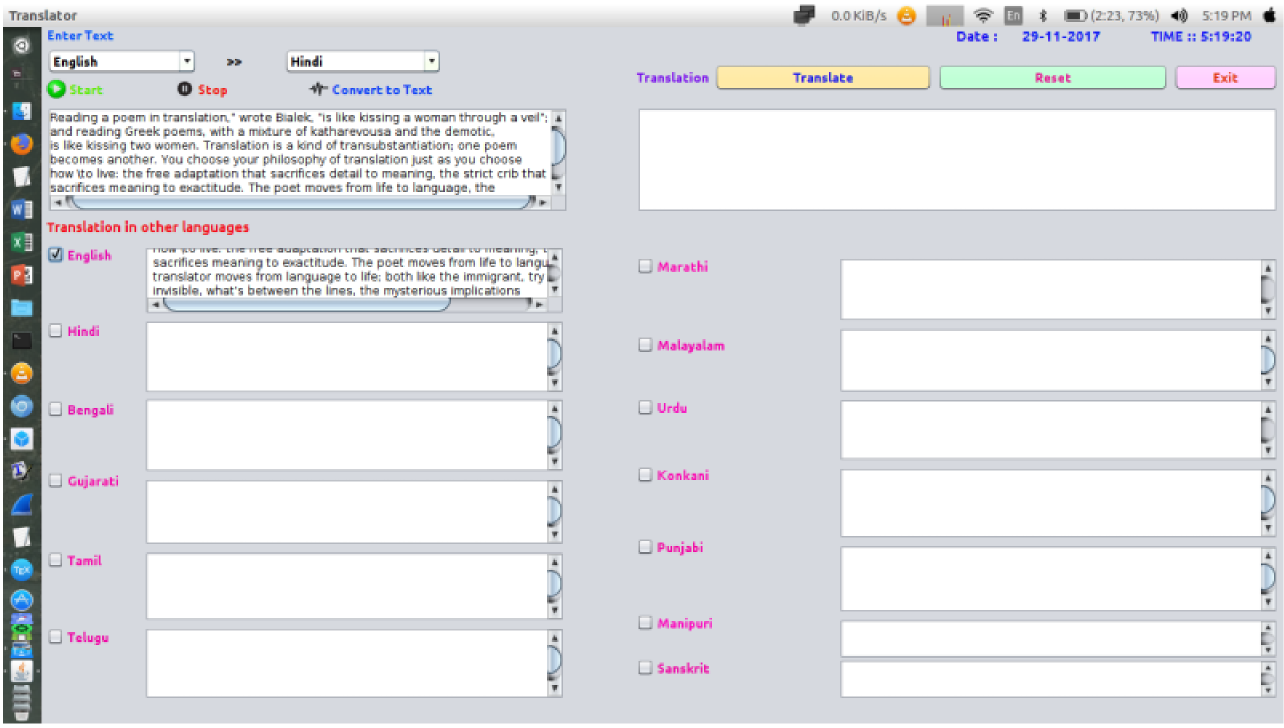
Reading a poem in translation," wrote Bialek, "is like kissing a woman through a veil"; a

रीडिंग एक कविता अनुवाद है , मसौदा Bialek , " चूम रहा महिला जरिए घूमट " ; और पढ़ने ग्रीक , साथ मिश्रण का katharevousa और demotic है

☑ English: y the invisible, what's between the lines, the mysterious implications

Date : 29-11-2017 TIME :: 5:17:11

Figure B.3: Testing for English to Hindi



English >> Hindi

Reading a poem in translation," wrote Bialek, "is like kissing a woman through a veil";
and reading Greek poems, with a mixture of katharevousa and the demotic,
is like kissing two women. Translation is a kind of transubstantiation; one poem
becomes another. You choose your philosophy of translation just as you choose
how \to live: the free adaptation that sacrifices detail to meaning, the strict crib that
sacrifices meaning to exactitude. The poet moves from life to language, the

☑ English:
how \to live: the free adaptation that sacrifices detail to meaning, t
sacrifices meaning to exactitude. The poet moves from life to langu
translator moves from language to life; both like the immigrant, try
invisible, what's between the lines, the mysterious implications

Date : 29-11-2017 TIME :: 5:19:20

Figure B.4: Translation Testing with Big Data

Bibliography

[1] Sata-anuvadak: Tackling multiway translation of indian languages