International Journal of Advances in Computer Networks and its Security

English to Hindi Statistical Machine Translation

Parteek Bhatia
Astt. Prof. :Computer Science and Engineering Department,
Thapar University
Thapar University: TU
Patiala, India.
parteek.bhatia@thapar.com

Nakul Sharma
Student, M.E. (S.E.): Computer Science and Engineering
Department, Thapar University,
Thapar University: TU.
Patiala, India.
nakul777@gmail.com

***Abstract*— Statistical Machine Translation (SMT) is uses the parallel, aligned corpus available in source and target languages. For undertaking efficient translations of high quality various issues need to addressed. The aim of this research paper is to undertake SMT using Moses tool-kit [1]. The English-Hindi translator is hence prepared. The quality of translation can be improved by increasing the size of corpus.**

***Keywords*— Statistical Machine Translation (SMT), Natural language Processing, (NLP)**

I. INTRODUCTION

Machine Translation (MT) entails training a computer for undertaking language conversion. There are various impediments for accurate translation.

English is one of the most utilised language in the world. Hindi is used by more than 400 million people [2, 3]. It is hence pertinent to remove the language barrier caused by lack of understanding of English language. The solution to this problem of language divide is given by Natural Language Processing (NLP). NLP is a field of research that involves analysing text, based upon the set of theories and set of technologies with the help of computers. Natural Language Processing is a theoretically motivated range of computational techniques, for analysing and representing naturally occurring texts at one or more levels of linguistic analysis, for the purpose of achieving human-like language processing for a range of tasks or applications [4, 5].

One problem is the inefficiency of the system by which translation is being performed. Another limitation is the complex nature of various natural languages. Statistical Machine Translation relies on use of parallel text for undertaking translation.

II. LITERATURE REVIEW

In their work, Knight Kevin and David Chiang , undertake Hiero hierarchical phrased based machine translation and syntax based machine translation for English to Chinese language pair. The algorithm utilized was Margin Infused Relaxed Algorithm of Crammer et al..They utilized BLEU score to evaluate the system [6] .

In their work,Antti-Veikko I. Rosti et.al. have given a combination of MT techniques in order to increase the accuracy of Machine Translation hence performed. Separately these systems do not provide high level of accuracy while undertaking translation. Their combination method operate on sentence, phrase and word level from N- best lists, system scores and target-to-source phrase alignments. Phrasal, hierarchical, syntax-based translation systems are combined [7].

Philip Koehn et al. explain in brief the various utilities got from Moses SMT toolkit. Various benefits got from Moses software include supporting linguistic dependent factoring. It also provide support for confusion network decoding and includes many efficient data formats for undertaking translations. They also provide a brief description of training and tuning the Moses SMT software [8].

Hao Xiong et al. describe creation of a ICT SMT system which in used to evaluate campaign of International Workshop on Spoken Language Translation 2010. Their experiments with the system had resulted in improvements with effective methods in several areas. This included refining data processing, reducing the number of Out-Of-Vocabulary (OOV) on the final test set, better inputs for the decoder, improving the overall performance of every decoder [9].

Durgesh Rao has elaborated on various types of Machine Translations systems mainly in regional languages of India. These include description of Anubharati, Anusaaraka, MaTra, Mantra, UCSG-based English-Kannada MT system, UNL based MT system between Marathi, English and Hindi. It also provides summary of major MT projects in India [3]

Goyal and G.S. Lehal describe a web based Hindi to Punjabi Machine Translation System. The system which have been developed makes use of lexicon based translation, transliteration and basic Word Sense Disambiguation. The

ACEE
Globalize The Research
Localize The World

system had provided high accuracy of 95%. This system's applicability in various areas is also mentioned [10].

III. MOTIVATION

Statistical machine translation is a part of Corpus based machine translation. Corpus Based Machine Translation Systems have advantages that, they are fully automatic and require significantly less human labor than traditional rule-based approaches. However, they require sentence-aligned parallel text for each language pair and cannot be used for language pairs for which such corpora does not exist. Corpus Based approach can be classified into Statistical and Example Based Machine Translation approaches.

Around 5000 lines parallel, aligned text of English-Hindi sentences were used to train the system. Moses decoder is an open-source software which is a phased-based SMT system. Phased-based SMT system enhances the performance of SMT system.

Development Environment

OS:Ubuntu 10.04 LTS.
RAM: 3GB.
HDD: 10GB.
Processor: C2D.

IV. STEPS

Following steps are necessary for undertaking SMT :-

1. Creation of Language Model (LM).
2. Invoking Translation Model (TM).
3. Calling decoder for final conversion.

1. In this step LM software is invoked. There are various software's such as SRILM, IRSRILM, CMU-Cam_Tookit. In our work we have used SRILM software. It is a product of SRI international and available for non-commercial purposes. SRILM is compatible with GIZA++. MGIZA++ Translation Model (TM) software. The format in which data is presented acceptable only by limited set of software. The installation entails downloading the binaries along with the source code and running the executable file. We had used executable ./ngram-count to undertake translation. The corpus is pre-processed before feeding it to LM software.
2. This step forms is now the part of step 3. this is because moses decoder embeds call to GIZA++ decoder while undertaking training.
3. The corpus along with language model and additional parameters is fed in to undertake training of Moses decoder. This step is essential before feeding into the system the training data.

Training Corpus & Moses.

Before undertaking training, Moses decoder must be properly installed onto operating system. Some additional scripts which are necessary for preparing corpus are downloadable from informatics home page server.
Preparing Data

- Tokenizing corpus:

It includes giving the parallel corpus as input to perl script, tokenizer.perl.

Command: zcat corpus_new4.en.gz |./tokenizer.perl -l en > corpusforRP/corpus_new4.tok.en
en

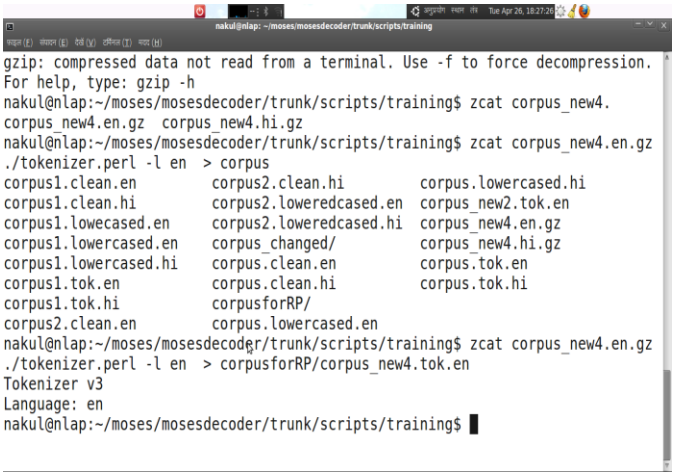

Figure 1 Tokenizing Corpus.

- cleaning corpus:

This step is necessary due to the fact that GIZA++ takes long time to train long sentences.

Command:./clean-corpus-n.perl corpusforRP/corpus_new4.tok en hi corpusforRP/corpus_new4.clean 1 40

ACEE
Globalize The Research
Localize The World

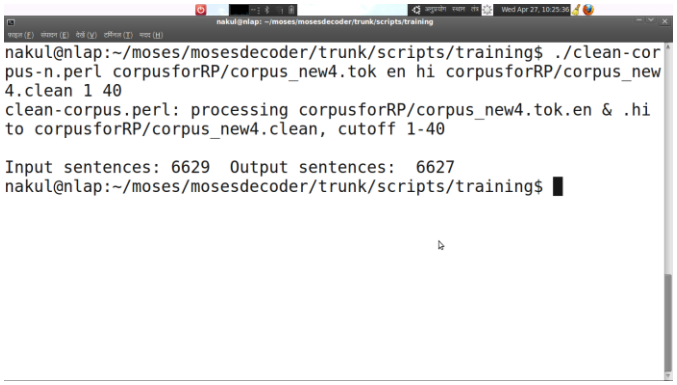

Figure 2 Cleaning Corpus

- Lower-casing corpus

It is performed with the intent of bringing in uniformity of case in corpus.

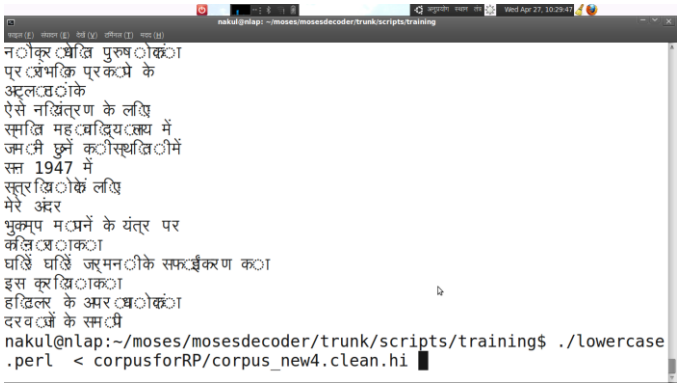

Figure 3 lowercasing Corpus

- LM creation:

This step entails creation of LM from executable ngram-count . It command and its screen-shot are as follows:-

Command:./ngram-count -order 3 -interpolate -kndiscount -unk -text corpus_new4.loweredcased.hi -lm hindi_lm.lm

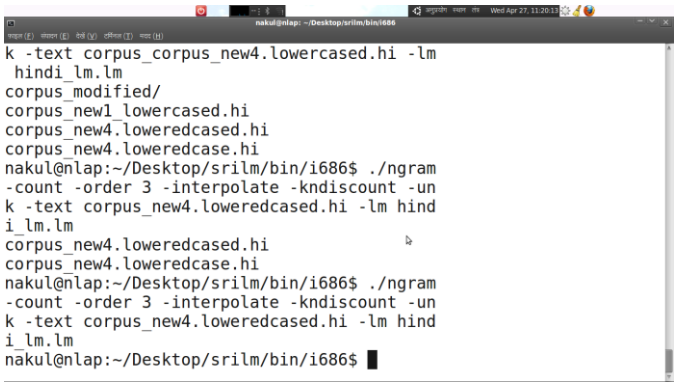

Figure 4 LM creation

- Training Moses software.

This entails use of perl script, train-factored-phrase-model to train the system. The command used is as follows:-

Command:./train-factored-phrase-model.perl -scripts-root-dir /home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/ -root-dir . -corpus corpusforRP/corpus_new4.lowercased -f en -e hi -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm 0:3:/home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/training/corpusforRP/hindi_lm.lm>& training_new4.out &

As the process executes in background, training_new4.out is updated and the last lines of this file indicate successful training with creation of moses.ini configuration file.

ACEE
Globalize The Research
Localize The World

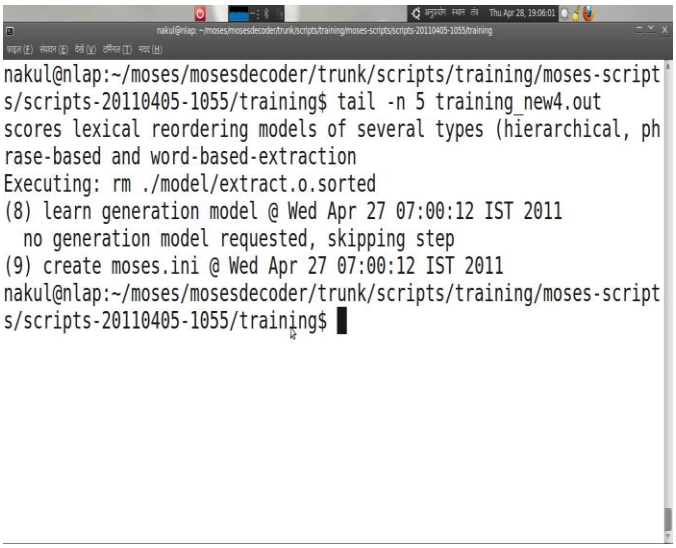

Figure 5 Training Moses software

- Running Moses software

The final execution of Moses software is done using following command:-

echo "good night" | TMP=/tmp /home/nakul/moses/mosesdecoder/trunk/moses-cmd/src/moses -f /home/nakul/moses/mosesdecoder/trunk/scripts/training/moses-scripts/scripts-20110405-1055/training/model/moses.ini -v 2

Herein moses.ini is created by undertaking training of the SMT system. For an input mentioned in quotes as "good night" the corresponding translation in Hindi language is given in the below figure:-

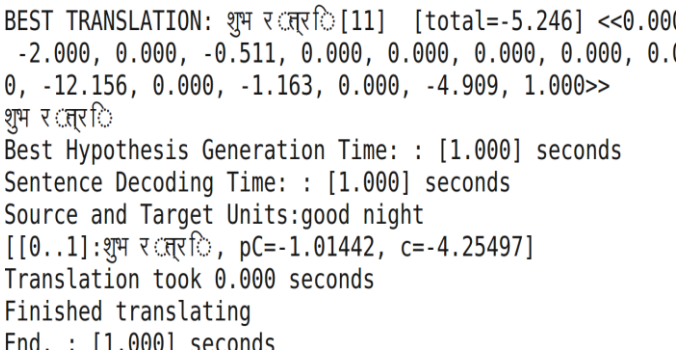

Figure 6 Sample Output

V. FUTURE WORK AND CONCLUSION

SMT can be applied in various domains such as Health-care industry for removing language gap. It requires less human effort. SMT can also help in reading documents which are available in selected languages provided it has been trained on a that set of language-pair. Since NLP is a emerging field, by combining various techniques of Machine Translation (MT), efficient high quality translations can be done.

SMT can also be applied in software engineering. This will allow documentations of various user-manuals and programs to be done in regional languages. Language bridge will be created henceforth and any gaps of understanding which hostage of languages will no longer exist.

REFERNCES

[1] http://statmt.org/moses.

[2] ISI ReWrite Decoder User's Manual, Version 0.2, available at *http://www.isi.edu/~germann/software/ReWrite-Decoder/isi-decoder-manual.html* accessed on 12/03/2010.

[3] Durgesh D Rao, "Machine Translation A Gentle Introduction", RESONANCE, July 1998.

[4] Hindi to Punjabi Translation system available at *http://h2p.learnpunjabi.org* accessed on 03/04/2010.

[5] Liddy, E. D., Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.

[6] Chiang D, Knight K, Wang W, "11,001 New Features for Statistical Machine Translation", Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 218–226, Boulder, Colorado, June 2009. 2009 Association for Computational Linguistics.

[7] Antti-Veikko I. Rosti, et al, "Combining Outputs from Multiple Machine Translation Systems", Proceedings of NAACL HLT 2007, pages 228–235, Rochester, NY, April 2007. 2007 Association for Computational Linguistics.

[8] Kohen P, et al, "Moses: Open Source Toolkit for Statistical Machine Translation." Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, Prague, June 2007. 2007 Association for Computational Linguistics.

[9] Xiong H, et al., "The ICT Statistical Machine Translation System for IWSLT 2010", Proceeding of the International Workshop on Spoken Language Translation 2010.

ACEE
Globalize The Research
Localize The World

[10] Goyal V., Lehal GS , "Web based hindi to punjabi machine translation system", Journal of Emerging Technologies in Web Intelligence, Vol 2, No 2, May 2010.

ACEE
Globalize The Research
Localize The World